Estimation and Inference in Modern Nonparametric Statistics

William George Underwood

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Operations Research and Financial Engineering
Adviser: Matias Damian Cattaneo

May 2024

Abstract

Nonparametric methods are central to modern statistics, enabling data analysis with minimal assumptions in a wide range of scenarios. While contemporary procedures such as random forests and kernel methods are popular due to their performance and flexibility, their statistical properties are often less well understood. The availability of sound inferential techniques is vital in the sciences, allowing researchers to quantify uncertainty in their models. We develop methodology for robust and practical statistical estimation and inference in some modern nonparametric settings involving complex estimators and nontraditional data.

We begin in the regression setting by studying the Mondrian random forest, a variant in which the partitions are drawn from a Mondrian process. We present a comprehensive analysis of the statistical properties of Mondrian random forests, including a central limit theorem for the estimated regression function and a characterization of the bias. We show how to conduct feasible and valid nonparametric inference by constructing confidence intervals, and further provide a debiasing procedure that enables minimax-optimal estimation rates for smooth function classes in arbitrary dimension.

Next, we turn our attention to nonparametric kernel density estimation with dependent dyadic network data. We present results for minimax-optimal estimation, including a novel lower bound for the dyadic uniform convergence rate, and develop methodology for uniform inference via confidence bands and counterfactual analysis. Our methods are based on strong approximations and are designed to be adaptive to potential dyadic degeneracy. We give empirical results with simulated and real-world economic trade data.

Finally, we develop some new probabilistic results with applications to nonparametric statistics. Coupling has become a popular approach for distributional analysis in recent years, and Yurinskii's method stands out for its wide applicability and explicit formulation. We present a generalization of Yurinskii's coupling, treating approximate martingale data under weaker conditions than previously imposed. We allow for Gaussian mixture coupling distributions, and a third-order method permits faster rates in certain situations. We showcase our results with applications to factor models and martingale empirical processes, as well as nonparametric partitioning-based and local polynomial regression procedures.

Acknowledgments

I am extremely fortunate to have been surrounded by many truly wonderful people over the course of my career, and without their support this dissertation would not have been possible. While it is impossible for me to identify every one of them individually, I would like to mention a few names in particular to recognize those who have been especially important to me during the last few years.

Firstly, I would like to express my utmost gratitude to my Ph.D. adviser, Matias Cattaneo. Working with Matias has been genuinely inspirational for me, and I could not have asked for a more rewarding start to my journey as a researcher. From the very beginning, he has guided me expertly through my studies, providing hands-on assistance when required while also allowing me the independence necessary to develop as an academic. I hope that, during the four years we have worked together, I have acquired just a fraction of his formidable mathematical intuition, keen attention to detail, boundless creativity, and inimitable pedagogical skill. Alongside his role as my adviser, Matias has been above all a friend, who has been in equal measure inspiring, insightful, dedicated, understanding, and kind.

Secondly, I would like to thank all of the faculty members at Princeton and beyond who have acted as my collaborators and mentors, without whom none of my work could have been realized. In particular, I express my gratitude to my tireless Ph.D. committee members and letter writers Jianqing Fan and Jason Klusowski, my coauthors Yingjie Feng and Ricardo Masini, my dissertation reader Boris Hanin, my teachers Amir Ali Ahmadi, Ramon van Handel, Miklós Rácz, and Mykhaylo Shkolnikov, my colleagues Sanjeev Kulkarni and Rocío Titiunik, and my former supervisor Mihai Cucuringu. I am also thankful for the staff members at Princeton who have been perpetually helpful, and I would like to identify Kim Lupinacci in particular; her assistance in all things administrative has been invaluable.

I am grateful to my fellow graduate students in the ORFE department for their technical expertise and generosity with their time, and for making Sherrerd Hall such a vibrant and exciting space, especially Jose Avilez, Pier Beneventano, Ben Budway, Rajita Chandak, Abraar Chaudhry, Stefan Clarke, Giulia Crippa, Gökçe Dayanıklı, Nicolas Garcia, Felix Hoefer, Erica Lai, Jackie Lok, Maya Mutic, Dan Rigobon, Till Saenger, Rajiv Sambharya, Boris Shigida,

Igor Silin, Giang Truong, and Rae Yu. Our regular social events made a contribution to my well-being which is difficult to overstate. My thanks extend also to the students I taught, as well as to my group of senior thesis undergraduates, for their commitment, patience, and responsiveness.

More broadly, I would like to thank all of my friends, near and far, for their unfailing support and reliability, and for helping to create so many of my treasured memories. In particular, Ole Agersnap, James Ashford, Christian Baehr, Chris Bambic, Kevin Beeson, James Broadhead, Alex Cox, Reece Edmends, Robin Franklin, Greg Henderson, Bonnie Ko, Grace Matthews, Dan Mead, Ben Musachio, Jacob Neis, Monika Papayova, Will Pedrick, Oliver Philcox, Nandita Rao, Alex Rice, Edward Rowe, David Snyder, Titi Sodimu, Nikitas Tampakis, and Anita Zhang. Thank you to the Princeton Chapel Choir for being such a wonderful community of musicians and a source of close friends, and to our directors, Nicole Aldrich and Penna Rose, and organist Eric Plutz.

Lastly, yet most importantly, I want to thank my family for their unwavering support throughout my studies. My visits back home have been a source of joy throughout my long and often challenging Ph.D., and I cherish every moment I have spent with my parents, sister, grandparents, and extended family.

Contents

List of Figures and Tables

Chapter 1

Introduction

Nonparametric estimation procedures are at the heart of many contemporary theoretical and methodological topics within the fields of statistics, data science, and machine learning. Where classical parametric techniques impose specific distributional and structural assumptions when modeling statistical problems, nonparametric methods instead take a more flexible approach, typically positing only high-level restrictions such as moment conditions, independence criteria, and smoothness assumptions. Examples of such procedures abound in modern data science and machine learning, encompassing histograms, kernel estimators, smoothing splines, decision trees, nearest neighbor methods, random forests, neural networks, and many more.

The benefits of the nonparametric framework are clear: statistical procedures can be formulated in cases where the stringent assumptions of parametric models are untestable, demonstrably violated, or simply unreasonable. As a consequence, the resulting methods often inherit desirable robustness properties against various forms of misspecification or misuse. The class of problems that can be formulated is correspondingly larger: arbitrary distributions and relationships can be characterized and estimated in a principled manner.

Nonetheless, these attractive properties do come at a price. In particular, as its name suggests, the nonparametric approach forgoes the ability to reduce a complex statistical problem to that of estimating a fixed, finite number of parameters. Rather, nonparametric procedures typically involve making inferences about a growing number of parameters simultaneously, as witnessed in high-dimensional regimes, or even directly handling infinite-dimensional objects

such as entire regression or density functions. As a consequence, nonparametric estimators are usually less efficient than their correctly specified parametric counterparts, when they are available; rates of convergence tend to be slower, and confidence sets more conservative. Another challenge is that theoretical mathematical analyses of nonparametric estimators are often significantly more demanding than those required for low-dimensional parametric settings, necessitating tools from contemporary developments in high-dimensional concentration phenomena, coupling and strong approximation theory, empirical processes, mathematical optimization, and stochastic calculus.

In addition to providing accurate point estimates of unknown (possibly high-dimensional or infinite-dimensional) quantities of interest, modern nonparametric procedures are also expected to come equipped with methodologies for conducting statistical inference. The availability of such inferential techniques is paramount, with contemporary nonparametric methods forming a ubiquitous component of modern data science tool kits. Valid uncertainty quantification is essential for hypothesis testing, error bar construction, assessing statistical significance, and performing power analyses. Inference is a central concept in classical statistics, and despite the rapid recent development of theory for modern nonparametric estimators, their applicability to statistical inference is in certain cases rather less well studied; theoretically sound and practically implementable inference procedures are sometimes absent in the literature.

In any statistical modeling problem, the selection and application of an estimator must naturally be tailored to the available data. Today, much of the data produced and analyzed does not necessarily fit neatly into the classical framework of independent and identically distributed samples, and instead might consist of time series, stochastic processes, networks, or high-dimensional or functional data, to name just a few. Therefore, it is important to understand how nonparametric methods might be adapted to correctly handle these data types, maintaining fast estimation rates and valid techniques for statistical inference. The technical challenges associated with such an endeavor are non-trivial; many standard techniques are ineffective in the presence of dependent or infinite-dimensional data, for example. As such, the development of new mathematical results in probability theory plays an important role in the comprehensive treatment of nonparametric statistics with complex data.

Overview of the dissertation

This dissertation presents a selection of topics relating to nonparametric estimation and inference, and the associated technical mathematical tools.

Chapter 2, titled "Inference with Mondrian Random Forests," is based on the work of Cattaneo, Klusowski, and Underwood (2023). Random forests are popular ensembling-based methods for classification and regression, which are well known for their good performance, flexibility, robustness, and efficiency. The majority of random forest models share the following common framework for producing estimates of a classification or regression function using covariates and a response variable. Firstly, the covariate space is partitioned in some algorithmic manner, possibly using a source of external randomness. Secondly, a local estimator of the classification or regression function is fitted to the responses in each cell separately, yielding a tree estimator. Finally, this process is repeated with many different partitions, and the resulting tree estimators are averaged to produce a random forest.

Many different variants of random forests have been proposed in recent years, typically with the aim of improving their statistical or computational properties, or simplifying their construction in order to permit a more detailed theoretical analysis. One interesting such example is that of the Mondrian random forest, in which the underlying partitions (or trees) are constructed independently of the data. Naturally, this restriction rules out many classical random forest models, which exhibit a complex and data-dependent partitioning scheme. Instead, trees are sampled from a canonical stochastic process known as the Mondrian process, which endows the resulting tree and forest estimators with various agreeable features.

We study the estimation and inference properties of Mondrian random forests in the nonparametric regression setting. In particular, we establish a novel central limit theorem for the estimates made by a Mondrian random forest which, when combined with a characterization of the bias and a consistent variance estimator, allows one to perform asymptotically valid statistical inference, such as constructing confidence intervals, on the unknown regression

function. We also provide a debiasing procedure for Mondrian random forests, which allows them to achieve minimax-optimal estimation rates with Hölder smooth regression functions, for any smoothness parameter and in arbitrary dimension.

Chapter 3, titled "Dyadic Kernel Density Estimators," is based on the work of Cattaneo, Feng, and Underwood (2024). Network data plays an important role in statistics, econometrics, and many other data science disciplines, providing a natural framework for modeling relationships between units, be they people, financial institutions, proteins, or economic entities. Of prominent interest is the task of performing statistical estimation and inference with data sampled from the edges of such networks, known as dyadic data. The archetypal lack of independence between edges in a network renders many classical statistical tools unsuited for direct application. As such, researchers must appeal to techniques tailored to dyadic data in order to accurately capture the complex structure present in the network.

We focus on nonparametric estimation and inference with dyadic data, and in particular we seek methods that are robust in the sense that our results should hold uniformly across the support of the data. Such uniformity guarantees allow for statistical inference in a broader range of settings, including specification testing and distributional counterfactual analysis. We specifically consider the problem of uniformly estimating a dyadic density function, focusing on kernel estimators taking the form of dyadic empirical processes.

Our main contributions include the minimax-optimal uniform convergence rate of the dyadic kernel density estimator, along with strong approximation results for the associated standardized and Studentized t-processes. A consistent variance estimator enables the construction of feasible uniform confidence bands for the unknown density function. We showcase the broad applicability of our results by developing novel counterfactual density estimation and inference methodology for dyadic data, which can be used for causal inference and program evaluation. A crucial feature of dyadic distributions is that they may be "degenerate" at certain points in the support of the data, a property that makes our analysis somewhat delicate. Nonetheless, our methods for uniform inference remain robust to the potential presence of such points. For implementation purposes, we discuss inference procedures based on positive semi-definite covariance estimators, mean squared error optimal bandwidth selec-

tors, and robust bias correction. We illustrate the empirical performance of our methods in simulations and with real-world trade data, for which we make comparisons between observed and counterfactual trade distributions in different years. Our technical results on strong approximations and maximal inequalities are of potential independent interest.

Finally, Chapter 4, titled "Yurinskii's Coupling for Martingales," is based on the work of Cattaneo, Masini, and Underwood (2022). Yurinskii's coupling is a popular theoretical tool for non-asymptotic distributional analysis in mathematical statistics and applied probability. Coupling theory, also known as strong approximation, provides an alternative framework to the more classical weak convergence approach to statistical analysis. Rather than merely approximating the distribution of a random variable, strong approximation techniques construct a sequence of random variables which are close almost surely or in probability, often with finite-sample guarantees.

Coupling allows distributional analysis in settings where weak convergence fails, including many applications to nonparametric or high-dimensional statistics; it is a key technical component in the main strong approximation results of our Chapter 3. The Yurinskii method specifically offers a Gaussian coupling with an explicit error bound under easily verified conditions; originally stated in ℓ^2-norm for sums of independent random vectors, it has recently been extended both to the ℓ^p-norm, for $1 \leq p \leq \infty$, and to vector-valued martingales in ℓ^2-norm, under some strong conditions.

We present as our main result a Yurinskii coupling for approximate martingales in ℓ^p-norm, under substantially weaker conditions than previously imposed. Our formulation allows the coupling variable to follow a general Gaussian mixture distribution, and we provide a novel third-order coupling method which gives tighter approximations in certain situations. We specialize our main result to mixingales, martingales, and independent data, and derive uniform Gaussian mixture strong approximations for martingale empirical processes. Applications to nonparametric partitioning-based and local polynomial regression procedures are provided.

Supplementary materials for Chapters 2, 3, and 4 are provided in Appendices A, B, and C respectively. These contain detailed proofs of the main results, additional technical contributions, and further discussion.

Chapter 2

Inference with Mondrian Random Forests

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

2.1 Introduction

Random forests, first introduced by Breiman (2001), are a workhorse in modern machine learning for classification and regression tasks. Their desirable traits include computational efficiency (via parallelization and greedy heuristics) in big data settings, simplicity of configuration and amenability to tuning parameter selection, ability to adapt to latent structure

in high-dimensional data sets, and flexibility in handling mixed data types. Random forests have achieved great empirical successes in many fields of study, including healthcare, finance, online commerce, text analysis, bioinformatics, image classification, and ecology.

Since Breiman introduced random forests over twenty years ago, the study of their statistical properties remains an active area of research: see Scornet, Biau, and Vert (2015), Chi, Vossler, Fan, and Lv (2022), Klusowski and Tian (2024), and references therein, for a sample of recent developments. Many fundamental questions about Breiman's random forests remain unanswered, owing in part to the subtle ingredients present in the estimation procedure which make standard analytical tools ineffective. These technical difficulties stem from the way the constituent trees greedily partition the covariate space, utilizing both the covariate and response data. This creates complicated dependencies on the data which are often exceedingly hard to untangle without overly stringent assumptions, thereby hampering theoretical progress.

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

Yet another variant of random forests, Mondrian random forests (Lakshminarayanan, Roy, and Teh, 2014), have received significant attention in the statistics and machine learning communities in recent years (Ma, Ghojogh, Samad, Zheng, and Crowley, 2020; Mourtada, Gaïffas, and Scornet, 2020; Scillitoe, Seshadri, and Girolami, 2021; Mourtada, Gaïffas, and Scornet, 2021; Vicuna, Khannouz, Kiar, Chatelain, and Glatard, 2021; Gao, Xu, and Zhou,

2022; O'Reilly and Tran, 2022). Like other purely random forest variants, Mondrian random forests offer a simplified modification of Breiman's original proposal in which the partition is generated independently of the data and according to a canonical stochastic process known as the Mondrian process (Roy and Teh, 2008). The Mondrian process takes a single parameter $\lambda > 0$ known as the "lifetime" and enjoys various mathematical properties. These probabilistic features allow Mondrian random forests to be fitted in an online manner as well as being subject to a rigorous statistical analysis, while also retaining some of the appealing features of other more traditional random forest methods.

This chapter studies the statistical properties of Mondrian random forests. We focus on this purely random forest variant not only because of its importance in the development of random forest theory in general, but also because the Mondrian process is, to date, the only known recursive tree mechanism involving randomization, pure or data-dependent, for which the resulting random forest is minimax-optimal for point estimation over a class of smooth regression functions in arbitrary dimension (Mourtada et al., 2020). In fact, when the covariate dimension exceeds one, the aforementioned centered and median random forests are both minimax-*suboptimal*, due to their large biases, over the class of Lipschitz smooth regression functions (Klusowski, 2021). It is therefore natural to focus our study of inference for random forests on versions that at the very least exhibit competitive bias and variance, as this will have important implications for the trade-off between precision and confidence.

Despite their recent popularity, relatively little is known about the formal statistical properties of Mondrian random forests. Focusing on nonparametric regression, Mourtada et al. (2020) recently showed that Mondrian forests containing just a single tree (called a Mondrian tree) can be minimax-optimal in integrated mean squared error whenever the regression function is β-Hölder continuous for some $\beta \in (0, 1]$. The authors also showed that, when appropriately tuned, large Mondrian random forests can be similarly minimax-optimal for $\beta \in (0, 2]$, while the constituent trees cannot. See also O'Reilly and Tran (2022) for analogous results for more general Mondrian tree and forest constructions. These results

formally demonstrate the value of ensembling with random forests from a point estimation perspective. No results are currently available in the literature for statistical inference using Mondrian random forests.

This chapter contributes to the literature on the foundational statistical properties of Mondrian random forest regression estimation with two main results. Firstly, we give a central limit theorem for the classical Mondrian random forest point estimator, and propose valid large-sample inference procedures employing a consistent standard error estimator. We establish this result by deploying a martingale central limit theorem (Hall and Heyde, 1980, Theorem 3.2) because we need to handle delicate probabilistic features of the Mondrian random forest estimator. In particular, we deal with the existence of Mondrian cells which are "too small" and lead to a reduced effective (local) sample size for some trees in the forest. Such pathological cells are in fact typical in Mondrian random forests and complicate the probability limits of certain sample averages; in fact, small Mondrian random forests (or indeed single Mondrian trees) remain random even in the limit due to the lack of ensembling. The presence of small cells renders inapplicable prior distributional approximation results for partitioning-based estimators in the literature (Huang, 2003; Cattaneo, Farrell, and Feng, 2020), since the commonly required quasi-uniformity assumption on the underlying partitioning scheme is violated by cells generated using the Mondrian process. We circumvent this technical challenge by establishing new theoretical results for Mondrian partitions and their associated Mondrian trees and forests, which may be of independent interest.

The second main contribution of the chapter is to propose a debiasing approach for the Mondrian random forest point estimator. We accomplish this by first precisely characterizing the probability limit of the large sample conditional bias, and then applying a debiasing procedure based on the generalized jackknife (Schucany and Sommers, 1977). We thus exhibit a Mondrian random forest variant which is minimax-optimal in pointwise mean squared error when the regression function is β-Hölder for any $\beta > 0$. Our method works by generating an ensemble of Mondrian random forests carefully chosen to have smaller misspecification bias when extra smoothness is available, resulting in minimax optimality even for $\beta > 2$. This result complements Mourtada et al. (2020) by demonstrating the existence of a class of

Mondrian random forests that can efficiently exploit the additional smoothness of the unknown regression function for minimax-optimal point estimation. Our proposed debiasing procedure is also useful when conducting statistical inference because it provides a principled method for ensuring that the bias is negligible relative to the standard deviation of the estimator. More specifically, we use our debiasing approach to construct valid inference procedures based on robust bias correction (Calonico, Cattaneo, and Farrell, 2018, 2022).

This chapter is structured as follows. In Section 2.2 we introduce the Mondrian process and give our assumptions on the data generating process, using a Hölder smoothness condition on the regression function to control the bias of various estimators. We define the Mondrian random forest estimator and present our assumptions on its lifetime parameter and the number of trees. We give our notation for the following sections in this chapter.

Section 2.3 presents our first set of main results, beginning with a central limit theorem for the centered Mondrian random forest estimator (Theorem 2.3.1), in which we characterize the limiting variance. Theorem 2.3.2 complements this result by precisely calculating the limiting bias of the estimator, with the aim of subsequently applying a debiasing procedure. To enable valid feasible statistical inference, we provide a consistent variance estimator in Theorem 2.3.3 and briefly discuss implications for lifetime parameter selection.

In Section 2.4 we provide a brief overview of the proofs of these first main results. We focus on the technical innovations and general strategic approach, giving some insight into the challenges involved, and refer the reader to Section A.2 for detailed proofs.

In Section 2.5 we define debiased Mondrian random forests, a collection of estimators based on linear combinations of Mondrian random forests with varying lifetime parameters. These parameters are carefully chosen to annihilate leading terms in our bias characterization, yielding an estimator with provably superior bias properties (Theorem 2.5.2). In Theorem 2.5.1 we verify that a central limit theorem continues to hold for the debiased Mondrian random forest. We again state the limiting variance, discuss the implications for the lifetime parameter, and provide a consistent variance estimator (Theorem 2.5.3) for constructing confidence intervals

(Theorem 2.5.4). As a final corollary of the improved bias properties, we demonstrate in Theorem 2.5.5 that the debiased Mondrian random forest estimator is minimax-optimal in pointwise mean squared error for all $\beta > 0$, provided that β is known a priori.

Section 2.6 discusses tuning parameter selection, beginning with a data-driven approach to selecting the crucial lifetime parameter using polynomial estimation, alongside other practical suggestions including generalized cross-validation. We also give advice on choosing the number of trees, and other parameters associated with the debiasing procedure.

In Section 2.7 we present an illustrative example application of our proposed methodology for estimation and inference in the setting of weather forecasting in Australia. We demonstrate the use of our debiased Mondrian random forest estimator and our generalized cross-validation procedure for lifetime parameter selection, as well as the construction of point estimates and confidence intervals.

Concluding remarks are given in Section 2.8, while Appendix A contains all the mathematical proofs of our theoretical contributions, along with some other technical probabilistic results on the Mondrian process which may be of interest.

2.1.1 Notation

We write $\|\cdot\|_2$ for the usual Euclidean ℓ^2-norm on $\mathbb{R}^d$. The natural numbers are $\mathbb{N} = \{0, 1, 2, \ldots\}$. We use $a \wedge b$ for the minimum and $a \vee b$ for the maximum of two real numbers. For a set A, we use A^{c} for the complement whenever the background space is clear from context. We use C to denote a positive constant whose value may change from line to line. For non-negative sequences a_n and b_n, write $a_n \lesssim b_n$ or $a_n = O(b_n)$ to indicate that a_n/b_n is bounded for $n \geq 1$. Write $a_n \ll b_n$ or $a_n = o(b_n)$ if $a_n/b_n \to 0$. If $a_n \lesssim b_n \lesssim a_n$, write $a_n \asymp b_n$. For random non-negative sequences A_n and B_n, similarly write $A_n \lesssim_{\mathbb{P}} B_n$ or $A_n = O_{\mathbb{P}}(B_n)$ if A_n/B_n is bounded in probability, and $A_n = o_{\mathbb{P}}(B_n)$ if $A_n/B_n \to 0$ in probability. Convergence of random variables X_n in distribution to a law $\mathbb{P}$ is denoted by $X_n \rightsquigarrow \mathbb{P}$.

2.2 Setup

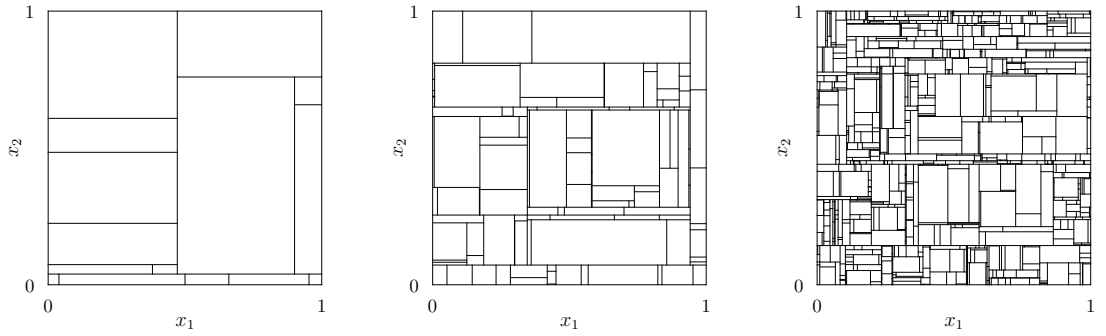
When using a Mondrian random forest, there are two sources of randomness. The first is of course the data, and here we consider the nonparametric regression setting with d-dimensional covariates. The second source is a collection of independent trees drawn from a Mondrian process, which we define in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

2.2.1 The Mondrian process

The Mondrian process was introduced by Roy and Teh (2008) and offers a canonical method for generating random rectangular partitions, which can be used as the trees for a random forest (Lakshminarayanan et al., 2014; Lakshminarayanan, Roy, and Teh, 2016). For the reader's convenience, we give a brief description of this process here; see Mourtada et al. (2020, Section 3) for a more complete definition.

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

To sample a partition T from the Mondrian process $\mathcal{M}\big([0,1]^d, \lambda\big)$ we start at time $t = 0$ with the trivial partition of $[0,1]^d$ which has no splits. We then repeatedly apply the following procedure to each cell D in the partition. Let t_D be the time at which the cell was formed, and sample $E_D \sim \mathrm{Exp}\left(|D|_1\right)$. If $t_D + E_D \leq \lambda$, then we split D. This is done by first selecting a split dimension J with $\mathbb{P}(J = j) = |D_j|/|D|_1$, and then sampling a split location $S_J \sim \mathrm{Unif}\left[D_J^-, D_J^+\right]$. The cell D splits into the two new cells $\{x \in D : x_J \leq S_J\}$ and $\{x \in D : x_J > S_J\}$, each with formation time $t_D + E_D$. The final outcome is the partition T consisting of the cells D which were not split because $t_D + E_D > \lambda$. The cell in T containing a point $x \in [0,1]^d$ is written $T(x)$. Figure 2.1 shows typical realizations of $T \sim \mathcal{M}\big([0,1]^d, \lambda\big)$ for $d = 2$ and with different lifetime parameters λ.



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

Figure 2.1: The Mondrian process $T \sim \mathcal{M}\big([0,1]^d, \lambda\big)$ with $d = 2$ and lifetime parameters λ.

2.2.2 Data generation

Throughout this chapter, we assume that the data satisfies Assumption 2.2.1. We begin with a definition of Hölder continuity which will be used for controlling the bias of various estimators.

Definition 2.2.1 (Hölder continuity)

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

Assumption 2.2.1 (Data generation)

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

Some comments are in order surrounding Assumption 2.2.1. The requirement that the covariate density $f(x)$ be strictly positive on all of $[0,1]^d$ may seem strong, particularly when d is moderately large. However, since our theory is presented pointwise in x, it is sufficient for

this to hold only on some neighborhood of x. To see this, note that continuity implies the density is positive on some hypercube containing x. Upon rescaling the covariates, we can map this hypercube onto $[0,1]^d$. The same argument of course holds for the Hölder smoothness assumptions and the upper and lower bounds on the conditional variance function.

2.2.3 Mondrian random forests

We define the basic Mondrian random forest estimator (2.1) as in Lakshminarayanan et al. (2014) and Mourtada et al. (2020), and will later extend it to a debiased version in Section 2.5. For a lifetime parameter $\lambda > 0$ and forest size $B \geq 1$, let $\mathbf{T} = (T_1, \ldots, T_B)$ be a Mondrian forest where $T_b \sim \mathcal{M}\big([0,1]^d, \lambda\big)$ are i.i.d. Mondrian trees which are independent of the data. For $x \in [0,1]^d$, write $N_b(x) = \sum_{i=1}^n \mathbb{I}\left\{X_i \in T_b(x)\right\}$ for the number of samples in $T_b(x)$, with $\mathbb{I}$ denoting an indicator function. Then the Mondrian random forest estimator of $\mu(x)$ is

$$\hat{\mu}(x) = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{i=1}^n Y_i\, \mathbb{I}\big\{X_i \in T_b(x)\big\}}{N_b(x)}. \tag{2.1}$$

If there are no samples X_i in $T_b(x)$ then $N_b(x) = 0$, so we define $0/0 = 0$ (see Section A.2 for details). To ensure the bias and variance of the Mondrian random forest estimator converge to zero (see Section 2.3), and to avoid boundary issues, we impose some basic conditions on x, λ, and B in Assumption 2.2.2.

Assumption 2.2.2 (Mondrian random forest estimator)
Suppose $x \in (0,1)^d$ is an interior point of the support of X_i, $\frac{\lambda^d}{n} \to 0$, $\log \lambda \asymp \log n$, and $B \asymp n^\xi$ for some $\xi \in (0,1)$, which may depend on the dimension d and smoothness β.

Assumption 2.2.2 implies that the size of the forest B grows with n. For the purpose of mitigating the computational burden, we suggest the sub-linear polynomial growth $B \asymp n^\xi$, satisfying the conditions imposed in our main results. Large forests usually do not present computational challenges in practice as the ensemble estimator is easily parallelizable over the trees. We emphasize places where this "large forest" condition is important to our theory as they arise throughout the chapter.

2.3 Inference with Mondrian random forests

We begin with a bias–variance decomposition for the Mondrian random forest estimator:

$$\begin{aligned}\hat{\mu}(x) - \mu(x) &= \Big(\hat{\mu}(x) - \mathbb{E}\big[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\big]\Big) + \Big(\mathbb{E}\big[\hat{\mu}(x) \mid \mathbf{X}, \mathbf{T}\big] - \mu(x)\Big) \\ &= \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{i=1}^{n} \varepsilon_i \, \mathbb{I}\big\{X_i \in T_b(x)\big\}}{N_b(x)} \\ &\quad + \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{i=1}^{n} \big(\mu(X_i) - \mu(x)\big) \, \mathbb{I}\big\{X_i \in T_b(x)\big\}}{N_b(x)}. \end{aligned} \tag{2.2}$$

Our approach to inference is summarized as follows. Firstly, we provide a central limit theorem (weak convergence to a Gaussian) for the first "variance" term in (2.2). Secondly, we precisely compute the probability limit of the second "bias" term. By ensuring that the standard deviation dominates the bias, a corresponding central limit theorem holds for the Mondrian random forest. With an appropriate estimator for the limiting variance, we establish procedures for valid and feasible statistical inference on the unknown regression function $\mu(x)$.

We begin with the aforementioned central limit theorem, which forms the core of our methodology for performing statistical inference. Before stating our main result, we highlight some of the challenges involved. At first glance, the summands in the first term in (2.2) seem to be independent over $1 \leq i \leq n$, conditional on the forest $\mathbf{T}$, depending only on X_i and ε_i. However, the $N_b(x)$ appearing in the denominator depends on all X_i simultaneously, violating this independence assumption and rendering classical central limit theorems inapplicable. A natural preliminary attempt to resolve this issue is to observe that

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

with high probability. One could attempt to use this by approximating the estimator with an average of i.i.d. random variables, or by employing a central limit theorem conditional on $\mathbf{X}$ and $\mathbf{T}$. However, such an approach fails because $\mathbb{E}\left[\frac{1}{|T_b(x)|^2}\right] = \infty$; the possible existence of

small cells causes the law of the inverse cell volume to have heavy tails. For similar reasons, attempts to directly establish a central limit theorem based on $2+\delta$ moments, such as the Lyapunov central limit theorem, are ineffective.

We circumvent these problems by directly analyzing $\frac{\mathbb{I}\{N_b(x)\geq 1\}}{N_b(x)}$. We establish concentration properties for this non-linear function of X_i via the Efron–Stein inequality (Boucheron, Lugosi, and Massart, 2013, Section 3.1) along with a sequence of somewhat delicate preliminary lemmas regarding inverse moments of truncated (conditional) binomial random variables. In particular, we show that $\mathbb{E}\left[\frac{\mathbb{I}\{N_b(x)\geq 1\}}{N_b(x)}\right] \lesssim \frac{\lambda^d}{n}$ and $\mathbb{E}\left[\frac{\mathbb{I}\{N_b(x)\geq 1\}}{N_b(x)^2}\right] \lesssim \frac{\lambda^{2d}\log n}{n^2}$. Asymptotic normality is then established using a central limit theorem for martingale difference sequences (Hall and Heyde, 1980, Theorem 3.2) with respect to an appropriate filtration. Section 2.4 gives an overview our proof strategy in which we further discuss the underlying challenges, while Section A.2 gives all the technical details.

2.3.1 Central limit theorem

Theorem 2.3.1 gives our first main result.

Theorem 2.3.1 (Central limit theorem for the centered Mondrian random forest estimator) *Suppose Assumptions 2.2.1 and 2.2.2 hold, and further assume that $\mathbb{E}[Y_i^4 \mid X_i]$ is bounded almost surely and $\frac{\lambda^d \log n}{n} \to 0$. Then*

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

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

We take this opportunity to contrast Mondrian random forests with more classical kernel-based smoothing methods. The lifetime λ plays a similar role to the inverse bandwidth in determining the effective sample size n/λ^d, and thus the associated rate of convergence.

However, due to the Mondrian process construction, some cells are typically "too small" (equivalent to an insufficiently large bandwidth) to give an appropriate effective sample size. Similarly, classical methods based on non-random partitioning such as spline estimators (Huang, 2003; Cattaneo et al., 2020) typically impose a quasi-uniformity assumption to ensure all the cells are of comparable size, a property which does not hold for the Mondrian process (not even with probability approaching one).

Bias characterization

We turn to the second term in (2.2), which captures the bias of the Mondrian random forest estimator conditional on the covariates $\mathbf{X}$ and the forest $\mathbf{T}$. As such, it is a random quantity which, as we will demonstrate, converges in probability. We precisely characterize the limiting non-random bias, including high-degree polynomials in λ which for now may seem ignorable. Indeed the magnitude of the bias is determined by its leading term, typically of order $1/\lambda^2$ whenever $\beta \geq 2$, and this suffices for ensuring a negligible contribution from the bias with an appropriate choice of lifetime parameter. However, the advantage of specifying higher-order bias terms is made apparent in Section 2.5 when we construct a debiased Mondrian random forest estimator. There, we target and annihilate the higher-order terms in order to furnish superior estimation and inference properties. Theorem 2.3.2 gives our main result on the bias of the Mondrian random forest estimator.

Theorem 2.3.2 (Bias of the Mondrian random forest estimator)
Suppose Assumptions 2.2.1 and 2.2.2 hold. Then for each $1 \leq r \leq \lfloor \beta/2 \rfloor$ *there exists* $B_r(x) \in \mathbb{R}$*, which is a function only of the derivatives of* f *and* μ *at* x *up to order* $2r$*, with*

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

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

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

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

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

In Theorem 2.3.2 we give some explicit examples of calculating the limiting bias if $\beta > 2$ or when X_i are uniformly distributed. The general form of $B_r(x)$ is provided in Section A.2 but is somewhat unwieldy except in specific situations. Nonetheless the most important properties are that $B_r(x)$ are non-random and do not depend on the lifetime λ, crucial facts for our debiasing procedure given in Section 2.5. If the forest size B does not diverge to infinity then we suffer the first-order bias term $\frac{1}{\lambda\sqrt{B}}$. This phenomenon was explained by Mourtada et al. (2020), who noted that it allows single Mondrian trees to achieve minimax optimality only when $\beta \in (0, 1]$. Large forests remove this first-order bias and are optimal for all $\beta \in (0, 2]$.

Using Theorem 2.3.1 and Theorem 2.3.2 together, along with an appropriate choice of lifetime parameter λ, gives a central limit theorem for the Mondrian random forest estimator which can be used, for example, to build confidence intervals for the unknown regression function $\mu(x)$ whenever the bias shrinks faster than the standard deviation. In general this will require $\frac{1}{\lambda^2} + \frac{1}{\lambda^\beta} + \frac{1}{\lambda\sqrt{B}} \ll \sqrt{\frac{\lambda^d}{n}}$, which can be satisfied by imposing the restrictions $\lambda \gg n^{\frac{1}{d+2(2\wedge\beta)}}$ and $B \gg n^{\frac{2(2\wedge\beta)-2}{d+2(2\wedge\beta)}}$ on the lifetime λ and forest size B. If instead we aim for optimal point estimation, then balancing the bias and standard deviation requires $\frac{1}{\lambda^2} + \frac{1}{\lambda^\beta} + \frac{1}{\lambda\sqrt{B}} \asymp \sqrt{\frac{\lambda^d}{n}}$, which can be satisfied by $\lambda \asymp n^{\frac{1}{d+2(2\wedge\beta)}}$ and $B \gtrsim n^{\frac{2(2\wedge\beta)-2}{d+2(2\wedge\beta)}}$. Such a choice of λ gives the convergence rate $n^{\frac{-(2\wedge\beta)}{d+2(2\wedge\beta)}}$ which is the minimax-optimal rate of convergence (Stone, 1982) for β-Hölder functions with $\beta \in (0, 2]$ as shown by Mourtada et al. (2020, Theorem 2). In Section 2.5 we will show how the Mondrian random forest estimator can be debiased, giving both weaker lifetime conditions for inference and also improved rates of convergence, under additional smoothness assumptions.

Variance estimation

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

$$\hat{\sigma}^2(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}\frac{\big(Y_i - \hat{\mu}(x)\big)^2\,\mathbb{I}\{X_i \in T_b(x)\}}{N_b(x)}, \tag{2.3}$$

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

In Theorem 2.3.3 we show that this estimator is consistent, and establish its rate of convergence.

Theorem 2.3.3 (Variance estimation)

Grant Assumptions 2.2.1 and 2.2.2, and suppose $\mathbb{E}[Y_i^4 \mid X_i]$ is bounded almost surely. Then

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

2.3.2 Confidence intervals

Theorem 2.3.4 shows how to construct valid confidence intervals for the regression function $\mu(x)$ under the lifetime and forest size assumptions previously discussed. For details on feasible and practical selection of the lifetime parameter λ, see Section 2.6.

Theorem 2.3.4 (Feasible confidence intervals using a Mondrian random forest)

Suppose that Assumptions 2.2.1 and 2.2.2 hold, $\mathbb{E}[Y_i^4 \mid X_i]$ is bounded almost surely, and $\frac{\lambda^d \log n}{n} \to 0$. Assume that $\lambda \gg n^{\frac{1}{d+2(2\wedge\beta)}}$ and $B \gg n^{\frac{2(2\wedge\beta)-2}{d+2(2\wedge\beta)}}$. For a confidence level $\alpha \in (0,1)$, let $q_{1-\alpha/2}$ be the normal quantile satisfying $\mathbb{P}\big(\mathcal{N}(0,1) \le q_{1-\alpha/2}\big) = 1-\alpha/2$. Then

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

When coupled with an appropriate lifetime selection method, Theorem 2.3.4 gives a fully feasible procedure for uncertainty quantification in Mondrian random forests. Our procedure requires no adjustment of the original Mondrian random forest estimator beyond ensuring that the bias is negligible, and in particular does not rely on sample splitting. The construction of confidence intervals is just one corollary of the weak convergence result given in Theorem 2.3.1, and follows immediately from Slutsky's theorem (Pollard, 2002, Chapter 7) with a consistent variance estimator. Other applications include hypothesis testing on the value of $\mu(x)$ at a design point x by inversion of the confidence interval, as well as parametric specification testing by comparison with a $\sqrt{n}$-consistent parametric regression estimator. The construction of simultaneous confidence intervals for finitely many points $x_1, \ldots, x_D$ can be accomplished either using standard multiple testing corrections or by first establishing a multivariate central limit theorem using the Cramér–Wold device (Pollard, 2002, Chapter 8) and formulating a consistent multivariate variance estimator.

2.4 Overview of proof strategy

This section provides some insight into the general approach we use to establish the main results in the preceding sections. We focus on the technical innovations forming the core of our arguments, and refer the reader to Section A.2 for detailed proofs, including those for the debiased estimator discussed in the upcoming Section 2.5.

Preliminary results

The starting point for our proofs is a characterization of the exact distribution of the shape of a Mondrian cell $T(x)$. This property is a direct consequence of the fact that the restriction of a Mondrian process to a subcell remains Mondrian (Mourtada et al., 2020, Fact 2). We have

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

for all $1 \leq j \leq d$, recalling that $T(x)_j$ is the side of the cell $T(x)$ aligned with axis j, and where E_{j1} and E_{j2} are mutually independent Exp(1) random variables. Our assumptions that $x \in (0,1)$ and $\lambda \to \infty$ make the boundary terms x_j and $1 - x_j$ eventually ignorable so

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

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

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

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

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

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

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

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

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

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

Central limit theorem

To prove our main central limit theorem result (Theorem 2.3.1), we use the martingale central limit theorem given by Hall and Heyde (1980, Theorem 3.2). For each $1 \leq i \leq n$ define $\mathcal{H}_{ni}$ to be the filtration generated by $\mathbf{T}$, $\mathbf{X}$, and $(\varepsilon_j : 1 \leq j \leq i)$, noting that $\mathcal{H}_{ni} \subseteq \mathcal{H}_{(n+1)i}$ because B increases as n increases. Define the $\mathcal{H}_{ni}$-measurable and square integrable variables

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

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

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

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

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

which is the variance of a non-linear function of the i.i.d. variables (X_i, ε_i), and so we apply the Efron–Stein inequality. The important insight here is that replacing a sample (X_i, ε_i) with an independent copy $(\tilde{X}_i, \tilde{\varepsilon}_i)$ can change the value of $N_b(x)$ by at most one. Further, this can happen only on the event $\{X_i \in T_b(x)\} \cup \{\tilde{X}_i \in T_b(x)\}$, which occurs with probability on the order $1/\lambda^d$ (the expected cell volume).

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

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

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

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

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

Bias characterization

Our second substantial technical result is the bias characterization given as Theorem 2.3.2, in which we precisely characterize the probability limit of the conditional bias

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

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

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

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

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

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

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

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

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

$$\mathbb{E}\left[\hat{\mu}(x)\right] - \mu(x) \approx \mathbb{E}\left[\frac{\int_{T_b(x)} (\mu(s) - \mu(x)) f(s)\,\mathrm{d}s}{\int_{T_b(x)} f(s)\,\mathrm{d}s}\right]. \tag{2.4}$$

Next we apply the multivariate version of Taylor's theorem to the integrands in both the numerator and the denominator in (2.4), and then apply the Maclaurin series of $\frac{1}{1+x}$ and the multinomial theorem to recover a single polynomial in $1/\lambda$. The error term is on the order of $1/\lambda^{\beta}$ and depends on the smoothness of μ and f, and the polynomial coefficients are given by various expectations involving exponential random variables. The final step is to verify using symmetry of Mondrian cells that all the odd monomial coefficients are zero, and to calculate some explicit examples of the form of the limiting bias.

2.5 Debiased Mondrian random forests

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

We use the explicit form of the bias given in Theorem 2.3.2 to construct a debiased version of the Mondrian forest estimator. Let $J \geq 0$ be the bias correction order. As such, with $J = 0$ we retain the original Mondrian forest estimator, with $J = 1$ we remove second-order bias, and with $J = \lfloor \underline{\beta}/2 \rfloor$ we remove bias terms up to and including order $2\lfloor \underline{\beta}/2 \rfloor$, giving the maximum possible bias reduction achievable in the Hölder class $\mathcal{H}^{\beta}$. As such, only bias terms of order $1/\lambda^{\beta}$ will remain.

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

$$\hat{\mu}_{\mathrm{d}}(x) = \sum_{r=0}^{J} \omega_r \hat{\mu}_r(x) = \sum_{r=0}^{J} \omega_r \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{i=1}^{n} Y_i \, \mathbb{I}\big\{X_i \in T_{rb}(x)\big\}}{N_{rb}(x)}. \tag{2.5}$$

This ensemble estimator retains the "forest" structure of the original estimators, but with varying lifetime parameters λ_r and coefficients ω_r. Thus by Theorem 2.3.2 we want to solve

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

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

2.5.1 Central limit theorem

In Theorem 2.5.1, we verify that a central limit theorem holds for the debiased random forest estimator $\hat{\mu}_{\mathrm{d}}(x)$ and give its limiting variance. The strategy and challenges associated with proving Theorem 2.5.1 are identical to those discussed earlier surrounding Theorem 2.3.1. In fact in Section A.2 we provide a direct proof only for Theorem 2.5.1 and deduce Theorem 2.3.1 as a special case.

Theorem 2.5.1 (Central limit theorem for the debiased Mondrian random forest estimator) *Suppose Assumptions 2.2.1 and 2.2.2 hold, $\mathbb{E}[Y_i^4 \mid X_i]$ is bounded, and $\frac{\lambda^d \log n}{n} \to 0$. Then*

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

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

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

It is easy to verify that in the case of no debiasing we have $J = 0$ and $a_0 = \omega_0 = 1$, yielding $\Sigma_{\mathrm{d}}(x) = \Sigma(x)$, and recovering Theorem 2.3.1.

Bias characterization

In Theorem 2.5.2 we verify that this debiasing procedure does indeed annihilate the desired bias terms, and its proof is a consequence of Theorem 2.3.2 and the construction of the debiased Mondrian random forest estimator $\hat{\mu}_{\mathrm{d}}(x)$.

Theorem 2.5.2 (Bias of the debiased Mondrian random forest estimator) *Grant Assumptions 2.2.1 and 2.2.2. In the notation of Theorem 2.3.2 with $\bar{\omega} = \sum_{r=0}^{J} \omega_r a_r^{-2J-2}$,*

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

Theorem 2.5.2 has the following consequence: the leading bias term is characterized in terms of $B_{J+1}(x)$ whenever $J < \beta/2 - 1$, or equivalently $J < \lfloor \beta/2 \rfloor$, that is, the debiasing order J does not exhaust the Hölder smoothness β. If this condition does not hold, then the estimator is fully debiased, and the resulting leading bias term is bounded above by $1/\lambda^\beta$ up to constants, but its form is left unspecified.

Variance estimation

As before, we propose a variance estimator in order to conduct feasible inference and show that it is consistent. With $\hat{\sigma}^2(x)$ as in (2.3) in Section 2.3, define the estimator

$$\hat{\Sigma}_{\mathrm{d}}(x) = \hat{\sigma}^2(x) \frac{n}{\lambda^d} \sum_{i=1}^{n} \left(\sum_{r=0}^{J} \omega_r \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{I}\{X_i \in T_{rb}(x)\}}{N_{rb}(x)} \right)^2. \tag{2.6}$$

Theorem 2.5.3 (Variance estimation)

Grant Assumptions 2.2.1 and 2.2.2, and suppose $\mathbb{E}[Y_i^4 \mid X_i]$ *is bounded almost surely. Then*

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

2.5.2 Confidence intervals

In analogy to Section 2.3, we now demonstrate the construction of feasible valid confidence intervals using the debiased Mondrian random forest estimator in Theorem 2.5.4. Once again we must ensure that the bias (now significantly reduced due to our debiasing procedure) is negligible when compared to the standard deviation (which is of the same order as before). We assume for simplicity that the estimator has been fully debiased by setting $J \geq \lfloor \beta/2 \rfloor$ to yield a leading bias of order $1/\lambda^\beta$, but intermediate "partially debiased" versions can easily be provided, with leading bias terms of order $1/\lambda^{\beta \wedge (2J+2)}$ in general. We thus require $\frac{1}{\lambda^\beta} + \frac{1}{\lambda\sqrt{B}} \ll \sqrt{\frac{\lambda^d}{n}}$, which can be satisfied by imposing the restrictions $\lambda \gg n^{\frac{1}{d+2\beta}}$ and $B \gg n^{\frac{2\beta-2}{d+2\beta}}$ on the lifetime parameter λ and forest size B.

Theorem 2.5.4 (Feasible confidence intervals using a debiased Mondrian random forest) *Suppose Assumptions 2.2.1 and 2.2.2 hold,* $\mathbb{E}[Y_i^4 \mid X_i]$ *is bounded, and* $\frac{\lambda^d \log n}{n} \to 0$. *Fix* $J \geq \lfloor \underline{\beta}/2 \rfloor$ *and assume that* $\lambda \gg n^{\frac{1}{d+2\beta}}$ *and* $B \gg n^{\frac{2\beta-2}{d+2\beta}}$. *For a confidence level* $\alpha \in (0,1)$, *let* $q_{1-\alpha/2}$ *be as in Theorem 2.3.4. Then*

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

One important benefit of our debiasing technique is made clear in Theorem 2.5.4: the restrictions imposed on the lifetime parameter λ are substantially relaxed, especially in smooth classes with large β. As well as the high-level of benefit of relaxed conditions, this is also useful for practical selection of appropriate lifetimes for estimation and inference respectively; see Section 2.6 for more details. Nonetheless, such improvements do not come without concession. The limiting variance $\Sigma_{\mathrm{d}}(x)$ of the debiased estimator is larger than that of the unbiased version (the extent of this increase depends on the choice of the debiasing parameters a_r), leading to wider confidence intervals and larger estimation error in small samples despite the theoretical asymptotic improvements.

2.5.3 Minimax optimality

Our final result Theorem 2.5.5 shows that, when using an appropriate sequence of lifetime parameters λ, the debiased Mondrian random forest estimator achieves, up to constants, the minimax-optimal rate of convergence for estimating a regression function $\mu \in \mathcal{H}^\beta$ in d dimensions (Stone, 1982). This result holds for all $d \geq 1$ and all $\beta > 0$, complementing a previous result established only for $\beta \in (0,2]$ by Mourtada et al. (2020).

Theorem 2.5.5 (Minimax optimality of the debiased Mondrian random forest estimator) *Grant Assumptions 2.2.1 and 2.2.2, and let* $J \geq \lfloor \underline{\beta}/2 \rfloor$, $\lambda \asymp n^{\frac{1}{d+2\beta}}$, *and* $B \gtrsim n^{\frac{2\beta-2}{d+2\beta}}$. *Then*

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

The sequence of lifetime parameters λ required in Theorem 2.5.5 are chosen to balance the bias and standard deviation bounds implied by Theorem 2.5.2 and Theorem 2.5.1 respectively, in order to minimize the pointwise mean squared error. While selecting an optimal debiasing order J needs only knowledge of an upper bound on the smoothness β, choosing an optimal sequence of λ values does assume that β is known a priori. The problem of adapting to β from data is challenging and beyond the scope of this chapter; we provide some practical advice for tuning parameter selection in Section 2.6.

Theorem 2.5.5 complements the minimaxity results proven by Mourtada et al. (2020) for Mondrian trees (with $\beta \leq 1$) and for Mondrian random forests (with $\beta \leq 2$), with one modification: our version is stated in pointwise rather than integrated mean squared error. This is because our debiasing procedure is designed to handle interior smoothing bias and so does not provide any correction for boundary bias. We leave the development of such boundary corrections to future work, but constructions similar to higher-order boundary-correcting kernels should be possible. If the region of integration is a compact set in the interior of $[0,1]^d$, then we do obtain an optimal integrated mean squared error bound: if $\delta \in (0, 1/2)$ is fixed then under the same conditions as Theorem 2.5.5,

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

The debiased Mondrian random forest estimator defined in (2.5) is a linear combination of Mondrian random forests, and as such contains both a sum over $0 \leq r \leq J$, representing the debiasing procedure, and a sum over $1 \leq b \leq B$, representing the forest averaging. We have thus far been interpreting this estimator as a debiased version of the standard Mondrian random forest given in (2.1), but it is equally valid to swap the order of these sums. This gives rise to an alternative point of view: we replace each Mondrian random tree with a "debiased" version, and then take a forest of such modified trees. This perspective is more in line with existing techniques for constructing randomized ensembles, where the outermost operation

represents a B-fold average of randomized base learners, not necessarily locally constant decision trees, each of which has a small bias component (Caruana, Niculescu-Mizil, Crew, and Ksikes, 2004; Zhou and Feng, 2019; Friedberg, Tibshirani, Athey, and Wager, 2020).

2.6 Tuning parameter selection

We discuss various procedures for selecting the parameters involved in fitting a debiased Mondrian random forest; namely the base lifetime parameter λ, the number of trees in each forest B, the bias correction order J, and the debiasing scale parameters a_r for $0 \leq r \leq J$.

2.6.1 Selecting the base lifetime parameter λ

The most important parameter is the base Mondrian lifetime parameter λ, which plays the role of a complexity parameter and thus governs the overall bias–variance trade-off of the estimator. Correct tuning of λ is especially important in two main respects: firstly, in order to use the central limit theorem established in Theorem 2.5.1, we must have that the bias converges to zero, requiring $\lambda \gg n^{\frac{1}{d+2\beta}}$. Secondly, the minimax optimality result of Theorem 2.5.5 is valid only in the regime $\lambda \asymp n^{\frac{1}{d+2\beta}}$, and thus requires careful determination in the more realistic finite-sample setting. For clarity, in this section we use the notation $\hat{\mu}_{\mathrm{d}}(x; \lambda, J)$ for the debiased Mondrian random forest with lifetime λ and debiasing order J as in (2.5). Similarly, write $\hat{\Sigma}_{\mathrm{d}}(x; \lambda, J)$ for the associated variance estimator given in (2.6).

For minimax-optimal point estimation when β is known, choose any sequence $\lambda \asymp n^{\frac{1}{d+2\beta}}$ and use $\hat{\mu}_{\mathrm{d}}(x; \lambda, J)$ with $J = \lfloor \beta/2 \rfloor$, following the theory given in Theorem 2.5.5. For an explicit example of how to choose the lifetime, one can instead use $\hat{\mu}_{\mathrm{d}}\big(x; \hat{\lambda}_{\mathrm{AIMSE}}(J-1), J-1\big)$ so that the leading bias is explicitly characterized by Theorem 2.5.2, and with $\hat{\lambda}_{\mathrm{AIMSE}}(J-1)$ as defined below. This is no longer minimax-optimal as $J - 1 < J$ does not satisfy the conditions of Theorem 2.5.5.

For performing inference, a more careful procedure is required; we suggest the following method assuming $\beta > 2$. Set $J = \lfloor \beta/2 \rfloor$ as before, and use $\hat{\mu}_{\mathrm{d}}\big(x; \hat{\lambda}_{\mathrm{AIMSE}}(J-1), J\big)$ and $\hat{\Sigma}_{\mathrm{d}}\big(x; \hat{\lambda}_{\mathrm{AIMSE}}(J-1), J\big)$ to construct a confidence interval. The reasoning for this is that we select a lifetime tailored for a more biased estimator than we actually use. This results in an

inflated lifetime estimate, guaranteeing the resulting bias is negligible when it is plugged into the fully debiased estimator. This approach to tuning parameter selection and debiasing for valid nonparametric inference corresponds to an application of robust bias correction (Calonico et al., 2018, 2022), where the point estimator is bias-corrected and the robust standard error estimator incorporates the additional sampling variability introduced by the bias correction. This leads to a more refined distributional approximation but does not necessarily exhaust the underlying smoothness of the regression function. An alternative inference approach based on Lepskii's method (Lepskii, 1992; Birgé, 2001) could be developed with the latter goal in mind.

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

Lifetime selection with polynomial estimation

Firstly, suppose $X_i \sim \mathrm{Unif}\left([0,1]^d\right)$ and that the leading bias of $\hat{\mu}_{\mathrm{d}}(x)$ is well approximated by an additively separable function so that, writing $\partial_j^{2J+2}\mu(x)$ for $\partial_j^{2J+2}\mu(x)/\partial x_j^{2J+2}$,

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

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

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

The asymptotic integrated mean squared error (AIMSE) is

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

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

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

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

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

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

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

and let $P_x = \big(1, x_1, x_1^2, \ldots, x_1^{2J+4}, x_2, x_2^2, \ldots, x_2^{2J+4}, \ldots, x_d, x_d^2, \ldots, x_d^{2J+4}\big)$. Then we define the derivative estimator as

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

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

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

Lifetime selection with cross-validation

As an alternative to the analytic plug-in methods described above, one can use a cross-validation approach. While leave-one-out cross-validation (LOOCV) can be applied directly (Fan et al., 2020), the linear smoother structure of the (debiased) Mondrian random forest estimator allows a computationally simpler formulation. Writing $\hat{\mu}_{\mathrm{d}}^{-i}(x)$ for a debiased

Mondrian random forest estimator fitted without the ith data sample, it is easy to show that

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

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

$$\mathrm{GCV}(\lambda, J) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{\mu}_{\mathrm{d}}(X_i)}{1-\bar{a}^d\lambda^d/n}\right)^2. \tag{2.7}$$

The lifetime can then be selected by computing either $\hat{\lambda}_{\mathrm{LOOCV}} \in \arg\min_\lambda \mathrm{LOOCV}(\lambda, J)$ or $\hat{\lambda}_{\mathrm{GCV}} \in \arg\min_\lambda \mathrm{GCV}(\lambda, J)$. See Section 2.7 for a practical illustration.

2.6.2 Choosing the other parameters

The number B of trees in each forest

If no debiasing is applied, we suggest $B = \sqrt{n}$ to satisfy Theorem 2.3.4. If debiasing is used then we recommend $B = n^{\frac{2J-1}{2J}}$, consistent with Theorem 2.5.4 and Theorem 2.5.5.

The debiasing order J

When debiasing a Mondrian random forest, one must decide how many orders of bias to remove. This requires some oracle knowledge of the Hölder smoothness of μ and f, which is difficult to estimate statistically. As such, we recommend removing only the first one or two bias terms, taking $J \in \{0, 1, 2\}$ to avoid overly inflating the variance of the estimator.

The debiasing coefficients a_r

As in Section 2.5, we take a_r to be a fixed geometric or arithmetic sequence. For example, one could set $a_r = (1+\gamma)^r$ or $a_r = 1 + \gamma r$ for some $\gamma > 0$. We suggest taking $a_r = 1.05^r$.

2.7 Illustrative example: weather forecasting

To demonstrate our methodology for estimation and inference with Mondrian random forests, we consider a simple application to a weather forecasting problem. We emphasize that the main aim of this section is to provide intuition and understanding for how a Mondrian random forest may be used in practice, and we refrain from an in-depth analysis of the specific results obtained. Indeed, our assumption of i.i.d. data is certainly violated with weather data, due to the time-series structure of sequential observations. Nonetheless, we use data from the Bureau of Meteorology, Australian Government (2017), containing daily weather information from 2007–2017, at 49 different locations across Australia, with $n = 125\,927$ samples.

We consider the classification problem of predicting whether or not it will rain on the following day using two covariates: the percentage relative humidity, and the pressure in mbar, both at 3pm on the current day. For the purpose of framing this as a nonparametric regression problem, we consider estimating the probability of rain as the regression function by setting $Y_i = 1$ if there is rain on the following day and $Y_i = 0$ otherwise. Outliers with pressure less than 985 mbar or more than 1040 mbar are removed to justify the assertion in Assumption 2.2.1 that the density of the covariates should be bounded away from zero, and

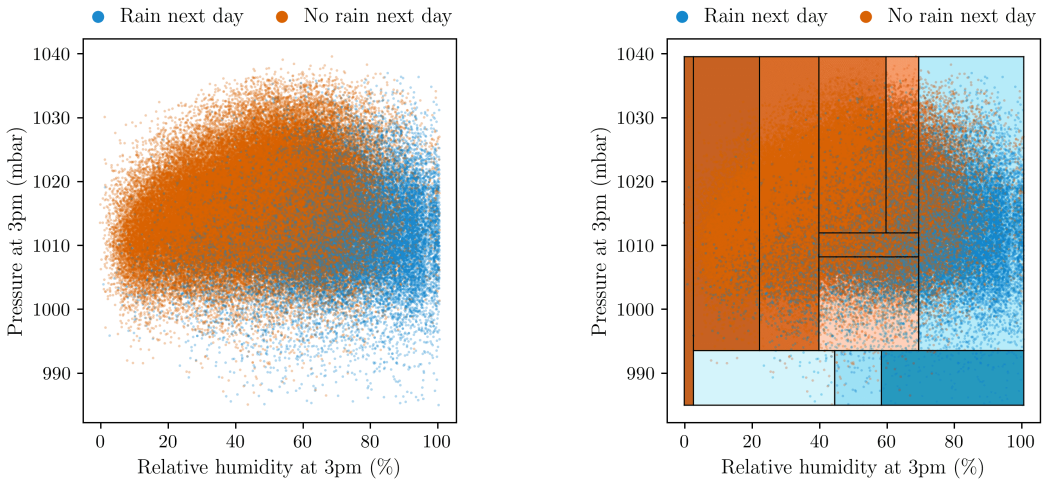


Figure 2.2: Australian weather forecasting data. Left: colors indicate the response variable of dry (orange) or wet (blue) on the following day. Right: the data is overlaid with a Mondrian random tree, fitted with a lifetime of $\lambda = 5$ selected by generalized cross-validation. Cell colors represent the response proportions.

the features are linearly scaled to provide normalized samples $(X_i, Y_i) \in [0,1]^2 \times \{0,1\}$. We fit a Mondrian random forest to the data as defined in Section 2.2.3, selecting the lifetime parameter with the generalized cross-validation (GCV) method detailed in Section 2.6.1.

Figure 2.2 plots the data, using colors to indicate the response values, and illustrates how a single Mondrian tree is fitted by sampling from an independent Mondrian process and then computing local averages (equivalent to response proportions in this special setting with binary outcomes) within each cell. The general pattern of rain being predicted by high humidity and low pressure is apparent, with the preliminary tree estimator taking the form of a step function on axis-aligned rectangles. This illustrates the first-order bias of Mondrian random trees discussed in Section 2.3.1, with the piecewise constant estimator providing a poor approximation for the smooth true regression function.

Figure 2.3 adds more trees to the estimator, demonstrating the effect of increasing the forest size first to $B = 2$ and then to $B = 40$. As more trees are included in the Mondrian random forest, the regression estimate $\hat{\mu}(x)$ becomes smoother and therefore also enjoys improved bias properties as shown in Theorem 2.3.2, assuming a correct model specification.

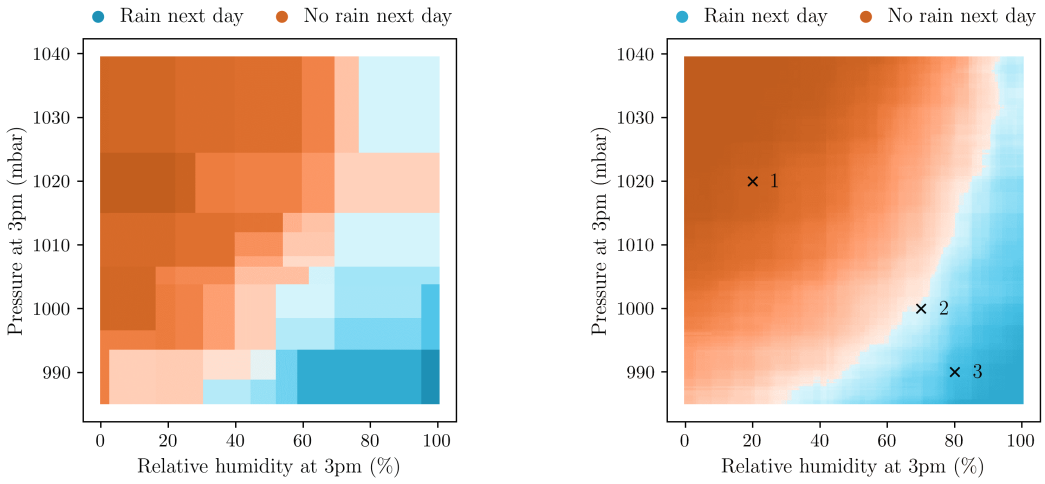


Figure 2.3: Fitting Mondrian random forests to the Australian weather data. Left: with $B = 2$ trees, individual cells are clearly visible and the step function persists. Right: with $B = 40$ trees, the estimate is much smoother as the individual tree estimates average out. Three design points are identified for further analysis.

We also choose three specific design points in the (humidity, pressure) covariate space, namely (20%, 1020 mbar), (70%, 1000 mbar), and (80%, 990 mbar), at which to conduct inference by constructing confidence intervals. See Table 2.5 for the results.

In Figure 2.4 we show the mean squared error and GCV scores computed using (2.7) with $B = 400$ trees for several candidate lifetime parameters λ. As expected, the mean squared error decreases monotonically as λ increases and the model overfits, but the GCV score is minimized at a value which appropriately balances the bias and variance; we take $\lambda = 5$. We then fit a debiased Mondrian forest with bias correction order $J = 1$ as described in Section 2.5, using $B = 20$ trees at each debiasing level $r \in \{0, 1\}$ for a total of 40 trees. We continue to use the same lifetime parameter $\lambda = 5$ selected through GCV without debiasing, following the approach recommended in Section 2.6.1 to ensure valid inference through negligible bias. The resulting debiased Mondrian random forest estimate is noticeably less smooth than the version without bias correction. This is expected due to both the inflated variance resulting from the debiasing procedure, and the undersmoothing enacted by selecting a lifetime parameter using GCV on the original estimator without debiasing.

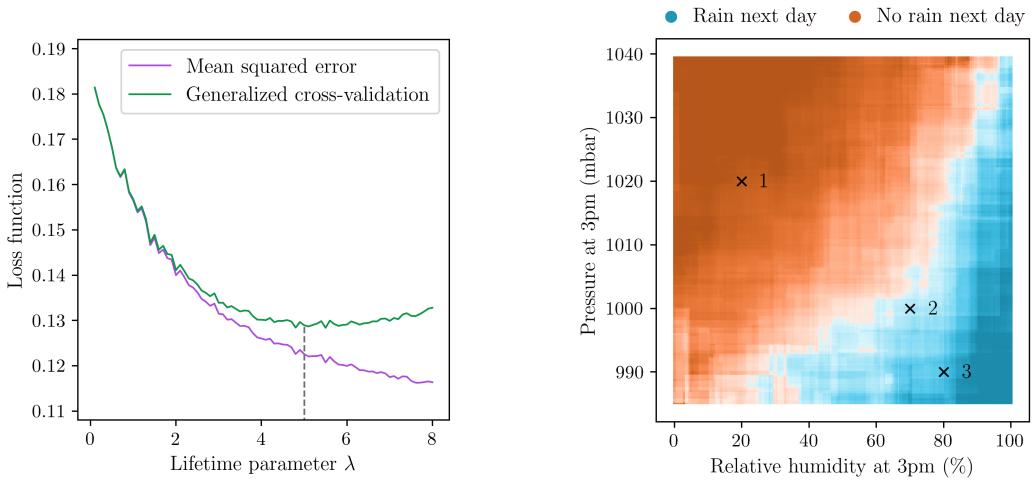


Figure 2.4: Left: mean squared error and generalized cross-validation scores for Mondrian random forests with the Australian weather data. Right: a debiased Mondrian random forest with $B = 20$, giving 40 trees in total. Three design points are identified for further analysis.

Point	Humidity	Pressure	No debiasing, $J = 0$		Debiasing, $J = 1$	
			$\hat{\mu}(x)$	95% CI	$\hat{\mu}(x)$	95% CI
1	20%	1020 mbar	4.2%	3.9% – 4.5%	2.0%	1.6% – 2.4%
2	70%	1000 mbar	52.6%	51.7% – 53.6%	59.8%	57.8% – 61.9%
3	80%	990 mbar	78.1%	75.0% – 81.2%	93.2%	86.7% – 99.6%

Table 2.5: Results for the Australian weather data at three specified design points.

Table 2.5 presents numerical results for estimation and inference at the three specified design points. We first give the outcomes without debiasing, using a Mondrian random forest with $B = 400$ trees and $\lambda = 5$ selected by GCV. We then show the results with a first-order ($J = 1$) debiased Mondrian random forest using $B = 200$ (again a total of 400 trees) and the same value of $\lambda = 5$. The predicted chance of rain $\hat{\mu}(x)$ is found to vary substantially across different covariate values, and the resulting confidence intervals (CI) are generally narrow due to the large sample size and moderate lifetime parameter. The forest with debiasing exhibits more extreme predictions away from 50% and wider confidence intervals in general, in line with the illustration in Figure 2.4. Interestingly, the confidence intervals for the non-debiased and debiased estimators do not intersect, indicating that the original estimator is severely biased, and providing further justification for our modified debiased random forest estimator.

2.8 Conclusion

We gave a central limit theorem for the Mondrian random forest estimator and showed how to perform statistical inference on an unknown nonparametric regression function. We introduced debiased versions of the Mondrian random forest, and demonstrated their advantages for statistical inference and minimax-optimal estimation. We discussed tuning parameter selection, enabling a fully feasible and practical methodology. An application to weather forecasting was presented as an illustrative example. Implementations of this chapter's methodology and empirical results are provided by a Julia package at `github.com/wgunderwood/MondrianForests.jl`. This work is based on Cattaneo et al. (2023), and has been presented by Underwood at the University of Illinois Statistics Seminar (2024), the University of Michigan Statistics Seminar (2024), and the University of Pittsburgh Statistics Seminar (2024).

Chapter 3

Dyadic Kernel Density Estimators

Dyadic data is often encountered when quantities of interest are associated with the edges of a network. As such, it plays an important role in statistics, econometrics, and many other data science disciplines. We consider the problem of uniformly estimating a dyadic Lebesgue density function, focusing on nonparametric kernel-based estimators taking the form of dyadic empirical processes. The main contributions of this chapter include the minimax-optimal uniform convergence rate of the dyadic kernel density estimator, along with strong approximation results for the associated standardized and Studentized t-processes. A consistent variance estimator enables the construction of valid and feasible uniform confidence bands for the unknown density function. We showcase the broad applicability of our results by developing novel counterfactual density estimation and inference methodology for dyadic data, which can be used for causal inference and program evaluation. A crucial feature of dyadic distributions is that they may be "degenerate" at certain points in the support of the data, a property making our analysis somewhat delicate. Nonetheless our methods for uniform inference remain robust to the potential presence of such points. For implementation purposes, we discuss inference procedures based on positive semi-definite covariance estimators, mean squared error optimal bandwidth selectors, and robust bias correction techniques. We illustrate the empirical finite-sample performance of our methods both in simulations and with

real-world trade data, for which we make comparisons between observed and counterfactual trade distributions in different years. Our technical results concerning strong approximations and maximal inequalities are of potential independent interest.

3.1 Introduction

Dyadic data, also known as graphon data, plays an important role in the statistical, social, behavioral, and biomedical sciences. In network settings, this type of dependent data captures interactions between the units of study, and its analysis is of interest in statistics (Kolaczyk, 2009), economics (Graham, 2020), psychology (Kenny, Kashy, and Cook, 2020), public health (Luke and Harris, 2007), and many other data science disciplines. For $n \geq 2$, a dyadic data set contains $\frac{1}{2}n(n-1)$ observed real-valued random variables

$$\mathbf{W}_n = (W_{ij} : 1 \leq i < j \leq n), \qquad W_{ij} = W(A_i, A_j, V_{ij}),$$

where W is an unknown function, $\mathbf{A}_n = (A_i : 1 \leq i \leq n)$ are independent and identically distributed (i.i.d.) latent random variables, and $\mathbf{V}_n = (V_{ij} : 1 \leq i < j \leq n)$ are i.i.d. latent random variables independent of $\mathbf{A}_n$. A natural interpretation of this data is as a complete undirected network on n vertices, with the latent variable A_i associated with node i and the observed variable W_{ij} associated with the edge between nodes i and j. The data generating process above is justified without loss of generality by the celebrated Aldous–Hoover representation theorem for exchangeable arrays (Aldous, 1981; Hoover, 1979).

Various distributional features of dyadic data are of interest in applications. Most of the statistical literature focuses on parametric analysis, almost exclusively considering moments of (transformations of) the identically distributed W_{ij}. See Davezies, D'Haultfœuille, and Guyonvarch (2021), Gao and Ma (2021), and Matsushita and Otsu (2021) for contemporary contributions and overviews. More recently, however, a few nonparametric procedures for dyadic data have been proposed in the literature (Graham, Niu, and Powell, 2021, 2024).

With the aim of estimating functions associated with W_{ij} using nonparametric kernel methods, we investigate the statistical properties of the class of local stochastic processes

$$w \mapsto \hat{f}_W(w) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k_h(W_{ij}, w), \tag{3.1}$$

where $k_h(s, w)$ is a kernel function that can change with the n-varying bandwidth parameter $h = h(n)$ and the evaluation point $w \in \mathcal{W} \subseteq \mathbb{R}$. For each $w \in \mathcal{W}$ and with an appropriate choice of the kernel function (e.g. $k_h(s, w) = K((s - w)/h)/h$ for an interior point w of $\mathcal{W}$ and a fixed symmetric integrable kernel function K), the statistic $\hat{f}_W(w)$ becomes a kernel density estimator for the Lebesgue density function $f_W(w) = \mathbb{E}\big[f_{W|AA}(w \mid A_i, A_j)\big]$, where $f_{W|AA}(w \mid A_i, A_j)$ denotes the conditional Lebesgue density of W_{ij} given A_i and A_j. Setting $k_h(s, w) = K((s-w)/h)/h$, Graham et al. (2024) recently introduced the dyadic point estimator $\hat{f}_W(w)$ and studied its large sample properties pointwise in $w \in \mathcal{W} = \mathbb{R}$, while Chiang and Tan (2023) established its rate of convergence uniformly in $w \in \mathcal{W}$ for a compact interval $\mathcal{W}$ strictly contained in the support of the dyadic data W_{ij}. Chiang, Kato, and Sasaki (2023) obtained a distributional approximation for the supremum statistic $\sup_{w \in \mathcal{W}} \big|\hat{f}_W(w)\big|$ over a finite collection $\mathcal{W}$ of design points. More generally, as we discuss below, the estimand $f_W(w)$ is useful in different applications because it forms the basis for counterfactual distributional analysis (Section 3.7) and other nonparametric and semiparametric methods (Section 3.8). While we assume throughout that the network is complete, our approach generalizes in a straightforward way to networks with missing edges, as in Section 3.7.1. This can be seen by setting $W_{ij} = -\infty$ whenever the edge $\{i, j\}$ is not present, so that the law of W_{ij} is a mixture between a continuous distribution and a point mass at $-\infty$. We apply our methodology to recover the continuous component of this distribution, following Chiang et al. (2023).

We contribute to the emerging literature on nonparametric smoothing methods for dyadic data with two main technical results. Firstly, we derive the minimax rate of uniform convergence for density estimation with dyadic data, and show that the estimator $\hat{f}_W$ in (3.1) is minimax-optimal under appropriate conditions. Secondly, we present a set of uniform distributional approximation results for the *entire* stochastic process $\big(\hat{f}_W(w) : w \in \mathcal{W}\big)$.

Furthermore, we illustrate the usefulness of our main results with two distinct substantive statistical applications: (i) confidence bands for f_W (Section 3.5), and (ii) estimation and inference for counterfactual dyadic distributions (Section 3.7). Our main results also lay the foundation for studying the uniform distributional properties of other nonparametric and semiparametric tests and estimators based on dyadic data (Section 3.8). Importantly, our inference results cannot be deduced from the existing U-statistic, empirical process and U-process theory available in the literature (van der Vaart and Wellner, 1996; Giné and Nickl, 2021) because, as explained in detail below, $\hat{f}_W(w)$ is not a standard U-statistic, nor is (a suitable rescaling of) the stochastic process $\hat{f}_W$ Donsker in general, and the underlying dyadic data $\mathbf{W}_n$ exhibits statistical dependence due to its network structure.

Section 3.2 outlines the setup and presents the main assumptions imposed throughout this chapter. We demonstrate in Theorem 3.2.1 how the smoothing bias of the dyadic kernel density estimator can be controlled, and then discuss a Hoeffding-type decomposition of the U-statistic-like $\hat{f}_W$ in Lemma 3.2.1. This is more general than the standard Hoeffding decomposition for second-order U-statistics due to the intrinsic dyadic data structure. In particular, (3.2) shows that $\hat{f}_W(w)$ decomposes into a sum of the four terms $B_n(w)$, $L_n(w)$, $E_n(w)$, and $Q_n(w)$, where $E_n(w)$ is not present in the classical second-order U-statistic theory. The first term $B_n(w)$ captures the usual smoothing bias, the second term $L_n(w)$ is akin to the Hájek projection for second-order U-statistics, the third term $E_n(w)$ is a mean-zero double average of conditionally independent terms, and the fourth term $Q_n(w)$ is a negligible totally degenerate second-order U-process. The leading stochastic fluctuations of the process $\hat{f}_W$ are captured by L_n and E_n, both of which are known to be asymptotically distributed as Gaussian random variables pointwise in $w \in \mathcal{W}$ (Graham et al., 2024). However, the Hájek projection term L_n will often be "degenerate" at some or possibly all evaluation points $w \in \mathcal{W}$. The three possible types of degeneracy are detailed in Lemma 3.2.2, and we establish bounds in probability for each term in the Hoeffding-type decomposition in Lemma 3.2.3. We give an example of a simple family of dyadic distributions exhibiting all three degeneracy types.

Section 3.3 studies minimax convergence rates for point estimation of f_W uniformly over $\mathcal{W}$ and gives precise conditions under which the estimator $\hat{f}_W$ is minimax-optimal. Firstly, in Theorem 3.3.1 we establish the uniform rate of convergence of $\hat{f}_W$ for f_W. This result improves upon the recent paper of Chiang and Tan (2023) by allowing for compactly supported dyadic data and generic kernel-like functions k_h (including boundary-adaptive kernels), while also explicitly accounting for possible degeneracy of the Hájek projection term L_n at some or possibly all points $w \in \mathcal{W}$. Secondly, in Theorem 3.3.2 we derive the minimax uniform convergence rate for estimating f_W, again allowing for possible degeneracy, and verify that it is achieved by $\hat{f}_W$. This result appears to be new to the literature, complementing recent work on parametric moment estimation using graphon data (Gao and Ma, 2021) and on nonparametric kernel-based regression using dyadic data (Graham et al., 2021).

Section 3.4 presents a distributional analysis of the stochastic process $\hat{f}_W$ uniformly in $w \in \mathcal{W}$. Because the t-process based on $\hat{f}_W$ is not asymptotically tight in general, it does not converge weakly in the space of uniformly bounded real functions supported on $\mathcal{W}$ and equipped with the uniform norm (van der Vaart and Wellner, 1996), and hence is non-Donsker. To circumvent this problem, we employ strong approximation methods to characterize its distributional properties. Up to the smoothing bias term B_n and the negligible term Q_n, it suffices to consider the stochastic process $w \mapsto L_n(w) + E_n(w)$. Since L_n can be degenerate at some or possibly all points $w \in \mathcal{W}$, and also because under some bandwidth choices both L_n and E_n can be of comparable order, it is crucial to analyze the joint distributional properties of L_n and E_n. To do so, we employ a carefully crafted conditioning approach where we first establish an unconditional strong approximation for L_n and a conditional-on-$\mathbf{A}_n$ strong approximation for E_n. We then combine these to obtain a strong approximation for $L_n + E_n$.

The stochastic process L_n is an empirical process indexed by an n-varying class of functions depending only on the i.i.d. random variables $\mathbf{A}_n$. Thus we use the celebrated Hungarian construction (Komlós, Major, and Tusnády, 1975), building on ideas in Giné, Koltchinskii, and Sakhanenko (2004) and Giné and Nickl (2010). The resulting rate of strong approximation

is optimal, and follows from a generic strong approximation result of potential independent interest given in Section B.2. Our main result for L_n is given as Lemma 3.4.1, and makes explicit the potential presence of degenerate points.

The stochastic process E_n is an empirical process depending on the dyadic variables W_{ij} and indexed by an n-varying class of functions. When conditioning on $\mathbf{A}_n$, the variables W_{ij} are independent but not necessarily identically distributed (i.n.i.d.), and thus we establish a conditional-on-$\mathbf{A}_n$ strong approximation for E_n based on the Yurinskii coupling (Yurinskii, 1978), leveraging a refinement obtained by Belloni, Chernozhukov, Chetverikov, and Fernández-Val (2019, Lemma 38). This result follows from a generic strong approximation result which gives a novel rate of strong approximation for (local) empirical processes based on i.n.i.d. data, given in Section B.2. Lemma 3.4.2 gives our conditional strong approximation for E_n.

Once the unconditional strong approximation for L_n and the conditional-on-$\mathbf{A}_n$ strong approximation for E_n are established, we show how to properly "glue" them together to deduce a final unconditional strong approximation for $L_n + E_n$ and hence also for $\hat{f}_W$ and its associated t-process. This final step requires some additional technical work. Firstly, building on our conditional strong approximation for E_n, we establish an unconditional strong approximation for E_n in Lemma 3.4.3. We then employ a generalization of the celebrated Vorob'ev–Berkes–Philipp theorem (Dudley, 1999), given in given in Section B.2, to deduce a *joint* strong approximation for (L_n, E_n) and, in particular, for $L_n + E_n$. Thus we obtain our main result in Theorem 3.4.1, which establishes a valid strong approximation for the t-process associated with $\hat{f}_W$. This uniform inference result complements the recent contribution of Davezies et al. (2021), which is not applicable here as the t-process is non-Donsker.

We illustrate the applicability of our strong approximation results for $\hat{f}_W$ and its associated t-process by constructing valid standardized uniform confidence bands for the unknown density function f_W in Theorem 3.4.2. Instead of relying on extreme value theory (as in Giné, Koltchinskii, and Sakhanenko, 2004), we employ anti-concentration methods, following Chernozhukov, Chetverikov, and Kato (2014a). This illustration improves on the recent work of Chiang et al. (2023), which obtained simultaneous confidence intervals for the dyadic density f_W based on a high-dimensional central limit theorem over rectangles, following prior work

by Chernozhukov, Chetverikov, and Kato (2017a). The distributional approximation therein is applied to the Hájek projection term L_n only, whereas our main construction leading to Theorem 3.4.1 gives a strong approximation for the entire U-process-like $\hat{f}_W$ and its associated t-process, uniformly on $\mathcal{W}$. As a consequence, our uniform inference theory is robust to potential unknown degeneracies in L_n by virtue of our strong approximation for $L_n + E_n$ and the use of proper standardization, delivering a "rate-adaptive" inference procedure. Our result appears to be the first to provide confidence bands that are valid uniformly over $w \in \mathcal{W}$ rather than merely over a finite collection of design points. Moreover, they provide distributional approximations for the whole t-statistic process, which can be useful in applications where functionals other than the supremum are of interest.

Section 3.5 addresses outstanding issues of implementation. Firstly, we discuss estimation of the covariance function of the Gaussian process underlying our strong approximation results. We present two estimators, one based on a plug-in method, and the other on a positive semi-definite regularization thereof (Laurent and Rendl, 2005). We derive the uniform convergence rates for both estimators in Lemma 3.5.1, which we then use to justify Studentization of $\hat{f}_W$ and a feasible simulation-based approximation of the infeasible Gaussian process underlying our strong approximation results. Secondly, we discuss integrated mean squared error (IMSE) bandwidth selection and provide a simple rule-of-thumb implementation for applications (Wand and Jones, 1994; Simonoff, 1996). Thirdly, we provide feasible, valid uniform inference methods for f_W by employing robust bias correction (Calonico et al., 2018, 2022). Algorithm 1 summarizes our entire feasible uniform inference methodology.

Section 3.6 reports empirical evidence for our proposed feasible robust bias-corrected confidence bands for f_W. We use simulations to show that these confidence bands are robust to potential unknown degenerate points in the underlying dyadic distribution.

Section 3.7 presents novel results for counterfactual dyadic density estimation and inference, offering an application of our general theory to a substantive problem in statistics and other data science disciplines. Counterfactual distributions are important for causal inference and policy evaluation (DiNardo, Fortin, and Lemieux, 1996; Chernozhukov, Fernández-Val, and Melly, 2013b), and in the context of network data, such analysis can be used to answer

empirical questions such as "what would the international trade distribution have been if the gross domestic product (GDP) of the countries had remained the same as in a previous year?" We formally show how our theory for kernel-based dyadic estimators can be used to infer the counterfactual density function of dyadic data had some monadic covariates followed a different distribution. We propose a two-step semiparametric reweighting approach in which we first estimate the Radon–Nikodym derivative between the observed and counterfactual covariate distributions using a simple parametric estimator, and then use this to construct a weighted dyadic kernel density estimator. We present uniform consistency, strong approximation, and feasible inference results for this dyadic counterfactual density estimator. Finally, we illustrate our methods with a real dyadic data set recording bilateral trade between countries from 1995 to 2005, using GDP as a covariate for the counterfactual analysis.

Section 3.8 discusses further statistical applications of our main results, including dyadic density testing and nonparametric and semiparametric dyadic regression. Section 3.9 concludes. Appendix B includes other technical and methodological results, proofs, and additional details omitted here to conserve space. Section B.2 may be of independent interest, containing two generic strong approximation theorems for empirical processes, a generalized Vorob'ev–Berkes–Philipp theorem, and a maximal inequality for i.n.i.d. random variables.

3.1.1 Notation

The total variation norm of a real-valued function g of a single real variable is written as $\|g\|_{\mathrm{TV}} = \sup_{n\geq 1} \sup_{x_1\leq\cdots\leq x_n} \sum_{i=1}^{n-1} |g(x_{i+1}) - g(x_i)|$. For an integer $m \geq 0$, denote by $\mathcal{C}^m(\mathcal{X})$ the space of all functions from $\mathbb{R}$ to $\mathbb{R}$ which are m times continuously differentiable on a subset $\mathcal{X} \subseteq \mathbb{R}$. For $C > 0$, define the Hölder class with smoothness parameter $\beta > 0$ to be $\mathcal{H}^\beta_C(\mathcal{X}) = \big\{g \in \mathcal{C}^{\underline{\beta}}(\mathcal{X}) : \max_{1\leq r\leq \underline{\beta}} \big|g^{(r)}(x)\big| \leq C, \big|g^{(\underline{\beta})}(x) - g^{(\underline{\beta})}(x')\big| \leq C|x - x'|^{\beta-\underline{\beta}}, \forall x, x' \in \mathcal{X}\big\}$, where $\underline{\beta}$ denotes the largest integer which is strictly less than β. Note that $\mathcal{H}^1_C(\mathcal{X})$ is the class of C-Lipschitz functions on $\mathcal{X}$. For $a \in \mathbb{R}$ and $b \geq 0$, we write $[a \pm b]$ for the interval $[a - b, a + b]$. For non-negative sequences a_n and b_n, write $a_n \lesssim b_n$ or $a_n = O(b_n)$ to indicate that a_n/b_n is bounded for $n \geq 1$. Write $a_n \ll b_n$ or $a_n = o(b_n)$ if $a_n/b_n \to 0$. If

$a_n \lesssim b_n \lesssim a_n$, write $a_n \asymp b_n$. For random non-negative sequences A_n and B_n, write $A_n \lesssim_{\mathbb{P}} B_n$ or $A_n = O_{\mathbb{P}}(B_n)$ if A_n/B_n is bounded in probability. Write $A_n = o_{\mathbb{P}}(B_n)$ if $A_n/B_n \to 0$ in probability. For $a, b \in \mathbb{R}$, define $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

3.2 Setup

We impose the following two assumptions throughout this chapter, which concern firstly the dyadic data generating process, and secondly the choice of kernel and bandwidth sequence.

Assumption 3.2.1 (Data generation)
Let $\mathbf{A}_n = (A_i : 1 \leq i \leq n)$ *be i.i.d. random variables supported on* $\mathcal{A} \subseteq \mathbb{R}$ *and let* $\mathbf{V}_n = (V_{ij} : 1 \leq i < j \leq n)$ *be i.i.d. random variables with a Lebesgue density* f_V *on* $\mathbb{R}$*, with* $\mathbf{A}_n$ *independent of* $\mathbf{V}_n$*. Let* $W_{ij} = W(A_i, A_j, V_{ij})$ *and* $\mathbf{W}_n = (W_{ij} : 1 \leq i < j \leq n)$*, where* W *is an unknown real-valued function which is symmetric in its first two arguments. Let* $\mathcal{W} \subseteq \mathbb{R}$ *be a compact interval with positive Lebesgue measure* $\mathrm{Leb}(\mathcal{W})$*. The conditional distribution of* W_{ij} *given* A_i *and* A_j *admits a Lebesgue density* $f_{W|AA}(w \mid A_i, A_j)$*. For* $C_{\mathrm{H}} > 0$ *and* $\beta \geq 1$*, take* $f_W \in \mathcal{H}^{\beta}_{C_{\mathrm{H}}}(\mathcal{W})$ *where* $f_W(w) = \mathbb{E}\left[f_{W|AA}(w \mid A_i, A_j)\right]$ *and* $f_{W|AA}(\cdot \mid a, a') \in \mathcal{H}^{1}_{C_{\mathrm{H}}}(\mathcal{W})$ *for all* $a, a' \in \mathcal{A}$*. Suppose* $\sup_{w \in \mathcal{W}} \|f_{W|A}(w \mid \cdot)\|_{\mathrm{TV}} < \infty$ *where* $f_{W|A}(w \mid a) = \mathbb{E}\left[f_{W|AA}(w \mid A_i, a)\right]$.

In Assumption 3.2.1 we require the density f_W be in a β-smooth Hölder class of functions on the compact interval $\mathcal{W}$. Hölder classes are well established in the minimax estimation literature (Stone, 1982; Giné and Nickl, 2021), with the smoothness parameter β appearing in the minimax-optimal rate of convergence. If the Hölder condition is satisfied only piecewise, then our results remain valid provided that the boundaries between the pieces are known and treated as boundary points.

If $W(a_1, a_2, v)$ is strictly monotonic and continuously differentiable in its third argument, we can give the conditional density of W_{ij} explicitly using the usual change-of-variables formula: with $w = W(a_1, a_2, v)$, we have $f_{W|AA}(w \mid a_1, a_2) = f_V(v)\left|\partial W(a_1, a_2, v)/\partial v\right|^{-1}$.

Assumption 3.2.2 (Kernels and bandwidth)

Let $h = h(n) > 0$ be a sequence of bandwidths satisfying $h \log n \to 0$ and $\frac{\log n}{n^2 h} \to 0$. For each $w \in \mathcal{W}$, let $k_h(\cdot, w)$ be a real-valued function supported on $[w \pm h] \cap \mathcal{W}$. For an integer $p \geq 1$, let k_h belong to a family of boundary bias-corrected kernels of order p, i.e.,

$$\int_{\mathcal{W}} (s-w)^r k_h(s,w)\,\mathrm{d}s \quad \begin{cases} = 1 & \text{for all } w \in \mathcal{W} \text{ if } r = 0, \\ = 0 & \text{for all } w \in \mathcal{W} \text{ if } 1 \leq r \leq p-1, \\ \neq 0 & \text{for some } w \in \mathcal{W} \text{ if } r = p. \end{cases}$$

Also, for $C_{\mathrm{L}} > 0$, suppose $k_h(s,\cdot) \in \mathcal{H}^1_{C_{\mathrm{L}} h^{-2}}(\mathcal{W})$ for all $s \in \mathcal{W}$.

This assumption allows for all standard compactly supported and possibly boundary-corrected kernel functions (Wand and Jones, 1994; Simonoff, 1996), constructed for example by taking polynomials on a compact interval and solving a linear system for the coefficients. Assumption 3.2.2 implies (see Lemma B.3.15 in Appendix B) that if $h \leq 1$ then k_h is uniformly bounded by $C_{\mathrm{k}} h^{-1}$ where $C_{\mathrm{k}} := 2C_{\mathrm{L}} + 1 + 1/\operatorname{Leb}(\mathcal{W})$.

3.2.1 Bias characterization

We begin by characterizing and bounding the bias $B_n(w) = \mathbb{E}\big[\hat{f}_W(w)\big] - f_W(w)$. Theorem 3.2.1 is a standard result for the non-random smoothing bias in kernel density estimation with higher-order kernels and boundary bias correction, and does not rely on the dyadic structure.

Theorem 3.2.1 (Bias bound)

Suppose that Assumptions 3.2.1 and 3.2.2 hold. For $w \in \mathcal{W}$ define the leading bias term as

$$b_p(w) = \frac{f_W^{(p)}(w)}{p!} \int_{\mathcal{W}} k_h(s,w) \left(\frac{s-w}{h}\right)^p \mathrm{d}s.$$

for $1 \leq p \leq \underline{\beta}$. Then we have the following bias bounds.

(i) If $p \leq \underline{\beta} - 1$, then $\sup_{w \in \mathcal{W}} |B_n(w) - h^p b_p(w)| \leq \frac{2C_{\mathrm{k}} C_{\mathrm{H}}}{(p+1)!} h^{p+1}$.

(ii) If $p = \underline{\beta}$, then $\sup_{w \in \mathcal{W}} |B_n(w) - h^p b_p(w)| \leq \frac{2C_{\mathrm{k}} C_{\mathrm{H}}}{\underline{\beta}!} h^{\beta}$.

(iii) If $p \geq \underline{\beta} + 1$, then $\sup_{w \in \mathcal{W}} |B_n(w)| \leq \frac{2C_{\mathrm{k}} C_{\mathrm{H}}}{\underline{\beta}!} h^{\beta}$.

Noting that $\sup_{\mathcal{W}} |b_p(w)| \leq 2C_{\mathrm{k}}C_{\mathrm{H}}/p!$*, we deduce that for* $h \leq 1$*,*

$$\sup_{w \in \mathcal{W}} |B_n(w)| \leq \frac{4C_{\mathrm{k}}C_{\mathrm{H}}}{(p \wedge \beta)!} h^{p \wedge \beta} \lesssim h^{p \wedge \beta}.$$

3.2.2 Hoeffding-type decomposition and degeneracy

Our next step is to consider the stochastic part $\hat{f}_W(w) - \mathbb{E}\big[\hat{f}_W(w)\big]$ of the classical bias–variance decomposition. This term is akin to a U-statistic and thus admits a Hoeffding-type decomposition, presented in Lemma 3.2.1, which is a key element in our analysis.

Lemma 3.2.1 (Hoeffding-type decomposition for $\hat{f}_W$)

Suppose that Assumptions 3.2.1 and 3.2.2 hold. Define the linear, quadratic, and error terms

$$L_n(w) = \frac{2}{n} \sum_{i=1}^{n} l_i(w), \qquad Q_n(w) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} q_{ij}(w),$$
$$E_n(w) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} e_{ij}(w)$$

respectively, where

$$\begin{aligned}
l_i(w) &= \mathbb{E}\left[k_h(W_{ij}, w) \mid A_i\right] - \mathbb{E}\left[k_h(W_{ij}, w)\right], \\
q_{ij}(w) &= \mathbb{E}\left[k_h(W_{ij}, w) \mid A_i, A_j\right] - \mathbb{E}\left[k_h(W_{ij}, w) \mid A_i\right] - \mathbb{E}\left[k_h(W_{ij}, w) \mid A_j\right] + \mathbb{E}\left[k_h(W_{ij}, w)\right], \\
e_{ij}(w) &= k_h(W_{ij}, w) - \mathbb{E}\left[k_h(W_{ij}, w) \mid A_i, A_j\right].
\end{aligned}$$

Then, recalling the bias term B_n *from Section 3.2.1, we have the Hoeffding-type decomposition*

$$\hat{f}_W(w) - f_W(w) = L_n(w) + Q_n(w) + E_n(w) + B_n(w). \tag{3.2}$$

The processes L_n*,* Q_n*, and* E_n *are mean-zero with* $\mathbb{E}\big[L_n(w)\big] = \mathbb{E}\big[Q_n(w)\big] = \mathbb{E}\big[E_n(w)\big] = 0$ *for all* $w \in \mathcal{W}$*. They are also orthogonal, satisfying* $\mathbb{E}\big[L_n(w)Q_n(w')\big] = \mathbb{E}\big[L_n(w)E_n(w')\big] = \mathbb{E}\big[Q_n(w)E_n(w')\big] = 0$ *for all* $w, w' \in \mathcal{W}$*.*

The process L_n is the Hájek projection of a U-process, which can exhibit degeneracy if $\mathrm{Var}[L_n(w)] = 0$ at some or all points $w \in \mathcal{W}$. To characterize the different possible degeneracy types in Lemma 3.2.2, we first introduce the following lower and upper degeneracy constants:

$$D_{\mathrm{lo}}^2 := \inf_{w \in \mathcal{W}} \mathrm{Var}\left[f_{W|A}(w \mid A_i)\right] \qquad \text{and} \qquad D_{\mathrm{up}}^2 := \sup_{w \in \mathcal{W}} \mathrm{Var}\left[f_{W|A}(w \mid A_i)\right].$$

Lemma 3.2.2 (Trichotomy of degeneracy)

Grant Assumptions 3.2.1 and 3.2.2. Then the type of degeneracy exhibited by $\hat{f}_W(w)$ is precisely one of the following three possibilities.

(i) Total degeneracy: $D_{\mathrm{up}} = D_{\mathrm{lo}} = 0$. Then $L_n(w) = 0$ for all $w \in \mathcal{W}$ almost surely.

(ii) No degeneracy: $D_{\mathrm{lo}} > 0$. Then $\inf_{w \in \mathcal{W}} \mathrm{Var}[L_n(w)] \geq \frac{2D_{\mathrm{lo}}}{n}$ for all large enough n.

(iii) Partial degeneracy: $D_{\mathrm{up}} > D_{\mathrm{lo}} = 0$. There exists $w \in \mathcal{W}$ with $\mathrm{Var}\left[f_{W|A}(w \mid A_i)\right] = 0$; such a point is labeled degenerate *and satisfies $\mathrm{Var}[L_n(w)] \leq 64 C_{\mathrm{k}} C_{\mathrm{H}} C_{\mathrm{d}} \frac{h}{n}$. There is also a point $w' \in \mathcal{W}$ with $\mathrm{Var}\left[f_{W|A}(w' \mid A_i)\right] > 0$; such a point is labeled* non-degenerate *and satisfies $\mathrm{Var}[L_n(w')] \geq \frac{2}{n} \mathrm{Var}\left[f_{W|A}(w' \mid A_i)\right]$ for all large enough n.*

The following lemma describes the uniform stochastic order of the different terms in the Hoeffding-type decomposition, explicitly accounting for potential degeneracy.

Lemma 3.2.3 (Uniform concentration)

Suppose Assumptions 3.2.1 and 3.2.2 hold. Then

$$\mathbb{E}\left[\sup_{w \in \mathcal{W}} |L_n(w)|\right] \lesssim \frac{D_{\mathrm{up}}}{\sqrt{n}}, \qquad \mathbb{E}\left[\sup_{w \in \mathcal{W}} |Q_n(w)|\right] \lesssim \frac{1}{n}, \qquad \mathbb{E}\left[\sup_{w \in \mathcal{W}} |E_n(w)|\right] \lesssim \sqrt{\frac{\log n}{n^2 h}}.$$

Lemma 3.2.3 captures the potential total degeneracy of L_n by illustrating how if $D_{\mathrm{up}} = 0$ then $L_n = 0$ everywhere on $\mathcal{W}$ almost surely. The following lemma captures the potential partial degeneracy of L_n, where $D_{\mathrm{up}} > D_{\mathrm{lo}} = 0$. For $w, w' \in \mathcal{W}$, define the covariance function

$$\Sigma_n(w, w') = \mathbb{E}\Big[\Big(\hat{f}_W(w) - \mathbb{E}\big[\hat{f}_W(w)\big]\Big)\Big(\hat{f}_W(w') - \mathbb{E}\big[\hat{f}_W(w')\big]\Big)\Big].$$

Lemma 3.2.4 (Variance bounds)

Suppose that Assumptions 3.2.1 and 3.2.2 hold. Then for sufficiently large n,

$$\frac{D_{\text{lo}}^2}{n} + \frac{1}{n^2 h} \inf_{w \in \mathcal{W}} f_W(w) \lesssim \inf_{w \in \mathcal{W}} \Sigma_n(w,w) \leq \sup_{w \in \mathcal{W}} \Sigma_n(w,w) \lesssim \frac{D_{\text{up}}^2}{n} + \frac{1}{n^2 h}.$$

As a simple example of the different types of degeneracy, consider the family of dyadic distributions $\mathbb{P}_\pi$ indexed by $\pi = (\pi_1, \pi_2, \pi_3)$ with $\sum_{i=1}^3 \pi_i = 1$ and $\pi_i \geq 0$, generated by $W_{ij} = A_i A_j + V_{ij}$, where A_i equals -1 with probability π_1, equals 0 with probability π_2 and equals $+1$ with probability π_3, and V_{ij} is standard Gaussian. This model induces a latent "community structure" where community membership is determined by the value of A_i for each node i, and the interaction outcome W_{ij} is a function only of the communities which i and j belong to and some idiosyncratic noise. Unlike the stochastic block model (Kolaczyk, 2009), our setup assumes that community membership has no impact on edge existence, as we work with fully connected networks; see Section 3.7.1 for a discussion of how to handle missing edges in practice. Also note that the parameter of interest in this chapter is the Lebesgue density of a continuous random variable W_{ij} rather than the probability of network edge existence, which is the focus of the graphon estimation literature (Gao and Ma, 2021).

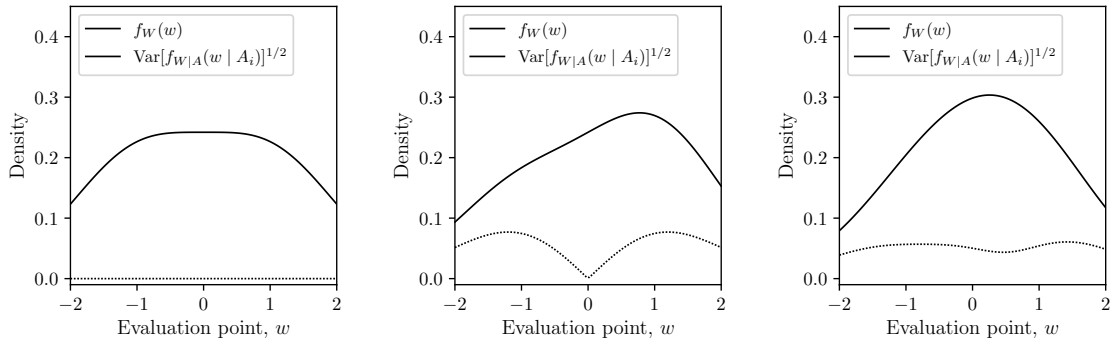
In line with Assumption 3.2.1, $\mathbf{A}_n$ and $\mathbf{V}_n$ are i.i.d. sequences independent of each other. Then $f_{W|AA}(w \mid A_i, A_j) = \phi(w - A_i A_j)$, $f_{W|A}(w \mid A_i) = \pi_1 \phi(w + A_i) + \pi_2 \phi(w) + \pi_3 \phi(w - A_i)$, and $f_W(w) = (\pi_1^2 + \pi_3^2)\phi(w-1) + \pi_2(2 - \pi_2)\phi(w) + 2\pi_1\pi_3\phi(w+1)$, where ϕ denotes the probability density function of the standard normal distribution. Note that $f_W(w)$ is strictly positive for all $w \in \mathbb{R}$. Consider the parameter choices:

(i) $\pi = \left(\frac{1}{2}, 0, \frac{1}{2}\right)$: $\mathbb{P}_\pi$ is degenerate at all $w \in \mathbb{R}$,

(ii) $\pi = \left(\frac{1}{4}, 0, \frac{3}{4}\right)$: $\mathbb{P}_\pi$ is degenerate only at $w = 0$,

(iii) $\pi = \left(\frac{1}{5}, \frac{1}{5}, \frac{3}{5}\right)$: $\mathbb{P}_\pi$ is non-degenerate for all $w \in \mathbb{R}$.

Figure 3.1 demonstrates these phenomena, plotting the density f_W and the standard deviation of the conditional density $f_{W|A}$ over $\mathcal{W} = [-2, 2]$ for each choice of the parameter π.



(a) Total degeneracy, $\pi = \left(\frac{1}{2}, 0, \frac{1}{2}\right)$.

(b) Partial degeneracy, $\pi = \left(\frac{1}{4}, 0, \frac{3}{4}\right)$.

(c) No degeneracy, $\pi = \left(\frac{1}{5}, \frac{1}{5}, \frac{3}{5}\right)$.

Figure 3.1: Density f_W and standard deviation of $f_{W|A}$ for the family of distributions $\mathbb{P}_\pi$.

The trichotomy of total/partial/no degeneracy is useful for understanding the distributional properties of the dyadic kernel density estimator $\hat{f}_W(w)$. Crucially, our need for uniformity in w complicates the simpler degeneracy/no degeneracy dichotomy observed previously in the literature (Graham et al., 2024). From a pointwise-in-w perspective, partial degeneracy causes no issues, while it is a fundamental problem when conducting inference uniformly over $w \in \mathcal{W}$. We develop methods that are valid regardless of the presence of partial or total degeneracy.

3.3 Point estimation results

Using the bias bound from Theorem 3.2.1 and the concentration results from Lemma 3.2.3, the next theorem establishes an upper bound on the uniform convergence rate of $\hat{f}_W$.

Theorem 3.3.1 (Uniform convergence rate)

Suppose that Assumptions 3.2.1 and 3.2.2 hold. Then

$$\mathbb{E}\left[\sup_{w \in \mathcal{W}} \left|\hat{f}_W(w) - f_W(w)\right|\right] \lesssim h^{p \wedge \beta} + \frac{D_{\mathrm{up}}}{\sqrt{n}} + \sqrt{\frac{\log n}{n^2 h}}.$$

The implicit constant in Theorem 3.3.1 depends only on $\mathcal{W}$, β, C_{H}, and the choice of kernel. We interpret this result in light of the degeneracy trichotomy from Lemma 3.2.2. These results generalize Chiang and Tan (2023, Theorem 1) by allowing for compactly supported data and more general kernels $k_h(\cdot, w)$, enabling boundary-adaptive estimation.

(i) Partial or no degeneracy: $D_{\text{up}} > 0$. Any bandwidths satisfying $n^{-1}\log n \lesssim h \lesssim n^{-\frac{1}{2(p\wedge\beta)}}$ yield $\mathbb{E}\big[\sup_{w\in\mathcal{W}}\big|\hat{f}_W(w) - f_W(w)\big|\big] \lesssim \frac{1}{\sqrt{n}}$, the "parametric" bandwidth-independent rate noted by Graham et al. (2024).

(ii) Total degeneracy: $D_{\text{up}} = 0$. Minimizing the bound in Theorem 3.3.1 with $h \asymp \left(\frac{\log n}{n^2}\right)^{\frac{1}{2(p\wedge\beta)+1}}$ yields $\mathbb{E}\big[\sup_{w\in\mathcal{W}}\big|\hat{f}_W(w) - f_W(w)\big|\big] \lesssim \left(\frac{\log n}{n^2}\right)^{\frac{p\wedge\beta}{2(p\wedge\beta)+1}}$.

3.3.1 Minimax optimality

We establish the minimax rate under the supremum norm for density estimation with dyadic data. This implies minimax optimality of the kernel density estimator $\hat{f}_W$, regardless of the degeneracy type of the dyadic distribution.

Theorem 3.3.2 (Uniform minimax optimality)

Fix $\beta \geq 1$ and $C_{\text{H}} > 0$, and take $\mathcal{W}$ a compact interval with positive Lebesgue measure. Define $\mathcal{P} = \mathcal{P}(\mathcal{W}, \beta, C_{\text{H}})$ as the class of dyadic distributions satisfying Assumption 3.2.1. Define $\mathcal{P}_{\text{d}}$ as the subclass of $\mathcal{P}$ containing only those distributions which are totally degenerate on $\mathcal{W}$ in the sense that $\sup_{w\in\mathcal{W}} \operatorname{Var}\big[f_{W|A}(w \mid A_i)\big] = 0$. Then

$$\inf_{\tilde{f}_W}\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\sup_{w\in\mathcal{W}}\big|\tilde{f}_W(w) - f_W(w)\big|\right] \asymp \frac{1}{\sqrt{n}},$$
$$\inf_{\tilde{f}_W}\sup_{\mathbb{P}\in\mathcal{P}_{\text{d}}} \mathbb{E}_{\mathbb{P}}\left[\sup_{w\in\mathcal{W}}\big|\tilde{f}_W(w) - f_W(w)\big|\right] \asymp \left(\frac{\log n}{n^2}\right)^{\frac{\beta}{2\beta+1}},$$

where $\tilde{f}_W$ is any estimator depending only on the data $\mathbf{W}_n = (W_{ij} : 1 \leq i < j \leq n)$ distributed according to the dyadic law $\mathbb{P}$. The constants in $\asymp$ depend only on $\mathcal{W}$, β, and C_{H}.

Theorem 3.3.2 shows that the uniform convergence rate of $n^{-1/2}$ obtained in Theorem 3.3.1 (coming from the L_n term) is minimax-optimal in general. When attention is restricted to totally degenerate dyadic distributions, $\hat{f}_W$ also achieves the minimax rate of uniform convergence (assuming a kernel of sufficiently high order $p \geq \beta$), which is on the order of $\left(\frac{\log n}{n^2}\right)^{\frac{\beta}{2\beta+1}}$ and is determined by the bias B_n and the leading variance term E_n in (3.2).

Combining Theorems 3.3.1 and 3.3.2, we conclude that $\hat{f}_W(w)$ achieves the minimax-optimal rate for uniformly estimating $f_W(w)$ if $h \asymp \left(\frac{\log n}{n^2}\right)^{\frac{1}{2\beta+1}}$ and a kernel of sufficiently high order ($p \geq \beta$) is used, whether or not there are any degenerate points in the underlying data generating process. This result appears to be new to the literature on nonparametric estimation with dyadic data. See Gao and Ma (2021) for a contemporaneous review.

3.4 Distributional results

We investigate the distributional properties of the standardized t-statistic process

$$T_n(w) = \frac{\hat{f}_W(w) - f_W(w)}{\sqrt{\Sigma_n(w,w)}},$$

which is not necessarily asymptotically tight. Therefore, to approximate the distribution of the entire t-statistic process, as well as specific functionals thereof, we rely on a novel strong approximation approach outlined in this section. Our results can be used to perform valid uniform inference irrespective of the degeneracy type.

This section is largely concerned with distributional properties and thus frequently requires copies of stochastic processes. For succinctness of notation, we will not differentiate between a process and its copy, but details are available in Section B.2.

3.4.1 Strong approximation

By the Hoeffding-type decomposition (3.2) and Lemma 3.2.3, it suffices to consider the distributional properties of the stochastic process $L_n + E_n$. Our approach combines the Kómlos–Major–Tusnády (KMT) approximation (Komlós et al., 1975) to obtain a strong approximation of L_n with a Yurinskii approximation (Yurinskii, 1978) to obtain a *conditional* (on $\mathbf{A}_n$) strong approximation of E_n. The latter is necessary because E_n is akin to a local empirical process of i.n.i.d. random variables, conditional on $\mathbf{A}_n$, and therefore the KMT approximation is not applicable. These approximations are then combined to give a final (unconditional) strong approximation for $L_n + E_n$, and thus for the t-statistic process T_n.

The following lemma is an application of our generic KMT approximation result for empirical processes, given in Section B.2, which builds on earlier work by Giné et al. (2004) and Giné and Nickl (2010) and may be of independent interest.

Lemma 3.4.1 (Strong approximation of L_n)

Suppose that Assumptions 3.2.1 and 3.2.2 hold. For each n there exists a mean-zero Gaussian process Z_n^L indexed on $\mathcal{W}$ satisfying $\mathbb{E}\big[\sup_{w\in\mathcal{W}}\big|\sqrt{n}L_n(w) - Z_n^L(w)\big|\big] \lesssim \frac{D_{\text{up}}\log n}{\sqrt{n}}$, where $\mathbb{E}[Z_n^L(w)Z_n^L(w')] = n\mathbb{E}[L_n(w)L_n(w')]$ for all $w, w' \in \mathcal{W}$. The process Z_n^L is a function only of $\mathbf{A}_n$ and some random noise independent of $(\mathbf{A}_n, \mathbf{V}_n)$.

The strong approximation result in Lemma 3.4.1 would be sufficient to develop valid and even optimal uniform inference procedures whenever both $D_{\text{lo}} > 0$ (no degeneracy in L_n) and $nh \gg \log n$ (L_n is leading). In this special case, the recent Donsker-type results of Davezies et al. (2021) can be applied to analyze the limiting distribution of the stochastic process $\hat{f}_W$. Alternatively, again only when L_n is non-degenerate and leading, standard empirical process methods could also be used. However, even in the special case when $\hat{f}_W(w)$ is asymptotically Donsker, our result in Lemma 3.4.1 improves upon the literature by providing a rate-optimal strong approximation for $\hat{f}_W$ as opposed to only a weak convergence result. See Theorem 3.4.2 and the subsequent discussion below.

More importantly, as illustrated above, it is common in the literature to find dyadic distributions which exhibit partial or total degeneracy, making the process $\hat{f}_W$ non-Donsker. Thus approximating only L_n is in general insufficient for valid uniform inference, and it is necessary to capture the distributional properties of E_n as well. The following lemma is an application of our strong approximation result for empirical processes based on the Yurinskii approximation, which builds on a refinement by Belloni et al. (2019).

Lemma 3.4.2 (Conditional strong approximation of E_n)

Suppose Assumptions 3.2.1 and 3.2.2 hold and take any $R_n \to \infty$. For each n there exists $\tilde{Z}_n^E$ a mean-zero Gaussian process conditional on $\mathbf{A}_n$ satisfying $\sup_{w\in\mathcal{W}}\big|\sqrt{n^2h}E_n(w) - \tilde{Z}_n^E(w)\big| \lesssim_{\mathbb{P}} \frac{(\log n)^{3/8}R_n}{n^{1/4}h^{3/8}}$, where $\mathbb{E}[\tilde{Z}_n^E(w)\tilde{Z}_n^E(w') \mid \mathbf{A}_n] = n^2h\mathbb{E}[E_n(w)E_n(w') \mid \mathbf{A}_n]$ for all $w, w' \in \mathcal{W}$.

The process $\tilde{Z}_n^E$ is a Gaussian process conditional on $\mathbf{A}_n$ but is not in general a Gaussian process unconditionally. The following lemma constructs an unconditional Gaussian process Z_n^E that approximates $\tilde{Z}_n^E$.

Lemma 3.4.3 (Unconditional strong approximation of E_n)
Suppose that Assumptions 3.2.1 and 3.2.2 hold. For each n there exists a mean-zero Gaussian process Z_n^E satisfying $\mathbb{E}\big[\sup_{w\in\mathcal{W}}\big|\tilde{Z}_n^E(w) - Z_n^E(w)\big|\big] \lesssim \frac{(\log n)^{2/3}}{n^{1/6}}$, where Z_n^E is independent of $\mathbf{A}_n$ and $\mathbb{E}[Z_n^E(w)Z_n^E(w')] = \mathbb{E}[\tilde{Z}_n^E(w)\tilde{Z}_n^E(w')] = n^2h\,\mathbb{E}[E_n(w)E_n(w')]$ for all $w, w' \in \mathcal{W}$.

Combining Lemmas 3.4.2 and 3.4.3, we obtain an unconditional strong approximation for E_n. The resulting rate of approximation may not be optimal, due to the Yurinskii coupling, but to the best of our knowledge it is the first in the literature for the process E_n, and hence for $\hat{f}_W$ and its associated t-process in the context of dyadic data. The approximation rate is sufficiently fast to allow for optimal bandwidth choices; see Section 3.5 for more details. Strong approximation results for local empirical processes (e.g. Giné and Nickl, 2010) are not applicable here because the summands in the non-negligible E_n are not (conditionally) i.i.d. Likewise, neither standard empirical process and U-process theory (van der Vaart and Wellner, 1996; Giné and Nickl, 2021) nor the recent results in Davezies et al. (2021) are applicable to the non-Donsker process E_n.

The previous lemmas showed that L_n is $\sqrt{n}$-consistent while E_n is $\sqrt{n^2h}$-consistent (pointwise in w), showcasing the importance of careful standardization (cf. Studentization in Section 3.5) for the purpose of rate adaptivity to the unknown degeneracy type. In other words, a challenge in conducting uniform inference is that the finite-dimensional distributions of the stochastic process $L_n + E_n$, and hence those of $\hat{f}_W$ and its associated t-process T_n, may converge at different rates at different points $w \in \mathcal{W}$. The following theorem provides an (infeasible) inference procedure which is fully adaptive to such potential unknown degeneracy.

Theorem 3.4.1 (Strong approximation of T_n)

Suppose that Assumptions 3.2.1 and 3.2.2 hold and $f_W(w) > 0$ on $\mathcal{W}$, and take any $R_n \to \infty$. Then for each n there exists a centered Gaussian process Z_n^T such that

$$\sup_{w\in\mathcal{W}} \left|T_n(w) - Z_n^T(w)\right| \lesssim_{\mathbb{P}} \frac{n^{-1}\log n + n^{-5/4}h^{-7/8}(\log n)^{3/8}R_n + n^{-7/6}h^{-1/2}(\log n)^{2/3} + h^{p\wedge\beta}}{D_{\text{lo}}/\sqrt{n} + 1/\sqrt{n^2h}},$$

where $\mathbb{E}[Z_n^T(w)Z_n^T(w')] = \mathbb{E}[T_n(w)T_n(w')]$ for all $w, w' \in \mathcal{W}$.

The first term in the numerator corresponds to the strong approximation for L_n in Lemma 3.4.1 and the error introduced by Q_n. The second and third terms correspond to the conditional and unconditional strong approximation errors for E_n in Lemmas 3.4.2 and 3.4.3. The fourth term is from the smoothing bias result in Theorem 3.2.1. The denominator is the lower bound on the standard deviation $\Sigma_n(w,w)^{1/2}$ formulated in Lemma 3.2.4.

In the absence of degenerate points ($D_{\text{lo}} > 0$) and if $nh^{7/2} \gtrsim 1$, Theorem 3.4.1 offers a strong approximation of the t-process at the rate $(\log n)/\sqrt{n} + \sqrt{n}h^{p\wedge\beta}$, which matches the celebrated KMT approximation rate for i.i.d. data plus the smoothing bias. Therefore, our novel t-process strong approximation can achieve the optimal KMT rate for non-degenerate dyadic distributions provided that $p \wedge \beta \geq 3.5$. This is achievable if a fourth-order (boundary-adaptive) kernel is used and f_W is sufficiently smooth.

In the presence of partial or total degeneracy ($D_{\text{lo}} = 0$), Theorem 3.4.1 provides a strong approximation for the t-process at the rate $\sqrt{h}\log n + n^{-1/4}h^{-3/8}(\log n)^{3/8}R_n + n^{-1/6}(\log n)^{2/3} + nh^{1/2+p\wedge\beta}$. If, for example, $nh^{p\wedge\beta} \lesssim 1$, then our result can achieve a strong approximation rate of $n^{-1/7}$ up to $\log n$ terms. Theorem 3.4.1 appears to be the first in the dyadic literature which is also robust to the presence of degenerate points in the underlying dyadic distribution.

3.4.2 Application: confidence bands

Theorem 3.4.2 constructs standardized confidence bands for f_W which are infeasible as they depend on the unknown population variance Σ_n. In Section 3.5 we will make this inference procedure feasible by proposing a valid estimator of the covariance function Σ_n for Studentization, as well as developing bandwidth selection and robust bias correction methods.

Before presenting our result on valid infeasible uniform confidence bands, we first impose in Assumption 3.4.1 some extra restrictions on the bandwidth sequence, which depend on the degeneracy type of the dyadic distribution, to ensure the coverage rate converges.

Assumption 3.4.1 (Rate restriction for uniform confidence bands)
Assume that one of the following holds:

(i) No degeneracy ($D_{\text{lo}} > 0$): $n^{-6/7} \log n \ll h \ll (n \log n)^{-\frac{1}{2(p \wedge \beta)}}$,

(ii) Partial or total degeneracy ($D_{\text{lo}} = 0$): $n^{-2/3} (\log n)^{7/3} \ll h \ll (n^2 \log n)^{-\frac{1}{2(p \wedge \beta)+1}}$.

We now construct the infeasible uniform confidence bands. For $\alpha \in (0,1)$, let $q_{1-\alpha}$ be the quantile satisfying $\mathbb{P}\left(\sup_{w \in \mathcal{W}} \left|Z_n^T(w)\right| \leq q_{1-\alpha}\right) = 1 - \alpha$. The following result employs the anti-concentration idea due to Chernozhukov et al. (2014a) to deduce valid standardized confidence bands, where we approximate the quantile of the unknown finite sample distribution of $\sup_{w \in \mathcal{W}} |T_n(w)|$ by the quantile $q_{1-\alpha}$ of $\sup_{w \in \mathcal{W}} |Z_n^T(w)|$. This approach offers a better rate of convergence than relying on extreme value theory for the distributional approximation, hence improving the finite sample performance of the proposed confidence bands.

Theorem 3.4.2 (Infeasible uniform confidence bands)
Suppose that Assumptions 3.2.1, 3.2.2, and 3.4.1 hold and $f_W(w) > 0$ on $\mathcal{W}$. Then

$$\mathbb{P}\left(f_W(w) \in \left[\hat{f}_W(w) \pm q_{1-\alpha} \sqrt{\Sigma_n(w,w)}\right] \text{ for all } w \in \mathcal{W}\right) \to 1 - \alpha.$$

By Theorem 3.3.1, the asymptotically optimal choice of bandwidth for uniform convergence is $h \asymp ((\log n)/n^2)^{\frac{1}{2(p \wedge \beta)+1}}$. As discussed in the next section, the approximate IMSE-optimal bandwidth is $h \asymp (1/n^2)^{\frac{1}{2(p \wedge \beta)+1}}$. Both bandwidth choices satisfy Assumption 3.4.1 only in the case of no degeneracy. The degenerate cases in Assumption 3.4.1(ii), which require $p \wedge \beta > 1$, exhibit behavior more similar to that of standard nonparametric kernel-based estimation and so the aforementioned optimal bandwidth choices will lead to a non-negligible smoothing bias in the distributional approximation for T_n. Different approaches are available in the literature to address this issue, including undersmoothing or ignoring the bias (Hall and Kang, 2001), bias correction (Hall, 1992), robust bias correction (Calonico et al., 2018, 2022), and Lepskii's

method (Lepskii, 1992; Birgé, 2001), among others. In the next section we develop a feasible uniform inference procedure, based on robust bias correction methods, which amounts to first selecting an optimal bandwidth for the point estimator $\hat{f}_W$ using a pth-order kernel, and then correcting the bias of the point estimator while also adjusting the standardization (Studentization) when forming the t-statistic T_n.

Importantly, regardless of the specific implementation details, Theorem 3.4.2 shows that any bandwidth sequence h satisfying both (i) and (ii) in Assumption 3.4.1 leads to valid uniform inference which is robust and adaptive to the (unknown) degeneracy type.

3.5 Implementation

We address outstanding implementation details to make our main uniform inference results feasible. In Section 3.5.1 we propose a covariance estimator along with a modified version which is guaranteed to be positive semi-definite. This allows for the construction of fully feasible confidence bands in Section 3.5.2. In Section 3.5.3 we discuss bandwidth selection and formalize our procedure for robust bias correction inference.

3.5.1 Covariance function estimation

Define the following plug-in covariance function estimator of Σ_n. For $w, w' \in \mathcal{W}$, let $S_i(w) = \frac{1}{n-1}\big(\sum_{j=1}^{i-1} k_h(W_{ji}, w) + \sum_{j=i+1}^{n} k_h(W_{ij}, w)\big)$ estimate $\mathbb{E}[k_h(W_{ij}, w) \mid A_i]$ and take

$$\begin{aligned}\hat{\Sigma}_n(w, w') = \frac{4}{n^2}\sum_{i=1}^{n} S_i(w)S_i(w') - \frac{4}{n^2(n-1)^2}\sum_{i<j} k_h(W_{ij}, w)k_h(W_{ij}, w') \\ - \frac{4n-6}{n(n-1)}\hat{f}_W(w)\hat{f}_W(w').\end{aligned}$$

Though $\hat{\Sigma}_n(w, w')$ is consistent in an appropriate sense as shown in Lemma 3.5.1, it is not necessarily positive semi-definite, even in the limit. We therefore propose a modified covariance estimator which is guaranteed to be positive semi-definite. Specifically, consider the following

optimization problem where C_{k} and C_{L} are as in Section 3.2.

$$\begin{aligned} \text{minimize:} \quad & \sup_{w,w' \in \mathcal{W}} \left| \frac{M(w,w') - \hat{\Sigma}_n(w,w')}{\sqrt{\hat{\Sigma}_n(w,w) + \hat{\Sigma}_n(w',w')}} \right| \quad \text{over } M : \mathcal{W} \times \mathcal{W} \to \mathbb{R} \\ \text{subject to:} \quad & M \text{ is symmetric and positive semi-definite,} \\ & \left| M(w,w') - M(w,w'') \right| \leq \frac{4}{nh^3} C_{\mathrm{k}} C_{\mathrm{L}} |w' - w''| \text{ for all } w, w', w'' \in \mathcal{W}. \end{aligned} \tag{3.3}$$

Denote by $\hat{\Sigma}_n^+$ any (approximately) optimal solution to (3.3). The following lemma establishes uniform convergence rates for both $\hat{\Sigma}_n$ and $\hat{\Sigma}_n^+$. We then use $\hat{\Sigma}_n^+$ to construct feasible versions of T_n and its associated Gaussian approximation Z_n^T defined in Theorem 3.4.1.

Lemma 3.5.1 (Consistency of $\hat{\Sigma}_n$ and $\hat{\Sigma}_n^+$)

Suppose Assumptions 3.2.1 and 3.2.2 hold and that $nh \gtrsim \log n$ *and* $f_W(w) > 0$ *on* $\mathcal{W}$. *Then*

$$\sup_{w,w' \in \mathcal{W}} \left| \frac{\hat{\Sigma}_n(w,w') - \Sigma_n(w,w')}{\sqrt{\Sigma_n(w,w) + \Sigma_n(w',w')}} \right| \lesssim_{\mathbb{P}} \frac{\sqrt{\log n}}{n}.$$

The optimization problem (3.3) *is a semi-definite program (SDP, Laurent and Rendl, 2005) and has an approximately optimal solution* $\hat{\Sigma}_n^+$ *satisfying*

$$\sup_{w,w' \in \mathcal{W}} \left| \frac{\hat{\Sigma}_n^+(w,w') - \Sigma_n(w,w')}{\sqrt{\Sigma_n(w,w) + \Sigma_n(w',w')}} \right| \lesssim_{\mathbb{P}} \frac{\sqrt{\log n}}{n}.$$

In practice we take $w, w' \in \mathcal{W}_d$ where $\mathcal{W}_d$ is a finite subset of $\mathcal{W}$, typically taken to be an equally-spaced grid. This yields finite-dimensional covariance matrices, for which (3.3) can be solved in polynomial time in $|\mathcal{W}_d|$ using a general-purpose SDP solver (e.g. by interior point methods, Laurent and Rendl, 2005). The number of points in $\mathcal{W}_d$ should be taken as large as is computationally practical in order to generate confidence bands rather than merely simultaneous confidence intervals. It is worth noting that the complexity of solving (3.3) does not depend on the number of vertices n, and so does not influence the ability of our methodology to handle large and possibly sparse networks.

The bias-corrected variance estimator in Matsushita and Otsu (2021, Section 3.2) takes a similar form to our estimator $\hat{\Sigma}_n$ but in the parametric setting, and is therefore also not guaranteed to be positive semi-definite in finite samples. Our approach addresses this issue, ensuring a positive semi-definite estimator $\hat{\Sigma}_n^+$ is always available.

3.5.2 Feasible confidence bands

Given a choice of the kernel order p and a bandwidth h, we construct a valid confidence band that is implementable in practice. Define the Studentized t-statistic process

$$\hat{T}_n(w) = \frac{\hat{f}_W(w) - f_W(w)}{\sqrt{\hat{\Sigma}_n^+(w,w)}}.$$

Let $\hat{Z}_n^T(w)$ be a process which, conditional on the data $\mathbf{W}_n$, is mean-zero and Gaussian, whose conditional covariance structure is $\mathbb{E}\big[\hat{Z}_n^T(w)\hat{Z}_n^T(w') \;\big|\; \mathbf{W}_n\big] = \frac{\hat{\Sigma}_n^+(w,w')}{\sqrt{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')}}$. For $\alpha \in (0,1)$, let $\hat{q}_{1-\alpha}$ be the conditional quantile satisfying $\mathbb{P}\big(\sup_{w\in\mathcal{W}} \big|\hat{Z}_n^T(w)\big| \leq \hat{q}_{1-\alpha} \;\big|\; \mathbf{W}_n\big) = 1-\alpha$, which is shown to be well defined in Section B.3.

Theorem 3.5.1 (Feasible uniform confidence bands)

Suppose that Assumptions 3.2.1, 3.2.2, and 3.4.1 hold and $f_W(w) > 0$ on $\mathcal{W}$. Then

$$\mathbb{P}\left(f_W(w) \in \left[\hat{f}_W(w) \pm \hat{q}_{1-\alpha}\sqrt{\hat{\Sigma}_n^+(w,w)}\right] \text{ for all } w \in \mathcal{W}\right) \to 1-\alpha.$$

Recently, Chiang et al. (2023) derived high-dimensional central limit theorems over rectangles for exchangeable arrays and applied them to construct simultaneous confidence intervals for a sequence of design points. Their inference procedure relies on the multiplier bootstrap, and their conditions for valid inference depend on the number of design points considered. In contrast, Theorem 3.5.1 constructs a feasible uniform confidence band over the entire domain of inference $\mathcal{W}$ based on our strong approximation results for the whole t-statistic process and the covariance estimator $\hat{\Sigma}_n^+$. The required rate condition specified in

Assumption 3.4.1 does not depend on the number of design points. Furthermore, our proposed inference methods are robust to potential unknown degenerate points in the underlying dyadic data generating process.

In practice, suprema over $\mathcal{W}$ can be replaced by maxima over sufficiently many design points in $\mathcal{W}$. The conditional quantile $\hat{q}_{1-\alpha}$ can be estimated by Monte Carlo simulation, resampling from the Gaussian process defined by the law of $\hat{Z}_n^T \mid \mathbf{W}_n$.

The bandwidth restrictions in Theorem 3.5.1 are the same as those for the infeasible version given in Theorem 3.4.2, namely those imposed in Assumption 3.4.1. This follows from the rates of convergence obtained in Lemma 3.5.1, coupled with some careful technical work given in Section B.3 to handle the potential presence of degenerate points in Σ_n.

3.5.3 Bandwidth selection and robust bias-corrected inference

We give practical suggestions for selecting the bandwidth parameter h. Let $\nu(w)$ be a non-negative real-valued function on $\mathcal{W}$ and suppose we use a kernel of order $p < \beta$ of the form $k_h(s,w) = K\big((s-w)/h\big)/h$. The ν-weighted asymptotic IMSE (AIMSE) is minimized by

$$h^*_{\mathrm{AIMSE}} = \left(\frac{p!(p-1)!\Big(\int_{\mathcal{W}} f_W(w)\nu(w)\,\mathrm{d}w\Big)\Big(\int_{\mathbb{R}} K(w)^2\,\mathrm{d}w\Big)}{2\Big(\int_{\mathcal{W}} f_W^{(p)}(w)^2\nu(w)\,\mathrm{d}w\Big)\Big(\int_{\mathbb{R}} w^p K(w)\,\mathrm{d}w\Big)^2} \right)^{\frac{1}{2p+1}} \left(\frac{n(n-1)}{2} \right)^{-\frac{1}{2p+1}}.$$

This is akin to the AIMSE-optimal bandwidth choice for traditional monadic kernel density estimation with a sample size of $\frac{1}{2}n(n-1)$. The choice h^*_{AIMSE} is slightly undersmoothed (up to a polynomial $\log n$ factor) relative to the uniform minimax-optimal bandwidth choice discussed in Section 3.3, but it is easier to implement in practice.

To implement the AIMSE-optimal bandwidth choice, we propose a simple rule-of-thumb (ROT) approach based on Silverman's rule. Suppose $p \wedge \beta = 2$ and let $\hat{\sigma}^2$ and $\hat{I}$ be the sample variance and sample interquartile range respectively of the data $\mathbf{W}_n$. Then $\hat{h}_{\mathrm{ROT}} = C(K)\big(\hat{\sigma} \wedge \frac{\hat{I}}{1.349}\big)\big(\frac{n(n-1)}{2}\big)^{-1/5}$, where we have $C(K) = 2.576$ for the triangular kernel $K(w) = (1-|w|)\vee 0$, and $C(K) = 2.435$ for the Epanechnikov kernel $K(w) = \frac{3}{4}(1-w^2)\vee 0$.

The AIMSE-optimal bandwidth selector $h^*_{\mathrm{AIMSE}} \asymp n^{-\frac{2}{2p+1}}$ and any of its feasible estimators only satisfy Assumption 3.4.1 in the case of no degeneracy ($D_{\mathrm{lo}} > 0$). Under partial or total degeneracy, such bandwidths are not valid due to the usual leading smoothing (or misspecification) bias of the distributional approximation. To circumvent this problem and construct feasible uniform confidence bands for f_W, we employ the following robust bias correction approach.

Firstly, estimate the bandwidth $h^*_{\mathrm{AIMSE}} \asymp n^{-\frac{2}{2p+1}}$ using a kernel of order p, which leads to an AIMSE-optimal point estimator $\hat{f}_W$ in an $L^2(\nu)$ sense. Then use this bandwidth and a kernel of order $p' > p$ to construct the statistic $\hat{T}_n$ and the confidence band as detailed in Section 3.5.2. Importantly, both $\hat{f}_W$ and $\hat{\Sigma}_n^+$ are recomputed with the new higher-order kernel. The change in centering is equivalent to a bias correction of the original AIMSE-optimal point estimator, while the change in scale captures the additional variability introduced by the bias correction itself. As shown formally in Calonico et al. (2018, 2022) for the case of kernel-based density estimation with i.i.d. data, this approach leads to higher-order refinements in the distributional approximation whenever additional smoothness is available ($p' \leq \beta$). In the present dyadic setting, this procedure is valid so long as $n^{-2/3}(\log n)^{7/3} \ll n^{-\frac{2}{2p+1}} \ll (n^2 \log n)^{-\frac{1}{2p'+1}}$, which is equivalent to $2 \leq p < p'$. For concreteness, we recommend taking $p = 2$ and $p' = 4$, and using the rule-of-thumb bandwidth choice $\hat{h}_{\mathrm{ROT}}$ defined above. In particular, this

Algorithm 1: Feasible uniform confidence bands

1 Choose a kernel k_h of order $p \geq 2$ satisfying Assumption 3.2.2.
2 Select a bandwidth $h \approx h^*_{\mathrm{AIMSE}}$ for k_h as in Section 3.5.3, perhaps using $h = \hat{h}_{\mathrm{ROT}}$.
3 Choose another kernel k'_h of order $p' > p$ satisfying Assumption 3.2.2.
4 For $d \geq 1$, choose a set of d distinct evaluation points $\mathcal{W}_d$.
5 For each $w \in \mathcal{W}_d$, construct the density estimate $\hat{f}_W(w)$ using k'_h as in Section 3.1.
6 For $w, w' \in \mathcal{W}_d$, estimate the covariance $\hat{\Sigma}_n(w, w')$ using k'_h as in Section 3.5.1.
7 Construct positive semi-definite covariance estimate $\hat{\Sigma}_n^+$ as in Section 3.5.1.
8 For $B \geq 1$, let $(\hat{Z}^T_{n,r} : 1 \leq r \leq B)$ be i.i.d. from $\hat{Z}^T_n$ as in Section 3.5.2.
9 For $\alpha \in (0, 1)$, set $\hat{q}_{1-\alpha} = \inf_{q \in \mathbb{R}}\{q : \#\{r : \max_{w \in \mathcal{W}_d} |\hat{Z}^T_{n,r}(w)| \leq q\} \geq B(1-\alpha)\}$.
10 Construct $\left[\hat{f}_W(w) \pm \hat{q}_{1-\alpha}\hat{\Sigma}_n^+(w, w)^{1/2}\right]$ for each $w \in \mathcal{W}_d$.

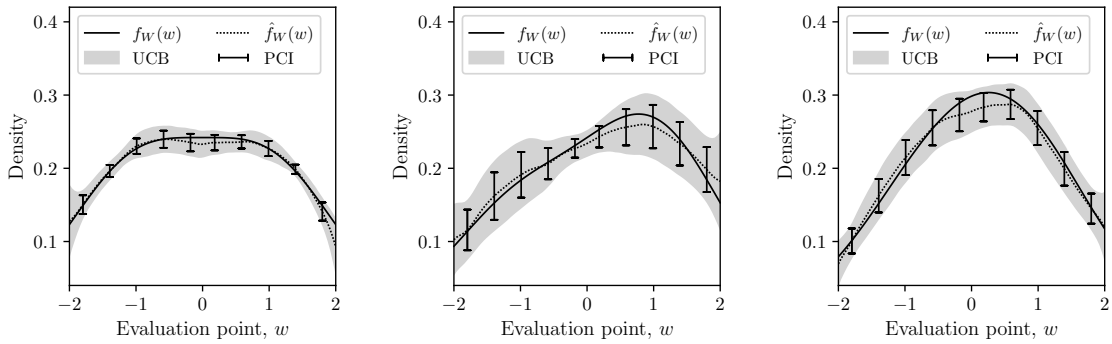
approach automatically delivers a KMT-optimal strong approximation whenever there are no degeneracies in the underlying dyadic data generating process. Our feasible robust bias correction method based on AIMSE-optimal dyadic kernel density estimation for constructing uniform confidence bands for f_W is summarized in Algorithm 1.

3.6 Simulations

We investigate the empirical finite-sample performance of the kernel density estimator with dyadic data using simulations. The family of dyadic distributions defined in Section 3.2.2, with its three parameterizations, is used to generate data sets with different degeneracy types.

We use two different boundary bias-corrected Epanechnikov kernels of orders $p = 2$ and $p = 4$ respectively, on the inference domain $\mathcal{W} = [-2, 2]$. We select an optimal bandwidth for $p = 2$ as recommended in Section 3.5.3, using the rule-of-thumb with $C(K) = 2.435$. The semi-definite program in Section 3.5.1 is solved with the MOSEK interior point optimizer (MOSEK ApS, 2021), ensuring positive semi-definite covariance estimates. Gaussian vectors are resampled $B = 10\,000$ times.

In Figure 3.2 we plot a typical outcome for each of the three degeneracy types (total, partial, none), using the Epanechnikov kernel of order $p = 2$, with sample size $n = 100$ (so $N = 4950$ pairs of nodes) and with $d = 100$ equally-spaced evaluation points. Each plot contains the true density function f_W, the dyadic kernel density estimate $\hat{f}_W$ and two



(a) Total degeneracy, $\pi = \left(\frac{1}{2}, 0, \frac{1}{2}\right)$.

(b) Partial degeneracy, $\pi = \left(\frac{1}{4}, 0, \frac{3}{4}\right)$.

(c) No degeneracy, $\pi = \left(\frac{1}{5}, \frac{1}{5}, \frac{3}{5}\right)$.

Figure 3.2: Typical outcomes for three different values of the parameter π.

different approximate 95% confidence bands for f_W. The first is the uniform confidence band (UCB) constructed using one of our main results, Theorem 3.5.1. The second is a sequence of pointwise confidence intervals (PCI) constructed by finding a confidence interval for each evaluation point separately. We show only 10 pointwise confidence intervals for clarity. In general, the PCIs are too narrow as they fail to provide simultaneous (uniform) coverage over the evaluation points. Note that under partial degeneracy the confidence band narrows near the degenerate point $w = 0$.

Next, Table 3.3 presents numerical results. For each degeneracy type (total, partial, none) and each kernel order ($p = 2$, $p = 4$), we run 2000 repeats with sample size $n = 3000$ (giving $N = 4\,498\,500$ pairs of nodes) and with $d = 50$ equally-spaced evaluation points. We record the average rule-of-thumb bandwidth $\hat{h}_{\mathrm{ROT}}$ and the average root integrated mean squared error (RIMSE). For both the uniform confidence bands (UCB) and the pointwise confidence intervals (PCI), we report the coverage rate (CR) and the average width (AW). The lower-order kernel ($p = 2$) ignores the bias, leading to good RIMSE performance and acceptable UCB coverage under partial or no degeneracy, but gives invalid inference under total degeneracy. In contrast, the higher-order kernel ($p = 4$) provides robust bias correction and hence improves the coverage of the UCB in every regime, particularly under total degeneracy, at the cost of increasing both the RIMSE and the average widths of the confidence bands. As expected, the pointwise (in $w \in \mathcal{W}$) confidence intervals (PCIs) severely undercover in every regime. Thus

π	Degeneracy type	$\hat{h}_{\mathrm{ROT}}$	p	RIMSE	UCB		PCI	
					CR	AW	CR	AW
$\left(\frac{1}{2}, 0, \frac{1}{2}\right)$	Total	0.161	2	0.00048	87.1%	0.0028	6.5%	0.0017
			4	0.00068	95.2%	0.0042	9.7%	0.0025
$\left(\frac{1}{4}, 0, \frac{3}{4}\right)$	Partial	0.158	2	0.00228	94.5%	0.0112	75.6%	0.0083
			4	0.00234	94.7%	0.0124	65.3%	0.0087
$\left(\frac{1}{5}, \frac{1}{5}, \frac{3}{5}\right)$	None	0.145	2	0.00201	94.2%	0.0106	73.4%	0.0077
			4	0.00202	95.6%	0.0117	64.3%	0.0080

Table 3.3: Numerical results for three values of the parameter π.

our simulation results show that the proposed feasible inference methods based on robust bias correction and proper Studentization deliver valid uniform inference which is robust to unknown degenerate points in the underlying dyadic distribution.

3.7 Counterfactual dyadic density estimation

To further showcase the applicability of our main results, we develop a kernel density estimator for dyadic counterfactual distributions. The aim of such counterfactual analysis is to estimate the distribution of an outcome variable had some covariates followed a distribution different from the actual one, and it is important in causal inference and program evaluation settings (DiNardo et al., 1996; Chernozhukov et al., 2013b).

For each $r \in \{0, 1\}$, let $\mathbf{W}_n^r$, $\mathbf{A}_n^r$, and $\mathbf{V}_n^r$ be random variables as defined in Assumption 3.2.1 and $\mathbf{X}_n^r = (X_1^r, \dots, X_n^r)$ be some covariates. We assume that (A_i^r, X_i^r) are independent over $1 \le i \le n$ and that $\mathbf{X}_n^r$ is independent of $\mathbf{V}_n^r$, that $W_{ij}^r \mid X_i^r, X_j^r$ has a conditional Lebesgue density $f_{W|XX}^r(\cdot \mid x_1, x_2) \in \mathcal{H}_{C_{\mathrm{H}}}^{\beta}(\mathcal{W})$, that X_i^r follows a distribution function F_X^r on a common support $\mathcal{X}$, and that $(\mathbf{A}_n^0, \mathbf{V}_n^0, \mathbf{X}_n^0)$ is independent of $(\mathbf{A}_n^1, \mathbf{V}_n^1, \mathbf{X}_n^1)$.

We interpret r as an index for two populations, labeled 0 and 1. The counterfactual density of population 1 had it followed the same covariate distribution as population 0 is

$$\begin{aligned} f_W^{1\triangleright 0}(w) &= \mathbb{E}\left[f_{W|XX}^1\big(w \mid X_1^0, X_2^0\big) \right] \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} f_{W|XX}^1(w \mid x_1, x_2) \psi(x_1) \psi(x_2) \, \mathrm{d}F_X^1(x_1) \, \mathrm{d}F_X^1(x_2), \end{aligned}$$

where $\psi(x) = \mathrm{d}F_X^0(x)/\mathrm{d}F_X^1(x)$ for $x \in \mathcal{X}$ is a Radon–Nikodym derivative. If X_i^0 and X_i^1 have Lebesgue densities, it is natural to consider a parametric model of the form $\mathrm{d}F_X^r(x) = f_X^r(x; \theta) \, \mathrm{d}x$ for some finite-dimensional parameter θ. Alternatively, if the covariates X_n^r are discrete and have a positive probability mass function $p_X^r(x)$ on a finite support $\mathcal{X}$, the object of interest becomes $f_W^{1\triangleright 0}(w) = \sum_{x_1 \in \mathcal{X}} \sum_{x_2 \in \mathcal{X}} f_{W|XX}^1(w \mid x_1, x_2) \psi(x_1) \psi(x_2) p_X^1(x_1) p_X^1(x_2)$, where $\psi(x) = p_X^0(x)/p_X^1(x)$ for $x \in \mathcal{X}$. We consider discrete covariates for simplicity, and

hence the counterfactual dyadic kernel density estimator is

$$\hat{f}_W^{1\triangleright 0}(w) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \hat{\psi}(X_i^1)\hat{\psi}(X_j^1) k_h(W_{ij}^1, w),$$

where $\hat{\psi}(x) = \hat{p}_X^0(x)/\hat{p}_X^1(x)$ and $\hat{p}_X^r(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{X_i^r = x\}$, with $\mathbb{I}$ the indicator function.

Section B.3.3 provides technical details: we show how an asymptotic linear representation for $\hat{\psi}(x)$ leads to a Hoeffding-type decomposition of $\hat{f}_W^{1\triangleright 0}(w)$, which is then used to establish that $\hat{f}_W^{1\triangleright 0}$ is uniformly consistent for $f_W^{1\triangleright 0}(w)$ and also admits a Gaussian strong approximation, with the same rates of convergence as for the standard density estimator. Furthermore, define the covariance function of $\hat{f}_W^{1\triangleright 0}(w)$ as $\Sigma_n^{1\triangleright 0}(w,w') = \mathrm{Cov}\big[\hat{f}_W^{1\triangleright 0}(w), \hat{f}_W^{1\triangleright 0}(w')\big]$, which can be estimated as follows. First let $\hat{\kappa}(X_i^0, X_i^1, x) = \frac{\mathbb{I}\{X_i^0 = x\} - \hat{p}_X^0(x)}{\hat{p}_X^1(x)} - \frac{\hat{p}_X^0(x)}{\hat{p}_X^1(x)} \frac{\mathbb{I}\{X_i^1 = x\} - \hat{p}_X^1(x)}{\hat{p}_X^1(x)}$ be a plug-in estimate of the influence function for $\hat{\psi}(x)$ and define the leave-one-out conditional expectation estimators $S_i^{1\triangleright 0}(w) = \frac{1}{n-1}\big(\sum_{j=1}^{i-1} k_h(W_{ji}^1, w)\hat{\psi}(X_j^1) + \sum_{j=i+1}^{n} k_h(W_{ij}^1, w)\hat{\psi}(X_j^1)\big)$ and $\tilde{S}_i^{1\triangleright 0}(w) = \frac{1}{n-1}\sum_{j=1}^n \mathbb{I}\{j \neq i\}\hat{\kappa}(X_i^0, X_i^1, X_j^1) S_j^{1\triangleright 0}(w)$. Define the covariance estimator

$$\begin{aligned}\hat{\Sigma}_n^{1\triangleright 0}(w,w') = {} & \frac{4}{n^2}\sum_{i=1}^{n} \big(\hat{\psi}(X_i^1)S_i^{1\triangleright 0}(w) + \tilde{S}_i^{1\triangleright 0}(w)\big)\big(\hat{\psi}(X_i^1)S_i^{1\triangleright 0}(w') + \tilde{S}_i^{1\triangleright 0}(w')\big) \\ & - \frac{4}{n^3(n-1)}\sum_{i<j} k_h(W_{ij}^1, w)k_h(W_{ij}^1, w')\hat{\psi}(X_i^1)^2\hat{\psi}(X_j^1)^2 - \frac{4}{n}\hat{f}_W^{1\triangleright 0}(w)\hat{f}_W^{1\triangleright 0}(w').\end{aligned}$$

We use a positive semi-definite approximation to $\hat{\Sigma}_n^{1\triangleright 0}$, denoted by $\hat{\Sigma}_n^{+,1\triangleright 0}$, as in Section 3.5.1. To construct feasible uniform confidence bands, define a process $\hat{Z}_n^{T,1\triangleright 0}(w)$ which is conditionally mean-zero and Gaussian given the data $\mathbf{W}_n^1$, $\mathbf{X}_n^0$, and $\mathbf{X}_n^1$, and whose conditional covariance structure is $\mathbb{E}\big[\hat{Z}_n^{T,1\triangleright 0}(w)\hat{Z}_n^{T,1\triangleright 0}(w') \mid \mathbf{W}_n^1, \mathbf{X}_n^0, \mathbf{X}_n^1\big] = \frac{\hat{\Sigma}_n^{+,1\triangleright 0}(w,w')}{\sqrt{\hat{\Sigma}_n^{+,1\triangleright 0}(w,w)\hat{\Sigma}_n^{+,1\triangleright 0}(w',w')}}$. For $\alpha \in (0,1)$, define $\hat{q}_{1-\alpha}^{1\triangleright 0}$ as the quantile satisfying $\mathbb{P}\big(\sup_{w\in\mathcal{W}} \big|\hat{Z}_n^{T,1\triangleright 0}(w)\big| \leq \hat{q}_{1-\alpha}^{1\triangleright 0} \mid \mathbf{W}_n^1, \mathbf{X}_n^0, \mathbf{X}_n^1\big) = 1-\alpha$. Then if the covariance estimator is appropriately consistent,

$$\mathbb{P}\left(f_W^{1\triangleright 0}(w) \in \left[\hat{f}_W^{1\triangleright 0}(w) \pm \hat{q}_{1-\alpha}^{1\triangleright 0}\sqrt{\hat{\Sigma}_n^{+,1\triangleright 0}(w,w)}\right] \text{ for all } w \in \mathcal{W}\right) \to 1-\alpha,$$

giving feasible uniform inference methods, which are robust to unknown degeneracies, for counterfactual distribution analysis in dyadic data settings.

3.7.1 Application to trade data

We illustrate the performance of our estimation and inference methods with a real-world data set. We use international bilateral trade data from the International Monetary Fund's Direction of Trade Statistics (DOTS), previously analyzed by Head and Mayer (2014) and Chiang et al. (2023). This data set contains information about the yearly trade flows among $n = 207$ economies ($N = 21\,321$ pairs), and we focus on the years 1995, 2000, and 2005.

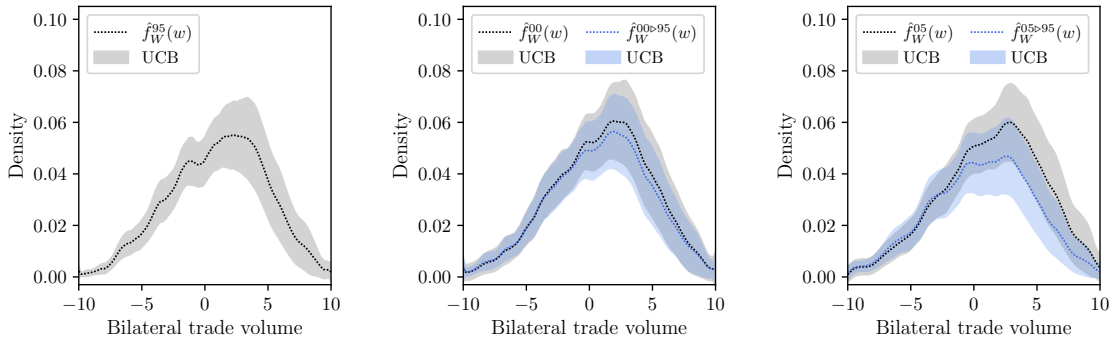
We define the *trade volume* between countries i and j as the logarithm of the sum of the trade flow (in billions of US dollars) from i to j and the trade flow from j to i. In each year several pairs of countries did not trade directly, yielding trade flows of zero and hence a trade volume of $-\infty$. We therefore assume that the distribution of trade volumes is a mixture of a point mass at $-\infty$ and a Lebesgue density on $\mathbb{R}$. The local nature of our estimator means that observations taking the value of $-\infty$ can simply be removed from the data set. Table 3.4 gives summary statistics for these trade networks, and shows how the networks become more connected over time, with edge density, average degree, and clustering coefficient increasing.

For counterfactual analysis we use the gross domestic product (GDP) of each country as a covariate, using 10%-percentiles to group the values into 10 different levels for ease of estimation. This allows for a comparison of the observed distribution of trade at each year with, for example, the counterfactual distribution of trade had the GDP distribution remained as it was in 1995. As such, we can measure how much of the change in trade distribution is attributable to a shift in the GDP distribution.

To estimate the trade volume density function we use Algorithm 1 with $d = 100$ equally-spaced evaluation points in $[-10, 10]$, using the rule-of-thumb bandwidth selector $\hat{h}_{\mathrm{ROT}}$ from Section 3.5.3 with $p = 2$ and $C(K) = 2.435$. For inference we use an Epanechnikov kernel

Year	Nodes	Edges	Edge density	Average degree	Clustering coefficient
1995	207	11 603	0.5442	112.1	0.7250
2000	207	12 528	0.5876	121.0	0.7674
2005	207	12 807	0.6007	123.7	0.7745

Table 3.4: Summary statistics for the DOTS trade networks.

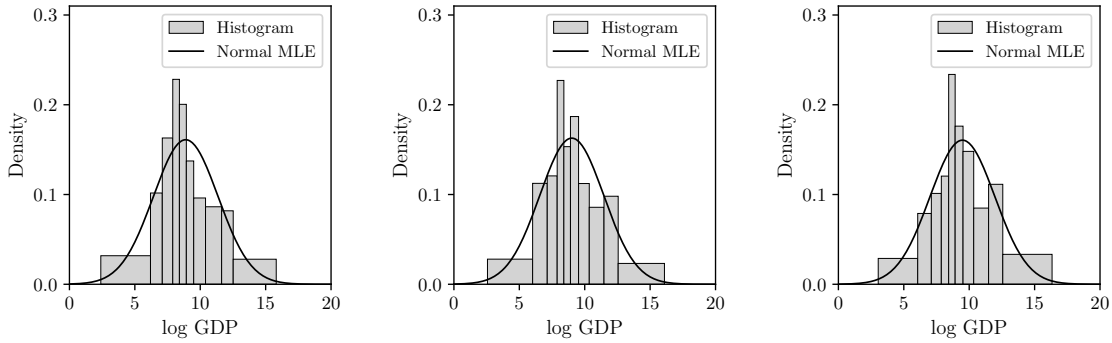


(a) Year 1995, $\hat{h}_{\mathrm{ROT}} = 1.27$. (b) Year 2000, $\hat{h}_{\mathrm{ROT}} = 1.31$. (c) Year 2005, $\hat{h}_{\mathrm{ROT}} = 1.37$.

Figure 3.5: Real and counterfactual density estimates and confidence bands for the DOTS data with histogram-based covariate estimation.

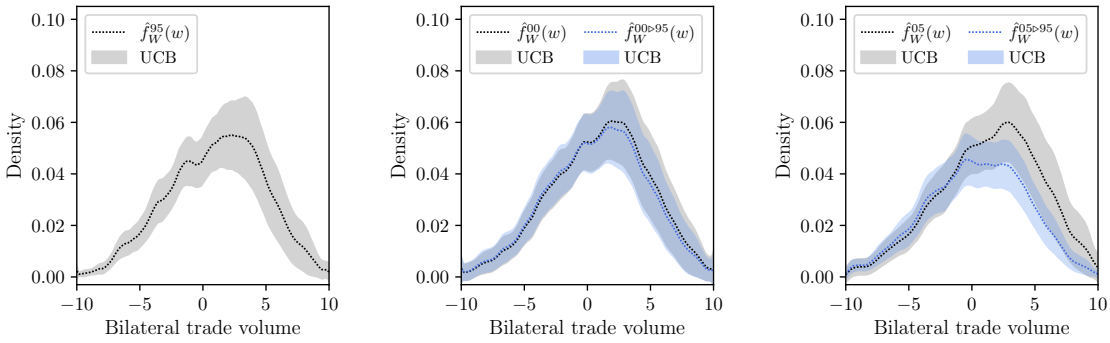
of order $p = 4$ and resample the Gaussian process $B = 10\,000$ times. We also estimate the counterfactual trade distributions in 2000 and 2005 respectively, replacing the GDP distribution with that from 1995. For each year, Figure 3.5 plots the real and counterfactual density estimates along with their respective uniform confidence bands (UCB) at the nominal coverage rate of 95%. Our empirical results show that the counterfactual distribution drifts further from the truth in 2005 compared with 2000, indicating a shift in the GDP distribution.

In Figure 3.6 we illustrate how, in the preliminary step of the counterfactual analysis, the distribution of log GDP is approximated using the histogram estimators $\hat{p}_X^0$ and $\hat{p}_X^1$ defined in Section 3.7. We also plot the density function of a normal distribution, fitted using maximum



(a) Year 1995 (b) Year 2000 (c) Year 2005

Figure 3.6: Estimated GDP distributions for the DOTS data using histograms and normal likelihood maximization.



(a) Year 1995, $\hat{h}_{\mathrm{ROT}} = 1.27$. (b) Year 2000, $\hat{h}_{\mathrm{ROT}} = 1.31$. (c) Year 2005, $\hat{h}_{\mathrm{ROT}} = 1.37$.

Figure 3.7: Real and counterfactual density estimates and confidence bands for the DOTS data with parametric covariate estimation.

likelihood estimation, and this seems to capture the distribution of log GDP reasonably well. Such a parametric approach to the preliminary step may be favored in cases where a choice of model is clear or where the histogram estimators perform poorly.

To demonstrate the relative robustness of our counterfactual analysis to the choice of preliminary estimation step, we provide results using a parametric estimator of the distribution of GDP. Figure 3.7 repeats the procedure used for Figure 3.5, but this time replacing the histogram estimators by parametric estimators of the log GDP based on normal likelihood maximization. The point estimates are qualitatively similar, with the counterfactual distribution drifting in the same direction over time. The confidence bands are also similar, with the band based on the parametric fit being slightly narrower in general. This could be due to the more stringent model specification leading to less estimated variance in the fitted values.

3.8 Other applications and future work

To emphasize the broad applicability of our methods to network science problems, we present three application scenarios. The first concerns comparison of networks (Kolaczyk, 2009), while the second and third involve nonparametric and semiparametric dyadic regression respectively.

Firstly, consider the setting where there are two independent networks with continuous dyadic covariates $\mathbf{W}_n^0$ and $\mathbf{W}_m^1$ respectively. Practitioners may wish to test if these two dyadic distributions are the same, that is, whether their density functions f_W^0 and f_W^1 are equal on

their common support $\mathcal{W} \subseteq \mathbb{R}$. We present a family of hypothesis tests for this scenario based on dyadic kernel density estimation. Let $\hat{f}_W^0(w)$ and $\hat{f}_W^1(w)$ be the associated (bias-corrected) dyadic kernel density estimators. Consider the test statistics τ_p for $1 \leq p \leq \infty$ where

$$\begin{aligned} \tau_p^p &= \int_{-\infty}^{\infty} \left| \hat{f}_W^1(w) - \hat{f}_W^0(w) \right|^p \, \mathrm{d}w \quad \text{for } p < \infty, \\ \tau_\infty &= \sup_{w \in \mathcal{W}} \left| \hat{f}_W^1(w) - \hat{f}_W^0(w) \right|. \end{aligned} \tag{3.4}$$

Clearly, we should reject the null hypothesis that $f_W^0 = f_W^1$ whenever the test statistic τ_p is sufficiently large. To estimate the critical value, let $\hat{\Sigma}_n^{+,0}(w, w')$ and $\hat{\Sigma}_m^{+,1}(w, w')$ be the positive semi-definite estimators defined in Section 3.5.1 and let $\hat{Z}_n^0(w)$ and $\hat{Z}_m^1(w)$ be zero-mean Gaussian processes with covariance structures $\hat{\Sigma}_n^{+,0}(w, w')$ and $\hat{\Sigma}_m^{+,1}(w, w')$ respectively, which are independent conditional on the data. Define the approximate null test statistic $\hat{\tau}_p$ by replacing $\hat{f}_W^0(w)$ and $\hat{f}_W^1(w)$ with $\hat{Z}_n^0(w)$ and $\hat{Z}_m^1(w)$ respectively in (3.4). For a significance level $\alpha \in (0, 1)$, the critical value is $\hat{C}_\alpha$ where $\mathbb{P}\big(\hat{\tau}_p \geq \hat{C}_\alpha \mid \mathbf{W}_n^0, \mathbf{W}_n^1\big) = \alpha$. This is estimated by Monte Carlo simulation, resampling from the conditional law of $\hat{Z}_n^0(w)$ and $\hat{Z}_m^1(w)$ and replacing integrals and suprema by sums and maxima over a finite partition of $\mathcal{W}$.

While our focus has been on density estimation with dyadic data, our uniform dyadic estimation and inference results are readily applicable to the settings of nonparametric and semiparametric dyadic regression. For a second example, suppose $Y_{ij} = Y(X_i, X_j, A_i, A_j, V_{ij})$, where only $\mathbf{X}_n$ and $\mathbf{Y}_n$ are observed and $\mathbf{V}_n$ is independent of $(\mathbf{X}_n, \mathbf{A}_n)$, with $\mathbf{X}_n = (X_i : 1 \leq i \leq n)$, $\mathbf{A}_n = (A_i : 1 \leq i \leq n)$, $\mathbf{Y}_n = (Y_{ij} : 1 \leq i < j \leq n)$, and $\mathbf{V}_n = (V_{ij} : 1 \leq i < j \leq n)$. A parameter of interest is the regression function $\mu(x_1, x_2) = \mathbb{E}[Y_{ij} \mid X_i = x_1, X_j = x_2]$, which can be used to analyze average or partial effects of changing the node attributes X_i and X_j on the edge variable Y_{ij}. This conditional expectation could be estimated using local polynomial methods: suppose that X_i takes values in $\mathbb{R}^m$ and let $r(x_1, x_2)$ be a monomial basis up to degree $\gamma \geq 0$ on $\mathbb{R}^m \times \mathbb{R}^m$. Then, for some bandwidth $h > 0$ and a kernel function k_h on $\mathbb{R}^m \times \mathbb{R}^m$, the local polynomial regression estimator of $\mu(x_1, x_2)$ is $\hat{\mu}(x_1, x_2) = e_1^\mathsf{T} \hat{\beta}(x_1, x_2)$

where e_1 is the first standard unit vector in $\mathbb{R}^q$ for $q = \binom{2m+\gamma}{\gamma}$ and

$$\begin{aligned}\hat{\beta}(x_1, x_2) &= \underset{\beta \in \mathbb{R}^q}{\arg\min} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(Y_{ij} - r(X_i - x_1, X_j - x_2)^{\mathsf{T}} \beta\right)^2 k_h(X_i - x_1, X_j - x_2) \\ &= \left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k_{ij} r_{ij} r_{ij}^{\mathsf{T}}\right)^{-1} \left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k_{ij} r_{ij} Y_{ij}\right), \end{aligned} \tag{3.5}$$

with $k_{ij} = k_h(X_i - x_1, X_j - x_2)$ and $r_{ij} = r(X_i - x_1, X_j - x_2)$. Graham et al. (2021) established pointwise distribution theory for the special case of the dyadic Nadaraya–Watson kernel regression estimator ($\gamma = 0$), but no uniform analogues have yet been given. It can be shown that the "denominator" matrix in (3.5) converges uniformly to its expectation, while the U-process-like "numerator" matrix can be handled the same way as we analyzed $\hat{f}_W(w)$ in this chapter, through a Hoeffding-type decomposition and strong approximation methods, along with standard bias calculations. Such distributional approximation results can be used to construct valid uniform confidence bands for the regression function $\mu(x_1, x_2)$, as well as to conduct hypothesis testing for parametric specifications or shape constraints.

As a third example, we consider applying our results to semiparametric semi-linear regression problems. The dyadic semi-linear regression model is $\mathbb{E}[Y_{ij} \mid W_{ij}, X_i, X_j] = \theta^{\mathsf{T}} W_{ij} + g(X_i, X_j)$ where θ is the finite-dimensional parameter of interest and $g(X_i, X_j)$ is an unknown function of the covariates (X_i, X_j). Local polynomial (or other) methods can be used to estimate θ and g, where the estimator of the nonparametric component g takes a similar form to (3.5), that is, a ratio of two kernel-based estimators as in (3.1). Consequently, the strong approximation techniques presented in this chapter can be appropriately modified to develop valid uniform inference procedures for g and $\mathbb{E}[Y_{ij} \mid W_{ij} = w, X_i = x_1, X_j = x_2]$, as well as functionals thereof.

3.9 Conclusion

We studied the uniform estimation and inference properties of the dyadic kernel density estimator $\hat{f}_W$ given in (3.1), which forms a class of U-process-like estimators indexed by the n-varying kernel function k_h on $\mathcal{W}$. We established uniform minimax-optimal point

estimation results and uniform distributional approximations for this estimator based on novel strong approximation strategies. We then applied these results to derive valid and feasible uniform confidence bands for the dyadic density estimand f_W, and also developed a substantive application of our theory to counterfactual dyadic density analysis. We gave some other statistical applications of our methodology as well as potential avenues for future research. From a technical perspective, Appendix B contains several generic results concerning strong approximation methods and maximal inequalities for empirical processes that may be of independent interest. Implementations of this chapter's methodology, along with replication files for the empirical results, are provided by a Julia package available at `github.com/wgunderwood/DyadicKDE.jl`. This work is based on Cattaneo et al. (2024), and has been presented by Cattaneo at the Columbia University Biostatistics Colloquium Seminar (2022) and the Georgia Institute of Technology Statistics Seminar (2022), by Feng at the Renmin University Econometrics Seminar (2022), the Xiamen University Symposium on Modern Statistics (2022), the Peking University Econometrics Seminar (2023), and the Asian Meeting of the Econometric Society in East and Southeast Asia, Singapore (2023), and by Underwood at the University of Illinois Statistics Seminar (2024), the University of Michigan Statistics Seminar (2024), and the University of Pittsburgh Statistics Seminar (2024).

Chapter 4

Yurinskii's Coupling for Martingales

Yurinskii's coupling is a popular theoretical tool for non-asymptotic distributional analysis in mathematical statistics and applied probability, offering a Gaussian strong approximation with an explicit error bound under easily verified conditions. Originally stated in ℓ^2-norm for sums of independent random vectors, it has recently been extended both to the ℓ^p-norm, for $1 \leq p \leq \infty$, and to vector-valued martingales in ℓ^2-norm, under some strong conditions. We present as our main result a Yurinskii coupling for approximate martingales in ℓ^p-norm, under substantially weaker conditions than those previously imposed. Our formulation further allows for the coupling variable to follow a more general Gaussian mixture distribution, and we provide a novel third-order coupling method which gives tighter approximations in certain settings. We specialize our main result to mixingales, martingales, and independent data, and derive uniform Gaussian mixture strong approximations for martingale empirical processes. Substantive applications of our theory to nonparametric partitioning-based and local polynomial regression procedures are provided.

4.1 Introduction

Yurinskii's coupling (Yurinskii, 1978) has proven to be an important theoretical tool for developing non-asymptotic distributional approximations in mathematical statistics and applied probability. For a sum S of n independent zero-mean d-dimensional random vectors, this coupling technique constructs (on a suitably enlarged probability space) a zero-mean d-dimensional Gaussian vector T with the same covariance matrix as S and which is close to S in probability, bounding the discrepancy $\|S - T\|$ as a function of n, d, the choice of the norm, and some features of the underlying distribution. See, for example, Pollard (2002, Chapter 10) for a textbook introduction.

When compared to other coupling approaches, such as the celebrated Hungarian construction (Komlós et al., 1975) or Zaitsev's coupling (Zaitsev, 1987a,b), Yurinskii's approach stands out for its simplicity, robustness, and wider applicability, while also offering tighter couplings in some applications (see below for more discussion and examples). These features have led many scholars to use Yurinskii's coupling to study the distributional features of high-dimensional statistical procedures in a variety of settings, often with the end goal of developing uncertainty quantification or hypothesis testing methods. For example, in recent years, Yurinskii's coupling has been used to construct Gaussian approximations for the suprema of empirical processes (Chernozhukov, Chetverikov, and Kato, 2014b); to establish distribution theory for non-Donsker stochastic t-processes generated in nonparametric series regression (Belloni, Chernozhukov, Chetverikov, and Kato, 2015); to prove distributional approximations for high-dimensional ℓ^p-norms (Biau and Mason, 2015); to develop distribution theory for vector-valued martingales (Belloni and Oliveira, 2018; Li and Liao, 2020); to derive a law of the iterated logarithm for stochastic gradient descent optimization methods (Anastasiou, Balasubramanian, and Erdogdu, 2019); to establish uniform distributional results for nonparametric high-dimensional quantile processes (Belloni et al., 2019); to develop distribution theory for non-Donsker stochastic t-processes generated in partitioning-based series regression (Cattaneo et al., 2020); to deduce Bernstein–von Mises theorems in high-dimensional settings (Ray and van der Vaart, 2021); and to develop distribution theory for non-Donsker U-processes based

on dyadic network data (Cattaneo et al., 2024). There are also many other early applications of Yurinskii's coupling: Dudley and Philipp (1983) and Dehling (1983) establish invariance principles for Banach space-valued random variables, and Le Cam (1988) and Sheehy and Wellner (1992) obtain uniform Donsker results for empirical processes, to name just a few.

This chapter presents a new Yurinskii coupling which encompasses and improves upon all of the results previously available in the literature, offering four new features:

(i) It applies to vector-valued *approximate martingale* data.

(ii) It allows for a *Gaussian mixture* coupling distribution.

(iii) It imposes *no restrictions on degeneracy* of the data covariance matrix.

(iv) It establishes a *third-order* coupling to improve the approximation in certain situations.

Closest to our work are the unpublished manuscript by Belloni and Oliveira (2018) and the recent paper by Li and Liao (2020), which both investigated distribution theory for martingale data using Yurinskii's coupling and related methods. Specifically, Li and Liao (2020) established a Gaussian ℓ^2-norm Yurinskii coupling for mixingales and martingales under the assumption that the covariance structure has a minimum eigenvalue bounded away from zero. As formally demonstrated in this chapter (Section 4.3.1), such eigenvalue assumptions can be prohibitively strong in practically relevant applications. In contrast, our Yurinskii coupling does not impose any restrictions on covariance degeneracy (iii), in addition to offering several other new features not present in Li and Liao (2020), including (i), (ii), (iv), and applicability to general ℓ^p-norms. In addition, we correct a slight technical inaccuracy in their proof relating to the derivation of bounds in probability (Remark 4.2.1). Belloni and Oliveira (2018) did not establish a Yurinskii coupling for martingales, but rather a central limit theorem for smooth functions of high-dimensional martingales using the celebrated second-order Lindeberg method (see Chatterjee, 2006, and references therein), explicitly accounting for covariance degeneracy. As a consequence, their result could be leveraged to deduce a Yurinskii coupling for martingales with additional, non-trivial technical work (see Section C.1 in Appendix C for details). Nevertheless, a Yurinskii coupling derived from Belloni and Oliveira (2018) would not feature (i), (ii), (iv), or general ℓ^p-norms, as our results

do. We discuss further the connections between our work and the related literature in the upcoming sections, both when introducing our main theoretical results and when presenting the examples and statistical applications.

The most general coupling result of this chapter (Theorem 4.2.1) is presented in Section 4.2, where we also specialize it to a slightly weaker yet more user-friendly formulation (Proposition 4.2.1). Our Yurinskii coupling for approximate martingales is a strict generalization of all previous Yurinskii couplings available in the literature, offering a Gaussian mixture strong approximation for approximate martingale vectors in ℓ^p-norm, with an improved rate of approximation when the third moments of the data are negligible, and with no assumptions on the spectrum of the data covariance matrix. A key technical innovation underlying the proof of Theorem 4.2.1 is that we explicitly account for the possibility that the minimum eigenvalue of the variance may be zero, or its lower bound may be unknown, with the argument proceeding using a carefully tailored regularization. Establishing a coupling to a Gaussian mixture distribution is achieved by an appropriate conditioning argument, leveraging a conditional version of Strassen's theorem established by Chen and Kato (2020), along with some related technical work detailed in Section C.1. A third-order coupling is obtained via a modification of a standard smoothing technique for Borel sets from classical versions of Yurinskii's coupling, enabling improved approximation errors whenever third moments are negligible.

In Proposition 4.2.1, we explicitly tune the parameters of the aforementioned regularization to obtain a simpler, parameter-free version of Yurinskii's coupling for approximate martingales, again offering Gaussian mixture coupling distributions and an improved third-order approximation error. This specialization of our main result takes an agnostic approach to potential singularities in the data covariance matrix and, as such, may be improved in specific applications where additional knowledge of the covariance structure is available. Section 4.2 also presents some further refinements when additional structure is imposed, deriving Yurinskii couplings for mixingales, martingales, and independent data as Corollaries 4.2.1, 4.2.2, and 4.2.3, respectively. We take the opportunity to discuss and correct in Remark 4.2.1 a technical issue which is often neglected (Pollard, 2002; Li and Liao, 2020) when using Yurinskii's cou-

pling to derive bounds in probability. Section 4.2.5 presents a stylized example portraying the relevance of our main technical results in the context of canonical factor models, illustrating the importance of each of our new Yurinskii coupling features (i)–(iv).

Section 4.3 considers a substantive application of our main results: strong approximation of martingale empirical processes. We begin with the motivating example of canonical kernel density estimation, demonstrating how Yurinskii's coupling can be applied, and showing in Lemma 4.3.1 why it is essential that we do not place any conditions on the minimum eigenvalue of the variance matrix (iii). We then present a general-purpose strong approximation for martingale empirical processes in Proposition 4.3.1, combining classical results in the empirical process literature (van der Vaart and Wellner, 1996) with our Corollary 4.2.2. This statement appears to be the first of its kind for martingale data, and when specialized to independent (and not necessarily identically distributed) data, it is shown to be superior to the best known comparable strong approximation result available in the literature (Berthet and Mason, 2006). Our improvement comes from using Yurinskii's coupling for the ℓ^∞-norm, where Berthet and Mason (2006) apply Zaitsev's coupling (Zaitsev, 1987a,b) with the larger ℓ^2-norm.

Section 4.4 further illustrates the applicability of our results through two examples in nonparametric regression estimation. Firstly, we deduce a strong approximation for partitioning-based least squares series estimators with time series data, applying Corollary 4.2.2 directly and additionally imposing only a mild mixing condition on the regressors. We show that our Yurinskii coupling for martingale vectors delivers the same distributional approximation rate as the best known result for independent data, and discuss how this can be leveraged to yield a feasible statistical inference procedure. We also show that if the residuals have vanishing conditional third moment, an improved rate of Gaussian approximation can be established. Secondly, we deduce a strong approximation for local polynomial estimators with time series data, using our result on martingale empirical processes (Proposition 4.3.1) and again imposing a mixing assumption. Appealing to empirical process theory is essential here as, in contrast with series estimators, local polynomials do not possess certain additive separability properties. The bandwidth restrictions we require are relatively mild, and, as far as we know, they have not been improved upon even with independent data.

Section 4.5 concludes the chapter. All proofs are collected in Appendix C, which also includes other technical lemmas of potential independent interest, alongside some further results on applications of our theory to deriving high-dimensional central limit theorems for martingales in Section C.2.

4.1.1 Notation

We write $\|x\|_p$ for $p \in [1, \infty]$ to denote the ℓ^p-norm if x is a (possibly random) vector or the induced operator ℓ^p–ℓ^p-norm if x is a matrix. For X a real-valued random variable and an Orlicz function ψ, we use $\|\|X\|\|_\psi$ to denote the Orlicz ψ-norm (van der Vaart and Wellner, 1996, Section 2.2) and $\|\|X\|\|_p$ for the $L^p(\mathbb{P})$-norm where $p \in [1, \infty]$. For a matrix M, we write $\|M\|_{\max}$ for the maximum absolute entry and $\|M\|_{\mathrm{F}}$ for the Frobenius norm. We denote positive semi-definiteness by $M \succeq 0$ and write I_d for the $d \times d$ identity matrix.

For scalar sequences x_n and y_n, we write $x_n \lesssim y_n$ if there exists a positive constant C such that $|x_n| \leq C|y_n|$ for sufficiently large n. We write $x_n \asymp y_n$ to indicate both $x_n \lesssim y_n$ and $y_n \lesssim x_n$. Similarly, for random variables X_n and Y_n, we write $X_n \lesssim_{\mathbb{P}} Y_n$ if for every $\varepsilon > 0$ there exists a positive constant C such that $\mathbb{P}(|X_n| \leq C|Y_n|) \leq \varepsilon$, and write $X_n \to_{\mathbb{P}} X$ for limits in probability. For real numbers a and b we use $a \vee b = \max\{a, b\}$. We write $\kappa \in \mathbb{N}^d$ for a multi-index, where $d \in \mathbb{N} = \{0, 1, 2, \ldots\}$, and define $|\kappa| = \sum_{j=1}^d \kappa_j$ and $x^\kappa = \prod_{j=1}^d x_j^{\kappa_j}$ for $x \in \mathbb{R}^d$, and $\kappa! = \prod_{j=1}^d \kappa_j!$.

Since our results concern couplings, some statements must be made on a new or enlarged probability space. We omit the details of this for clarity of notation, but technicalities are handled by the Vorob'ev–Berkes–Philipp Theorem (Dudley, 1999, Theorem 1.1.10).

4.2 Main results

We begin with our most general result: an ℓ^p-norm Yurinskii coupling of a sum of vector-valued approximate martingale differences to a Gaussian mixture-distributed random vector. The general result is presented in Theorem 4.2.1, while Proposition 4.2.1 gives a simplified and slightly weaker version which is easier to use in applications. We then further specialize Proposition 4.2.1 to three scenarios with successively stronger assumptions, namely mixingales,

martingales, and independent data in Corollaries 4.2.1, 4.2.2, and 4.2.3 respectively. In each case we allow for possibly random quadratic variations (cf. mixing convergence), thereby establishing a Gaussian mixture coupling in the general setting. In Remark 4.2.1 we comment on and correct an often overlooked technicality relating to the derivation of bounds in probability from Yurinskii's coupling. As a first illustration of the power of our generalized ℓ^p-norm Yurinskii coupling, we present in Section 4.2.5 a simple factor model example relating to all three of the aforementioned scenarios.

Theorem 4.2.1 (Strong approximation for vector-valued approximate martingales)
Take a complete probability space with a countably generated filtration $\mathcal{H}_0, \dots, \mathcal{H}_n$ for $n \geq 1$, supporting the $\mathbb{R}^d$-valued square-integrable variables $X_1, \dots, X_n$. Let $S = \sum_{i=1}^n X_i$ and define

$$\tilde{X}_i = \sum_{r=1}^{n} \big(\mathbb{E}[X_r \mid \mathcal{H}_i] - \mathbb{E}[X_r \mid \mathcal{H}_{i-1}]\big) \quad \text{and} \quad U = \sum_{i=1}^{n} \big(X_i - \mathbb{E}[X_i \mid \mathcal{H}_n] + \mathbb{E}[X_i \mid \mathcal{H}_0]\big).$$

Let $V_i = \operatorname{Var}[\tilde{X}_i \mid \mathcal{H}_{i-1}]$ and define $\Omega = \sum_{i=1}^n V_i - \Sigma$ where Σ is an almost surely positive semi-definite $\mathcal{H}_0$-measurable $d \times d$ matrix. Then, for each $\eta > 0$ and $p \in [1, \infty]$, there exists, on an enlarged probability space, an $\mathbb{R}^d$-valued random vector T with $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$ and

$$\begin{aligned}\mathbb{P}\big(\|S - T\|_p > 6\eta\big) \leq \inf_{t>0} \left\{ 2\mathbb{P}\big(\|Z\|_p > t\big) + \min\left\{ \frac{\beta_{p,2} t^2}{\eta^3}, \frac{\beta_{p,3} t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3} \right\} \right\} \\ + \inf_{M \succeq 0} \Big\{ 2\mathbb{P}\big(\Omega \not\preceq M\big) + \delta_p(M, \eta) + \varepsilon_p(M, \eta) \Big\} + \mathbb{P}\big(\|U\|_p > \eta\big), \quad (4.1)\end{aligned}$$

where $Z, Z_1, \dots, Z_n$ are i.i.d. standard Gaussian random variables on $\mathbb{R}^d$ independent of $\mathcal{H}_n$, the second infimum is taken over all positive semi-definite $d \times d$ non-random matrices M,

$$\beta_{p,k} = \sum_{i=1}^{n} \mathbb{E}\left[\|\tilde{X}_i\|_2^k \|\tilde{X}_i\|_p + \|V_i^{1/2} Z_i\|_2^k \|V_i^{1/2} Z_i\|_p \right], \quad \pi_3 = \sum_{i=1}^{n} \sum_{|\kappa|=3} \mathbb{E}\Big[\big| \mathbb{E}[\tilde{X}_i^\kappa \mid \mathcal{H}_{i-1}] \big| \Big]$$

for $k \in \{2, 3\}$, with $\pi_3 = \infty$ if the associated conditional expectation does not exist, and with

$$\begin{aligned}\delta_p(M, \eta) &= \mathbb{P}\left(\left\| \big((\Sigma + M)^{1/2} - \Sigma^{1/2}\big) Z \right\|_p \geq \eta \right), \\ \varepsilon_p(M, \eta) &= \mathbb{P}\left(\left\| (M - \Omega)^{1/2} Z \right\|_p \geq \eta,\ \Omega \preceq M \right).\end{aligned}$$

This theorem offers four novel contributions to the literature on coupling theory and strong approximation, as discussed in the introduction. Firstly (i), it allows for approximate vector-valued martingales, with the variables $\tilde{X}_i$ forming martingale differences with respect to $\mathcal{H}_i$ by construction, and U quantifying the associated martingale approximation error. Such martingale approximation techniques for sequences of dependent random vectors are well established and have been used in a range of scenarios: see, for example, Wu and Woodroofe (2004), Dedecker, Merlevède, and Volnỳ (2007), Zhao and Woodroofe (2008), Peligrad (2010), Atchadé and Cattaneo (2014), Cuny and Merlevède (2014), Magda and Zhang (2018), and references therein. In Section 4.2.2 we demonstrate how this approximation can be established in practice by restricting our general theorem to the special case of mixingales, while the upcoming example in Section 4.2.5 provides an illustration in the context of auto-regressive factor models.

Secondly (ii), Theorem 4.2.1 allows for the resulting coupling variable T to follow a multivariate Gaussian distribution only conditionally, and thus we offer a useful analog of mixing convergence in the context of strong approximation. To be more precise, the random matrix $\sum_{i=1}^n V_i$ is the quadratic variation of the constructed martingale $\sum_{i=1}^n \tilde{X}_i$, and we approximate it using the $\mathcal{H}_0$-measurable random matrix Σ. This yields the coupling variable $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$, which can alternatively be written as $T = \Sigma^{1/2} Z$ with $Z \sim \mathcal{N}(0, I_d)$ independent of $\mathcal{H}_0$. The errors in this quadratic variation approximation are accounted for by the terms $\mathbb{P}(\Omega \not\preceq M)$, $\delta_p(M, \eta)$, and $\varepsilon_p(M, \eta)$, utilizing a regularization argument through the free matrix parameter M. If a non-random Σ is used, then T is unconditionally Gaussian, and one can take $\mathcal{H}_0$ to be the trivial σ-algebra. As demonstrated in our proof, our approach to establishing a mixing approximation is different from naively taking an unconditional version of Yurinskii's coupling and applying it conditionally on $\mathcal{H}_0$, which will not deliver the same coupling as in Theorem 4.2.1 for a few reasons. To begin with, we explicitly indicate in the conditions of Theorem 4.2.1 where conditioning is required. Next, our error of approximation is given unconditionally, involving only marginal expectations and probabilities. Finally, we provide a rigorous account of the construction of the conditionally Gaussian coupling variable T via a conditional version of Strassen's theorem (Chen and Kato, 2020). Section 4.2.3

illustrates how a strong approximation akin to mixing convergence can arise when the data forms an exact martingale, and Section 4.2.5 gives a simple example relating to factor modeling in statistics and data science.

As a third contribution to the literature (iii), and of particular importance for applications, Theorem 4.2.1 makes no requirements on the minimum eigenvalue of the quadratic variation of the approximating martingale sequence. Instead, our proof technique employs a careful regularization scheme designed to account for any such exact or approximate rank degeneracy in Σ. This capability is fundamental in some applications, a fact which we illustrate in Section 4.3.1 by demonstrating the significant improvements in strong approximation errors delivered by Theorem 4.2.1 relative to those obtained using prior results in the literature.

Finally (iv), Theorem 4.2.1 gives a third-order strong approximation alongside the usual second-order version considered in all prior literature. More precisely, we observe that an analog of the term $\beta_{p,2}$ is present in the classical Yurinskii coupling and comes from a Lindeberg telescoping sum argument, replacing random variables by Gaussians with the same mean and variance to match the first and second moments. Whenever the third moments of $\tilde{X}_i$ are negligible (quantified by π_3), this moment-matching argument can be extended to third-order terms, giving a new term $\beta_{p,3}$. In certain settings, such as when the data is symmetrically distributed around zero, using $\beta_{p,3}$ rather than $\beta_{p,2}$ can give smaller approximation errors in the coupling given in (4.1). Such a refinement can be viewed as a strong approximation counterpart to classical Edgeworth expansion methods. We illustrate this phenomenon in our upcoming applications to nonparametric inference (Section 4.4).

4.2.1 User-friendly formulation of the main result

The result in Theorem 4.2.1 is given in a somewhat implicit manner, involving infima over the free parameters $t > 0$ and $M \succeq 0$, and it is not clear how to compute these in general. In the upcoming Proposition 4.2.1, we set $M = \nu^2 I_d$ and approximately optimize over $t > 0$ and $\nu > 0$, resulting in a simplified and slightly weaker version of our main general result. In specific applications, where there is additional knowledge of the quadratic variation structure, other choices of regularization schemes may be more appropriate. Nonetheless, the choice

$M = \nu^2 I_d$ leads to arguably the principal result of our work, due to its simplicity and utility in statistical applications. For convenience, define the functions $\phi_p : \mathbb{N} \to \mathbb{R}$ for $p \in [0, \infty]$,

$$\phi_p(d) = \begin{cases} \sqrt{pd^{2/p}} & \text{if } p \in [1, \infty), \\ \sqrt{2 \log 2d} & \text{if } p = \infty, \end{cases}$$

which are related to tail probabilities of the ℓ^p-norm of a standard Gaussian.

Proposition 4.2.1 (Simplified strong approximation for approximate martingales) *Assume the setup and notation of Theorem 4.2.1. For each $\eta > 0$ and $p \in [1, \infty]$, there exists a random vector $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$ satisfying*

$$\mathbb{P}\big(\|S - T\|_p > \eta\big) \leq 24 \left(\frac{\beta_{p,2} \phi_p(d)^2}{\eta^3}\right)^{1/3} + 17 \left(\frac{\mathbb{E}\left[\|\Omega\|_2\right] \phi_p(d)^2}{\eta^2}\right)^{1/3} + \mathbb{P}\left(\|U\|_p > \frac{\eta}{6}\right).$$

If further $\pi_3 = 0$ then

$$\mathbb{P}\big(\|S - T\|_p > \eta\big) \leq 24 \left(\frac{\beta_{p,3} \phi_p(d)^3}{\eta^4}\right)^{1/4} + 17 \left(\frac{\mathbb{E}\left[\|\Omega\|_2\right] \phi_p(d)^2}{\eta^2}\right)^{1/3} + \mathbb{P}\left(\|U\|_p > \frac{\eta}{6}\right).$$

Proposition 4.2.1 makes clear the potential benefit of a third-order coupling when $\pi_3 = 0$, as in this case the bound features $\beta_{p,3}^{1/4}$ rather than $\beta_{p,2}^{1/3}$. If π_3 is small but non-zero, an analogous result can easily be derived by adjusting the optimal choices of t and ν, but we omit this for clarity of notation. In applications (see Section 4.4.1), this reduction of the exponent can provide a significant improvement in terms of the dependence of the bound on the sample size n, the dimension d, and other problem-specific quantities. When using our results for strong approximation, it is usual to set $p = \infty$ to bound the maximum discrepancy over the entries of a vector (to construct uniform confidence sets, for example). In this setting, we have that $\phi_\infty(d) = \sqrt{2 \log 2d}$ has a sub-Gaussian slow-growing dependence on the dimension. The remaining term depends on $\mathbb{E}[\|\Omega\|_2]$ and requires that the matrix Σ be a good approximation of $\sum_{i=1}^n V_i$, while remaining $\mathcal{H}_0$-measurable. In some applications (such as factor modeling; see Section 4.2.5), it can be shown that the quadratic variation $\sum_{i=1}^n V_i$ remains random and $\mathcal{H}_0$-measurable even in large samples, giving a natural choice for Σ.

In the next few sections, we continue to refine Proposition 4.2.1, presenting a sequence of results with increasingly strict assumptions on the dependence structure of the data X_i. These allow us to demonstrate the broad applicability of our main results, providing more explicit bounds in settings which are likely to be of special interest. In particular, we consider mixingales, martingales, and independent data, comparing our derived results with those in the existing literature.

4.2.2 Mixingales

In our first refinement, we provide a natural method for bounding the martingale approximation error term U. Suppose that X_i form an ℓ^p-mixingale in $L^1(\mathbb{P})$ in the sense that there exist non-negative $c_1, \ldots, c_n$ and $\zeta_0, \ldots, \zeta_n$ such that for all $1 \leq i \leq n$ and $0 \leq r \leq i$,

$$\mathbb{E}\left[\left\|\mathbb{E}\left[X_i \mid \mathcal{H}_{i-r}\right]\right\|_p\right] \leq c_i \zeta_r, \tag{4.2}$$

and for all $1 \leq i \leq n$ and $0 \leq r \leq n - i$,

$$\mathbb{E}\left[\left\|X_i - \mathbb{E}\left[X_i \mid \mathcal{H}_{i+r}\right]\right\|_p\right] \leq c_i \zeta_{r+1}. \tag{4.3}$$

These conditions are satisfied, for example, if X_i are integrable strongly α-mixing random variables (McLeish, 1975), or if X_i are generated by an auto-regressive or auto-regressive moving average process (see Section 4.2.5), among many other possibilities (Bradley, 2005). Then, in the notation of Theorem 4.2.1, we have by Markov's inequality that

$$\mathbb{P}\left(\|U\|_p > \frac{\eta}{6}\right) \leq \frac{6}{\eta} \sum_{i=1}^{n} \mathbb{E}\left[\left\|X_i - \mathbb{E}\left[X_i \mid \mathcal{H}_n\right]\right\|_p + \left\|\mathbb{E}\left[X_i \mid \mathcal{H}_0\right]\right\|_p\right] \leq \frac{\zeta}{\eta},$$

with $\zeta = 6 \sum_{i=1}^{n} c_i(\zeta_i + \zeta_{n-i+1})$. Combining Proposition 4.2.1 with this martingale error bound yields the following result for mixingales.

Corollary 4.2.1 (Strong approximation for vector-valued mixingales)

Assume the setup and notation of Theorem 4.2.1, and suppose the mixingale conditions (4.2) *and* (4.3) *hold. For each* $\eta > 0$ *and* $p \in [1, \infty]$ *there is a random vector* $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$ *with*

$$\mathbb{P}\big(\|S - T\|_p > \eta\big) \leq 24 \left(\frac{\beta_{p,2}\phi_p(d)^2}{\eta^3}\right)^{1/3} + 17 \left(\frac{\mathbb{E}\left[\|\Omega\|_2\right]\phi_p(d)^2}{\eta^2}\right)^{1/3} + \frac{\zeta}{\eta}.$$

If further $\pi_3 = 0$ *then*

$$\mathbb{P}\big(\|S - T\|_p > \eta\big) \leq 24 \left(\frac{\beta_{p,3}\phi_p(d)^3}{\eta^4}\right)^{1/4} + 17 \left(\frac{\mathbb{E}\left[\|\Omega\|_2\right]\phi_p(d)^2}{\eta^2}\right)^{1/3} + \frac{\zeta}{\eta}.$$

The closest antecedent to Corollary 4.2.1 is found in Li and Liao (2020, Theorem 4), who also considered Yurinskii's coupling for mixingales. Our result improves on this work in the following manner: it removes any requirements on the minimum eigenvalue of the quadratic variation of the mixingale sequence; it allows for general ℓ^p-norms with $p \in [1, \infty]$; it establishes a coupling to a multivariate Gaussian mixture distribution in general; and it permits third-order couplings (when $\pi_3 = 0$). These improvements have important practical implications as demonstrated in Sections 4.2.5 and 4.4, where significantly better coupling approximation errors are demonstrated for a variety of statistical applications. On the technical side, our result is rigorously established using a conditional version of Strassen's theorem (Chen and Kato, 2020), a carefully crafted regularization argument, and a third-order Lindeberg method (see Chatterjee, 2006, and references therein, for more discussion on the standard second-order Lindeberg method). Furthermore, as explained in Remark 4.2.1, we clarify a technical issue in Li and Liao (2020) surrounding the derivation of valid probability bounds for $\|S - T\|_p$.

Corollary 4.2.1 focused on mixingales for simplicity, but, as previously discussed, any method for constructing a martingale approximation $\tilde{X}_i$ and bounding the resulting error U could be used instead in Proposition 4.2.1 to derive a similar result.

4.2.3 Martingales

For our second refinement, suppose that X_i form martingale differences with respect to $\mathcal{H}_i$. In this case, $\mathbb{E}[X_i \mid \mathcal{H}_n] = X_i$ and $\mathbb{E}[X_i \mid \mathcal{H}_0] = 0$, so $U = 0$, and the martingale approximation error term vanishes. Applying Proposition 4.2.1 in this setting directly yields the following result.

Corollary 4.2.2 (Strong approximation for vector-valued martingales)
With the setup and notation of Theorem 4.2.1, suppose that X_i is $\mathcal{H}_i$-measurable satisfying $\mathbb{E}[X_i \mid \mathcal{H}_{i-1}] = 0$ for $1 \le i \le n$. Then, for each $\eta > 0$ and $p \in [1, \infty]$, there is a random vector $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$ with

$$\mathbb{P}\big(\|S - T\|_p > \eta\big) \le 24 \left(\frac{\beta_{p,2} \phi_p(d)^2}{\eta^3} \right)^{1/3} + 17 \left(\frac{\mathbb{E}\left[\|\Omega\|_2\right] \phi_p(d)^2}{\eta^2} \right)^{1/3}. \tag{4.4}$$

If further $\pi_3 = 0$ then

$$\mathbb{P}\big(\|S - T\|_p > \eta\big) \le 24 \left(\frac{\beta_{p,3} \phi_p(d)^3}{\eta^4} \right)^{1/4} + 17 \left(\frac{\mathbb{E}\left[\|\Omega\|_2\right] \phi_p(d)^2}{\eta^2} \right)^{1/3}. \tag{4.5}$$

The closest antecedents to Corollary 4.2.2 are Belloni and Oliveira (2018) and Li and Liao (2020), who also implicitly or explicitly considered Yurinskii's coupling for martingales. More specifically, Li and Liao (2020, Theorem 1) established an explicit ℓ^2-norm Yurinskii coupling for martingales under a strong assumption on the minimum eigenvalue of the martingale quadratic variation, while Belloni and Oliveira (2018, Theorem 2.1) established a central limit theorem for vector-valued martingale sequences employing the standard second-order Lindeberg method, implying that their proof could be adapted to deduce a Yurinskii coupling for martingales with the help of a conditional version of Strassen's theorem (Chen and Kato, 2020) and some additional nontrivial technical work.

Corollary 4.2.2 improves over this prior work as follows. With respect to Li and Liao (2020), our result establishes an ℓ^p-norm Gaussian mixture Yurinskii coupling for martingales without any requirements on the minimum eigenvalue of the martingale quadratic variation, and permits a third-order coupling if $\pi_3 = 0$. The first probability bound (4.4) in Corollary 4.2.2

gives the same rate of strong approximation as that in Theorem 1 of Li and Liao (2020) when $p = 2$, with non-random Σ, and when the eigenvalues of a normalized version of Σ are bounded away from zero. In Section 4.3.1 we demonstrate the crucial importance of removing this eigenvalue lower bound restriction in applications involving nonparametric kernel estimators, while in Section 4.4.1 we demonstrate how the availability of a third-order coupling (4.5) can give improved approximation rates in applications involving nonparametric series estimators with conditionally symmetrically distributed residual errors. Finally, our technical work improves on Li and Liao (2020) in two respects: (i) we employ a conditional version of Strassen's theorem (see Lemma C.1.1 in the appendix) to appropriately handle the conditioning arguments; and (ii) we deduce valid probability bounds for $\|S - T\|_p$, as the following Remark 4.2.1 makes clear.

Remark 4.2.1 (Yurinskii's coupling and bounds in probability)
Given a sequence of random vectors S_n, Yurinskii's method provides a coupling in the following form: for each n and any $\eta > 0$, there exists a random vector T_n with $\mathbb{P}\big(\|S_n - T_n\| > \eta\big) < r_n(\eta)$, where $r_n(\eta)$ is the approximation error. Crucially, each coupling variable T_n is a function of the desired approximation level η and, as such, deducing bounds in probability on $\|S_n - T_n\|$ requires some extra care. One option is to select a sequence $R_n \to \infty$ and note that $\mathbb{P}\big(\|S_n - T_n\| > r_n^{-1}(1/R_n)\big) < 1/R_n \to 0$ and hence $\|S_n - T_n\| \lesssim_{\mathbb{P}} r_n^{-1}(1/R_n)$. In this case, T_n depends on the choice of R_n, which can in turn typically be chosen to diverge slowly enough to cause no issues in applications.

Technicalities akin to those outlined in Remark 4.2.1 have been both addressed and neglected alike in the prior literature. Pollard (2002, Chapter 10.4, Example 16) apparently misses this subtlety, providing an inaccurate bound in probability based on the Yurinskii coupling. Li and Liao (2020) seem to make the same mistake in the proof of their Lemma A2, which invalidates the conclusion of their Theorem 1. In contrast, Belloni et al. (2015) and Belloni et al. (2019) directly provide bounds in $o_{\mathbb{P}}$ instead of $O_{\mathbb{P}}$, circumventing these issues in a manner similar to our approach involving a diverging sequence R_n.

To see how this phenomenon applies to our main results, observe that the second-order martingale coupling given as (4.4) in Corollary 4.2.2 implies that for any $R_n \to \infty$,

$$\|S - T\|_p \lesssim_{\mathbb{P}} \beta_{p,2}^{1/3} \phi_p(d)^{2/3} R_n + \mathbb{E}[\|\Omega\|_2]^{1/2} \phi_p(d) R_n.$$

This bound is comparable to that obtained by Li and Liao (2020, Theorem 1) with $p = 2$, albeit with their formulation missing the R_n correction terms. In Section 4.4.1 we discuss further their (amended) result, in the setting of nonparametric series estimation. Our approach using $p = \infty$ obtains superior distributional approximation rates, alongside exhibiting various other improvements such as the aforementioned third-order coupling.

Turning to the comparison with Belloni and Oliveira (2018), our Corollary 4.2.2 again offers the same improvements, with the only exception being that the authors did account for the implications of a possibly vanishing minimum eigenvalue. However, their results exclusively concern high-dimensional central limit theorems for vector-valued martingales, and therefore while their findings could in principle enable the derivation of a result similar to our Corollary 4.2.2, this would require additional technical work on their behalf in multiple ways (see Appendix C): (i) a correct application of a conditional version of Strassen's theorem (Lemma C.1.1); (ii) the development of a third-order Borel set smoothing technique and associated ℓ^p-norm moment control (Lemmas C.1.2, C.1.3, and C.1.4); (iii) a careful truncation scheme to account for $\Omega \not\asymp 0$; and (iv) a valid third-order Lindeberg argument (Lemma C.1.8), among others.

4.2.4 Independence

As a final refinement, suppose that X_i are independent and zero-mean conditionally on $\mathcal{H}_0$, and take $\mathcal{H}_i$ to be the filtration generated by $X_1, \ldots, X_i$ and $\mathcal{H}_0$ for $1 \leq i \leq n$. Then, taking $\Sigma = \sum_{i=1}^n V_i$ gives $\Omega = 0$, and hence Corollary 4.2.2 immediately yields the following result.

Corollary 4.2.3 (Strong approximation for sums of independent vectors)

Take the setup of Theorem 4.2.1, and let X_i be independent given $\mathcal{H}_0$, with $\mathbb{E}[X_i \mid \mathcal{H}_0] = 0$. Then, for each $\eta > 0$ and $p \in [1, \infty]$, with $\Sigma = \sum_{i=1}^n V_i$, there is $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$ with

$$\mathbb{P}\big(\|S - T\|_p > \eta\big) \leq 24 \left(\frac{\beta_{p,2} \phi_p(d)^2}{\eta^3} \right)^{1/3}. \tag{4.6}$$

If further $\pi_3 = 0$ then

$$\mathbb{P}\big(\|S - T\|_p > \eta\big) \leq 24 \left(\frac{\beta_{p,3} \phi_p(d)^3}{\eta^4} \right)^{1/4}.$$

Taking $\mathcal{H}_0$ to be trivial, (4.6) provides an ℓ^p-norm approximation analogous to that presented in Belloni et al. (2019). By further restricting to $p = 2$, we recover the original Yurinskii coupling as presented in Le Cam (1988, Theorem 1) and Pollard (2002, Theorem 10). Thus, in the independent data setting, our result improves on prior work as follows: (i) it establishes a coupling to a multivariate Gaussian mixture distribution; and (ii) it permits a third-order coupling if $\pi_3 = 0$.

4.2.5 Stylized example: factor modeling

In this section, we present a simple statistical example of how our improvements over prior coupling results can have important theoretical and practical implications. Consider the stylized factor model

$$X_i = L f_i + \varepsilon_i, \qquad 1 \leq i \leq n,$$

with random variables L taking values in $\mathbb{R}^{d \times m}$, f_i in $\mathbb{R}^m$, and ε_i in $\mathbb{R}^d$. We interpret f_i as a latent factor variable and L as a random factor loading, with idiosyncratic disturbances ε_i. See Fan et al. (2020), and references therein, for a textbook review of factor analysis in statistics and econometrics.

We employ the above factor model to give a first illustration of the applicability of our main result Theorem 4.2.1, the user-friendly Proposition 4.2.1, and their specialized Corollaries 4.2.1–4.2.3. We consider three different sets of conditions to demonstrate the applicability of each of our corollaries for mixingales, martingales, and independent data, respectively. We assume throughout that $(\varepsilon_1, \ldots, \varepsilon_n)$ is zero-mean and finite variance, and that $(\varepsilon_1, \ldots, \varepsilon_n)$ is independent of L and $(f_1, \ldots, f_n)$. Let $\mathcal{H}_i$ be the σ-algebra generated by L, $(f_1, \ldots, f_i)$, and $(\varepsilon_1, \ldots, \varepsilon_i)$, with $\mathcal{H}_0$ the σ-algebra generated by L alone.

- *Independent data.* Suppose that the factors $(f_1, \ldots, f_n)$ are independent conditional on L and satisfy $\mathbb{E}[f_i \mid L] = 0$. Then, since X_i are independent conditional on $\mathcal{H}_0$ and with $\mathbb{E}[X_i \mid \mathcal{H}_0] = \mathbb{E}[Lf_i + \varepsilon_i \mid L] = 0$, we can apply Corollary 4.2.3 to $\sum_{i=1}^n X_i$. In general, we will obtain a coupling variable which has the Gaussian mixture distribution $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = \sum_{i=1}^n (L \operatorname{Var}[f_i \mid L] L^\mathsf{T} + \operatorname{Var}[\varepsilon_i])$. In the special case where L is non-random and $\mathcal{H}_0$ is trivial, the coupling is Gaussian. Further, if $f_i \mid L$ and ε_i are symmetric about zero and bounded, then $\pi_3 = 0$, and the coupling is improved.

- *Martingales.* Suppose instead that we assume only a martingale condition on the latent factor variables so that $\mathbb{E}[f_i \mid L, f_1, \ldots, f_{i-1}] = 0$. Then $\mathbb{E}[X_i \mid \mathcal{H}_{i-1}] = L\,\mathbb{E}[f_i \mid \mathcal{H}_{i-1}] = 0$ and Corollary 4.2.2 is applicable to $\sum_{i=1}^n X_i$. The preceding comments on Gaussian mixture distributions and third-order couplings continue to apply.

- *Mixingales.* Finally, assume that the factors follow the auto-regressive model $f_i = Af_{i-1} + u_i$ where $A \in \mathbb{R}^{m \times m}$ is non-random and $(u_1, \ldots, u_n)$ are zero-mean, independent, and independent of $(\varepsilon_1, \ldots, \varepsilon_n)$. Then $\mathbb{E}[f_i \mid f_0] = A^i f_0$, so taking $p \in [1, \infty]$ we see that $\mathbb{E}\big[\|\mathbb{E}[f_i \mid f_0]\|_p\big] = \mathbb{E}\big[\|A^i f_0\|_p\big] \leq \|A\|_p^i\, \mathbb{E}[\|f_0\|_p]$, and that clearly $f_i - \mathbb{E}[f_i \mid \mathcal{H}_n] = 0$. Thus, whenever $\|A\|_p < 1$, the geometric sum formula implies that we can apply the mixingale result from Corollary 4.2.1 to $\sum_{i=1}^n X_i$. The conclusions on Gaussian mixture distributions and third-order couplings parallel the previous cases.

This simple application to factor modeling gives a preliminary illustration of the power of our main results, encompassing settings which could not be handled by employing Yurinskii couplings available in the existing literature. Even with independent data, we offer new

Yurinskii couplings to Gaussian mixture distributions (due to the presence of the common random factor loading L), which could be further improved whenever the factors and residuals possess symmetric (conditional) distributions. Furthermore, our results do not impose any restrictions on the minimum eigenvalue of Σ, thereby allowing for more general factor structures. These improvements are maintained in the martingale, mixingale, and weakly dependent stationary data settings.

4.3 Strong approximation for martingale empirical processes

In this section, we demonstrate how our main results can be applied to some more substantive problems in statistics. Having until this point studied only finite-dimensional (albeit potentially high-dimensional) random vectors, we now turn our attention to infinite-dimensional stochastic processes. Specifically, we consider empirical processes of the form $S(f) = \sum_{i=1}^{n} f(X_i)$ for $f \in \mathcal{F}$ a problem-specific class of real-valued functions, where each $f(X_i)$ forms a martingale difference sequence with respect to an appropriate filtration. We construct (conditionally) Gaussian processes $T(f)$ for which an upper bound on the uniform coupling error $\sup_{f\in\mathcal{F}} |S(f) - T(f)|$ is precisely quantified. We control the complexity of $\mathcal{F}$ using metric entropy under Orlicz norms.

The novel strong approximation results which we present concern the entire martingale empirical process $(S(f) : f \in \mathcal{F})$, as opposed to just the scalar supremum of the empirical process, $\sup_{f\in\mathcal{F}} |S(f)|$. This distinction has been carefully noted by Chernozhukov et al. (2014b), who studied Gaussian approximation of empirical process suprema in the independent data setting and wrote (p. 1565): "A related but different problem is that of approximating *whole* empirical processes by a sequence of Gaussian processes in the sup-norm. This problem is more difficult than [approximating the supremum of the empirical process]." Indeed, the results we establish in this section are for a strong approximation for the entire empirical process by a sequence of Gaussian mixture processes in the supremum norm, when the data has a martingale difference structure (cf. Corollary 4.2.2). Our results can be further generalized to approximate martingale empirical processes (cf. Corollary 4.2.1), but we do not consider this extension to reduce notation and the technical burden.

4.3.1 Motivating example: kernel density estimation

We begin with a brief study of a canonical example of an empirical process which is non-Donsker (thus precluding the use of uniform central limit theorems) due to the presence of a function class whose complexity increases with the sample size: the kernel density estimator with i.i.d. scalar data. We give an overview of our general strategy for strong approximation of stochastic processes via discretization, and show explicitly in Lemma 4.3.1 how it is crucial that we do not impose lower bounds on the eigenvalues of the discretized covariance matrix. Detailed calculations for this section are relegated to Appendix C for conciseness.

Let $X_1, \dots, X_n$ be i.i.d. Unif$[0,1]$, take $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ the Gaussian kernel and let $h \in (0,1]$ be a bandwidth. Then, for $a \in (0, 1/4]$ and $x \in \mathcal{X} = [a, 1-a]$ to avoid boundary issues, the kernel density estimator of the true density function $g(x) = 1$ is

$$\hat{g}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x), \qquad K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right).$$

Consider establishing a strong approximation for the stochastic process $(\hat{g}(x) - \mathbb{E}[\hat{g}(x)] : x \in \mathcal{X})$ which is, upon rescaling, non-Donsker whenever the bandwidth decreases to zero in large samples. To match notation with the upcoming general result for empirical processes, set $f_x(u) = \frac{1}{n}(K_h(u-x) - \mathbb{E}[K_h(X_i - x)])$ so $S(x) := S(f_x) = \hat{g}(x) - \mathbb{E}[\hat{g}(x)]$. The next step is standard: a mesh separates the local oscillations of the processes from the finite-dimensional coupling. For $\delta \in (0, 1/2)$, set $N = \lfloor 1 + \frac{1-2a}{\delta} \rfloor$ and $\mathcal{X}_\delta = (a + (j-1)\delta : 1 \le j \le N)$. Letting $T(x)$ be the approximating stochastic process to be constructed, consider the decomposition

$$\sup_{x \in \mathcal{X}} \big|S(x) - T(x)\big| \le \sup_{|x - x'| \le \delta} \big|S(x) - S(x')\big| + \max_{x \in \mathcal{X}_\delta} |S(x) - T(x)| + \sup_{|x-x'| \le \delta} \big|T(x) - T(x')\big|.$$

Writing $S(\mathcal{X}_\delta)$ for $\big(S(x) : x \in \mathcal{X}_\delta\big) \in \mathbb{R}^N$, noting that this is a sum of i.i.d. random vectors, we apply Corollary 4.2.3 as $\max_{x \in \mathcal{X}_\delta} |S(x) - T(x)| = \|S(\mathcal{X}_\delta) - T(\mathcal{X}_\delta)\|_\infty$. We obtain that for each $\eta > 0$ there is a Gaussian vector $T(\mathcal{X}_\delta)$ with the same covariance matrix as $S(\mathcal{X}_\delta)$

satisfying

$$\mathbb{P}\left(\|S(\mathcal{X}_\delta) - T(\mathcal{X}_\delta)\|_\infty > \eta\right) \leq 31\left(\frac{N\log 2N}{\eta^3 n^2 h^2}\right)^{1/3}$$

assuming that $1/h \geq \log 2N$. By the Vorob'ev–Berkes–Philipp theorem (Dudley, 1999, Theorem 1.1.10), $T(\mathcal{X}_\delta)$ extends to a Gaussian process $T(x)$ defined for all $x \in \mathcal{X}$ and with the same covariance structure as $S(x)$.

Next, chaining with the Bernstein–Orlicz and sub-Gaussian norms (van der Vaart and Wellner, 1996, Section 2.2) shows that if $\log(N/h) \lesssim \log n$ and $nh \gtrsim \log n$,

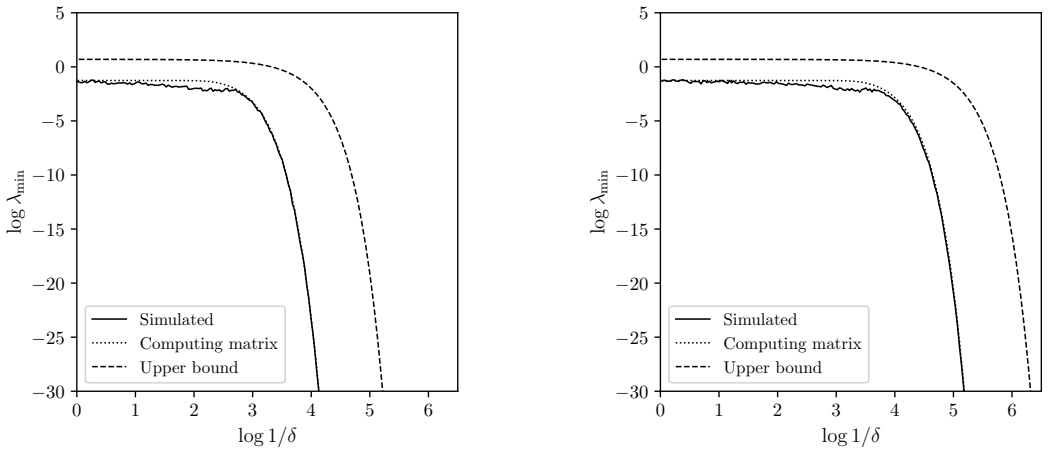
$$\sup_{|x-x'|\leq\delta} \left\|S(x) - S(x')\right\|_\infty \lesssim_{\mathbb{P}} \delta\sqrt{\frac{\log n}{nh^3}} \quad \text{and} \quad \sup_{|x-x'|\leq\delta} \left\|T(x) - T(x')\right\|_\infty \lesssim_{\mathbb{P}} \delta\sqrt{\frac{\log n}{nh^3}}.$$

Finally, for any $R_n \to \infty$ (see Remark 4.2.1), the resulting bound on the coupling error is

$$\sup_{x\in\mathcal{X}} \left|S(x) - T(x)\right| \lesssim_{\mathbb{P}} \left(\frac{N\log 2N}{n^2h^2}\right)^{1/3} R_n + \delta\sqrt{\frac{\log n}{nh^3}},$$

where the mesh size δ can then be approximately optimized to obtain the tightest possible strong approximation.

The discretization strategy outlined above is at the core of the proof strategy for our upcoming Proposition 4.3.1. Since we will consider martingale empirical processes, our proof will rely on Corollary 4.2.2, which, unlike the martingale Yurinskii coupling established by Li and Liao (2020), does not require a lower bound on the minimum eigenvalue of Σ. Using the simple kernel density example just discussed, we now demonstrate precisely the crucial importance of removing such eigenvalue conditions. The following Lemma 4.3.1 shows that the discretized covariance matrix $\Sigma = nh\operatorname{Var}[S(\mathcal{X}_\delta)]$ has exponentially small eigenvalues, which in turn will negatively affect the strong approximation bound if the Li and Liao (2020) coupling were to be used instead of the results in this dissertation.



(a) $h = 0.03$ (b) $h = 0.01$

Figure 4.1: Upper bounds on the minimum eigenvalue of the discretized covariance matrix in kernel density estimation, with $n = 100$ and $a = 0.2$. Simulated: the kernel density estimator is simulated, resampling the data 100 times to estimate its covariance. Computing matrix: the minimum eigenvalue of the limiting covariance matrix Σ is computed explicitly. Upper bound: the bound derived in Lemma 4.3.1 is shown.

Lemma 4.3.1 (Minimum eigenvalue of a kernel density estimator covariance matrix)

The minimum eigenvalue of $\Sigma = nh \operatorname{Var}[S(\mathcal{X}_\delta)] \in \mathbb{R}^{N \times N}$ *satisfies the upper bound*

$$\lambda_{\min}(\Sigma) \leq 2e^{-h^2/\delta^2} + \frac{h}{\pi a \delta} e^{-a^2/h^2}.$$

Figure 4.1 shows how the upper bound in Lemma 4.3.1 captures the behavior of the simulated minimum eigenvalue of Σ. In particular, the smallest eigenvalue decays exponentially fast in the discretization level δ and the bandwidth h. As seen in the calculations above, the coupling rate depends on δ/h, while the bias will generally depend on h, implying that both δ and h must converge to zero to ensure valid statistical inference. In general, this will lead to Σ possessing extremely small eigenvalues, rendering strong approximation approaches such as that of Li and Liao (2020) ineffective in such scenarios.

The discussion in this section focuses on the strong approximation of the centered process $\hat{g}(x) - \mathbb{E}[\hat{g}(x)]$. In practice, the goal is often rather to approximate the feasible process $\hat{g}(x) - g(x)$. The difference between these is captured by the smoothing bias $\mathbb{E}[\hat{g}(x)] - g(x)$, which is straightforward to control in this case with $\sup_{x \in \mathcal{X}} \left|\mathbb{E}[\hat{g}(x)] - g(x)\right| \lesssim \frac{h}{a} e^{-a^2/(2h^2)}$. See Section 4.4 for further comments.

4.3.2 General result for martingale empirical processes

We now give our general result on a strong approximation for martingale empirical processes, obtained by applying the first result (4.4) in Corollary 4.2.2 with $p = \infty$ to a discretization of the empirical process, as in Section 4.3.1. We then control the increments in the stochastic processes using chaining with Orlicz norms, but note that other tools are available, including generalized entropy with bracketing (van de Geer, 2000) and sequential symmetrization (Rakhlin, Sridharan, and Tewari, 2015).

A class of functions is said to be *pointwise measurable* if it contains a countable subclass which is dense under the pointwise convergence topology. For a finite class $\mathcal{F}$, write $\mathcal{F}(x) = \big(f(x) : f \in \mathcal{F}\big)$. Define the set of Orlicz functions

$$\Psi = \left\{ \psi : [0,\infty) \to [0,\infty) \text{ convex increasing, } \psi(0) = 0,\ \limsup_{x,y\to\infty} \tfrac{\psi(x)\psi(y)}{\psi(Cxy)} < \infty \text{ for } C > 0 \right\}$$

and, for real-valued Y, the Orlicz norm $\|\!|Y|\!\|_\psi = \inf\{C > 0 : \mathbb{E}\left[\psi(|Y|/C) \le 1\right]\}$ as in van der Vaart and Wellner (1996, Section 2.2).

Proposition 4.3.1 (Strong approximation for martingale empirical processes)
Let X_i be random variables for $1 \le i \le n$ taking values in a measurable space $\mathcal{X}$, and $\mathcal{F}$ be a pointwise measurable class of functions from $\mathcal{X}$ to $\mathbb{R}$. Let $\mathcal{H}_0, \ldots, \mathcal{H}_n$ be a filtration such that each X_i is $\mathcal{H}_i$-measurable, with $\mathcal{H}_0$ the trivial σ-algebra, and suppose that $\mathbb{E}[f(X_i) \mid \mathcal{H}_{i-1}] = 0$ for all $f \in \mathcal{F}$. Define $S(f) = \sum_{i=1}^n f(X_i)$ for $f \in \mathcal{F}$ and let $\Sigma : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ be an almost surely positive semi-definite $\mathcal{H}_0$-measurable random function. Suppose that for a non-random metric d on $\mathcal{F}$, constant L, and $\psi \in \Psi$,

$$\Sigma(f,f) - 2\Sigma(f,f') + \Sigma(f',f') + \big\|\!\big|S(f) - S(f')\big|\!\big|_\psi^2 \le L^2 d(f,f')^2 \quad a.s. \tag{4.7}$$

Then for each $\eta > 0$ there is a process $T(f)$ which, conditional on $\mathcal{H}_0$, is zero-mean and Gaussian, satisfying $\mathbb{E}\big[T(f)T(f') \mid \mathcal{H}_0\big] = \Sigma(f, f')$ for all $f, f' \in \mathcal{F}$, and for all $t > 0$ has

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \big|S(f) - T(f)\big| \geq C_\psi(t+\eta)\right) \leq C_\psi \inf_{\delta>0} \inf_{\mathcal{F}_\delta} \left\{ \frac{\beta_\delta^{1/3}(\log 2|\mathcal{F}_\delta|)^{1/3}}{\eta} + \left(\frac{\sqrt{\log 2|\mathcal{F}_\delta|}\sqrt{\mathbb{E}\left[\|\Omega_\delta\|_2\right]}}{\eta}\right)^{2/3} + \psi\left(\frac{t}{LJ_\psi(\delta)}\right)^{-1} + \exp\left(\frac{-t^2}{L^2 J_2(\delta)^2}\right) \right\}$$

where $\mathcal{F}_\delta$ is any finite δ-cover of $(\mathcal{F}, d)$ and C_ψ is a constant depending only on ψ, with

$$\beta_\delta = \sum_{i=1}^n \mathbb{E}\left[\|\mathcal{F}_\delta(X_i)\|_2^2 \|\mathcal{F}_\delta(X_i)\|_\infty + \|V_i(\mathcal{F}_\delta)^{1/2} Z_i\|_2^2 \|V_i(\mathcal{F}_\delta)^{1/2} Z_i\|_\infty\right],$$

$$V_i(\mathcal{F}_\delta) = \mathbb{E}\big[\mathcal{F}_\delta(X_i)\mathcal{F}_\delta(X_i)^\mathsf{T} \mid \mathcal{H}_{i-1}\big], \qquad \Omega_\delta = \sum_{i=1}^n V_i(\mathcal{F}_\delta) - \Sigma(\mathcal{F}_\delta),$$

$$J_\psi(\delta) = \int_0^\delta \psi^{-1}\big(N_\varepsilon\big)\,\mathrm{d}\varepsilon + \delta\psi^{-1}\big(N_\delta^2\big), \qquad J_2(\delta) = \int_0^\delta \sqrt{\log N_\varepsilon}\,\mathrm{d}\varepsilon,$$

where $N_\delta = N(\delta, \mathcal{F}, d)$ is the δ-covering number of $(\mathcal{F}, d)$ and Z_i are i.i.d. $\mathcal{N}\big(0, I_{|\mathcal{F}_\delta|}\big)$ independent of $\mathcal{H}_n$. If $\mathcal{F}_\delta$ is a minimal δ-cover of $(\mathcal{F}, d)$, then $|\mathcal{F}_\delta| = N_\delta$.

Proposition 4.3.1 is given in a rather general form to accommodate a range of different settings and applications. In particular, consider the following well-known Orlicz functions.

Polynomial: $\psi(x) = x^a$ for $a \geq 2$ has $\|\!|X|\!\|_2 \leq \|\!|X|\!\|_\psi$ and $\sqrt{\log x} \leq \sqrt{a}\psi^{-1}(x)$.

Exponential: $\psi(x) = \exp(x^a) - 1$ for $a \in [1, 2]$ has $\|\!|X|\!\|_2 \leq 2\|\!|X|\!\|_\psi$ and $\sqrt{\log x} \leq \psi^{-1}(x)$.

Bernstein: $\psi(x) = \exp\left(\left(\frac{\sqrt{1+2ax}-1}{a}\right)^2\right) - 1$ for $a > 0$ has $\|\!|X|\!\|_2 \leq (1+a)\|\!|X|\!\|_\psi$ and $\sqrt{\log x} \leq \psi^{-1}(x)$.

For these Orlicz functions and when $\Sigma(f, f') = \mathrm{Cov}[S(f), S(f')]$ is non-random, the terms involving Σ in (4.7) can be controlled by the Orlicz ψ-norm term; similarly, J_2 is bounded by J_ψ. Further, C_ψ can be replaced by a universal constant C which does not depend on the parameter a. See Section 2.2 in van der Vaart and Wellner (1996) for details. If the conditional third moments of $f(X_i)$ given $\mathcal{H}_{i-1}$ are all zero (if f and X_i are appropriately

symmetric, for example), then the second inequality in Corollary 4.2.2 can be applied to obtain a tighter coupling inequality; the details of this are omitted for brevity, and the proof would proceed in exactly the same manner.

In general, however, Proposition 4.3.1 allows for a random covariance function, yielding a coupling to a stochastic process that is Gaussian only conditionally. Such a process can equivalently be viewed as a mixture of Gaussian processes, writing $T = \Sigma^{1/2} Z$ with an operator square root and where Z is a Gaussian white noise on $\mathcal{F}$ independent of $\mathcal{H}_0$. This extension is in contrast with much of the existing strong approximation and empirical process literature, which tends to focus on couplings and weak convergence results with marginally Gaussian processes (Settati, 2009; Chernozhukov, Chetverikov, and Kato, 2016).

A similar approach was taken by Berthet and Mason (2006), who used a Gaussian coupling due to Zaitsev (1987a,b) along with a discretization method to obtain strong approximations for empirical processes with independent data. They handled fluctuations in the stochastic processes with uniform L^2 covering numbers and bracketing numbers where we opt instead for chaining with Orlicz norms. Our version using the martingale Yurinskii coupling can improve upon theirs in approximation rate even for independent data in certain circumstances. Suppose the setup of Proposition 1 in Berthet and Mason (2006); that is, $X_1, \dots, X_n$ are i.i.d. and $\sup_{\mathcal{F}} \|f\|_\infty \leq M$, with the VC-type assumption $\sup_{\mathbb{Q}} N(\varepsilon, \mathcal{F}, d_{\mathbb{Q}}) \leq c_0 \varepsilon^{-\nu_0}$ where $d_{\mathbb{Q}}(f, f')^2 = \mathbb{E}_{\mathbb{Q}}\big[(f - f')^2\big]$ for a measure $\mathbb{Q}$ on $\mathcal{X}$ and M, c_0, ν_0 are constants. Using uniform L^2 covering numbers rather than Orlicz chaining in our Proposition 4 gives the following. Firstly, as X_i are i.i.d., take $\Sigma(f, f') = \operatorname{Cov}[S(f), S(f')]$ so $\Omega_\delta = 0$. Let $\mathcal{F}_\delta$ be a minimal δ-cover of $(\mathcal{F}, d_{\mathbb{P}})$ with cardinality $N_\delta \lesssim \delta^{-\nu_0}$ where $\delta \to 0$. It is easy to show that $\beta_\delta \lesssim n\delta^{-\nu_0}\sqrt{\log(1/\delta)}$. Theorem 2.2.8 and Theorem 2.14.1 in van der Vaart and Wellner (1996) then give

$$\begin{aligned}\mathbb{E}\left[\sup_{d_{\mathbb{P}}(f,f')\leq\delta}\Big(|S(f) - S(f')| + |T(f) - T(f')|\Big)\right] &\lesssim \sup_{\mathbb{Q}} \int_0^\delta \sqrt{n \log N(\varepsilon, \mathcal{F}, d_{\mathbb{Q}})}\,\mathrm{d}\varepsilon \\ &\lesssim \delta\sqrt{n\log(1/\delta)},\end{aligned}$$

where we used the VC-type property to bound the entropy integral. So by our Proposition 4.3.1, for any sequence $R_n \to \infty$ (see Remark 4.2.1),

$$\sup_{f \in \mathcal{F}} \left| S(f) - T(f) \right| \lesssim_{\mathbb{P}} n^{1/3} \delta^{-\nu_0/3} \sqrt{\log(1/\delta)} R_n + \delta \sqrt{n \log(1/\delta)} \lesssim_{\mathbb{P}} n^{\frac{2+\nu_0}{6+2\nu_0}} \sqrt{\log n} R_n,$$

where we minimized over δ in the last step. Berthet and Mason (2006, Proposition 1) achieved

$$\sup_{f \in \mathcal{F}} \left| S(f) - T(f) \right| \lesssim_{\mathbb{P}} n^{\frac{5\nu_0}{4+10\nu_0}} (\log n)^{\frac{4+5\nu_0}{4+10\nu_0}},$$

showing that our approach achieves a better approximation rate whenever $\nu_0 > 4/3$. In particular, our method is superior in richer function classes with larger VC-type dimension. For example, if $\mathcal{F}$ is smoothly parameterized by $\theta \in \Theta \subseteq \mathbb{R}^d$ where Θ contains an open set, then $\nu_0 > 4/3$ corresponds to $d \geq 2$ and our rate is better as soon as the parameter space is more than one-dimensional. The difference in approximation rate is due to Zaitsev's coupling having better dependence on the sample size but worse dependence on the dimension. In particular, Zaitsev's coupling is stated only in ℓ^2-norm and hence Berthet and Mason (2006, Equation 5.3) are compelled to use the inequality $\| \cdot \|_\infty \leq \| \cdot \|_2$ in the coupling step, a bound which is loose when the dimension of the vectors (here on the order of $\delta^{-\nu_0}$) is even moderately large. We use the fact that our version of Yurinskii's coupling applies directly to the supremum norm, giving sharper dependence on the dimension.

In Section 4.4.2 we apply Proposition 4.3.1 to obtain strong approximations for local polynomial estimators in the nonparametric regression setting. In contrast with the series estimators of the upcoming Section 4.4.1, local polynomial estimators are not linearly separable and hence cannot be analyzed directly using the finite-dimensional Corollary 4.2.2.

4.4 Applications to nonparametric regression

We illustrate the applicability of our previous strong approximation results with two substantial and classical examples in nonparametric regression estimation. Firstly, we present an analysis of partitioning-based series estimators, where we can apply Corollary 4.2.2 directly due to an intrinsic linear separability property. Secondly, we consider local polynomial estimators, this time using Proposition 4.3.1 due to a non-linearly separable martingale empirical process.

4.4.1 Partitioning-based series estimators

Partitioning-based least squares methods are essential tools for estimation and inference in nonparametric regression, encompassing splines, piecewise polynomials, compactly supported wavelets and decision trees as special cases. See Cattaneo et al. (2020) for further details and references throughout this section. We illustrate the usefulness of Corollary 4.2.2 by deriving a Gaussian strong approximation for partitioning series estimators based on multivariate martingale data. Proposition 4.4.1 shows how we achieve the best known rate of strong approximation for independent data by imposing an additional mild α-mixing condition to control the time series dependence of the regressors.

Consider the nonparametric regression setup with martingale difference residuals defined by $Y_i = \mu(W_i) + \varepsilon_i$ for $1 \leq i \leq n$ where the regressors W_i have compact connected support $\mathcal{W} \subseteq \mathbb{R}^m$, $\mathcal{H}_i$ is the σ-algebra generated by $(W_1, \ldots, W_{i+1}, \varepsilon_1, \ldots, \varepsilon_i)$, $\mathbb{E}[\varepsilon_i \mid \mathcal{H}_{i-1}] = 0$ and $\mu : \mathcal{W} \to \mathbb{R}$ is the estimand. Let $p(w)$ be a k-dimensional vector of bounded basis functions on $\mathcal{W}$ which are locally supported on a quasi-uniform partition (Cattaneo et al., 2020, Assumption 2). Under minimal regularity conditions, the least-squares partitioning-based series estimator is $\hat{\mu}(w) = p(w)^{\mathsf{T}} \hat{H}^{-1} \sum_{i=1}^{n} p(W_i) Y_i$ with $\hat{H} = \sum_{i=1}^{n} p(W_i) p(W_i)^{\mathsf{T}}$. The approximation power of the estimator $\hat{\mu}(w)$ derives from letting $k \to \infty$ as $n \to \infty$. The assumptions made on $p(w)$ are mild enough to accommodate splines, wavelets, piecewise polynomials, and certain types of decision trees. For such a tree, $p(w)$ is comprised of indicator functions over k axis-aligned rectangles forming a partition of $\mathcal{W}$ (a Haar basis), provided that the partitions are constructed using independent data (e.g., with sample splitting).

Our goal is to approximate the law of the stochastic process $(\hat{\mu}(w) - \mu(w) : w \in \mathcal{W})$, which upon rescaling is typically not asymptotically tight as $k \to \infty$ and thus does not converge weakly. Nevertheless, exploiting the intrinsic linearity of the estimator $\hat{\mu}(w)$, we can apply Corollary 4.2.2 directly to construct a Gaussian strong approximation. Specifically, we write

$$\hat{\mu}(w) - \mu(w) = p(w)^{\mathsf{T}} H^{-1} S + p(w)^{\mathsf{T}} \big(\hat{H}^{-1} - H^{-1}\big) S + \operatorname{Bias}(w),$$

where $H = \sum_{i=1}^{n} \mathbb{E}\left[p(W_i)p(W_i)^{\mathsf{T}}\right]$ is the expected outer product matrix, $S = \sum_{i=1}^{n} p(W_i)\varepsilon_i$ is the score vector, and $\operatorname{Bias}(w) = p(w)^{\mathsf{T}} \hat{H}^{-1} \sum_{i=1}^{n} p(W_i)\mu(W_i) - \mu(w)$. Imposing some mild time series restrictions and assuming stationarity, it is not difficult to show (see Section C.1) that $\|\hat{H} - H\|_1 \lesssim_{\mathbb{P}} \sqrt{nk}$ and $\sup_{w \in \mathcal{W}} |\operatorname{Bias}(w)| \lesssim_{\mathbb{P}} k^{-\gamma}$ for some $\gamma > 0$, depending on the specific structure of the basis functions, the dimension m of the regressors, and the smoothness of the regression function μ. It remains to study the k-dimensional mean-zero martingale S by applying Corollary 4.2.2 with $X_i = p(W_i)\varepsilon_i$. Controlling the convergence of the quadratic variation term $\mathbb{E}[\|\Omega\|_2]$ requires some time series dependence assumptions; we impose an α-mixing condition on $(W_1, \ldots, W_n)$ for illustration (Bradley, 2005).

Proposition 4.4.1 (Strong approximation for partitioning series estimators)
Consider the nonparametric regression setup described above and further assume the following:

(i) $(W_i, \varepsilon_i)_{1 \le i \le n}$ *is strictly stationary.*

(ii) $W_1, \ldots, W_n$ *is* α*-mixing with mixing coefficients satisfying* $\sum_{j=1}^{\infty} \alpha(j) < \infty$.

(iii) W_i *has a Lebesgue density on* $\mathcal{W}$ *which is bounded above and away from zero.*

(iv) $\mathbb{E}\big[|\varepsilon_i|^3\big] < \infty$ *and* $\mathbb{E}\big[\varepsilon_i^2 \mid \mathcal{H}_{i-1}\big] = \sigma^2(W_i)$ *is bounded away from zero.*

(v) $p(w)$ *is a basis with* k *features satisfying Assumptions 2 and 3 in Cattaneo et al. (2020).*

Then, for any sequence $R_n \to \infty$*, there is a zero-mean Gaussian process* $G(w)$ *indexed on* $\mathcal{W}$ *with* $\operatorname{Var}[G(w)] \asymp \frac{k}{n}$ *satisfying* $\operatorname{Cov}[G(w), G(w')] = \operatorname{Cov}[p(w)^{\mathsf{T}} H^{-1} S,\, p(w')^{\mathsf{T}} H^{-1} S]$ *and*

$$\sup_{w \in \mathcal{W}} |\hat{\mu}(w) - \mu(w) - G(w)| \lesssim_{\mathbb{P}} \sqrt{\frac{k}{n}} \left(\frac{k^3 (\log k)^3}{n}\right)^{1/6} R_n + \sup_{w \in \mathcal{W}} |\operatorname{Bias}(w)|$$

assuming the number of basis functions satisfies $k^3/n \to 0$. *If further* $\mathbb{E}\left[\varepsilon_i^3 \mid \mathcal{H}_{i-1}\right] = 0$ *then*

$$\sup_{w \in \mathcal{W}} |\hat{\mu}(w) - \mu(w) - G(w)| \lesssim_{\mathbb{P}} \sqrt{\frac{k}{n}} \left(\frac{k^3 (\log k)^2}{n}\right)^{1/4} R_n + \sup_{w \in \mathcal{W}} |\operatorname{Bias}(w)|.$$

The core concept in the proof of Proposition 4.4.1 is to apply Corollary 4.2.2 with $S = \sum_{i=1}^n p(W_i)\varepsilon_i$ and $p = \infty$ to construct $T \sim \mathcal{N}\big(0, \operatorname{Var}[S]\big)$ such that $\|S - T\|_\infty$ is small, and then setting $G(w) = p(w)^\mathsf{T} H^{-1} T$. So long as the bias can be appropriately controlled, this result allows for uniform inference procedures such as uniform confidence bands or shape specification testing. The condition $k^3/n \to 0$ is the same (up to logs) as that imposed by Cattaneo et al. (2020) for i.i.d. data, which gives the best known strong approximation rate for this problem. Thus, Proposition 4.4.1 gives the same best approximation rate without requiring any extra restrictions for α-mixing time series data.

Our results improve substantially on Li and Liao (2020, Theorem 1): using the notation of our Corollary 4.2.2, and with any sequence $R_n \to \infty$, a valid (see Remark 4.2.1) version of their martingale Yurinskii coupling is

$$\|S - T\|_2 \lesssim_{\mathbb{P}} d^{1/2} r_n^{1/2} + (B_n d)^{1/3} R_n,$$

where $B_n = \sum_{i=1}^n \mathbb{E}[\|X_i\|_2^3]$ and r_n is a term controlling the convergence of the quadratic variation, playing a similar role to our term $\mathbb{E}[\|\Omega\|_2]$. Under the assumptions of our Proposition 4.4.1, applying this result with $S = \sum_{i=1}^n p(W_i)\varepsilon_i$ yields a rate no better than $\|S - T\|_2 \lesssim_{\mathbb{P}} (nk)^{1/3} R_n$. As such, they attain a rate of strong approximation no faster than

$$\sup_{w \in \mathcal{W}} |\hat{\mu}(w) - \mu(w) - G(w)| \lesssim_{\mathbb{P}} \sqrt{\frac{k}{n}} \left(\frac{k^5}{n}\right)^{1/6} R_n + \sup_{w \in \mathcal{W}} |\operatorname{Bias}(w)|.$$

Hence, for this approach to yield a valid strong approximation, the number of basis functions must satisfy $k^5/n \to 0$, a more restrictive assumption than our $k^3/n \to 0$ (up to logs). This difference is due to Li and Liao (2020) using the ℓ^2-norm version of Yurinskii's coupling rather than the recently established ℓ^∞ version. Further, our approach allows for an improved rate of distributional approximation whenever the residuals have zero conditional third moment.

To illustrate the statistical applicability of Proposition 4.4.1, consider constructing a feasible uniform confidence band for the regression function μ, using standardization and Studentization for statistical power improvements. We assume throughout that the bias is negligible. Proposition 4.4.1 and anti-concentration for Gaussian suprema (Chernozhukov et al., 2014a, Corollary 2.1) yield a distributional approximation for the supremum statistic whenever $k^3(\log n)^6/n \to 0$, giving

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\frac{\hat{\mu}(w)-\mu(w)}{\sqrt{\rho(w,w)}}\right|\leq t\right)-\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\frac{G(w)}{\sqrt{\rho(w,w)}}\right|\leq t\right)\right|\to 0,$$

where $\rho(w,w')=\mathbb{E}[G(w)G(w')]$. Further, by a Gaussian–Gaussian comparison result (Chernozhukov, Chetverikov, and Kato, 2013a, Lemma 3.1) and anti-concentration, we show (see the proof of Proposition 4.4.1) that with $\mathbf{W}=(W_1,\ldots,W_n)$ and $\mathbf{Y}=(Y_1,\ldots,Y_n)$,

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\frac{\hat{\mu}(w)-\mu(w)}{\sqrt{\hat{\rho}(w,w)}}\right|\leq t\right)-\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\frac{\hat{G}(w)}{\sqrt{\hat{\rho}(w,w)}}\right|\leq t\;\middle|\;\mathbf{W},\mathbf{Y}\right)\right|\to_{\mathbb{P}} 0,$$

where $\hat{G}(w)$ is a zero-mean Gaussian process conditional on $\mathbf{W}$ and $\mathbf{Y}$ with conditional covariance function $\hat{\rho}(w,w')=\mathbb{E}\big[\hat{G}(w)\hat{G}(w')\mid\mathbf{W},\mathbf{Y}\big]=p(w)^{\mathsf{T}}\hat{H}^{-1}\hat{V}\hat{H}^{-1}p(w')$ for some estimator $\hat{V}$ satisfying $\frac{k(\log n)^2}{n}\big\|\hat{V}-\mathrm{Var}[S]\big\|_2\to_{\mathbb{P}}0$. For example, one could use the plug-in estimator $\hat{V}=\sum_{i=1}^n p(W_i)p(W_i)^{\mathsf{T}}\hat{\sigma}^2(W_i)$ where $\hat{\sigma}^2(w)$ satisfies $(\log n)^2\sup_{w\in\mathcal{W}}|\hat{\sigma}^2(w)-\sigma^2(w)|\to_{\mathbb{P}}0$. This leads to the following feasible and asymptotically valid $100(1-\tau)\%$ uniform confidence band for partitioning-based series estimators based on martingale data.

Proposition 4.4.2 (Feasible uniform confidence bands for partitioning series estimators) *Assume the setup of the preceding section. Then*

$$\mathbb{P}\Big(\mu(w)\in\Big[\hat{\mu}(w)\pm\hat{q}(\tau)\sqrt{\hat{\rho}(w,w)}\Big]\ \textit{for all}\ w\in\mathcal{W}\Big)\to 1-\tau,$$

where

$$\hat{q}(\tau)=\inf\left\{t\in\mathbb{R}:\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\frac{\hat{G}(w)}{\sqrt{\hat{\rho}(w,w)}}\right|\leq t\;\middle|\;\mathbf{W},\mathbf{Y}\right)\geq\tau\right\}$$

is the conditional quantile of the supremum of the Studentized Gaussian process. This can be estimated by resampling the conditional law of $\hat{G}(w) \mid \mathbf{W}, \mathbf{Y}$ *with a discretization of* $w \in \mathcal{W}$.

4.4.2 Local polynomial estimators

As a second example application we consider nonparametric regression estimation with martingale data employing local polynomial methods (Fan and Gijbels, 1996). In contrast with the partitioning-based series methods of Section 4.4.1, local polynomials induce stochastic processes which are not linearly separable, allowing us to showcase the empirical process result given in Proposition 4.3.1.

As before, suppose that $Y_i = \mu(W_i) + \varepsilon_i$ for $1 \leq i \leq n$ where W_i has compact connected support $\mathcal{W} \subseteq \mathbb{R}^m$, $\mathcal{H}_i$ is the σ-algebra generated by $(W_1, \ldots, W_{i+1}, \varepsilon_1, \ldots, \varepsilon_i)$, $\mathbb{E}[\varepsilon_i \mid \mathcal{H}_{i-1}] = 0$, and $\mu : \mathcal{W} \to \mathbb{R}$ is the estimand. Let K be a kernel function on $\mathbb{R}^m$ and $K_h(w) = h^{-m} K(w/h)$ for some bandwidth $h > 0$. Take $\gamma \geq 0$ a fixed polynomial order and let $k = (m+\gamma)!/(m!\gamma!)$ be the number of monomials up to order γ. Using multi-index notation, let $p(w)$ be the k-dimensional vector collecting the monomials $w^\kappa/\kappa!$ for $0 \leq |\kappa| \leq \gamma$, and set $p_h(w) = p(w/h)$. The local polynomial regression estimator of $\mu(w)$ is, with $e_1 = (1, 0, \ldots, 0)^\mathsf{T} \in \mathbb{R}^k$ the first standard unit vector,

$$\hat{\mu}(w) = e_1^\mathsf{T} \hat{\beta}(w) \qquad \text{where} \qquad \hat{\beta}(w) = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} \Big(Y_i - p_h(W_i - w)^\mathsf{T} \beta \Big)^2 K_h(W_i - w).$$

Our goal is again to approximate the distribution of the entire stochastic process, $(\hat{\mu}(w) - \mu(w) : w \in \mathcal{W})$, which upon rescaling is non-Donsker if $h \to 0$, and decomposes as follows:

$$\hat{\mu}(w) - \mu(w) = e_1^\mathsf{T} H(w)^{-1} S(w) + e_1^\mathsf{T} \big(\hat{H}(w)^{-1} - H(w)^{-1} \big) S(w) + \text{Bias}(w)$$

where $\hat{H}(w) = \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) p_h(W_i - w)^\mathsf{T}$, $H(w) = \mathbb{E}\big[\hat{H}(w)\big]$, $S(w) = \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) \varepsilon_i$, and $\text{Bias}(w) = e_1^\mathsf{T} \hat{H}(w)^{-1} \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) \mu(W_i) - \mu(w)$. A key distinctive feature of local polynomial regression is that both $\hat{H}(w)$ and $S(w)$ are functions of the evaluation point $w \in \mathcal{W}$; contrast this with the partitioning-based series

estimator discussed in Section 4.4.1, for which neither $\hat{H}$ nor S depend on w. Therefore we use Proposition 4.3.1 to obtain a Gaussian strong approximation for the martingale empirical process directly.

Under mild regularity conditions, including stationarity for simplicity and an α-mixing assumption on the time-dependence of the data, we show $\sup_{w\in\mathcal{W}} \|\hat{H}(w) - H(w)\|_2 \lesssim_{\mathbb{P}} \sqrt{nh^{-2m}\log n}$. Further, $\sup_{w\in\mathcal{W}} |\operatorname{Bias}(w)| \lesssim_{\mathbb{P}} h^{\gamma}$ provided that the regression function is sufficiently smooth. It remains to analyze the martingale empirical process given by $\big(e_1^{\mathsf{T}} H(w)^{-1} S(w) : w \in \mathcal{W}\big)$ via Proposition 4.3.1 by setting

$$\mathcal{F} = \Big\{ (W_i, \varepsilon_i) \mapsto e_1^{\mathsf{T}} H(w)^{-1} K_h(W_i - w) p_h(W_i - w) \varepsilon_i : w \in \mathcal{W} \Big\}.$$

With this approach, we obtain the following result.

Proposition 4.4.3 (Strong approximation for local polynomial estimators)
Under the nonparametric regression setup described above, assume further that

(i) $(W_i, \varepsilon_i)_{1\le i\le n}$ *is strictly stationary.*

(ii) $(W_i, \varepsilon_i)_{1\le i\le n}$ *is* α*-mixing with mixing coefficients* $\alpha(j) \le e^{-2j/C_\alpha}$ *for some* $C_\alpha > 0$.

(iii) W_i *has a Lebesgue density on* $\mathcal{W}$ *which is bounded above and away from zero.*

(iv) $\mathbb{E}\big[e^{|\varepsilon_i|/C_\varepsilon}\big] < \infty$ *for* $C_\varepsilon > 0$ *and* $\mathbb{E}\left[\varepsilon_i^2 \mid \mathcal{H}_{i-1}\right] = \sigma^2(W_i)$ *is bounded away from zero.*

(v) K *is a non-negative Lipschitz compactly supported kernel with* $\int K(w)\,\mathrm{d}w = 1$.

Then for any $R_n \to \infty$, *there is a zero-mean Gaussian process* $T(w)$ *on* $\mathcal{W}$ *with* $\operatorname{Var}[T(w)] \asymp \frac{1}{nh^m}$ *satisfying* $\operatorname{Cov}[T(w), T(w')] = \operatorname{Cov}[e_1^{\mathsf{T}} H(w)^{-1} S(w),\, e_1^{\mathsf{T}} H(w')^{-1} S(w')]$ *and*

$$\sup_{w\in\mathcal{W}} |\hat{\mu}(w) - \mu(w) - T(w)| \lesssim_{\mathbb{P}} \frac{R_n}{\sqrt{nh^m}} \left(\frac{(\log n)^{m+4}}{nh^{3m}}\right)^{\frac{1}{2m+6}} + \sup_{w\in\mathcal{W}} |\operatorname{Bias}(w)|,$$

provided that the bandwidth sequence satisfies $nh^{3m} \to \infty$.

If the residuals further satisfy $\mathbb{E}\left[\varepsilon_i^3 \mid \mathcal{H}_{i-1}\right] = 0$, then a third-order Yurinskii coupling delivers an improved rate of strong approximation for Proposition 4.4.3; this is omitted here for brevity. For completeness, the proof of Proposition 4.4.3 verifies that if the regression

function $\mu(w)$ is γ times continuously differentiable on $\mathcal{W}$ then $\sup_w |\operatorname{Bias}(w)| \lesssim_{\mathbb{P}} h^\gamma$. Further, the assumption that $p(w)$ is a vector of monomials is unnecessary in general; any collection of bounded linearly independent functions which exhibit appropriate approximation power will suffice (Eggermont and LaRiccia, 2009). As such, we can encompass local splines and wavelets, as well as polynomials, and also choose whether or not to include interactions between the regressor variables. The bandwidth restriction of $nh^{3m} \to \infty$ is analogous to that imposed in Proposition 4.4.1 for partitioning-based series estimators, and as far as we know, has not been improved upon for non-i.i.d. data.

Applying an anti-concentration result for Gaussian process suprema, such as Corollary 2.1 in Chernozhukov et al. (2014a), allows one to write a Kolmogorov–Smirnov bound comparing the law of $\sup_{w\in\mathcal{W}} |\hat{\mu}(w) - \mu(w)|$ to that of $\sup_{w\in\mathcal{W}} |T(w)|$. With an appropriate covariance estimator, we can further replace $T(w)$ by a feasible version $\hat{T}(w)$ or its Studentized counterpart, enabling procedures for uniform inference analogous to the confidence bands constructed in Section 4.4.1. We omit the details of this to conserve space but note that our assumptions on W_i and ε_i ensure that Studentization is possible even when the discretized covariance matrix has small eigenvalues (Section 4.3.1), as we normalize only by the diagonal entries. Chernozhukov et al. (2014b, Remark 3.1) achieve better rates for approximating the supremum of the t-process based on i.i.d. data in Kolmogorov–Smirnov distance by bypassing the step where we first approximate the entire stochastic process (see Section 4.3 for a discussion). Nonetheless, our approach targeting the entire process allows for a potential future treatment of other functionals as well as the supremum.

We finally remark that in this setting of kernel-based local empirical processes, it is essential that our initial strong approximation result (Corollary 4.2.2) does not impose a lower bound on the eigenvalues of the variance matrix Σ. This effect was demonstrated by Lemma 4.3.1, Figure 4.1, and their surrounding discussion in Section 4.3.1. As such, the result of Li and Liao (2020) is unsuited for this application, even in its simplest formulation, due to the strong minimum eigenvalue assumption.

4.5 Conclusion

In this chapter we introduced as our main result a new version of Yurinskii's coupling which strictly generalizes all previously known forms of the result. Our formulation gave a Gaussian mixture coupling for approximate martingale vectors in ℓ^p-norm where $1 \leq p \leq \infty$, with no restrictions on the minimum eigenvalues of the associated covariance matrices. We further showed how to obtain an improved approximation whenever third moments of the data are negligible. We demonstrated the applicability of this main result by first deriving a user-friendly version, and then specializing it to mixingales, martingales, and independent data, illustrating the benefits with a collection of simple factor models. We then considered the problem of constructing uniform strong approximations for martingale empirical processes, demonstrating how our new Yurinskii coupling can be employed in a stochastic process setting. As substantive illustrative applications of our theory to some well-established problems in statistical methodology, we showed how to use our coupling results for both vector-valued and empirical process-valued martingales in developing uniform inference procedures for partitioning-based series estimators and local polynomial models in nonparametric regression. At each stage we addressed issues of feasibility, compared our work with the existing literature, and provided implementable statistical inference procedures. The work in this chapter is based on Cattaneo et al. (2022).

Appendix A

Supplement to Inference with Mondrian Random Forests

In this section we present the full proofs of all our results, and also state some useful technical preliminary and intermediate lemmas, along with some further properties of the Mondrian process not required for our primary analysis. See Section 2.4 in the main text for an overview of the main proof strategies and a discussion of the challenges involved. We use the following simplified notation for convenience, whenever it is appropriate. We write $\mathbb{I}_{ib}(x) = \mathbb{I}\{X_i \in T_b(x)\}$ and $N_b(x) = \sum_{i=1}^n \mathbb{I}_{ib}(x)$, as well as $\mathbb{I}_b(x) = \mathbb{I}\{N_b(x) \geq 1\}$.

A.1 Preliminary lemmas

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

Lemma A.1.1 (Upper bound on the largest cell in a Mondrian forest)
Let $T_1, \ldots, T_b \sim \mathcal{M}\big([0,1]^d, \lambda\big)$ and take $x \in (0,1)^d$. Then for all $t > 0$

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

Proof (Lemma A.1.1)

We use the distribution of the Mondrian cell shape (Mourtada et al., 2020, Proposition 1). We have $|T_b(x)_j| = \left(\frac{E_{bj1}}{\lambda} \wedge x_j\right) + \left(\frac{E_{bj2}}{\lambda} \wedge (1 - x_j)\right)$ where E_{bj1} and E_{bj2} are i.i.d. Exp(1) variables for $1 \leq b \leq B$ and $1 \leq j \leq d$. Thus $|T_b(x)_j| \leq \frac{E_{bj1}+E_{bj2}}{\lambda}$ so by a union bound

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

Next is another localization result, showing that the union of the cells $T_b(x)$ containing x does not contain "too many" samples X_i. Thus the Mondrian random forest estimator fitted at x only depends on n/λ^d (the effective sample size) data points up to logarithmic terms.

Lemma A.1.2 (Upper bound on the number of active data points)
Suppose Assumptions 2.2.1 and 2.2.2 hold, and define $N_{\cup}(x) = \sum_{i=1}^n \mathbb{I}\left\{X_i \in \bigcup_{b=1}^B T_b(x)\right\}$. *Then for* $t > 0$ *and sufficiently large* n, *with* $\|f\|_\infty = \sup_{x\in[0,1]^d} f(x)$,

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

Proof (Lemma A.1.2)

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

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

By Lemma A.1.1, $\mathbb{P}\left(\max_{1\leq b\leq B}\max_{1\leq j\leq d}|T_j(x)| > \frac{t}{\lambda}\right) \leq 2dBe^{-t/2}$. Hence

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

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

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

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

Lemma A.1.3 (An inverse moment bound for the binomial distribution)
For $n \geq 1$ and $p \in [0,1]$, let $N \sim \text{Bin}(n,p)$ and $a_1, \ldots, a_k \geq 0$. Then

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

Proof (Lemma A.1.3)
By Bernstein's inequality, $\mathbb{P}\left(N \leq np - t\right) \leq \exp\left(-\frac{t^2/2}{np(1-p)+t/3}\right) \leq \exp\left(-\frac{3t^2}{6np+2t}\right)$. Therefore we have $\mathbb{P}\left(N \leq np/4\right) \leq \exp\left(-\frac{27n^2p^2/16}{6np+3np/2}\right) = e^{-9np/40}$. Partitioning by this event gives

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

Our next result is probably the most technically involved, allowing one to bound moments of (products of) $\frac{\mathbb{I}_b(x)}{N_b(x)}$ by the corresponding moments of (products of) $\frac{1}{n|T_b(x)|}$, again based on the heuristic that $N_b(x)$ is conditionally binomial so concentrates around its conditional expectation $n\int_{T_b(x)} f(x)\,\mathrm{d}s \asymp n|T_b(x)|$. By independence of the trees, the latter expected products then factorize since the dependence on the data X_i has been eliminated. The proof is complicated, and relies on the following induction procedure. First we consider the common refinement consisting of the subcells $\mathcal{R}$ generated by all possible intersections of $T_b(x)$ over the selected trees (say $T_b(x), T_{b'}(x), T_{b''}(x)$ though there could be arbitrarily many). Note that $N_b(x)$ is the sum of the number of samples X_i in each such subcell in $\mathcal{R}$. We then apply Lemma A.1.3 repeatedly to each subcell in $\mathcal{R}$ in turn, replacing the number of samples X_i in that subcell with its volume multiplied by n, and controlling the error incurred at each step. We record the subcells which have been "checked" in this manner using the class $\mathcal{D} \subseteq \mathcal{R}$ and proceed by finite induction, beginning with $\mathcal{D} = \emptyset$ and ending at $\mathcal{D} = \mathcal{R}$.

Lemma A.1.4 (Generalized moment bound for Mondrian random forest denominators) *Suppose Assumptions 2.2.1 and 2.2.2 hold. Let $T_b \sim \mathcal{M}\big([0,1]^d, \lambda\big)$ be independent and $k_b \geq 1$ for $1 \leq b \leq B_0$. Then with $k = \sum_{b=1}^{B_0} k_b$, for sufficiently large n,*

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

Proof (Lemma A.1.4)

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

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

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

that $N_b(x) = \sum_{D \in \mathcal{A}(b)} N_b(D)$. Note that for any $D \in \mathcal{R} \setminus \mathcal{D}$,

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

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

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

conditional on $\mathbf{T}$ and $N_b(D') : D' \in \mathcal{R} \setminus (\mathcal{D} \cup \{D\})$. For sufficiently large t by Lemma A.1.2

$$\mathbb{P}\left(\sum_{D' \in \mathcal{R} \setminus (\mathcal{D} \cup \{D\})} N_b(D') > t^{d+1} \frac{n}{\lambda^d} \|f\|_\infty\right) \leq \mathbb{P}\left(N_\cup(x) > t^{d+1} \frac{n}{\lambda^d} \|f\|_\infty\right) \leq 4 d B_0 e^{-t/4}.$$

Thus $N_b(D) \geq \mathrm{Bin}(n/2, |D| \inf_x f(x))$ conditional on $\{\mathbf{T}, N_b(D') : D' \in \mathcal{R} \setminus (\mathcal{D} \cup \{D\})\}$ with probability at least $1 - 4dB_0 e^{\frac{-\sqrt{\lambda}}{8\|f\|_\infty}}$. So by Lemma A.1.3,

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

Therefore plugging this back into the marginal expectation yields

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

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

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

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

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

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

Now that moments of (products of) $\frac{\mathbb{I}_b(x)}{N_b(x)}$ have been bounded by moments of (products of) $\frac{1}{n|T_b(x)|}$, we establish further explicit bounds for these in the next result. Note that the problem has been reduced to determining properties of Mondrian cells, so once again we return to the exact cell shape distribution given by Mourtada et al. (2020), and evaluate the appropriate expectations by integration. Note that the truncation by taking the minimum with one inside the expectation is essential here, as otherwise second moment of the inverse Mondrian cell volume is not even finite. As such, there is a "penalty" of $\log n$ when bounding truncated second moments, and the upper bound for the kth moment is significantly larger than the naive assumption of $(\lambda^d/n)^k$ whenever $k \geq 3$. This "small cell" phenomenon in which the inverse volumes of Mondrian cells have heavy tails is a recurring challenge.

Lemma A.1.5 (Inverse moments of the volume of a Mondrian cell)
Suppose Assumption 2.2.2 holds and let $T \sim \mathcal{M}\big([0,1]^d, \lambda\big)$. *Then for sufficiently large* n,

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

Proof (Lemma A.1.5)

By Mourtada et al. (2020, Proposition 1), $|T(x)| = \prod_{j=1}^{d} \left(\left(\frac{1}{\lambda} E_{j1}\right) \wedge x_j + \left(\frac{1}{\lambda} E_{j2}\right) \wedge (1 - x_j) \right)$ where E_{j1} and E_{j2} are mutually independent Exp(1) random variables. Thus for $0 < t < 1$, using the fact that $E_{j1} + E_{j2} \sim \mathrm{Gam}(2, 1)$,

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

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

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

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

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

Lower bounds which match up to constants for the first moment and up to logarithmic terms for the second moment are obtained as $\mathbb{E}\left[1 \wedge \frac{1}{(n|T(x)|)^2}\right] \geq \mathbb{E}\left[1 \wedge \frac{1}{n|T(x)|}\right]^2$ by Jensen, and

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

The endeavor to bound moments of (products of) $\frac{\mathbb{I}_b(x)}{N_b(x)}$ is concluded with the next result, combining the previous two lemmas to give a bound without expectations on the right.

Lemma A.1.6 (Simplified generalized moment bound for Mondrian forest denominators)
Suppose Assumptions 2.2.1 and 2.2.2 hold. Let $T_b \sim \mathcal{M}\big([0,1]^d, \lambda\big)$ *and* $k_b \geq 1$ *for* $1 \leq b \leq B_0$. *Then with* $k = \sum_{b=1}^{B_0} k_b$,

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

for sufficiently large n.

Proof (Lemma A.1.6)

This follows directly from Lemmas A.1.4 and A.1.5. □

Our final preliminary lemma is concerned with further properties of the inverse truncated binomial distribution, again with the aim of analyzing $\frac{\mathbb{I}_b(x)}{N_b(x)}$. This time, instead of merely upper bounding the moments, we aim to give convergence results for those moments, again in terms of moments of $\frac{1}{n|T_b(x)|}$. This time we only need to handle the first and second moment, so this result does not strictly generalize Lemma A.1.3 except in simple cases. The proof is by Taylor's theorem and the Cauchy–Schwarz inequality, using explicit expressions for moments of the binomial distribution and bounds from Lemma A.1.3.

Lemma A.1.7 (Expectation inequalities for the binomial distribution)
Let $N \sim \mathrm{Bin}(n,p)$ *and take* $a, b \geq 1$. *Then*

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

Proof (Lemma A.1.7)

For the first result, Taylor's theorem with Lagrange remainder for $N \mapsto \frac{1}{N+a}$ around np gives

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

for some ξ between np and N. The second term in the expectation is zero-mean, showing the non-negativity part, and the Cauchy–Schwarz inequality for the remaining term gives

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

Now we use $\mathbb{E}\big[(N-np)^4\big] \le np(1+3np)$ and apply Lemma A.1.3 to see that

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

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

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

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

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

Now we use $\mathbb{E}\big[(N-np)^4\big] \leq np(1+3np)$ and apply Lemma A.1.3 to see that

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

A.2 Proofs of main results

A.2.1 Mondrian random forests

We give rigorous proofs of the central limit theorem, bias characterization, and variance estimation results for the Mondrian random forest estimator without debiasing. See Section 2.4 in the main text for details on our approaches to these proofs.

Proof (Theorem 2.3.1)

From the debiased version (Theorem 2.5.1) with $J = 0$, $a_0 = 1$, and $\omega_0 = 1$. $\square$

Proof (Theorem 2.3.2)

Part 1: removing the dependence on the trees

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

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

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

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

using the law of $T(x)_j$ from Mourtada et al. (2020, Proposition 1). By Chebyshev's inequality,

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

Part 2: showing the conditional bias converges in probability

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

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

For the first term in (A.1) to be non-zero, we must have $|N_b(x) - \tilde{N}_{jb}(x)| = 1$. Writing $N_{-jb}(x) = \sum_{i\neq j} \mathbb{I}_{ib}(x)$, assume by symmetry that $\tilde{N}_{jb}(x) = N_{-jb}(x)$ and $N_b(x) = N_{-jb}(x)+1$, and $\mathbb{I}_{jb}(x) = 1$. As f is bounded and μ is Lipschitz, writing $\mathbb{I}_{-jb}(x) = \mathbb{I}\{N_{-jb}(x) \geq 1\}$,

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

For $t > 0$, partition by $\{\max_{1\le l\le d} |T_b(x)_l| \ge t/\lambda\}$ and apply Lemma A.1.1 and Lemma A.1.6:

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

where we set $t = 4\log n$. For the second term in (A.1) we have

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

in the same manner. Hence

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

and so by Chebyshev's inequality,

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

Part 3: computing the limiting bias

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

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

By Lemma A.1.7, as $N_{-ib}(x) \sim \mathrm{Bin}\left(n-1, \int_{T_b(x)} f(s)\,\mathrm{d}s\right)$ given $\mathbf{T}$ and f is bounded below,

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

and also

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

So by Lemmas A.1.1 and A.1.5, since f is Lipschitz and bounded, using Cauchy–Schwarz,

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

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

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

Note that since f and μ are Lipschitz, and by integrating the tail probability given in Lemma A.1.1, the Maclaurin remainder term is bounded by

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

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

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

We evaluate the expectation. By Taylor's theorem, with ν a multi-index, as $f \in \mathcal{H}^\beta$,

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

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

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

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

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

Now by integrating the tail probabilities in Lemma A.1.1, $\mathbb{E}\left[\max_{1\le l\le d}|T_b(x)_l|^\beta\right] \lesssim \frac{1}{\lambda^\beta}$. Therefore, by Lemma A.1.5, writing $T_b(x)^\nu$ for $\int_{T_b(x)}(s-x)^\nu\,\mathrm{d}s$,

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

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

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

So by independence over j,

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

The final step is to replace E_{1j*} by $E_{1j} \sim \text{Exp}(1)$ and similarly for E_{2j*}. For some $C > 0$,

$$\mathbb{P}\left(\bigcup_{j=1}^d (\{E_{1j*} \ne E_{1j}\} \cup \{E_{2j*} \ne E_{2j}\}) \right) \le 2d \, \mathbb{P}\left(\text{Exp}(1) \ge \lambda \min_{1 \le j \le d} (x_j \wedge (1-x_j)) \right) \le 2de^{-C\lambda}.$$

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

$$\begin{aligned} &\mathbb{E}\left[\hat{\mu}(x)\right] - \mu(x) \\ &\quad = \sum_{|\nu'|=1}^{\beta} \sum_{|\nu''|=0}^{\beta} \sum_{|u|=0}^{\beta} \frac{\partial^{\nu'} \mu(x)}{\nu'!} \frac{\partial^{\nu''} f(x)}{\nu''!} \left(\frac{\partial^\nu f(x)}{\nu!} \right)^u \binom{|u|}{u} \frac{(-1)^{|u|}}{f(x)^{|u|+1}} \lambda^{-|\nu'|-|\nu''|-|\nu|\cdot u} \\ &\qquad \times \prod_{j=1}^d \mathbb{E}\left[\frac{E_{2j}^{\nu_j'+\nu_j''+1} + (-1)^{\nu_j'+\nu_j''} E_{1j}^{\nu_j'+\nu_j''+1}}{(\nu_j'+\nu_j''+1)(E_{2j}+E_{1j})} \frac{\left(E_{2j}^{\nu_j+1} + (-1)^{\nu_j} E_{1j}^{\nu_j+1}\right)^u}{(\nu_j+1)^u (E_{2j}+E_{1j})^{|u|}} \right] \\ &\qquad + O\left(\frac{\log n}{\lambda} \frac{\lambda^d}{n} \right) + O\left(\frac{1}{\lambda^\beta} \right), \end{aligned} \tag{A.3}$$

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

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

$$\frac{E_{2j}^{\nu'_j+\nu''_j+1} - (-E_{1j})^{\nu'_j+\nu''_j+1}}{E_{2j} + E_{1j}} \tag{A.4}$$

and

$$\frac{\left(E_{2j}^{\nu_j+1} - (-E_{1j})^{\nu_j+1}\right)^u}{(E_{2j} + E_{1j})^{|u|}} = \prod_{\substack{|\nu|=1 \\ \nu_j u_\nu \text{ even}}}^{\beta} \frac{\left(E_{2j}^{\nu_j+1} - (-E_{1j})^{\nu_j+1}\right)^{u_\nu}}{(E_{2j} + E_{1j})^{u_\nu}} \prod_{\substack{|\nu|=1 \\ \nu_j u_\nu \text{ odd}}}^{\beta} \frac{\left(E_{2j}^{\nu_j+1} - (-E_{1j})^{\nu_j+1}\right)^{u_\nu}}{(E_{2j} + E_{1j})^{u_\nu}}. \tag{A.5}$$

Clearly, $\nu'_j + \nu''_j$ being odd inverts the sign of (A.4). For (A.5), each term in the first product has either ν_j even or u_ν even, so its sign is preserved. Every term in the second product of (A.5) has its sign inverted due to both ν_j and u_ν being odd, but there are an even number of terms, preserving the overall sign. Therefore the expected product of (A.4) and (A.5) is zero by symmetry.

If however $\nu'_j + \nu''_j$ is even, then $\sum_{|\nu|=1}^{\beta} \nu_j u_\nu$ is odd so $|\{\nu : \nu_j u_\nu \text{ is odd}\}|$ is odd. Clearly, the sign of (A.4) is preserved. Again the sign of the first product in (A.5) is preserved, and the sign of every term in (A.5) is inverted. However there are now an odd number of terms in the second product, so its overall sign is inverted. Therefore the expected product of (A.4) and (A.5) is again zero.

Part 4: calculating the second-order bias

Next we calculate some special cases, beginning with the form of the leading second-order bias, where the exponent in λ is $|\nu'| + |\nu''| + u \cdot |\nu| = 2$, proceeding by cases on the values of $|\nu'|$, $|\nu''|$, and $|u|$. Firstly, if $|\nu'| = 2$ then $|\nu''| = |u| = 0$. Note that if any $\nu'_j = 1$ then the expectation in (A.3) is zero. Hence we can assume $\nu'_j \in \{0, 2\}$, yielding

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

where we used that E_{1j} and E_{2j} are independent Exp(1). Next we consider $|\nu'| = 1$ and $|\nu''| = 1$, so $|u| = 0$. Note that if $\nu'_j = \nu''_{j'} = 1$ with $j \neq j'$ then the expectation in (A.3) is zero. So we need only consider $\nu'_j = \nu''_j = 1$, giving

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

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

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

Hence the second-order bias term is

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

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

If $X_i \sim \text{Unif}\left([0,1]^d\right)$ then $f(x) = 1$ and the bias expansion from (A.3) becomes

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

This is zero if any ν_j' is odd, so we group these terms based on the exponent of λ to see

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

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

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

so finally

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

Proof (Theorem 2.3.3)

This follows from the debiased version in Theorem 2.5.3 with $J = 0$, $a_0 = 1$, and $\omega_0 = 1$. $\square$

Proof (Theorem 2.3.4)

By Theorem 2.3.2 and Theorem 2.3.3,

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

The first term now converges weakly to $\mathcal{N}(0,1)$ by Slutsky's theorem, Theorem 2.3.1, and Theorem 2.3.3, while the second term is $o_{\mathbb{P}}(1)$ by assumption. Validity of the confidence interval follows immediately. □

A.2.2 Debiased Mondrian random forests

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

Proof (Theorem 2.5.1)

We use the martingale central limit theorem given by Hall and Heyde (1980, Theorem 3.2). For each $1 \leq i \leq n$ define $\mathcal{H}_{ni}$ to be the filtration generated by $\mathbf{T}$, $\mathbf{X}$, and $(\varepsilon_j : 1 \leq j \leq i)$, noting that $\mathcal{H}_{ni} \subseteq \mathcal{H}_{(n+1)i}$ because B increases weakly as n increases. Let $\mathbb{I}_{ibr}(x) = \mathbb{I}\{X_i \in T_{br}(x)\}$ where $T_{br}(x)$ is the cell containing x in tree b used to construct $\hat{\mu}_r(x)$, and similarly let $N_{br}(x) = \sum_{i=1}^{n} \mathbb{I}_{ibr}(x)$ and $\mathbb{I}_{br}(x) = \mathbb{I}\{N_{br}(x) \geq 1\}$. Define the $\mathcal{H}_{ni}$-measurable and square integrable variables

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

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

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

By Hall and Heyde (1980, Theorem 3.2) it suffices to check that (i) $\max_i |S_i(x)| \to 0$ in probability, (ii) $\mathbb{E}\left[\max_i S_i(x)^2\right] \lesssim 1$, and (iii) $\sum_i S_i(x)^2 \to \Sigma_{\mathrm{d}}(x)$ in probability.

Part 1: checking condition (i)

Since J is fixed and $\mathbb{E}[|\varepsilon_i|^3 \mid X_i]$ is bounded, by Jensen's inequality and Lemma A.1.6,

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

Part 2: checking condition (ii)

Since $\mathbb{E}[\varepsilon_i^2 \mid X_i]$ is bounded and by Lemma A.1.6,

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

Part 3: checking condition (iii)

Next, we have

$$\begin{aligned}\sum_{i=1}^{n} S_i(x)^2 &= \sum_{i=1}^{n}\left(\sqrt{\frac{n}{\lambda^d}}\sum_{r=0}^{J}\omega_r\frac{1}{B}\sum_{b=1}^{B}\frac{\mathbb{I}_{ibr}(x)\varepsilon_i}{N_{br}(x)}\right)^2 \\ &= \frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{r=0}^{J}\sum_{r'=0}^{J}\omega_r\omega_{r'}\sum_{b=1}^{B}\sum_{b'=1}^{B}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)} \\ &= \frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{r=0}^{J}\sum_{r'=0}^{J}\omega_r\omega_{r'}\sum_{b=1}^{B}\left(\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ibr'}(x)\varepsilon_i^2}{N_{br}(x)N_{br'}(x)}+\sum_{b'\neq b}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\right).\end{aligned} \tag{A.6}$$

By boundedness of $\mathbb{E}[\varepsilon_i^2 \mid X_i]$ and Lemma A.1.6, the first term in (A.6) vanishes as

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

For the second term in (A.6), the law of total variance gives

$$\begin{aligned}&\mathrm{Var}\left[\frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{r=0}^{J}\sum_{r'=0}^{J}\omega_r\omega_{r'}\sum_{b=1}^{B}\sum_{b'\neq b}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\right] \\ &\le (J+1)^4\max_{0\le r,r'\le J}\omega_r\omega_{r'}\,\mathrm{Var}\left[\frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{b=1}^{B}\sum_{b'\neq b}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\right] \\ &\lesssim \max_{0\le r,r'\le J}\mathbb{E}\left[\mathrm{Var}\left[\frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{b=1}^{B}\sum_{b'\neq b}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\;\Big|\;\mathbf{X},\mathbf{Y}\right]\right] \\ &\quad+\max_{0\le r,r'\le J}\mathrm{Var}\left[\mathbb{E}\left[\frac{n}{\lambda^d}\frac{1}{B^2}\sum_{i=1}^{n}\sum_{b=1}^{B}\sum_{b'\neq b}\frac{\mathbb{I}_{ibr}(x)\mathbb{I}_{ib'r'}(x)\varepsilon_i^2}{N_{br}(x)N_{b'r'}(x)}\;\Big|\;\mathbf{X},\mathbf{Y}\right]\right]\end{aligned} \tag{A.7}$$

For the first term in (A.7),

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

Since T_{br} is independent of $T_{b'r'}$ given $\mathbf{X},\mathbf{Y}$, the summands are zero whenever $\big|\{b,b',\tilde{b},\tilde{b}'\}\big|=4$. Since $\mathbb{E}[\varepsilon_i^2\mid X_i]$ is bounded and by the Cauchy–Schwarz inequality and Lemma A.1.6,

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

For the second term in (A.7), the random variable inside the variance is a nonlinear function of the i.i.d. variables (X_i,ε_i), so we apply the Efron–Stein inequality (Efron and Stein, 1981). Let $(\tilde{X}_{ij},\tilde{Y}_{ij})=(X_i,Y_i)$ if $i\neq j$ and be an independent copy of (X_j,Y_j), denoted $(\tilde{X}_j,\tilde{Y}_j)$, if $i=j$, and define $\tilde{\varepsilon}_{ij}=\tilde{Y}_{ij}-\mu(\tilde{X}_{ij})$. Write $\tilde{\mathbb{I}}_{ijbr}(x)=\mathbb{I}\big\{\tilde{X}_{ij}\in T_{br}(x)\big\}$ and $\tilde{\mathbb{I}}_{jbr}(x)=\mathbb{I}\big\{\tilde{X}_j\in T_{br}(x)\big\}$, and also $\tilde{N}_{jbr}(x)=\sum_{i=1}^{n}\tilde{\mathbb{I}}_{ijbr}(x)$. We use the leave-one-out notation $N_{-jbr}(x)=\sum_{i\neq j}\mathbb{I}_{ibr}(x)$

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

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

For the first term in the above display, note that

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

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

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

where we used Lemma A.1.6. So $\sum_{i=1}^{n} S_i(x)^2 - n\,\mathbb{E}\left[S_i(x)^2\right] = O_{\mathbb{P}}\left(\frac{1}{\sqrt{B}} + \sqrt{\frac{\lambda^d\log n}{n}}\right) = o_{\mathbb{P}}(1)$.

Part 4: calculating the limiting variance

Thus by Hall and Heyde (1980, Theorem 3.2) we conclude that

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

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

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

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

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

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

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

Since σ^2 is Lipschitz and $\mathbb{P}\left(\max_{1\le l\le d}|T_b(x)_l| \ge t/\lambda\right) \le 2de^{-t/2}$ by Lemma A.1.1,

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

by Lemma A.1.6, where we set $t = 4\log n$. Therefore

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

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

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

arguing using Lemma A.1.1, the Lipschitz property of $f(x)$, and Lemma A.1.6. So

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

Now we apply the binomial result in Lemma A.1.7 to approximate the expectation. With $N_{-ib'r'\setminus br}(x) = \sum_{j\neq i} \mathbb{I}\{X_j \in T_{b'r'}(x) \setminus T_{br}(x)\}$,

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

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

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

We bound these parameters above and below. Firstly, by Lemma A.1.2 with $B=1$,

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

for some $C>0$ and sufficiently large t. Next, if f is L-Lipschitz in ℓ^2, by Lemma A.1.1,

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

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

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

So by Lemma A.1.7 conditionally on $\mathbf{T}$, $N_{-ib'r'\cap br}(x)$, and $N_{-ibr\setminus b'r'}(x)$, we have with probability at least $1 - e^{-t/C}$ that

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

Therefore, by the same approach as the proof of Lemma A.1.4, taking $t = 3C \log n$,

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

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

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

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

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

Thus far we have proven that

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

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

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

Therefore, by Lemma A.1.7 applied conditionally on $\mathbf{T}$, with probability at least $1 - e^{-t/C}$,

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

Now by Lemma A.1.5, with $t = 3C \log n$,

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

This allows us to deduce that

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

Now that we have reduced the limiting variance to an expression only involving the sizes of Mondrian cells, we can exploit their exact distribution to compute this expectation. Recall from Mourtada et al. (2020, Proposition 1) that we can write

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

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

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

and note that

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

for some $C>0$ and sufficiently large λ. So by Cauchy–Schwarz and Lemma A.1.5,

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

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

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

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

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

Secondly, we have in exactly the same manner that

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

Therefore

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

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

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

as $\int_0^\infty \frac{e^{-t}}{a+t}\,\mathrm{d}t = e^a\Gamma(0,a)$ with $\Gamma(0,a) = \int_a^\infty \frac{e^{-t}}{t}\,\mathrm{d}t$. Now

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

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

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

Finally, we conclude by giving the limiting variance.

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

So the limit exists, and with $\ell_{rr'} = \frac{2a_r}{3}\left(1 - \frac{a_r}{a_{r'}}\log\left(\frac{a_{r'}}{a_r}+1\right)\right)$, the limiting variance is

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

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

Proof (Theorem 2.5.2)

By the definition of the debiased estimator and Theorem 2.3.2, since J and a_r are fixed,

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

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

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

Proof (Theorem 2.5.3)

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

Recall that

$$\hat{\sigma}^2(x) = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{i=1}^n Y_i^2 \mathbb{I}\{X_i \in T_b(x)\}}{\sum_{i=1}^n \mathbb{I}\{X_i \in T_b(x)\}} - \hat{\mu}(x)^2. \tag{A.8}$$

The first term in (A.8) is simply a Mondrian forest estimator of $\mathbb{E}[Y_i^2 \mid X_i = x] = \sigma^2(x)+\mu(x)^2$, which is bounded and Lipschitz, where $\mathbb{E}[Y_i^4 \mid X_i]$ is bounded almost surely. So its conditional bias is controlled by Theorem 2.3.2 and is at most $O_{\mathbb{P}}\left(\frac{1}{\lambda} + \frac{\log n}{\lambda}\sqrt{\lambda^d/n}\right)$. Its variance is at most $\frac{\lambda^d}{n}$ by Theorem 2.5.1. Consistency of the second term in (A.8) follows directly from Theorems 2.3.2 and 2.5.1 with the same bias and variance bounds. Therefore

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

Part 2: consistency of the sum

Note that

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

This is exactly the same as the quantity in (A.6), if we were to take ε_i to be ± 1 with equal probability. Thus we immediately have convergence in probability by the proof of Theorem 2.5.1:

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

Part 3: conclusion

By the proof of Theorem 2.5.1 with ε_i being ± 1 with equal probability, and by previous parts,

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

Proof (Theorem 2.5.4)

By Theorem 2.5.2 and Theorem 2.5.3,

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

The first term converges weakly to $\mathcal{N}(0,1)$ by Slutsky's theorem and Theorems 2.5.1 and 2.5.3, while the second is $o_{\mathbb{P}}(1)$ by assumption. Validity of the confidence interval follows. $\square$

Proof (Theorem 2.5.5)

Theorem 2.5.2 and the proof of Theorem 2.5.1 with $J = \lfloor \beta/2 \rfloor$ gives

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

We use here an L^2 version of Theorem 2.5.2 which is immediate from the proof of Theorem 2.3.2, since we leveraged Chebyshev's inequality. Now since $\lambda \asymp n^{\frac{1}{d+2\beta}}$ and $B \gtrsim n^{\frac{2\beta-2}{d+2\beta}}$,

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

A.3 Further properties of the Mondrian process

In section, we state and prove a collection of lemmas concerning various properties of the Mondrian process. While they are not used directly in our analysis of Mondrian random forest estimators, we believe that these results, along with the techniques displayed during their proofs, may be of potential independent interest.

Our analysis of Mondrian random forest estimators in the main text is for the most part conducted pointwise, in the sense that we first fix $x \in [0,1]^d$ and then analyze $\hat{\mu}(x)$. This means that we interact with the Mondrian process only through $T(x)$; that is, the cell in T which contains the point x. As such, we rely only on local properties of T, and may consider just a single Mondrian cell. The lemmas in this section take a more global approach to analyzing the Mondrian process, and we make statements about the entire process T, rather than individual cells $T(x)$. Such results may be useful for a future investigation of the uniform properties of Mondrian forest estimators, as well as being interesting in their own right.

We begin with a tail bound for the number of cells appearing in a Mondrian tree, offering a multiplicative exponential inequality which complements the exact expectation result given in Mourtada et al. (2020, Proposition 2). The resulting bound in probability is the same up to logarithmic terms, and the sharp tail decay is useful in combination with union bounds in our upcoming results.

Lemma A.3.1 (Tail bound for the number of cells in a Mondrian tree)
Let $D \subseteq \mathbb{R}^d$ be a rectangle and $T \sim \mathcal{M}(D, \lambda)$. Writing $\#T$ for the number of cells in T,

$$\mathbb{P}\left(\#T > 3(1+\lambda|D|_1)^d(t+1+d\log(1+\lambda|D|_1))\right) \leq e^{-t}.$$

Proof (Lemma A.3.1)
We refer to this method as the "subcell trick" and attribute it to Mourtada, Gaïffas, and Scornet (2017). For $\varepsilon > 0$, partition D into at most $(1+1/\varepsilon)^d$ cells $D' \in \mathcal{D}_\varepsilon$ with side lengths at most $(|D_1|\varepsilon, \ldots, |D_d|\varepsilon)$. Denote the restriction of a tree T to a subcell D' by $T \cap D'$. Since

a split in T induces a split in at least one $T \cap D'$, by a union bound

$$\mathbb{P}\left(\#T > t\right) \leq \mathbb{P}\left(\sum_{D' \in \mathcal{D}_\varepsilon} \#(T \cap D') > t\right) \leq \sum_{D' \in \mathcal{D}_\varepsilon} \mathbb{P}\left(\#(T \cap D') > \frac{t}{\#\mathcal{D}_\varepsilon}\right).$$

Now $\#(T \cap D')$ is dominated by a Yule process with parameter $|D'|_1$ stopped at time λ (Mourtada et al., 2017, proof of Lemma 2), so using that fact that if $X \sim \text{Yule}(a)$ then $\mathbb{P}(X_t > n) \leq (1 - e^{-at})^{n-1}$,

$$\mathbb{P}\left(\#T > t\right) \leq \#\mathcal{D}_\varepsilon \left(1 - e^{-\lambda |D|_1 \varepsilon}\right)^{t/\#\mathcal{D}_\varepsilon - 1} \leq (1 + 1/\varepsilon)^d (1 - e^{-\lambda |D|_1 \varepsilon})^{t(1+1/\varepsilon)^{-d} - 1}.$$

Set $\varepsilon = \frac{1}{\lambda |D|_1}$, note $1 - 1/e \leq e^{-1/3}$ and replace t by $3(1 + \lambda|D|_1)^d(t + 1 + d\log(1 + \lambda|D|_1))$:

$$\mathbb{P}\left(\#T > t\right) \leq (1 + \lambda|D|_1)^d (1 - 1/e)^{t(1+\lambda|D|_1)^{-d} - 1} \leq 2(1 + \lambda|D|_1)^d e^{-t(1+\lambda|D|_1)^{-d}/3},$$
$$\mathbb{P}\left(\#T > 3(1 + \lambda|D|_1)^d(t + 1 + d\log(1 + \lambda|D|_1))\right) \leq e^{-t}. \qquad \square$$

Next we provide a rigorous justification to the observation that the cells in a Mondrian process should have the same shape distribution, though of course they are not independent. To state and prove this result, we need a way to identify a particular cell by endowing the cells in a Mondrian tree with a natural order.

Definition A.3.1 (Canonical order of cells in a Mondrian tree)
Let $T \sim \mathcal{M}(D, \lambda)$. Each cell in a fixed realization of T can be described by a word from the alphabet $\{l, r\}$, where l indicates the cell to the left of a split and r indicates the cell to the right. For example, if there are no splits we have one cell described by the empty word. After one split there are two cells, denoted l and r. Now suppose that the cell r splits again, giving two splits and three cells, denoted l, rl, and rr. Define the canonical ordering of the cells of T by applying the lexicographic order to their words, with $l < r$. Note that it does not matter which coordinate each split occurs in: in two dimensions, l might refer to the "left" or "bottom" and r to the "right" or "top" cell.

Lemma A.3.2 (Cells in a Mondrian tree have identically distributed shapes)
Let $T \sim \mathcal{M}(D, \lambda)$ with ordered cells $D'_1, \ldots, D'_{\#T}$. For $\varepsilon_1, \ldots, \varepsilon_d \geq 0$ and $1 \leq i \leq k$,

$$\mathbb{P}\left(|D'_{i1}| \leq \varepsilon_1, \ldots, |D'_{id}| \leq \varepsilon_d, \#T = k\right) = \mathbb{P}\left(|D'_{11}| \leq \varepsilon_1, \ldots, |D'_{1d}| \leq \varepsilon_d, \#T = k\right).$$

Marginalizing over $\#T$ with E_j i.i.d. Exp(1)*, Mourtada et al. (2020, Proposition 1) gives*

$$\mathbb{P}\left(|D'_{i1}| > \varepsilon_1, \ldots, |D'_{id}| > \varepsilon_d\right) = \prod_{j=1}^{d} \mathbb{P}\left(\frac{E_j}{\lambda} \wedge |D_j| > \varepsilon_j\right) = \prod_{j=1}^{d} \mathbb{I}\{|D_j| > \varepsilon_j\} e^{-\lambda \varepsilon_j}.$$

We observe a version of the famous Poisson process inspection or waiting time paradox in the sizes of Mondrian cells. The above Lemma A.3.2 shows that for a large enough lifetime λ, the volume of any cell D has the same distribution as the volume of a corner cell, and is asymptotically $\mathbb{E}[|D|] \asymp \mathbb{E}\left[\prod_{j=1}^{d}(E_j/\lambda)\right] = 1/\lambda^d$. This is consistent with Mourtada et al. (2020, Proposition 2) who give $\mathbb{E}[\#T] \asymp \lambda^d$. However, if instead of selecting a cell directly, we instead select a fixed interior point x and query the cell $T(x)$ which contains it, we find that $\mathbb{E}[|T(x)|] \asymp \mathbb{E}\left[\prod_{j=1}^{d}((E_{1j} + E_{2j})/\lambda)\right] = 2^d/\lambda^d$, where E_{1j}, E_{2j} are i.i.d. Exp(1), by Mourtada et al. (2020, Proposition 1). Since $T(x)$ contains x by construction, a size-biasing phenomenon occurs and we see that $T(x)$ is on average larger than a typical Mondrian cell.

Proof (Lemma A.3.2)
Let w be the word associated with the cell $D_i \in T$. Note that $i = 1$ if and only if $r \notin w$, as then D_i is the left child of every split. So suppose $r \in w$. Let $\tilde{w}$ be the word obtained by replacing all occurrences of r in w with an l. Each such replacement corresponds to a split in T. Let $\tilde{T}$ be the same process as T but with the following modification: for each split where a replacement was made, change the uniform random variable S (from the definition of T, see Section 2.2.1) to $1 - S$. Since S is independent of everything else in the construction of T, we observe that $\tilde{T} \sim \mathcal{M}(D, \lambda)$ also. Further, there is almost surely exactly one cell in $\tilde{T}$ which has the same shape as D, as the uniform distribution has no atoms. Denote this cell by $\tilde{D}$ and note that the replacements imply that its word in $\tilde{T}$ is $\tilde{w}$. Thus $\tilde{D} = \tilde{D}_1$ in $\tilde{T}$ and so $(|D_{i1}|, \ldots, |D_{id}|, \#T) = (|\tilde{D}_{11}|, \ldots, |\tilde{D}_{1d}|, \#\tilde{T})$. Equality of the distributions follows. □

As our next result we provide a tail bound for the size of the largest Mondrian cell. The cells within a Mondrian tree are of course not independent, and in fact there should intuitively be some negative correlation between their sizes, due to the fact that they must all fit within the original cell D.

Lemma A.3.3 (Tail bound on largest Mondrian cell)
Let $T \sim \mathcal{M}(D, \lambda)$. For any $\varepsilon > 0$,

$$\mathbb{P}\left(\max_{D' \in T} \max_{1 \le j \le d} |D'_j| > \varepsilon\right) \le 5d(1 + \lambda|D|_1)^{d+1} e^{-\lambda\varepsilon}.$$

Proof (Lemma A.3.3)
Let D_i be the ordered cells of T and take $k \ge 1$. By union bounds and Lemma A.3.2,

$$\begin{aligned}
\mathbb{P}\left(\max_{D' \in T} \max_{1 \le j \le d} |D'_j| > \varepsilon\right) &\le \sum_{l=1}^{k} \mathbb{P}\left(\max_{1 \le i \le l} \max_{1 \le j \le d} |D_{ij}| > \varepsilon, \#T = l\right) + \mathbb{P}\left(\#T > k\right) \\
&\le \sum_{l=1}^{k} \sum_{i=1}^{l} \sum_{j=1}^{d} \mathbb{P}\big(|D_{ij}| > \varepsilon, \#T = l\big) + \mathbb{P}\left(\#T > k\right) \\
&\le \sum_{l=1}^{k} ld\, \mathbb{P}\big(|D_{1j}| > \varepsilon, \#T = l\big) + \mathbb{P}\left(\#T > k\right) \\
&\le kd\, \mathbb{P}\big(|D_{1j}| > \varepsilon\big) + \mathbb{P}\left(\#T > k\right).
\end{aligned}$$

For the first term we use the exact distribution of D_1 from Lemma A.3.2 and for the second term we apply Lemma A.3.1.

$$\begin{aligned}
\mathbb{P}\left(\max_{D' \in T} \max_{1 \le j \le d} |D'_j| > \varepsilon\right) &\le kd\, \mathbb{P}\big(|D_{1j}| > \varepsilon\big) + \mathbb{P}\left(\#T > k\right) \\
&\le kd\, e^{-\lambda\varepsilon} + 2(1 + \lambda|D|_1)^d e^{-k(1+\lambda|D|_1)^{-d}/3}.
\end{aligned}$$

Finally, set $k = \lceil 3\lambda\varepsilon(1+\lambda|D|_1)^d \rceil$ and note the bound is trivial unless $\varepsilon \leq |D|_1$.

$$\begin{aligned}
\mathbb{P}\left(\max_{D'\in T}\max_{1\leq j\leq d}|D'_j| > \varepsilon\right) &\leq \big(3\lambda\varepsilon(1+\lambda|D|_1)^d + 1\big)d\, e^{-\lambda\varepsilon} + 2(1+\lambda|D|_1)^d e^{-\lambda\varepsilon} \\
&\leq 3d(1+\lambda|D|_1)^{d+1}e^{-\lambda\varepsilon} + 2(1+\lambda|D|_1)^d e^{-\lambda\varepsilon} \\
&\leq 5d(1+\lambda|D|_1)^{d+1}e^{-\lambda\varepsilon}. \qquad \square
\end{aligned}$$

For the remainder of this section, we turn our attention to the partitions generated by Mondrian random forests. In particular, we study the refinement generated by overlaying B independent Mondrian processes with possibly different lifetime parameters, and intersecting their resulting individual partitions.

Definition A.3.2 (Partition refinement)

Let $T_1, \ldots, T_B$ be partitions of a set. Their common refinement is

$$\bigwedge_{b=1}^{B} T_b = \left\{ \bigcap_{b=1}^{B} D_b : D_b \in T_b \right\} \setminus \{\emptyset\}.$$

We begin our analysis of Mondrian forest refinements with a pair of simple inequalities for bounding the total number of refined cells in Lemma A.3.4. This result does not depend on the probabilistic structure of the Mondrian process, and holds for any rectangular partitions.

Lemma A.3.4 (Inequalities for refinements of rectangular partitions)

Let $T_1, \ldots, T_B$ be rectangular partitions of a d-dimensional rectangle D. Then

$$\#\bigwedge_{b=1}^{B} T_b \leq \prod_{b=1}^{B} \#T_b, \tag{A.9}$$

and for all $B \leq d$ there exist T_b such that (A.9) *holds with equality. If $\#T_{bj}$ denotes the number of splits made by T_b in dimension j, then*

$$\#\bigwedge_{b=1}^{B} T_b \leq \prod_{j=1}^{d}\left(1 + \sum_{b=1}^{B} \#T_{bj}\right), \tag{A.10}$$

and for all $B \geq d$ there exist T_b such that (A.10) *holds with equality.*

Proof (Lemma A.3.4)

The first inequality (A.9) follows because every cell in $\bigwedge_b T_b$ is the intersection of cells $D_b \in T_b$ for $1 \leq b \leq B$, and there at at most $\prod_{b=1}^{B} \# T_b$ ways to choose these. This bound is achievable when $B \leq d$ by setting T_b to be a tree with splits only in dimension b, so that every such intersection of cells gives a cell in the refinement.

For the second inequality (A.10), we construct a new forest of trees. In particular, for each $1 \leq j \leq d$ define A_j to be the set of locations in D_j where a tree T_b makes a split in dimension j for some b. Define T'_j to be a tree which has splits only in dimension j and at the locations prescribed by A_j. Clearly, since every split in T'_j comes from a split in some T_b in dimension j, we have $\# T'_j \leq 1 + \sum_b \# T_{bj}$. Applying the first inequality to this new forest yields $\# \bigwedge_j T'_j \leq \prod_j \# T'_j \leq \prod_j \left(1 + \sum_b \# T_{bj}\right)$. Finally, note that $\bigwedge_j T'_j$ is a refinement of $\bigwedge_b T_b$ and the result follows. This bound is achievable when $B \geq d$ by letting T_b have splits only in dimension b when $b \leq d$ and to be the trivial partition otherwise. □

The inequalities in Lemma A.3.4 provide rather crude bounds for the number of cells in a Mondrian forest refinement as they do not take into account the random structure. Indeed, it should be clear that the "worst case" scenarios, involving trees which contain splits only in a single direction, should be extremely unlikely under the Mondrian law. In Lemma A.3.5 we confirm this intuition and provide an exact value for the expected number of cells in a Mondrian refinement by direct calculation. This result strictly generalizes the single tree version provided as Mourtada et al. (2020, Proposition 2).

Lemma A.3.5 (Expected number of cells in a Mondrian forest refinement)

Let D be a d-dimensional rectangle and take $\lambda_b > 0$ for $1 \leq b \leq B$. Let $T_b \sim \mathcal{M}(D, \lambda_b)$ be independent. Then the expected number of cells in their refinement is exactly

$$\mathbb{E}\left[\# \bigwedge_{b=1}^{B} T_b\right] = \prod_{j=1}^{d} \left(1 + |D_j| \sum_{b=1}^{B} \lambda_b\right).$$

Proof (Lemma A.3.5)

By Mourtada et al. (2020, Proposition 2) we have the result for a single tree:

$$\mathbb{E}\left[\#T_b\right] = \prod_{j=1}^{d} \left(1 + |D_j|\lambda_b\right). \tag{A.11}$$

We proceed by induction on B. By the tower law,

$$\mathbb{E}\left[\#\bigwedge_{b=1}^{B} T_b\right] = \mathbb{E}\left[\sum_{D' \in T_B} \#\bigwedge_{b=1}^{B-1} (T_b \cap D')\right] = \mathbb{E}\left[\sum_{D' \in T_B} \mathbb{E}\left[\#\bigwedge_{b=1}^{B-1} (T_b \cap D') \;\middle|\; T_B\right]\right].$$

Now by the restriction property of Mondrian processes (Mourtada et al., 2020, Fact 2), observe that $T_b \cap D' \sim \mathcal{M}(D', \lambda_b)$ conditional on T_B. Then by the induction hypothesis,

$$\mathbb{E}\left[\#\bigwedge_{b=1}^{B-1} (T_b \cap D') \;\middle|\; T_B\right] = \prod_{j=1}^{d} \left(1 + |D'_j| \sum_{b=1}^{B-1} \lambda_b\right) = \mathbb{E}\big[\#T_{D'} \mid T_B\big]$$

where $T_{D'} \sim \mathcal{M}\big(D', \sum_{b=1}^{B-1} \lambda_B\big)$ conditional on T_B, by the result for a single tree (A.11). The restriction property finally shows that there exist realizations of $T_{D'}$ which ensure that $\sum_{D' \in T_B} \#T_{D'}$ is equal in distribution to $\#T$, where $T \sim \mathcal{M}(D, \sum_{b=1}^{B} \lambda_b)$, so by (A.11),

$$\mathbb{E}\left[\#\bigwedge_{b=1}^{B} T_b\right] = \mathbb{E}\left[\sum_{D' \in T_B} \mathbb{E}\big[\#T_{D'} \mid T_B\big]\right] = \mathbb{E}\big[\#T\big] = \prod_{j=1}^{d} \left(1 + |D_j| \sum_{b=1}^{B} \lambda_b\right). \qquad \square$$

While the exact expectation calculation in Lemma A.3.5 is neat, sharper control on the tail behavior of the number of cells in a Mondrian refinement is desired. Lemma A.3.6 provides this, again making use of the subcell trick to convert a crude bound based on Lemma A.3.4 into a useful tail inequality. We assume for simplicity that all of the lifetimes are identical.

Lemma A.3.6 (Tail bound on the number of cells in a Mondrian forest refinement)
Let $T_b \sim \mathcal{M}(D, \lambda)$ be i.i.d. for $1 \le b \le B$. Then

$$\mathbb{P}\left(\#\bigwedge_{b=1}^{B} T_b > 3^d 2^{d^2} B^d (1 + \lambda|D|_1)^d t^d\right) \le 2^{d+1} dB (1 + \lambda|D|_1)^d e^{-t}.$$

Proof (Lemma A.3.6)

We begin with a coarse estimate and refine it with the subcell trick. By Lemma A.3.4 (A.10), for any $t > 0$, recalling that $\#T_{bj}$ is the number of splits made by T_b in dimension j,

$$\begin{aligned}\mathbb{P}\left(\#\bigwedge_{b=1}^{B} T_b > t\right) &\le \mathbb{P}\left(\prod_{j=1}^{d}\left(1+\sum_{b=1}^{B}\#T_{bj}\right) > t\right) \le \sum_{j=1}^{d}\mathbb{P}\left(1+\sum_{b=1}^{B}\#T_{bj} > t^{1/d}\right)\\ &\le d\,\mathbb{P}\left(\sum_{b=1}^{B}\#T_b > t^{1/d}\right) \le dB\,\mathbb{P}\left(\#T_b > \frac{t^{1/d}}{B}\right). \qquad \text{(A.12)}\end{aligned}$$

By the subcell trick, partition D into at most $(1+1/\varepsilon)^d$ cells $D' \in \mathcal{D}_\varepsilon$ with side lengths at most $(|D_1|\varepsilon, \ldots, |D_d|\varepsilon)$. As every cell in $\bigwedge_b T_b$ corresponds to at least one cell in $\bigwedge_b (T_b \cap D')$,

$$\mathbb{P}\left(\#\bigwedge_{b=1}^{B} T_b > t\right) \le \mathbb{P}\left(\sum_{D'\in\mathcal{D}_\varepsilon}\#\bigwedge_{b=1}^{B}(T_b\cap D') > t\right) \le \sum_{D'\in\mathcal{D}_\varepsilon}\mathbb{P}\left(\#\bigwedge_{b=1}^{B}(T_b\cap D') > \frac{t}{\#\mathcal{D}_\varepsilon}\right).$$

Applying the coarse estimate (A.12) to $\#\bigwedge_b (T_b \cap D')$ gives

$$\mathbb{P}\left(\#\bigwedge_{b=1}^{B} T_b > t\right) \le dB\sum_{D'\in\mathcal{D}_\varepsilon}\mathbb{P}\left(\#(T_b\cap D') > \frac{t^{1/d}}{B\#\mathcal{D}_\varepsilon^{1/d}}\right).$$

Now apply Lemma A.3.1 and set $\varepsilon = \frac{1}{\lambda|D|_1}$ to obtain

$$\begin{aligned}\mathbb{P}\left(\#\bigwedge_{b=1}^{B} T_b > t\right) &\le dB\sum_{D'\in\mathcal{D}_\varepsilon}\mathbb{P}\left(\#(T_b\cap D') > \frac{t^{1/d}}{B\#\mathcal{D}_\varepsilon^{1/d}}\right)\\ &\le dB\sum_{D'\in\mathcal{D}_\varepsilon} 2(1+\lambda|D'|_1)^d e^{-t^{1/d}\#\mathcal{D}_\varepsilon^{-1/d}B^{-1}(1+\lambda|D'|_1)^{-d}/3}\\ &\le 2dB(1+1/\varepsilon)^d(1+\lambda\varepsilon|D|_1)^d e^{-t^{1/d}(1+1/\varepsilon)^{-1}B^{-1}(1+\lambda\varepsilon|D|_1)^{-d}/3}\\ &\le 2^{d+1}dB(1+\lambda|D|_1)^d e^{-t^{1/d}(1+\lambda|D|_1)^{-1}B^{-1}2^{-d}/3}.\end{aligned}$$

Finally, replacing t by $3^d 2^{d^2} B^d (1+\lambda|D|_1)^d t^d$ we have

$$\mathbb{P}\left(\#\bigwedge_{b=1}^{B} T_b > 3^d 2^{d^2} B^d (1+\lambda|D|_1)^d t^d\right) \le 2^{d+1}dB(1+\lambda|D|_1)^d e^{-t}. \qquad \square$$

Appendix B

Supplement to Dyadic Kernel Density Estimators

This section contains complementary detailed expositions of some of our main results, along with additional technical lemmas which may be of independent interest. We also provide full proofs for all of our theoretical contributions.

B.1 Supplementary main results

In this first section we provide more detailed versions of some of the results presented in the main text, alongside some intermediate lemmas which were skipped for conciseness. We begin with some extra notation used throughout this appendix.

For real vectors, $\|\cdot\|_p$ is the standard ℓ^p-norm defined for $p \in [1, \infty]$. For real square matrices, $\|\cdot\|_p$ is the operator norm induced by the corresponding vector norm. In particular, $\|\cdot\|_1$ is the maximum absolute column sum, $\|\cdot\|_\infty$ is the maximum absolute row sum, and $\|\cdot\|_2$ is the maximum singular value. For real symmetric matrices, $\|\cdot\|_2$ coincides with the maximum absolute eigenvalue. We use $\|\cdot\|_{\max}$ to denote the largest absolute entry of a real matrix. For real-valued functions, $\|\cdot\|_\infty$ denotes the (essential) supremum norm. For a bounded set $\mathcal{X} \subseteq \mathbb{R}$ and $a \geq 0$ we use $[\mathcal{X} \pm a]$ to denote the compact interval $[\inf \mathcal{X} - a,\ \sup \mathcal{X} + a]$. For

measurable subsets of $\mathbb{R}^d$ we use Leb to denote the Lebesgue measure, and for finite sets we use $|\cdot|$ for the cardinality. Write $\sum_i$ for $\sum_{i=1}^n$ when clear from context. Similarly, use $\sum_{i<j}$ for $\sum_{i=1}^{n-1}\sum_{j=i+1}^n$ and $\sum_{i<j<r}$ for $\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{r=j+1}^n$.

B.1.1 Strong approximation

We give a detailed construction of the strong approximation of the dyadic empirical process $\hat{f}_W$. We begin by using the Kómlos–Major–Tusnády (KMT) approximation to obtain a strong approximation for L_n in Lemma B.1.1. Since E_n is an empirical process of i.n.i.d. variables, the KMT approximation is not valid. Instead we apply a conditional version of Yurinskii's coupling to obtain a conditional strong approximation for E_n in Lemma B.1.2, and then construct an unconditional strong approximation for E_n in Lemma B.1.3. These approximations are combined to give a strong approximation for $\hat{f}_W$ in Theorem B.1.1. We do not need to approximate the negligible Q_n.

This section is largely concerned with distributional properties, and, as such, will frequently involve *copies* of processes. We say that X' is a copy of a random variable X if they have the same distribution, though they may be defined on different probability spaces. To ensure that all of the joint distributional properties of such processes are preserved, we also carry over a copy of the latent variables $(\mathbf{A}_n, \mathbf{V}_n)$ to the new space.

Many of the technical details regarding the copying and embedding of stochastic processes are covered by the Vorob'ev–Berkes–Philipp theorem, which is stated and discussed in Lemma B.2.5. In particular, this theorem can be used for random vectors or for stochastic processes indexed by a compact rectangle in $\mathbb{R}^d$ with a.s. continuous sample paths.

We present more detailed versions of Lemmas 3.4.1, 3.4.2, and 3.4.3 as Lemmas B.1.1, B.1.2, and B.1.3 respectively, taking care to describe how copies of the stochastic processes are constructed, and also providing smoothness properties for the resulting sample paths.

Lemma B.1.1 (Strong approximation of L_n)

Suppose that Assumptions 3.2.1 and 3.2.2 hold. For each $n \geq 2$ there exists on some probability space a copy of $\left(\mathbf{A}_n, \mathbf{V}_n, L_n\right)$, denoted $\left(\mathbf{A}'_n, \mathbf{V}'_n, L'_n\right)$, and a mean-zero Gaussian process $Z_n^{L\prime}$ indexed on $\mathcal{W}$ satisfying

$$\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\sqrt{n}L'_n(w) - Z_n^{L\prime}(w)\right| > D_{\text{up}}\frac{t + C_1\log n}{\sqrt{n}}\right) \leq C_2 e^{-C_3 t},$$

for some positive constants C_1, C_2, C_3, and for all $t > 0$. By integration of tail probabilities,

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}}\left|\sqrt{n}L'_n(w) - Z_n^{L\prime}(w)\right|\right] \lesssim \frac{D_{\text{up}}\log n}{\sqrt{n}}.$$

Further, $Z_n^{L\prime}$ has the same covariance structure as $\sqrt{n}L'_n$ in the sense that for all $w, w' \in \mathcal{W}$,

$$\mathbb{E}\left[Z_n^{L\prime}(w)Z_n^{L\prime}(w')\right] = n\mathbb{E}\left[L'_n(w)L'_n(w')\right].$$

It also satisfies the following trajectory regularity property for any $\delta_n \in (0, 1/2]$:

$$\mathbb{E}\left[\sup_{|w-w'|\leq\delta_n}\left|Z_n^{L\prime}(w) - Z_n^{L\prime}(w')\right|\right] \lesssim D_{\text{up}}\delta_n\sqrt{\log 1/\delta_n},$$

and has continuous trajectories. The process $Z_n^{L\prime}$ is a function only of $\mathbf{A}'_n$ and some random noise which is independent of $(\mathbf{A}'_n, \mathbf{V}'_n)$.

Lemma B.1.2 (Conditional strong approximation of E_n)

Suppose Assumptions 3.2.1 and 3.2.2 hold. For $n \geq 2$ and $t_n > 0$ with $|\log t_n| \lesssim \log n$, there exists on some probability space a copy of $\left(\mathbf{A}_n, \mathbf{V}_n, E_n\right)$, denoted $\left(\mathbf{A}'_n, \mathbf{V}'_n, E'_n\right)$, and a process $\tilde{Z}_n^{E\prime}$ which is Gaussian conditional on $\mathbf{A}'_n$ and mean-zero conditional on $\mathbf{A}'_n$, satisfying

$$\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\sqrt{n^2h}E'_n(w) - \tilde{Z}_n^{E\prime}(w)\right| > t_n \;\middle|\; \mathbf{A}'_n\right) \leq C_1 t_n^{-2} n^{-1/2} h^{-3/4}(\log n)^{3/4},$$

$\mathbf{A}'_n$-*almost surely for some constant* $C_1 > 0$. *Setting* $t_n = n^{-1/4}h^{-3/8}(\log n)^{3/8}R_n$ *for any sequence* $R_n \to \infty$ *and taking an expectation gives*

$$\sup_{w\in\mathcal{W}} \left|\sqrt{n^2h}E'_n(w) - \tilde{Z}_n^{E\prime}(w)\right| \lesssim_{\mathbb{P}} n^{-1/4}h^{-3/8}(\log n)^{3/8}R_n.$$

Further, $\tilde{Z}_n^{E\prime}$ *has the same conditional covariance as* $\sqrt{n^2h}E'_n$ *in that for all* $w, w' \in \mathcal{W}$,

$$\mathbb{E}\left[\tilde{Z}_n^{E\prime}(w)\tilde{Z}_n^{E\prime}(w') \mid \mathbf{A}'_n\right] = n^2h\mathbb{E}\left[E'_n(w)E'_n(w') \mid \mathbf{A}'_n\right].$$

It also satisfies the following trajectory regularity property for any $\delta_n \in (0, 1/(2h)]$:

$$\mathbb{E}\left[\sup_{|w-w'|\le\delta_n} \left|\tilde{Z}_n^{E\prime}(w) - \tilde{Z}_n^{E\prime}(w')\right|\right] \lesssim \frac{\delta_n}{h}\sqrt{\log\frac{1}{h\delta_n}},$$

and has continuous trajectories.

Lemma B.1.3 (Unconditional strong approximation of E_n)
Suppose Assumptions 3.2.1 and 3.2.2 hold. Let $\left(\mathbf{A}'_n, \mathbf{V}'_n, \tilde{Z}_n^{E\prime}\right)$ *be defined as in Lemma B.1.2. For each* $n \ge 2$ *there exists (on some probability space) a copy of* $\left(\mathbf{A}'_n, \mathbf{V}'_n, \tilde{Z}_n^{E\prime}\right)$, *denoted* $\left(\mathbf{A}''_n, \mathbf{V}''_n, \tilde{Z}_n^{E\prime\prime}\right)$, *and a centered Gaussian process* $Z_n^{E\prime\prime}$ *satisfying*

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}} \left|\tilde{Z}_n^{E\prime\prime}(w) - Z_n^{E\prime\prime}(w)\right|\right] \lesssim n^{-1/6}(\log n)^{2/3}.$$

Further, $Z_n^{E\prime\prime}$ *has the same (unconditional) covariance structure as* $\tilde{Z}_n^{E\prime\prime}$ *and* $\sqrt{n^2h}E_n$ *in the sense that for all* $w, w' \in \mathcal{W}$,

$$\mathbb{E}\left[Z_n^{E\prime\prime}(w)Z_n^{E\prime\prime}(w')\right] = \mathbb{E}\left[\tilde{Z}_n^{E\prime\prime}(w)\tilde{Z}_n^{E\prime\prime}(w')\right] = n^2h\,\mathbb{E}\left[E_n(w)E_n(w')\right].$$

It also satisfies the following trajectory regularity property for any $\delta_n \in (0, 1/(2h)]$:

$$\mathbb{E}\left[\sup_{|w-w'|\le\delta_n} \left|Z_n^{E\prime\prime}(w) - Z_n^{E\prime\prime}(w')\right|\right] \lesssim \frac{\delta_n}{h}\sqrt{\log\frac{1}{h\delta_n}}.$$

Finally, $Z_n^{E\prime\prime}$ *is independent of* $\mathbf{A}''_n$ *and has continuous trajectories.*

We combine these strong approximations to deduce a coupling for $\hat{f}_W$ in Theorem B.1.1, taking care with independence to ensure the approximating processes are jointly Gaussian.

Theorem B.1.1 (Strong approximation of $\hat{f}_W$)
Suppose that Assumptions 3.2.1 and 3.2.2 hold. For each $n \geq 2$ and any sequence $R_n \to \infty$ there exists on some probability space a centered Gaussian process $Z_n^{f\prime}$ and a copy of $\hat{f}_W$, denoted $\hat{f}'_W$, satisfying

$$\begin{aligned}&\sup_{w\in\mathcal{W}} \left|\hat{f}'_W(w) - \mathbb{E}[\hat{f}'_W(w)] - Z_n^{f\prime}(w)\right| \\ &\quad\lesssim_{\mathbb{P}} n^{-1}\log n + n^{-5/4}h^{-7/8}(\log n)^{3/8}R_n + n^{-7/6}h^{-1/2}(\log n)^{2/3}.\end{aligned}$$

Further, $Z_n^{f\prime}$ has the same covariance structure as $\hat{f}'_W(w)$ in the sense that for all $w, w' \in \mathcal{W}$,

$$\mathbb{E}\big[Z_n^{f\prime}(w)Z_n^{f\prime}(w')\big] = \mathrm{Cov}\left[\hat{f}'_W(w), \hat{f}'_W(w')\right] = \Sigma_n(w, w').$$

It has continuous trajectories satisfying the following regularity property for any $\delta_n \in (0, 1/2]$:

$$\mathbb{E}\left[\sup_{|w-w'|\leq\delta_n}\left|Z_n^{f\prime}(w) - Z_n^{f\prime}(w')\right|\right] \lesssim \frac{D_{\mathrm{up}}}{\sqrt{n}}\delta_n\sqrt{\log\frac{1}{\delta_n}} + \frac{1}{\sqrt{n^2h}}\frac{\delta_n}{h}\sqrt{\log\frac{1}{h\delta_n}}.$$

The main result Theorem 3.4.1 now follows easily using Theorem B.1.1, the bias bound from Theorem 3.2.1, and properties of Σ_n established in Lemma 3.2.4.

B.1.2 Covariance estimation

In this section we carefully construct a consistent estimator for the covariance function Σ_n. Firstly, we characterize Σ_n in Lemma B.1.4. In Lemma B.1.5 we define the estimator and demonstrate that it converges in probability in a suitable sense. In Lemma B.1.6 we give an alternative representation which is more amenable to computation.

Lemma B.1.4 (Covariance structure)

Suppose Assumptions 3.2.1 and 3.2.2 hold. Then Σ_n, as defined in Section 3.2.2, admits the following representations, where $1 \le i < j < r \le n$.

$$\begin{aligned}\Sigma_n(w,w') &= \frac{2}{n(n-1)}\operatorname{Cov}\big[k_h(W_{ij},w),k_h(W_{ij},w')\big] + \frac{4(n-2)}{n(n-1)}\operatorname{Cov}\big[k_h(W_{ij},w),k_h(W_{ir},w')\big]\\ &= \frac{2}{n(n-1)}\operatorname{Cov}\big[k_h(W_{ij},w),k_h(W_{ij},w')\big]\\ &\quad + \frac{4(n-2)}{n(n-1)}\operatorname{Cov}\big[\mathbb{E}[k_h(W_{ij},w)\mid A_i],\mathbb{E}[k_h(W_{ij},w')\mid A_i]\big],\end{aligned}$$

Lemma B.1.5 (Covariance estimation)

Grant Assumptions 3.2.1 and 3.2.2, and suppose $nh \gtrsim \log n$ and $f_W(w) > 0$ on $\mathcal{W}$. Define

$$\begin{aligned}S_{ijr}(w,w') &= \frac{1}{6}\Big(k_h(W_{ij},w)k_h(W_{ir},w') + k_h(W_{ij},w)k_h(W_{jr},w') + k_h(W_{ir},w)k_h(W_{ij},w')\\ &\quad + k_h(W_{ir},w)k_h(W_{jr},w') + k_h(W_{jr},w)k_h(W_{ij},w') + k_h(W_{jr},w)k_h(W_{ir},w')\Big),\\ \hat\Sigma_n(w,w') &= \frac{4}{n^2(n-1)^2}\sum_{i<j}k_h(W_{ij},w)k_h(W_{ij},w') + \frac{24}{n^2(n-1)^2}\sum_{i<j<r}S_{ijr}(w,w')\\ &\quad - \frac{4n-6}{n(n-1)}\hat f_W(w)\hat f_W(w').\end{aligned}$$

Then $\hat\Sigma_n$ is uniformly entrywise-consistent in the sense that

$$\sup_{w,w'\in\mathcal{W}}\left|\frac{\hat\Sigma_n(w,w')-\Sigma_n(w,w')}{\sqrt{\Sigma_n(w,w)+\Sigma_n(w',w')}}\right| \lesssim_{\mathbb{P}} \frac{\sqrt{\log n}}{n}.$$

Lemma B.1.6 (Alternative covariance estimator representation)

Suppose that Assumptions 3.2.1 and 3.2.2 hold, and let $\hat\Sigma_n$ be the covariance estimator defined in Lemma B.1.5. Then the following alternative representation for $\hat\Sigma_n$ holds, which may be easier to compute as it does not involve any triple summations over the data. Let $S_i(w) = \frac{1}{n-1}\sum_{j=1}^{i-1}k_h(W_{ji},w) + \frac{1}{n-1}\sum_{j=i+1}^{n}k_h(W_{ij},w)$ estimate $\mathbb{E}[k_h(W_{ij},w)\mid A_i]$.

$$\begin{aligned}\hat\Sigma_n(w,w') &= \frac{4}{n^2}\sum_{i=1}^{n}S_i(w)S_i(w') - \frac{4}{n^2(n-1)^2}\sum_{i<j}k_h(W_{ij},w)k_h(W_{ij},w')\\ &\quad - \frac{4n-6}{n(n-1)}\hat f_W(w)\hat f_W(w').\end{aligned}$$

We show how to obtain a positive semi-definite estimator $\hat{\Sigma}_n^+$ which is uniformly entrywise-consistent for Σ_n. Define $\hat{\Sigma}_n$ as in Lemma B.1.5 and consider the following optimization problem over bivariate functions.

$$\begin{aligned} \text{minimize:} \quad & \sup_{w,w' \in \mathcal{W}} \left| \frac{M(w,w') - \hat{\Sigma}_n(w,w')}{\sqrt{\hat{\Sigma}_n(w,w) + \hat{\Sigma}_n(w',w')}} \right| \quad \text{over } M : \mathcal{W} \times \mathcal{W} \to \mathbb{R} \\ \text{subject to:} \quad & M \text{ is symmetric and positive semi-definite,} \\ & \left| M(w,w') - M(w,w'') \right| \leq \frac{4}{nh^3} C_{\mathrm{k}} C_{\mathrm{L}} |w' - w''| \text{ for all } w, w', w'' \in \mathcal{W}. \end{aligned} \tag{B.1}$$

Lemma B.1.7 (Consistency of $\hat{\Sigma}_n^+$)

Suppose that Assumptions 3.2.1 and 3.2.2 hold, and that $nh \gtrsim \log n$ *and* $f_W(w) > 0$ *on* $\mathcal{W}$. *Then the optimization problem* (B.1) *has an approximately optimal solution* $\hat{\Sigma}_n^+$ *which is uniformly entrywise-consistent for* Σ_n *in the sense that*

$$\sup_{w,w' \in \mathcal{W}} \left| \frac{\hat{\Sigma}_n^+(w,w') - \Sigma_n(w,w')}{\sqrt{\Sigma_n(w,w) + \Sigma_n(w',w')}} \right| \lesssim_{\mathbb{P}} \frac{\sqrt{\log n}}{n}.$$

The optimization problem (B.1) is stated for functions rather than matrices so is infinite-dimensional. However, when restricting to finite-size matrices, Lemma B.1.7 still holds and does not depend on the size of the matrices. Furthermore, the problem then becomes a semi-definite program and so can be solved to arbitrary precision in polynomial time in the size of the matrices (Laurent and Rendl, 2005).

The Lipschitz-type constraint in the optimization problem (B.1) ensures that $\hat{\Sigma}_n^+$ is sufficiently smooth and is a technicality required by some of the later proofs. In practice this constraint is readily verified.

Lemma B.1.8 (Positive semi-definite variance estimator bounds)

Suppose that Assumptions 3.2.1 and 3.2.2 hold, and that $nh \gtrsim \log n$ *and* $f_W(w) > 0$ *on* $\mathcal{W}$. *Then* $\hat{\Sigma}_n^+(w,w) \geq 0$ *almost surely for all* $w \in \mathcal{W}$ *and*

$$\frac{D_{\mathrm{lo}}^2}{n} + \frac{1}{n^2 h} \lesssim_{\mathbb{P}} \inf_{w \in \mathcal{W}} \hat{\Sigma}_n^+(w,w) \leq \sup_{w \in \mathcal{W}} \hat{\Sigma}_n^+(w,w) \lesssim_{\mathbb{P}} \frac{D_{\mathrm{up}}^2}{n} + \frac{1}{n^2 h}.$$

B.1.3 Feasible uniform confidence bands

We use the strong approximation derived in Section B.1.1 and the positive semi-definite covariance estimator introduced in Section B.1.2 to construct feasible uniform confidence bands. We drop the prime notation for copies of processes in the interest of clarity.

Lemma B.1.9 (Proximity of the standardized and studentized t-statistics)
Let Assumptions 3.2.1 and 3.2.2 hold, and suppose that $nh \gtrsim \log n$ and $f_W(w) > 0$ on $\mathcal{W}$. Define for $w \in \mathcal{W}$ the Studentized t-statistic process

$$\hat{T}_n(w) = \frac{\hat{f}_W(w) - f_W(w)}{\sqrt{\hat{\Sigma}_n^+(w,w)}}.$$

Then

$$\sup_{w\in\mathcal{W}} \left|\hat{T}_n(w) - T_n(w)\right| \lesssim_{\mathbb{P}} \sqrt{\frac{\log n}{n}} \left(\sqrt{\log n} + \frac{\sqrt{n}h^{p\wedge\beta}}{D_{\mathrm{lo}} + 1/\sqrt{nh}}\right) \frac{1}{D_{\mathrm{lo}} + 1/\sqrt{nh}}.$$

Lemma B.1.10 (Feasible Gaussian approximation of the infeasible Gaussian process)
Let Assumptions 3.2.1 and 3.2.2 hold, and suppose that $nh \gtrsim \log n$ and $f_W(w) > 0$ on $\mathcal{W}$. Define a process $\hat{Z}_n^T(w)$ which, conditional on the data $\mathbf{W}_n$, is conditionally mean-zero and conditionally Gaussian, and whose conditional covariance structure is

$$\mathbb{E}\left[\hat{Z}_n^T(w)\hat{Z}_n^T(w') \mid \mathbf{W}_n\right] = \frac{\hat{\Sigma}_n^+(w,w')}{\sqrt{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')}}$$

Then the following conditional Kolmogorov–Smirnov result holds.

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\left(\sup_{w\in\mathcal{W}} \left|Z_n^T(w)\right| \le t\right) - \mathbb{P}\left(\sup_{w\in\mathcal{W}} \left|\hat{Z}_n^T(w)\right| \le t \;\middle|\; \mathbf{W}_n\right)\right| \lesssim_{\mathbb{P}} \frac{n^{-1/6}(\log n)^{5/6}}{D_{\mathrm{lo}}^{1/3} + (nh)^{-1/6}}.$$

Lemma B.1.11 (Feasible Gaussian approximation of the studentized t-statistic)
Let Assumptions 3.2.1, 3.2.2 and 3.4.1 hold, and suppose that $f_W(w) > 0$ on $\mathcal{W}$. Then

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\left(\sup_{w\in\mathcal{W}} \left|\hat{T}_n(w)\right| \le t\right) - \mathbb{P}\left(\sup_{w\in\mathcal{W}} \left|\hat{Z}_n^T(w)\right| \le t \;\middle|\; \mathbf{W}_n\right)\right| \ll_{\mathbb{P}} 1.$$

These intermediate lemmas can be used to establish the valid and feasible uniform confidence bands presented in Theorem 3.5.1 in the main text. See Section B.3 for details.

B.1.4 Counterfactual dyadic density estimation

In this section we give a detailed analysis of the counterfactual estimator of Section 3.7. We begin with an assumption describing the counterfactual setup.

Assumption B.1.1 (Counterfactual data generation)
For each $r \in \{0,1\}$, let $\mathbf{W}_n^r$, $\mathbf{A}_n^r$, and $\mathbf{V}_n^r$ be as in Assumption 3.2.1. Let X_i^r be finitely-supported variables, setting $\mathbf{X}_n^r = (X_1^r, \ldots, X_n^r)$. Suppose that (A_i^r, X_i^r) are independent over $1 \le i \le n$ and that $\mathbf{X}_n^r$ is independent of $\mathbf{V}_n^r$. Assume that $W_{ij}^r \mid X_i^r, X_j^r$ has a Lebesgue density $f_{W|XX}^r(\cdot \mid x_1, x_2) \in \mathcal{H}_{C_{\mathrm{H}}}^{\beta}(\mathcal{W})$ and that X_i^r has positive probability mass function $p_X^r(x)$ on a common support $\mathcal{X}$. Suppose that $(\mathbf{A}_n^0, \mathbf{V}_n^0, \mathbf{X}_n^0)$ and $(\mathbf{A}_n^1, \mathbf{V}_n^1, \mathbf{X}_n^1)$ are independent.

The counterfactual density of W_{ij} in population 1 had X_i, X_j followed population 0 is

$$f_W^{1\triangleright 0}(w) = \mathbb{E}\left[f_{W|XX}^1\big(w \mid X_1^0, X_2^0\big)\right] = \sum_{x_1 \in \mathcal{X}} \sum_{x_2 \in \mathcal{X}} f_{W|XX}^1(w \mid x_1, x_2)\psi(x_1)\psi(x_2)p_X^1(x_1)p_X^1(x_2),$$

with $\psi(x) = p_X^0(x)/p_X^1(x)$ for $x \in \mathcal{X}$. Define the counterfactual dyadic kernel density estimator

$$\hat{f}_W^{1\triangleright 0}(w) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \hat{\psi}(X_i^1)\hat{\psi}(X_j^1)k_h(W_{ij}^1, w),$$

where $\hat{\psi}(x) = \mathbb{I}\{\hat{p}_X^1(x) > 0\}\hat{p}_X^0(x)/\hat{p}_X^1(x)$ and $\hat{p}_X^r(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{X_i^r = x\}$. Since $p_X^r(x) > 0$,

$$\begin{aligned}
\hat{\psi}(x) - \psi(x) &= \frac{\hat{p}_X^0(x) - p_X^0(x)}{p_X^1(x)} - \frac{p_X^0(x)}{p_X^1(x)}\frac{\hat{p}_X^1(x) - p_X^1(x)}{p_X^1(x)} \\
&\quad + \frac{\hat{p}_X^1(x) - p_X^1(x)}{p_X^1(x)}\frac{\hat{p}_X^1(x)p_X^0(x) - \hat{p}_X^0(x)p_X^1(x)}{\hat{p}_X^1(x)p_X^1(x)} \\
&= \frac{1}{n}\sum_{r=1}^{n} \kappa(X_r^0, X_r^1, x) + O_{\mathbb{P}}\left(\frac{1}{n}\right)
\end{aligned}$$

is an asymptotic linear representation where

$$\kappa(X_i^0, X_i^1, x) = \frac{\mathbb{I}\{X_i^0 = x\} - p_X^0(x)}{p_X^1(x)} - \frac{p_X^0(x)}{p_X^1(x)} \frac{\mathbb{I}\{X_i^1 = x\} - p_X^1(x)}{p_X^1(x)}$$

satisfies $\mathbb{E}[\kappa(X_i^0, X_i^1, x)] = 0$. We now establish uniform consistency and feasible strong approximation results for the counterfactual density estimator.

Lemma B.1.12 (Bias of $\hat{f}_W^{1\triangleright 0}$)

Suppose that Assumptions 3.2.1, 3.2.2, and B.1.1 hold. Then

$$\sup_{w \in \mathcal{W}} \left| \mathbb{E}\big[\hat{f}_W^{1\triangleright 0}(w)\big] - f_W^{1\triangleright 0}(w) \right| \lesssim h^{p \wedge \beta} + \frac{1}{n}.$$

Lemma B.1.13 (Hoeffding-type decomposition for $\hat{f}_W^{1\triangleright 0}$)

Suppose that Assumptions 3.2.1, 3.2.2, and B.1.1 hold. With $k_{ij} = k_h(W_{ij}^1, w)$, $\kappa_{ri} = \kappa(X_r^0, X_r^1, X_i^1)$ and $\psi_i = \psi(X_i^1)$, define the projections

$$\begin{aligned}
u &= \mathbb{E}\left[k_{ij}\psi_i\psi_j\right], \\
u_i &= \frac{2}{3}\psi_i \mathbb{E}\left[k_{ij}\psi_j \mid A_i^1\right] + \frac{2}{3}\mathbb{E}\left[k_{jr}\psi_j\kappa_{ir} \mid X_i^0, X_i^1\right] - \frac{2}{3}u, \\
u_{ij} &= \frac{1}{3}\psi_i\psi_j \mathbb{E}\left[k_{ij} \mid A_i^1, A_j^1\right] + \frac{1}{3}\psi_i \mathbb{E}\left[k_{ir}\psi_r \mid A_i^1\right] + \frac{1}{3}\psi_i \mathbb{E}\left[k_{ir}\kappa_{jr} \mid A_i^1, X_j^0, X_j^1\right] \\
&\quad + \frac{1}{3}\kappa_{ji} \mathbb{E}\left[k_{ir}\psi_r \mid A_i^1\right] + \frac{1}{3}\psi_j \mathbb{E}\left[k_{jr}\psi_r \mid A_j^1\right] + \frac{1}{3}\psi_j \mathbb{E}\left[k_{jr}\kappa_{ir} \mid X_i^0, X_i^1, A_j^1\right] \\
&\quad + \frac{1}{3}\kappa_{ij} \mathbb{E}\left[k_{jr}\psi_r \mid A_j^1\right] - u_i - u_j + u, \\
u_{ijr} &= \frac{1}{3}\psi_i\psi_j \mathbb{E}\left[k_{ij} \mid A_i^1, A_j^1\right] + \frac{1}{3}\psi_i\kappa_{rj} \mathbb{E}\left[k_{ij} \mid A_i^1, A_j^1\right] + \frac{1}{3}\psi_j\kappa_{ri} \mathbb{E}\left[k_{ij} \mid A_i^1, A_j^1\right] \\
&\quad + \frac{1}{3}\psi_i\psi_r \mathbb{E}\left[k_{ir} \mid A_i^1, A_r^1\right] + \frac{1}{3}\psi_i\kappa_{jr} \mathbb{E}\left[k_{ir} \mid A_i^1, A_r^1\right] + \frac{1}{3}\psi_r\kappa_{ji} \mathbb{E}\left[k_{ir} \mid A_i^1, A_r^1\right] \\
&\quad + \frac{1}{3}\psi_j\psi_r \mathbb{E}\left[k_{jr} \mid A_j^1, A_r^1\right] + \frac{1}{3}\psi_j\kappa_{ir} \mathbb{E}\left[k_{jr} \mid A_j^1, A_r^1\right] + \frac{1}{3}\psi_r\kappa_{ij} \mathbb{E}\left[k_{jr} \mid A_j^1, A_r^1\right] \\
&\quad - u_{ij} - u_{ir} - u_{jr} + u_i + u_j + u_r - u, \\
v_{ijr} &= \frac{1}{3}k_{ij}\big(\psi_i\psi_j + \psi_i\kappa_{rj} + \psi_j\kappa_{ri}\big) + \frac{1}{3}k_{ir}\big(\psi_i\psi_r + \psi_i\kappa_{jr} + \psi_r\kappa_{ji}\big) \\
&\quad + \frac{1}{3}k_{jr}\big(\psi_j\psi_r + \psi_j\kappa_{ir} + \psi_r\kappa_{ij}\big).
\end{aligned}$$

With $l_i^{1\triangleright 0}(w) = u_i$ *and* $e_{ijr}^{1\triangleright 0}(w) = v_{ijr} - u_{ijr}$*, set*

$$L_n^{1\triangleright 0}(w) = \frac{3}{n}\sum_{i=1}^{n} l_i^{1\triangleright 0}(w) \qquad \text{and} \qquad E_n^{1\triangleright 0}(w) = \frac{6}{n(n-1)(n-2)}\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{r=i+1}^{n} e_{ijr}^{1\triangleright 0}(w).$$

Then the following Hoeffding-type decomposition holds, where $O_{\mathbb{P}}(1/n)$ *is uniform in* $w \in \mathcal{W}$*.*

$$\hat{f}_W^{1\triangleright 0}(w) = \mathbb{E}\big[\hat{f}_W^{1\triangleright 0}(w)\big] + L_n^{1\triangleright 0}(w) + E_n^{1\triangleright 0}(w) + O_{\mathbb{P}}\left(\frac{1}{n}\right).$$

Further, the stochastic processes $L_n^{1\triangleright 0}$ *and* $E_n^{1\triangleright 0}$ *are mean-zero and orthogonal in* $L^2(\mathbb{P})$*. Define the upper and lower degeneracy constants as*

$$D_{\text{up}}^{1\triangleright 0} = \limsup_{n\to\infty}\sup_{w\in\mathcal{W}} \text{Var}\left[l_i^{1\triangleright 0}(w)\right]^{1/2} \qquad \text{and} \qquad D_{\text{lo}}^{1\triangleright 0} = \liminf_{n\to\infty}\inf_{w\in\mathcal{W}} \text{Var}\left[l_i^{1\triangleright 0}(w)\right]^{1/2}.$$

Lemma B.1.14 (Uniform consistency of $\hat{f}_W^{1\triangleright 0}$)

Suppose that Assumptions 3.2.1, 3.2.2, and B.1.1 hold. Then

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}}\left|\hat{f}_W^{1\triangleright 0}(w) - f_W^{1\triangleright 0}(w)\right|\right] \lesssim h^{p\wedge\beta} + \frac{D_{\text{up}}^{1\triangleright 0}}{\sqrt{n}} + \sqrt{\frac{\log n}{n^2 h}}.$$

Lemma B.1.15 (Strong approximation of $\hat{f}_W^{1\triangleright 0}$)

On an appropriately enlarged probability space and for any sequence $R_n \to \infty$*, there exists a mean-zero Gaussian process* $Z_n^{f,1\triangleright 0}$ *with the same covariance structure as* $\hat{f}_W^{1\triangleright 0}(w)$ *satisfying*

$$\begin{aligned}&\sup_{w\in\mathcal{W}}\left|\hat{f}_W^{1\triangleright 0}(w) - \mathbb{E}\big[\hat{f}_W^{1\triangleright 0}(w)\big] - Z_n^{f,1\triangleright 0}(w)\right| \\ &\quad\lesssim_{\mathbb{P}} n^{-1}\log n + n^{-5/4}h^{-7/8}(\log n)^{3/8}R_n + n^{-7/6}h^{-1/2}(\log n)^{2/3}.\end{aligned}$$

Lemma B.1.16 (Counterfactual covariance structure)

Writing k'_{ij} for $k_h(W^1_{ij}, w')$ etc., the counterfactual covariance function is

$$\begin{aligned}
\Sigma_n^{1\triangleright 0}(w,w') &= \operatorname{Cov}\left[\hat{f}_W^{1\triangleright 0}(w), \hat{f}_W^{1\triangleright 0}(w')\right] \\
&= \frac{4}{n}\mathbb{E}\left[\left(\psi_i \mathbb{E}\left[k_{ij}\psi_j \mid A_i^1\right] + \mathbb{E}\left[k_{rj}\psi_r\kappa_{ij} \mid X_i^0, X_i^1\right]\right)\right. \\
&\qquad \left.\times\left(\psi_i \mathbb{E}\left[k'_{ij}\psi_j \mid A_i^1\right] + \mathbb{E}\left[k'_{rj}\psi_r\kappa_{ij} \mid X_i^0, X_i^1\right]\right)\right] \\
&\quad + \frac{2}{n^2}\mathbb{E}\left[k_{ij}k'_{ij}\psi_i^2\psi_j^2\right] - \frac{4}{n}\mathbb{E}\left[k_{ij}\psi_i\psi_j\right]\mathbb{E}\left[k'_{ij}\psi_i\psi_j\right] + O\left(\frac{1}{n^{3/2}} + \frac{1}{\sqrt{n^4 h}}\right).
\end{aligned}$$

Lemma B.1.17 (Gaussian approximation of the standardized counterfactual t-statistic)

Let Assumptions 3.2.1, 3.2.2, and B.1.1 hold, and suppose $f_W^{1\triangleright 0}(w) > 0$ on $\mathcal{W}$. Define

$$T_n^{1\triangleright 0}(w) = \frac{\hat{f}_W^{1\triangleright 0}(w) - f_W^{1\triangleright 0}(w)}{\sqrt{\Sigma_n^{1\triangleright 0}(w,w)}} \quad \text{and} \quad Z_n^{T,1\triangleright 0}(w) = \frac{Z_n^{f,1\triangleright 0}(w)}{\sqrt{\Sigma_n^{1\triangleright 0}(w,w)}}.$$

Then with $R_n \to \infty$ as in Lemma B.1.15,

$$\begin{aligned}
&\sup_{w\in\mathcal{W}}\left|T_n^{1\triangleright 0}(w) - Z_n^{T,1\triangleright 0}(w)\right| \\
&\quad \lesssim_{\mathbb{P}} \frac{n^{-1/2}\log n + n^{-3/4}h^{-7/8}(\log n)^{3/8}R_n + n^{-2/3}h^{-1/2}(\log n)^{2/3} + n^{1/2}h^{p\wedge\beta}}{D_{\mathrm{lo}}^{1\triangleright 0} + 1/\sqrt{nh}}.
\end{aligned}$$

Theorem B.1.2 (Infeasible counterfactual uniform confidence bands)

Let Assumptions 3.2.1, 3.2.2, 3.4.1, and B.1.1 hold and suppose that $f_W^{1\triangleright 0}(w) > 0$ on $\mathcal{W}$. Let $\alpha \in (0,1)$ be a confidence level and define $q_{1-\alpha}^{1\triangleright 0}$ as the quantile satisfying

$$\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|Z_n^{T,1\triangleright 0}(w)\right| \le q_{1-\alpha}^{1\triangleright 0}\right) = 1-\alpha.$$

Then

$$\mathbb{P}\left(f_W^{1\triangleright 0}(w) \in \left[\hat{f}_W^{1\triangleright 0}(w) \pm q_{1-\alpha}^{1\triangleright 0}\sqrt{\Sigma_n^{1\triangleright 0}(w,w)}\right] \text{ for all } w \in \mathcal{W}\right) \to 1-\alpha.$$

We propose an estimator for the counterfactual covariance function $\Sigma_n^{1\triangleright 0}$. First let

$$\hat{\kappa}(X_i^0, X_i^1, x) = \frac{\mathbb{I}\{X_i^0 = x\} - \hat{p}_X^0(x)}{\hat{p}_X^1(x)} - \frac{\hat{p}_X^0(x)}{\hat{p}_X^1(x)} \frac{\mathbb{I}\{X_i^1 = x\} - \hat{p}_X^1(x)}{\hat{p}_X^1(x)},$$

and define the leave-out conditional expectation estimators

$$\begin{aligned}
S_i^{1\triangleright 0}(w) &= \hat{\mathbb{E}}\left[k_h(W_{ij}^1, w)\psi(X_j^1) \mid A_i^1\right] \\
&= \frac{1}{n-1}\left(\sum_{j=1}^{i-1} k_h(W_{ji}^1, w)\hat{\psi}(X_j^1) + \sum_{j=i+1}^{n} k_h(W_{ij}^1, w)\hat{\psi}(X_j^1)\right), \\
\tilde{S}_i^{1\triangleright 0}(w) &= \hat{\mathbb{E}}\left[k_h(W_{rj}^1, w)\psi(X_r^1)\kappa(X_i^0, X_i^1, X_j^1) \mid X_i^0, X_i^1\right] \\
&= \frac{1}{n-1}\sum_{j=1}^{n} \mathbb{I}\{j \neq i\}\hat{\kappa}(X_i^0, X_i^1, X_j^1) S_j^{1\triangleright 0}(w).
\end{aligned}$$

Then set

$$\begin{aligned}
\hat{\Sigma}_n^{1\triangleright 0}(w, w') = \frac{4}{n^2}\sum_{i=1}^{n}\left(\hat{\psi}(X_i^1)S_i^{1\triangleright 0}(w) + \tilde{S}_i^{1\triangleright 0}(w)\right)\left(\hat{\psi}(X_i^1)S_i^{1\triangleright 0}(w') + \tilde{S}_i^{1\triangleright 0}(w')\right) \\
- \frac{4}{n^3(n-1)}\sum_{i<j} k_h(W_{ij}^1, w)k_h(W_{ij}^1, w')\hat{\psi}(X_i^1)^2\hat{\psi}(X_j^1)^2 - \frac{4}{n}\hat{f}_W^{1\triangleright 0}(w)\hat{f}_W^{1\triangleright 0}(w').
\end{aligned}$$

We use a positive semi-definite approximation to $\hat{\Sigma}_n^{1\triangleright 0}$, denoted by $\hat{\Sigma}_n^{+,1\triangleright 0}$, and omit the proof of consistency for brevity. To construct feasible uniform confidence bands, define a process $\hat{Z}_n^{T,1\triangleright 0}(w)$ which, conditional on the data $\mathbf{W}_n^1$, $\mathbf{X}_n^0$, and $\mathbf{X}_n^1$ is conditionally mean-zero and conditionally Gaussian, and whose conditional covariance structure is

$$\mathbb{E}\left[\hat{Z}_n^{T,1\triangleright 0}(w)\hat{Z}_n^{T,1\triangleright 0}(w') \mid \mathbf{W}_n^1, \mathbf{X}_n^0, \mathbf{X}_n^1\right] = \frac{\hat{\Sigma}_n^{+,1\triangleright 0}(w, w')}{\sqrt{\hat{\Sigma}_n^{+,1\triangleright 0}(w, w)\hat{\Sigma}_n^{+,1\triangleright 0}(w', w')}}.$$

Let $\alpha \in (0, 1)$ be a confidence level and define $\hat{q}_{1-\alpha}^{1\triangleright 0}$ as the conditional quantile satisfying

$$\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{Z}_n^{T,1\triangleright 0}(w)\right| \leq \hat{q}_{1-\alpha}^{1\triangleright 0} \;\middle|\; \mathbf{W}_n^1, \mathbf{X}_n^0, \mathbf{X}_n^1\right) = 1 - \alpha.$$

Then assuming that the covariance estimator is appropriately consistent, we have that

$$\mathbb{P}\left(f_W^{1\triangleright 0}(w) \in \left[\hat{f}_W^{1\triangleright 0}(w) \pm \hat{q}_{1-\alpha}^{1\triangleright 0}\sqrt{\hat{\Sigma}_n^{+,1\triangleright 0}(w,w)}\right] \text{ for all } w \in \mathcal{W}\right) \to 1-\alpha.$$

B.2 Technical lemmas

We present some lemmas which provide the technical foundations for our main results. These lemmas are stated in as much generality as is reasonable, and may be of independent interest.

B.2.1 Maximal inequalities for i.n.i.d. empirical processes

Firstly, we provide a maximal inequality for empirical processes of independent but not necessarily identically distributed (i.n.i.d.) random variables, indexed by a class of functions. This result is an extension of Theorem 5.2 from Chernozhukov et al. (2014b), which only covers i.i.d. random variables, and is proven in the same manner. Such a result is useful in the study of dyadic data because when conditioning on latent variables, we may encounter random variables which are conditionally independent but which do not necessarily follow the same conditional distribution.

Lemma B.2.1 (A maximal inequality for i.n.i.d. empirical processes)
Let $X_1, \ldots, X_n$ be independent but not necessarily identically distributed (i.n.i.d.) random variables taking values in a measurable space $(S, \mathcal{S})$. Denote the joint distribution of $X_1, \ldots, X_n$ by $\mathbb{P}$ and the marginal distribution of X_i by $\mathbb{P}_i$, and let $\bar{\mathbb{P}} = n^{-1}\sum_i \mathbb{P}_i$. Let $\mathcal{F}$ be a class of Borel measurable functions from S to $\mathbb{R}$ which is pointwise measurable (i.e. it contains a countable subclass which is dense under pointwise convergence). Let F be a strictly positive measurable envelope function for $\mathcal{F}$ (i.e. $|f(s)| \leq |F(s)|$ for all $f \in \mathcal{F}$ and $s \in S$). For a distribution $\mathbb{Q}$ and some $q \geq 1$, define the $(\mathbb{Q}, q)$-norm of $f \in \mathcal{F}$ as $\|f\|_{\mathbb{Q},q}^q = \mathbb{E}_{X\sim\mathbb{Q}}[f(X)^q]$ and suppose $\|F\|_{\bar{\mathbb{P}},2} < \infty$. For $f \in \mathcal{F}$ define the empirical process

$$G_n(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \big(f(X_i) - \mathbb{E}[f(X_i)]\big).$$

Let $\sigma > 0$ *satisfy* $\sup_{f\in\mathcal{F}} \|f\|_{\bar{\mathbb{P}},2} \le \sigma \le \|F\|_{\bar{\mathbb{P}},2}$ *and* $M = \max_{1\le i\le n} F(X_i)$. *Then with* $\delta = \sigma/\|F\|_{\bar{\mathbb{P}},2} \in (0,1]$,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|G_n(f)\right|\right] \lesssim \|F\|_{\bar{\mathbb{P}},2}\, J\big(\delta,\mathcal{F},F\big) + \frac{\|M\|_{\mathbb{P},2}\, J(\delta,\mathcal{F},F)^2}{\delta^2\sqrt{n}},$$

where $\lesssim$ *is up to a universal constant, and* $J(\delta,\mathcal{F},F)$ *is the covering integral*

$$J\big(\delta,\mathcal{F},F\big) = \int_0^\delta \sqrt{1+\sup_{\mathbb{Q}} \log N(\mathcal{F},\rho_{\mathbb{Q}},\varepsilon\|F\|_{\mathbb{Q},2})}\,\mathrm{d}\varepsilon,$$

with the supremum taken over finite discrete probability measures $\mathbb{Q}$ *on* $(S,\mathcal{S})$.

Lemma B.2.2 (A VC class maximal inequality for i.n.i.d. empirical processes)
Assume the same setup as in Lemma B.2.1, and suppose that $\mathcal{F}$ *forms a VC-type class in that*

$$\sup_{\mathbb{Q}} N(\mathcal{F},\rho_{\mathbb{Q}},\varepsilon\|F\|_{\mathbb{Q},2}) \le (C_1/\varepsilon)^{C_2}$$

for all $\varepsilon\in(0,1]$, *for some constants* $C_1 \ge e$ *(where* e *is the standard exponential constant) and* $C_2 \ge 1$. *Then for* $\delta\in(0,1]$ *we have the covering integral bound* $J\big(\delta,\mathcal{F},F\big) \le 3\delta\sqrt{C_2\log(C_1/\delta)}$, *and so by Lemma B.2.1, up to a universal constant,*

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|G_n(f)\right|\right] &\lesssim \sigma\sqrt{C_2\log(C_1/\delta)} + \frac{\|M\|_{\mathbb{P},2} C_2\log(C_1/\delta)}{\sqrt{n}} \\ &\lesssim \sigma\sqrt{C_2\log\big(C_1\|F\|_{\bar{\mathbb{P}},2}/\sigma\big)} + \frac{\|M\|_{\mathbb{P},2} C_2\log\big(C_1\|F\|_{\bar{\mathbb{P}},2}/\sigma\big)}{\sqrt{n}}.\end{aligned}$$

B.2.2 Strong approximation results

Next we provide two strong approximation results. The first is a corollary of the KMT approximation (Komlós et al., 1975) which applies to bounded-variation functions of i.i.d. variables. The second is an extension of the Yurinskii coupling (Belloni et al., 2019) which applies to Lipschitz functions of i.n.i.d. variables.

Lemma B.2.3 (A KMT approximation corollary)

For $n \geq 1$ let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables and $g_n : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a function satisfying the total variation bound $\sup_{x \in \mathbb{R}} \|g_n(\cdot, x)\|_{\mathrm{TV}} < \infty$. Then on some probability space there exist independent copies of $X_1, \ldots, X_n$, denoted $X_1', \ldots, X_n'$, and a mean-zero Gaussian process $Z_n(x)$ such that if we define the empirical process

$$G_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(g_n(X_i', x) - \mathbb{E}\big[g_n(X_i', x)\big] \Big),$$

then for some universal positive constants C_1, C_2, and C_3,

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} \big| G_n(x) - Z_n(x) \big| > \sup_{x \in \mathbb{R}} \|g_n(\cdot, x)\|_{\mathrm{TV}} \, \frac{t + C_1 \log n}{\sqrt{n}} \right) \leq C_2 e^{-C_3 t}.$$

Further, Z_n has the same covariance structure as G_n in the sense that for all x, $x' \in \mathbb{R}$,

$$\mathbb{E}\big[Z_n(x) Z_n(x')\big] = \mathbb{E}\big[G_n(x) G_n(x')\big].$$

By independently sampling from the law of Z_n conditional on $X_1', \ldots, X_n'$, we can assume that Z_n is a function only of $X_1', \ldots, X_n'$ and some independent random noise.

Lemma B.2.4 (Yurinskii coupling for Lipschitz i.n.i.d. empirical processes)

For $n \geq 2$ let $X_1, \ldots, X_n$ be independent but not necessarily identically distributed (i.n.i.d.) random variables taking values in a measurable space $(S, \mathcal{S})$ and let $\mathcal{X}_n \subseteq \mathbb{R}$ be a compact interval with $|\log \mathrm{Leb}(\mathcal{X}_n)| \leq C_1 \log n$ where $C_1 > 0$ is a constant. Let g_n be measurable on $S \times \mathcal{X}_n$ satisfying $\sup_{\xi \in S} \sup_{x \in \mathcal{X}_n} |g_n(\xi, x)| \leq M_n$ and $\sup_{x \in \mathcal{X}_n} \max_{1 \leq i \leq n} \mathrm{Var}[g_n(X_i, x)] \leq \sigma_n^2$, with $|\log M_n| \leq C_1 \log n$ and $\big|\log \sigma_n^2\big| \leq C_1 \log n$. Suppose that g_n satisfies the following uniform Lipschitz condition:

$$\sup_{\xi \in S} \sup_{x, x' \in \mathcal{X}_n} \left| \frac{g_n(\xi, x) - g_n(\xi, x')}{x - x'} \right| \leq l_{n,\infty},$$

and also the following L^2 Lipschitz condition:

$$\sup_{x,x' \in \mathcal{X}_n} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\frac{g_n(X_i,x)-g_n(X_i,x')}{x-x'}\right|^2\right]^{1/2} \le l_{n,2},$$

where $0 < l_{n,2} \le l_{n,\infty}$, $|\log l_{n,2}| \le C_1 \log n$, and $|\log l_{n,\infty}| \le C_1 \log n$. Then for any $t_n > 0$ with $|\log t_n| \le C_1 \log n$, there is a probability space carrying independent copies of $X_1, \ldots, X_n$ denoted $X_1', \ldots, X_n'$ and a mean-zero Gaussian process $Z_n(x)$ such that if we define the empirical process $G_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(g_n(X_i', x) - \mathbb{E}[g_n(X_i', x)]\big)$, then

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x \in \mathcal{X}_n} \big|G_n(x) - Z_n(x)\big| > t_n\right) \\
&\quad \le \frac{C_2 \sigma_n \sqrt{\mathrm{Leb}(\mathcal{X}_n)} \sqrt{\log n} \sqrt{M_n + \sigma_n \sqrt{\log n}}}{n^{1/4} t_n^2} \sqrt{l_{n,2}\sqrt{\log n} + \frac{l_{n,\infty}}{\sqrt{n}} \log n}
\end{aligned}$$

where $C_2 > 0$ is a constant depending only on C_1. Further, Z_n has the same covariance structure as G_n in the sense that for all $x, x' \in \mathcal{X}_n$,

$$\mathbb{E}\big[Z_n(x) Z_n(x')\big] = \mathbb{E}\big[G_n(x) G_n(x')\big].$$

B.2.3 The Vorob'ev–Berkes–Philipp theorem

We present a generalization of the Vorob'ev–Berkes–Philipp theorem (Dudley, 1999) which allows one to "glue" multiple random variables or stochastic processes onto the same probability space, while preserving some pairwise distributions. We begin with some definitions.

Definition B.2.1 (Tree)

A tree *is a finite undirected graph which is connected and contains no cycles or self-loops.*

Definition B.2.2 (Polish Borel probability space)

A Polish Borel probability space *is a triple $(\mathcal{X}, \mathcal{F}, \mathbb{P})$, where $\mathcal{X}$ is a Polish space (a topological space metrizable by a complete separable metric), $\mathcal{F}$ is the Borel σ-algebra induced on $\mathcal{X}$ by its topology, and $\mathbb{P}$ is a probability measure on $(\mathcal{X}, \mathcal{F})$. Important examples of Polish spaces*

include $\mathbb{R}^d$ and the Skorokhod space $\mathcal{D}[0,1]^d$ for $d \geq 1$. In particular, one can consider vectors of real-valued random variables or stochastic processes indexed by compact subsets of $\mathbb{R}^d$ which have almost surely continuous trajectories.

Definition B.2.3 (Projection of a law)

Let $(\mathcal{X}_1, \mathcal{F}_1)$ and $(\mathcal{X}_2, \mathcal{F}_2)$ be measurable spaces, and let $\mathbb{P}_{12}$ be a law on the product space $(\mathcal{X}_1 \times \mathcal{X}_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$. The projection *of $\mathbb{P}_{12}$ onto $\mathcal{X}_1$ is the law $\mathbb{P}_1$ defined on $(\mathcal{X}_1, \mathcal{F}_1)$ by $\mathbb{P}_1 = \mathbb{P}_{12} \circ \pi_1^{-1}$ where $\pi_1(x_1, x_2) = x_1$ is the first-coordinate projection.*

Lemma B.2.5 (Vorob'ev–Berkes–Philipp theorem, tree form)

Let $\mathcal{T}$ be a tree with vertex set $\mathcal{V} = \{1, \ldots, n\}$ and edge set $\mathcal{E}$. Suppose that attached to each vertex i is a Polish Borel probability space $(\mathcal{X}_i, \mathcal{F}_i, \mathbb{P}_i)$. Suppose that attached to each edge $(i,j) \in \mathcal{E}$ (where $i < j$ without loss of generality) is a law $\mathbb{P}_{ij}$ on $(\mathcal{X}_i \times \mathcal{X}_j, \mathcal{F}_i \otimes \mathcal{F}_j)$. Assume that these laws are pairwise-consistent in the sense that the projection of $\mathbb{P}_{ij}$ onto $\mathcal{X}_i$ (resp. $\mathcal{X}_j$) is $\mathbb{P}_i$ (resp. $\mathbb{P}_j$) for each $(i,j) \in \mathcal{E}$. Then there exists a law $\mathbb{P}$ on

$$\left(\prod_{i=1}^{n} \mathcal{X}_i, \bigotimes_{i=1}^{n} \mathcal{F}_i \right)$$

such that the projection of $\mathbb{P}$ onto $\mathcal{X}_i \times \mathcal{X}_j$ is $\mathbb{P}_{ij}$ for each $(i,j) \in \mathcal{E}$, and therefore also the projection of $\mathbb{P}$ onto $\mathcal{X}_i$ is $\mathbb{P}_i$ for each $i \in \mathcal{V}$.

Remark B.2.1

The requirement that $\mathcal{T}$ must contain no cycles is necessary in general. To see this, consider the Polish Borel probability spaces given by $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_3 = \{0,1\}$, their respective Borel σ-algebras, and the pairwise-consistent probability measures:

$$1/2 = \mathbb{P}_1(0) = \mathbb{P}_2(0) = \mathbb{P}_3(0)$$

$$1/2 = \mathbb{P}_{12}(0,1) = \mathbb{P}_{12}(1,0) = \mathbb{P}_{13}(0,1) = \mathbb{P}_{13}(1,0) = \mathbb{P}_{23}(0,1) = \mathbb{P}_{23}(1,0).$$

Each measure $\mathbb{P}_i$ places equal mass on 0 and 1, while $\mathbb{P}_{ij}$ asserts that each pair of realizations is a.s. not equal. The graph of these laws forms a triangle, which is not a tree. Suppose that (X_1, X_2, X_3) has distribution given by $\mathbb{P}$, where $X_i \sim \mathbb{P}_i$ and $(X_i, X_j) \sim \mathbb{P}_{ij}$ for each i, j. But then by definition of $\mathbb{P}_{ij}$ we have $X_1 = 1 - X_2 = X_3 = 1 - X_1$ a.s., which is a contradiction.

Remark B.2.2

Two important applications of Lemma B.2.5 include the embedding of a random vector into a stochastic process and the coupling of stochastic processes onto the same probability space:

(i) Let X_1 and X_2 be stochastic processes with trajectories in $\mathcal{D}[0,1]$. For $x_1, \ldots, x_n \in [0,1]$ let $\tilde{X}_1 = (X_1(x_1), \ldots, X_1(x_n))$ be a random vector and suppose that $\tilde{X}_1'$ is a copy of $\tilde{X}_1$. Then there is a law $\mathbb{P}$ on $\mathcal{D}[0,1] \times \mathbb{R}^n \times \mathcal{D}[0,1]$ such that restriction of $\mathbb{P}$ to $\mathcal{D}[0,1] \times \mathbb{R}^n$ is the law of $(X_1, \tilde{X}_1)$, while the restriction of $\mathbb{P}$ to $\mathbb{R}^n \times \mathcal{D}[0,1]$ is the law of $(\tilde{X}_1', X_2)$. In other words, we can embed the vector $\tilde{X}_1'$ into a stochastic process X_1 while maintaining the joint distribution of $\tilde{X}_1'$ and X_2.

(ii) Let $X_1, X_1', \ldots, X_n, X_n'$ be stochastic processes with trajectories in $\mathcal{D}[0,1]$, where X_i' is a copy of X_i for each $1 \leq i \leq n-1$. Suppose that $\mathbb{P}\big(\|X_{i+1} - X_i'\| > t\big) \leq r_i$ for each $1 \leq i \leq n-1$, where $\|\cdot\|$ is a norm on $\mathcal{D}[0,1]$. Then there exist copies of $X_1, \ldots, X_n$ denoted $X_1'', \ldots, X_n''$ satisfying $\mathbb{P}\big(\|X_{i+1}'' - X_i''\| > t\big) \leq r_i$ for each $1 \leq i \leq n$. That is, all of the inequalities can be satisfied simultaneously on the same probability space.

B.3 Proofs

We present full proofs of all the results stated in Chapter 3 and Appendix B.

B.3.1 Preliminary lemmas

In this section we list some results in probability and U-statistic theory which are used in proofs of our main results. Other auxiliary lemmas will be introduced when they are needed.

Lemma B.3.1 (Bernstein's inequality for independent random variables)

Let $X_1, \ldots, X_n$ be independent real-valued random variables with $\mathbb{E}[X_i] = 0$, $|X_i| \leq M$, and $\mathbb{E}[X_i^2] \leq \sigma^2$, where M and σ are non-random. Then for all $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2 n}{2\sigma^2 + \frac{2}{3}Mt}\right).$$

Proof (Lemma B.3.1)

See for example Lemma 2.2.9 in van der Vaart and Wellner (1996). □

Lemma B.3.2 (The matrix Bernstein inequality)

For $1 \leq i \leq n$ let X_i be independent symmetric $d \times d$ real random matrices with expected values $\mu_i = \mathbb{E}[X_i]$. Suppose that $\|X_i - \mu_i\|_2 \leq M$ almost surely for all $1 \leq i \leq n$ where M is non-random, and define $\sigma^2 = \left\|\sum_i \mathbb{E}[(X_i - \mu_i)^2]\right\|_2$. Then there exists a universal constant $C > 0$ such that for any $t > 0$ and $q \geq 1$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n}(X_i - \mu_i)\right\|_2 \geq 2\sigma\sqrt{t} + \frac{4}{3}Mt\right) \leq 2de^{-t},$$

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n}(X_i - \mu_i)\right\|_2^q\right]^{1/q} \leq C\sigma\sqrt{q + \log 2d} + CM(q + \log 2d).$$

Another simplified version of this is as follows: suppose that $\|X_i\|_2 \leq M$ almost surely, so that $\|X_i - \mu_i\|_2 \leq 2M$. Then since $\sigma^2 \leq nM^2$, we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n}(X_i - \mu_i)\right\|_2 \geq 4M\big(t + \sqrt{nt}\big)\right) \leq 2de^{-t},$$

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n}(X_i - \mu_i)\right\|_2^q\right]^{1/q} \leq CM\big(q + \log 2d + \sqrt{n(q + \log 2d)}\big).$$

Proof (Lemma B.3.2)

See Lemma 3.2 in Minsker and Wei (2019). □

Lemma B.3.3 (A maximal inequality for Gaussian vectors)

Take $n \geq 2$. Let $X_i \sim \mathcal{N}(0, \sigma_i^2)$ for $1 \leq i \leq n$ with $\sigma_i^2 \leq \sigma^2$. Then

$$\mathbb{E}\left[\max_{1\leq i\leq n} X_i\right] \leq \sigma\sqrt{2\log n}, \tag{B.2}$$

$$\mathbb{E}\left[\max_{1\leq i\leq n} |X_i|\right] \leq 2\sigma\sqrt{\log n}. \tag{B.3}$$

If Σ_1 and Σ_2 are constant positive semi-definite $n \times n$ matrices and $N \sim \mathcal{N}(0, I_n)$, then

$$\mathbb{E}\Big[\big\|\Sigma_1^{1/2}N - \Sigma_2^{1/2}N\big\|_\infty\Big] \leq 2\sqrt{\log n}\,\big\|\Sigma_1 - \Sigma_2\big\|_2^{1/2}. \tag{B.4}$$

If further Σ_1 is positive definite, then

$$\mathbb{E}\Big[\big\|\Sigma_1^{1/2}N - \Sigma_2^{1/2}N\big\|_\infty\Big] \leq \sqrt{\log n}\,\lambda_{\min}(\Sigma_1)^{-1/2}\,\big\|\Sigma_1 - \Sigma_2\big\|_2. \tag{B.5}$$

Proof (Lemma B.3.3)

For $t > 0$, Jensen's inequality on the concave logarithm function gives

$$\begin{aligned}\mathbb{E}\left[\max_{1\leq i\leq n} X_i\right] &= \frac{1}{t}\mathbb{E}\left[\log\exp\max_{1\leq i\leq n} tX_i\right] \leq \frac{1}{t}\log\mathbb{E}\left[\exp\max_{1\leq i\leq n} tX_i\right] \leq \frac{1}{t}\log\sum_{i=1}^n \mathbb{E}\left[\exp tX_i\right] \\ &= \frac{1}{t}\log\sum_{i=1}^n \exp\left(\frac{t^2\sigma_i^2}{2}\right) \leq \frac{1}{t}\log n + \frac{t\sigma^2}{2},\end{aligned}$$

by the Gaussian moment generating function. Minimizing with $t = \sqrt{2\log n}/\sigma$ yields (B.2). For (B.3), we use the symmetry of the Gaussian distribution:

$$\mathbb{E}\left[\max_{1\leq i\leq n} |X_i|\right] = \mathbb{E}\left[\max_{1\leq i\leq n}\{X_i, -X_i\}\right] \leq \sigma\sqrt{2\log 2n} \leq 2\sigma\sqrt{\log n}.$$

For (B.4) and (B.5), note that $\Sigma_1^{1/2}N - \Sigma_2^{1/2}N$ is Gaussian with covariance matrix $\left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right)^2$. The variances of of its components are the diagonal elements of this matrix, namely

$$\sigma_i^2 = \mathrm{Var}\left[\left(\Sigma_1^{1/2}N - \Sigma_2^{1/2}N\right)_i\right] = \left(\left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right)^2\right)_{ii}.$$

Note that if e_i is the ith standard unit basis vector, then for any real symmetric matrix A, we have $e_i^\mathsf{T} A^2 e_i = (A^2)_{ii}$, so in particular $(A^2)_{ii} \le \|A\|_2^2$. Therefore

$$\sigma_i^2 \le \left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_2^2 =: \sigma^2.$$

Applying (B.3) then gives

$$\mathbb{E}\Big[\left\|\Sigma_1^{1/2} N - \Sigma_2^{1/2} N\right\|_\infty\Big] \le 2\sqrt{\log n}\left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_2.$$

By Theorem X.1.1 in Bhatia (1997), we can deduce

$$\left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_2 \le \left\|\Sigma_1 - \Sigma_2\right\|_2^{1/2},$$

giving (B.4). If Σ_1 is positive definite, Theorem X.3.8 in Bhatia (1997) gives (B.5):

$$\left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_2 \le \frac{1}{2}\lambda_{\min}(\Sigma_1)^{-1/2}\left\|\Sigma_1 - \Sigma_2\right\|_2. \qquad \square$$

Lemma B.3.4 (Maximal inequalities for Gaussian processes)
Let Z be a separable mean-zero Gaussian process indexed by $x \in \mathcal{X}$. Recall that Z is separable for example if $\mathcal{X}$ is Polish and Z has continuous trajectories. Define its covariance structure on $\mathcal{X} \times \mathcal{X}$ by $\Sigma(x, x') = \mathbb{E}[Z(x)Z(x')]$, and the corresponding semimetric on $\mathcal{X}$ by

$$\rho(x, x') = \mathbb{E}\big[\big(Z(x) - Z(x')\big)^2\big]^{1/2} = \big(\Sigma(x,x) - 2\Sigma(x,x') + \Sigma(x',x')\big)^{1/2}.$$

Let $N(\varepsilon, \mathcal{X}, \rho)$ denote the ε-covering number of $\mathcal{X}$ with respect to the semimetric ρ. Define $\sigma = \sup_x \Sigma(x,x)^{1/2}$. Then there exists a universal constant $C > 0$ such that for any $\delta > 0$,

$$\mathbb{E}\left[\sup_{x\in\mathcal{X}} |Z(x)|\right] \le C\sigma + C\int_0^{2\sigma} \sqrt{\log N(\varepsilon, \mathcal{X}, \rho)}\,\mathrm{d}\varepsilon,$$
$$\mathbb{E}\left[\sup_{\rho(x,x')\le\delta} |Z(x) - Z(x')|\right] \le C\int_0^{\delta} \sqrt{\log N(\varepsilon, \mathcal{X}, \rho)}\,\mathrm{d}\varepsilon.$$

Proof (Lemma B.3.4)

See Corollary 2.2.8 in van der Vaart and Wellner (1996), noting that for any $x, x' \in \mathcal{X}$, we have $\mathbb{E}[|Z(x)|] \lesssim \sigma$ and $\rho(x, x') \leq 2\sigma$, implying that $\log N(\varepsilon, \mathcal{X}, \rho) = 0$ for all $\varepsilon > 2\sigma$. □

Lemma B.3.5 (Anti-concentration for Gaussian process absolute suprema)

Let Z be a separable mean-zero Gaussian process indexed by a semimetric space $\mathcal{X}$ with $\mathbb{E}[Z(x)^2] = 1$ for all $x \in \mathcal{X}$. Then for any $\varepsilon > 0$,

$$\sup_{t \in \mathbb{R}} \mathbb{P}\left(\left| \sup_{x \in \mathcal{X}} \left| Z(x) \right| - t \right| \leq \varepsilon \right) \leq 4\varepsilon \left(1 + \mathbb{E}\left[\sup_{x \in \mathcal{X}} \left| Z(x) \right| \right] \right).$$

Proof (Lemma B.3.5)

See Corollary 2.1 in Chernozhukov et al. (2014a). □

Lemma B.3.6 (No slowest rate of convergence in probability)

Let X_n be a sequence of real-valued random variables with $X_n = o_{\mathbb{P}}(1)$. Then there exists a deterministic sequence $\varepsilon_n \to 0$ such that $\mathbb{P}\big(|X_n| > \varepsilon_n\big) \leq \varepsilon_n$ for all $n \geq 1$.

Proof (Lemma B.3.6)

Define the following deterministic sequence for $k \geq 1$.

$$\tau_k = \sup\big\{ n \geq 1 : \mathbb{P}\big(|X_n| > 1/k\big) > 1/k \big\} \vee (\tau_{k-1} + 1)$$

with $\tau_0 = 0$. Since $X_n = o_{\mathbb{P}}(1)$, each τ_k is finite and so we can define $\varepsilon_n = \frac{1}{k}$ where $\tau_k < n \leq \tau_{k+1}$. Then, noting that $\varepsilon_n \to 0$, we have $\mathbb{P}\big(|X_n| > \varepsilon_n\big) = \mathbb{P}\big(|X_n| > 1/k\big) \leq 1/k = \varepsilon_n$. □

Lemma B.3.7 (General second-order Hoeffding-type decomposition)

Let $\mathcal{U}$ be a vector space. Let $u_{ij} \in \mathcal{U}$ be defined for $1 \leq i, j \leq n$ and $i \neq j$. Suppose that $u_{ij} = u_{ji}$ for all i, j. Then for any $u_i \in \mathcal{U}$ (for $1 \leq i \leq n$) and any $u \in \mathcal{U}$, the following decomposition holds:

$$\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \big(u_{ij} - u\big) = 2(n-1) \sum_{i=1}^{n} \big(u_i - u\big) + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \big(u_{ij} - u_i - u_j + u\big).$$

Proof (Lemma B.3.7)

We compute the left hand side minus the right hand side, beginning by observing that all of the u_{ij} and u terms clearly cancel.

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j\neq i}^{n}\big(u_{ij}-u\big)-2(n-1)\sum_{i=1}^{n}\big(u_i-u\big)-\sum_{i=1}^{n}\sum_{j\neq i}\big(u_{ij}-u_i-u_j+u\big)\\
&\quad=-2(n-1)\sum_{i=1}^{n}u_i-\sum_{i=1}^{n}\sum_{j\neq i}^{n}\big(-u_i-u_j\big)=-2(n-1)\sum_{i=1}^{n}u_i+\sum_{i=1}^{n}\sum_{j\neq i}^{n}u_i+\sum_{j=1}^{n}\sum_{i\neq j}^{n}u_j\\
&\quad=-2(n-1)\sum_{i=1}^{n}u_i+(n-1)\sum_{i=1}^{n}u_i+(n-1)\sum_{j=1}^{n}u_j=0.
\end{aligned}$$

□

Lemma B.3.8 (A U-statistic concentration inequality)

Let $(S,\mathcal{S})$ be a measurable space and $X_1,\ldots,X_n$ be i.i.d. S-valued random variables. Let $H:S^m\to\mathbb{R}$ be a function of m variables satisfying the symmetry property $H(x_1,\ldots,x_m)=H(x_{\tau(1)},\ldots,x_{\tau(m)})$ for any m-permutation τ. Suppose also that $\mathbb{E}[H(X_1,\ldots,X_m)]=0$. Let $M=\|H\|_\infty$ and $\sigma^2=\mathbb{E}\big[\mathbb{E}[H(X_1,\ldots,X_m)\mid X_1]^2\big]$. Define the U-statistic

$$U_n=\frac{m!(n-m)!}{n!}\sum_{1\leq i_1<\cdots<i_m\leq n}H(X_1,\ldots,X_n).$$

Then for any $t>0$, with $C_1(m)$, $C_2(m)$ positive constants depending only on m,

$$\mathbb{P}\left(|U_n|>t\right)\leq 4\exp\left(-\frac{nt^2}{C_1(m)\sigma^2+C_2(m)Mt}\right).$$

Proof (Lemma B.3.8)

See Theorem 2 in Arcones (1995). □

Lemma B.3.9 (A second-order U-process maximal inequality)

Let $X_1 \ldots, X_n$ be i.i.d. random variables taking values in a measurable space $(S, \mathcal{S})$ with distribution $\mathbb{P}$. Let $\mathcal{F}$ be a class of measurable functions from $S \times S$ to $\mathbb{R}$ which is also pointwise measurable. Define the degenerate second-order U-process

$$U_n(f) = \frac{2}{n(n-1)} \sum_{i<j} \Big(f(X_i, X_j) - \mathbb{E}\big[f(X_i, X_j) \mid X_i\big] - \mathbb{E}\big[f(X_i, X_j) \mid X_j\big] + \mathbb{E}\big[f(X_i, X_j)\big] \Big)$$

for $f \in \mathcal{F}$. Suppose that each $f \in \mathcal{F}$ is symmetric in the sense that $f(s_1, s_2) = f(s_2, s_1)$ for all $s_1, s_2 \in S$. Let F be a measurable envelope function for $\mathcal{F}$ satisfying $|f(s_1, s_2)| \leq F(s_1, s_2)$ for all $s_1, s_2 \in S$. For a law $\mathbb{Q}$ on $(S \times S, \mathcal{S} \otimes \mathcal{S})$, define the $(\mathbb{Q}, q)$-norm of $f \in \mathcal{F}$ by $\|f\|_{\mathbb{Q},q}^q = \mathbb{E}_{\mathbb{Q}}[|f|^q]$. Assume that $\mathcal{F}$ is VC-type in the following manner.

$$\sup_{\mathbb{Q}} N\big(\mathcal{F}, \|\cdot\|_{\mathbb{Q},2}, \varepsilon \|F\|_{\mathbb{Q},2}\big) \leq (C_1/\varepsilon)^{C_2}$$

for some constants $C_1 \geq e$ and $C_2 \geq 1$, and for all $\varepsilon \in (0, 1]$, where $\mathbb{Q}$ ranges over all finite discrete laws on $S \times S$. Let $\sigma > 0$ be any deterministic value satisfying $\sup_{f \in \mathcal{F}} \|f\|_{\mathbb{P},2} \leq \sigma \leq \|F\|_{\mathbb{P},2}$, and define the random variable $M = \max_{i,j} |F(X_i, X_j)|$. Then there exists a universal constant $C_3 > 0$ satisfying

$$n\mathbb{E}\left[\sup_{f \in \mathcal{F}} \big|U_n(f)\big|\right] \leq C_3 \sigma \Big(C_2 \log\big(C_1 \|F\|_{\mathbb{P},2}/\sigma\big)\Big) + \frac{C_3 \|M\|_{\mathbb{P},2}}{\sqrt{n}} \Big(C_2 \log\big(C_1 \|F\|_{\mathbb{P},2}/\sigma\big)\Big)^2.$$

Proof (Lemma B.3.9)

Apply Corollary 5.3 from Chen and Kato (2020) with the order of the U-statistic fixed at $r = 2$, and with $k = 2$. □

Lemma B.3.10 (A U-statistic matrix concentration inequality)

Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in a measurable space $(S, \mathcal{S})$. Suppose $H : S^2 \to \mathbb{R}^{d \times d}$ is a measurable matrix-valued function of two variables satisfying the following:

(i) $H(X_1, X_2)$ is an almost surely symmetric matrix.

(ii) $\|H(X_1, X_2)\|_2 \leq M$ almost surely.

(iii) H is a symmetric function in its arguments in that $H(X_1, X_2) = H(X_2, X_1)$.

(iv) H is degenerate in the sense that $\mathbb{E}[H(X_1, x_2)] = 0$ for all $x_2 \in S$.

Let $U_n = \sum_i \sum_{j \neq i} H(X_i, X_j)$ be a U-statistic, and define the variance-type constant

$$\sigma^2 = \mathbb{E}\left[\left\|\mathbb{E}\left[H(X_i, X_j)^2 \mid X_j\right]\right\|_2\right].$$

Then for a universal constant $C > 0$ and for all $t > 0$,

$$\mathbb{P}\left(\|U_n\|_2 \geq C\sigma n(t + \log d) + CM\sqrt{n}(t + \log d)^{3/2}\right) \leq Ce^{-t}.$$

By Jensen's inequality, $\sigma^2 \leq \mathbb{E}[\|H(X_i, X_j)^2\|_2] = \mathbb{E}[\|H(X_i, X_j)\|_2^2] \leq M^2$, giving the simpler

$$\mathbb{P}\left(\|U_n\|_2 \geq 2CMn(t + \log d)^{3/2}\right) \leq Ce^{-t}.$$

From this last inequality we deduce a moment bound by integration of tail probabilities:

$$\mathbb{E}\left[\|U_n\|_2\right] \lesssim Mn(\log d)^{3/2}.$$

Proof (Lemma B.3.10)

We apply results from Minsker and Wei (2019).

Part 1: decoupling

Let $\bar{U}_n = \sum_{i=1}^n \sum_{j=1}^n H(X_i^{(1)}, X_j^{(2)})$ be a decoupled matrix U-statistic, where $X^{(1)}$ and $X^{(2)}$ are i.i.d. copies of the sequence $X_1, \ldots, X_n$. By Lemma 5.2 in Minsker and Wei (2019), since we are only stating this result for degenerate U-statistics of order 2, there exists a universal constant D_2 such that for any $t > 0$, we have

$$\mathbb{P}\left(\|U_n\|_2 \geq t\right) \leq D_2 \mathbb{P}\left(\|\bar{U}_n\|_2 \geq t/D_2\right).$$

Part 2: concentration of the decoupled U-statistic

By Equation 11 in Minsker and Wei (2019), we have the following concentration inequality for decoupled degenerate U-statistics. For some universal constant C_1 and for any $t > 0$,

$$\mathbb{P}\left(\|\bar{U}_n\|_2 \geq C_1\sigma n(t + \log d) + C_1 M\sqrt{n}(t + \log d)^{3/2}\right) \leq e^{-t}.$$

Part 3: concentration of the original U-statistic

Hence we have

$$\begin{aligned}&\mathbb{P}\left(\|U_n\|_2 \geq C_1 D_2\sigma n(t + \log d) + C_1 D_2 M\sqrt{n}(t + \log d)^{3/2}\right)\\ &\quad\leq D_2\mathbb{P}\left(\|\bar{U}_n\|_2 \geq C_1\sigma n(t + \log d) + C_1 M\sqrt{n}(t + \log d)^{3/2}\right) \leq D_2 e^{-t}.\end{aligned}$$

The main result follows by setting $C = C_1 + C_1 D_2$.

Part 4: moment bound

We now obtain a moment bound for the simplified version. We already have that

$$\mathbb{P}\left(\|U_n\|_2 \geq 2CMn(t + \log d)^{3/2}\right) \leq Ce^{-t}.$$

This implies that for any $t \geq \log d$, we have

$$\mathbb{P}\left(\|U_n\|_2 \geq 8CMnt^{3/2}\right) \leq Ce^{-t}.$$

Defining $s = 8CMnt^{3/2}$ so $t = \left(\frac{s}{8CMn}\right)^{2/3}$ shows that for any $s \geq 8CMn(\log d)^{3/2}$,

$$\mathbb{P}\left(\|U_n\|_2 \geq s\right) \leq Ce^{-\left(\frac{s}{8CMn}\right)^{2/3}}.$$

Hence the moment bound is obtained:

$$\begin{aligned}
\mathbb{E}\left[\|U_n\|_2\right] &= \int_0^\infty \mathbb{P}\left(\|U_n\|_2 \geq s\right) \mathrm{d}s \\
&= \int_0^{8CMn(\log d)^{3/2}} \mathbb{P}\left(\|U_n\|_2 \geq s\right) \mathrm{d}s + \int_{8CMn(\log d)^{3/2}}^\infty \mathbb{P}\left(\|U_n\|_2 \geq s\right) \mathrm{d}s \\
&\leq 8CMn(\log d)^{3/2} + \int_0^\infty Ce^{-\left(\frac{s}{8CMn}\right)^{2/3}} \mathrm{d}s \\
&= 8CMn(\log d)^{3/2} + 8CMn \int_0^\infty e^{s^{-2/3}} \mathrm{d}s \lesssim Mn(\log d)^{3/2}. \qquad \square
\end{aligned}$$

B.3.2 Technical lemmas

Before presenting the proof of Lemma B.2.1, we give some auxiliary lemmas; namely a symmetrization inequality (Lemma B.3.11), a Rademacher contraction principle (Lemma B.3.12), and a Hoffman–Jørgensen inequality (Lemma B.3.13). Recall that the Rademacher distribution places probability mass of $1/2$ on each of the points -1 and 1.

Lemma B.3.11 (A symmetrization inequality for i.n.i.d. variables)
Let $(S, \mathcal{S})$ be a measurable space and $\mathcal{F}$ a class of Borel-measurable functions from S to $\mathbb{R}$ which is pointwise measurable (i.e. it contains a countable dense subset under pointwise convergence). Let $X_1, \ldots, X_n$ be independent but not necessarily identically distributed S-valued random variables. Let $a_1, \ldots, a_n$ be arbitrary points in S and ϕ a non-negative non-decreasing convex function from $\mathbb{R}$ to $\mathbb{R}$. Define $\varepsilon_1, \ldots, \varepsilon_n$ as independent Rademacher random variables, independent of $X_1, \ldots, X_n$. Then

$$\mathbb{E}\left[\phi\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \Big(f(X_i) - \mathbb{E}[f(X_i)]\Big)\right|\right)\right] \leq \mathbb{E}\left[\phi\left(2\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i\Big(f(X_i) - a_i\Big)\right|\right)\right].$$

Note that in particular this holds with $a_i = 0$ and also holds with $\phi(t) = t \vee 0$.

Proof (Lemma B.3.11)
See Lemma 2.3.6 in van der Vaart and Wellner (1996). $\square$

Lemma B.3.12 (A Rademacher contraction principle)

Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables and $\mathcal{T}$ be a bounded subset of $\mathbb{R}^n$. Define $M = \sup_{t\in\mathcal{T}} \max_{1\leq i\leq n} |t_i|$. Then, noting that the supremum is measurable because $\mathcal{T}$ is a subset of a separable metric space and is therefore itself separable,

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i t_i^2\right|\right] \leq 4M\,\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i t_i\right|\right].$$

This gives the following corollary. Let $X_1, \ldots, X_n$ be mutually independent and also independent of $\varepsilon_1, \ldots, \varepsilon_n$. Let $\mathcal{F}$ be a pointwise measurable class of functions from a measurable space $(S, \mathcal{S})$ to $\mathbb{R}$, with measurable envelope F. Define $M = \max_i F(X_i)$. Then we obtain

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\varepsilon_i f(X_i)^2\right|\right] \leq 4\mathbb{E}\left[M\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right].$$

Proof (Lemma B.3.12)

Apply Theorem 4.12 from Ledoux and Talagrand (1991) with F the identity function and

$$\psi_i(s) = \psi(s) = \min\left(\frac{s^2}{2M}, \frac{M}{2}\right).$$

This is a weak contraction (i.e. 1-Lipschitz) because it is continuous, differentiable on $(-M, M)$ with derivative bounded by $|\psi'(s)| \leq |s|/M \leq 1$, and constant outside $(-M, M)$. Note that since $|t_i| \leq M$ by definition, we have $\psi_i(t_i) = t_i^2/(2M)$. Hence by Theorem 4.12 from Ledoux and Talagrand (1991),

$$\mathbb{E}\left[F\left(\frac{1}{2}\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i\psi_i(t_i)\right|\right)\right] \leq \mathbb{E}\left[F\left(\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i t_i\right|\right)\right],$$

$$\mathbb{E}\left[\frac{1}{2}\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i\frac{t_i^2}{2M}\right|\right] \leq \mathbb{E}\left[\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i t_i\right|\right],$$

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i t_i^2\right|\right] \leq 4M\,\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i t_i\right|\right].$$

For the corollary, set $\mathcal{T} = \big\{\big(f(X_1), \ldots, f(X_n)\big) : f \in \mathcal{F}\big\}$. For a fixed realization $X_1, \ldots, X_n$,

$$\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i)^2 \right| \right] = \mathbb{E}_\varepsilon \left[\sup_{t \in \mathcal{T}} \left| \sum_{i=1}^n \varepsilon_i t_i^2 \right| \right]$$
$$\leq 4\mathbb{E}_\varepsilon \left[M \sup_{t \in \mathcal{T}} \left| \sum_{i=1}^n \varepsilon_i t_i \right| \right] = 4\mathbb{E}_\varepsilon \left[M \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right].$$

Taking an expectation over $X_1, \ldots, X_n$ and applying Fubini's theorem yields the result. □

Lemma B.3.13 (A Hoffmann–Jørgensen inequality)

Let $(S, \mathcal{S})$ be a measurable space and $X_1, \ldots, X_n$ be S-valued random variables. Suppose that $\mathcal{F}$ is a pointwise measurable class of functions from S to $\mathbb{R}$ with finite envelope F. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher variables independent of $X_1, \ldots, X_n$. For $q \in (1, \infty)$,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|^q \right]^{1/q} \leq C_q \left(\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right] + \mathbb{E}\left[\max_{1 \leq i \leq n} \sup_{f \in \mathcal{F}} \big|f(X_i)\big|^q \right]^{1/q} \right),$$

where C_q is a positive constant depending only on q.

Proof (Lemma B.3.13)

We use Talagrand's formulation of a Hoffmann–Jørgensen inequality. Consider the independent $\ell^\infty(\mathcal{F})$-valued random functionals u_i defined by $u_i(f) = \varepsilon_i f(X_i)$, where $\ell^\infty(\mathcal{F})$ is the Banach space of bounded functions from $\mathcal{F}$ to $\mathbb{R}$, equipped with the norm $\|u\|_\mathcal{F} = \sup_{f \in \mathcal{F}} |u(f)|$. Then Remark 3.4 in Kwapien and Szulga (1991) gives

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n u_i(f) \right|^q \right]^{1/q} \leq C_q \left(\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n u_i(f) \right| \right] + \mathbb{E}\left[\max_{1 \leq i \leq n} \sup_{f \in \mathcal{F}} |u_i(f)|^q \right]^{1/q} \right)$$
$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|^q \right]^{1/q} \leq C_q \left(\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right] + \mathbb{E}\left[\max_{1 \leq i \leq n} \sup_{f \in \mathcal{F}} \big|f(X_i)\big|^q \right]^{1/q} \right). \quad \square$$

Proof (Lemma B.2.1)

We follow the proof of Theorem 5.2 from Chernozhukov et al. (2014b), using our i.n.i.d. versions of the symmetrization inequality (Lemma B.3.11), Rademacher contraction principle (Lemma B.3.12), and Hoffmann–Jørgensen inequality (Lemma B.3.13).

Without loss of generality, we may assume that $J(1, \mathcal{F}, F) < \infty$ as otherwise there is nothing to prove, and that $F > 0$ everywhere on S. Let $\mathbb{P}_n = n^{-1} \sum_i \delta_{X_i}$ be the empirical distribution of X_i, and define the empirical variance bound $\sigma_n^2 = \sup_{\mathcal{F}} n^{-1} \sum_i f(X_i)^2$. By the i.n.i.d. symmetrization inequality (Lemma B.3.11),

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|G_n(f)\right|\right] = \frac{1}{\sqrt{n}}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \Big(f(X_i) - \mathbb{E}[f(X_i)]\Big)\right|\right] \leq \frac{2}{\sqrt{n}}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right],$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent Rademacher random variables, independent of $X_1, \ldots, X_n$. Then the standard entropy integral inequality from the proof of Theorem 5.2 in the supplemental materials for Chernozhukov et al. (2014b) gives for a universal constant $C_1 > 0$,

$$\frac{1}{\sqrt{n}}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \varepsilon_i f(X_i)\right| \;\middle|\; X_1, \ldots, X_n\right] \leq C_1 \|F\|_{\mathbb{P}_n,2}\, J(\sigma_n/\|F\|_{\mathbb{P}_n,2}, \mathcal{F}, F).$$

Taking marginal expectations and applying Jensen's inequality along with a convexity result for the covering integral, as in Lemma A.2 in Chernozhukov et al. (2014b), gives

$$Z := \frac{1}{\sqrt{n}}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] \leq C_1 \|F\|_{\bar{\mathbb{P}},2}\, J(\mathbb{E}[\sigma_n^2]^{1/2}/\|F\|_{\bar{\mathbb{P}},2}, \mathcal{F}, F).$$

Now use symmetrization (Lemma B.3.11), the contraction principle (Lemma B.3.12), the Cauchy–Schwarz inequality, and the Hoffmann–Jørgensen inequality (Lemma B.3.13) to deduce that

$$\begin{aligned}
\mathbb{E}[\sigma_n^2] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(X_i)^2\right] \leq \sup_{f\in\mathcal{F}} \mathbb{E}_{\bar{\mathbb{P}}}\left[f(X_i)^2\right] + \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n f(X_i)^2 - \mathbb{E}\left[f(X_i)^2\right]\right|\right] \\
&\leq \sigma^2 + \frac{2}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)^2\right|\right] \leq \sigma^2 + \frac{8}{n}\mathbb{E}\left[M\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] \\
&\leq \sigma^2 + \frac{8}{n}\mathbb{E}\left[M^2\right]^{1/2}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|^2\right]^{1/2} \\
&\leq \sigma^2 + \frac{8}{n}\|M\|_{\mathbb{P},2}\, C_2\left(\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] + \mathbb{E}\left[\max_{1\leq i\leq n}\sup_{f\in\mathcal{F}}\left|f(X_i)\right|^2\right]^{1/2}\right) \\
&= \sigma^2 + \frac{8C_2}{n}\|M\|_{\mathbb{P},2}\left(\sqrt{n}Z + \|M\|_{\mathbb{P},2}\right) \lesssim \sigma^2 + \frac{\|M\|_{\mathbb{P},2}Z}{\sqrt{n}} + \frac{\|M\|_{\bar{\mathbb{P}},2}^2}{n},
\end{aligned}$$

where $\lesssim$ indicates a bound up to a universal constant. Hence taking a square root we see that, following the notation from the proof of Theorem 5.2 in the supplemental materials to Chernozhukov et al. (2014b),

$$\sqrt{\mathbb{E}[\sigma_n^2]} \lesssim \sigma + \|M\|_{\mathbb{P},2}^{1/2} Z^{1/2} n^{-1/4} + \|M\|_{\mathbb{P},2} n^{-1/2} \lesssim \|F\|_{\bar{\mathbb{P}},2}\left(\Delta \vee \sqrt{DZ}\right),$$

where $\Delta^2 = \|F\|_{\bar{\mathbb{P}},2}^{-2}\big(\sigma^2 \vee (\|M\|_{\bar{\mathbb{P}},2}^2/n)\big) \geq \delta^2$ and $D = \|M\|_{\mathbb{P},2} n^{-1/2}\|F\|_{\bar{\mathbb{P}},2}^{-2}$. Thus returning to our bound on Z, we now have

$$Z \lesssim \|F\|_{\bar{\mathbb{P}},2}\, J(\Delta \vee \sqrt{DZ}, \mathcal{F}, F).$$

The final steps proceed as in the proof of Theorem 5.2 from Chernozhukov et al. (2014b), considering cases separately for $\Delta \geq \sqrt{DZ}$ and $\Delta < \sqrt{DZ}$, and applying convexity properties of the entropy integral J. □

Proof (Lemma B.2.2)

We assume the VC-type condition $\sup_{\mathbb{Q}} N(\mathcal{F}, \rho_{\mathbb{Q}}, \varepsilon\|F\|_{\mathbb{Q},2}) \leq (C_1/\varepsilon)^{C_2}$ for all $\varepsilon \in (0,1]$, with constants $C_1 \geq e$ and $C_2 \geq 1$. Hence for $\delta \in (0,1]$, the entropy integral can be bounded as

$$\begin{aligned} J\big(\delta, \mathcal{F}, F\big) &= \int_0^\delta \sqrt{1 + \sup_{\mathbb{Q}} \log N(\mathcal{F}, \rho_{\mathbb{Q}}, \varepsilon\|F\|_{\mathbb{Q},2})}\,\mathrm{d}\varepsilon \leq \int_0^\delta \sqrt{1 + C_2 \log(C_1/\varepsilon)}\,\mathrm{d}\varepsilon \\ &\leq \int_0^\delta \Big(1 + \sqrt{C_2 \log(C_1/\varepsilon)}\Big)\,\mathrm{d}\varepsilon = \delta + \sqrt{C_2}\int_0^\delta \sqrt{\log(C_1/\varepsilon)}\,\mathrm{d}\varepsilon \\ &\leq \delta + \sqrt{\frac{C_2}{\log(C_1/\delta)}}\int_0^\delta \log(C_1/\varepsilon)\,\mathrm{d}\varepsilon = \delta + \sqrt{\frac{C_2}{\log(C_1/\delta)}}\big(\delta + \delta\log(C_1/\delta)\big) \\ &\leq 3\delta\sqrt{C_2 \log(C_1/\delta)}. \end{aligned}$$

The remaining equations now follow by Lemma B.2.1. □

Before proving Lemma B.2.3, we give a bounded-variation characterization (Lemma B.3.14).

Lemma B.3.14 (A characterization of bounded-variation functions)

Let $\mathcal{V}_1$ be the class of real-valued functions on $[0,1]$ which are 0 at 1 and have total variation bounded by 1. Also define the class of half-interval indicator functions $\mathcal{I} = \{\mathbb{I}[0,t] : t \in [0,1]\}$. For any topological vector space $\mathcal{X}$, define the symmetric convex hull of a subset $\mathcal{Y} \subseteq \mathcal{X}$ as

$$\operatorname{symconv} \mathcal{Y} = \left\{ \sum_{i=1}^n \lambda_i y_i : \sum_{i=1}^n \lambda_i = 1,\ \lambda_i \geq 0,\ y_i \in \mathcal{Y} \cup -\mathcal{Y},\ n \in \mathbb{N} \right\}.$$

Denote its closure by $\overline{\operatorname{symconv}}\ \mathcal{Y}$. Under the pointwise convergence topology, $\mathcal{V}_1 \subseteq \overline{\operatorname{symconv}}\ \mathcal{I}$.

Proof (Lemma B.3.14)

Firstly, let $\mathcal{D} \subseteq \mathcal{V}_1$ be the class of real-valued functions on $[0,1]$ which are 0 at 1, have total variation exactly 1, and are weakly monotone decreasing. Therefore, for $g \in \mathcal{D}$, we have $\|g\|_{\mathrm{TV}} = g(0) = 1$. Let $S = \{s_1, s_2, \dots\} \subseteq [0,1]$ be the countable set of discontinuity points of g. We want to find a sequence of convex combinations of elements of $\mathcal{I}$ which converges pointwise to g. To do this, first define the sequence of meshes

$$A_n = \{s_k : 1 \leq k \leq n\} \cup \{k/n : 0 \leq k \leq n\},$$

which satisfies $\bigcup_n A_n = S \cup ([0,1] \cap \mathbb{Q})$. Endow A_n with the ordering induced by the canonical order on $\mathbb{R}$, giving $A_n = \{a_1, a_2, \ldots\}$, and define the sequence of functions

$$g_n(x) = \sum_{k=1}^{|A_n|-1} \mathbb{I}[0, a_k]\big(g(a_k) - g(a_{k+1})\big),$$

where clearly $\mathbb{I}[0, a_k] \in \mathcal{I}$, $g(a_k) - g(a_{k+1}) \geq 0$, and $\sum_{k=1}^{|A_n|-1} \big(g(a_k) - g(a_{k+1})\big) = g(0) - g(1) = 1$. Therefore g_n is a convex combination of elements of $\mathcal{I}$. Further, note that for $a_k \in A_n$,

$$g_n(a_k) = \sum_{j=k}^{|A_n|-1} \big(g(a_j) - g(a_{j+1})\big) = g(a_k) - g(a_{|A_n|}) = g(a_k) - g(1) = g(a_k).$$

Hence if $x \in S$, then eventually $x \in A_n$ so $g_n(x) \to g(x)$. Alternatively, if $x \notin S$, then g is continuous at x. But $g_n \to g$ on the dense set $\bigcup_n A_n$, so also $g_n(x) \to g(x)$. Hence $g_n \to g$ pointwise on $[0, 1]$.

Now take $f \in \mathcal{V}_1$. By the Jordan decomposition for total variation functions (Royden and Fitzpatrick, 1988), we can write $f = f^+ - f^-$, with f^+ and f^- weakly decreasing, $f^+(1) = f^-(1) = 0$, and $\|f^+\|_{\mathrm{TV}} + \|f^-\|_{\mathrm{TV}} = \|f\|_{\mathrm{TV}}$. Supposing that both $\|f^+\|_{\mathrm{TV}}$ and $\|f^-\|_{\mathrm{TV}}$ are strictly positive, let g_n^+ approximate the unit-variation function $f^+/\|f^+\|_{\mathrm{TV}}$ and g_n^- approximate $f^-/\|f^-\|_{\mathrm{TV}}$ as above. Then since trivially

$$f = \|f^+\|_{\mathrm{TV}} f^+/\|f^+\|_{\mathrm{TV}} - \|f^-\|_{\mathrm{TV}} f^-/\|f^-\|_{\mathrm{TV}} + \big(1 - \|f^+\|_{\mathrm{TV}} - \|f^-\|_{\mathrm{TV}}\big) \cdot 0,$$

we have that the convex combination

$$g_n^+ \|f^+\|_{\mathrm{TV}} - g_n^- \|f^-\|_{\mathrm{TV}} + \big(1 - \|f^+\|_{\mathrm{TV}} - \|f^-\|_{\mathrm{TV}}\big) \cdot 0$$

converges pointwise to f. This also holds if either of the total variations $\|f^\pm\|_{\mathrm{TV}}$ are zero, since then the corresponding sequence $g_n^\pm$ need not be defined. Now note that each of g_n^+, $-g_n^-$, and 0 are in $\operatorname{symconv} \mathcal{I}$, so $f \in \overline{\operatorname{symconv}}\ \mathcal{I}$ under pointwise convergence. $\square$

Proof (Lemma B.2.3)

We follow the Gaussian approximation method given in Section 2 of Giné et al. (2004). The KMT approximation theorem (Komlós et al., 1975) asserts the existence of a probability space carrying n i.i.d. uniform random variables $\xi_1, \ldots, \xi_n \sim \text{Unif}[0,1]$ and a standard Brownian motion $B_n(s) : s \in [0,1]$ such that if

$$\alpha_n(s) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(\mathbb{I}\{\xi_i \le s\} - s \big), \qquad \beta_n(s) := B_n(s) - sB_n(1),$$

then for some universal positive constants C_1, C_2, C_3, and for all $t > 0$,

$$\mathbb{P}\left(\sup_{s \in [0,1]} \big| \alpha_n(s) - \beta_n(s) \big| > \frac{t + C_1 \log n}{\sqrt{n}} \right) \le C_2 e^{-C_3 t}.$$

We can view α_n and β_n as random functionals defined on the class of half-interval indicator functions $\mathcal{I} = \big\{ \mathbb{I}[0,s] : s \in [0,1] \big\}$ in the following way.

$$\begin{aligned} \alpha_n(\mathbb{I}[0,s]) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(\mathbb{I}[0,s](\xi_i) - \mathbb{E}[\mathbb{I}[0,s](\xi_i)] \big), \\ \beta_n(\mathbb{I}[0,s]) &= \int_0^1 \mathbb{I}[0,s](u) \, \mathrm{d}B_n(u) - B_n(1) \int_0^1 \mathbb{I}[0,s](u) \, \mathrm{d}u, \end{aligned}$$

where the integrals are defined as Itô and Riemann–Stieltjes integrals in the usual way for stochastic integration against semimartingales (Le Gall, 2016, Chapter 5). Now we extend their definitions to the class $\mathcal{V}_1$ of functions on $[0,1]$ which are 0 at 1 and have total variation bounded by 1. This is achieved by noting that by Lemma B.3.14, we have $\mathcal{V}_1 \subseteq \overline{\text{symconv}}\, \mathcal{I}$ where $\overline{\text{symconv}}\, \mathcal{I}$ is the smallest symmetric convex class containing $\mathcal{I}$ which is closed under pointwise convergence. Thus by the dominated convergence theorem, every function in $\mathcal{V}_1$ is approximated in L^2 by finite convex combinations of functions in $\pm\mathcal{I}$, and the extension to $g \in \mathcal{V}_1$ follows by linearity and L^2 convergence of (stochastic) integrals:

$$\alpha_n(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(g(\xi_i) - \mathbb{E}[g(\xi_i)] \big), \qquad \beta_n(g) = \int_0^1 g(s) \, \mathrm{d}B_n(s) - B_n(1) \int_0^1 g(s) \, \mathrm{d}s.$$

Now we show that the norm induced on $(\alpha_n - \beta_n)$ by the function class $\mathcal{V}_1$ is a.s. identical to the supremum norm. Writing the sums as integrals and using integration by parts for finite-variation Lebesgue–Stieltjes and Itô integrals, and recalling that $g(1) = \alpha_n(0) = B_n(0) = 0$,

$$\begin{aligned}
\sup_{g\in\mathcal{V}_1} \big|\alpha_n(g) - \beta_n(g)\big| &= \sup_{g\in\mathcal{V}_1} \left| \int_0^1 g(s)\,\mathrm{d}\alpha_n(s) - \int_0^1 g(s)\,\mathrm{d}B_n(s) + B_n(1)\int_0^1 g(s)\,\mathrm{d}s \right| \\
&= \sup_{g\in\mathcal{V}_1} \left| \int_0^1 \alpha_n(s)\,\mathrm{d}g(s) - \int_0^1 B_n(s)\,\mathrm{d}g(s) + B_n(1)\int_0^1 s\,\mathrm{d}g(s) \right| \\
&= \sup_{g\in\mathcal{V}_1} \left| \int_0^1 \big(\alpha_n(s) - \beta_n(s)\big)\,\mathrm{d}g(s) \right| = \sup_{s\in[0,1]} \big|\alpha_n(s) - \beta_n(s)\big|,
\end{aligned}$$

where in the last line the upper bound is because $\|g\|_{\mathrm{TV}} \leq 1$, and the lower bound is by taking $g_\varepsilon = \pm\mathbb{I}[0, s_\varepsilon]$ where $|\alpha_n(s_\varepsilon) - \beta_n(s_\varepsilon)| \geq \sup_s |\alpha_n(s) - \beta_n(s)| - \varepsilon$. Hence we obtain

$$\mathbb{P}\left(\sup_{g\in\mathcal{V}_1} \big|\alpha_n(g) - \beta_n(g)\big| > \frac{t + C_1 \log n}{\sqrt{n}} \right) \leq C_2 e^{-C_3 t}. \tag{B.6}$$

Now define $V_n = \sup_{x\in\mathbb{R}} \|g_n(\cdot, x)\|_{\mathrm{TV}}$, noting that if $V_n = 0$ then the result is trivially true by setting $Z_n = 0$. Let F_X be the common c.d.f. of X_i, and define the quantile function $F_X^{-1}(s) = \inf\{u : F_X(u) \geq s\}$ for $s \in [0,1]$, writing $\inf\emptyset = \infty$ and $\inf\mathbb{R} = -\infty$. Consider the function class

$$\mathcal{G}_n = \big\{ V_n^{-1} g_n\big(F_X^{-1}(\cdot), x\big) - V_n^{-1} g_n\big(F_X^{-1}(1), x\big) : x \in \mathbb{R} \big\},$$

noting that $g_n(\cdot, x)$ is finite-variation so $g_n(\pm\infty, x)$ can be interpreted as the relevant limit. By monotonicity of F_X and the definition of V_n, the members of $\mathcal{G}_n$ have total variation of at most 1 and are 0 at 1, implying that $\mathcal{G}_n \subseteq \mathcal{V}_1$. Noting that α_n and β_n are random linear operators which a.s. annihilate constant functions, define

$$Z_n(x) = \beta_n\Big(g_n\big(F_X^{-1}(\cdot), x\big)\Big) = V_n \beta_n\Big(V_n^{-1} g_n\big(F_X^{-1}(\cdot), x\big) - V_n^{-1} g_n\big(F_X^{-1}(1), x\big)\Big),$$

which is a mean-zero continuous Gaussian process. Its covariance structure is

$$
\begin{aligned}
&\mathbb{E}[Z_n(x)Z_n(x')] \\
&= \mathbb{E}\Big[\Big(\int_0^1 g_n\big(F_X^{-1}(s), x\big)\, \mathrm{d}B_n(s) - B_n(1) \int_0^1 g_n\big(F_X^{-1}(s), x\big)\, \mathrm{d}s \Big) \\
&\quad \times \Big(\int_0^1 g_n\big(F_X^{-1}(s), x'\big)\, \mathrm{d}B_n(s) - B_n(1) \int_0^1 g_n\big(F_X^{-1}(s), x'\big)\, \mathrm{d}s \Big) \Big] \\
&= \mathbb{E}\left[\int_0^1 g_n\big(F_X^{-1}(s), x\big)\, \mathrm{d}B_n(s) \int_0^1 g_n\big(F_X^{-1}(s), x'\big)\, \mathrm{d}B_n(s) \right] \\
&\quad - \int_0^1 g_n\big(F_X^{-1}(s), x\big)\, \mathrm{d}s\ \mathbb{E}\left[B_n(1) \int_0^1 g_n\big(F_X^{-1}(s), x'\big)\, \mathrm{d}B_n(s) \right] \\
&\quad - \int_0^1 g_n\big(F_X^{-1}(s), x'\big)\, \mathrm{d}s\ \mathbb{E}\left[B_n(1) \int_0^1 g_n\big(F_X^{-1}(s), x\big)\, \mathrm{d}B_n(s) \right] \\
&\quad + \int_0^1 g_n\big(F_X^{-1}(s), x\big)\, \mathrm{d}s \int_0^1 g_n\big(F_X^{-1}(s), x'\big)\, \mathrm{d}s\ \mathbb{E}\left[B_n(1)^2 \right] \\
&= \int_0^1 g_n\big(F_X^{-1}(s), x\big) g_n\big(F_X^{-1}(s), x'\big)\, \mathrm{d}s - \int_0^1 g_n\big(F_X^{-1}(s), x\big)\, \mathrm{d}s \int_0^1 g_n\big(F_X^{-1}(s), x'\big)\, \mathrm{d}s \\
&= \mathbb{E}\Big[g_n\big(F_X^{-1}(\xi_i), x\big) g_n\big(F_X^{-1}(\xi_i), x'\big) \Big] - \mathbb{E}\Big[g_n\big(F_X^{-1}(\xi_i), x\big) \Big] \mathbb{E}\Big[g_n\big(F_X^{-1}(\xi_i), x'\big) \Big] \\
&= \mathbb{E}\Big[g_n\big(X_i, x\big) g_n\big(X_i, x'\big) \Big] - \mathbb{E}\Big[g_n\big(X_i, x\big) \Big] \mathbb{E}\Big[g_n\big(X_i, x'\big) \Big] = \mathbb{E}\big[G_n(x) G_n(x') \big]
\end{aligned}
$$

as desired, by the Itô isometry for stochastic integrals, writing $B_n(1) = \int_0^1 \mathrm{d}B_n(s)$; and noting that $F_X^{-1}(\xi_i)$ has the same distribution as X_i. Finally, note that

$$
G_n(x) = \alpha_n\Big(g_n\big(F_X^{-1}(\cdot), x\big) \Big) = V_n \alpha_n\Big(V_n^{-1} g_n\big(F_X^{-1}(\cdot), x\big) - V_n^{-1} g_n\big(F_X^{-1}(1), x\big) \Big),
$$

and so by (B.6)

$$
\begin{aligned}
\mathbb{P}\left(\sup_{x \in \mathbb{R}} \Big| G_n(x) - Z_n(x) \Big| > V_n \frac{t + C_1 \log n}{\sqrt{n}} \right) &\leq \mathbb{P}\left(\sup_{g \in \mathcal{V}_1} \big| \alpha_n(g) - \beta_n(g) \big| > \frac{t + C_1 \log n}{\sqrt{n}} \right) \\
&\leq C_2 e^{-C_3 t}. \qquad \square
\end{aligned}
$$

Proof (Lemma B.2.4)

Take $0 < \delta_n \leq \text{Leb}(\mathcal{X}_n)$ and let $\mathcal{X}_n^\delta = \{x_1, \ldots, x_{|\mathcal{X}_n^\delta|}\}$ be a δ_n-covering of $\mathcal{X}_n$ with cardinality $|\mathcal{X}_n^\delta| \leq \text{Leb}(\mathcal{X}_n)/\delta_n$. Suppose that $|\log \delta_n| \lesssim C_1 \log n$ up to a universal constant. We first use the Yurinskii coupling to construct a Gaussian process Z_n which is close to G_n on this finite cover. Then we bound the fluctuations in G_n and in Z_n using entropy methods.

Part 1: Yurinskii coupling

Define the i.n.i.d. and mean-zero variables

$$h_i(x) = \frac{1}{\sqrt{n}}\Big(g_n(X_i', x) - \mathbb{E}[g_n(X_i', x)]\Big),$$

where $X_1', \ldots, X_n'$ are independent copies of $X_1, \ldots, X_n$ on some new probability space, so that we have $G_n(x) = \sum_{i=1}^n h_i(x)$ in distribution. Also define the length-$|\mathcal{X}_n^\delta|$ random vector

$$h_i^\delta = \big(h_i(x) : x \in \mathcal{X}_n^\delta\big).$$

By an extension of Yurinskii's coupling to general norms (Belloni et al., 2019, supplemental materials, Lemma 38), there exists on the new probability space a Gaussian length-$|\mathcal{X}_n^\delta|$ vector Z_n^δ which is mean-zero and with the same covariance structure as $\sum_{i=1}^n h_i^\delta$ satisfying

$$\mathbb{P}\left(\left\|\sum_{i=1}^n h_i^\delta - Z_n^\delta\right\|_\infty > 3t_n\right) \leq \min_{s>0}\left(2\mathbb{P}\big(\|N\|_\infty > s\big) + \frac{\beta s^2}{t_n^3}\right),$$

where

$$\beta = \sum_{i=1}^n \Big(\mathbb{E}\big[\|h_i^\delta\|_2^2 \, \|h_i^\delta\|_\infty\big] + \mathbb{E}\big[\|z_i\|_2^2 \, \|z_i\|_\infty\big]\Big),$$

with $z_i \sim \mathcal{N}(0, \text{Var}[h_i^\delta])$ independent and $N \sim \mathcal{N}(0, I_{|\mathcal{X}_n^\delta|})$. By the bounds on g_n,

$$\mathbb{E}\big[\|h_i^\delta\|_2^2 \, \|h_i^\delta\|_\infty\big] \leq \frac{M_n}{\sqrt{n}}\mathbb{E}\big[\|h_i^\delta\|_2^2\big] = \frac{M_n}{\sqrt{n}}\sum_{x \in \mathcal{X}_n^\delta}\mathbb{E}\big[h_i(x)^2\big] \leq \frac{M_n}{\sqrt{n}}\frac{|\mathcal{X}_n^\delta|\sigma_n^2}{n} \leq \frac{M_n \sigma_n^2 \, \text{Leb}(\mathcal{X}_n)}{n^{3/2}\delta_n}.$$

By the fourth moment bound for Gaussian variables,

$$\begin{aligned}\mathbb{E}\big[\|z_i\|_2^4\big] &\le |\mathcal{X}_n^\delta|\,\mathbb{E}\big[\|z_i\|_4^4\big] \le |\mathcal{X}_n^\delta|^2 \max_j \mathbb{E}\big[(z_i^{(j)})^4\big] \le 3|\mathcal{X}_n^\delta|^2 \max_j \mathbb{E}\big[(z_i^{(j)})^2\big]^2 \\ &= 3|\mathcal{X}_n^\delta|^2 \max_{x\in\mathcal{X}_n^\delta} \mathbb{E}\big[h_i(x)^2\big]^2 \le \frac{3\sigma_n^4 \operatorname{Leb}(\mathcal{X}_n)^2}{n^2\delta_n^2}.\end{aligned}$$

Also by Jensen's inequality and for $|\mathcal{X}_n^\delta| \ge 2$, assuming $C_1 > 1$ without loss of generality,

$$\begin{aligned}\mathbb{E}\big[\|z_i\|_\infty^2\big] &\le \frac{4\sigma_n^2}{n}\log \mathbb{E}\big[e^{\|z_i\|_\infty^2/(4\sigma_n^2/n)}\big] \le \frac{4\sigma_n^2}{n}\log \mathbb{E}\left[\sum_{j=1}^{|\mathcal{X}_n^\delta|} e^{(z_i^{(j)})^2/(4\sigma_n^2/n)}\right] \le \frac{4\sigma_n^2}{n}\log\big(2|\mathcal{X}_n^\delta|\big) \\ &\le \frac{4\sigma_n^2}{n}\left(\log 2 + \log \operatorname{Leb}(\mathcal{X}_n) - \log\delta_n\right) \le \frac{12C_1\sigma_n^2\log n}{n},\end{aligned}$$

where we used the moment generating function of a χ_1^2 random variable. Therefore we can apply the Cauchy–Schwarz inequality to obtain

$$\begin{aligned}\mathbb{E}\big[\|z_i\|_2^2\,\|z_i\|_\infty\big] &\le \sqrt{\mathbb{E}\big[\|z_i\|_2^4\big]}\sqrt{\mathbb{E}\big[\|z_i\|_\infty^2\big]} \le \sqrt{\frac{3\sigma_n^4\operatorname{Leb}(\mathcal{X}_n)^2}{n^2\delta_n^2}}\sqrt{\frac{12C_1\sigma_n^2\log n}{n}} \\ &\le \frac{6\sigma_n^3\operatorname{Leb}(\mathcal{X}_n)\sqrt{C_1\log n}}{n^{3/2}\delta_n}.\end{aligned}$$

Now summing over the n samples gives

$$\beta \le \frac{M_n\sigma_n^2\operatorname{Leb}(\mathcal{X}_n)}{\sqrt{n}\delta_n} + \frac{6\sigma_n^3\operatorname{Leb}(\mathcal{X}_n)\sqrt{C_1\log n}}{\sqrt{n}\delta_n} = \frac{\sigma_n^2\operatorname{Leb}(\mathcal{X}_n)}{\sqrt{n}\delta_n}\Big(M_n + 6\sigma_n\sqrt{C_1\log n}\Big).$$

By a union bound and Gaussian tail probabilities, we have that $\mathbb{P}\big(\|N\|_\infty > s\big) \le 2|\mathcal{X}_n^\delta|e^{-s^2/2}$. Thus we get the following Yurinskii coupling inequality for all $s > 0$:

$$\mathbb{P}\left(\left\|\sum_{i=1}^n h_i^\delta - Z_n^\delta\right\|_\infty > t_n\right) \le \frac{4\operatorname{Leb}(\mathcal{X}_n)}{\delta_n}e^{-s^2/2} + \frac{\sigma_n^2\operatorname{Leb}(\mathcal{X}_n)s^2}{\sqrt{n}\delta_n t_n^3}\Big(M_n + 6\sigma_n\sqrt{C_1\log n}\Big).$$

Note that Z_n^δ now extends by the Vorob'ev–Berkes–Philipp theorem (Lemma B.2.5) to a mean-zero Gaussian process Z_n on the compact interval $\mathcal{X}_n$ with covariance structure

$$\mathbb{E}\big[Z_n(x)Z_n(x')\big] = \mathbb{E}\big[G_n(x)G_n(x')\big],$$

satisfying for any $s' > 0$

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}_n^\delta}\left|G_n(x) - Z_n(x)\right| > t_n\right) \le \frac{4\,\mathrm{Leb}(\mathcal{X}_n)}{\delta_n}e^{-s^2/2} + \frac{\sigma_n^2\,\mathrm{Leb}(\mathcal{X}_n)s^2}{\sqrt{n}\delta_n t_n^3}\Big(M_n + 6\sigma_n\sqrt{C_1\log n}\Big).$$

Part 2: regularity of G_n

Next we bound the fluctuations in the empirical process G_n. Consider the following classes of functions on S and their associated (constant) envelope functions. By continuity of g_n, each class is pointwise measurable (to see this, restrict the index sets to rationals).

$$\mathcal{G}_n = \big\{g_n(\cdot, x) : x \in \mathcal{X}_n\big\}, \qquad \mathrm{Env}(\mathcal{G}_n) = M_n,$$
$$\mathcal{G}_n^\delta = \big\{g_n(\cdot, x) - g_n(\cdot, x') : x, x' \in \mathcal{X}_n, |x - x'| \le \delta_n\big\}, \qquad \mathrm{Env}(\mathcal{G}_n^\delta) = l_{n,\infty}\delta_n.$$

We first show these are VC-type. By the uniform Lipschitz assumption,

$$\big\|g_n(\cdot, x) - g_n(\cdot, x')\big\|_\infty \le l_{n,\infty}|x - x'|$$

for all $x, x' \in \mathcal{X}_n$. Therefore, with $\mathbb{Q}$ ranging over the finitely-supported distributions on $(S, \mathcal{S})$, noting that any $\|\cdot\|_\infty$-cover is a $\rho_\mathbb{Q}$-cover,

$$\sup_\mathbb{Q} N\big(\mathcal{G}_n, \rho_\mathbb{Q}, \varepsilon l_{n,\infty}\mathrm{Leb}(\mathcal{X}_n)\big) \le N\big(\mathcal{G}_n, \|\cdot\|_\infty, \varepsilon l_{n,\infty}\mathrm{Leb}(\mathcal{X}_n)\big) \le N\big(\mathcal{X}_n, |\cdot|, \varepsilon\mathrm{Leb}(\mathcal{X}_n)\big) \le 1/\varepsilon.$$

Replacing ε by $\varepsilon M_n/(l_{n,\infty}\,\mathrm{Leb}(\mathcal{X}_n))$ gives

$$\sup_\mathbb{Q} N\big(\mathcal{G}_n, \rho_\mathbb{Q}, \varepsilon M_n\big) \le \frac{l_{n,\infty}\,\mathrm{Leb}(\mathcal{X}_n)}{\varepsilon M_n},$$

and so $\mathcal{G}_n$ is a VC-type class. To see that $\mathcal{G}_n^\delta$ is also a VC-type class, we construct a cover in the following way. Let $\mathcal{F}_n$ be an ε-cover for $(\mathcal{G}_n, \|\cdot\|_\infty)$. By the triangle inequality, $\mathcal{F}_n - \mathcal{F}_n$ is a 2ε-cover for $(\mathcal{G}_n - \mathcal{G}_n, \|\cdot\|_\infty)$ of cardinality at most $|\mathcal{F}_n|^2$, where the subtractions are set

subtractions. Since $\mathcal{G}_n^\delta \subseteq \mathcal{G}_n - \mathcal{G}_n$, we see that $\mathcal{F}_n - \mathcal{F}_n$ is a 2ε-external cover for $\mathcal{G}_n^\delta$. Thus

$$\begin{aligned}\sup_{\mathbb{Q}} N\big(\mathcal{G}_n^\delta, \rho_{\mathbb{Q}}, \varepsilon l_{n,\infty} \operatorname{Leb}(\mathcal{X}_n)\big) &\le N\big(\mathcal{G}_n^\delta, \|\cdot\|_\infty, \varepsilon l_{n,\infty} \operatorname{Leb}(\mathcal{X}_n)\big) \\ &\le N\big(\mathcal{G}_n, \|\cdot\|_\infty, \varepsilon l_{n,\infty} \operatorname{Leb}(\mathcal{X}_n)\big)^2 \le 1/\varepsilon^2.\end{aligned}$$

Replacing ε by $\varepsilon\delta_n / \operatorname{Leb}(\mathcal{X}_n)$ gives

$$\sup_{\mathbb{Q}} N\big(\mathcal{G}_n^\delta, \rho_{\mathbb{Q}}, \varepsilon l_{n,\infty}\delta_n\big) \le \frac{\operatorname{Leb}(\mathcal{X}_n)^2}{\varepsilon^2 \delta_n^2} \le (C_{1,n}/\varepsilon)^2$$

with $C_{1,n} = \operatorname{Leb}(\mathcal{X}_n)/\delta_n$, demonstrating that $\mathcal{G}_n^\delta$ forms a VC-type class. We now apply the maximal inequality for i.n.i.d. data given in Lemma B.2.2. To do this, note that $\sup_{\mathcal{G}_n^\delta} \|g\|_{\bar{\mathbb{P}},2} \le l_{n,2}\delta_n$ by the L^2 Lipschitz condition, and recall $\operatorname{Env}(\mathcal{G}_n^\delta) = l_{n,\infty}\delta_n$. Therefore Lemma B.2.2 with $\|F\|_{\bar{\mathbb{P}},2} = l_{n,\infty}\delta_n$, $\|M\|_{\mathbb{P},2} = l_{n,\infty}\delta_n$, and $\sigma = l_{n,2}\delta_n$ gives, up to universal constants

$$\begin{aligned}&\mathbb{E}\left[\sup_{g\in\mathcal{G}_n^\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(g(X_i) - \mathbb{E}[g(X_i)]\Big)\right|\right] \\ &\lesssim \sigma\sqrt{2\log\big(C_{1,n}\|F\|_{\bar{\mathbb{P}},2}/\sigma\big)} + \frac{\|M\|_{\mathbb{P},2} 2\log\big(C_{1,n}\|F\|_{\bar{\mathbb{P}},2}/\sigma\big)}{\sqrt{n}} \\ &\lesssim l_{n,2}\delta_n\sqrt{C_1\log n} + \frac{l_{n,\infty}\delta_n}{\sqrt{n}} C_1 \log n,\end{aligned}$$

and hence by Markov's inequality,

$$\begin{aligned}&\mathbb{P}\left(\sup_{|x-x'|\le\delta_n}\big|G_n(x) - G_n(x')\big| > t_n\right) \\ &= \mathbb{P}\left(\sup_{|x-x'|\le\delta_n}\frac{1}{\sqrt{n}}\left|\sum_{i=1}^n \Big(g_n(X_i,x) - \mathbb{E}[g_n(X_i,x)] - g_n(X_i,x') + \mathbb{E}[g_n(X_i,x')]\Big)\right| > t_n\right) \\ &= \mathbb{P}\left(\sup_{g\in\mathcal{G}_n^\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(g(X_i) - \mathbb{E}[g(X_i)]\Big)\right| > t_n\right) \le \frac{1}{t}\mathbb{E}\left[\sup_{g\in\mathcal{G}_n^\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(g(X_i) - \mathbb{E}[g(X_i)]\Big)\right|\right] \\ &\lesssim \frac{l_{n,2}\delta_n}{t_n}\sqrt{C_1\log n} + \frac{l_{n,\infty}\delta_n}{t_n\sqrt{n}} C_1 \log n.\end{aligned}$$

Part 3: regularity of Z_n

Next we bound the fluctuations in the Gaussian process Z_n. Let ρ be the following semimetric:

$$\begin{aligned}\rho(x,x')^2 &= \mathbb{E}\big[\big(Z_n(x)-Z_n(x')\big)^2\big] = \mathbb{E}\big[\big(G_n(x)-G_n(x')\big)^2\big] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[\big(h_i(x)-h_i(x')\big)^2\big] \le l_{n,2}^2\,|x-x'|^2.\end{aligned}$$

Hence $\rho(x,x') \le l_{n,2}\,|x-x'|$. By the Gaussian process maximal inequality from Lemma B.3.4, we obtain that

$$\begin{aligned}&\mathbb{E}\bigg[\sup_{|x-x'|\le\delta_n}\big|Z_n(x)-Z_n(x')\big|\bigg] \lesssim \mathbb{E}\bigg[\sup_{\rho(x,x')\le l_{n,2}\delta_n}\big|Z_n(x)-Z_n(x')\big|\bigg] \\ &\le \int_0^{l_{n,2}\delta_n}\sqrt{\log N(\varepsilon,\mathcal{X}_n,\rho)}\,\mathrm{d}\varepsilon \le \int_0^{l_{n,2}\delta_n}\sqrt{\log N(\varepsilon/l_{n,2},\mathcal{X}_n,|\cdot|)}\,\mathrm{d}\varepsilon \\ &\le \int_0^{l_{n,2}\delta_n}\sqrt{\log\left(1+\frac{\mathrm{Leb}(\mathcal{X}_n)l_{n,2}}{\varepsilon}\right)}\,\mathrm{d}\varepsilon \le \int_0^{l_{n,2}\delta_n}\sqrt{\log\left(\frac{2\,\mathrm{Leb}(\mathcal{X}_n)l_{n,2}}{\varepsilon}\right)}\,\mathrm{d}\varepsilon \\ &\le \log\left(\frac{2\,\mathrm{Leb}(\mathcal{X}_n)}{\delta_n}\right)^{-1/2}\int_0^{l_{n,2}\delta_n}\log\left(\frac{2\,\mathrm{Leb}(\mathcal{X}_n)l_{n,2}}{\varepsilon}\right)\,\mathrm{d}\varepsilon \\ &= \log\left(\frac{2\,\mathrm{Leb}(\mathcal{X}_n)}{\delta_n}\right)^{-1/2}\left(l_{n,2}\delta_n\log\left(2\,\mathrm{Leb}(\mathcal{X}_n)l_{n,2}\right)+l_{n,2}\delta_n+l_{n,2}\delta_n\log\left(\frac{1}{l_{n,2}\delta_n}\right)\right) \\ &= \log\left(\frac{2\,\mathrm{Leb}(\mathcal{X}_n)}{\delta_n}\right)^{-1/2} l_{n,2}\delta_n\left(1+\log\left(\frac{2\,\mathrm{Leb}(\mathcal{X}_n)}{\delta_n}\right)\right) \lesssim l_{n,2}\delta_n\sqrt{\log\left(\frac{\mathrm{Leb}(\mathcal{X}_n)}{\delta_n}\right)} \\ &\lesssim l_{n,2}\delta_n\sqrt{C_1\log n},\end{aligned}$$

where we used that $\delta_n \le \mathrm{Leb}(\mathcal{X}_n)$. So by Markov's inequality,

$$\mathbb{P}\left(\sup_{|x-x'|\le\delta_n}\big|Z_n(x)-Z_n(x')\big| > t_n\right) \lesssim t_n^{-1}l_{n,2}\delta_n\sqrt{C_1\log n}.$$

Part 4: conclusion

By the results of the previous parts, we have up to universal constants that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x\in\mathcal{X}_n}\left|G_n(x)-Z_n(x)\right|>t_n\right)\\
&\quad\le \mathbb{P}\left(\sup_{x\in\mathcal{X}_n^\delta}\left|G_n(x)-Z_n(x)\right|>t_n/3\right)+\mathbb{P}\left(\sup_{|x-x'|\le\delta_n}\left|G_n(x)-G_n(x')\right|>t_n/3\right)\\
&\qquad+\mathbb{P}\left(\sup_{|x-x'|\le\delta_n}\left|Z_n(x)-Z_n(x')\right|>t_n/3\right)\\
&\quad\lesssim \frac{4\,\mathrm{Leb}(\mathcal{X}_n)}{\delta_n}e^{-s^2/2}+\frac{\sigma_n^2\,\mathrm{Leb}(\mathcal{X}_n)s^2}{\sqrt{n}\delta_n t_n^3}\Big(M_n+6\sigma_n\sqrt{C_1\log n}\Big)\\
&\qquad+\frac{l_{n,2}\delta_n}{t_n}\sqrt{C_1\log n}+\frac{l_{n,\infty}\delta_n}{t_n\sqrt{n}}C_1\log n.
\end{aligned}$$

Choosing an approximately optimal mesh size of

$$\delta_n=\sqrt{\frac{\sigma_n^2\,\mathrm{Leb}(\mathcal{X}_n)\log n}{\sqrt{n}t_n^3}\Big(M_n+\sigma_n\sqrt{\log n}\Big)}\Bigg/\sqrt{t_n^{-1}l_{n,2}\sqrt{\log n}\left(1+\frac{l_{n,\infty}\sqrt{\log n}}{l_{n,2}\sqrt{n}}\right)}$$

gives $\log|\delta_n|\lesssim C_1\log n$ for a universal constant, so with s a large enough multiple of $\sqrt{\log n}$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x\in\mathcal{X}_n}\left|G_n(x)-Z_n(x)\right|>t_n\right)\\
&\quad\lesssim \frac{4\,\mathrm{Leb}(\mathcal{X}_n)}{\delta_n}e^{-s^2/2}+\frac{\sigma_n^2\,\mathrm{Leb}(\mathcal{X}_n)s^2}{\sqrt{n}\delta_n t_n^3}\Big(M_n+6\sigma_n\sqrt{C_1\log n}\Big)\\
&\qquad+\frac{l_{n,2}\delta_n}{t_n}\sqrt{C_1\log n}+\frac{l_{n,\infty}\delta_n}{t_n\sqrt{n}}C_1\log n\\
&\quad\lesssim \delta_n\frac{l_{n,2}\sqrt{\log n}}{t_n}\left(1+\frac{l_{n,\infty}\sqrt{\log n}}{l_{n,2}\sqrt{n}}\right)\\
&\quad\lesssim \frac{\sigma_n\sqrt{\mathrm{Leb}(\mathcal{X}_n)}\sqrt{\log n}\sqrt{M_n+\sigma_n\sqrt{\log n}}}{n^{1/4}t_n^2}\sqrt{l_{n,2}\sqrt{\log n}+\frac{l_{n,\infty}}{\sqrt{n}}\log n}.
\end{aligned}$$

□

Proof (Lemma B.2.5)

The proof is by induction on the number of vertices in the tree. Let $\mathcal{T}$ have n vertices, and suppose that vertex n is a leaf connected to vertex $n-1$ by an edge, relabeling the vertices if necessary. By the induction hypothesis we assume that there is a probability measure $\mathbb{P}^{(n-1)}$

on $\prod_{i=1}^{n-1} \mathcal{X}_i$ whose projections onto $\mathcal{X}_i$ are $\mathbb{P}_i$ and whose projections onto $\mathcal{X}_i \times \mathcal{X}_j$ are $\mathbb{P}_{ij}$, for $i, j \leq n-1$. Now apply the original Vorob'ev–Berkes–Philipp theorem, which can be found as Theorem 1.1.10 in Dudley (1999), to the spaces $\prod_{i=1}^{n-2} \mathcal{X}_i$, $\mathcal{X}_{n-1}$, and $\mathcal{X}_n$; and to the laws $\mathbb{P}^{(n-1)}$ and $\mathbb{P}_{n-1,n}$. This gives a law $\mathbb{P}^{(n)}$ which agrees with $\mathbb{P}_i$ at every vertex by definition, and agrees with $\mathbb{P}_{ij}$ for all $i, j \leq n-1$. It also agrees with $\mathbb{P}_{n-1,n}$, and this is the only edge touching vertex n. Hence $\mathbb{P}^{(n)}$ satisfies the desired properties. □

B.3.3 Main results

We give supplementary details for our main results on consistency, minimax optimality, strong approximation, covariance estimation, feasible inference and counterfactual estimation. We begin with a basic fact about Lipschitz functions.

Lemma B.3.15 (Lipschitz kernels are bounded)
Let $\mathcal{X} \subseteq \mathbb{R}$ be a connected set. Let $f : \mathcal{X} \to \mathbb{R}$ satisfy the Lipschitz condition $|f(x) - f(x')| \leq C|x - x'|$ for some $C > 0$ and all $x, x' \in \mathcal{X}$. Suppose also that f is a kernel in the sense that $\int_{\mathcal{X}} f(x)\,\mathrm{d}x = 1$. Then we have

$$\sup_{x \in \mathcal{X}} |f(x)| \leq C \operatorname{Leb}(\mathcal{X}) + \frac{1}{\operatorname{Leb}(\mathcal{X})}.$$

Now let $g : \mathcal{X} \to [0, \infty)$ satisfy $|g(x) - g(x')| \leq C|x - x'|$ for some $C > 0$ and all $x, x' \in \mathcal{X}$. Suppose g is a sub-kernel with $\int_{\mathcal{X}} g(x)\,\mathrm{d}x \leq 1$. Then for any $M \in \big(0, \operatorname{Leb}(\mathcal{X})\big]$, we have

$$\sup_{x \in \mathcal{X}} f(x) \leq CM + \frac{1}{M}.$$

Applying Lemma B.3.15 to the density and kernel functions defined in Assumptions 3.2.1 and 3.2.2 yields the following. Firstly, since $k_h(\cdot, w)$ is C_{L}/h^2-Lipschitz on $[w \pm h] \cap \mathcal{W}$ and integrates to one, we have by the first inequality in Lemma B.3.15 that

$$|k_h(s, w)| \leq \frac{2C_{\mathrm{L}} + 1}{h} + \frac{1}{\operatorname{Leb}(\mathcal{W})}.$$

Since each of $f_{W|AA}(\cdot \mid a, a')$, $f_{W|A}(\cdot \mid a)$, and f_W is non-negative, and C_{H}-Lipschitz on $\mathcal{W}$ and integrates to at most one over $\mathcal{W}$, taking $M = \frac{1}{\sqrt{C_{\mathrm{H}}}} \wedge \mathrm{Leb}(\mathcal{W})$ in the second inequality in Lemma B.3.15 gives

$$\begin{aligned} f_{W|AA}(w \mid a, a') &\leq 2\sqrt{C_{\mathrm{H}}} + \frac{1}{\mathrm{Leb}(\mathcal{W})}, \\ f_{W|A}(w \mid a) &\leq 2\sqrt{C_{\mathrm{H}}} + \frac{1}{\mathrm{Leb}(\mathcal{W})}, \\ f_W(w) &\leq 2\sqrt{C_{\mathrm{H}}} + \frac{1}{\mathrm{Leb}(\mathcal{W})}. \end{aligned}$$

Proof (Lemma B.3.15)

We begin with the first inequality. Note that if $\mathrm{Leb}(\mathcal{X}) = \infty$ there is nothing to prove. Suppose for contradiction that $|f(x)| > C\,\mathrm{Leb}(\mathcal{X}) + \frac{1}{\mathrm{Leb}(\mathcal{X})}$ for some $x \in \mathcal{X}$. If $f(x) \geq 0$ then by the Lipschitz property, for any $y \in \mathcal{X}$,

$$f(y) \geq f(x) - C|y - x| > C\,\mathrm{Leb}(\mathcal{X}) + \frac{1}{\mathrm{Leb}(\mathcal{X})} - C\,\mathrm{Leb}(\mathcal{X}) = \frac{1}{\mathrm{Leb}(\mathcal{X})}.$$

Similarly, if $f(x) \leq 0$ then

$$f(y) \leq f(x) + C|y - x| < -C\,\mathrm{Leb}(\mathcal{X}) - \frac{1}{\mathrm{Leb}(\mathcal{X})} + C\,\mathrm{Leb}(\mathcal{X}) = -\frac{1}{\mathrm{Leb}(\mathcal{X})}.$$

But then either $\int_{\mathcal{X}} f(x)\,\mathrm{d}x > \int_{\mathcal{X}} 1/\mathrm{Leb}(\mathcal{X})\,\mathrm{d}x = 1$ or $\int_{\mathcal{X}} f(x)\,\mathrm{d}x < \int_{\mathcal{X}} -1/\mathrm{Leb}(\mathcal{X})\,\mathrm{d}x = -1 < 1$, giving a contradiction.

For the second inequality, assume that f is non-negative on $\mathcal{X}$, and take $M \in \big(0, \mathrm{Leb}(\mathcal{X})\big]$. Suppose for contradiction that $f(x) > CM + \frac{1}{M}$ for some $x \in \mathcal{X}$. Then by the Lipschitz property, $f(y) \geq 1/M$ for all y such that $|y - x| \leq M$. Since $\mathcal{X}$ is connected, we have $\mathrm{Leb}(\mathcal{X} \cap [x \pm M]) \geq M$ and so we deduce that $\int_{\mathcal{X}} f(x)\,\mathrm{d}x > M/M = 1$ which is a contradiction. □

Proof (Theorem 3.2.1)

Begin by defining

$$P_p(s, w) = \sum_{r=0}^{p} \frac{f_W^{(r)}(w)}{r!}(s - w)^r$$

for $s, w \in \mathcal{W}$ as the degree-p Taylor polynomial of f_W, centered at w and evaluated at s. Note that for $p \leq \underline{\beta} - 1$, by Taylor's theorem with Lagrange remainder,

$$f_W(s) - P_p(s,w) = \frac{f_W^{(p+1)}(w')}{(p+1)!}(s-w)^{p+1}$$

for some w' between w and s. Also note that for any p,

$$\int_{\mathcal{W}} k_h(s,w)\big(P_p(s,w) - P_{p-1}(s,w)\big)\,\mathrm{d}s = \int_{\mathcal{W}} k_h(s,w)\frac{f_W^{(p)}(w)}{p!}(s-w)^p\,\mathrm{d}s = h^p b_p(w).$$

Further, by the order of the kernel,

$$\begin{aligned}\mathbb{E}\big[\hat{f}_W(w)\big] - f_W(w) &= \int_{\mathcal{W}} k_h(s,w) f_W(s)\,\mathrm{d}s - f_W(w) = \int_{\mathcal{W}} k_h(s,w)\big(f_W(s) - f_W(w)\big)\,\mathrm{d}s \\ &= \int_{\mathcal{W}} k_h(s,w)\big(f_W(s) - P_{p-1}(s,w)\big)\,\mathrm{d}s.\end{aligned}$$

Part 1: low-order kernel

Suppose that $p \leq \underline{\beta} - 1$. Then

$$\begin{aligned}&\sup_{w\in\mathcal{W}}\big|\mathbb{E}[\hat{f}_W(w)] - f_W(w) - h^p b_p(w)\big| \\ &\quad= \sup_{w\in\mathcal{W}}\left|\int_{\mathcal{W}} k_h(s,w)\big(f_W(s) - P_{p-1}(s,w)\big)\,\mathrm{d}s - h^p b_p(w)\right| \\ &\quad= \sup_{w\in\mathcal{W}}\left|\int_{\mathcal{W}} k_h(s,w)\big(f_W(s) - P_p(s,w) + P_p(s,w) - P_{p-1}(s,w)\big)\,\mathrm{d}s - h^p b_p(w)\right| \\ &\quad= \sup_{w\in\mathcal{W}}\left|\int_{\mathcal{W}} k_h(s,w)\big(f_W(s) - P_p(s,w)\big)\,\mathrm{d}s\right| = \sup_{w\in\mathcal{W}}\left|\int_{\mathcal{W}} k_h(s,w)\frac{f_W^{(p+1)}(w')}{(p+1)!}(s-w)^{p+1}\,\mathrm{d}s\right| \\ &\quad\leq \sup_{w\in\mathcal{W}}\left|\int_{[w\pm h]} \frac{C_\mathrm{k}}{h}\frac{C_\mathrm{H}}{(p+1)!}h^{p+1}\,\mathrm{d}s\right| \leq \frac{2C_\mathrm{k}C_\mathrm{H}}{(p+1)!}h^{p+1}.\end{aligned}$$

Part 2: order of kernel matches smoothness

Suppose that $p = \underline{\beta}$. Then

$$\begin{aligned}
&\sup_{w\in\mathcal{W}} \left|\mathbb{E}[\hat{f}_W(w)] - f_W(w) - h^p b_p(w)\right| \\
&\quad= \sup_{w\in\mathcal{W}} \left| \int_{\mathcal{W}} k_h(s,w)\big(f_W(s) - P_{\underline{\beta}-1}(s,w)\big)\,\mathrm{d}s - h^p b_p(w) \right| \\
&\quad= \sup_{w\in\mathcal{W}} \left| \int_{\mathcal{W}} k_h(s,w)\big(f_W(s) - P_{\underline{\beta}}(s,w) + P_{\underline{\beta}}(s,w) - P_{\underline{\beta}-1}(s,w)\big)\,\mathrm{d}s - h^{\underline{\beta}} b_{\underline{\beta}}(w) \right| \\
&\quad= \sup_{w\in\mathcal{W}} \left| \int_{\mathcal{W}} k_h(s,w)\big(f_W(s) - P_{\underline{\beta}}(s,w)\big)\,\mathrm{d}s \right| \\
&\quad= \sup_{w\in\mathcal{W}} \left| \int_{\mathcal{W}} k_h(s,w) \frac{f_W^{(\underline{\beta})}(w') - f_W^{(\underline{\beta})}(w)}{\underline{\beta}!}(s-w)^{\underline{\beta}}\,\mathrm{d}s \right| \\
&\quad\leq \sup_{w\in\mathcal{W}} \left| \int_{[w\pm h]} \frac{C_{\mathrm{k}}}{h} \frac{C_{\mathrm{H}} h^{\beta-\underline{\beta}}}{\underline{\beta}!} h^{\underline{\beta}}\,\mathrm{d}s \right| \leq \frac{2C_{\mathrm{k}}C_{\mathrm{H}}}{\underline{\beta}!} h^{\beta}.
\end{aligned}$$

Part 3: high-order kernel

Suppose that $p \geq \underline{\beta} + 1$. Then as in the previous part

$$\sup_{w\in\mathcal{W}} \left|\mathbb{E}[\hat{f}_W(w)] - f_W(w)\right| = \sup_{w\in\mathcal{W}} \left| \int_{[w\pm h]\cap\mathcal{W}} k_h(s,w)\big(f_W(s) - P_{\underline{\beta}}(s,w)\big)\,\mathrm{d}s \right| \leq \frac{2C_{\mathrm{k}}C_{\mathrm{H}}}{\underline{\beta}!} h^{\beta}. \quad \square$$

Proof (Lemma 3.2.1)

Part 1: Hoeffding-type decomposition

$$\begin{aligned}
\hat{f}_W(w) - E_n(w) - \mathbb{E}[\hat{f}_W(w)] &= \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \Big(\mathbb{E}[k_h(W_{ij},w) \mid A_i, A_j] - \mathbb{E}[k_h(W_{ij},w)]\Big) \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j\neq i} \Big(\mathbb{E}[k_h(W_{ij},w) \mid A_i, A_j] - \mathbb{E}[k_h(W_{ij},w)]\Big),
\end{aligned}$$

and apply Lemma B.3.7 with

$$u_{ij} = \frac{1}{n(n-1)}\mathbb{E}\big[k_h(W_{ij}, w) \mid A_i, A_j\big], \qquad u_i = \frac{1}{n(n-1)}\mathbb{E}\big[k_h(W_{ij}, w) \mid A_i\big],$$
$$u = \frac{1}{n(n-1)}\mathbb{E}\big[k_h(W_{ij}, w)\big],$$

to see

$$\begin{aligned}\hat{f}_W(w) - E_n(w) - \mathbb{E}[\hat{f}_W(w)] &= \frac{2}{n}\sum_{i=1}^{n}\big(u_i - u\big) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\big(u_{ij} - u_i - u_j + u\big)\\ &= \frac{2}{n}\sum_{i=1}^{n} l_i(w) + \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n} q_{ij}(w) = L_n + Q_n.\end{aligned}$$

Part 2: expectation and covariance of L_n, Q_n, and E_n

L_n, Q_n, and E_n are clearly mean-zero. For orthogonality, note that their summands have the following properties, for any $1 \leq i < j \leq n$ and $1 \leq r < s \leq n$, and for any $w, w' \in \mathcal{W}$:

$$\mathbb{E}\big[l_i(w)q_{rs}(w')\big] = \mathbb{E}\big[l_i(w)\mathbb{E}\big[q_{rs}(w') \mid A_i\big]\big] = 0,$$

$$\begin{aligned}\mathbb{E}\big[l_i(w)e_{rs}(w')\big] &= \begin{cases}\mathbb{E}\big[l_i(w)\big]\mathbb{E}\big[e_{rs}(w')\big], \text{ if } i \notin \{r,s\},\\ \mathbb{E}\big[l_i(w)\mathbb{E}\big[e_{rs}(w') \mid A_r, A_s\big]\big], \text{ if } i \in \{r,s\},\end{cases}\\ &= 0,\end{aligned}$$

$$\begin{aligned}\mathbb{E}\big[q_{ij}(w)e_{rs}(w')\big] &= \begin{cases}\mathbb{E}\big[q_{ij}(w)\big]\mathbb{E}\big[e_{rs}(w')\big], \text{ if } \{i,j\} \cap \{r,s\} = \emptyset,\\ \mathbb{E}\big[\mathbb{E}\big[q_{ij}(w) \mid A_i\big]\mathbb{E}\big[e_{rs}(w') \mid A_i\big]\big], \text{ if } \{i,j\} \cap \{r,s\} = \{i\},\\ \mathbb{E}\big[\mathbb{E}\big[q_{ij}(w) \mid A_j\big]\mathbb{E}\big[e_{rs}(w') \mid A_j\big]\big], \text{ if } \{i,j\} \cap \{r,s\} = \{j\},\\ \mathbb{E}\big[q_{ij}(w)\mathbb{E}\big[e_{rs}(w') \mid A_r, A_s\big]\big], \text{ if } \{i,j\} = \{r,s\},\end{cases}\\ &= 0,\end{aligned}$$

by independence of $\mathbf{A}_n$ and $\mathbf{V}_n$ and as $\mathbb{E}[q_{rs}(w) \mid A_i] = 0$ and $\mathbb{E}[e_{ij}(w) \mid A_i, A_j] = 0$. □

Proof (Lemma 3.2.2)

Part 1: total degeneracy

Suppose $D_{\mathrm{lo}} = 0$, so $\mathrm{Var}[f_{W|A}(w \mid A_i)] = 0$ for all $w \in \mathcal{W}$. Therefore, for all $w \in \mathcal{W}$, we have $f_{W|A}(w) = f_W(w)$ almost surely. By taking a union over $\mathcal{W} \cap \mathbb{Q}$ and by continuity of $f_{W|A}$ and f_W, this implies that $f_{W|A}(w) = f_W(w)$ for all $w \in \mathcal{W}$ almost surely. Thus

$$\mathbb{E}\left[k_h(W_{ij}, w) \mid A_i\right] = \int_{\mathcal{W}} k_h(s, w) f_{W|A}(s \mid A_i)\,\mathrm{d}s = \int_{\mathcal{W}} k_h(s, w) f_W(s)\,\mathrm{d}s = \mathbb{E}\left[k_h(W_{ij}, w)\right]$$

for all $w \in \mathcal{W}$ almost surely. Hence $l_i(w) = 0$ and so $L_n(w) = 0$ for all $w \in \mathcal{W}$ almost surely.

Part 2: no degeneracy

Suppose $D_{\mathrm{lo}} > 0$. As $f_{W|A}(\cdot \mid a)$ is C_{H}-Lipschitz for all $a \in \mathcal{A}$ and since $|k_h| \leq C_{\mathrm{k}}/h$,

$$\begin{aligned}
&\sup_{w \in \mathcal{W}} \left|\mathbb{E}[k_h(W_{ij}, w) \mid A_i] - f_{W|A}(w \mid A_i)\right| \\
&\quad = \sup_{w \in \mathcal{W}} \left| \int_{\mathcal{W}} k_h(s, w) f_{W|A}(s \mid A_i)\,\mathrm{d}s - f_{W|A}(w \mid A_i) \right| \\
&\quad = \sup_{w \in \mathcal{W}} \left| \int_{\mathcal{W} \cap [w \pm h]} k_h(s, w) \left(f_{W|A}(s \mid A_i) - f_{W|A}(w \mid A_i)\right)\,\mathrm{d}s \right| \\
&\quad \leq 2h \frac{C_{\mathrm{k}}}{h} C_{\mathrm{H}} h \leq 2 C_{\mathrm{k}} C_{\mathrm{H}} h
\end{aligned}$$

almost surely. Therefore, since $f_{W|A}(w \mid a) \leq C_{\mathrm{d}}$, we have

$$\sup_{w \in \mathcal{W}} \left|\mathrm{Var}\left[\mathbb{E}[k_h(W_{ij}, w) \mid A_i]\right] - \mathrm{Var}\left[f_{W|A}(w \mid A_i)\right]\right| \leq 16 C_{\mathrm{k}} C_{\mathrm{H}} C_{\mathrm{d}} h$$

whenever h is small enough that $2C_{\mathrm{k}} C_{\mathrm{H}} h \leq C_{\mathrm{d}}$. Thus

$$\inf_{w \in \mathcal{W}} \mathrm{Var}\left[\mathbb{E}[k_h(W_{ij}, w) \mid A_i]\right] \geq \inf_{w \in \mathcal{W}} \mathrm{Var}[f_{W|A}(w \mid A_i)] - 16 C_{\mathrm{k}} C_{\mathrm{H}} C_{\mathrm{d}} h.$$

Therefore, if $D_{\mathrm{lo}} > 0$, then eventually $\inf_{w \in \mathcal{W}} \mathrm{Var}\left[\mathbb{E}[k_h(W_{ij}, w) \mid A_i]\right] \geq D_{\mathrm{lo}}/2$. Finally,

$$\inf_{w \in \mathcal{W}} \mathrm{Var}[L_n(w)] = \frac{4}{n} \inf_{w \in \mathcal{W}} \mathrm{Var}\left[\mathbb{E}[k_h(W_{ij}, w) \mid A_i]\right] \geq \frac{2 D_{\mathrm{lo}}}{n}.$$

Part 3: partial degeneracy

Since $f_{W|A}(w \mid A_i)$ is bounded by C_{d} and C_{H}-Lipschitz in w, we have that $\mathrm{Var}[f_{W|A}(w \mid A_i)]$ is continuous on $\mathcal{W}$. Thus if $D_{\mathrm{lo}} = 0$, there is at least one point $w \in \mathcal{W}$ for which $\mathrm{Var}[f_{W|A}(w \mid A_i)] = 0$ by compactness. Let w be any such degenerate point. Then by the previous part,

$$\mathrm{Var}[L_n(w)] = \frac{4}{n} \mathrm{Var}\left[\mathbb{E}[k_h(W_{ij}, w) \mid A_i]\right] \leq 64 C_{\mathrm{k}} C_{\mathrm{H}} C_{\mathrm{d}} \frac{h}{n}.$$

If conversely w is not a degenerate point then $\mathrm{Var}[f_{W|A}(w \mid A_i)] > 0$ so eventually

$$\mathrm{Var}[L_n(w)] = \frac{4}{n} \mathrm{Var}\left[\mathbb{E}[k_h(W_{ij}, w) \mid A_i]\right] \geq \frac{2}{n} \mathrm{Var}[f_{W|A}(w \mid A_i)]. \qquad \square$$

Proof (Lemma 3.2.3)

We establish VC-type properties of function classes and apply empirical process theory.

Part 1: establishing VC-type classes

Consider the following function classes:

$$\begin{aligned}
\mathcal{F}_1 &= \Big\{ W_{ij} \mapsto k_h(W_{ij}, w) : w \in \mathcal{W} \Big\}, \\
\mathcal{F}_2 &= \Big\{ (A_i, A_j) \mapsto \mathbb{E}\big[k_h(W_{ij}, w) \mid A_i, A_j\big] : w \in \mathcal{W} \Big\}, \\
\mathcal{F}_3 &= \Big\{ A_i \mapsto \mathbb{E}\big[k_h(W_{ij}, w) \mid A_i\big] : w \in \mathcal{W} \Big\}.
\end{aligned}$$

For $\mathcal{F}_1$, take $0 < \varepsilon \leq \mathrm{Leb}(\mathcal{W})$ and $\mathcal{W}_\varepsilon$ an ε-cover of $\mathcal{W}$ of cardinality at most $\mathrm{Leb}(\mathcal{W})/\varepsilon$. As

$$\sup_{s,w,w' \in \mathcal{W}} \left| \frac{k_h(s, w) - k_h(s, w')}{w - w'} \right| \leq \frac{C_{\mathrm{L}}}{h^2}$$

almost surely, we see that

$$\sup_{\mathbb{Q}} N\left(\mathcal{F}_1, \rho_{\mathbb{Q}}, \frac{C_{\mathrm{L}}}{h^2}\varepsilon\right) \leq N\left(\mathcal{F}_1, \|\cdot\|_\infty, \frac{C_{\mathrm{L}}}{h^2}\varepsilon\right) \leq \frac{\mathrm{Leb}(\mathcal{W})}{\varepsilon},$$

where $\mathbb{Q}$ ranges over Borel probability measures on $\mathcal{W}$. Since $\frac{C_{\mathrm{k}}}{h}$ is an envelope for $\mathcal{F}_1$,

$$\sup_{\mathbb{Q}} N\left(\mathcal{F}_1, \rho_{\mathbb{Q}}, \frac{C_{\mathrm{k}}}{h}\varepsilon\right) \le \frac{C_{\mathrm{L}}}{C_{\mathrm{k}}} \frac{\mathrm{Leb}(\mathcal{W})}{h\varepsilon}.$$

Thus for all $\varepsilon \in (0, 1]$,

$$\sup_{\mathbb{Q}} N\left(\mathcal{F}_1, \rho_{\mathbb{Q}}, \frac{C_{\mathrm{k}}}{h}\varepsilon\right) \le \frac{C_{\mathrm{L}}}{C_{\mathrm{k}}} \frac{\mathrm{Leb}(\mathcal{W}) \vee 1}{h\varepsilon} \le (C_1/(h\varepsilon))^{C_2},$$

where $C_1 = \frac{C_{\mathrm{L}}}{C_{\mathrm{k}}}(\mathrm{Leb}(\mathcal{W}) \vee 1)$ and $C_2 = 1$. Next, $\mathcal{F}_2$ forms a smoothly parameterized class of functions since for $w, w' \in \mathcal{W}$ we have by the uniform Lipschitz properties of $f_{W|AA}(\cdot \mid A_i, A_j)$ and $k_h(s, \cdot)$, with $|w - w'| \le h$,

$$\begin{aligned}
&\big|\mathbb{E}\big[k_h(W_{ij}, w) \mid A_i, A_j\big] - \mathbb{E}\big[k_h(W_{ij}, w') \mid A_i, A_j\big]\big| \\
&\quad = \left|\int_{[w\pm h]\cap\mathcal{W}} k_h(s, w) f_{W|AA}(s \mid A_i, A_j)\,\mathrm{d}s - \int_{[w'\pm h]\cap\mathcal{W}} k_h(s, w') f_{W|AA}(s \mid A_i, A_j)\,\mathrm{d}s\right| \\
&\quad = \left|\int_{[w\pm 2h]\cap\mathcal{W}} \big(k_h(s, w) - k_h(s, w')\big) f_{W|AA}(s \mid A_i, A_j)\,\mathrm{d}s\right| \\
&\quad = \left|\int_{[w\pm 2h]\cap\mathcal{W}} \big(k_h(s, w) - k_h(s, w')\big)\big(f_{W|AA}(s \mid A_i, A_j) - f_{W|AA}(w \mid A_i, A_j)\big)\,\mathrm{d}s\right| \\
&\quad \le 4h\frac{C_{\mathrm{L}}}{h^2}|w - w'| 2C_{\mathrm{H}}h \le 8C_{\mathrm{L}}C_{\mathrm{H}}|w - w'| \le C_3|w - w'|,
\end{aligned}$$

where $C_3 = 8C_{\mathrm{L}}C_{\mathrm{H}}$. The same holds for $|w - w'| > h$ as the Lipschitz property is local. By taking $\mathbb{E}[\cdot \mid A_i]$, it can be seen by the contraction property of conditional expectation that the same holds for the singly-conditioned terms:

$$\big|\mathbb{E}\big[k_h(W_{ij}, w) \mid A_i\big] - \mathbb{E}\big[k_h(W_{ij}, w') \mid A_i\big]\big| \le C_3|w - w'|.$$

Therefore $\mathcal{F}_3$ is also smoothly parameterized in exactly the same manner. Let

$$\begin{aligned} C_4 &= \sup_{w\in\mathcal{W}} \operatorname*{ess\,sup}_{A_i,A_j} \left|\mathbb{E}\big[k_h(W_{ij},w) \mid A_i, A_j\big]\right| \\ &= \sup_{w\in\mathcal{W}} \operatorname*{ess\,sup}_{A_i,A_j} \left|\int_{[w\pm h]\cap\mathcal{W}} k_h(s,w) f_{W|AA}(s \mid A_i, A_j)\,\mathrm{d}s\right| \\ &\le 2h\frac{C_\mathrm{k}}{h}C_\mathrm{d} \le 2C_\mathrm{k}C_\mathrm{d}. \end{aligned}$$

For $\varepsilon \in (0,1]$, take an $(\varepsilon C_4/C_3)$-cover of $\mathcal{W}$ of cardinality at most $C_3\,\mathrm{Leb}(\mathcal{W})/(\varepsilon C_4)$. By the above parameterization properties, this cover induces an εC_4-cover for both $\mathcal{F}_2$ and $\mathcal{F}_3$:

$$\begin{aligned} \sup_{\mathbb{Q}} N\big(\mathcal{F}_2, \rho_{\mathbb{Q}}, \varepsilon C_4\big) &\le N\big(\mathcal{F}_2, \|\cdot\|_\infty, \varepsilon C_4\big) \le C_3\,\mathrm{Leb}(\mathcal{W})/(\varepsilon C_4), \\ \sup_{\mathbb{Q}} N\big(\mathcal{F}_3, \rho_{\mathbb{Q}}, \varepsilon C_4\big) &\le N\big(\mathcal{F}_3, \|\cdot\|_\infty, \varepsilon C_4\big) \le C_3\,\mathrm{Leb}(\mathcal{W})/(\varepsilon C_4). \end{aligned}$$

Hence $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$ form VC-type classes with envelopes $F_1 = C_\mathrm{k}/h$ and $F_2 = F_3 = C_4$:

$$\sup_{\mathbb{Q}} N\left(\mathcal{F}_1, \rho_{\mathbb{Q}}, \varepsilon C_\mathrm{k}/h\right) \le (C_1/(h\varepsilon))^{C_2}, \qquad \sup_{\mathbb{Q}} N\big(\mathcal{F}_2, \rho_{\mathbb{Q}}, \varepsilon C_4\big) \le (C_1/\varepsilon)^{C_2},$$
$$\sup_{\mathbb{Q}} N\big(\mathcal{F}_3, \rho_{\mathbb{Q}}, \varepsilon C_4\big) \le (C_1/\varepsilon)^{C_2},$$

for some constants $C_1 \ge e$ and $C_2 \ge 1$, where we augment the constants if necessary.

Part 2: controlling L_n

Observe that $\sqrt{n}L_n$ is the empirical process of the i.i.d. variables A_i indexed by $\mathcal{F}_3$. We apply Lemma B.2.2 with $\sigma = C_4$:

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}} \left|\sqrt{n}L_n(w)\right|\right] \lesssim C_4\sqrt{C_2 \log C_1} + \frac{C_4 C_2 \log C_1}{\sqrt{n}} \lesssim 1.$$

By Lemma 3.2.2, the left hand side is zero whenever $D_\mathrm{up} = 0$, so we can also write

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}} \left|\sqrt{n}L_n(w)\right|\right] \lesssim D_\mathrm{up}.$$

Part 3: controlling Q_n

Observe that nQ_n is the completely degenerate second-order U-process of the i.i.d. variables A_i indexed by $\mathcal{F}_2$. This function class is again uniformly bounded and VC-type, so applying the U-process maximal inequality from Lemma B.3.9 yields with $\sigma = C_4$

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}}\left|nQ_n(w)\right|\right] \lesssim C_4C_2\log C_1 + \frac{C_4(C_2\log C_1)^2}{\sqrt{n}} \lesssim 1.$$

Part 4: controlling E_n

Conditional on $\mathbf{A}_n$, note that nE_n is the empirical process of the conditionally i.n.i.d. variables W_{ij} indexed by $\mathcal{F}_1$. We apply Lemma B.2.2 conditionally with

$$\begin{aligned}\sigma^2 &= \sup_{w\in\mathcal{W}}\mathbb{E}\Big[\big(k_h(W_{ij},w) - \mathbb{E}[k_h(W_{ij},w)\mid A_i,A_j]\big)^2 \mid A_i,A_j\Big] \leq \sup_{w\in\mathcal{W}}\mathbb{E}\Big[k_h(W_{ij},w)^2\mid A_i,A_j\Big] \\ &\leq \sup_{w\in\mathcal{W}}\int_{[w\pm h]\cap\mathcal{W}} k_h(s,w)^2 f_{W|AA}(s\mid A_i,A_j)\,\mathrm{d}s \leq 2h\frac{C_{\mathrm{k}}^2}{h^2} \lesssim 1/h\end{aligned}$$

and noting that we have a sample size of $\frac{1}{2}n(n-1)$, giving

$$\begin{aligned}\mathbb{E}\left[\sup_{w\in\mathcal{W}}\left|nE_n(w)\right|\right] &\lesssim \sigma\sqrt{C_2\log\big((C_1/h)F_1/\sigma\big)} + \frac{F_1C_2\log\big((C_1/h)F_1/\sigma\big)}{n} \\ &\lesssim \frac{1}{\sqrt{h}}\sqrt{C_2\log\big((C_1/h)(C_{\mathrm{k}}/h)\sqrt{h}\big)} + \frac{(C_{\mathrm{k}}/h)C_2\log\big((C_1/h)(C_{\mathrm{k}}/h)\sqrt{h}\big)}{n} \\ &\lesssim \sqrt{\frac{\log 1/h}{h}} + \frac{\log\big(1/h\big)}{nh} \lesssim \sqrt{\frac{\log n}{h}},\end{aligned}$$

where the last line follows by the bandwidth assumption of $\frac{\log n}{n^2h} \to 0$. □

Proof (Theorem 3.3.1)

This follows from Theorem 3.2.1 and Lemma 3.2.3. □

Before proving Theorem 3.3.2 we first give a lower bound result for parametric point estimation in Lemma B.3.16.

Lemma B.3.16 (A Neyman–Pearson result for Bernoulli random variables)
Recall that the Bernoulli distribution $\mathrm{Ber}(\theta)$ *places mass* θ *at* 1 *and mass* $1-\theta$ *at* 0. *Define* $\mathbb{P}_\theta^n$ *as the law of* $(A_1, A_2, \ldots, A_n, V)$, *where* $A_1, \ldots, A_n$ *are i.i.d.* $\mathrm{Ber}(\theta)$, *and* V *is an* $\mathbb{R}^d$*-valued random variable for some* $d \geq 1$ *which is independent of the* A *variables and with a fixed distribution that does not depend on* θ. *Let* $\theta_0 = \frac{1}{2}$ *and* $\theta_{1,n} = \frac{1}{2} + \frac{1}{\sqrt{8n}}$. *Then for any estimator* $\tilde{\theta}_n$ *which is a function of* $(A_1, A_2, \ldots, A_n, V)$ *only,*

$$\mathbb{P}_{\theta_0}^n\left(\left|\tilde{\theta}_n - \theta_0\right| \geq \frac{1}{\sqrt{32n}}\right) + \mathbb{P}_{\theta_{1,n}}^n\left(\left|\tilde{\theta}_n - \theta_{1,n}\right| \geq \frac{1}{\sqrt{32n}}\right) \geq \frac{1}{2}.$$

Proof (Lemma B.3.16)
Let $f : \{0,1\}^n \to \{0,1\}$ be any function. Considering this function as a statistical test, the Neyman–Pearson lemma and Pinsker's inequality (Giné and Nickl, 2021) give

$$\begin{aligned}
\mathbb{P}_{\theta_0}^n\big(f=1\big) + \mathbb{P}_{\theta_{1,n}}^n\big(f=0\big) &\geq 1 - \mathrm{TV}\left(\mathbb{P}_{\theta_0}^n, \mathbb{P}_{\theta_{1,n}}^n\right) \geq 1 - \sqrt{\frac{1}{2}\,\mathrm{KL}\left(\mathbb{P}_{\theta_0}^n \,\middle\|\, \mathbb{P}_{\theta_{1,n}}^n\right)} \\
&= 1 - \sqrt{\frac{n}{2}\,\mathrm{KL}\left(\mathrm{Ber}(\theta_0) \,\middle\|\, \mathrm{Ber}(\theta_{1,n})\right) + \frac{n}{2}\,\mathrm{KL}\left(V \,\middle\|\, V\right)} \\
&= 1 - \sqrt{\frac{n}{2}\,\mathrm{KL}\left(\mathrm{Ber}(\theta_0) \,\middle\|\, \mathrm{Ber}(\theta_{1,n})\right)},
\end{aligned}$$

where TV is the total variation distance and KL is the Kullback–Leibler divergence. In the penultimate line we used the tensorization of Kullback–Leibler divergence (Giné and Nickl, 2021), noting that the law of V is fixed and hence does not contribute. We now evaluate this Kullback–Leibler divergence at the specified parameter values.

$$\begin{aligned}
\mathbb{P}_{\theta_0}^n\big(f=1\big) + \mathbb{P}_{\theta_{1,n}}^n\big(f=0\big) &\geq 1 - \sqrt{\frac{n}{2}\,\mathrm{KL}\left(\mathrm{Ber}(\theta_0) \,\middle\|\, \mathrm{Ber}(\theta_{1,n})\right)} \\
&= 1 - \sqrt{\frac{n}{2}}\sqrt{\theta_0 \log\frac{\theta_0}{\theta_{1,n}} + (1-\theta_0)\log\frac{1-\theta_0}{1-\theta_{1,n}}} \\
&= 1 - \sqrt{\frac{n}{2}}\sqrt{\frac{1}{2}\log\frac{1/2}{1/2+1/\sqrt{8n}} + \frac{1}{2}\log\frac{1/2}{1/2-1/\sqrt{8n}}} \\
&= 1 - \frac{\sqrt{n}}{2}\sqrt{\log\frac{1}{1-1/(2n)}} \geq 1 - \frac{\sqrt{n}}{2}\sqrt{\frac{1}{n}} = \frac{1}{2},
\end{aligned}$$

where in the penultimate line we used that $\log \frac{1}{1-x} \leq 2x$ for $x \in [0, 1/2]$. Now define a test f by $f = 1$ if $\tilde{\theta}_n > \frac{1}{2} + \frac{1}{\sqrt{32n}}$ and $f = 0$ otherwise, to see

$$\mathbb{P}_{\theta_0}^n \left(\tilde{\theta}_n > \frac{1}{2} + \frac{1}{\sqrt{32n}} \right) + \mathbb{P}_{\theta_{1,n}}^n \left(\tilde{\theta}_n \leq \frac{1}{2} + \frac{1}{\sqrt{32n}} \right) \geq \frac{1}{2}.$$

By the triangle inequality, recalling that $\theta_0 = \frac{1}{2}$ and $\theta_{1,n} = \frac{1}{2} + \frac{1}{\sqrt{8n}}$, we have

$$\begin{aligned} \left\{ \tilde{\theta}_n > \frac{1}{2} + \frac{1}{\sqrt{32n}} \right\} &\subseteq \left\{ \left| \tilde{\theta}_n - \theta_0 \right| \geq \frac{1}{\sqrt{32n}} \right\} \\ \left\{ \tilde{\theta}_n \leq \frac{1}{2} + \frac{1}{\sqrt{32n}} \right\} &\subseteq \left\{ \left| \tilde{\theta}_n - \theta_{1,n} \right| \geq \frac{1}{\sqrt{32n}} \right\}. \end{aligned}$$

Thus by the monotonicity of measures,

$$\mathbb{P}_{\theta_0}^n \left(\left| \tilde{\theta}_n - \theta_0 \right| \geq \frac{1}{\sqrt{32n}} \right) + \mathbb{P}_{\theta_{1,n}}^n \left(\left| \tilde{\theta}_n - \theta_{1,n} \right| \geq \frac{1}{\sqrt{32n}} \right) \geq \frac{1}{2}. \qquad \square$$

Proof (Theorem 3.3.2)

Part 1: lower bound for $\mathcal{P}$

By translation and scaling of the data, we may assume without loss of generality that $\mathcal{W} = [-1, 1]$. We may also assume that $C_{\mathrm{H}} \leq 1/2$, since reducing C_{H} can only shrink the class of distributions. Define the dyadic distribution $\mathbb{P}_\theta$ with parameter $\theta \in [1/2, 1]$ as follows: $A_1, \ldots, A_n$ are i.i.d. Ber(θ), while V_{ij} for $1 \leq i < j \leq n$ are i.i.d. and independent of $\mathbf{A}_n$. The distribution of V_{ij} is given by its density function $f_V(v) = \frac{1}{2} + C_{\mathrm{H}} v$ on $[-1, 1]$. Finally, generate $W_{ij} = W(A_i, A_j, V_{ij}) := (2A_i A_j - 1)V_{ij}$. Note that the function W does not depend

on θ. The conditional and marginal densities of W_{ij} are for $w \in [-1,1]$

$$f_{W|AA}(w \mid A_i, A_j) = \begin{cases} \frac{1}{2} + C_{\mathrm{H}} w & \text{if } A_i = A_j = 1, \\ \frac{1}{2} - C_{\mathrm{H}} w & \text{if } A_i = 0 \text{ or } A_j = 0, \end{cases}$$

$$f_{W|A}(w \mid A_i) = \begin{cases} \frac{1}{2} + (2\theta - 1) C_{\mathrm{H}} w & \text{if } A_i = 1, \\ \frac{1}{2} - C_{\mathrm{H}} w & \text{if } A_i = 0, \end{cases}$$

$$f_W(w) = \frac{1}{2} + (2\theta^2 - 1) C_{\mathrm{H}} w.$$

Clearly, $f_W \in \mathcal{H}^{\beta}_{C_{\mathrm{H}}}(\mathcal{W})$ and $f_{W|AA}(\cdot \mid a, a') \in \mathcal{H}^{1}_{C_{\mathrm{H}}}(\mathcal{W})$. Also $\sup_{w \in \mathcal{W}} \|f_{W|A}(w \mid \cdot)\|_{\mathrm{TV}} \leq 1$. Therefore $\mathbb{P}_\theta$ satisfies Assumption 3.2.1 and so $\left\{\mathbb{P}_\theta : \theta \in [1/2, 1]\right\} \subseteq \mathcal{P}$.

Note that $f_W(1) = \frac{1}{2} + (2\theta^2 - 1) C_{\mathrm{H}}$, so $\theta^2 = \frac{1}{2C_{\mathrm{H}}}(f_W(1) - 1/2 + C_{\mathrm{H}})$. Thus if $\tilde{f}_W$ is some density estimator depending only on the data $\mathbf{W}_n$, we define the parameter estimator

$$\tilde{\theta}_n^2 := \frac{1}{2C_{\mathrm{H}}} \left(\tilde{f}_W(1) - \frac{1}{2} + C_{\mathrm{H}} \right) \vee 0.$$

This gives the inequality

$$\begin{aligned} \left|\tilde{\theta}_n^2 - \theta^2\right| &= \left| \frac{1}{2C_{\mathrm{H}}} \left(\tilde{f}_W(1) - \frac{1}{2} + C_{\mathrm{H}} \right) \vee 0 - \frac{1}{2C_{\mathrm{H}}} \left(f_W(1) - \frac{1}{2} + C_{\mathrm{H}} \right) \right| \\ &\leq \frac{1}{2C_{\mathrm{H}}} \sup_{w \in \mathcal{W}} \left| \tilde{f}_W(w) - f_W(w) \right|. \end{aligned}$$

Therefore, since also $\tilde{\theta} \geq 0$ and $\theta \geq \frac{1}{2}$,

$$\left|\tilde{\theta}_n - \theta\right| = \frac{\left|\tilde{\theta}_n^2 - \theta^2\right|}{\tilde{\theta}_n + \theta} \leq \frac{1}{C_{\mathrm{H}}} \sup_{w \in \mathcal{W}} \left| \tilde{f}_W(w) - f_W(w) \right|.$$

Now we apply the point estimation lower bound from Lemma B.3.16, setting $\theta_0 = \frac{1}{2}$ and $\theta_{1,n} = \frac{1}{2} + \frac{1}{\sqrt{8n}}$, noting that the estimator $\tilde{\theta}_n$ is a function of $\mathbf{W}_n$ only, thus is a function of $\mathbf{A}_n$ and $\mathbf{V}_n$ only and so satisfies the conditions.

$$\begin{aligned}
&\mathbb{P}_{\theta_0}\left(\sup_{w\in\mathcal{W}}\left|\tilde{f}_W(w) - f_W^{(0)}(w)\right| \geq \frac{1}{C\sqrt{n}}\right) + \mathbb{P}_{\theta_{1,n}}\left(\sup_{w\in\mathcal{W}}\left|\tilde{f}_W(w) - f_W^{(1)}(w)\right| \geq \frac{1}{C\sqrt{n}}\right) \\
&\quad \geq \mathbb{P}_{\theta_0}\left(\left|\tilde{\theta}_n - \theta_0\right| \geq \frac{1}{CC_{\mathrm{H}}\sqrt{n}}\right) + \mathbb{P}_{\theta_{1,n}}\left(\left|\tilde{\theta}_n - \theta_{1,n}\right| \geq \frac{1}{CC_{\mathrm{H}}\sqrt{n}}\right) \\
&\quad \geq \mathbb{P}_{\theta_0}\left(\left|\tilde{\theta}_n - \theta_0\right| \geq \frac{1}{\sqrt{32n}}\right) + \mathbb{P}_{\theta_{1,n}}\left(\left|\tilde{\theta}_n - \theta_{1,n}\right| \geq \frac{1}{\sqrt{32n}}\right) \geq \frac{1}{2},
\end{aligned}$$

where we set $C \geq \frac{\sqrt{32}}{C_{\mathrm{H}}}$. Therefore we deduce that

$$\inf_{\tilde{f}_W}\sup_{\mathbb{P}\in\mathcal{P}}\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\tilde{f}_W(w) - f_W(w)\right| \geq \frac{1}{C\sqrt{n}}\right) \geq \frac{1}{4}$$

and so

$$\inf_{\tilde{f}_W}\sup_{\mathbb{P}\in\mathcal{P}}\mathbb{E}_{\mathbb{P}}\left[\sup_{w\in\mathcal{W}}\left|\tilde{f}_W(w) - f_W(w)\right|\right] \geq \frac{1}{4C\sqrt{n}}.$$

Part 2: lower bound for $\mathcal{P}_{\mathrm{d}}$

For the subclass of totally degenerate distributions, we rely on the main theorem from Khasminskii (1978). Let $\mathcal{P}_0$ be the subclass of $\mathcal{P}_{\mathrm{d}}$ consisting of the distributions which satisfy $A_1 = \cdots = A_n = 0$ and $W_{ij} := A_i + A_j + V_{ij} = V_{ij}$, so that W_{ij} are i.i.d. with common density $f_W = f_V$. Define the class

$$\mathcal{F} = \left\{f \text{ density function on } \mathbb{R},\ f \in \mathcal{H}^{\beta}_{C_{\mathrm{H}}}(\mathcal{W})\right\}.$$

Write $\mathbb{E}_f$ for the expectation under W_{ij} having density f. Then by Khasminskii (1978),

$$\liminf_{n\to\infty}\inf_{\tilde{f}_W}\sup_{f\in\mathcal{F}}\mathbb{E}_f\left[\left(\frac{n^2}{\log n}\right)^{\frac{\beta}{2\beta+1}}\sup_{w\in\mathcal{W}}\left|\tilde{f}_W(w) - f_W(w)\right|\right] > 0,$$

where $\tilde{f}_W$ is any density estimator depending only on the $\frac{1}{2}n(n-1)$ i.i.d. data samples $\mathbf{W}_n$. Now every density function in $\mathcal{H}^{\beta}_{C_{\mathrm{H}}}(\mathcal{W})$ corresponds to a distribution in $\mathcal{P}_0$ and therefore to a distribution in $\mathcal{P}_{\mathrm{d}}$. Thus for large enough n and some positive constant C,

$$\inf_{\tilde{f}_W} \sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{d}}} \mathbb{E}_{\mathbb{P}}\left[\sup_{w\in\mathcal{W}} \left|\tilde{f}_W(w) - f_W(w)\right|\right] \geq \frac{1}{C}\left(\frac{\log n}{n^2}\right)^{\frac{\beta}{2\beta+1}}.$$

Part 3: upper bounds

The upper bounds follow by using a dyadic kernel density estimator $\hat{f}_W$ with a boundary bias-corrected Lipschitz kernel of order $p \geq \beta$ and a bandwidth of h. Theorem 3.2.1 gives

$$\sup_{\mathbb{P}\in\mathcal{P}} \sup_{w\in\mathcal{W}} \left|\mathbb{E}_{\mathbb{P}}\big[\hat{f}_W(w)\big] - f_W(w)\right| \leq \frac{4C_{\mathrm{k}}C_{\mathrm{H}}}{\underline{\beta}!}h^{\beta}.$$

Then, treating the degenerate and non-degenerate cases separately and noting that all inequalities hold uniformly over $\mathcal{P}$ and $\mathcal{P}_{\mathrm{d}}$, the proof of Lemma 3.2.3 shows that

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\sup_{w\in\mathcal{W}} \left|\hat{f}_W(w) - \mathbb{E}_{\mathbb{P}}[\hat{f}_W(w)]\right|\right] \lesssim \frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{n^2 h}},$$
$$\sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{d}}} \mathbb{E}_{\mathbb{P}}\left[\sup_{w\in\mathcal{W}} \left|\hat{f}_W(w) - \mathbb{E}_{\mathbb{P}}[\hat{f}_W(w)]\right|\right] \lesssim \sqrt{\frac{\log n}{n^2 h}}.$$

Thus combining these yields that

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\sup_{w\in\mathcal{W}} \left|\hat{f}_W(w) - f_W(w)\right|\right] \lesssim h^{\beta} + \frac{1}{\sqrt{n}} + \sqrt{\frac{\log n}{n^2 h}},$$
$$\sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{d}}} \mathbb{E}_{\mathbb{P}}\left[\sup_{w\in\mathcal{W}} \left|\hat{f}_W(w) - f_W(w)\right|\right] \lesssim h^{\beta} + \sqrt{\frac{\log n}{n^2 h}}.$$

Set $h = \left(\frac{\log n}{n^2}\right)^{\frac{1}{2\beta+1}}$ and note that $\beta \geq 1$ implies that $\left(\frac{\log n}{n^2}\right)^{\frac{\beta}{2\beta+1}} \ll \frac{1}{\sqrt{n}}$. So for $C > 0$,

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\sup_{w\in\mathcal{W}} \left|\hat{f}_W(w) - f_W(w)\right|\right] \lesssim \frac{1}{\sqrt{n}} + \left(\frac{\log n}{n^2}\right)^{\frac{\beta}{2\beta+1}} \leq \frac{C}{\sqrt{n}},$$
$$\sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{d}}} \mathbb{E}_{\mathbb{P}}\left[\sup_{w\in\mathcal{W}} \left|\hat{f}_W(w) - f_W(w)\right|\right] \leq C\left(\frac{\log n}{n^2}\right)^{\frac{\beta}{2\beta+1}}. \qquad \square$$

Proof (Lemma B.1.4)

We write k_{ij} for $k_h(W_{ij}, w)$ and k'_{ij} for $k_h(W_{ij}, w')$, in the interest of brevity.

$$\begin{aligned}
\Sigma_n(w, w') &= \mathbb{E}\Big[\big(\hat{f}_W(w) - \mathbb{E}[\hat{f}_W(w)]\big)\big(\hat{f}_W(w') - \mathbb{E}[\hat{f}_W(w')]\big)\Big] \\
&= \mathbb{E}\left[\left(\frac{2}{n(n-1)}\sum_{i<j}\big(k_{ij} - \mathbb{E}k_{ij}\big)\right)\left(\frac{2}{n(n-1)}\sum_{r<s}\big(k'_{rs} - \mathbb{E}k'_{rs}\big)\right)\right] \\
&= \frac{4}{n^2(n-1)^2}\sum_{i<j}\sum_{r<s}\mathbb{E}\left[\big(k_{ij} - \mathbb{E}k_{ij}\big)\big(k'_{rs} - \mathbb{E}k'_{rs}\big)\right] \\
&= \frac{4}{n^2(n-1)^2}\sum_{i<j}\sum_{r<s}\operatorname{Cov}\left[k_{ij}, k'_{rs}\right].
\end{aligned}$$

Note first that for i, j, r, s all distinct, k_{ij} is independent of k'_{rs} and so the covariance is zero. By a counting argument, it can be seen that there are $n(n-1)/2$ summands where $|\{i, j, r, s\}| = 2$, and $n(n-1)(n-2)$ summands where $|\{i, j, r, s\}| = 3$. Therefore, since the samples are identically distributed, the value of the summands depends only on the number of distinct indices and we have the decomposition

$$\begin{aligned}
\Sigma_n(w, w') &= \frac{4}{n^2(n-1)^2}\left(\frac{n(n-1)}{2}\operatorname{Cov}[k_{ij}, k'_{ij}] + n(n-1)(n-2)\operatorname{Cov}[k_{ij}, k'_{ir}]\right) \\
&= \frac{2}{n(n-1)}\operatorname{Cov}[k_{ij}, k'_{ij}] + \frac{4(n-2)}{n(n-1)}\operatorname{Cov}[k_{ij}, k'_{ir}],
\end{aligned}$$

giving the first representation. To obtain the second representation, note that since W_{ij} and W_{ir} are independent conditional on A_i,

$$\begin{aligned}
\operatorname{Cov}\left[k_{ij}k'_{ir}\right] &= \mathbb{E}\big[k_{ij}k'_{ir}\big] - \mathbb{E}[k_{ij}]\mathbb{E}[k'_{ir}] = \mathbb{E}\big[\mathbb{E}\big[k_{ij}k'_{ir} \mid A_i\big]\big] - \mathbb{E}[k_{ij}]\mathbb{E}[k'_{ir}] \\
&= \mathbb{E}\big[\mathbb{E}[k_{ij} \mid A_i]\mathbb{E}[k'_{ir} \mid A_i]\big] - \mathbb{E}[k_{ij}]\mathbb{E}[k'_{ir}] = \operatorname{Cov}\big[\mathbb{E}[k_{ij} \mid A_i], \mathbb{E}[k'_{ir} \mid A_i]\big]. \quad \square
\end{aligned}$$

Proof (Lemma 3.2.4)

By Lemma B.1.4, the diagonal elements of Σ_n are

$$\Sigma_n(w, w) = \frac{2}{n(n-1)}\operatorname{Var}\big[k_h(W_{ij}, w)\big] + \frac{4(n-2)}{n(n-1)}\operatorname{Var}\big[\mathbb{E}[k_h(W_{ij}, w) \mid A_i]\big].$$

We bound each of the two terms separately. Firstly, note that since k_h is bounded by C_{k}/h,

$$\operatorname{Var}\left[k_h(W_{ij}, w)\right] \leq \mathbb{E}\left[k_h(W_{ij}, w)^2\right] = \int_{\mathcal{W}\cap[w\pm h]} k_h(s, w)^2 f_W(s)\,\mathrm{d}s \leq 2C_{\mathrm{k}}^2/h.$$

Since $\left|\mathbb{E}[k_h(W_{ij}, w)]\right| = \left|\int_{[w\pm h]\cap\mathcal{W}} k_h(s, w) f_W(s)\,\mathrm{d}s\right| \leq 2C_{\mathrm{k}}C_{\mathrm{d}}$, Jensen's inequality shows

$$\begin{aligned}
\operatorname{Var}\left[k_h(W_{ij}, w)\right] &\geq \int_{\mathcal{W}\cap[w\pm h]} k_h(s, w)^2 f_W(s)\,\mathrm{d}s - 4C_{\mathrm{k}}^2C_{\mathrm{d}}^2 \\
&\geq \inf_{w\in\mathcal{W}} f_W(w) \frac{1}{2h}\left(\int_{\mathcal{W}\cap[w\pm h]} k_h(s, w)\,\mathrm{d}s\right)^2 - 4C_{\mathrm{k}}^2C_{\mathrm{d}}^2 \\
&\geq \frac{1}{2h}\inf_{w\in\mathcal{W}} f_W(w) - 4C_{\mathrm{k}}^2C_{\mathrm{d}}^2 \geq \frac{1}{4h}\inf_{w\in\mathcal{W}} f_W(w)
\end{aligned}$$

for small enough h, noting that this is trivial if the infimum is zero. For the other term,

$$\operatorname{Var}\left[\mathbb{E}[k_h(W_{ij}, w) \mid A_i]\right] \leq \operatorname{Var}\left[f_{W|A}(w \mid A_i)\right] + 16C_{\mathrm{H}}C_{\mathrm{k}}C_{\mathrm{d}}h \leq 2D_{\mathrm{up}}^2$$

for small enough h, by a result from the proof of Lemma 3.2.2. Also

$$\operatorname{Var}\left[\mathbb{E}[k_h(W_{ij}, w) \mid A_i]\right] \geq \operatorname{Var}\left[f_{W|A}(w \mid A_i)\right] - 16C_{\mathrm{H}}C_{\mathrm{k}}C_{\mathrm{d}}h \geq \frac{D_{\mathrm{lo}}^2}{2}$$

for small enough h. Combining these four inequalities yields that for all large enough n,

$$\begin{aligned}
\frac{2}{n(n-1)}\frac{1}{4h}\inf_{w\in\mathcal{W}} f_W(w) + \frac{4(n-2)}{n(n-1)}\frac{D_{\mathrm{lo}}^2}{2} &\leq \inf_{w\in\mathcal{W}} \Sigma_n(w, w) \\
\leq \sup_{w\in\mathcal{W}} \Sigma_n(w, w) &\leq \frac{2}{n(n-1)}\frac{2C_{\mathrm{k}}^2}{h} + \frac{4(n-2)}{n(n-1)}2D_{\mathrm{up}}^2,
\end{aligned}$$

so that

$$\frac{D_{\mathrm{lo}}^2}{n} + \frac{1}{n^2h}\inf_{w\in\mathcal{W}} f_W(w) \lesssim \inf_{w\in\mathcal{W}} \Sigma_n(w, w) \leq \sup_{w\in\mathcal{W}} \Sigma_n(w, w) \lesssim \frac{D_{\mathrm{up}}^2}{n} + \frac{1}{n^2h}. \qquad \square$$

Proof (Lemma B.1.1)

For the strong approximation, apply the KMT corollary from Lemma B.2.3. Define

$$k_h^A(a,w) = 2\mathbb{E}[k_h(W_{ij},w) \mid A_i = a],$$

which are of bounded variation in a uniformly over w since

$$\begin{aligned}
\sup_{w\in\mathcal{W}} \|k_h^A(\cdot,w)\|_{\mathsf{T}} &= 2\sup_{w\in\mathcal{W}}\sup_{m\in\mathbb{N}}\sup_{a_0\le\cdots\le a_m}\sum_{i=1}^{m}\left|k_h^A(a_i,w)-k_h^A(a_{i-1},w)\right| \\
&= 2\sup_{w\in\mathcal{W}}\sup_{m\in\mathbb{N}}\sup_{a_0\le\cdots\le a_m}\sum_{i=1}^{m}\left|\int_{[w\pm h]\cap\mathcal{W}} k_h(s,w)\big(f_{W|A}(s\mid a_i)-f_{W|A}(s\mid a_{i-1})\big)\,\mathrm{d}s\right| \\
&\le 2\sup_{w\in\mathcal{W}}\int_{[w\pm h]\cap\mathcal{W}}|k_h(s,w)|\sup_{m\in\mathbb{N}}\sup_{a_0\le\cdots\le a_m}\sum_{i=1}^{m}\left|f_{W|A}(s\mid a_i)-f_{W|A}(s\mid a_{i-1})\right|\mathrm{d}s \\
&\le 2\sup_{w\in\mathcal{W}}\int_{[w\pm h]\cap\mathcal{W}}|k_h(s,w)|\left\|f_{W|A}(w\mid\cdot)\right\|_{\mathrm{TV}}\mathrm{d}s \\
&\le 4C_{\mathrm{k}}\sup_{w\in\mathcal{W}}\left\|f_{W|A}(w\mid\cdot)\right\|_{\mathrm{TV}} \lesssim D_{\mathrm{up}},
\end{aligned}$$

where the last line is by observing that the total variation is zero whenever $D_{\mathrm{up}} = 0$. Hence by Lemma B.2.3 there exist (on some probability space) n independent copies of A_i, denoted A_i' and a centered Gaussian process $Z_n^{L\prime}$ such that if we define

$$L_n'(w) = \frac{1}{n}\sum_{i=1}^{n}\big(k_h^A(A_i',w)-\mathbb{E}[k_h^A(A_i',w)]\big),$$

then for positive constants C_1, C_2, C_3, by defining the processes as zero outside $\mathcal{W}$ we have

$$\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\sqrt{n}L_n'(w)-Z_n^{L\prime}(w)\right| > D_{\mathrm{up}}\frac{t+C_1\log n}{\sqrt{n}}\right) \le C_2 e^{-C_3 t}.$$

Integrating tail probabilities shows that

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}}\left|\sqrt{n}L_n'(w)-Z_n^{L\prime}(w)\right|\right] \le D_{\mathrm{up}}\frac{C_1\log n}{\sqrt{n}} + \int_0^\infty \frac{D_{\mathrm{up}}}{\sqrt{n}}C_2 e^{-C_3 t}\,\mathrm{d}t \lesssim \frac{D_{\mathrm{up}}\log n}{\sqrt{n}}.$$

Further, $Z_n^{L\prime}$ has the same covariance structure as $G_n^{L\prime}$ in the sense that for all $w, w' \in \mathcal{W}$,

$$\mathbb{E}\big[Z_n^{L\prime}(w) Z_n^{L\prime}(w')\big] = \mathbb{E}\big[G_n^{L\prime}(w) G_n^{L\prime}(w')\big],$$

and clearly L_n' is equal in distribution to L_n. To obtain the trajectory regularity property of $Z_n^{L\prime}$, note that it was shown in the proof of Lemma 3.2.3 that for all $w, w' \in \mathcal{W}$,

$$\big|k_h^A(A_i, w) - k_h^A(A_i, w')\big| \leq C|w - w'|$$

for some constant $C > 0$. Therefore, since the A_i are i.i.d.,

$$\begin{aligned}
\mathbb{E}\left[\big|Z_n^{L\prime}(w) - Z_n^{L\prime}(w')\big|^2\right]^{1/2} &= \sqrt{n}\mathbb{E}\left[\big|L_n(w) - L_n(w')\big|^2\right]^{1/2} \\
&= \sqrt{n}\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \Big(k_h^A(A_i, w) - k_h^A(A_i, w') - \mathbb{E}\big[k_h^A(A_i, w)\big] + \mathbb{E}\big[k_h^A(A_i, w')\big]\Big)\right|^2\right]^{1/2} \\
&= \mathbb{E}\left[\Big|k_h^A(A_i, w) - k_h^A(A_i, w') - \mathbb{E}\big[k_h^A(A_i, w)\big] + \mathbb{E}\big[k_h^A(A_i, w')\big]\Big|^2\right]^{1/2} \lesssim |w - w'|.
\end{aligned}$$

Therefore, by the regularity result for Gaussian processes in Lemma B.3.4, with $\delta_n \in (0, 1/2]$:

$$\mathbb{E}\left[\sup_{|w-w'|\leq\delta_n} \big|Z_n^{L\prime}(w) - Z_n^{L\prime}(w')\big|\right] \lesssim \int_0^{\delta_n} \sqrt{\log 1/\varepsilon}\,\mathrm{d}\varepsilon \lesssim \delta_n\sqrt{\log 1/\delta_n} \lesssim D_{\mathrm{up}}\delta_n\sqrt{\log 1/\delta_n},$$

where the last inequality is because $Z_n^{L\prime} \equiv 0$ whenever $D_{\mathrm{up}} = 0$. There is a modification of $Z_n^{L\prime}$ with continuous trajectories by Kolmogorov's continuity criterion (Le Gall, 2016, Theorem 2.9). Note that L_n' is $\mathbf{A}_n'$-measurable and so by Lemma B.2.3 we can assume that $Z_n^{L\prime}$ depends only on $\mathbf{A}_n'$ and some random noise which is independent of $(\mathbf{A}_n', \mathbf{V}_n')$. Finally, in order to have $\mathbf{A}_n', \mathbf{V}_n', L_n'$, and $Z_n^{L\prime}$ all defined on the same probability space, we note that $\mathbf{A}_n$ and $\mathbf{V}_n$ are random vectors while L_n' and $Z_n^{L\prime}$ are stochastic processes with continuous sample paths indexed on the compact interval $\mathcal{W}$. Hence the Vorob'ev–Berkes–Philipp theorem (Lemma B.2.5) allows us to "glue" them together in the desired way on another new probability space, giving $\big(\mathbf{A}_n', \mathbf{V}_n', L_n', Z_n^{L\prime}\big)$, retaining the single prime notation for clarity. □

Proof (Lemma 3.4.1)

See Lemma B.1.1 □

Proof (Lemma B.1.2)

We apply Lemma B.2.4 conditional on $\mathbf{A}_n$. While this lemma is not in its current form stated for conditional distributions, the Yurinskii coupling on which it depends can be readily extended by following the proof of Belloni et al. (2019, Lemma 38), using a conditional version of Strassen's theorem (Chen and Kato, 2020, Theorem B.2). Care must similarly be taken in embedding the conditionally Gaussian vectors into a conditionally Gaussian process, using the Vorob'ev–Berkes–Philipp theorem (Lemma B.2.5).

By the mutual independence of A_i and V_{ij}, we have that the observations W_{ij} are independent (but not necessarily identically distributed) conditionally on $\mathbf{A}_n$. Note that $\sup_{s,w\in\mathcal{W}} |k_h(s,w)| \lesssim M_n = h^{-1}$ and $\mathbb{E}[k_h(W_{ij},w)^2 \mid \mathbf{A}_n] \lesssim \sigma_n^2 = h^{-1}$. The following uniform Lipschitz condition holds with $l_{n,\infty} = C_{\mathrm{L}} h^{-2}$, by the Lipschitz property of the kernels:

$$\sup_{s,w,w'\in\mathcal{W}} \left| \frac{k_h(s,w) - k_h(s,w')}{w - w'} \right| \leq l_{n,\infty}.$$

Also, the following L^2 Lipschitz condition holds uniformly with $l_{n,2} = 2C_{\mathrm{L}}\sqrt{C_{\mathrm{d}}}h^{-3/2}$:

$$\begin{aligned}
&\mathbb{E}\big[\big|k_h(W_{ij},w) - k_h(W_{ij},w')\big|^2 \mid \mathbf{A}_n\big]^{1/2} \\
&\quad\leq \frac{C_{\mathrm{L}}}{h^2}|w-w'| \left(\int_{([w\pm h]\cup[w'\pm h])\cap\mathcal{W}} f_{W|AA}(s \mid \mathbf{A}_n)\,\mathrm{d}s \right)^{1/2} \\
&\quad\leq \frac{C_{\mathrm{L}}}{h^2}|w-w'|\sqrt{4hC_{\mathrm{d}}} \leq l_{n,2}|w-w'|.
\end{aligned}$$

So we apply Lemma B.2.4 conditionally on $\mathbf{A}_n$ to the $\frac{1}{2}n(n-1)$ observations, noting that

$$\sqrt{n^2h}E_n(w) = \sqrt{\frac{2nh}{n-1}}\sqrt{\frac{2}{n(n-1)}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\Big(k_h(W_{ij},w) - \mathbb{E}[k_h(W_{ij},w) \mid A_i, A_j]\Big),$$

to deduce that for $t_n > 0$ there exist (an enlarged probability space) conditionally mean-zero and conditionally Gaussian processes $\tilde{Z}_n^{E\prime}(w)$ with the same conditional covariance structure as $\sqrt{n^2h}E_n(w)$ and satisfying

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\sqrt{n^2h}E_n(w)-\tilde{Z}_n^{E\prime}(w)\right|>t_n \;\middle|\; \mathbf{A}_n'\right)\\
&=\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\sqrt{\frac{n(n-1)}{2}}E_n(w)-\sqrt{\frac{n-1}{2nh}}\tilde{Z}_n^{E\prime}(w)\right|>\sqrt{\frac{n-1}{2nh}}t_n \;\middle|\; \mathbf{A}_n'\right)\\
&\lesssim \frac{\sigma_n\sqrt{\mathrm{Leb}(\mathcal{W})}\sqrt{\log n}\sqrt{M_n+\sigma_n\sqrt{\log n}}}{n^{1/2}t_n^2/h}\sqrt{l_{n,2}\sqrt{\log n}+\frac{l_{n,\infty}}{n}\log n}\\
&\lesssim \frac{h^{-1/2}\sqrt{\log n}\sqrt{h^{-1}+h^{-1/2}\sqrt{\log n}}}{n^{1/2}t_n^2/h}\sqrt{h^{-3/2}\sqrt{\log n}+\frac{h^{-2}}{n}\log n}\\
&\lesssim \sqrt{\frac{\log n}{n}}\frac{\sqrt{1+\sqrt{h\log n}}}{t_n^2}\sqrt{\sqrt{\frac{\log n}{h^3}}\left(1+\sqrt{\frac{\log n}{n^2h}}\right)}\\
&\lesssim \sqrt{\frac{\log n}{n}}\frac{1}{t_n^2}\left(\frac{\log n}{h^3}\right)^{1/4}\lesssim t_n^{-2}n^{-1/2}h^{-3/4}(\log n)^{3/4},
\end{aligned}
$$

where we used $h\lesssim 1/\log n$ and $\frac{\log n}{n^2h}\lesssim 1$. To obtain the trajectory regularity property of $\tilde{Z}_n^{E\prime}$, note that for $w,w'\in\mathcal{W}$, by conditional independence,

$$
\begin{aligned}
&\mathbb{E}\left[\left|\tilde{Z}_n^{E\prime}(w)-\tilde{Z}_n^{E\prime}(w')\right|^2 \mid \mathbf{A}_n'\right]^{1/2}=\sqrt{n^2h}\,\mathbb{E}\left[\left|E_n(w)-E_n(w')\right|^2 \mid \mathbf{A}_n\right]^{1/2}\\
&\lesssim\sqrt{n^2h}\,\mathbb{E}\left[\left|\frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\Big(k_h(W_{ij},w)-k_h(W_{ij},w')\Big)\right|^2 \;\middle|\; \mathbf{A}_n\right]^{1/2}\\
&\lesssim\sqrt{h}\,\mathbb{E}\left[\left|k_h(W_{ij},w)-k_h(W_{ij},w')\right|^2 \mid \mathbf{A}_n\right]^{1/2}\lesssim h^{-1}|w-w'|.
\end{aligned}
$$

So by the regularity result for Gaussian processes in Lemma B.3.4, with $\delta_n\in(0,1/(2h)]$:

$$
\mathbb{E}\left[\sup_{|w-w'|\le\delta_n}\left|\tilde{Z}_n^{E\prime}(w)-\tilde{Z}_n^{E\prime}(w')\right| \mid \mathbf{A}_n'\right]\lesssim\int_0^{\delta_n/h}\sqrt{\log(\varepsilon^{-1}h^{-1})}\,\mathrm{d}\varepsilon\lesssim\frac{\delta_n}{h}\sqrt{\log\frac{1}{h\delta_n}},
$$

and there exists a modification with continuous trajectories. Finally, in order to have $\mathbf{A}_n',\mathbf{V}_n',E_n'$, and $\tilde{Z}_n^{E\prime}$ all defined on the same probability space, we note that $\mathbf{A}_n$ and $\mathbf{V}_n$ are random vectors while E_n' and $\tilde{Z}_n^{E\prime}$ are stochastic processes with continuous sample paths indexed on the compact interval $\mathcal{W}$. Hence the Vorob'ev–Berkes–Philipp theorem

(Lemma B.2.5) allows us to "glue together" $\left(\mathbf{A}_n, \mathbf{V}_n, E_n\right)$ and $\left(E'_n, \tilde{Z}_n^{E\prime}\right)$ in the desired way on another new probability space, giving $\left(\mathbf{A}'_n, \mathbf{V}'_n, E'_n, \tilde{Z}_n^{E\prime}\right)$, retaining the single prime notation for clarity.

The trajectories of the conditionally Gaussian processes $\tilde{Z}_n^{E\prime}$ depend on the choice of t_n, necessitating the use of a divergent sequence R_n to establish bounds in probability. □

Proof (Lemma 3.4.2)

See Lemma B.1.2 □

Proof (Lemma B.1.3)

Part 1: defining $Z_n^{E\prime\prime}$

Pick $\delta_n \to 0$ with $\log 1/\delta_n \lesssim \log n$. Let $\mathcal{W}_\delta$ be a δ_n-covering of $\mathcal{W}$ with cardinality $\mathrm{Leb}(\mathcal{W})/\delta_n$ which is also a δ_n-packing. Let $\tilde{Z}_{n,\delta}^{E\prime}$ be the restriction of $\tilde{Z}_n^{E\prime}$ to $\mathcal{W}_\delta$. Let $\tilde{\Sigma}_n^E(w, w') = \mathbb{E}\left[\tilde{Z}_n^{E\prime}(w)\tilde{Z}_n^{E\prime}(w') \mid \mathbf{A}'_n\right]$ be the conditional covariance function of $\tilde{Z}_n^{E\prime}$, and define $\Sigma_n^E(w, w') = \mathbb{E}\left[\tilde{\Sigma}_n^E(w, w')\right]$. Let $\tilde{\Sigma}_{n,\delta}^E$ and $\Sigma_{n,\delta}^E$ be the restriction matrices of $\tilde{\Sigma}_n^E$ and Σ_n^E to $\mathcal{W}_\delta \times \mathcal{W}_\delta$, noting that, as (conditional) covariance matrices, these are (almost surely) positive semi-definite.

Let $N \sim \mathcal{N}(0, I_{|\mathcal{W}_\delta|})$ be independent of $\mathbf{A}'_n$, and define using the matrix square root $\tilde{Z}_{n,\delta}^{E\prime\prime} = \left(\tilde{\Sigma}_{n,\delta}^E\right)^{1/2} N$, which has the same distribution as $\tilde{Z}_{n,\delta}^{E\prime}$, conditional on $\mathbf{A}'_n$. Extend it using the Vorob'ev–Berkes–Philipp theorem (Lemma B.2.5) to the compact interval $\mathcal{W}$, giving a conditionally Gaussian process $\tilde{Z}_n^{E\prime\prime}$ which has the same distribution as $\tilde{Z}_n^{E\prime}$, conditional on $\mathbf{A}'_n$. Define $Z_{n,\delta}^{E\prime\prime} = \left(\Sigma_{n,\delta}^E\right)^{1/2} N$, noting that this is independent of $\mathbf{A}'_n$, and extend it using the Vorob'ev–Berkes–Philipp theorem (Lemma B.2.5) to a Gaussian process $Z_n^{E\prime\prime}$ on the compact interval $\mathcal{W}$, which is independent of $\mathbf{A}'_n$ and has covariance structure given by Σ_n^E.

Part 2: closeness of $Z_n^{E\prime\prime}$ and $\tilde{Z}_n^{E\prime\prime}$ on the mesh

Note that conditionally on $\mathbf{A}'_n$, $\tilde{Z}_{n,\delta}^{E\prime\prime} - Z_{n,\delta}^{E\prime\prime}$ is a length-$|\mathcal{W}_\delta|$ Gaussian random vector with covariance matrix $\left(\left(\tilde{\Sigma}_{n,\delta}^E\right)^{1/2} - \left(\Sigma_{n,\delta}^E\right)^{1/2}\right)^2$. So by the Gaussian maximal inequality in Lemma B.3.3 applied conditionally on $\mathbf{A}'_n$,

$$\mathbb{E}\left[\max_{w \in \mathcal{W}_\delta} \left|\tilde{Z}_n^{E\prime\prime}(w) - Z_n^{E\prime\prime}(w)\right| \;\middle|\; \mathbf{A}'_n\right] \lesssim \sqrt{\log n}\left\|\tilde{\Sigma}_{n,\delta}^E - \Sigma_{n,\delta}^E\right\|_2^{1/2},$$

since $\log|\mathcal{W}_\delta| \lesssim \log n$. Next, we apply some U-statistic theory to $\tilde{\Sigma}^E_{n,\delta} - \Sigma^E_{n,\delta}$, with the aim of applying the matrix concentration result for second-order U-statistics presented in Lemma B.3.10. Firstly, we note that since the conditional covariance structures of $\tilde{Z}^{E\prime}_n$ and $\sqrt{n^2h}E_n$ are equal in distribution, we have, writing $E_n(\mathcal{W}_\delta)$ for the vector $\big(E_n(w) : w \in \mathcal{W}_\delta\big)$ and similarly for $k_h(W_{ij}, \mathcal{W}_\delta)$,

$$\begin{aligned}
\tilde{\Sigma}^E_{n,\delta} &= n^2h\mathbb{E}[E_n(\mathcal{W}_\delta)E_n(\mathcal{W}_\delta)^\mathsf{T} \mid \mathbf{A}_n] \\
&= n^2h\frac{4}{n^2(n-1)^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}\Big[\Big(k_h(W_{ij},\mathcal{W}_\delta) - \mathbb{E}\left[k_h(W_{ij},\mathcal{W}_\delta) \mid \mathbf{A}_n\right]\Big) \\
&\qquad \times\Big(k_h(W_{ij},\mathcal{W}_\delta) - \mathbb{E}\left[k_h(W_{ij},\mathcal{W}_\delta) \mid \mathbf{A}_n\right]\Big)^\mathsf{T} \mid \mathbf{A}_n\Big] \\
&= \frac{4h}{(n-1)^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} u(A_i, A_j),
\end{aligned}$$

where we define the random $|\mathcal{W}_\delta| \times |\mathcal{W}_\delta|$ matrices

$$u(A_i, A_j) = \mathbb{E}\Big[k_h(W_{ij},\mathcal{W}_\delta)k_h(W_{ij},\mathcal{W}_\delta)^\mathsf{T} \mid \mathbf{A}_n\Big] - \mathbb{E}[k_h(W_{ij},\mathcal{W}_\delta) \mid \mathbf{A}_n]\,\mathbb{E}[k_h(W_{ij},\mathcal{W}_\delta) \mid \mathbf{A}_n]^\mathsf{T}.$$

Let $u(A_i) = \mathbb{E}[u(A_i, A_j) \mid A_i]$ and $u = \mathbb{E}[u(A_i, A_j)]$. The decomposition $\tilde{\Sigma}^E_{n,\delta} - \Sigma^E_{n,\delta} = \tilde{L} + \tilde{Q}$ holds by Lemma B.3.7, where

$$\tilde{L} = \frac{4h}{n-1}\sum_{i=1}^{n}\big(u(A_i) - u\big), \quad \tilde{Q} = \frac{4h}{(n-1)^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\big(u(A_i, A_j) - u(A_i) - u(A_j) + u\big).$$

Next, we seek an almost sure upper bound on $\|u(A_i, A_j)\|_2$. Since this is a symmetric matrix, we have by Hölder's inequality

$$\|u(A_i, A_j)\|_2 \leq \|u(A_i, A_j)\|_1^{1/2}\|u(A_i, A_j)\|_\infty^{1/2} = \max_{1\leq k\leq|\mathcal{W}_\delta|}\sum_{l=1}^{|\mathcal{W}_\delta|}|u(A_i, A_j)_{kl}|.$$

The terms on the right hand side can be bounded as follows, writing w, w' for the kth and lth points in $\mathcal{W}_\delta$ respectively:

$$\begin{aligned}
|u(A_i, A_j)_{kl}| &= \left|\mathbb{E}\left[k_h(W_{ij}, w)k_h(W_{ij}, w') \mid \mathbf{A}_n\right] - \mathbb{E}\left[k_h(W_{ij}, w) \mid \mathbf{A}_n\right]\mathbb{E}\left[k_h(W_{ij}, w') \mid \mathbf{A}_n\right]\right| \\
&\lesssim \mathbb{E}\left[|k_h(W_{ij}, w)k_h(W_{ij}, w')| \mid \mathbf{A}_n\right] + \mathbb{E}\left[|k_h(W_{ij}, w)| \mid \mathbf{A}_n\right]\mathbb{E}\left[|k_h(W_{ij}, w')| \mid \mathbf{A}_n\right] \\
&\lesssim h^{-1}\mathbb{I}\big\{|w - w'| \leq 2h\big\} + 1 \lesssim h^{-1}\mathbb{I}\big\{|k - l| \leq 2h/\delta_n\big\} + 1,
\end{aligned}$$

where we used that $|w - w'| \geq |k - l|\delta_n$ because $\mathcal{W}_\delta$ is a δ_n-packing. Hence

$$\begin{aligned}
\|u(A_i, A_j)\|_2 &\leq \max_{1\leq k\leq|\mathcal{W}_\delta|}\sum_{l=1}^{|\mathcal{W}_\delta|}|u(A_i, A_j)_{kl}| \lesssim \max_{1\leq k\leq|\mathcal{W}_\delta|}\sum_{l=1}^{|\mathcal{W}_\delta|}\Big(h^{-1}\mathbb{I}\big\{|k - l| \leq 2h/\delta_n\big\} + 1\Big) \\
&\lesssim 1/\delta_n + 1/h + |\mathcal{W}_\delta| \lesssim 1/\delta_n + 1/h.
\end{aligned}$$

Clearly, the same bound holds for $\|u(A_i)\|_2$ and $\|u\|_2$, by Jensen's inequality. Therefore, applying the matrix Bernstein inequality (Lemma B.3.2) to the zero-mean matrix $\tilde{L}$ gives

$$\mathbb{E}\left[\left\|\tilde{L}\right\|_2\right] \lesssim \frac{h}{n}\left(\frac{1}{\delta_n} + \frac{1}{h}\right)\left(\log|\mathcal{W}_\delta| + \sqrt{n\log|\mathcal{W}_\delta|}\right) \lesssim \left(\frac{h}{\delta_n} + 1\right)\sqrt{\frac{\log n}{n}}.$$

The matrix U-statistic concentration inequality (Lemma B.3.10) with $\tilde{Q}$ gives

$$\mathbb{E}\left[\left\|\tilde{Q}\right\|_2\right] \lesssim \frac{h}{n^2}n\left(\frac{1}{\delta_n} + \frac{1}{h}\right)(\log|\mathcal{W}_\delta|)^{3/2} \lesssim \left(\frac{h}{\delta_n} + 1\right)\frac{(\log n)^{3/2}}{n}.$$

Hence taking a marginal expectation and applying Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\max_{w\in\mathcal{W}_\delta}\left|\tilde{Z}_n^{E\prime\prime}(w) - Z_n^{E\prime\prime}(w)\right|\right] \\
&\quad\lesssim \sqrt{\log n}\,\mathbb{E}\left[\left\|\tilde{\Sigma}_{n,\delta}^E - \Sigma_{n,\delta}^E\right\|_2^{1/2}\right] \lesssim \sqrt{\log n}\,\mathbb{E}\left[\left\|\tilde{\Sigma}_{n,\delta}^E - \Sigma_{n,\delta}^E\right\|_2\right]^{1/2} \\
&\quad\lesssim \sqrt{\log n}\,\mathbb{E}\left[\left\|\tilde{L} + \tilde{Q}\right\|_2\right]^{1/2} \lesssim \sqrt{\log n}\,\mathbb{E}\left[\left\|\tilde{L}\right\|_2 + \left\|\tilde{Q}\right\|_2\right]^{1/2} \\
&\quad\lesssim \sqrt{\log n}\left(\left(\frac{h}{\delta_n} + 1\right)\sqrt{\frac{\log n}{n}} + \left(\frac{h}{\delta_n} + 1\right)\frac{(\log n)^{3/2}}{n}\right)^{1/2} \\
&\quad\lesssim \sqrt{\frac{h}{\delta_n} + 1}\frac{(\log n)^{3/4}}{n^{1/4}}.
\end{aligned}$$

Part 3: regularity of Z_n^E and $\tilde{Z}_n^{E\prime}$

Define the semimetrics

$$\rho(w,w')^2 = \mathbb{E}\left[\left|Z_n^{E\prime\prime}(w) - Z_n^{E\prime\prime}(w')\right|^2\right], \quad \tilde{\rho}(w,w')^2 = \mathbb{E}\left[\left|\tilde{Z}_n^{E\prime\prime}(w) - \tilde{Z}_n^{E\prime\prime}(w')\right|^2 \mid \mathbf{A}_n\right].$$

We bound $\tilde{\rho}$ as follows, since $\tilde{Z}_n^{E\prime\prime}$ and $\sqrt{n^2h}E_n$ have the same conditional covariance structure:

$$\begin{aligned}\tilde{\rho}(w,w') &= \mathbb{E}\left[\left|\tilde{Z}_n^{E\prime\prime}(w) - \tilde{Z}_n^{E\prime\prime}(w')\right|^2 \mid \mathbf{A}_n'\right]^{1/2} \\ &= \sqrt{n^2h}\,\mathbb{E}\left[\left|E_n(w) - E_n(w')\right|^2 \mid \mathbf{A}_n'\right]^{1/2} \lesssim h^{-1}|w-w'|,\end{aligned}$$

uniformly in $\mathbf{A}_n'$, where the last line was shown in the proof of Lemma B.1.2. Note that also

$$\rho(w,w') = \sqrt{\mathbb{E}[\tilde{\rho}(w,w')^2]} \lesssim h^{-1}|w-w'|.$$

Thus Lemma B.3.4 applies directly to Z_n^E and conditionally to $\tilde{Z}_n^{E\prime}$, with $\delta_n \in (0, 1/(2h)]$, demonstrating that

$$\begin{aligned}\mathbb{E}\left[\sup_{|w-w'|\le\delta_n}\left|\tilde{Z}_n^{E\prime\prime}(w) - \tilde{Z}_n^{E\prime\prime}(w')\right| \mid \mathbf{A}_n'\right] &\lesssim \int_0^{\delta_n/h}\sqrt{\log(1/(\varepsilon h))}\,\mathrm{d}\varepsilon \lesssim \frac{\delta_n}{h}\sqrt{\log\frac{1}{h\delta_n}}, \\ \mathbb{E}\left[\sup_{|w-w'|\le\delta_n}\left|Z_n^{E\prime\prime}(w) - Z_n^{E\prime\prime}(w')\right|\right] &\lesssim \int_0^{\delta_n/h}\sqrt{\log(1/(\varepsilon h))}\,\mathrm{d}\varepsilon \lesssim \frac{\delta_n}{h}\sqrt{\log\frac{1}{h\delta_n}}.\end{aligned}$$

Continuity of trajectories follows from this.

Part 4: conclusion

We use the previous parts to deduce that

$$\begin{aligned}&\mathbb{E}\left[\sup_{w\in\mathcal{W}}\left|\tilde{Z}_n^{E\prime\prime}(w) - Z_n^{E\prime\prime}(w)\right|\right] \\ &\quad\lesssim \mathbb{E}\left[\max_{w\in\mathcal{W}_\delta}\left|\tilde{Z}_n^{E\prime\prime}(w) - Z_n^{E\prime\prime}(w)\right|\right] \\ &\qquad + \mathbb{E}\left[\sup_{|w-w'|\le\delta_n}\left\{\left|\tilde{Z}_n^{E\prime\prime}(w) - \tilde{Z}_n^{E\prime\prime}(w')\right| + \left|Z_n^{E\prime\prime}(w) - Z_n^{E\prime\prime}(w')\right|\right\}\right] \\ &\quad\lesssim \sqrt{\frac{h}{\delta_n}+1}\,\frac{(\log n)^{3/4}}{n^{1/4}} + \frac{\delta_n\sqrt{\log n}}{h}.\end{aligned}$$

Setting $\delta_n = h\left(\frac{\log n}{n}\right)^{1/6}$ gives

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}}\left|\tilde{Z}_n^{E\prime\prime}(w) - Z_n^{E\prime\prime}(w)\right|\right] \lesssim n^{-1/6}(\log n)^{2/3}.$$

Independence of $Z_n^{E\prime\prime}$ and $\mathbf{A}_n''$ follows by applying the Vorob'ev–Berkes–Philipp theorem (Lemma B.2.5), conditionally on $\mathbf{A}_n'$, to the variables $\left(\mathbf{A}_n', \tilde{Z}_n^{E\prime}\right)$ and $\left(\tilde{Z}_n^{E\prime\prime}, Z_n^{E\prime\prime}\right)$. □

Proof (Lemma 3.4.3)

See Lemma B.1.3 □

Proof (Theorem B.1.1)

We add together the strong approximations for the L_n and E_n terms, and then add an independent Gaussian process to account for the variance of Q_n.

Part 1: gluing together the strong approximations

Let $\left(\mathbf{A}_n', \mathbf{V}_n', L_n', Z_n^{L\prime}\right)$ be the strong approximation for L_n derived in Lemma B.1.1. Let $\left(\mathbf{A}_n'', \mathbf{V}_n'', E_n'', \tilde{Z}_n^{E\prime\prime}\right)$ and $\left(\mathbf{A}_n''', \mathbf{V}_n''', \tilde{Z}_n^{E\prime\prime\prime}, Z_n^{E\prime\prime\prime}\right)$ be the conditional and unconditional strong approximations for E_n given in Lemmas B.1.2 and B.1.3 respectively. The first step is to define copies of these variables and processes on the same probability space. This is achieved by applying the Vorob'ev–Berkes–Philipp theorem (Lemma B.2.5). Dropping the prime notation for clarity, we construct $\left(\mathbf{A}_n, \mathbf{V}_n, L_n, Z_n^L, E_n, \tilde{Z}_n^E, Z_n^E\right)$ with the following properties:

(i) $\sup_{w\in\mathcal{W}}\left|\sqrt{n}L_n(w) - Z_n^L(w)\right| \lesssim_{\mathbb{P}} n^{-1/2}\log n$,

(ii) $\sup_{w\in\mathcal{W}}\left|\sqrt{n^2h}E_n(w) - \tilde{Z}_n^E(w)\right| \lesssim_{\mathbb{P}} n^{-1/4}h^{-3/8}(\log n)^{3/8}R_n$,

(iii) $\sup_{w\in\mathcal{W}}\left|\tilde{Z}_n^E(w) - Z_n^E(w)\right| \lesssim_{\mathbb{P}} n^{-1/6}(\log n)^{2/3}$,

(iv) Z_n^L is independent of Z_n^E.

Note that the independence of Z_n^L and Z_n^E follows since Z_n^L depends only on $\mathbf{A}_n$ and some independent random noise, while Z_n^E is independent of $\mathbf{A}_n$. Therefore (Z_n^L, Z_n^E) are jointly Gaussian. To get the strong approximation result for $\hat{f}_W$, define the Gaussian process

$$Z_n^f(w) = \frac{1}{\sqrt{n}} Z_n^L(w) + \frac{1}{n} Z_n^Q(w) + \frac{1}{\sqrt{n^2 h}} Z_n^E(w),$$

where $Z_n^Q(w)$ is a mean-zero Gaussian process independent of everything else with covariance

$$\mathbb{E}\big[Z_n^Q(w) Z_n^Q(w')\big] = n^2 \mathbb{E}\big[Q_n(w) Q_n(w')\big].$$

As shown in the proof of Lemma 3.2.3, the process $Q_n(w)$ is uniformly Lipschitz and uniformly bounded in w. Thus by Lemma B.3.4, we have $\mathbb{E}\big[\sup_{w\in\mathcal{W}} |Z_n^Q(w)|\big] \lesssim 1$. Therefore the uniform approximation error is given by

$$\begin{aligned}
&\sup_{w\in\mathcal{W}} \big|\hat{f}_W(w) - \mathbb{E}[\hat{f}_W(w)] - Z_n^f(w)\big| \\
&\quad = \sup_{w\in\mathcal{W}} \left| \frac{1}{\sqrt{n}} Z_n^L(w) + \frac{1}{n} Z_n^Q(w) + \frac{1}{\sqrt{n^2 h}} Z_n^E(w) - \Big(L_n(w) + Q_n(w) + E_n(w)\Big) \right| \\
&\quad \le \sup_{w\in\mathcal{W}} \left(\frac{1}{\sqrt{n}} \big|Z_n^L(w) - \sqrt{n} L_n(w)\big| + \frac{1}{\sqrt{n^2 h}} \left|\tilde{Z}_n^E(w) - \sqrt{n^2 h} E_n(w)\right| \right. \\
&\qquad \left. + \frac{1}{\sqrt{n^2 h}} \left|Z_n^E(w) - \tilde{Z}_n^E(w)\right| \big|Q_n(w)\big| + \frac{1}{n} \big|Z_n^Q(w)\big| \right) \\
&\quad \lesssim_{\mathbb{P}} n^{-1} \log n + n^{-5/4} h^{-7/8} (\log n)^{3/8} R_n + n^{-7/6} h^{-1/2} (\log n)^{2/3}.
\end{aligned}$$

Part 2: covariance structure

Since L_n, Q_n, and E_n are mutually orthogonal in L^2 (as shown in Lemma 3.2.1), we have the following covariance structure:

$$\begin{aligned}
\mathbb{E}\big[Z_n^f(w) Z_n^f(w')\big] &= \frac{1}{n} \mathbb{E}\big[Z_n^L(w) Z_n^L(w')\big] + \frac{1}{n^2} \mathbb{E}\big[Z_n^Q(w) Z_n^Q(w')\big] + \frac{1}{n^2 h} \mathbb{E}\big[Z_n^E(w) Z_n^E(w')\big] \\
&= \mathbb{E}\big[L_n(w) L_n(w')\big] + \mathbb{E}\big[Q_n(w) Q_n(w')\big] + \mathbb{E}\big[E_n(w) E_n(w')\big] \\
&= \mathbb{E}\big[\big(\hat{f}_W(w) - \mathbb{E}[\hat{f}_W(w)]\big)\big(\hat{f}_W(w') - \mathbb{E}[\hat{f}_W(w')]\big)\big].
\end{aligned}$$

Part 3: trajectory regularity

The trajectory regularity of the process Z_n^f follows directly by adding the regularities of the processes $\frac{1}{\sqrt{n}}Z_n^L$, $\frac{1}{n}Z_n^Q$, and $\frac{1}{\sqrt{n^2h}}Z_n^E$. Similarly, Z_n^f has continuous trajectories. □

Proof (Theorem 3.4.1)

Define $Z_n^T(w) = \frac{Z_n^f(w)}{\sqrt{\Sigma_n(w,w)}}$ so that

$$\left|T_n(w) - Z_n^T(w)\right| = \frac{\left|\hat{f}_W(w) - f_W(w) - Z_n^f(w)\right|}{\sqrt{\Sigma_n(w,w)}}.$$

By Theorems B.1.1 and 3.2.1, the numerator can be bounded above by

$$\begin{aligned}
&\sup_{w\in\mathcal{W}}\left|\hat{f}_W(w) - f_W(w) - Z_n^f(w)\right| \\
&\quad\leq \sup_{w\in\mathcal{W}}\left|\hat{f}_W(w) - \mathbb{E}\big[\hat{f}_W(w)\big] - Z_n^f(w)\right| + \sup_{w\in\mathcal{W}}\left|\mathbb{E}\big[\hat{f}_W(w)\big] - f_W(w)\right| \\
&\quad\lesssim_{\mathbb{P}} n^{-1}\log n + n^{-5/4}h^{-7/8}(\log n)^{3/8}R_n + n^{-7/6}h^{-1/2}(\log n)^{2/3} + h^{p\wedge\beta}.
\end{aligned}$$

By Lemma 3.2.4 with $\inf_{\mathcal{W}} f_W(w) > 0$, the denominator is bounded below by

$$\inf_{w\in\mathcal{W}}\sqrt{\Sigma_n(w,w)} \gtrsim \frac{D_{\text{lo}}}{\sqrt{n}} + \frac{1}{\sqrt{n^2h}},$$

and the result follows. □

Proof (Theorem 3.4.2)

Note that the covariance structure of Z_n^T is given by

$$\text{Cov}\left[Z_n^T(w), Z_n^T(w')\right] = \frac{\Sigma_n(w,w')}{\sqrt{\Sigma_n(w,w)\Sigma_n(w',w')}}.$$

We apply an anti-concentration result to establish that all quantiles of $\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right|$ exist. To do this, we must first establish regularity properties of Z_n^T.

Part 1: L^2 regularity of Z_n^T

Writing k'_{ij} for $k_h(W_{ij}, w')$ etc., note that by Lemma B.1.4,

$$\begin{aligned}
&\left|\Sigma_n(w,w') - \Sigma_n(w,w'')\right| \\
&\quad= \left| \frac{2}{n(n-1)} \operatorname{Cov}\left[k_{ij}, k'_{ij}\right] + \frac{4(n-2)}{n(n-1)} \operatorname{Cov}\left[k_{ij}, k'_{ir}\right] \right. \\
&\qquad \left. - \frac{2}{n(n-1)} \operatorname{Cov}\left[k_{ij}, k''_{ij}\right] - \frac{4(n-2)}{n(n-1)} \operatorname{Cov}\left[k_{ij}, k''_{ir}\right] \right| \\
&\quad\le \frac{2}{n(n-1)} \left| \operatorname{Cov}\left[k_{ij}, k'_{ij} - k''_{ij}\right] \right| + \frac{4(n-2)}{n(n-1)} \left| \operatorname{Cov}\left[k_{ij}, k'_{ir} - k''_{ir}\right] \right| \\
&\quad\le \frac{2}{n(n-1)} \|k_{ij}\|_\infty \|k'_{ij} - k''_{ij}\|_\infty + \frac{4(n-2)}{n(n-1)} \|k_{ij}\|_\infty \|k'_{ir} - k''_{ir}\|_\infty \\
&\quad\le \frac{4}{nh^3} C_{\mathrm{k}} C_{\mathrm{L}} |w' - w''| \lesssim n^{-1}h^{-3}|w' - w''|
\end{aligned}$$

uniformly in $w, w', w'' \in \mathcal{W}$. Therefore, by Lemma 3.2.4, with $\delta_n \le n^{-2}h^2$, we have

$$\begin{aligned}
\inf_{|w-w'|\le\delta_n} \Sigma_n(w,w') &\gtrsim \frac{D_{\mathrm{lo}}^2}{n} + \frac{1}{n^2h} - n^{-1}h^{-3}\delta_n \gtrsim \frac{D_{\mathrm{lo}}^2}{n} + \frac{1}{n^2h} - \frac{1}{n^3h} \gtrsim \frac{D_{\mathrm{lo}}^2}{n} + \frac{1}{n^2h}, \\
\sup_{|w-w'|\le\delta_n} \Sigma_n(w,w') &\lesssim \frac{D_{\mathrm{up}}^2}{n} + \frac{1}{n^2h} + n^{-1}h^{-3}\delta_n \lesssim \frac{D_{\mathrm{up}}^2}{n} + \frac{1}{n^2h} + \frac{1}{n^3h} \lesssim \frac{D_{\mathrm{up}}^2}{n} + \frac{1}{n^2h}.
\end{aligned}$$

The L^2 regularity of Z_n^T is

$$\mathbb{E}\left[\left(Z_n^T(w) - Z_n^T(w')\right)^2\right] = 2 - 2\frac{\Sigma_n(w,w')}{\sqrt{\Sigma_n(w,w)\Sigma_n(w',w')}}.$$

Applying the elementary result that for $a, b, c > 0$,

$$1 - \frac{a}{\sqrt{bc}} = \frac{b(c-a) + a(b-a)}{\sqrt{bc}\left(\sqrt{bc} + a\right)},$$

with $a = \Sigma_n(w,w')$, $b = \Sigma_n(w,w)$, and $c = \Sigma_n(w',w')$, and noting $|c-a| \lesssim n^{-1}h^{-3}|w-w'|$ and $|b-a| \lesssim n^{-1}h^{-3}|w-w'|$ and $\frac{D_{\mathrm{lo}}^2}{n} + \frac{1}{n^2h} \lesssim a, b, c \lesssim \frac{D_{\mathrm{up}}^2}{n} + \frac{1}{n^2h}$, yields

$$\begin{aligned}
\mathbb{E}\left[\left(Z_n^T(w) - Z_n^T(w')\right)^2\right] &\lesssim \frac{(D_{\mathrm{up}}^2/n + 1/(n^2h))n^{-1}h^{-3}|w-w'|}{(D_{\mathrm{lo}}^2/n + 1/(n^2h))^2} \\
&\lesssim \frac{n^2h^{-4}|w-w'|}{n^{-4}h^{-2}} \lesssim n^2h^{-2}|w-w'|.
\end{aligned}$$

Thus the semimetric induced by Z_n^T on $\mathcal{W}$ is

$$\rho(w, w') := \mathbb{E}\left[\left(Z_n^T(w) - Z_n^T(w')\right)^2\right]^{1/2} \lesssim nh^{-1}\sqrt{|w - w'|}.$$

Part 2: trajectory regularity of Z_n^T

By the bound on ρ from the previous part, we deduce the covering number bound

$$\begin{aligned} N(\varepsilon, \mathcal{W}, \rho) &\lesssim N\big(\varepsilon, \mathcal{W}, nh^{-1}\sqrt{|\cdot|}\big) \lesssim N\big(n^{-1}h\varepsilon, \mathcal{W}, \sqrt{|\cdot|}\big) \\ &\lesssim N\big(n^{-2}h^2\varepsilon^2, \mathcal{W}, |\cdot|\big) \lesssim n^2h^{-2}\varepsilon^{-2}. \end{aligned}$$

Now apply the Gaussian process regularity result from Lemma B.3.4.

$$\begin{aligned} \mathbb{E}\left[\sup_{\rho(w,w')\le\delta} \left|Z_n^T(w) - Z_n^T(w')\right|\right] &\lesssim \int_0^\delta \sqrt{\log N(\varepsilon, \mathcal{W}, \rho)}\,\mathrm{d}\varepsilon \lesssim \int_0^\delta \sqrt{\log(n^2h^{-2}\varepsilon^{-2})}\,\mathrm{d}\varepsilon \\ &\lesssim \int_0^\delta \left(\sqrt{\log n} + \sqrt{\log 1/\varepsilon}\right)\,\mathrm{d}\varepsilon \lesssim \delta\left(\sqrt{\log n} + \sqrt{\log 1/\delta}\right), \end{aligned}$$

and so

$$\mathbb{E}\left[\sup_{|w-w'|\le\delta_n} \left|Z_n^T(w) - Z_n^T(w')\right|\right] \lesssim \mathbb{E}\left[\sup_{\rho(w,w')\le nh^{-1}\delta_n^{1/2}} \left|Z_n^T(w) - Z_n^T(w')\right|\right] \lesssim nh^{-1}\sqrt{\delta_n \log n},$$

whenever $1/\delta_n$ is at most polynomial in n.

Part 3: existence of the quantile

Apply the Gaussian anti-concentration result from Lemma B.3.5, noting that Z_n^T is separable, mean-zero, and has unit variance:

$$\sup_{t\in\mathbb{R}} \mathbb{P}\left(\left|\sup_{w\in\mathcal{W}} \left|Z_n^T(w)\right| - t\right| \le 2\varepsilon_n\right) \le 8\varepsilon_n\left(1 + \mathbb{E}\left[\sup_{w\in\mathcal{W}} \left|Z_n^T(w)\right|\right]\right).$$

To bound the supremum on the right hand side, apply the Gaussian process maximal inequality from Lemma B.3.4 with $\sigma \le 1$ and $N(\varepsilon, \mathcal{W}, \rho) \lesssim n^2h^{-2}\varepsilon^{-2}$:

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}} \left|Z_n^T(w)\right|\right] \lesssim 1 + \int_0^2 \sqrt{\log(n^2h^{-2}\varepsilon^{-2})}\,\mathrm{d}\varepsilon \lesssim \sqrt{\log n}.$$

Therefore

$$\sup_{t\in\mathbb{R}} \mathbb{P}\left(\left|\sup_{w\in\mathcal{W}} \left|Z_n^T(w)\right| - t\right| \leq \varepsilon\right) \lesssim \varepsilon\sqrt{\log n}.$$

Letting $\varepsilon \to 0$ shows that the distribution function of $\sup_{w\in\mathcal{W}} \left|Z_n^T(w)\right|$ is continuous, and therefore all of its quantiles exist.

Part 4: validity of the infeasible uniform confidence band

Under Assumption 3.4.1 and with a sufficiently slowly diverging sequence R_n, the strong approximation rate established in Theorem 3.4.1 is

$$\begin{aligned}
&\sup_{w\in\mathcal{W}} \left|T_n(w) - Z_n^T(w)\right| \\
&\quad\lesssim_{\mathbb{P}} \frac{n^{-1/2}\log n + n^{-3/4}h^{-7/8}(\log n)^{3/8}R_n + n^{-2/3}h^{-1/2}(\log n)^{2/3} + n^{1/2}h^{p\wedge\beta}}{D_{\text{lo}} + 1/\sqrt{nh}} \ll \frac{1}{\sqrt{\log n}}.
\end{aligned}$$

So by Lemma B.3.6, take ε_n such that

$$\mathbb{P}\left(\sup_{w\in\mathcal{W}} \left|T_n(w) - Z_n^T(w)\right| > \varepsilon_n\right) \leq \varepsilon_n\sqrt{\log n}$$

and $\varepsilon_n\sqrt{\log n} \to 0$. So by the previously established anti-concentration result,

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{f}_W(w) - f_W(w)\right| \leq q_{1-\alpha}\sqrt{\Sigma_n(w,w)} \text{ for all } w \in \mathcal{W}\right) \\
&\quad= \mathbb{P}\left(\sup_{w\in\mathcal{W}} |T_n(w)| \leq q_{1-\alpha}\right) \\
&\quad\leq \mathbb{P}\left(\sup_{w\in\mathcal{W}} \left|Z_n^T(w)\right| \leq q_{1-\alpha} + \varepsilon_n\right) + \mathbb{P}\left(\sup_{w\in\mathcal{W}} \left|T_n(w) - Z_n^T(w)\right| > \varepsilon_n\right) \\
&\quad\leq \mathbb{P}\left(\sup_{w\in\mathcal{W}} \left|Z_n^T(w)\right| \leq q_{1-\alpha}\right) + \mathbb{P}\left(\left|\sup_{w\in\mathcal{W}} \left|Z_n^T(w)\right| - q_{1-\alpha}\right| \leq \varepsilon_n\right) + \varepsilon_n\sqrt{\log n} \\
&\quad\leq 1 - \alpha + 2\varepsilon_n\sqrt{\log n}.
\end{aligned}$$

The lower bound follows analogously:

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{f}_W(w) - f_W(w)\right| \leq q_{1-\alpha}\sqrt{\Sigma_n(w,w)} \text{ for all } w \in \mathcal{W}\right) \\
&\quad \geq \mathbb{P}\left(\sup_{w\in\mathcal{W}} \left|Z_n^T(w)\right| \leq q_{1-\alpha} - \varepsilon_n\right) - \varepsilon_n\sqrt{\log n} \\
&\quad \geq \mathbb{P}\left(\sup_{w\in\mathcal{W}} \left|Z_n^T(w)\right| \leq q_{1-\alpha}\right) - \mathbb{P}\left(\left|\sup_{w\in\mathcal{W}} \left|Z_n^T(w)\right| - q_{1-\alpha}\right| \leq \varepsilon_n\right) - \varepsilon_n\sqrt{\log n} \\
&\quad \leq 1 - \alpha - 2\varepsilon_n\sqrt{\log n}.
\end{aligned}$$

Finally, we apply $\varepsilon_n\sqrt{\log n} \to 0$ to see

$$\left|\mathbb{P}\left(\left|\hat{f}_W(w) - f_W(w)\right| \leq q_{1-\alpha}\sqrt{\Sigma_n(w,w)} \text{ for all } w \in \mathcal{W}\right) - (1-\alpha)\right| \to 0. \qquad \square$$

Before proving Lemma B.1.5, we provide the following useful concentration inequality. This is essentially a corollary of the U-statistic concentration inequality given in Theorem 3.3 in Giné, Latała, and Zinn (2000).

Lemma B.3.17 (A concentration inequality)
Let X_{ij} be mutually independent for $1 \leq i < j \leq n$ taking values in a measurable space $\mathcal{X}$. Let h_1, h_2 be measurable functions from $\mathcal{X}$ to $\mathbb{R}$ satisfying the following for all i and j.

$$\begin{aligned}
\mathbb{E}\big[h_1(X_{ij})\big] &= 0, & \mathbb{E}\big[h_2(X_{ij})\big] &= 0, \\
\mathbb{E}\big[h_1(X_{ij})^2\big] &\leq \sigma^2, & \mathbb{E}\big[h_2(X_{ij})^2\big] &\leq \sigma^2, \\
\big|h_1(X_{ij})\big| &\leq M, & \big|h_2(X_{ij})\big| &\leq M.
\end{aligned}$$

Consider the sum

$$S_n = \sum_{1\leq i<j<r\leq n} h_1(X_{ij})h_2(X_{ir}).$$

Then S_n satisfies the concentration inequality

$$\mathbb{P}\big(|S_n| \geq t\big) \leq C \exp\left(-\frac{1}{C} \min\left\{\frac{t^2}{n^3\sigma^4}, \frac{t}{\sqrt{n^3\sigma^4}}, \frac{t^{2/3}}{(nM\sigma)^{2/3}}, \frac{t^{1/2}}{M}\right\}\right)$$

for some universal constant $C > 0$ and for all $t > 0$.

Proof (Lemma B.3.17)

We proceed in three main steps. Firstly, we write S_n as a second-order U-statistic where we use double indices instead of single indices. Then we use a decoupling result to introduce extra independence. Finally, a concentration result is applied to the decoupled U-statistic.

Part 1: writing S_n as a second-order U-statistic

Note that we can write S_n as the second-order U-statistic

$$S_n = \sum_{1 \leq i < j \leq n} \sum_{1 \leq q < r \leq n} h_{ijqr}(X_{ij}, X_{qr}),$$

where

$$h_{ijqr}(a, b) = h_1(a) h_2(b)\, \mathbb{I}\{j < r,\, q = i\}.$$

Although this may look like a fourth-order U-statistic, it is in fact second-order. This is due to independence of the variables X_{ij}, and by treating (i, j) as a single index.

Part 2: decoupling

By the decoupling result of Theorem 1 from de la Peña and Montgomery-Smith (1995), there exists a universal constant $C_1 > 0$ satisfying $\mathbb{P}\big(|S_n| \geq t\big) \leq C_1 \mathbb{P}\big(C_1 |\tilde{S}_n| \geq t\big)$, where $\tilde{S}_n = \sum_{1 \leq i < j \leq n} \sum_{1 \leq q < r \leq n} h_{ijqr}(X_{ij}, X'_{qr})$, with (X'_{ij}) an independent copy of (X_{ij}).

Part 3: U-statistic concentration

The U-statistic kernel $h_{ijqr}(X_{ij}, X'_{qr})$ is totally degenerate in that $\mathbb{E}[h_{ijqr}(X_{ij}, X'_{qr}) \mid X_{ij}] = \mathbb{E}[h_{ijqr}(X_{ij}, X'_{qr}) \mid X'_{qr}] = 0$. Define and bound the following quantities:

$$
\begin{aligned}
A &= \max_{ijqr} \|h_{ijqr}(X_{ij}, X'_{qr})\|_\infty \le M^2, \\
B &= \max\left\{ \left\| \sum_{1\le i<j\le n} \mathbb{E}\Big[h_{ijqr}(X_{ij}, X'_{qr})^2 \mid X_{ij}\Big] \right\|_\infty , \left\| \sum_{1\le q<r\le n} \mathbb{E}\Big[h_{ijqr}(X_{ij}, X'_{qr})^2 \mid X'_{qr}\Big] \right\|_\infty \right\}^{1/2} \\
&= \max\left\{ \left\| \sum_{1\le i<j\le n} h_1(X_{ij})^2 \mathbb{E}\big[h_2(X'_{qr})^2\big] \mathbb{I}\{j<r, q=i\} \right\|_\infty , \right. \\
&\qquad \left. \left\| \sum_{1\le q<r\le n} \mathbb{E}\big[h_1(X_{ij})^2\big] h_2(X'_{qr})^2 \mathbb{I}\{j<r, q=i\} \right\|_\infty \right\}^{1/2} \\
&\le \max\left\{n^2M^2\sigma^2, nM^2\sigma^2\right\}^{1/2} = nM\sigma, \\
C &= \left(\sum_{1\le i<j\le n} \sum_{1\le q<r\le n} \mathbb{E}\big[h_{ijqr}(X_{ij}, X'_{qr})^2\big] \right)^{1/2} = \left(\sum_{1\le i<j<r\le n} \mathbb{E}\big[h_1(X_{ij})^2 h_2(X'_{ir})^2\big] \right)^{1/2} \le \sqrt{n^3\sigma^4}, \\
D &= \sup_{f,g} \left\{ \sum_{1\le i<j\le n} \sum_{1\le q<r\le n} \mathbb{E}\big[h_{ijqr}(X_{ij}, X'_{qr}) f_{ij}(X_{ij}) g_{qr}(X'_{qr})\big] \ : \right. \\
&\qquad \left. \sum_{1\le i<j\le n} \mathbb{E}\big[f_{ij}(X_{ij})^2\big] \le 1, \ \sum_{1\le q<r\le n} \mathbb{E}\big[g_{qr}(X'_{qr})^2\big] \le 1 \right\} \\
&= \sup_{f,g} \left\{ \sum_{1\le i<j<r\le n} \mathbb{E}\big[h_1(X_{ij}) f_{ij}(X_{ij})\big] \mathbb{E}\big[h_2(X'_{ir}) g_{ir}(X'_{ir})\big] \ : \right. \\
&\qquad \left. \sum_{1\le i<j\le n} \mathbb{E}\big[f_{ij}(X_{ij})^2\big] \le 1, \ \sum_{1\le q<r\le n} \mathbb{E}\big[g_{qr}(X'_{qr})^2\big] \le 1 \right\} \\
&\le \sup_{f,g} \left\{ \sum_{1\le i<j<r\le n} \mathbb{E}\big[h_1(X_{ij})^2\big]^{1/2} \mathbb{E}\big[f_{ij}(X_{ij})^2\big]^{1/2} \mathbb{E}\big[h_2(X'_{ir})^2\big]^{1/2} \mathbb{E}\big[g_{ir}(X'_{ir})^2\big]^{1/2} \ : \right. \\
&\qquad \left. \sum_{1\le i<j\le n} \mathbb{E}\big[f_{ij}(X_{ij})^2\big] \le 1, \ \sum_{1\le q<r\le n} \mathbb{E}\big[g_{qr}(X'_{qr})^2\big] \le 1 \right\} \\
&\le \sigma^2 \sup_{f,g} \left\{ \sum_{1\le i<j\le n} \mathbb{E}\big[f_{ij}(X_{ij})^2\big]^{1/2} \sum_{1\le r\le n} \mathbb{E}\big[g_{ir}(X'_{ir})^2\big]^{1/2} \ : \right. \\
&\qquad \left. \sum_{1\le i<j\le n} \mathbb{E}\big[f_{ij}(X_{ij})^2\big] \le 1, \ \sum_{1\le q<r\le n} \mathbb{E}\big[g_{qr}(X'_{qr})^2\big] \le 1 \right\} \\
&\le \sigma^2 \sup_{f,g} \left\{ \left(n^2 \sum_{1\le i<j\le n} \mathbb{E}\big[f_{ij}(X_{ij})^2\big] \right)^{1/2} \left(n \sum_{1\le r\le n} \mathbb{E}\big[g_{ir}(X'_{ir})^2\big] \right)^{1/2} \ : \right. \\
&\qquad \left. \sum_{1\le i<j\le n} \mathbb{E}\big[f_{ij}(X_{ij})^2\big] \le 1, \ \sum_{1\le q<r\le n} \mathbb{E}\big[g_{qr}(X'_{qr})^2\big] \le 1 \right\} \le \sqrt{n^3\sigma^4}.
\end{aligned}
$$

By Theorem 3.3 in Giné et al. (2000), for some universal constant $C_2 > 0$ and for all $t > 0$,

$$\begin{aligned}\mathbb{P}\left(|\tilde{S}_n| \geq t\right) &\leq C_2 \exp\left(-\frac{1}{C_2}\min\left\{\frac{t^2}{C^2}, \frac{t}{D}, \frac{t^{2/3}}{B^{2/3}}, \frac{t^{1/2}}{A^{1/2}}\right\}\right) \\ &\leq C_2 \exp\left(-\frac{1}{C_2}\min\left\{\frac{t^2}{n^3\sigma^4}, \frac{t}{\sqrt{n^3\sigma^4}}, \frac{t^{2/3}}{(nM\sigma)^{2/3}}, \frac{t^{1/2}}{M}\right\}\right).\end{aligned}$$

Part 4: Conclusion

By the previous parts and absorbing constants into a new constant $C > 0$, we therefore have

$$\begin{aligned}\mathbb{P}\left(|S_n| \geq t\right) &\leq C_1 \mathbb{P}\left(C_1|\tilde{S}_n| \geq t\right) \\ &\leq C_1 C_2 \exp\left(-\frac{1}{C_2}\min\left\{\frac{t^2}{n^3\sigma^4 C_1^2}, \frac{t}{\sqrt{n^3\sigma^4 C_1}}, \frac{t^{2/3}}{(nM\sigma C_1)^{2/3}}, \frac{t^{1/2}}{MC_1^{1/2}}\right\}\right) \\ &\leq C \exp\left(-\frac{1}{C}\min\left\{\frac{t^2}{n^3\sigma^4}, \frac{t}{\sqrt{n^3\sigma^4}}, \frac{t^{2/3}}{(nM\sigma)^{2/3}}, \frac{t^{1/2}}{M}\right\}\right). \qquad \square\end{aligned}$$

Proof (Lemma B.1.5)

Throughout this proof we will write k_{ij} for $k_h(W_{ij}, w)$ and k'_{ij} for $k_h(W_{ij}, w')$, in the interest of brevity. Similarly, we write S_{ijr} to denote $S_{ijr}(w, w')$. The estimand and estimator are reproduced below for clarity.

$$\begin{aligned}\Sigma_n(w, w') &= \frac{2}{n(n-1)}\mathbb{E}[k_{ij}k'_{ij}] + \frac{4(n-2)}{n(n-1)}\mathbb{E}[k_{ij}k'_{ir}] - \frac{4n-6}{n(n-1)}\mathbb{E}[k_{ij}]\mathbb{E}[k'_{ij}] \\ \hat{\Sigma}_n(w, w') &= \frac{2}{n(n-1)}\frac{2}{n(n-1)}\sum_{i<j} k_{ij}k'_{ij} + \frac{4(n-2)}{n(n-1)}\frac{6}{n(n-1)(n-2)}\sum_{i<j<r} S_{ijr} \\ &\quad - \frac{4n-6}{n(n-1)}\hat{f}_W(w)\hat{f}_W(w'),\end{aligned}$$

where $S_{ijr} = \frac{1}{6}\big(k_{ij}k'_{ir} + k_{ij}k'_{jr} + k_{ir}k'_{ij} + k_{ir}k'_{jr} + k_{jr}k'_{ij} + k_{jr}k'_{ir}\big)$. We will prove uniform consistency of each of the three terms separately.

Part 1: uniform consistency of the $\hat{f}_W(w)\hat{f}_W(w')$ term

By boundedness of f_W and Theorem 3.3.1, $\hat{f}_W$ is uniformly bounded in probability. Noting that $\mathbb{E}[\hat{f}_W(w)] = \mathbb{E}[k_{ij}]$ and by Lemma 3.2.4,

$$\begin{aligned}
&\sup_{w,w'\in\mathcal{W}} \left| \frac{\hat{f}_W(w)\hat{f}_W(w') - \mathbb{E}\big[k_{ij}\big]\mathbb{E}\big[k_{ij'}\big]}{\sqrt{\Sigma_n(w,w)+\Sigma_n(w',w')}} \right| = \sup_{w,w'\in\mathcal{W}} \left| \frac{\hat{f}_W(w)\hat{f}_W(w') - \mathbb{E}\big[\hat{f}_W(w)\big]\mathbb{E}\big[\hat{f}_W(w')\big]}{\sqrt{\Sigma_n(w,w)+\Sigma_n(w',w')}} \right| \\
&\quad \leq \sup_{w,w'\in\mathcal{W}} \left| \frac{\hat{f}_W(w) - \mathbb{E}\big[\hat{f}_W(w)\big]}{\sqrt{\Sigma_n(w,w)}} \hat{f}_W(w') + \frac{\hat{f}_W(w') - \mathbb{E}\big[\hat{f}_W(w')\big]}{\sqrt{\Sigma_n(w',w')}} \mathbb{E}\big[\hat{f}_W(w)\big] \right| \\
&\quad \lesssim_{\mathbb{P}} \sup_{w\in\mathcal{W}} \left| \frac{\hat{f}_W(w) - \mathbb{E}\big[\hat{f}_W(w)\big]}{\sqrt{\Sigma_n(w,w)}} \right| \\
&\quad \lesssim_{\mathbb{P}} \sup_{w\in\mathcal{W}} \left| \frac{L_n(w)}{\sqrt{\Sigma_n(w,w)}} \right| + \sqrt{n^2h} \sup_{w\in\mathcal{W}} |Q_n(w)| + \sqrt{n^2h} \sup_{w\in\mathcal{W}} |E_n(w)| \\
&\quad \lesssim_{\mathbb{P}} \sup_{w\in\mathcal{W}} \left| \frac{L_n(w)}{\sqrt{\Sigma_n(w,w)}} \right| + \sqrt{n^2h}\frac{1}{n} + \sqrt{n^2h}\sqrt{\frac{\log n}{n^2h}} \lesssim_{\mathbb{P}} \sup_{w\in\mathcal{W}} \left| \frac{L_n(w)}{\sqrt{\Sigma_n(w,w)}} \right| + \sqrt{\log n}.
\end{aligned}$$

Now consider the function class

$$\mathcal{F} = \left\{ a \mapsto \frac{\mathbb{E}\big[k_h(W_{ij},w) \mid A_i = a\big] - \mathbb{E}\big[k_h(W_{ij},w)\big]}{\sqrt{n\Sigma_n(w,w)}} : w \in \mathcal{W} \right\},$$

noting that

$$\frac{L_n(w)}{\Sigma_n(w,w)^{1/2}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_w(A_i)$$

is an empirical process evaluated at $g_w \in \mathcal{F}$. By the lower bound on $\Sigma_n(w,w)$ from Lemma 3.2.4 with $\inf_{\mathcal{W}} f_W(w) > 0$ and since $nh \gtrsim \log n$, the class $\mathcal{F}$ has a constant envelope function given by $F(a) \lesssim \sqrt{nh}$. Clearly, $M = \sup_a F(a) \lesssim \sqrt{nh}$. Also by definition of Σ_n and orthogonality of L_n, Q_n, and E_n, we have $\sup_{f\in\mathcal{F}} \mathbb{E}[f(A_i)^2] \leq \sigma^2 = 1$. To verify a VC-type condition on $\mathcal{F}$ we need to establish the regularity of the process. By Lipschitz properties of L_n and Σ_n

derived in the proofs of Lemma 3.2.3 and Theorem 3.4.2 respectively, we have

$$\begin{aligned}\left|\frac{L_n(w)}{\sqrt{\Sigma_n(w,w)}}-\frac{L_n(w')}{\sqrt{\Sigma_n(w',w')}}\right| &\lesssim \frac{\left|L_n(w)-L_n(w')\right|}{\sqrt{\Sigma_n(w,w)}}+\left|L_n(w')\right|\left|\frac{1}{\sqrt{\Sigma_n(w,w)}}-\frac{1}{\sqrt{\Sigma_n(w',w')}}\right| \\ &\lesssim \sqrt{n^2h}|w-w'|+\left|\frac{\Sigma_n(w,w)-\Sigma_n(w',w')}{\Sigma_n(w,w)\sqrt{\Sigma_n(w',w')}}\right| \\ &\lesssim \sqrt{n^2h}|w-w'|+(n^2h)^{3/2}\left|\Sigma_n(w,w)-\Sigma_n(w',w')\right| \\ &\lesssim \sqrt{n^2h}|w-w'|+(n^2h)^{3/2}n^{-1}h^{-3}|w-w'|\lesssim n^4|w-w'|,\end{aligned}$$

uniformly over $w, w' \in \mathcal{W}$. By compactness of $\mathcal{W}$ we have the covering number bound $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \lesssim N(\mathcal{W}, |\cdot|, n^{-4}\varepsilon) \lesssim n^4\varepsilon^{-1}$. Thus by Lemma B.2.2,

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}}\left|\frac{L_n(w)}{\sqrt{\Sigma_n(w,w)}}\right|\right] \lesssim \sqrt{\log n}+\frac{\sqrt{nh}\log n}{\sqrt{n}} \lesssim \sqrt{\log n}.$$

Therefore

$$\sup_{w,w'\in\mathcal{W}}\left|\frac{\hat{f}_W(w)\hat{f}_W(w')-\mathbb{E}\big[k_{ij}\big]\mathbb{E}\big[k_{ij'}\big]}{\sqrt{\Sigma_n(w,w)+\Sigma_n(w',w')}}\right| \lesssim_{\mathbb{P}} \sqrt{\log n}.$$

Part 2: decomposition of the S_{ijr} term

We first decompose the S_{ijr} term into two parts, and obtain a pointwise concentration result for each. This is extended to a uniform concentration result by considering the regularity of the covariance estimator process. Note that $\mathbb{E}[S_{ijr}] = \mathbb{E}[k_{ij}k'_{ir}]$, and hence

$$\begin{aligned}&\frac{6}{n(n-1)(n-2)}\sum_{i<j<r}\left(S_{ijr}-\mathbb{E}[k_{ij}k'_{ir}]\right) \\ &\quad=\frac{6}{n(n-1)(n-2)}\sum_{i<j<r}S^{(1)}_{ijr}+\frac{6}{n(n-1)(n-2)}\sum_{i<j<r}S^{(2)}_{ijr},\end{aligned}$$

where $S^{(1)}_{ijr} = S_{ijr} - \mathbb{E}[S_{ijr} \mid \mathbf{A}_n]$ and $S^{(2)}_{ijr} = \mathbb{E}[S_{ijr} \mid \mathbf{A}_n] - \mathbb{E}[S_{ijr}]$.

Part 3: pointwise concentration of the $S^{(1)}_{ijr}$ term

By symmetry in i, j, and r it is sufficient to consider only the first summand in the definition of S_{ijr}. By conditional independence properties, we have the decomposition

$$\begin{aligned}
&\frac{6}{n(n-1)(n-2)} \sum_{i<j<r} \Big(k_{ij}k'_{ir} - \mathbb{E}[k_{ij}k'_{ir} \mid \mathbf{A}_n] \Big) \\
&\quad = \frac{6}{n(n-1)(n-2)} \sum_{i<j<r} \Big(k_{ij}k'_{ir} - \mathbb{E}[k_{ij} \mid \mathbf{A}_n]\mathbb{E}[k'_{ir} \mid \mathbf{A}_n] \Big) \\
&\quad = \frac{6}{n(n-1)(n-2)} \sum_{i<j<r} \Big(\big(k_{ij} - \mathbb{E}[k_{ij} \mid \mathbf{A}_n]\big)\big(k'_{ir} - \mathbb{E}[k'_{ir} \mid \mathbf{A}_n]\big) \\
&\qquad + \big(k_{ij} - \mathbb{E}[k_{ij} \mid \mathbf{A}_n]\big)\mathbb{E}[k'_{ir} \mid \mathbf{A}_n] + \big(k'_{ir} - \mathbb{E}[k'_{ir} \mid \mathbf{A}_n]\big)\mathbb{E}[k_{ij} \mid \mathbf{A}_n] \Big) \\
&\quad = \frac{6}{n(n-1)(n-2)} \sum_{i<j<r} \Big(k_{ij} - \mathbb{E}[k_{ij} \mid \mathbf{A}_n] \Big)\Big(k'_{ir} - \mathbb{E}[k'_{ir} \mid \mathbf{A}_n] \Big) \qquad \text{(B.7)} \\
&\qquad + \frac{2}{(n-1)(n-2)} \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \Big(k_{ij} - \mathbb{E}[k_{ij} \mid \mathbf{A}_n] \Big) \cdot \frac{3}{n} \sum_{r=j+1}^{n} \mathbb{E}[k'_{ir} \mid \mathbf{A}_n] \qquad \text{(B.8)} \\
&\qquad + \frac{2}{(n-1)(n-2)} \sum_{i=1}^{n-2} \sum_{r=i+2}^{n} \Big(k'_{ir} - \mathbb{E}[k'_{ir} \mid \mathbf{A}_n] \Big) \cdot \frac{3}{n} \sum_{j=i+1}^{r-1} \mathbb{E}[k_{ij} \mid \mathbf{A}_n]. \qquad \text{(B.9)}
\end{aligned}$$

For the term in (B.7), note that conditional on $\mathbf{A}_n$, we have that $k_{ij} - \mathbb{E}[k_{ij} \mid \mathbf{A}_n]$ are conditionally mean-zero and conditionally independent, as the only randomness is from $\mathbf{V}_n$. Also $\mathrm{Var}[k_{ij} \mid \mathbf{A}_n] \lesssim \sigma^2 := 1/h$ and $|k_{ij}| \lesssim M := 1/h$ uniformly. The same is true for k'_{ij}. Thus by Lemma B.3.17 for some universal constant $C_1 > 0$:

$$\begin{aligned}
&\mathbb{P}\left(\left| \sum_{i<j<r} \Big(k_{ij} - \mathbb{E}[k_{ij} \mid \mathbf{A}_n] \Big)\Big(k'_{ir} - \mathbb{E}[k'_{ir} \mid \mathbf{A}_n] \Big) \right| > t \;\middle|\; \mathbf{A}_n \right) \\
&\quad \le C_1 \exp\left(-\frac{1}{C_1} \min\left\{ \frac{t^2}{n^3\sigma^4}, \frac{t}{\sqrt{n^3\sigma^4}}, \frac{t^{2/3}}{(nM\sigma)^{2/3}}, \frac{t^{1/2}}{M} \right\} \right) \\
&\quad \le C_1 \exp\left(-\frac{1}{C_1} \min\left\{ t^2h^2, \frac{th}{\sqrt{n^3}}, \frac{t^{2/3}h}{n^{2/3}}, t^{1/2}h \right\} \right),
\end{aligned}$$

and therefore with $t \geq 1$ and since $nh \gtrsim \log n$, introducing and adjusting a new constant C_2 where necessary,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{6}{n(n-1)(n-2)}\sum_{i<j<r}\Big(k_{ij}-\mathbb{E}[k_{ij}\mid \mathbf{A}_n]\Big)\Big(k'_{ir}-\mathbb{E}[k'_{ir}\mid \mathbf{A}_n]\Big)\right| > t\frac{\log n}{\sqrt{n^3h^2}} \;\Big|\; \mathbf{A}_n\right)\\
&\quad\leq \mathbb{P}\left(\left|\sum_{i<j<r}\Big(k_{ij}-\mathbb{E}[k_{ij}\mid \mathbf{A}_n]\Big)\Big(k'_{ir}-\mathbb{E}[k'_{ir}\mid \mathbf{A}_n]\Big)\right| > tn^{3/2}h^{-1}\log n/24 \;\Big|\; \mathbf{A}_n\right)\\
&\quad\leq C_2\exp\left(-\frac{1}{C_2}\min\Big\{(t\log n)^2, t\log n, (t\log n)^{2/3}(nh)^{1/3}, (tnh\log n)^{1/2}n^{1/4}\Big\}\right)\\
&\quad\leq C_2\exp\left(-\frac{1}{C_2}\min\Big\{t\log n, t\log n, t^{2/3}\log n, t^{1/2}n^{1/4}\log n\Big\}\right)\\
&\quad= C_2\exp\left(-\frac{t^{2/3}\log n}{C_2}\right) = C_2 n^{-t^{2/3}/C_2}.
\end{aligned}$$

Now for the term in (B.8), note that $\frac{3}{n}\sum_{r=j+1}^{n}\mathbb{E}[k'_{ir}\mid \mathbf{A}_n]$ is $\mathbf{A}_n$-measurable and bounded uniformly in i, j. Also, using the previously established conditional variance and almost sure bounds on k_{ij}, Bernstein's inequality (Lemma B.3.1) applied conditionally gives for some constant $C_3 > 0$

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{2}{(n-1)(n-2)}\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\Big(k_{ij}-\mathbb{E}[k_{ij}\mid \mathbf{A}_n]\Big)\cdot\frac{3}{n}\sum_{r=j+1}^{n}\mathbb{E}[k'_{ir}\mid \mathbf{A}_n]\right| > t\sqrt{\frac{\log n}{n^2h}} \;\Big|\; \mathbf{A}_n\right)\\
&\quad\leq 2\exp\left(-\frac{t^2n^2\log n/(n^2h)}{C_3/(2h)+C_3t\sqrt{\log n/(n^2h)}/(2h)}\right)\\
&\quad= 2\exp\left(-\frac{t^2\log n}{C_3/2+C_3t\sqrt{\log n/(n^2h)}/2}\right) \leq 2\exp\left(-\frac{t^2\log n}{C_3}\right) = 2n^{-t^2/C_3}.
\end{aligned}$$

The term in (B.9) is controlled in exactly the same way. Putting these together, noting the symmetry in i, j, r and taking a marginal expectation, we obtain the unconditional pointwise concentration inequality

$$\mathbb{P}\left(\left|\frac{6}{n(n-1)(n-2)}\sum_{i<j<r}S^{(1)}_{ijr}\right| > t\frac{\log n}{\sqrt{n^3h^2}}+t\sqrt{\frac{\log n}{n^2h}}\right) \leq C_2n^{-t^{2/3}/C_2}+4n^{-t^2/(4C_3)}.$$

Multiplying by $\big(\Sigma_n(w,w)+\Sigma_n(w',w')\big)^{-1/2} \lesssim \sqrt{n^2h}$ gives (adjusting constants if necessary)

$$\mathbb{P}\left(\left|\frac{6}{n(n-1)(n-2)}\sum_{i<j<r}\frac{S^{(1)}_{ijr}}{\sqrt{\Sigma_n(w,w)+\Sigma_n(w',w')}}\right| > t\frac{\log n}{\sqrt{nh}} + t\sqrt{\log n}\right)$$
$$\leq C_2 n^{-t^{2/3}/C_2} + 4n^{-t^2/(4C_3)}.$$

Part 4: pointwise concentration of the $S^{(2)}_{ijr}$ term

We apply the U-statistic concentration inequality from Lemma B.3.8. Note that the terms $\mathbb{E}[S_{ijr} \mid \mathbf{A}_n]$ are permutation-symmetric functions of the random variables A_i, A_j, and A_r only, making $S^{(2)}_{ijr}$ the summands of a (non-degenerate) mean-zero third-order U-statistic. While we could apply a third-order Hoeffding decomposition here to achieve degeneracy, it is unnecessary as Lemma B.3.8 is general enough to deal with the non-degenerate case directly. The quantity of interest here is

$$\frac{6}{n(n-1)(n-2)}\sum_{i<j<r} S^{(2)}_{ijr} = \frac{6}{n(n-1)(n-2)}\sum_{i<j<r}\Big(\mathbb{E}[S_{ijr} \mid \mathbf{A}_n] - \mathbb{E}[S_{ijr}]\Big).$$

Note that by conditional independence,

$$\big|\mathbb{E}\big[k_{ij}k_{ir} \mid \mathbf{A}_n\big]\big| = \big|\mathbb{E}\big[k_{ij} \mid \mathbf{A}_n\big]\mathbb{E}\big[k_{ir} \mid \mathbf{A}_n\big]\big| \lesssim 1,$$

and similarly for the other summands in S_{ijr}, giving the almost sure bound $|S^{(2)}_{ijr}| \lesssim 1$. Also,

$$\begin{aligned}\operatorname{Var}\left[\mathbb{E}[k_{ij} \mid A_i]\mathbb{E}[k'_{ir} \mid A_i]\right] &\lesssim \operatorname{Var}\left[\mathbb{E}[k_{ij} \mid A_i]\right] + \operatorname{Var}\left[\mathbb{E}[k'_{ir} \mid A_i]\right] \\ &\lesssim n\operatorname{Var}[L_n(w)] + n\operatorname{Var}[L_n(w')] \\ &\lesssim n\Sigma_n(w,w) + n\Sigma_n(w',w')\end{aligned}$$

and similarly for the other summands in S_{ijr}, giving the conditional variance bound

$$\mathbb{E}[\mathbb{E}[S^{(2)}_{ijr} \mid A_i]^2] \lesssim n\Sigma_n(w,w) + n\Sigma_n(w',w').$$

So Lemma B.3.8 and Lemma 3.2.4 give the pointwise concentration inequality

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{6}{n(n-1)(n-2)}\sum_{i<j<r} S^{(2)}_{ijr}\right| > t\sqrt{\log n}\sqrt{\Sigma_n(w,w)+\Sigma_n(w',w')}\right) \\
&\quad\le 4\exp\left(-\frac{nt^2(\Sigma_n(w,w)+\Sigma_n(w',w'))\log n}{C_4(n\Sigma_n(w,w)+n\Sigma_n(w',w'))+C_4t\sqrt{\Sigma_n(w,w)+\Sigma_n(w',w')}\sqrt{\log n}}\right) \\
&\quad\le 4\exp\left(-\frac{t^2\log n}{C_4+C_4t(\Sigma_n(w,w)+\Sigma_n(w',w'))^{-1/2}\sqrt{\log n}/n}\right) \\
&\quad\le 4\exp\left(-\frac{t^2\log n}{C_4+C_4t\sqrt{h}}\right) \le 4n^{-t^2/C_4}
\end{aligned}$$

for some universal constant $C_4 > 0$ (which may change from line to line), since the order of this U-statistic is fixed at three.

Part 5: concentration of the S_{ijr} term on a mesh

Pick $\delta_n \to 0$ with $\log 1/\delta_n \lesssim \log n$. Let $\mathcal{W}_\delta$ be a δ_n-covering of $\mathcal{W}$ with cardinality $O(1/\delta_n)$. Then $\mathcal{W}_\delta \times \mathcal{W}_\delta$ is a $2\delta_n$-covering of $\mathcal{W}\times\mathcal{W}$ with cardinality $O(1/\delta_n^2)$, under the Manhattan metric $d\big((w_1,w_1'),(w_2,w_2')\big) = |w_1-w_2|+|w_1'-w_2'|$. By the previous parts, we have that for fixed w and w':

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{6}{n(n-1)(n-2)}\sum_{i<j<r}\frac{S_{ijr}(w,w')-\mathbb{E}[S_{ijr}(w,w')]}{\sqrt{\Sigma_n(w,w)+\Sigma_n(w',w')}}\right| > t\frac{\log n}{\sqrt{nh}}+2t\sqrt{\log n}\right) \\
&\quad\le C_2n^{-t^{2/3}/C_2}+4n^{-t^2/(4C_3)}+4n^{-t^2/C_4}.
\end{aligned}$$

Taking a union bound over $\mathcal{W}_\delta\times\mathcal{W}_\delta$, noting that $nh \gtrsim \log n$ and adjusting constants gives

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w,w'\in\mathcal{W}_\delta}\left|\frac{6}{n(n-1)(n-2)}\sum_{i<j<r}\frac{S_{ijr}(w,w')-\mathbb{E}[S_{ijr}(w,w')]}{\sqrt{\Sigma_n(w,w)+\Sigma_n(w',w')}}\right| > t\sqrt{\log n}\right) \\
&\quad\lesssim \delta_n^{-2}\Big(C_2n^{-t^{2/3}/C_2}+4n^{-t^2/(4C_3)}+4n^{-t^2/C_4}\Big) \lesssim \delta_n^{-2}n^{-t^{2/3}/C_5},
\end{aligned}$$

for some constant $C_5 > 0$.

Part 6: regularity of the S_{ijr} term

Next we bound the fluctuations in $S_{ijr}(w, w')$. Writing $k_{ij}(w)$ for $k_h(W_{ij}, w)$, note that

$$\begin{aligned}\big|k_{ij}(w_1)k_{ir}(w_1') - k_{ij}(w_2)k_{ir}(w_2')\big| &\lesssim \frac{1}{h}\big|k_{ij}(w_1) - k_{ij}(w_2)\big| + \frac{1}{h}\big|k_{ir}(w_1') - k_{ir}(w_2')\big| \\ &\lesssim \frac{1}{h^3}\Big(|w_1 - w_2| + |w_1' - w_2'|\Big),\end{aligned}$$

by the Lipschitz property of the kernel, and similarly for the other summands in S_{ijr}. Therefore,

$$\sup_{|w_1-w_2|\le\delta_n}\sup_{|w_1'-w_2'|\le\delta_n}\big|S_{ijr}(w_1, w_1') - S_{ijr}(w_2, w_2')\big| \lesssim \delta_n h^{-3}.$$

Also as noted in the proof of Theorem 3.4.2,

$$\sup_{|w_1-w_2|\le\delta_n}\sup_{|w_1'-w_2'|\le\delta_n}\big|\Sigma_n(w_1, w_1') - \Sigma_n(w_2, w_2')\big| \lesssim \delta_n n^{-1}h^{-3}.$$

Therefore, since $\sqrt{\Sigma_n(w,w)} \gtrsim \sqrt{n^2h}$ and $|S_{ijr}| \lesssim h^{-2}$, using $\frac{a}{\sqrt{b}} - \frac{c}{\sqrt{d}} = \frac{a-c}{\sqrt{b}} + c\frac{d-b}{\sqrt{bd}\sqrt{b+d}}$,

$$\begin{aligned}&\sup_{|w_1-w_2|\le\delta_n}\sup_{|w_1'-w_2'|\le\delta_n}\left|\frac{S_{ijr}(w_1, w_1')}{\sqrt{\Sigma_n(w_1,w_1)+\Sigma_n(w_1',w_1')}} - \frac{S_{ijr}(w_2, w_2')}{\sqrt{\Sigma_n(w_2,w_2)+\Sigma_n(w_2',w_2')}}\right| \\ &\lesssim \delta_n h^{-3}\sqrt{n^2h} + h^{-2}\delta_n n^{-1}h^{-3}(n^2h)^{3/2} \lesssim \delta_n n h^{-5/2} + \delta_n n^2 h^{-7/2} \lesssim \delta_n n^6,\end{aligned}$$

where in the last line we use that $1/h \lesssim n$.

Part 7: uniform concentration of the S_{ijr} term

By setting $\delta_n = n^{-6}\sqrt{\log n}$, the fluctuations can be at most $\sqrt{\log n}$, so we have for $t \ge 1$

$$\begin{aligned}&\mathbb{P}\left(\sup_{w,w'\in\mathcal{W}}\left|\frac{6}{n(n-1)(n-2)}\sum_{i<j<r}\frac{S_{ijr}(w,w') - \mathbb{E}[S_{ijr}(w,w')]}{\sqrt{\Sigma_n(w,w)+\Sigma_n(w',w')}}\right| > 2t\sqrt{\log n}\right) \\ &\lesssim \delta_n^{-2}n^{-t^{2/3}/C_5} \lesssim n^{12-t^{2/3}/C_5}.\end{aligned}$$

This converges to zero for any sufficiently large t, so

$$\sup_{w,w' \in \mathcal{W}} \left| \frac{6}{n(n-1)(n-2)} \sum_{i<j<r} \frac{S_{ijr}(w,w') - \mathbb{E}[S_{ijr}(w,w')]}{\sqrt{\Sigma_n(w,w) + \Sigma_n(w',w')}} \right| \lesssim_{\mathbb{P}} \sqrt{\log n}.$$

Part 8: decomposition of the $k_{ij}k'_{ij}$ term

We move on to the final term in the covariance estimator. We have the decomposition

$$\frac{2}{n(n-1)} \sum_{i<j} \Big(k_{ij}k'_{ij} - \mathbb{E}\big[k_{ij}k'_{ij}\big] \Big) = \frac{2}{n(n-1)} \sum_{i<j} S_{ij}^{(1)} + \frac{2}{n(n-1)} \sum_{i<j} S_{ij}^{(2)},$$

where

$$S_{ij}^{(1)} = k_{ij}k'_{ij} - \mathbb{E}\big[k_{ij}k'_{ij} \mid \mathbf{A}_n\big], \qquad S_{ij}^{(2)} = \mathbb{E}\big[k_{ij}k'_{ij} \mid \mathbf{A}_n\big] - \mathbb{E}\big[k_{ij}k'_{ij}\big].$$

Part 9: pointwise concentration of the $S_{ij}^{(1)}$ term

Conditioning on $\mathbf{A}_n$, the variables $S_{ij}^{(1)}$ are conditionally independent and conditionally mean-zero. They further satisfy $|S_{ij}^{(1)}| \lesssim h^{-2}$ and the conditional variance bound $\mathbb{E}\big[\big(S_{ij}^{(1)}\big)^2 \mid \mathbf{A}_n\big] \lesssim h^{-3}$. Therefore applying Bernstein's inequality (Lemma B.3.1) conditional on $\mathbf{A}_n$, we obtain the pointwise in w, w' concentration inequality

$$\begin{aligned}
&\mathbb{P}\left(\left| \frac{2}{n(n-1)} \sum_{i<j} S_{ij}^{(1)} \right| > t\sqrt{\frac{\log n}{n^2 h^3}} \;\Big|\; \mathbf{A}_n \right) \\
&\quad \leq 2\exp\left(-\frac{t^2 n^2 \log n/(n^2h^3)}{C_6 h^{-3}/2 + C_6 t h^{-2}\sqrt{\log n/(n^2h^3)}/2} \right) \\
&\quad \leq 2\exp\left(-\frac{t^2 \log n}{C_6/2 + C_6 t\sqrt{\log n/(n^2 h)}/2} \right) \leq 2\exp\left(-\frac{t^2 \log n}{C_6} \right) = 2n^{-t^2/C_6},
\end{aligned}$$

where C_6 is a universal positive constant.

Part 10: pointwise concentration of the $S_{ij}^{(2)}$ term

We apply the U-statistic concentration inequality from Lemma B.3.8. Note that $S_{ij}^{(2)}$ are permutation-symmetric functions of the random variables A_i and A_j only, making them the summands of a (non-degenerate) mean-zero second-order U-statistic. Note that $\big|S_{ij}^{(2)}\big| \lesssim h^{-1}$

and so trivially $\mathbb{E}\big[\mathbb{E}[S_{ij}^{(2)} \mid A_i]^2\big] \lesssim h^{-2}$. Thus by Lemma B.3.8, since the order of this U-statistic is fixed at two, for some universal positive constant C_7 we have

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{2}{n(n-1)}\sum_{i<j} S_{ij}^{(2)}\right| > t\sqrt{\frac{\log n}{nh^2}}\right) &\le 2\exp\left(-\frac{t^2 n\log n/(nh^2)}{C_7 h^{-2}/2 + C_7 t h^{-1}\sqrt{\log n/(nh^2)}/2}\right)\\
&\le 2\exp\left(-\frac{t^2\log n}{C_7/2 + C_7 t\sqrt{\log n/n}/2}\right)\\
&\le 2\exp\left(-\frac{t^2\log n}{C_7}\right) = 2n^{-t^2/C_7}.
\end{aligned}$$

Part 11: concentration of the $k_{ij}k'_{ij}$ term on a mesh

As before, use a union bound on the mesh $\mathcal{W}_\delta \times \mathcal{W}_\delta$.

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w,w'\in\mathcal{W}_\delta}\left|\frac{2}{n(n-1)}\sum_{i<j}\Big(k_{ij}k'_{ij} - \mathbb{E}\big[k_{ij}k'_{ij}\big]\Big)\right| > t\sqrt{\frac{\log n}{n^2h^3}} + t\sqrt{\frac{\log n}{nh^2}}\right)\\
&\le \mathbb{P}\left(\sup_{w,w'\in\mathcal{W}_\delta}\left|\frac{2}{n(n-1)}\sum_{i<j} S_{ij}^{(1)}\right| > t\sqrt{\frac{\log n}{n^2h^3}}\right) + \mathbb{P}\left(\sup_{w,w'\in\mathcal{W}_\delta}\left|\frac{2}{n(n-1)}\sum_{i<j} S_{ij}^{(2)}\right| > t\sqrt{\frac{\log n}{nh^2}}\right)\\
&\lesssim \delta_n^{-2}n^{-t^2/C_6} + \delta_n^{-2}n^{-t^2/C_7}.
\end{aligned}$$

Part 12: regularity of the $k_{ij}k'_{ij}$ term

As for the S_{ijr} term, $\big|k_{ij}(w_1)k_{ij}(w'_1) - k_{ij}(w_2)k_{ij}(w'_2)\big| \lesssim \frac{1}{h^3}\Big(|w_1 - w_2| + |w'_1 - w'_2|\Big)$.

Part 13: uniform concentration of the $k_{ij}k'_{ij}$ term

Setting $\delta_n = h^3\sqrt{\log n/(nh^2)}$, the fluctuations are at most $\sqrt{\log n/(nh^2)}$, so for $t \ge 1$

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w,w'\in\mathcal{W}}\left|\frac{2}{n(n-1)}\sum_{i<j}\Big(k_{ij}k'_{ij} - \mathbb{E}\big[k_{ij}k'_{ij}\big]\Big)\right| > t\sqrt{\frac{\log n}{n^2h^3}} + 2t\sqrt{\frac{\log n}{nh^2}}\right)\\
&\quad\le \mathbb{P}\left(\sup_{w,w'\in\mathcal{W}_\delta}\left|\frac{2}{n(n-1)}\sum_{i<j}\Big(k_{ij}k'_{ij} - \mathbb{E}\big[k_{ij}k'_{ij}\big]\Big)\right| > t\sqrt{\frac{\log n}{n^2h^3}} + t\sqrt{\frac{\log n}{nh^2}}\right)\\
&\qquad + \mathbb{P}\left(\sup_{|w_1-w_2|\le\delta_n}\sup_{|w'_1-w'_2|\le\delta_n}\big|k_{ij}(w_1)k_{ij}(w'_1) - k_{ij}(w_2)k_{ij}(w'_2)\big| > t\sqrt{\frac{\log n}{nh^2}}\right)\\
&\quad\lesssim \delta_n^{-2}n^{-t^2/C_6} + \delta_n^{-2}n^{-t^2/C_7} \lesssim n^{1-t^2/C_6}h^{-4} + n^{1-t^2/C_7}h^{-4} \lesssim n^{5-t^2/C_8},
\end{aligned}$$

where $C_8 > 0$ is a constant and in the last line we use $1/h \lesssim n$. This converges to zero for any sufficiently large t, so by Lemma 3.2.4 we have

$$\sup_{w,w' \in \mathcal{W}} \left| \frac{2}{n(n-1)} \sum_{i<j} \frac{k_{ij}k'_{ij} - \mathbb{E}\big[k_{ij}k'_{ij}\big]}{\sqrt{\Sigma_n(w,w) + \Sigma_n(w',w')}} \right| \lesssim_{\mathbb{P}} \left(\sqrt{\frac{\log n}{n^2 h^3}} + \sqrt{\frac{\log n}{nh^2}} \right) \sqrt{n^2 h} \lesssim_{\mathbb{P}} \sqrt{\frac{n \log n}{h}}.$$

Part 14: conclusion

By the uniform bounds derived in the previous parts, and with $nh \gtrsim \log n$, we conclude that

$$\begin{aligned}
\sup_{w,w' \in \mathcal{W}} \left| \frac{\hat{\Sigma}_n(w,w') - \Sigma_n(w,w')}{\sqrt{\Sigma_n(w,w) + \Sigma_n(w',w')}} \right| &\leq \frac{2}{n(n-1)} \sup_{w,w' \in \mathcal{W}} \left| \frac{2}{n(n-1)} \sum_{i<j} \frac{k_{ij}k'_{ij} - \mathbb{E}\big[k_{ij}k'_{ij}\big]}{\sqrt{\Sigma_n(w,w) + \Sigma_n(w',w')}} \right| \\
&\quad + \frac{4(n-2)}{n(n-1)} \sup_{w,w' \in \mathcal{W}} \left| \frac{6}{n(n-1)(n-2)} \sum_{i<j<r} \frac{S_{ijr} - \mathbb{E}\big[k_{ij}k'_{ir}\big]}{\sqrt{\Sigma_n(w,w) + \Sigma_n(w',w')}} \right| \\
&\quad + \frac{4n-6}{n(n-1)} \sup_{w,w' \in \mathcal{W}} \left| \frac{\hat{f}_W(w)\hat{f}_W(w') - \mathbb{E}[k_{ij}]\mathbb{E}[k'_{ij}]}{\sqrt{\Sigma_n(w,w) + \Sigma_n(w',w')}} \right| \\
&\lesssim_{\mathbb{P}} \sqrt{\frac{\log n}{n^3 h}} + \frac{\sqrt{\log n}}{n} + \frac{\sqrt{\log n}}{n} \lesssim_{\mathbb{P}} \frac{\sqrt{\log n}}{n}.
\end{aligned}$$

□

Proof (Lemma B.1.6)

Write k_{ij} for $k_h(W_{ij}, w)$ if $i < j$ and $k_h(W_{ji}, w)$ if $j < i$, and use a prime to denote evaluation at w'. Thus we write $S_i(w) = \frac{1}{n-1} \sum_{j \neq i} k_{ij}$. Let $\sum_{i \neq j \neq r}$ indicate all indices are distinct.

$$\begin{aligned}
\frac{4}{n^2} \sum_{i=1}^{n} S_i(w) S_i(w') &= \frac{4}{n^2} \sum_{i=1}^{n} \frac{1}{n-1} \sum_{j \neq i} k_{ij} \frac{1}{n-1} \sum_{r \neq i} k'_{ir} = \frac{4}{n^2(n-1)^2} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{r \neq i} k_{ij}k'_{ir} \\
&= \frac{4}{n^2(n-1)^2} \sum_{i=1}^{n} \sum_{j \neq i} \left(\sum_{r \neq i, r \neq j} k_{ij}k'_{ir} + k_{ij}k'_{ij} \right) \\
&= \frac{4}{n^2(n-1)^2} \sum_{i \neq j \neq r} k_{ij}k'_{ir} + \frac{4}{n^2(n-1)^2} \sum_{i \neq j} k_{ij}k'_{ij} \\
&= \frac{24}{n^2(n-1)^2} \sum_{i<j<r} S_{ijr}(w,w') + \frac{8}{n^2(n-1)^2} \sum_{i<j} k_{ij}k'_{ij} \\
&= \hat{\Sigma}_n(w,w') + \frac{4}{n^2(n-1)^2} \sum_{i<j} k_{ij}k'_{ir} + \frac{4n-6}{n(n-1)} \hat{f}\hat{f}'.
\end{aligned}$$

□

Proof (Lemma B.1.7)

Firstly, we prove that the true covariance function Σ_n is feasible for the optimization problem (B.1) in the sense that it satisfies the constraints. As a covariance function, it is symmetric and positive semi-definite. The Lipschitz constraint is established in the proof of Theorem 3.4.2:

$$\left|\Sigma_n(w,w') - \Sigma_n(w,w'')\right| \leq \frac{4}{nh^3} C_{\mathrm{k}} C_{\mathrm{L}} |w' - w''|$$

for all $w, w', w'' \in \mathcal{W}$. Denote the (random) objective function in (B.1) by

$$\mathrm{obj}(M) = \sup_{w,w' \in \mathcal{W}} \left| \frac{M(w,w') - \hat{\Sigma}_n(w,w')}{\sqrt{\hat{\Sigma}_n(w,w) + \hat{\Sigma}_n(w',w')}} \right|.$$

By Lemma B.1.5 with $w = w'$ we deduce that $\sup_{w \in \mathcal{W}} \left| \frac{\hat{\Sigma}_n(w,w)}{\Sigma_n(w,w)} - 1 \right| \lesssim_{\mathbb{P}} \sqrt{h \log n}$ and so

$$\begin{aligned}
\mathrm{obj}(\Sigma_n) &= \sup_{w,w' \in \mathcal{W}} \left| \frac{\hat{\Sigma}_n(w,w') - \Sigma_n}{\sqrt{\Sigma_n(w,w) + \Sigma_n(w',w')}} \right| \sqrt{\frac{\Sigma_n(w,w) + \Sigma_n(w',w')}{\hat{\Sigma}_n(w,w) + \hat{\Sigma}_n(w',w')}} \\
&\lesssim_{\mathbb{P}} \frac{\sqrt{\log n}}{n} \left(1 - \frac{\left|\hat{\Sigma}_n(w,w) - \Sigma_n(w,w)\right|}{\Sigma_n(w,w)} - \frac{\left|\hat{\Sigma}_n(w',w') - \Sigma_n(w',w')\right|}{\Sigma_n(w',w')} \right)^{-1/2} \\
&\lesssim_{\mathbb{P}} \frac{\sqrt{\log n}}{n} \left(1 - \sqrt{h \log n} \right)^{-1/2} \lesssim_{\mathbb{P}} \frac{\sqrt{\log n}}{n}.
\end{aligned}$$

Since the objective function is non-negative and because we have established at least one feasible function M with an almost surely finite objective value, we can conclude the following. Let $\mathrm{obj}^* = \inf_M \mathrm{obj}(M)$, where the infimum is over feasible functions M. Then for all $\varepsilon > 0$ there exists a feasible function M_ε with $\mathrm{obj}(M_\varepsilon) \leq \mathrm{obj}^* + \varepsilon$, and we call such a solution ε-optimal. Let $\hat{\Sigma}_n^+$ be an n^{-1}-optimal solution. Then

$$\mathrm{obj}(\hat{\Sigma}_n^+) \leq \mathrm{obj}^* + n^{-1} \leq \mathrm{obj}(\Sigma_n) + n^{-1}.$$

Thus by the triangle inequality,

$$\sup_{w,w' \in \mathcal{W}} \left| \frac{\hat{\Sigma}_n^+(w,w') - \Sigma_n(w,w')}{\sqrt{\Sigma_n(w,w) + \Sigma_n(w',w')}} \right| \leq \mathrm{obj}(\hat{\Sigma}_n^+) + \mathrm{obj}(\Sigma_n) \leq 2\,\mathrm{obj}(\Sigma_n) + n^{-1} \lesssim_{\mathbb{P}} \frac{\sqrt{\log n}}{n}. \quad \square$$

Proof (Lemma B.1.8)

Since $\hat{\Sigma}_n^+$ is positive semi-definite, we must have $\hat{\Sigma}_n^+(w,w) \geq 0$. Now Lemma B.1.7 implies that for all $\varepsilon \in (0,1)$ there exists a C_ε such that

$$\mathbb{P}\left(\Sigma_n(w,w) - C_\varepsilon \frac{\sqrt{\log n}}{n}\sqrt{\Sigma_n(w,w)} \leq \hat{\Sigma}_n^+(w,w) \right.$$
$$\left. \leq \Sigma_n(w,w) + C_\varepsilon \frac{\sqrt{\log n}}{n}\sqrt{\Sigma_n(w,w)}, \quad \forall w \in \mathcal{W}\right) \geq 1-\varepsilon.$$

Consider the function $g_a(t) = t - a\sqrt{t}$ and note that it is increasing on $\{t \geq a^2/4\}$. Applying this with $t = \Sigma_n(w,w)$ and $a = \frac{\sqrt{\log n}}{n}$, noting that by Lemma 3.2.4 we have $t = \Sigma_n(w,w) \gtrsim \frac{1}{n^2 h} \gg \frac{\log n}{4n^2} = a^2/4$, shows that for n large enough,

$$\inf_{w\in\mathcal{W}} \Sigma_n(w,w) - \frac{\sqrt{\log n}}{n}\sqrt{\inf_{w\in\mathcal{W}} \Sigma_n(w,w)} \lesssim_{\mathbb{P}} \inf_{w\in\mathcal{W}} \hat{\Sigma}_n^+(w,w),$$
$$\sup_{w\in\mathcal{W}} \hat{\Sigma}_n^+(w,w) \lesssim_{\mathbb{P}} \sup_{w\in\mathcal{W}} \Sigma_n(w,w) + \frac{\sqrt{\log n}}{n}\sqrt{\sup_{w\in\mathcal{W}} \Sigma_n(w,w)}.$$

Applying the bounds from Lemma 3.2.4 yields

$$\frac{D_{\text{lo}}^2}{n} + \frac{1}{n^2 h} - \frac{\sqrt{\log n}}{n}\left(\frac{D_{\text{lo}}}{\sqrt{n}} + \frac{1}{\sqrt{n^2 h}}\right) \lesssim_{\mathbb{P}} \inf_{w\in\mathcal{W}} \hat{\Sigma}_n^+(w,w),$$
$$\sup_{w\in\mathcal{W}} \hat{\Sigma}_n^+(w,w) \lesssim_{\mathbb{P}} \frac{D_{\text{up}}^2}{n} + \frac{1}{n^2 h} + \frac{\sqrt{\log n}}{n}\left(\frac{D_{\text{up}}}{\sqrt{n}} + \frac{1}{\sqrt{n^2 h}}\right)$$

and so

$$\frac{D_{\text{lo}}^2}{n} + \frac{1}{n^2 h} \lesssim_{\mathbb{P}} \inf_{w\in\mathcal{W}} \hat{\Sigma}_n^+(w,w) \leq \sup_{w\in\mathcal{W}} \hat{\Sigma}_n^+(w,w) \lesssim_{\mathbb{P}} \frac{D_{\text{up}}^2}{n} + \frac{1}{n^2 h}. \qquad \square$$

Proof (Lemma 3.5.1)

See Lemma B.1.5 and Lemma B.1.7. $\square$

Proof (Lemma B.1.9)

We have

$$\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w)-T_n(w)\right| = \sup_{w\in\mathcal{W}}\left\{\left|\hat{f}_W(w)-f_W(w)\right|\cdot\left|\frac{1}{\hat{\Sigma}_n^+(w,w)^{1/2}}-\frac{1}{\Sigma_n(w,w)^{1/2}}\right|\right\}$$
$$\leq \sup_{w\in\mathcal{W}}\left|\frac{\hat{f}_W(w)-\mathbb{E}\big[\hat{f}_W(w)\big]}{\sqrt{\Sigma_n(w,w)}}+\frac{\mathbb{E}\big[\hat{f}_W(w)\big]-f_W(w)}{\sqrt{\Sigma_n(w,w)}}\right|\cdot\sup_{w\in\mathcal{W}}\left|\frac{\hat{\Sigma}_n^+(w,w)-\Sigma_n(w,w)}{\sqrt{\Sigma_n(w,w)\hat{\Sigma}_n^+(w,w)}}\right|.$$

Now from the proof of Lemma B.1.5 we have that $\sup_{w\in\mathcal{W}}\left|\frac{\hat{f}_W(w)-\mathbb{E}\left[\hat{f}_W(w)\right]}{\sqrt{\Sigma_n(w,w)}}\right| \lesssim_{\mathbb{P}} \sqrt{\log n}$, while Theorem 3.2.1 gives $\sup_{w\in\mathcal{W}}\left|\mathbb{E}\big[\hat{f}_W(w)\big]-f_W(w)\right| \lesssim h^{p\wedge\beta}$. By Lemma 3.2.4, note that $\sup_{w\in\mathcal{W}}\Sigma_n(w,w)^{-1/2} \lesssim \frac{1}{D_{\mathrm{lo}}/\sqrt{n}+1/\sqrt{n^2h}}$, and $\sup_{w\in\mathcal{W}}\hat{\Sigma}_n^+(w,w)^{-1/2} \lesssim_{\mathbb{P}} \frac{1}{D_{\mathrm{lo}}/\sqrt{n}+1/\sqrt{n^2h}}$ by Lemma B.1.8. Thus, applying Lemma B.1.7 to control the covariance estimation error,

$$\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w)-T_n(w)\right| \lesssim_{\mathbb{P}} \left(\sqrt{\log n}+\frac{h^{p\wedge\beta}}{D_{\mathrm{lo}}/\sqrt{n}+1/\sqrt{n^2h}}\right)\frac{\sqrt{\log n}}{n}\frac{1}{D_{\mathrm{lo}}/\sqrt{n}+1/\sqrt{n^2h}}$$
$$\lesssim_{\mathbb{P}} \sqrt{\frac{\log n}{n}}\left(\sqrt{\log n}+\frac{\sqrt{n}h^{p\wedge\beta}}{D_{\mathrm{lo}}+1/\sqrt{nh}}\right)\frac{1}{D_{\mathrm{lo}}+1/\sqrt{nh}}. \qquad \square$$

Proof (Lemma B.1.10)

Firstly, note that $\hat{Z}_n^T$ exists by noting that $\hat{\Sigma}_n^+(w,w')$ and therefore also $\frac{\hat{\Sigma}_n^+(w,w')}{\sqrt{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')}}$ are positive semi-definite functions and appealing to the Kolmogorov consistency theorem (Giné and Nickl, 2021). To obtain the desired Kolmogorov–Smirnov result we discretize and use the Gaussian–Gaussian comparison result found in Lemma 3.1 in Chernozhukov et al. (2013a).

Part 1: bounding the covariance discrepancy

Define the maximum discrepancy in the (conditional) covariances of $\hat{Z}_n^T$ and Z_n^T by

$$\Delta := \sup_{w,w'\in\mathcal{W}}\left|\frac{\hat{\Sigma}_n^+(w,w')}{\sqrt{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')}}-\frac{\Sigma_n(w,w')}{\sqrt{\Sigma_n(w,w)\Sigma_n(w',w')}}\right|.$$

This variable can be bounded in probability in the following manner. First note that by the Cauchy–Schwarz inequality for covariances, $|\Sigma_n(w,w')| \le \sqrt{\Sigma_n(w,w)\Sigma_n(w',w')}$. Hence

$$\begin{aligned}
\Delta &\le \sup_{w,w' \in \mathcal{W}} \left\{ \left| \frac{\hat{\Sigma}_n^+(w,w') - \Sigma_n(w,w')}{\sqrt{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')}} \right| + \left| \frac{\sqrt{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')} - \sqrt{\Sigma_n(w,w)\Sigma_n(w',w')}}{\sqrt{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')}} \right| \right\} \\
&\le \sup_{w,w' \in \mathcal{W}} \left\{ \sqrt{\frac{\Sigma_n(w,w) + \Sigma_n(w',w')}{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')}} \left| \frac{\hat{\Sigma}_n^+(w,w') - \Sigma_n(w,w')}{\sqrt{\Sigma_n(w,w) + \Sigma_n(w',w')}} \right| \right\} \\
&\quad + \sup_{w,w' \in \mathcal{W}} \left| \frac{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w') - \Sigma_n(w,w)\Sigma_n(w',w')}{\sqrt{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')\Sigma_n(w,w)\Sigma_n(w',w')}} \right|.
\end{aligned}$$

For the first term, note that $\inf_{w \in \mathcal{W}} \hat{\Sigma}_n^+(w,w) \gtrsim \frac{D_{\mathrm{lo}}^2}{n} + \frac{1}{n^2 h}$ by Lemma B.1.8 and also $\sup_{w \in \mathcal{W}} \left| \frac{\hat{\Sigma}_n(w,w)}{\Sigma_n(w,w)} - 1 \right| \lesssim_{\mathbb{P}} \sqrt{h \log n}$ by the proof of Lemma B.1.7. Thus by Lemma B.1.7,

$$\begin{aligned}
&\sup_{w,w' \in \mathcal{W}} \left\{ \sqrt{\frac{\Sigma_n(w,w) + \Sigma_n(w',w')}{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')}} \left| \frac{\hat{\Sigma}_n^+(w,w') - \Sigma_n(w,w')}{\sqrt{\Sigma_n(w,w) + \Sigma_n(w',w')}} \right| \right\} \\
&\lesssim_{\mathbb{P}} \frac{\sqrt{\log n}}{n} \frac{1}{D_{\mathrm{lo}}/\sqrt{n} + 1/\sqrt{n^2 h}} \lesssim_{\mathbb{P}} \sqrt{\frac{\log n}{n}} \frac{1}{D_{\mathrm{lo}} + 1/\sqrt{nh}}.
\end{aligned}$$

For the second term, we have by the same bounds

$$\begin{aligned}
&\sup_{w,w' \in \mathcal{W}} \left| \frac{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w') - \Sigma_n(w,w)\Sigma_n(w',w')}{\sqrt{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')\Sigma_n(w,w)\Sigma_n(w',w')}} \right| \\
&\le \sup_{w,w' \in \mathcal{W}} \left\{ \frac{\left|\hat{\Sigma}_n^+(w,w) - \Sigma_n(w,w)\right| \hat{\Sigma}_n^+(w',w')}{\sqrt{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')\Sigma_n(w,w)\Sigma_n(w',w')}} \right\} \\
&\quad + \sup_{w,w' \in \mathcal{W}} \left\{ \frac{\left|\hat{\Sigma}_n^+(w',w') - \Sigma_n(w',w')\right| \Sigma_n(w,w)}{\sqrt{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')\Sigma_n(w,w)\Sigma_n(w',w')}} \right\} \\
&\le \sup_{w,w' \in \mathcal{W}} \left\{ \frac{\left|\hat{\Sigma}_n^+(w,w) - \Sigma_n(w,w)\right|}{\sqrt{\Sigma_n(w,w)}} \frac{\sqrt{\hat{\Sigma}_n^+(w',w')}}{\sqrt{\hat{\Sigma}_n^+(w,w)\Sigma_n(w',w')}} \right\} \\
&\quad + \sup_{w,w' \in \mathcal{W}} \left\{ \frac{\left|\hat{\Sigma}_n^+(w',w') - \Sigma_n(w',w')\right|}{\sqrt{\Sigma_n(w',w')}} \frac{\sqrt{\Sigma_n(w,w)}}{\sqrt{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')}} \right\} \lesssim_{\mathbb{P}} \sqrt{\frac{\log n}{n}} \frac{1}{D_{\mathrm{lo}} + 1/\sqrt{nh}}.
\end{aligned}$$

Therefore $\Delta \lesssim_{\mathbb{P}} \sqrt{\frac{\log n}{n}} \frac{1}{D_{\mathrm{lo}} + 1/\sqrt{nh}}$.

Part 2: Gaussian comparison on a mesh

Let $\mathcal{W}_\delta$ be a δ_n-covering of $\mathcal{W}$ with cardinality $O(1/\delta_n)$, where $1/\delta_n$ is at most polynomial in n. The scaled (conditionally) Gaussian processes Z_n^T and $\hat{Z}_n^T$ both have pointwise (conditional) variances of 1. Therefore, by Lemma 3.1 in Chernozhukov et al. (2013a),

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sup_{w\in\mathcal{W}_\delta} Z_n^T(w)\le t\right)-\mathbb{P}\left(\sup_{w\in\mathcal{W}_\delta}\hat{Z}_n^T(w)\le t\;\middle|\;\mathbf{W}_n\right)\right|\lesssim \Delta^{1/3}\Big(1\vee\log\frac{1}{\Delta\delta_n}\Big)^{2/3}$$

uniformly in the data. By the previous part and since $x(\log 1/x)^2$ is increasing on $\big(0,e^{-2}\big)$,

$$\begin{aligned}&\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sup_{w\in\mathcal{W}_\delta} Z_n^T(w)\le t\right)-\mathbb{P}\left(\sup_{w\in\mathcal{W}_\delta}\hat{Z}_n^T(w)\le t\;\middle|\;\mathbf{W}_n\right)\right|\\ &\quad\lesssim_{\mathbb{P}}\left(\sqrt{\frac{\log n}{n}}\frac{1}{D_{\mathrm{lo}}+1/\sqrt{nh}}\right)^{1/3}(\log n)^{2/3}\lesssim_{\mathbb{P}}\frac{n^{-1/6}(\log n)^{5/6}}{D_{\mathrm{lo}}^{1/3}+(nh)^{-1/6}}.\end{aligned}$$

Part 3: trajectory regularity of Z_n^T

In the proof of Theorem 3.4.2 we established that Z_n^T satisfies the regularity property

$$\mathbb{E}\left[\sup_{|w-w'|\le\delta_n}\big|Z_n^T(w)-Z_n^T(w')\big|\right]\lesssim nh^{-1}\sqrt{\delta_n\log n},$$

whenever $1/\delta_n$ is at most polynomial in n.

Part 4: conditional L^2 regularity of $\hat{Z}_n^T$

By Lemma B.1.7, with $nh\gtrsim\log n$, we have uniformly in w,w',

$$\big|\hat{\Sigma}_n^+(w,w')-\hat{\Sigma}_n^+(w,w)\big|\lesssim n^{-1}h^{-3}|w-w'|.$$

Taking $\delta_n\le n^{-2}h^2$, Lemma B.1.8 gives

$$\inf_{|w-w'|\le\delta_n}\hat{\Sigma}_n^+(w,w')\gtrsim\frac{D_{\mathrm{lo}}^2}{n}+\frac{1}{n^2h}-n^{-1}h^{-3}\delta_n\gtrsim\frac{D_{\mathrm{lo}}^2}{n}+\frac{1}{n^2h}-\frac{1}{n^3h}\gtrsim\frac{D_{\mathrm{lo}}^2}{n}+\frac{1}{n^2h}.$$

The conditional L^2 regularity of $\hat{Z}_n^T$ is

$$\mathbb{E}\left[\left(\hat{Z}_n^T(w) - \hat{Z}_n^T(w')\right)^2 \mid \mathbf{W}_n\right] = 2 - 2\frac{\hat{\Sigma}_n^+(w,w')}{\sqrt{\hat{\Sigma}_n^+(w,w)\hat{\Sigma}_n^+(w',w')}}.$$

Applying the same elementary result as for Z_n^T in the proof of Theorem 3.4.2 yields

$$\mathbb{E}\left[\left(\hat{Z}_n^T(w) - \hat{Z}_n^T(w')\right)^2 \mid \mathbf{W}_n\right] \lesssim_{\mathbb{P}} n^2 h^{-2}|w - w'|.$$

Thus the conditional semimetric induced by $\hat{Z}_n^T$ on $\mathcal{W}$ is

$$\hat{\rho}(w,w') := \mathbb{E}\left[\left(\hat{Z}_n^T(w) - \hat{Z}_n^T(w')\right)^2 \mid \mathbf{W}_n\right]^{1/2} \lesssim_{\mathbb{P}} nh^{-1}\sqrt{|w - w'|}.$$

Part 5: conditional trajectory regularity of $\hat{Z}_n^T$

As for Z_n^T in the proof of Theorem 3.4.2, we apply Lemma B.3.4, now conditionally, to obtain

$$\mathbb{E}\left[\sup_{|w-w'|\le\delta_n}\left|\hat{Z}_n^T(w) - \hat{Z}_n^T(w')\right| \,\middle|\, \mathbf{W}_n\right] \lesssim_{\mathbb{P}} nh^{-1}\sqrt{\delta_n \log n},$$

whenever $1/\delta_n$ is at most polynomial in n.

Part 6: uniform Gaussian comparison

Now we use the trajectory regularity properties to extend the Gaussian–Gaussian comparison result from a finite mesh to all of $\mathcal{W}$. Write the previously established approximation rate as

$$r_n = \frac{n^{-1/6}(\log n)^{5/6}}{D_{\mathrm{lo}}^{1/3} + (nh)^{-1/6}}.$$

Take $\varepsilon_n > 0$ and observe that uniformly in $t \in \mathbb{R}$,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{Z}_n^T(w)\right| \le t \;\middle|\; \mathbf{W}_n\right) \\
&\quad\le \mathbb{P}\left(\sup_{w\in\mathcal{W}_\delta}\left|\hat{Z}_n^T(w)\right| \le t+\varepsilon_n \;\middle|\; \mathbf{W}_n\right) + \mathbb{P}\left(\sup_{|w-w'|\le\delta_n}\left|\hat{Z}_n^T(w)-\hat{Z}_n^T(w')\right| \ge \varepsilon_n \;\middle|\; \mathbf{W}_n\right) \\
&\quad\le \mathbb{P}\left(\sup_{w\in\mathcal{W}_\delta}\left|Z_n^T(w)\right| \le t+\varepsilon_n\right) + O_{\mathbb{P}}(r_n) + \mathbb{P}\left(\sup_{|w-w'|\le\delta_n}\left|\hat{Z}_n^T(w)-\hat{Z}_n^T(w')\right| \ge \varepsilon_n \;\middle|\; \mathbf{W}_n\right) \\
&\quad\le \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| \le t+2\varepsilon_n\right) + O_{\mathbb{P}}(r_n) + \mathbb{P}\left(\sup_{|w-w'|\le\delta_n}\left|Z_n^T(w)-Z_n^T(w')\right| \ge \varepsilon_n\right) \\
&\qquad + \mathbb{P}\left(\sup_{|w-w'|\le\delta_n}\left|\hat{Z}_n^T(w)-\hat{Z}_n^T(w')\right| \ge \varepsilon_n \;\middle|\; \mathbf{W}_n\right) \\
&\quad\le \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| \le t+2\varepsilon_n\right) + O_{\mathbb{P}}(r_n) + O_{\mathbb{P}}(\varepsilon_n^{-1}nh^{-1}\sqrt{\delta_n\log n}) \\
&\quad\le \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| \le t\right) + \mathbb{P}\left(\left|\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| - t\right| \le 2\varepsilon_n\right) \\
&\qquad + O_{\mathbb{P}}(r_n) + O_{\mathbb{P}}(\varepsilon_n^{-1}nh^{-1}\sqrt{\delta_n\log n}).
\end{aligned}
$$

The converse inequality is obtained analogously as follows:

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{Z}_n^T(w)\right| \le t \;\middle|\; \mathbf{W}_n\right) \\
&\quad\ge \mathbb{P}\left(\sup_{w\in\mathcal{W}_\delta}\left|\hat{Z}_n^T(w)\right| \le t-\varepsilon_n \;\middle|\; \mathbf{W}_n\right) - \mathbb{P}\left(\sup_{|w-w'|\le\delta_n}\left|\hat{Z}_n^T(w)-\hat{Z}_n^T(w')\right| \ge \varepsilon_n \;\middle|\; \mathbf{W}_n\right) \\
&\quad\ge \mathbb{P}\left(\sup_{w\in\mathcal{W}_\delta}\left|Z_n^T(w)\right| \le t-\varepsilon_n\right) - O_{\mathbb{P}}(r_n) - \mathbb{P}\left(\sup_{|w-w'|\le\delta_n}\left|\hat{Z}_n^T(w)-\hat{Z}_n^T(w')\right| \ge \varepsilon_n \;\middle|\; \mathbf{W}_n\right) \\
&\quad\ge \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| \le t-2\varepsilon_n\right) - O_{\mathbb{P}}(r_n) - \mathbb{P}\left(\sup_{|w-w'|\le\delta_n}\left|Z_n^T(w)-Z_n^T(w')\right| \ge \varepsilon_n\right) \\
&\qquad - \mathbb{P}\left(\sup_{|w-w'|\le\delta_n}\left|\hat{Z}_n^T(w)-\hat{Z}_n^T(w')\right| \ge \varepsilon_n \;\middle|\; \mathbf{W}_n\right) \\
&\quad\ge \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| \le t-2\varepsilon_n\right) - O_{\mathbb{P}}(r_n) - O_{\mathbb{P}}(\varepsilon_n^{-1}nh^{-1}\sqrt{\delta_n\log n}) \\
&\quad\ge \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| \le t\right) - \mathbb{P}\left(\left|\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| - t\right| \le 2\varepsilon_n\right) \\
&\qquad - O_{\mathbb{P}}(r_n) - O_{\mathbb{P}}(\varepsilon_n^{-1}nh^{-1}\sqrt{\delta_n\log n}).
\end{aligned}
$$

Combining these uniform upper and lower bounds gives

$$\begin{aligned}\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{Z}_n^T(w)\right|\le t\;\middle|\;\mathbf{W}_n\right)-\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right|\le t\right)\right|\\ \lesssim_{\mathbb{P}}\sup_{t\in\mathbb{R}}\mathbb{P}\left(\left|\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right|-t\right|\le 2\varepsilon_n\right)+r_n+\varepsilon_n^{-1}nh^{-1/2}\delta_n^{1/2}\sqrt{\log n}.\end{aligned}$$

For the remaining term, apply anti-concentration for Z_n^T from the proof of Theorem 3.4.2:

$$\sup_{t\in\mathbb{R}}\mathbb{P}\left(\left|\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right|-t\right|\le\varepsilon\right)\lesssim\varepsilon\sqrt{\log n}.$$

Therefore

$$\begin{aligned}\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{Z}_n^T(w)\right|\le t\;\middle|\;\mathbf{W}_n\right)-\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right|\le t\right)\right|\\ \lesssim_{\mathbb{P}}\varepsilon_n\sqrt{\log n}+r_n+\varepsilon_n^{-1}nh^{-1/2}\delta_n^{1/2}\sqrt{\log n}.\end{aligned}$$

Taking $\varepsilon=r_n/\sqrt{\log n}$ and then $\delta_n=n^{-2}hr_n^2\varepsilon_n^2/\log n$ yields

$$\left|\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{Z}_n^T(w)\right|\le t\;\middle|\;\mathbf{W}_n\right)-\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right|\le t\right)\right|\lesssim_{\mathbb{P}}r_n=\frac{n^{-1/6}(\log n)^{5/6}}{D_{\text{lo}}^{1/3}+(nh)^{-1/6}}.\qquad\square$$

Proof (Lemma B.1.11)

Part 1: Kolmogorov–Smirnov approximation

Let Z_n^T and $\hat{Z}_n^T$ be defined as in the proof of Lemma B.1.10. Write

$$r_n=\frac{n^{-1/6}(\log n)^{5/6}}{D_{\text{lo}}^{1/3}+(nh)^{-1/6}}$$

for the rate of approximation from Lemma B.1.10. For any $\varepsilon_n > 0$ and uniformly in $t \in \mathbb{R}$:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{Z}_n^T(w)\right| \le t \;\middle|\; \mathbf{W}_n\right) \le \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| \le t\right) + O_{\mathbb{P}}(r_n) \\
&\quad\le \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| \le t-\varepsilon_n\right) + \mathbb{P}\left(\left|\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| - t\right| \le \varepsilon_n\right) + O_{\mathbb{P}}(r_n) \\
&\quad\le \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w)\right| \le t\right) + \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w) - Z_n^T(w)\right| \ge \varepsilon_n\right) \\
&\qquad + \mathbb{P}\left(\left|\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| - t\right| \le \varepsilon_n\right) + O_{\mathbb{P}}(r_n) \\
&\quad\le \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w)\right| \le t\right) + \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w) - Z_n^T(w)\right| \ge \varepsilon_n\right) + \varepsilon_n\sqrt{\log n} + O_{\mathbb{P}}(r_n),
\end{aligned}$$

where in the last line we used the anti-concentration result from Lemma B.3.5 applied to Z_n^T, as in the proof of Lemma B.1.10. The corresponding lower bound is as follows:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{Z}_n^T(w)\right| \le t \;\middle|\; \mathbf{W}_n\right) \ge \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| \le t\right) - O_{\mathbb{P}}(r_n) \\
&\quad\ge \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| \le t+\varepsilon_n\right) - \mathbb{P}\left(\left|\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| - t\right| \le \varepsilon_n\right) - O_{\mathbb{P}}(r_n) \\
&\quad\ge \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w)\right| \le t\right) - \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w) - Z_n^T(w)\right| \ge \varepsilon_n\right) \\
&\qquad - \mathbb{P}\left(\left|\sup_{w\in\mathcal{W}}\left|Z_n^T(w)\right| - t\right| \le \varepsilon_n\right) - O_{\mathbb{P}}(r_n) \\
&\quad\ge \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w)\right| \le t\right) - \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w) - Z_n^T(w)\right| \ge \varepsilon_n\right) - \varepsilon_n\sqrt{\log n} - O_{\mathbb{P}}(r_n).
\end{aligned}$$

Part 2: t-statistic approximation

To control the remaining term, note that by Theorem 3.4.1 and Lemma B.1.9,

$$\begin{aligned}
&\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w) - Z_n^T(w)\right| \\
&\quad\le \sup_{w\in\mathcal{W}}\left|\hat{T}_n(w) - T_n(w)\right| + \sup_{w\in\mathcal{W}}\left|T_n(w) - Z_n^T(w)\right| \\
&\quad\lesssim_{\mathbb{P}} \sqrt{\frac{\log n}{n}}\left(\sqrt{\log n} + \frac{\sqrt{n}h^{p\wedge\beta}}{D_{\text{lo}} + 1/\sqrt{nh}}\right)\frac{1}{D_{\text{lo}} + 1/\sqrt{nh}} \\
&\qquad + \frac{n^{-1/2}\log n + n^{-3/4}h^{-7/8}(\log n)^{3/8}R_n + n^{-2/3}h^{-1/2}(\log n)^{2/3} + n^{1/2}h^{p\wedge\beta}}{D_{\text{lo}} + 1/\sqrt{nh}}
\end{aligned}$$

and denote this last quantity by r'_n. Then for any $\varepsilon_n \gg r'_n$, we have

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w)\right|\le t\right)-\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{Z}_n^T(w)\right|\le t\;\middle|\;\mathbf{W}_n\right)\right|\lesssim_{\mathbb{P}}\varepsilon_n\sqrt{\log n}+r_n+o(1).$$

Part 3: rate analysis

This rate is $o_{\mathbb{P}}(1)$ with an appropriate choice of ε_n whenever $r_n\to 0$ and $r'_n\sqrt{\log n}\to 0$, by Lemma B.3.6, along with a slowly diverging sequence R_n. Explicitly, we require the following.

$$\frac{n^{-1/2}(\log n)^{3/2}}{D_{\text{lo}}+1/\sqrt{nh}}\to 0, \qquad \frac{h^{p\wedge\beta}\log n}{D_{\text{lo}}^2+(nh)^{-1}}\to 0,$$
$$\frac{n^{-1/2}(\log n)^{3/2}}{D_{\text{lo}}+1/\sqrt{nh}}\to 0, \qquad \frac{n^{-3/4}h^{-7/8}(\log n)^{7/8}}{D_{\text{lo}}+1/\sqrt{nh}}\to 0,$$
$$\frac{n^{-2/3}h^{-1/2}(\log n)^{7/6}}{D_{\text{lo}}+1/\sqrt{nh}}\to 0, \qquad \frac{n^{1/2}h^{p\wedge\beta}(\log n)^{1/2}}{D_{\text{lo}}+1/\sqrt{nh}}\to 0,$$
$$\frac{n^{-1/6}(\log n)^{5/6}}{D_{\text{lo}}^{1/3}+(nh)^{-1/6}}\to 0.$$

Using the fact that $h\lesssim n^{-\varepsilon}$ for some $\varepsilon>0$ and removing trivial statements leaves us with

$$\frac{n^{-3/4}h^{-7/8}(\log n)^{7/8}}{D_{\text{lo}}+1/\sqrt{nh}}\to 0, \qquad \frac{n^{1/2}h^{p\wedge\beta}(\log n)^{1/2}}{D_{\text{lo}}+1/\sqrt{nh}}\to 0.$$

We analyze these based on the degeneracy and verify that they hold under Assumption 3.4.1.

(i) No degeneracy: if $D_{\text{lo}}>0$ then we need

$$n^{-3/4}h^{-7/8}(\log n)^{7/8}\to 0, \qquad n^{1/2}h^{p\wedge\beta}(\log n)^{1/2}\to 0.$$

These reduce to $n^{-6/7}\log n\ll h\ll(n\log n)^{-\frac{1}{2(p\wedge\beta)}}$.

(ii) Partial or total degeneracy: if $D_{\text{lo}}=0$ then we need

$$n^{-1/4}h^{-3/8}(\log n)^{7/8}\to 0, \qquad nh^{(p\wedge\beta)+1/2}(\log n)^{1/2}\to 0.$$

These reduce to $n^{-2/3}(\log n)^{7/3}\ll h\ll(n^2\log n)^{-\frac{1}{2(p\wedge\beta)+1}}$. □

Proof (Theorem 3.5.1)

Part 1: existence of the conditional quantile

We argue as in the proof of Lemma B.1.10, now also conditioning on the data. In particular, using the anti-concentration result from Lemma B.3.5, the regularity property of $\hat{Z}_n^T$, and the Gaussian process maximal inequality from Lemma B.3.4, we see that for any $\varepsilon > 0$,

$$\sup_{t\in\mathbb{R}} \mathbb{P}\left(\left|\sup_{w\in\mathcal{W}} \left|\hat{Z}_n^T(w)\right| - t\right| \leq 2\varepsilon \;\middle|\; \mathbf{W}_n\right) \leq 8\varepsilon\left(1 + \mathbb{E}\left[\sup_{w\in\mathcal{W}} \left|\hat{Z}_n^T(w)\right| \;\middle|\; \mathbf{W}_n\right]\right) \lesssim \varepsilon\sqrt{\log n}.$$

Thus letting $\varepsilon \to 0$ shows that the conditional distribution function of $\sup_{w\in\mathcal{W}} \left|\hat{Z}_n^T(w)\right|$ is continuous, and therefore all of its conditional quantiles exist.

Part 2: validity of the confidence band

Define the following (conditional) distribution functions.

$$F_Z(t \mid \mathbf{W}_n) = \mathbb{P}\left(\sup_{w\in\mathcal{W}} \left|\hat{Z}_n^T(w)\right| \leq t \;\middle|\; \mathbf{W}_n\right), \qquad F_T(t) = \mathbb{P}\left(\sup_{w\in\mathcal{W}} \left|\hat{T}_n(w)\right| \leq t\right),$$

along with their well-defined right-quantile functions,

$$F_Z^{-1}(p \mid \mathbf{W}_n) = \sup\left\{t \in \mathbb{R} \,:\, F_Z(t \mid \mathbf{W}_n) = p\right\}, \qquad F_T^{-1}(p) = \sup\left\{t \in \mathbb{R} \,:\, F_T(t) = p\right\}.$$

Note that $t \leq F_Z^{-1}(p \mid \mathbf{W}_n)$ if and only if $F_Z(t \mid \mathbf{W}_n) \leq p$. Take $\alpha \in (0,1)$ and define the quantile $\hat{q}_{1-\alpha} = F_Z^{-1}(1-\alpha \mid \mathbf{W}_n)$, so that $F_Z(\hat{q}_{1-\alpha} \mid \mathbf{W}_n) = 1-\alpha$. By Lemma B.1.11,

$$\sup_{t\in\mathbb{R}} \left|F_Z(t \mid \mathbf{W}_n) - F_T(t)\right| = o_{\mathbb{P}}(1).$$

Thus by Lemma B.3.6, this can be replaced by

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}} \left|F_Z(t \mid \mathbf{W}_n) - F_T(t)\right| > \varepsilon_n\right) \leq \varepsilon_n$$

for some $\varepsilon_n \to 0$. Therefore

$$\begin{aligned}
\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w)\right| \le \hat{q}_{1-\alpha}\right) &= \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w)\right| \le F_Z^{-1}(1-\alpha \mid \mathbf{W}_n)\right) \\
&= \mathbb{P}\left(F_Z\left(\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w)\right| \;\middle|\; \mathbf{W}_n\right) \le 1-\alpha\right) \\
&\le \mathbb{P}\left(F_T\left(\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w)\right|\right) \le 1-\alpha+\varepsilon_n\right) + \varepsilon_n \le 1-\alpha+3\varepsilon_n,
\end{aligned}$$

where we used the fact that for any real-valued random variable X with distribution function F, we have $\left|\mathbb{P}\big(F(X)\le t\big) - t\right| \le \Delta$, where Δ is the size of the largest jump discontinuity in F. By uniform integrability, $\sup_{t\in\mathbb{R}}\left|F_Z(t) - F_T(t)\right| = o(\varepsilon_n)$. Since F_Z has no jumps, we must have $\Delta \le \varepsilon_n$ for F_T. Finally, a lower bound is constructed in an analogous manner, giving

$$\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\hat{T}_n(w)\right| \le \hat{q}_{1-\alpha}\right) \ge 1-\alpha-3\varepsilon_n. \qquad \square$$

Proof (Lemma B.1.12)

Writing $k_{ij} = k_h(W_{ij}^1, w)$, $\psi_i = \psi(X_i^1)$, $\hat{\psi}_i = \hat{\psi}(X_i^1)$, and $\kappa_{ij} = \kappa(X_i^0, X_i^1, X_j^1)$,

$$\begin{aligned}
\mathbb{E}\big[\hat{f}_W^{1\triangleright 0}(w)\big] &= \mathbb{E}\left[\frac{2}{n(n-1)}\sum_{i<j}\hat{\psi}_i\hat{\psi}_j k_{ij}\right] \\
&= \frac{2}{n(n-1)(n-2)}\sum_{i<j}\sum_{r\notin\{i,j\}}\mathbb{E}\left[k_{ij}\Big(\psi_i\psi_j + \psi_i\kappa_{rj} + \psi_j\kappa_{ri}\Big)\right] + O\left(\frac{1}{n}\right) \\
&= \mathbb{E}\left[k_{ij}\psi_i\psi_j\right] + O\left(\frac{1}{n}\right) = \mathbb{E}\big[\psi_i\psi_j\mathbb{E}\left[k_h(W_{ij}^1, w) \mid X_i^1, X_j^1\right]\big] + O\left(\frac{1}{n}\right) \\
&= \mathbb{E}\big[\psi_i\psi_j f_{W|XX}^1(w \mid X_i^1, X_j^1) + O_{\mathbb{P}}\left(h^{p\wedge\beta}\right)\big] + O\left(\frac{1}{n}\right) \\
&= f_W^{1\triangleright 0}(w) + O\left(h^{p\wedge\beta} + \frac{1}{n}\right)
\end{aligned}$$

uniformly in w, by the proof of Theorem 3.2.1 and Hölder continuity of $f_{W|XX}^1$. $\square$

Proof (Lemma B.1.13)

$$\begin{aligned}
\hat{f}_W^{1\triangleright 0}(w) &= \frac{2}{n(n-1)} \sum_{i<j} \hat{\psi}_i \hat{\psi}_j k_{ij} \\
&= \frac{2}{n(n-1)} \sum_{i<j} \left(\psi_i + \frac{1}{n} \sum_{r=1}^{n} \kappa_{ri} \right) \left(\psi_j + \frac{1}{n} \sum_{r=1}^{n} \kappa_{rj} \right) k_{ij} + O_{\mathbb{P}}\left(\frac{1}{n}\right) \\
&= \frac{2}{n(n-1)} \sum_{i<j} \psi_i \psi_j k_{ij} + \frac{2}{n(n-1)} \sum_{i<j} \psi_i \frac{1}{n} \sum_{r \notin \{i,j\}}^{n} \kappa_{rj} k_{ij} \\
&\quad + \frac{2}{n(n-1)} \sum_{i<j} \psi_j \frac{1}{n} \sum_{r \notin \{i,j\}}^{n} \kappa_{ri} k_{ij} + O_{\mathbb{P}}\left(\frac{1}{n}\right) \\
&= \frac{2}{n(n-1)(n-2)} \sum_{i<j} \sum_{r \notin \{i,j\}} k_{ij} \Big(\psi_i \psi_j + \psi_i \kappa_{rj} + \psi_j \kappa_{ri} \Big) + O_{\mathbb{P}}\left(\frac{1}{n}\right) \\
&= \frac{6}{n(n-1)(n-2)} \sum_{i<j<r} v_{ijr} + O_{\mathbb{P}}\left(\frac{1}{n}\right)
\end{aligned}$$

where

$$\begin{aligned}
v_{ijr} = \frac{1}{3} k_{ij} \Big(\psi_i \psi_j + \psi_i \kappa_{rj} + \psi_j \kappa_{ri} \Big) + \frac{1}{3} k_{ir} \Big(\psi_i \psi_r + \psi_i \kappa_{jr} + \psi_r \kappa_{ji} \Big) \\
+ \frac{1}{3} k_{jr} \Big(\psi_j \psi_r + \psi_j \kappa_{ir} + \psi_r \kappa_{ij} \Big)
\end{aligned}$$

So by the Hoeffding decomposition for third-order U-statistics,

$$\begin{aligned}
\hat{f}_W^{1\triangleright 0}(w) &= u + \frac{3}{n} \sum_{i=1}^{n} u_i + \frac{6}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} u_{ij} + \frac{6}{n(n-1)(n-2)} \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{r=j+1}^{n} u_{ijr} \\
&\quad + \frac{6}{n(n-1)(n-2)} \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{r=j+1}^{n} \big(v_{ijr} - u_{ijr} \big) + O_{\mathbb{P}}\left(\frac{1}{n}\right) \\
&= \mathbb{E}\big[\hat{f}_W^{1\triangleright 0}(w) \big] + L_n^{1\triangleright 0}(w) + Q_n^{1\triangleright 0}(w) + T_n^{1\triangleright 0}(w) + E_n^{1\triangleright 0}(w) + O_{\mathbb{P}}\left(\frac{1}{n}\right).
\end{aligned}$$

Noting that ψ_i, κ_{ij} and $\mathbb{E}[k_{ij} \mid A_i^1, A_j^1]$ are all bounded and that $\mathbb{E}[k_{ij} \mid A_i^1, A_j^1]$ is Lipschitz in w, we deduce by Lemma B.3.9 and Proposition 2.3 of Arcones and Giné (1993) that $\sup_{w \in \mathcal{W}} |Q_n^{1\triangleright 0}(w) + T_n^{1\triangleright 0}(w)| \lesssim_{\mathbb{P}} \frac{1}{n}$. □

Proof (Lemma B.1.14)

By Lemma B.2.2, $\sup_{w\in\mathcal{W}}\left|L_n^{1\triangleright 0}(w)\right| \lesssim_{\mathbb{P}} \frac{1}{\sqrt{n}}$. In the proof of Lemma B.1.13 the terms $v_{ijr}-u_{ijr}$ depend only on V_{ij}, V_{ir}, and V_{jr} after conditioning on $\mathbf{A}_n^1$, $\mathbf{X}_n^0$, and $\mathbf{X}_n^1$. Thus $E_n^{1\triangleright 0}(w)$ is a degenerate second-order U-statistic so $\sup_{w\in\mathcal{W}}\left|E_n^{1\triangleright 0}(w)\right| \lesssim_{\mathbb{P}} \sqrt{\frac{\log n}{n^2 h}}$ by Lemma B.3.9. □

Proof (Lemma B.1.15)

Note that $L_n^{1\triangleright 0}(w) = \frac{3}{n}\sum_{i=1}^n l_i^{1\triangleright 0}(w)$ where $l_i^{1\triangleright 0}(w)$ depends only on A_i^1, X_i^0, and X_i^1. Let $\gamma : \mathcal{X}\times\mathcal{X} \to \{1,\ldots,|\mathcal{X}|^2\}$ be a bijection and define $\text{logistic}(x) = \frac{1}{1+e^{-x}}$. Let $\tilde{A}_i = \text{logistic}(A_i^1) + \gamma(X_i^0, X_i^1)$ so that $A_i^1 = \text{logistic}^{-1}\left(\tilde{A}_i - \lfloor \tilde{A}_i \rfloor\right)$ and $(X_i^0, X_i^1) = \gamma^{-1}(\lfloor \tilde{A}_i \rfloor)$. Thus $l_i^{1\triangleright 0}(w)$ is a bounded-variation function of $\tilde{A}_i$, uniformly in w, and so as in Lemma B.1.1 we have that on an appropriately enlarged probability space,

$$\mathbb{E}\left[\sup_{w\in\mathcal{W}}\left|\sqrt{n}L_n^{1\triangleright 0}(w) - Z_n^{L,1\triangleright 0}(w)\right|\right] \lesssim \frac{\log n}{\sqrt{n}}$$

where $Z_n^{L,1\triangleright 0}$ is a mean-zero Gaussian process with the same covariance as $\sqrt{n}L_n^{1\triangleright 0}$. For $E_n^{1\triangleright 0}(w)$, we first construct a strong approximation conditional on $\mathbf{A}_n$ and $\mathbf{X}_n$ as shown in Lemma B.1.2 and deduce an unconditional strong approximation as in Lemma B.1.3 to see

$$\sup_{w\in\mathcal{W}}\left|\sqrt{n^2 h}E_n^{1\triangleright 0}(w) - Z_n^{E,1\triangleright 0}(w)\right| \lesssim_{\mathbb{P}} n^{-1/4}h^{-3/8}(\log n)^{3/8}R_n + n^{-1/6}(\log n)^{2/3}$$

where $Z_n^{E,1\triangleright 0}$ is a mean-zero Gaussian process with the same covariance as $\sqrt{n^2 h}E_n^{1\triangleright 0}$. Arguing as in the proof of Theorem B.1.1 shows that the Gaussian processes are independent and can be summed to yield a single strong approximation. □

Proof (Lemma B.1.16)

Arguing by mean-zero properties and conditional independence,

$$
\begin{aligned}
\Sigma_n^{1\triangleright 0}(w,w') &= \operatorname{Cov}\left[\hat{f}_W^{1\triangleright 0}(w), \hat{f}_W^{1\triangleright 0}(w')\right] \\
&= \frac{1}{n^2(n-1)^2(n-2)^2} \sum_{i\neq j} \sum_{r\notin\{i,j\}} \sum_{i'\neq j'} \sum_{r'\notin\{i',j'\}} \mathbb{E}\Big[\Big(k_{ij}\psi_i\psi_j - \mathbb{E}[k_{ij}\psi_i\psi_j] + k_{ij}\psi_i\kappa_{rj} + k_{ij}\psi_j\kappa_{ri}\Big) \\
&\qquad \times \Big(k'_{i'j'}\psi_{i'}\psi_{j'} - \mathbb{E}[k'_{ij}\psi_i\psi_j] + k'_{i'j'}\psi_{i'}\kappa_{r'j'} + k'_{i'j'}\psi_{j'}\kappa_{r'i'}\Big)\Big] + O\left(\frac{1}{n^{3/2}} + \frac{1}{\sqrt{n^4h}}\right) \\
&= \frac{2}{n^2}\mathbb{E}\left[k_{ij}\psi_i\psi_j k'_{ij}\psi_i\psi_j\right] + \frac{4}{n}\mathbb{E}\left[k_{ij}\psi_i\psi_j k'_{ir}\psi_i\psi_r\right] - \frac{4}{n}\mathbb{E}\left[k_{ij}\psi_i\psi_j\right]\mathbb{E}\left[k'_{ij}\psi_i\psi_j\right] \\
&\quad + \frac{4}{n}\mathbb{E}\left[k_{ij}\psi_i\kappa_{i'j}k'_{i'j'}\psi_{i'}\psi_{j'}\right] + \frac{4}{n}\mathbb{E}\left[k_{ij}\psi_i\psi_j k'_{i'j'}\psi_{i'}\kappa_{ij'}\right] + \frac{4}{n}\mathbb{E}\left[k_{ij}k'_{i'j'}\psi_i\psi_{i'}\kappa_{rj}\kappa_{rj'}\right] \\
&\quad + O\left(\frac{1}{n^{3/2}} + \frac{1}{\sqrt{n^4h}}\right) \\
&= \frac{4}{n}\mathbb{E}\left[\Big(\psi_i\mathbb{E}\big[k_{ij}\psi_j \mid i\big] + \mathbb{E}\left[k_{rj}\psi_r\kappa_{ij} \mid i\right]\Big)\Big(\psi_i\mathbb{E}\big[k'_{ij}\psi_j \mid i\big] + \mathbb{E}\left[k'_{rj}\psi_r\kappa_{ij} \mid i\right]\Big)\right] \\
&\quad + \frac{2}{n^2}\mathbb{E}\left[k_{ij}k'_{ij}\psi_i^2\psi_j^2\right] - \frac{4}{n}\mathbb{E}\left[k_{ij}\psi_i\psi_j\right]\mathbb{E}\left[k'_{ij}\psi_i\psi_j\right] + O\left(\frac{1}{n^{3/2}} + \frac{1}{\sqrt{n^4h}}\right),
\end{aligned}
$$

where all indices are distinct. □

Proof (Lemma B.1.17)

The proof is exactly the same as the proof of Theorem 3.4.1. □

Proof (Theorem B.1.2)

This proof proceeds in the same manner as the proof of Theorem 3.4.2. □

Appendix C

Supplement to Yurinskii's Coupling for Martingales

C.1 Proofs of main results

C.1.1 Preliminary lemmas

We give a sequence of preliminary lemmas which are useful for establishing our main results. Firstly, we present a conditional version of Strassen's theorem for the ℓ^p-norm (Chen and Kato, 2020, Theorem B.2), stated for completeness as Lemma C.1.1.

Lemma C.1.1 (A conditional Strassen theorem for the ℓ^p-norm)
Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space supporting the $\mathbb{R}^d$-valued variable X for some $d \geq 1$. Let $\mathcal{H}'$ be a countably generated sub-σ-algebra of $\mathcal{H}$ and suppose there is a $\mathrm{Unif}[0,1]$ *random variable on $(\Omega, \mathcal{H}, \mathbb{P})$, independent of the σ-algebra generated by X and $\mathcal{H}'$. Take a regular conditional distribution $F(\cdot \mid \mathcal{H}')$ satisfying the following. Firstly, $F(A \mid \mathcal{H}')$ is an $\mathcal{H}'$-measurable variable for all Borel sets $A \in \mathcal{B}(\mathbb{R}^d)$. Secondly, $F(\cdot \mid \mathcal{H}')(\omega)$ is a Borel probability measure on $\mathbb{R}^d$ for all $\omega \in \Omega$. Taking $\eta, \rho > 0$ and $p \in [1, \infty]$, with $\mathbb{E}^*$ the outer expectation, if*

$$\mathbb{E}^*\left[\sup_{A \in \mathcal{B}(\mathbb{R}^d)} \left\{\mathbb{P}\big(X \in A \mid \mathcal{H}'\big) - F\big(A_p^\eta \mid \mathcal{H}'\big)\right\}\right] \leq \rho,$$

where $A_p^\eta = \{x \in \mathbb{R}^d : \|x - A\|_p \le \eta\}$ *and* $\|x - A\|_p = \inf_{x' \in A} \|x - x'\|_p$*, then there exists an* $\mathbb{R}^d$*-valued random variable* Y *with* $Y \mid \mathcal{H}' \sim F(\cdot \mid \mathcal{H}')$ *and* $\mathbb{P}(\|X - Y\|_p > \eta) \le \rho$*.*

Proof (Lemma C.1.1)

By Theorem B.2 in Chen and Kato (2020), noting that the σ-algebra generated by Z is countably generated and using the metric induced by the ℓ^p-norm. □

Next, we present in Lemma C.1.2 an analytic result concerning the smooth approximation of Borel set indicator functions, similar to that given in Belloni et al. (2019, Lemma 39).

Lemma C.1.2 (Smooth approximation of Borel indicator functions)

Let $A \subseteq \mathbb{R}^d$ *be a Borel set and* $Z \sim \mathcal{N}(0, I_d)$*. For* $\sigma, \eta > 0$ *and* $p \in [1, \infty]$*, define*

$$g_{A\eta}(x) = \left(1 - \frac{\|x - A^\eta\|_p}{\eta}\right) \vee 0 \qquad \text{and} \qquad f_{A\eta\sigma}(x) = \mathbb{E}\big[g_{A\eta}(x + \sigma Z)\big].$$

Then f *is infinitely differentiable and with* $\varepsilon = \mathbb{P}(\|Z\|_p > \eta/\sigma)$*, for all* $k \ge 0$*, any multi-index* $\kappa = (\kappa_1, \ldots, \kappa_d) \in \mathbb{N}^d$*, and all* $x, y \in \mathbb{R}^d$*, we have* $|\partial^\kappa f_{A\eta\sigma}(x)| \le \frac{\sqrt{\kappa!}}{\sigma^{|\kappa|}}$ *and*

$$\left| f_{A\eta\sigma}(x + y) - \sum_{|\kappa|=0}^{k} \frac{1}{\kappa!} \partial^\kappa f_{A\eta\sigma}(x) y^\kappa \right| \le \frac{\|y\|_p \|y\|_2^k}{\sigma^k \eta \sqrt{k!}},$$

$$(1 - \varepsilon)\mathbb{I}\big\{x \in A\big\} \le f_{A\eta\sigma}(x) \le \varepsilon + (1 - \varepsilon)\mathbb{I}\big\{x \in A^{3\eta}\big\}.$$

Proof (Lemma C.1.2)

Drop subscripts on $g_{A\eta}$ and $f_{A\eta\sigma}$. By Taylor's theorem with Lagrange remainder, for $t \in [0, 1]$,

$$\left| f(x + y) - \sum_{|\kappa|=0}^{k} \frac{1}{\kappa!} \partial^\kappa f(x) y^\kappa \right| \le \left| \sum_{|\kappa|=k} \frac{y^\kappa}{\kappa!} \big(\partial^\kappa f(x + ty) - \partial^\kappa f(x)\big) \right|.$$

Now with $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$,

$$f(x) = \mathbb{E}\big[g(x + \sigma W)\big] = \int_{\mathbb{R}^d} g(x + \sigma u) \prod_{j=1}^{d} \phi(u_j) \, \mathrm{d}u = \frac{1}{\sigma^d} \int_{\mathbb{R}^d} g(u) \prod_{j=1}^{d} \phi\left(\frac{u_j - x_j}{\sigma}\right) \mathrm{d}u$$

and since the integrand is bounded, we exchange differentiation and integration to compute

$$\begin{aligned}\partial^\kappa f(x) &= \frac{1}{\sigma^{d+|\kappa|}} \int_{\mathbb{R}^d} g(u) \prod_{j=1}^d \partial^{\kappa_j} \phi\left(\frac{u_j - x_j}{\sigma}\right) \mathrm{d}u = \left(\frac{-1}{\sigma}\right)^{|\kappa|} \int_{\mathbb{R}^d} g(x+\sigma u) \prod_{j=1}^d \partial^{\kappa_j} \phi(u_j)\, \mathrm{d}u \\ &= \left(\frac{-1}{\sigma}\right)^{|\kappa|} \mathbb{E}\left[g(x+\sigma Z) \prod_{j=1}^d \frac{\partial^{\kappa_j} \phi(Z_j)}{\phi(Z_j)} \right], \end{aligned} \tag{C.1}$$

where $Z \sim \mathcal{N}(0, I_d)$. Recalling that $|g(x)| \le 1$ and applying the Cauchy–Schwarz inequality,

$$|\partial^\kappa f(x)| \le \frac{1}{\sigma^{|\kappa|}} \prod_{j=1}^d \mathbb{E}\left[\left(\frac{\partial^{\kappa_j} \phi(Z_j)}{\phi(Z_j)} \right)^2 \right]^{1/2} \le \frac{1}{\sigma^{|\kappa|}} \prod_{j=1}^d \sqrt{\kappa_j!} = \frac{\sqrt{\kappa!}}{\sigma^{|\kappa|}},$$

as the expected square of the Hermite polynomial of degree κ_j against the standard Gaussian measure is $\kappa_j!$. By the reverse triangle inequality, $|g(x+ty) - g(x)| \le t\|y\|_p/\eta$, so by (C.1),

$$\begin{aligned} &\left| \sum_{|\kappa|=k} \frac{y^\kappa}{\kappa!} \big(\partial^\kappa f(x+ty) - \partial^\kappa f(x) \big) \right| \\ &\quad = \left| \sum_{|\kappa|=k} \frac{y^\kappa}{\kappa!} \frac{1}{\sigma^{|\kappa|}} \mathbb{E}\left[\big(g(x+ty+\sigma Z) - g(x+\sigma Z) \big) \prod_{j=1}^d \frac{\partial^{\kappa_j} \phi(Z_j)}{\phi(Z_j)} \right] \right| \\ &\quad \le \frac{t\|y\|_p}{\sigma^k \eta} \, \mathbb{E}\left[\left| \sum_{|\kappa|=k} \frac{y^\kappa}{\kappa!} \prod_{j=1}^d \frac{\partial^{\kappa_j} \phi(Z_j)}{\phi(Z_j)} \right| \right]. \end{aligned}$$

Therefore, by the Cauchy–Schwarz inequality,

$$\begin{aligned} \left(\sum_{|\kappa|=k} \frac{y^\kappa}{\kappa!} \big(\partial^\kappa f(x+ty) - \partial^\kappa f(x) \big) \right)^2 &\le \frac{t^2 \|y\|_p^2}{\sigma^{2k} \eta^2} \, \mathbb{E}\left[\left(\sum_{|\kappa|=k} \frac{y^\kappa}{\kappa!} \prod_{j=1}^d \frac{\partial^{\kappa_j} \phi(Z_j)}{\phi(Z_j)} \right)^2 \right] \\ &= \frac{t^2 \|y\|_p^2}{\sigma^{2k} \eta^2} \sum_{|\kappa|=k} \sum_{|\kappa'|=k} \frac{y^{\kappa+\kappa'}}{\kappa! \kappa'!} \prod_{j=1}^d \mathbb{E}\left[\frac{\partial^{\kappa_j} \phi(Z_j)}{\phi(Z_j)} \frac{\partial^{\kappa'_j} \phi(Z_j)}{\phi(Z_j)} \right]. \end{aligned}$$

Orthogonality of Hermite polynomials gives zero if $\kappa_j \ne \kappa'_j$. By the multinomial theorem,

$$\begin{aligned} \left| f(x+y) - \sum_{|\kappa|=0}^k \frac{1}{\kappa!} \partial^\kappa f(x) y^\kappa \right| &\le \frac{\|y\|_p}{\sigma^k \eta} \left(\sum_{|\kappa|=k} \frac{y^{2\kappa}}{\kappa!} \right)^{1/2} \le \frac{\|y\|_p}{\sigma^k \eta \sqrt{k!}} \left(\sum_{|\kappa|=k} \frac{k!}{\kappa!} y^{2\kappa} \right)^{1/2} \\ &\le \frac{\|y\|_p \|y\|_2^k}{\sigma^k \eta \sqrt{k!}}. \end{aligned}$$

For the final result, since $f(x) = \mathbb{E}\left[g(x + \sigma Z)\right]$ and $\mathbb{I}\{x \in A^{\eta}\} \leq g(x) \leq \mathbb{I}\{x \in A^{2\eta}\}$,

$$\begin{aligned}
f(x) &\leq \mathbb{P}\left(x + \sigma Z \in A^{2\eta}\right) \\
&\leq \mathbb{P}\left(\|Z\|_p > \frac{\eta}{\sigma}\right) + \mathbb{I}\left\{x \in A^{3\eta}\right\} \mathbb{P}\left(\|Z\|_p \leq \frac{\eta}{\sigma}\right) = \varepsilon + (1-\varepsilon)\mathbb{I}\left\{x \in A^{3\eta}\right\}, \\
f(x) &\geq \mathbb{P}\left(x + \sigma Z \in A^{\eta}\right) \geq \mathbb{I}\left\{x \in A\right\} \mathbb{P}\left(\|Z\|_p \leq \frac{\eta}{\sigma}\right) = (1-\varepsilon)\mathbb{I}\left\{x \in A\right\}. \qquad \square
\end{aligned}$$

We provide a useful Gaussian inequality in Lemma C.1.3 which helps bound the $\beta_{\infty,k}$ moment terms appearing in several places throughout the analysis.

Lemma C.1.3 (A useful Gaussian inequality)
Let $X \sim \mathcal{N}(0, \Sigma)$ where $\sigma_j^2 = \Sigma_{jj} \leq \sigma^2$ for all $1 \leq j \leq d$. Then

$$\mathbb{E}\left[\|X\|_2^2 \|X\|_\infty\right] \leq 4\sigma\sqrt{\log 2d} \sum_{j=1}^{d} \sigma_j^2 \quad \textit{and} \quad \mathbb{E}\left[\|X\|_2^3 \|X\|_\infty\right] \leq 8\sigma\sqrt{\log 2d} \left(\sum_{j=1}^{d} \sigma_j^2\right)^{3/2}.$$

Proof (Lemma C.1.3)
By Cauchy–Schwarz, with $k \in \{2, 3\}$, we have $\mathbb{E}\left[\|X\|_2^k \|X\|_\infty\right] \leq \mathbb{E}\big[\|X\|_2^{2k}\big]^{1/2} \mathbb{E}\big[\|X\|_\infty^2\big]^{1/2}$.
For the first term, by Hölder's inequality and the even moments of the normal distribution,

$$\begin{aligned}
\mathbb{E}\big[\|X\|_2^4\big] &= \mathbb{E}\left[\left(\sum_{j=1}^{d} X_j^2\right)^2\right] = \sum_{j=1}^{d}\sum_{k=1}^{d} \mathbb{E}\big[X_j^2 X_k^2\big] \leq \left(\sum_{j=1}^{d} \mathbb{E}\big[X_j^4\big]^{\frac{1}{2}}\right)^2 = 3\left(\sum_{j=1}^{d} \sigma_j^2\right)^2, \\
\mathbb{E}\big[\|X\|_2^6\big] &= \sum_{j=1}^{d}\sum_{k=1}^{d}\sum_{l=1}^{d} \mathbb{E}\big[X_j^2 X_k^2 X_l^2\big] \leq \left(\sum_{j=1}^{d} \mathbb{E}\big[X_j^6\big]^{\frac{1}{3}}\right)^3 = 15\left(\sum_{j=1}^{d} \sigma_j^2\right)^3.
\end{aligned}$$

For the second term, by Jensen's inequality and the χ^2 moment generating function,

$$\mathbb{E}\big[\|X\|_\infty^2\big] = \mathbb{E}\left[\max_{1 \leq j \leq d} X_j^2\right] \leq 4\sigma^2 \log \sum_{j=1}^{d} \mathbb{E}\left[e^{X_j^2/(4\sigma^2)}\right] \leq 4\sigma^2 \log \sum_{j=1}^{d} \sqrt{2} \leq 4\sigma^2 \log 2d. \qquad \square$$

We provide an ℓ^p-norm tail probability bound for Gaussian variables in Lemma C.1.4, motivating the definition of the term $\phi_p(d)$.

Lemma C.1.4 (Gaussian ℓ^p-norm bound)

Let $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{d\times d}$ is a positive semi-definite matrix. Then we have that $\mathbb{E}\left[\|X\|_p\right] \leq \phi_p(d) \max_{1\leq j\leq d} \sqrt{\Sigma_{jj}}$ with $\phi_p(d) = \sqrt{pd^{2/p}}$ for $p \in [1, \infty)$ and $\phi_\infty(d) = \sqrt{2\log 2d}$.

Proof (Lemma C.1.4)

For $p \in [1, \infty)$, as each X_j is Gaussian, we have $\left(\mathbb{E}\left[|X_j|^p\right]\right)^{1/p} \leq \sqrt{p\,\mathbb{E}[X_j^2]} = \sqrt{p\Sigma_{jj}}$. So

$$\mathbb{E}\left[\|X\|_p\right] \leq \left(\sum_{j=1}^d \mathbb{E}\left[|X_j|^p\right]\right)^{1/p} \leq \left(\sum_{j=1}^d p^{p/2}\Sigma_{jj}^{p/2}\right)^{1/p} \leq \sqrt{pd^{2/p}} \max_{1\leq j\leq d} \sqrt{\Sigma_{jj}}$$

by Jensen's inequality. For $p = \infty$, with $\sigma^2 = \max_j \Sigma_{jj}$, for $t > 0$,

$$\mathbb{E}\left[\|X\|_\infty\right] \leq t \log \sum_{j=1}^d \mathbb{E}\left[e^{|X_j|/t}\right] \leq t \log \sum_{j=1}^d \mathbb{E}\left[2e^{X_j/t}\right] \leq t \log\left(2de^{\sigma^2/(2t^2)}\right) \leq t\log 2d + \frac{\sigma^2}{2t},$$

again by Jensen's inequality. Setting $t = \frac{\sigma}{\sqrt{2\log 2d}}$ gives $\mathbb{E}\left[\|X\|_\infty\right] \leq \sigma\sqrt{2\log 2d}$. □

We give a Gaussian–Gaussian ℓ^p-norm approximation as Lemma C.1.5, useful for ensuring approximations remain valid upon substituting an estimator for the true variance matrix.

Lemma C.1.5 (Gaussian–Gaussian approximation in ℓ^p-norm)

Let $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$ be positive semi-definite and take $Z \sim \mathcal{N}(0, I_d)$. For $p \in [1, \infty]$ we have

$$\mathbb{P}\left(\left\|\left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right) Z\right\|_p > t\right) \leq 2d \exp\left(\frac{-t^2}{2d^{2/p}\left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_2^2}\right).$$

Proof (Lemma C.1.5)

Let $\Sigma \in \mathbb{R}^{d\times d}$ be positive semi-definite and write $\sigma_j^2 = \Sigma_{jj}$. For $p \in [1, \infty)$ by a union bound and Gaussian tail probabilities,

$$\begin{aligned}\mathbb{P}\left(\left\|\Sigma^{1/2}Z\right\|_p > t\right) &= \mathbb{P}\left(\sum_{j=1}^d \left|\left(\Sigma^{1/2}Z\right)_j\right|^p > t^p\right) \leq \sum_{j=1}^d \mathbb{P}\left(\left|\left(\Sigma^{1/2}Z\right)_j\right|^p > \frac{t^p\sigma_j^p}{\|\sigma\|_p^p}\right) \\ &= \sum_{j=1}^d \mathbb{P}\left(|\sigma_j Z_j|^p > \frac{t^p\sigma_j^p}{\|\sigma\|_p^p}\right) = \sum_{j=1}^d \mathbb{P}\left(|Z_j| > \frac{t}{\|\sigma\|_p}\right) \leq 2d\,\exp\left(\frac{-t^2}{2\|\sigma\|_p^2}\right).\end{aligned}$$

The same result holds for $p = \infty$ since

$$\mathbb{P}\left(\left\|\Sigma^{1/2}Z\right\|_\infty > t\right) = \mathbb{P}\left(\max_{1\le j\le d}\left|\left(\Sigma^{1/2}Z\right)_j\right| > t\right) \le \sum_{j=1}^d \mathbb{P}\left(\left|\left(\Sigma^{1/2}Z\right)_j\right| > t\right)$$
$$= \sum_{j=1}^d \mathbb{P}\left(|\sigma_j Z_j| > t\right) \le 2\sum_{j=1}^d \exp\left(\frac{-t^2}{2\sigma_j^2}\right) \le 2d\exp\left(\frac{-t^2}{2\|\sigma\|_\infty^2}\right).$$

Now we apply this to the matrix $\Sigma = \left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right)^2$. For $p \in [1,\infty)$,

$$\|\sigma\|_p^p = \sum_{j=1}^d (\Sigma_{jj})^{p/2} = \sum_{j=1}^d \left(\left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right)^2\right)_{jj}^{p/2} \le d\max_{1\le j\le d}\left(\left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right)^2\right)_{jj}^{p/2}$$
$$\le d\left\|\left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right)^2\right\|_2^{p/2} = d\left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_2^p$$

Similarly, for $p = \infty$ we have

$$\|\sigma\|_\infty = \max_{1\le j\le d}(\Sigma_{jj})^{1/2} = \max_{1\le j\le d}\left(\left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right)^2\right)_{jj}^{1/2} \le \left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_2.$$

Thus for all $p \in [1,\infty]$ we have $\|\sigma\|_p \le d^{1/p}\left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_2$, with $d^{1/\infty} = 1$. Hence

$$\mathbb{P}\left(\left\|\left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right)Z\right\|_p > t\right) \le 2d\exp\left(\frac{-t^2}{2\|\sigma\|_p^2}\right) \le 2d\exp\left(\frac{-t^2}{2d^{2/p}\left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_2^2}\right). \quad \square$$

We give a variance bound and an exponential inequality for α-mixing variables.

Lemma C.1.6 (Variance bounds for α-mixing random variables)

Let $X_1, \ldots, X_n$ be real-valued α-mixing random variables with mixing coefficients $\alpha(j)$. Then

(i) If for constants M_i we have $|X_i| \le M_i$ a.s. then

$$\mathrm{Var}\left[\sum_{i=1}^n X_i\right] \le 4\sum_{j=1}^\infty \alpha(j)\sum_{i=1}^n M_i^2.$$

(ii) If $\alpha(j) \le e^{-2j/C_\alpha}$ then for any $r > 2$ there is a constant C_r depending only on r with

$$\mathrm{Var}\left[\sum_{i=1}^n X_i\right] \le C_r C_\alpha \sum_{i=1}^n \mathbb{E}\left[|X_i|^r\right]^{2/r}.$$

Proof (Lemma C.1.6)

Define $\alpha^{-1}(t) = \inf\{j \in \mathbb{N} : \alpha(j) \le t\}$ and $Q_i(t) = \inf\{s \in \mathbb{R} : \mathbb{P}(|X_i| > s) \le t\}$. By Corollary 1.1 in Rio (2017) and Hölder's inequality for $r > 2$,

$$\begin{aligned}\operatorname{Var}\left[\sum_{i=1}^n X_i\right] &\le 4\sum_{i=1}^n \int_0^1 \alpha^{-1}(t) Q_i(t)^2 \,\mathrm{d}t \\ &\le 4\sum_{i=1}^n \left(\int_0^1 \alpha^{-1}(t)^{\frac{r}{r-2}} \,\mathrm{d}t\right)^{\frac{r-2}{r}} \left(\int_0^1 |Q_i(t)|^r \,\mathrm{d}t\right)^{\frac{2}{r}} \mathrm{d}t.\end{aligned}$$

Now note that if $U \sim \mathrm{Unif}[0,1]$ then $Q_i(U)$ has the same distribution as X_i. Therefore

$$\operatorname{Var}\left[\sum_{i=1}^n X_i\right] \le 4\left(\int_0^1 \alpha^{-1}(t)^{\frac{r}{r-2}} \,\mathrm{d}t\right)^{\frac{r-2}{r}} \sum_{i=1}^n \mathbb{E}[|X_i|^r]^{\frac{2}{r}}.$$

If $\alpha(j) \le e^{-2j/C_\alpha}$ then $\alpha^{-1}(t) \le \frac{-C_\alpha \log t}{2}$ so, for some constant C_r depending only on r,

$$\operatorname{Var}\left[\sum_{i=1}^n X_i\right] \le 2C_\alpha \left(\int_0^1 (-\log t)^{\frac{r}{r-2}} \,\mathrm{d}t\right)^{\frac{r-2}{r}} \sum_{i=1}^n \mathbb{E}[|X_i|^r]^{\frac{2}{r}} \le C_r C_\alpha \sum_{i=1}^n \mathbb{E}[|X_i|^r]^{\frac{2}{r}}.$$

Alternatively, if for constants M_i we have $|X_i| \le M_i$ a.s. then

$$\operatorname{Var}\left[\sum_{i=1}^n X_i\right] \le 4\int_0^1 \alpha^{-1}(t) \,\mathrm{d}t \sum_{i=1}^n M_i^2 \le 4\sum_{j=1}^\infty \alpha(j) \sum_{i=1}^n M_i^2. \qquad \square$$

Lemma C.1.7 (Exponential concentration inequalities for α-mixing random variables)
Let $X_1, \dots, X_n$ be zero-mean real-valued variables with α-mixing coefficients $\alpha(j) \le e^{-2j/C_\alpha}$.

(i) Suppose $|X_i| \le M$ a.s. for $1 \le i \le n$. Then for all $t > 0$ there is a constant C_1 with

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| > C_1 M\big(\sqrt{nt} + (\log n)(\log\log n)t\big)\right) \le C_1 e^{-t}.$$

(ii) If further $\sum_{j=1}^n |\operatorname{Cov}[X_i, X_j]| \le \sigma^2$, then for all $t > 0$ there is a constant C_2 with

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \ge C_2\big((\sigma\sqrt{n} + M)\sqrt{t} + M(\log n)^2 t\big)\right) \le C_2 e^{-t}.$$

Proof (Lemma C.1.7)

(i) By Theorem 1 in Merlevède, Peligrad, and Rio (2009),

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| > t\right) \leq \exp\left(-\frac{C_1 t^2}{nM^2 + Mt(\log n)(\log\log n)}\right).$$

Replace t by $M\sqrt{nt} + M(\log n)(\log\log n)t$.

(ii) By Theorem 2 in Merlevède et al. (2009),

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| > t\right) \leq \exp\left(-\frac{C_2 t^2}{n\sigma^2 + M^2 + Mt(\log n)^2}\right).$$

Replace t by $\sigma\sqrt{n}\sqrt{t} + M\sqrt{t} + M(\log n)^2 t$. □

C.1.2 Main results

To establish Theorem 4.2.1, we first give the analogous result for martingales as Lemma C.1.8. Our approach is similar to that used in modern versions of Yurinskii's coupling for independent data, as in Theorem 1 in Le Cam (1988) and Theorem 10 in Chapter 10 of Pollard (2002). The proof of Lemma C.1.8 relies on constructing a "modified" martingale, which is close to the original martingale, but which has an $\mathcal{H}_0$-measurable terminal quadratic variation.

Lemma C.1.8 (Strong approximation for vector-valued martingales)
Let $X_1, \ldots, X_n$ be $\mathbb{R}^d$-valued square-integrable random vectors adapted to a countably generated filtration $\mathcal{H}_0, \ldots, \mathcal{H}_n$. Suppose that $\mathbb{E}[X_i \mid \mathcal{H}_{i-1}] = 0$ for all $1 \leq i \leq n$ and define $S = \sum_{i=1}^n X_i$. Let $V_i = \mathrm{Var}[X_i \mid \mathcal{H}_{i-1}]$ and $\Omega = \sum_{i=1}^n V_i - \Sigma$ where Σ is a positive semi-definite $\mathcal{H}_0$-measurable $d \times d$ random matrix. For each $\eta > 0$ and $p \in [1, \infty]$ there is $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$ with

$$\mathbb{P}\big(\|S - T\|_p > 5\eta\big) \leq \inf_{t>0}\left\{2\mathbb{P}\big(\|Z\|_p > t\big) + \min\left\{\frac{\beta_{p,2}t^2}{\eta^3}, \frac{\beta_{p,3}t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3}\right\}\right\}$$
$$+ \inf_{M \succeq 0}\big\{2\gamma(M) + \delta_p(M,\eta) + \varepsilon_p(M,\eta)\big\},$$

where the second infimum is over all positive semi-definite $d \times d$ non-random matrices, and

$$\beta_{p,k} = \sum_{i=1}^{n} \mathbb{E}\left[\|X_i\|_2^k \|X_i\|_p + \|V_i^{1/2} Z_i\|_2^k \|V_i^{1/2} Z_i\|_p\right], \qquad \gamma(M) = \mathbb{P}\big(\Omega \not\preceq M\big),$$

$$\delta_p(M,\eta) = \mathbb{P}\left(\left\|\left((\Sigma+M)^{1/2} - \Sigma^{1/2}\right) Z\right\|_p \geq \eta\right), \qquad \pi_3 = \sum_{i=1}^{n+m} \sum_{|\kappa|=3} \mathbb{E}\Big[\big|\mathbb{E}\left[X_i^\kappa \mid \mathcal{H}_{i-1}\right]\big|\Big],$$

$$\varepsilon_p(M,\eta) = \mathbb{P}\left(\left\|(M-\Omega)^{1/2} Z\right\|_p \geq \eta,\ \Omega \preceq M\right),$$

for $k \in \{2,3\}$, with $Z, Z_1, \ldots, Z_n$ i.i.d. standard Gaussian on $\mathbb{R}^d$ independent of $\mathcal{H}_n$.

Proof (Lemma C.1.8)

Part 1: constructing a modified martingale

Take $M \succeq 0$ a fixed positive semi-definite $d \times d$ matrix. We start by constructing a new martingale based on S whose quadratic variation is $\Sigma + M$. Take $m \geq 1$ and define

$$H_k = \Sigma + M - \sum_{i=1}^{k} V_i, \qquad \tau = \sup\big\{k \in \{0,1,\ldots,n\} : H_k \succeq 0\big\},$$

$$\tilde{X}_i = X_i \mathbb{I}\{i \leq \tau\} + \frac{1}{\sqrt{m}} H_\tau^{1/2} Z_i \mathbb{I}\{n+1 \leq i \leq n+m\}, \qquad \tilde{S} = \sum_{i=1}^{n+m} \tilde{X}_i,$$

where $Z_{n+1}, \ldots, Z_{n+m}$ is an i.i.d. sequence of standard Gaussian vectors in $\mathbb{R}^d$ independent of $\mathcal{H}_n$, noting that $H_0 = \Sigma + M \succeq 0$ a.s. Define the filtration $\tilde{\mathcal{H}}_0, \ldots, \tilde{\mathcal{H}}_{n+m}$, where $\tilde{\mathcal{H}}_i = \mathcal{H}_i$ for $0 \leq i \leq n$ and is the σ-algebra generated by $\mathcal{H}_n$ and $Z_{n+1}, \ldots, Z_i$ for $n+1 \leq i \leq n+m$. Observe that τ is a stopping time with respect to $\tilde{\mathcal{H}}_i$ because $H_{i+1} - H_i = -V_{i+1} \preceq 0$ almost surely, so $\{\tau \leq i\} = \{H_{i+1} \not\succeq 0\}$ for $0 \leq i < n$. This depends only on $V_1, \ldots, V_{i+1}$ and Σ which are $\tilde{\mathcal{H}}_i$-measurable. Similarly, $\{\tau = n\} = \{H_n \succeq 0\} \in \tilde{\mathcal{H}}_{n-1}$. Let $\tilde{V}_i = V_i \mathbb{I}\{i \leq \tau\}$ for $1 \leq i \leq n$ and $\tilde{V}_i = H_\tau / m$ for $n+1 \leq i \leq n+m$. Note that $\tilde{X}_i$ is $\tilde{\mathcal{H}}_i$-measurable and $\tilde{V}_i$ is $\tilde{\mathcal{H}}_{i-1}$-measurable. Further, $\mathbb{E}\left[\tilde{X}_i \mid \tilde{\mathcal{H}}_{i-1}\right] = 0$ and $\mathbb{E}\left[\tilde{X}_i \tilde{X}_i^\mathsf{T} \mid \tilde{\mathcal{H}}_{i-1}\right] = \tilde{V}_i$.

Part 2: bounding the difference between the original and modified martingales

By the triangle inequality,

$$\|S - \tilde{S}\|_p \le \left\| \sum_{i=\tau+1}^{n} X_i \right\|_p + \left\| \frac{1}{\sqrt{m}} \sum_{i=n+1}^{m} H_\tau^{1/2} Z_i \right\|_p .$$

The first term on the right vanishes on $\{\tau = n\} = \{H_n \succeq 0\} = \{\Omega \preceq M\}$. For the second term, note that $\frac{1}{\sqrt{m}} \sum_{i=n+1}^{m} H_\tau^{1/2} Z_i$ is distributed as $H_\tau^{1/2} Z$, where Z is an independent standard Gaussian variable. Also $\mathbb{P}\big(\|H_\tau^{1/2} Z\|_p > \eta\big) \le \mathbb{P}\big(\|H_n^{1/2} Z\|_p > \eta,\, \Omega \preceq M\big) + \mathbb{P}\big(\Omega \not\preceq M\big)$, so

$$\mathbb{P}\big(\|S - \tilde{S}\|_p > \eta\big) \le 2\mathbb{P}\big(\Omega \not\preceq M\big) + \mathbb{P}\big(\|(M-\Omega)^{1/2} Z\|_p > \eta,\, \Omega \preceq M\big) = 2\gamma(M) + \varepsilon_p(M, \eta).$$

Part 3: strong approximation of the modified martingale

Let $\tilde{Z}_1, \dots, \tilde{Z}_{n+m}$ be i.i.d. $\mathcal{N}(0, I_d)$ and independent of $\tilde{\mathcal{H}}_{n+m}$. Define $\check{X}_i = \tilde{V}_i^{1/2} \tilde{Z}_i$ and $\check{S} = \sum_{i=1}^{n+m} \check{X}_i$. Fix a Borel set $A \subseteq \mathbb{R}^d$ and $\sigma, \eta > 0$ and let $f = f_{A\eta\sigma}$ be the function defined in Lemma C.1.2. By the Lindeberg method, write the telescoping sum

$$\mathbb{E}\Big[f\big(\tilde{S}\big) - f\big(\check{S}\big) \mid \mathcal{H}_0\Big] = \sum_{i=1}^{n+m} \mathbb{E}\Big[f\big(Y_i + \tilde{X}_i\big) - f\big(Y_i + \check{X}_i\big) \mid \mathcal{H}_0\Big]$$

where $Y_i = \sum_{j=1}^{i-1} \tilde{X}_j + \sum_{j=i+1}^{n+m} \check{X}_j$. By Lemma C.1.2 we have for $k \ge 0$

$$\begin{aligned} &\left| \mathbb{E}\big[f(Y_i + \tilde{X}_i) - f(Y_i + \check{X}_i) \mid \mathcal{H}_0\big] - \sum_{|\kappa|=0}^{k} \frac{1}{\kappa!} \mathbb{E}\left[\partial^\kappa f(Y_i)\left(\tilde{X}_i^\kappa - \check{X}_i^\kappa\right) \mid \mathcal{H}_0\right] \right| \\ &\quad \le \frac{1}{\sigma^k \eta \sqrt{k!}} \mathbb{E}\left[\|\tilde{X}_i\|_p \|\tilde{X}_i\|_2^k + \|\check{X}_i\|_p \|\check{X}_i\|_2^k \mid \mathcal{H}_0\right]. \end{aligned}$$

With $k \in \{2, 3\}$, we bound each summand. With $|\kappa| = 0$ we have $\tilde{X}_i^\kappa = \check{X}_i^\kappa$, so consider $|\kappa| = 1$. Noting that $\sum_{i=1}^{n+m} \tilde{V}_i = \Sigma + M$, define

$$\tilde{Y}_i = \sum_{j=1}^{i-1} \tilde{X}_j + \tilde{Z}_i \left(\sum_{j=i+1}^{n+m} \tilde{V}_j \right)^{1/2} = \sum_{j=1}^{i-1} \tilde{X}_j + \tilde{Z}_i \left(\Sigma + M - \sum_{j=1}^{i} \tilde{V}_j \right)^{1/2}$$

and let $\check{\mathcal{H}}_i$ be the σ-algebra generated by $\tilde{\mathcal{H}}_{i-1}$ and $\tilde{Z}_i$. Note that $\tilde{Y}_i$ is $\check{\mathcal{H}}_i$-measurable and that Y_i and $\tilde{Y}_i$ have the same distribution conditional on $\tilde{\mathcal{H}}_{n+m}$. So

$$\begin{aligned}
\sum_{|\kappa|=1} \frac{1}{\kappa!}\mathbb{E}\left[\partial^\kappa f(Y_i)\big(\tilde{X}_i^\kappa - \check{X}_i^\kappa\big) \mid \mathcal{H}_0\right] &= \mathbb{E}\left[\nabla f(Y_i)^\mathsf{T}\big(\tilde{X}_i - \tilde{V}_i^{1/2}\tilde{Z}_i\big) \mid \mathcal{H}_0\right] \\
&= \mathbb{E}\left[\nabla f(\tilde{Y}_i)^\mathsf{T}\tilde{X}_i \mid \mathcal{H}_0\right] - \mathbb{E}\left[\nabla f(Y_i)^\mathsf{T}\tilde{V}_i^{1/2}\tilde{Z}_i \mid \mathcal{H}_0\right] \\
&= \mathbb{E}\left[\nabla f(\tilde{Y}_i)^\mathsf{T}\mathbb{E}\left[\tilde{X}_i \mid \check{\mathcal{H}}_i\right] \mid \mathcal{H}_0\right] - \mathbb{E}\left[\tilde{Z}_i\right]\mathbb{E}\left[\nabla f(Y_i)^\mathsf{T}\tilde{V}_i^{1/2} \mid \mathcal{H}_0\right] \\
&= \mathbb{E}\left[\nabla f(\tilde{Y}_i)^\mathsf{T}\mathbb{E}\left[\tilde{X}_i \mid \tilde{\mathcal{H}}_{i-1}\right] \mid \mathcal{H}_0\right] - 0 = 0.
\end{aligned}$$

Next, if $|\kappa| = 2$ then

$$\begin{aligned}
&\sum_{|\kappa|=2} \frac{1}{\kappa!}\mathbb{E}\left[\partial^\kappa f(Y_i)\left(\tilde{X}_i^\kappa - \check{X}_i^\kappa\right) \mid \mathcal{H}_0\right] \\
&\quad= \frac{1}{2}\mathbb{E}\left[\tilde{X}_i^\mathsf{T}\nabla^2 f(Y_i)\tilde{X}_i - \tilde{Z}_i^\mathsf{T}\tilde{V}_i^{1/2}\nabla^2 f(Y_i)\tilde{V}_i^{1/2}\tilde{Z}_i \mid \mathcal{H}_0\right] \\
&\quad= \frac{1}{2}\mathbb{E}\left[\mathbb{E}\left[\operatorname{Tr}\nabla^2 f(\tilde{Y}_i)\tilde{X}_i\tilde{X}_i^\mathsf{T} \mid \check{\mathcal{H}}_i\right] \mid \mathcal{H}_0\right] - \frac{1}{2}\mathbb{E}\left[\operatorname{Tr}\tilde{V}_i^{1/2}\nabla^2 f(Y_i)\tilde{V}_i^{1/2} \mid \mathcal{H}_0\right]\mathbb{E}\left[\tilde{Z}_i\tilde{Z}_i^\mathsf{T}\right] \\
&\quad= \frac{1}{2}\mathbb{E}\left[\operatorname{Tr}\nabla^2 f(Y_i)\mathbb{E}\left[\tilde{X}_i\tilde{X}_i^\mathsf{T} \mid \tilde{\mathcal{H}}_{i-1}\right] \mid \mathcal{H}_0\right] - \frac{1}{2}\mathbb{E}\left[\operatorname{Tr}\nabla^2 f(Y_i)\tilde{V}_i \mid \mathcal{H}_0\right] = 0.
\end{aligned}$$

Finally, if $|\kappa| = 3$, then since $\check{X}_i \sim \mathcal{N}(0, \tilde{V}_i)$ conditional on $\tilde{\mathcal{H}}_{n+m}$, we have by symmetry of the Gaussian distribution and Lemma C.1.2,

$$\begin{aligned}
&\left|\sum_{|\kappa|=3} \frac{1}{\kappa!}\mathbb{E}\left[\partial^\kappa f(Y_i)\left(\tilde{X}_i^\kappa - \check{X}_i^\kappa\right) \mid \mathcal{H}_0\right]\right| \\
&\quad= \left|\sum_{|\kappa|=3} \frac{1}{\kappa!}\left(\mathbb{E}\left[\partial^\kappa f(\tilde{Y}_i)\mathbb{E}\left[\tilde{X}_i^\kappa \mid \check{\mathcal{H}}_i\right] \mid \mathcal{H}_0\right] - \mathbb{E}\left[\partial^\kappa f(Y_i)\,\mathbb{E}\left[\check{X}_i^\kappa \mid \tilde{\mathcal{H}}_{n+m}\right] \mid \mathcal{H}_0\right]\right)\right| \\
&\quad= \left|\sum_{|\kappa|=3} \frac{1}{\kappa!}\mathbb{E}\left[\partial^\kappa f(Y_i)\,\mathbb{E}\left[\tilde{X}_i^\kappa \mid \tilde{\mathcal{H}}_{i-1}\right] \mid \mathcal{H}_0\right]\right| \le \frac{1}{\sigma^3}\sum_{|\kappa|=3}\mathbb{E}\left[\left|\mathbb{E}\left[\tilde{X}_i^\kappa \mid \tilde{\mathcal{H}}_{i-1}\right]\right| \mid \mathcal{H}_0\right].
\end{aligned}$$

Combining these and summing over i with $k = 2$ shows

$$\mathbb{E}\left[f(\tilde{S}) - f(\check{S}) \mid \mathcal{H}_0\right] \le \frac{1}{\sigma^2\eta\sqrt{2}}\sum_{i=1}^{n+m}\mathbb{E}\left[\|\tilde{X}_i\|_p\|\tilde{X}_i\|_2^2 + \|\check{X}_i\|_p\|\check{X}_i\|_2^2 \mid \mathcal{H}_0\right]$$

On the other hand, taking $k = 3$ gives

$$\mathbb{E}\left[f(\tilde{S}) - f(\check{S}) \mid \mathcal{H}_0\right] \leq \frac{1}{\sigma^3\eta\sqrt{6}} \sum_{i=1}^{n+m} \mathbb{E}\left[\|\tilde{X}_i\|_p \|\tilde{X}_i\|_2^3 + \|\check{X}_i\|_p \|\check{X}_i\|_2^3 \mid \mathcal{H}_0\right] + \frac{1}{\sigma^3} \sum_{i=1}^{n+m} \sum_{|\kappa|=3} \mathbb{E}\left[\left|\mathbb{E}\left[\tilde{X}_i^\kappa \mid \tilde{\mathcal{H}}_{i-1}\right]\right| \mid \mathcal{H}_0\right].$$

For $1 \leq i \leq n$ we have $\|\tilde{X}_i\| \leq \|X_i\|$ and $\|\check{X}_i\| \leq \|V_i^{1/2}\tilde{Z}_i\|$. For $n+1 \leq i \leq n+m$ we have $\tilde{X}_i = H_\tau^{1/2} Z_i/\sqrt{m}$ and $\check{X}_i = H_\tau^{1/2}\tilde{Z}_i/\sqrt{m}$ which are equal in distribution given $\mathcal{H}_0$. So with

$$\tilde{\beta}_{p,k} = \sum_{i=1}^{n} \mathbb{E}\left[\|X_i\|_p \|X_i\|_2^k + \|V_i^{1/2} Z_i\|_p \|V_i^{1/2} Z_i\|_2^k \mid \mathcal{H}_0\right],$$

we have, since $k \in \{2, 3\}$,

$$\sum_{i=1}^{n+m} \mathbb{E}\left[\|\tilde{X}_i\|_p \|\tilde{X}_i\|_2^k + \|\check{X}_i\|_p \|\check{X}_i\|_2^k \mid \mathcal{H}_0\right] \leq \tilde{\beta}_{p,k} + \frac{2}{\sqrt{m}} \mathbb{E}\left[\|H_\tau^{1/2} Z\|_p \|H_\tau^{1/2} Z\|_2^k \mid \mathcal{H}_0\right].$$

Since H_i is weakly decreasing under the semi-definite partial order, we have $H_\tau \preceq H_0 = \Sigma + M$ implying that $|(H_\tau)_{jj}| \leq \|\Sigma + M\|_{\max}$ and $\mathbb{E}\big[|(H_\tau^{1/2} Z)_j|^3 \mid \mathcal{H}_0\big] \leq \sqrt{8/\pi}\, \|\Sigma + M\|_{\max}^{3/2}$. Hence as $p \geq 1$ and $k \in \{2, 3\}$,

$$\begin{aligned}\mathbb{E}\left[\|H_\tau^{1/2} Z\|_p \|H_\tau^{1/2} Z\|_2^k \mid \mathcal{H}_0\right] &\leq \mathbb{E}\left[\|H_\tau^{1/2} Z\|_1^{k+1} \mid \mathcal{H}_0\right] \leq d^{k+1} \max_{1 \leq j \leq d} \mathbb{E}\left[|(H_\tau^{1/2} Z)_j|^{k+1} \mid \mathcal{H}_0\right] \\ &\leq 3d^4 \|\Sigma + M\|_{\max}^{(k+1)/2} \leq 6d^4 \|\Sigma\|_{\max}^{(k+1)/2} + 6d^4 \|M\|.\end{aligned}$$

Assuming some X_i is not identically zero so the result is non-trivial, and supposing that Σ is bounded a.s. (replacing Σ by $\Sigma \cdot \mathbb{I}\{\|\Sigma\|_{\max} \leq C\}$ for an appropriately large C if necessary), take m large enough that

$$\frac{2}{\sqrt{m}} \mathbb{E}\left[\|H_\tau^{1/2} Z\|_p \|H_\tau^{1/2} Z\|_2^k \mid \mathcal{H}_0\right] \leq \frac{1}{4} \beta_{p,k}. \tag{C.2}$$

Further, if $|\kappa| = 3$ then $\big|\mathbb{E}\big[\tilde{X}_i^\kappa \mid \tilde{\mathcal{H}}_{i-1}\big]\big| \le \big|\mathbb{E}\left[X_i^\kappa \mid \mathcal{H}_{i-1}\right]\big|$ for $1 \le i \le n$ while by symmetry of the Gaussian distribution $\mathbb{E}\left[\tilde{X}_i^\kappa \mid \tilde{\mathcal{H}}_{i-1}\right] = 0$ for $n+1 \le i \le n+m$. Hence with

$$\tilde{\pi}_3 = \sum_{i=1}^{n+m} \sum_{|\kappa|=3} \mathbb{E}\Big[\big|\mathbb{E}\left[X_i^\kappa \mid \mathcal{H}_{i-1}\right]\big| \mid \mathcal{H}_0\Big],$$

we have

$$\mathbb{E}\left[f\big(\tilde{S}\big) - f\big(\check{S}\big) \mid \mathcal{H}_0\right] \le \min\left\{\frac{3\tilde{\beta}_{p,2}}{4\sigma^2\eta} + \frac{\beta_{p,2}}{4\sigma^2\eta}, \frac{3\tilde{\beta}_{p,3}}{4\sigma^3\eta} + \frac{\beta_{p,3}}{4\sigma^3\eta} + \frac{\tilde{\pi}_3}{\sigma^3}\right\}.$$

Along with Lemma C.1.2, and with $\sigma = \eta/t$ and $\varepsilon = \mathbb{P}(\|Z\|_p > t)$, we conclude that

$$\begin{aligned}
&\mathbb{P}(\tilde{S} \in A \mid \mathcal{H}_0) = \mathbb{E}\big[\mathbb{I}\{\tilde{S} \in A\} - f(\tilde{S}) \mid \mathcal{H}_0\big] + \mathbb{E}\big[f(\tilde{S}) - f\big(\check{S}\big) \mid \mathcal{H}_0\big] + \mathbb{E}\big[f\big(\check{S}\big) \mid \mathcal{H}_0\big] \\
&\le \varepsilon\mathbb{P}(\tilde{S} \in A \mid \mathcal{H}_0) + \min\left\{\frac{3\tilde{\beta}_{p,2}}{4\sigma^2\eta} + \frac{\beta_{p,2}}{4\sigma^2\eta}, \frac{3\tilde{\beta}_{p,3}}{4\sigma^3\eta} + \frac{\beta_{p,3}}{4\sigma^3\eta} + \frac{\tilde{\pi}_3}{\sigma^3}\right\} + \varepsilon + (1-\varepsilon)\mathbb{P}\big(\check{S} \in A_p^{3\eta} \mid \mathcal{H}_0\big) \\
&\le \mathbb{P}\big(\check{S} \in A_p^{3\eta} \mid \mathcal{H}_0\big) + 2\mathbb{P}(\|Z\|_p > t) + \min\left\{\frac{3\tilde{\beta}_{p,2}t^2}{4\eta^3} + \frac{\beta_{p,2}t^2}{4\eta^3}, \frac{3\tilde{\beta}_{p,3}t^3}{4\eta^4} + \frac{\beta_{p,3}t^3}{4\eta^4} + \frac{\tilde{\pi}_3 t^3}{\eta^3}\right\}.
\end{aligned}$$

Taking a supremum and an outer expectation yields with $\beta_{p,k} = \mathbb{E}\big[\tilde{\beta}_{p,k}\big]$ and $\pi_3 = \mathbb{E}[\tilde{\pi}_3]$,

$$\begin{aligned}
&\mathbb{E}^*\left[\sup_{A \in \mathcal{B}(\mathbb{R}^d)} \Big\{\mathbb{P}(\tilde{S} \in A \mid \mathcal{H}_0) - \mathbb{P}\big(\check{S} \in A_p^{3\eta} \mid \mathcal{H}_0\big)\Big\}\right] \\
&\qquad \le 2\mathbb{P}(\|Z\|_p > t) + \min\left\{\frac{\beta_{p,2}t^2}{\eta^3}, \frac{\beta_{p,3}t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3}\right\}.
\end{aligned}$$

Finally, since $\check{S} = \sum_{i=1}^n \tilde{V}_i^{1/2}\tilde{Z}_i \sim \mathcal{N}(0, \Sigma + M)$ conditional on $\mathcal{H}_0$, the conditional Strassen theorem in Lemma C.1.1 ensures the existence of $\tilde{S}$ and $\tilde{T} \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma + M)$ such that

$$\mathbb{P}\left(\|\tilde{S} - \tilde{T}\|_p > 3\eta\right) \le \inf_{t>0}\left\{2\mathbb{P}(\|Z\|_p > t) + \min\left\{\frac{\beta_{p,2}t^2}{\eta^3}, \frac{\beta_{p,3}t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3}\right\}\right\}, \tag{C.3}$$

since the infimum is attained by continuity of $\|Z\|_p$.

Part 4: conclusion

We show how to write $\tilde{T} = (\Sigma + M)^{1/2} W$ where $W \sim \mathcal{N}(0, I_d)$ and use this representation to construct $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$. By the spectral theorem, let $\Sigma + M = U \Lambda U^{\mathsf{T}}$ where U is a $d \times d$ orthogonal random matrix and Λ is a diagonal $d \times d$ random matrix with diagonal entries satisfying $\lambda_1 \geq \cdots \geq \lambda_r > 0$ and $\lambda_{r+1} = \cdots = \lambda_d = 0$ where $r = \operatorname{rank}(\Sigma + M)$. Let Λ^+ be the Moore–Penrose pseudo-inverse of Λ (obtained by inverting its non-zero elements) and define $W = U(\Lambda^+)^{1/2} U^{\mathsf{T}} \tilde{T} + U\tilde{W}$, where the first r elements of $\tilde{W}$ are zero and the last $d - r$ elements are i.i.d. $\mathcal{N}(0,1)$ independent from $\tilde{T}$. Then, it is easy to check that $W \sim \mathcal{N}(0, I_d)$ and that $\tilde{T} = (\Sigma + M)^{1/2} W$. Now define $T = \Sigma^{1/2} W$ so

$$\mathbb{P}\big(\|T - \tilde{T}\|_p > \eta\big) = \mathbb{P}\big(\big\|\big((\Sigma + M)^{1/2} - \Sigma^{1/2}\big) W\big\|_p > \eta\big) = \delta_p(M, \eta). \tag{C.4}$$

Finally (2), (C.3), (C.4), the triangle inequality, and a union bound conclude the proof since by taking an infimum over $M \succeq 0$, and by possibly reducing the constant of $1/4$ in (C.2) to account for this infimum being potentially unattainable,

$$\begin{aligned}
\mathbb{P}\big(\|S - T\|_p > 5\eta\big) &\leq \mathbb{P}\big(\|\tilde{S} - \tilde{T}\|_p > 3\eta\big) + \mathbb{P}\big(\|S - \tilde{S}\|_p > \eta\big) + \mathbb{P}\big(\|T - \tilde{T}\|_p > \eta\big) \\
&\leq \inf_{t > 0} \left\{ 2\mathbb{P}\big(\|Z\|_p > t\big) + \min\left\{ \frac{\beta_{p,2} t^2}{\eta^3}, \frac{\beta_{p,3} t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3} \right\} \right\} \\
&\quad + \inf_{M \succeq 0} \big\{ 2\gamma(M) + \delta_p(M, \eta) + \varepsilon_p(M, \eta) \big\}. \qquad \square
\end{aligned}$$

Lemma C.1.8 and the martingale approximation immediately yield Theorem 4.2.1.

Proof (Theorem 4.2.1)

Apply Lemma C.1.8 to the martingale $\sum_{i=1}^n \tilde{X}_i$, noting that $S - \sum_{i=1}^n \tilde{X}_i = U$. $\square$

Bounding the quantities in Theorem 4.2.1 gives a user-friendly version as Proposition 4.2.1.

Proof (Proposition 4.2.1)

Set $M = \nu^2 I_d$ and bound the terms appearing the main inequality in Proposition 4.2.1.

Part 1: bounding $\mathbb{P}(\|Z\|_p > t)$

By Markov's inequality and Lemma C.1.4, we have $\mathbb{P}(\|Z\|_p > t) \leq \mathbb{E}[\|Z\|_p]/t \leq \phi_p(d)/t$.

Part 2: bounding $\gamma(M)$

With $M = \nu^2 I_d$, by Markov's inequality, $\gamma(M) = \mathbb{P}\big(\Omega \not\preceq M\big) = \mathbb{P}\big(\|\Omega\|_2 > \nu^2\big) \le \nu^{-2}\mathbb{E}[\|\Omega\|_2]$.

Part 3: bounding $\delta(M, \eta)$

By Markov's inequality and Lemma C.1.4, using $\max_j |M_{jj}| \le \|M\|_2$ for $M \succeq 0$,

$$\delta_p(M,\eta) = \mathbb{P}\left(\left\|\left((\Sigma+M)^{1/2} - \Sigma^{1/2}\right)Z\right\|_p \ge \eta\right) \le \frac{\phi_p(d)}{\eta}\mathbb{E}\left[\left\|(\Sigma+M)^{1/2} - \Sigma^{1/2}\right\|_2\right].$$

For semi-definite matrices the eigenvalue operator commutes with smooth matrix functions so

$$\|(\Sigma+M)^{1/2} - \Sigma^{1/2}\|_2 = \max_{1\le j\le d}\left|\sqrt{\lambda_j(\Sigma)+\nu^2} - \sqrt{\lambda_j(\Sigma)}\right| \le \nu$$

and hence $\delta_p(M,\eta) \le \phi_p(d)\nu/\eta$.

Part 4: bounding $\varepsilon(M, \eta)$

Note that $(M-\Omega)^{1/2}Z$ is a centered Gaussian conditional on $\mathcal{H}_n$, on the event $\{\Omega \preceq M\}$. We thus have by Markov's inequality, Lemma C.1.4, and Jensen's inequality that

$$\begin{aligned}
\varepsilon_p(M,\eta) &= \mathbb{P}\left(\left\|(M-\Omega)^{1/2}Z\right\|_p \ge \eta,\ \Omega \preceq M\right) \le \frac{1}{\eta}\mathbb{E}\left[\mathbb{I}\{\Omega\preceq M\}\mathbb{E}\left[\left\|(M-\Omega)^{1/2}Z\right\|_p \mid \mathcal{H}_n\right]\right] \\
&\le \frac{\phi_p(d)}{\eta}\mathbb{E}\left[\mathbb{I}\{\Omega\preceq M\}\max_{1\le j\le d}\sqrt{(M-\Omega)_{jj}}\right] \le \frac{\phi_p(d)}{\eta}\mathbb{E}\left[\sqrt{\|M-\Omega\|_2}\right] \\
&\le \frac{\phi_p(d)}{\eta}\mathbb{E}\left[\sqrt{\|\Omega\|_2}+\nu\right] \le \frac{\phi_p(d)}{\eta}\left(\sqrt{\mathbb{E}[\|\Omega\|_2]}+\nu\right).
\end{aligned}$$

Thus by Theorem 4.2.1 and the previous parts,

$$\begin{aligned}
\mathbb{P}\big(\|S-T\|_p > 6\eta\big) &\le \inf_{t>0}\left\{2\mathbb{P}\big(\|Z\|_p > t\big) + \min\left\{\frac{\beta_{p,2}t^2}{\eta^3}, \frac{\beta_{p,3}t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3}\right\}\right\} \\
&\quad + \inf_{M\succeq 0}\big\{2\gamma(M) + \delta_p(M,\eta) + \varepsilon_p(M,\eta)\big\} + \mathbb{P}\big(\|U\|_p > \eta\big) \\
&\le \inf_{t>0}\left\{\frac{2\phi_p(d)}{t} + \min\left\{\frac{\beta_{p,2}t^2}{\eta^3}, \frac{\beta_{p,3}t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3}\right\}\right\} \\
&\quad + \inf_{\nu>0}\left\{\frac{2\mathbb{E}\left[\|\Omega\|_2\right]}{\nu^2} + \frac{2\phi_p(d)\nu}{\eta}\right\} + \frac{\phi_p(d)\sqrt{\mathbb{E}\left[\|\Omega\|_2\right]}}{\eta} + \mathbb{P}\big(\|U\|_p > \eta\big).
\end{aligned}$$

Set $t = 2^{1/3}\phi_p(d)^{1/3}\beta_{p,2}^{-1/3}\eta$ and $\nu = \mathbb{E}[\|\Omega\|_2]^{1/3}\phi_p(d)^{-1/3}\eta^{1/3}$, then replace η with $\eta/6$ to see

$$\mathbb{P}\big(\|S-T\|_p > 6\eta\big) \leq 24\left(\frac{\beta_{p,2}\phi_p(d)^2}{\eta^3}\right)^{1/3} + 17\left(\frac{\mathbb{E}\left[\|\Omega\|_2\right]\phi_p(d)^2}{\eta^2}\right)^{1/3} + \mathbb{P}\left(\|U\|_p > \frac{\eta}{6}\right).$$

Whenever $\pi_3 = 0$ we can set $t = 2^{1/4}\phi_p(d)^{1/4}\beta_{p,3}^{-1/4}\eta$, and with ν as above we obtain

$$\mathbb{P}\big(\|S-T\|_p > \eta\big) \leq 24\left(\frac{\beta_{p,3}\phi_p(d)^3}{\eta^4}\right)^{1/4} + 17\left(\frac{\mathbb{E}\left[\|\Omega\|_2\right]\phi_p(d)^2}{\eta^2}\right)^{1/3} + \mathbb{P}\left(\|U\|_p > \frac{\eta}{6}\right). \quad \square$$

After establishing Proposition 4.2.1, Corollaries 4.2.1, 4.2.2, and 4.2.3 follow easily.

Proof (Corollary 4.2.1)

Proposition 4.2.1 with $\mathbb{P}(\|U\|_p > \frac{\eta}{6}) \leq \frac{6}{\eta}\sum_{i=1}^n c_i(\zeta_i + \zeta_{n-i+1})$. $\square$

Proof (Corollary 4.2.2)

By Proposition 4.2.1 with $U = 0$ a.s. $\square$

Proof (Corollary 4.2.3)

By Corollary 4.2.2 with $\Omega = 0$ a.s. $\square$

We conclude this section with a discussion expanding on the comments made in Remark 4.2.1 on deriving bounds in probability from Yurinskii's coupling. Consider for illustration the independent data second-order result given in Corollary 4.2.3: for each $\eta > 0$, there exists $T_n \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$ satisfying

$$\mathbb{P}\big(\|S_n - T_n\|_p > \eta\big) \leq 24\left(\frac{\beta_{p,2}\phi_p(d)^2}{\eta^3}\right)^{1/3},$$

where here we make explicit the dependence on the sample size n for clarity. The naive approach to converting this into a probability bound for $\|S_n - T_n\|_p$ is to select η to ensure the right-hand side is of order 1, arguing that the probability can then be made arbitrarily small by taking, in this case, η to be a large enough multiple of $\beta_{p,2}^{1/3}\phi_p(d)^{2/3}$. However, the somewhat subtle mistake is in neglecting the fact that the realization of the coupling variable T_n will in general depend on η, rendering the resulting bound invalid. As an explicit example

of this phenomenon, take $\eta > 1$ and suppose $\|S_n - T_n(\eta)\| = \eta$ with probability $1 - 1/\eta$ and $\|S_n - T_n(\eta)\| = n$ with probability $1/\eta$. Then $\mathbb{P}\big(\|S_n - T_n(\eta)\| > \eta\big) = 1/\eta$ but it is not true for any η that $\|S_n - T_n(\eta)\| \lesssim_{\mathbb{P}} 1$.

We propose in Remark 4.2.1 the following fix. Instead of selecting η to ensure the right-hand side is of order 1, we instead choose it so the bound converges (slowly) to zero. This is easily achieved by taking the naive and incorrect bound and multiplying by some divergent sequence R_n. The resulting inequality reads, in the case of Corollary 4.2.3 with $\eta = \beta_{p,2}^{1/3}\phi_p(d)^{2/3}R_n$,

$$\mathbb{P}\Big(\|S_n - T_n\|_p > \beta_{p,2}^{1/3}\phi_p(d)^{2/3}R_n\Big) \leq \frac{24}{R_n} \to 0.$$

We thus recover, for the price of a rate which is slower by an arbitrarily small amount, a valid upper bound in probability, as we can immediately conclude that

$$\|S_n - T_n\|_p \lesssim_{\mathbb{P}} \beta_{p,2}^{1/3}\phi_p(d)^{2/3}R_n.$$

C.1.3 Strong approximation for martingale empirical processes

We begin by presenting some calculations omitted from the main text relating to the motivating example of kernel density estimation with i.i.d. data. First, the bias is bounded as

$$\big|\mathbb{E}\big[\hat{g}(x)\big] - g(x)\big| = \left|\int_{\frac{-x}{h}}^{\frac{1-x}{h}} K(\xi)\,\mathrm{d}\xi - 1\right| \leq 2\int_{\frac{a}{h}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}}\,\mathrm{d}\xi \leq \frac{h}{a}\sqrt{\frac{2}{\pi}}e^{-\frac{a^2}{2h^2}}.$$

Next, we do the calculations necessary to apply Corollary 4.2.3. Define $k_{ij} = \frac{1}{nh}K\left(\frac{X_i - x_j}{h}\right)$ and $k_i = (k_{ij} : 1 \leq j \leq N)$. Then $\|k_i\|_\infty \leq \frac{1}{nh\sqrt{2\pi}}$ a.s. and $\mathbb{E}[\|k_i\|_2^2] \leq \frac{N}{n^2h}\int_{-\infty}^{\infty} K(\xi)^2\,\mathrm{d}\xi \leq \frac{N}{2n^2h\sqrt{\pi}}$. Let $V = \mathrm{Var}[k_i] \in \mathbb{R}^{N\times N}$, so assuming that $1/h \geq \log 2N$, by Lemma C.1.3 we bound

$$\beta_{\infty,2} = n\mathbb{E}\left[\|k_i\|_2^2\|k_i\|_\infty\right] + n\mathbb{E}\left[\|V^{1/2}Z\|_2^2\|V^{1/2}Z\|_\infty\right] \leq \frac{N}{\sqrt{8}n^2h^2\pi} + \frac{4N\sqrt{\log 2N}}{\sqrt{8}n^2h^{3/2}\pi^{3/4}} \leq \frac{N}{n^2h^2}.$$

Finally, we verify the stochastic continuity bounds. By the Lipschitz property of K, it is easy to show that for $x, x' \in \mathcal{X}$ we have $\left|\frac{1}{h}K\left(\frac{X_i - x}{h}\right) - \frac{1}{h}K\left(\frac{X_i - x'}{h}\right)\right| \lesssim \frac{|x - x'|}{h^2}$ almost surely, and also that $\mathbb{E}\Big[\left|\frac{1}{h}K\left(\frac{X_i - x}{h}\right) - \frac{1}{h}K\left(\frac{X_i - x'}{h}\right)\right|^2\Big] \lesssim \frac{|x - x'|^2}{h^3}$. By chaining with the Bernstein–Orlicz

norm and polynomial covering numbers,

$$\sup_{|x-x'|\leq\delta} \left\|S(x) - S(x')\right\|_\infty \lesssim_{\mathbb{P}} \delta\sqrt{\frac{\log n}{nh^3}}$$

whenever $\log(N/h) \lesssim \log n$ and $nh \gtrsim \log n$. By a Gaussian process maximal inequality (van der Vaart and Wellner, 1996, Corollary 2.2.8) the same bound holds for $T(x)$ with

$$\sup_{|x-x'|\leq\delta} \left\|T(x) - T(x')\right\|_\infty \lesssim_{\mathbb{P}} \delta\sqrt{\frac{\log n}{nh^3}}.$$

Proof (Lemma 4.3.1)

For $x, x' \in [a, 1-a]$, the scaled covariance function of this nonparametric estimator is

$$\begin{aligned} nh\ \mathrm{Cov}\left[\hat{g}(x), \hat{g}(x')\right] &= \frac{1}{h}\mathbb{E}\left[K\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x'}{h}\right)\right] \\ &\quad - \frac{1}{h}\mathbb{E}\left[K\left(\frac{X_i - x}{h}\right)\right]\mathbb{E}\left[K\left(\frac{X_i - x'}{h}\right)\right] \\ &= \frac{1}{2\pi}\int_{\frac{-x}{h}}^{\frac{1-x}{h}} \exp\left(-\frac{t^2}{2}\right)\exp\left(-\frac{1}{2}\left(t + \frac{x-x'}{h}\right)^2\right)\,\mathrm{d}t - hI(x)I(x') \end{aligned}$$

where $I(x) = \frac{1}{\sqrt{2\pi}}\int_{-x/h}^{(1-x)/h} e^{-t^2/2}\,\mathrm{d}t$. Completing the square and a substitution gives

$$nh\ \mathrm{Cov}\left[\hat{g}(x), \hat{g}(x')\right] = \frac{1}{2\pi}\exp\left(-\frac{1}{4}\left(\frac{x-x'}{h}\right)^2\right)\int_{\frac{-x-x'}{2h}}^{\frac{2-x-x'}{2h}} \exp\left(-t^2\right)\,\mathrm{d}t - hI(x)I(x').$$

Now we show that since x, x' are not too close to the boundary of $[0,1]$, the limits in the above integral can be replaced by $\pm\infty$. Note that $\frac{-x-x'}{2h} \leq \frac{-a}{h}$ and $\frac{2-x-x'}{2h} \geq \frac{a}{h}$ so

$$\int_{-\infty}^{\infty} \exp\left(-t^2\right)\,\mathrm{d}t - \int_{\frac{-x-x'}{2h}}^{\frac{2-x-x'}{2h}} \exp\left(-t^2\right)\,\mathrm{d}t \leq 2\int_{a/h}^{\infty} \exp\left(-t^2\right)\,\mathrm{d}t \leq \frac{h}{a}\exp\left(-\frac{a^2}{h^2}\right).$$

Therefore, since $\int_{-\infty}^{\infty} e^{-t^2}\,\mathrm{d}t = \sqrt{\pi}$,

$$\left|nh\ \mathrm{Cov}\left[\hat{g}(x), \hat{g}(x')\right] - \frac{1}{2\sqrt{\pi}}\exp\left(-\frac{1}{4}\left(\frac{x-x'}{h}\right)^2\right) + hI(x)I(x')\right| \leq \frac{h}{2\pi a}\exp\left(-\frac{a^2}{h^2}\right).$$

Define the $N \times N$ matrix $\tilde{\Sigma}_{ij} = \frac{1}{2\sqrt{\pi}} \exp\left(-\frac{1}{4}\left(\frac{x_i - x_j}{h}\right)^2\right)$. By Baxter (1994, Proposition 2.4, Proposition 2.5, and Equation 2.10), with $\mathcal{B}_k = \left\{b \in \mathbb{R}^{\mathbb{Z}} : \sum_{i\in\mathbb{Z}} \mathbb{I}\{b_i \neq 0\} \leq k\right\}$,

$$\inf_{k\in\mathbb{N}} \inf_{b\in\mathbb{R}^k} \frac{\sum_{i=1}^k \sum_{j=1}^k b_i b_j \, e^{-\lambda(i-j)^2}}{\sum_{i=1}^k b_i^2} = \sqrt{\frac{\pi}{\lambda}} \sum_{i=-\infty}^{\infty} \exp\left(-\frac{(\pi e + 2\pi i)^2}{4\lambda}\right).$$

We use Riemann sums, noting that $\pi e + 2\pi x = 0$ at $x = -e/2 \approx -1.359$. Consider the substitutions $\mathbb{Z} \cap (-\infty, -3] \mapsto (-\infty, -2]$, $\{-2, -1\} \mapsto \{-2, -1\}$, and $\mathbb{Z} \cap [0, \infty) \mapsto [-1, \infty)$.

$$\begin{aligned}\sum_{i\in\mathbb{Z}} e^{-(\pi e+2\pi i)^2/4\lambda} \leq \int_{-\infty}^{-2} e^{-(\pi e+2\pi x)^2/4\lambda} \, \mathrm{d}x + e^{-(\pi e-4\pi)^2/4\lambda} \\ + e^{-(\pi e-2\pi)^2/4\lambda} + \int_{-1}^{\infty} e^{-(\pi e+2\pi x)^2/4\lambda} \, \mathrm{d}x.\end{aligned}$$

Now use the substitution $t = \frac{\pi e + 2\pi x}{2\sqrt{\lambda}}$ and suppose $\lambda < 1$, yielding

$$\begin{aligned}\sum_{i\in\mathbb{Z}} e^{-(\pi e+2\pi i)^2/4\lambda} &\leq \frac{\sqrt{\lambda}}{\pi} \int_{-\infty}^{\frac{\pi e-4\pi}{2\sqrt{\lambda}}} e^{-t^2} \, \mathrm{d}t + e^{-(\pi e-4\pi)^2/4\lambda} + e^{-(\pi e-2\pi)^2/4\lambda} + \frac{\sqrt{\lambda}}{\pi} \int_{\frac{\pi e-2\pi}{2\sqrt{\lambda}}}^{\infty} e^{-t^2} \, \mathrm{d}t \\ &\leq \left(1 + \frac{1}{\pi} \frac{\lambda}{4\pi - \pi e}\right) e^{-(\pi e-4\pi)^2/4\lambda} + \left(1 + \frac{1}{\pi} \frac{\lambda}{\pi e - 2\pi}\right) e^{-(\pi e-2\pi)^2/4\lambda} \\ &\leq \frac{13}{12} e^{-(\pi e-4\pi)^2/4\lambda} + \frac{8}{7} e^{-(\pi e-2\pi)^2/4\lambda} \leq \frac{9}{4} \exp\left(-\frac{5}{4\lambda}\right).\end{aligned}$$

Therefore

$$\inf_{k\in\mathbb{N}} \inf_{b\in\mathcal{B}_k} \frac{\sum_{i\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} b_i b_j \, e^{-\lambda(i-j)^2}}{\sum_{i\in\mathbb{Z}} b_i^2} < \frac{4}{\sqrt{\lambda}} \exp\left(-\frac{5}{4\lambda}\right) < 4e^{-1/\lambda}.$$

From this and since $\tilde{\Sigma}_{ij} = \frac{1}{2\sqrt{\pi}} e^{-\lambda(i-j)^2}$ with $\lambda = \frac{1}{4(N-1)^2 h^2} \leq \frac{\delta^2}{h^2}$, for each h and some $\delta \leq h$, we have $\lambda_{\min}(\tilde{\Sigma}) \leq 2e^{-h^2/\delta^2}$. Recall that

$$\left|\Sigma_{ij} - \tilde{\Sigma}_{ij} + hI(x_i)I(x_j)\right| \leq \frac{h}{2\pi a} \exp\left(-\frac{a^2}{h^2}\right).$$

For any positive semi-definite $N \times N$ matrices A and B and vector v we have $\lambda_{\min}(A - vv^{\mathsf{T}}) \leq \lambda_{\min}(A)$ and $\lambda_{\min}(B) \leq \lambda_{\min}(A) + \|B - A\|_2 \leq \lambda_{\min}(A) + N\|B - A\|_{\max}$. Hence with $I_i = I(x_i)$,

$$\lambda_{\min}(\Sigma) \leq \lambda_{\min}(\tilde{\Sigma} - hII^{\mathsf{T}}) + \frac{Nh}{2\pi a} \exp\left(-\frac{a^2}{h^2}\right) \leq 2e^{-h^2/\delta^2} + \frac{h}{\pi a \delta} e^{-a^2/h^2}. \qquad \square$$

Proof (Proposition 4.3.1)

Let $\mathcal{F}_\delta$ be a δ-cover of $(\mathcal{F}, d)$. Using a union bound, we can write

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left|S(f) - T(f)\right| \geq 2t + \eta\right) \leq \mathbb{P}\left(\sup_{f \in \mathcal{F}_\delta} \left|S(f) - T(f)\right| \geq \eta\right) + \mathbb{P}\left(\sup_{d(f,f') \leq \delta} \left|S(f) - S(f')\right| \geq t\right) + \mathbb{P}\left(\sup_{d(f,f') \leq \delta} \left|T(f) - T(f')\right| \geq t\right).$$

Part 1: bounding the difference on $\mathcal{F}_\delta$

We apply Corollary 4.2.2 with $p = \infty$ to the martingale difference sequence $\mathcal{F}_\delta(X_i) = \big(f(X_i) : f \in \mathcal{F}_\delta\big)$ which takes values in $\mathbb{R}^{|\mathcal{F}_\delta|}$. Square integrability can be assumed otherwise $\beta_\delta = \infty$. Note $\sum_{i=1}^n \mathcal{F}_\delta(X_i) = S(\mathcal{F}_\delta)$ and $\phi_\infty(\mathcal{F}_\delta) \leq \sqrt{2 \log 2|\mathcal{F}_\delta|}$. Therefore there exists a conditionally Gaussian vector $T(\mathcal{F}_\delta)$ with the same covariance structure as $S(\mathcal{F}_\delta)$ conditional on $\mathcal{H}_0$ satisfying

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}_\delta} \left|S(f) - T(f)\right| \geq \eta\right) \leq \frac{24 \beta_\delta^{\frac{1}{3}} (2 \log 2|\mathcal{F}_\delta|)^{\frac{1}{3}}}{\eta} + 17 \left(\frac{\sqrt{2 \log 2|\mathcal{F}_\delta|} \sqrt{\mathbb{E}\left[\|\Omega_\delta\|_2\right]}}{\eta}\right)^{\frac{2}{3}}.$$

Part 2: bounding the fluctuations in $S(f)$

Since $\big\|S(f) - S(f')\big\|_\psi \leq L d(f, f')$, by Theorem 2.2.4 in van der Vaart and Wellner (1996)

$$\left\| \sup_{d(f,f') \leq \delta} \left|S(f) - S(f')\right| \right\|_\psi \leq C_\psi L \left(\int_0^\delta \psi^{-1}(N_\varepsilon) \, \mathrm{d}\varepsilon + \delta \psi^{-1}(N_\delta^2) \right) = C_\psi L J_\psi(\delta).$$

Then, by Markov's inequality and the definition of the Orlicz norm,

$$\mathbb{P}\left(\sup_{d(f,f')\leq\delta}\left|S(f)-S(f')\right|\geq t\right)\leq\psi\left(\frac{t}{C_\psi L J_\psi(\delta)}\right)^{-1}.$$

Part 3: bounding the fluctuations in $T(f)$

By the Vorob'ev–Berkes–Philipp theorem (Dudley, 1999), $T(\mathcal{F}_\delta)$ extends to a conditionally Gaussian process $T(f)$. Firstly, since $\left\|\left|T(f)-T(f')\right|\right\|_2\leq Ld(f,f')$ conditionally on $\mathcal{H}_0$, and $T(f)$ is a conditional Gaussian process, we have $\left\|T(f)-T(f')\right\|_{\psi_2}\leq 2Ld(f,f')$ conditional on $\mathcal{H}_0$ by van der Vaart and Wellner (1996, Chapter 2.2, Complement 1), where $\psi_2(x)=\exp(x^2)-1$. Thus again by Theorem 2.2.4 in van der Vaart and Wellner (1996), again conditioning on $\mathcal{H}_0$,

$$\left\|\sup_{d(f,f')\leq\delta}\left|T(f)-T(f')\right|\right\|_{\psi_2}\leq C_1 L\int_0^\delta\sqrt{\log N_\varepsilon}\,\mathrm{d}\varepsilon=C_1 L J_2(\delta)$$

for some universal constant $C_1>0$, where we used $\psi_2^{-1}(x)=\sqrt{\log(1+x)}$ and monotonicity of covering numbers. Then by Markov's inequality and the definition of the Orlicz norm,

$$\mathbb{P}\left(\sup_{d(f,f')\leq\delta}\left|T(f)-T(f')\right|\geq t\right)\leq\left(\exp\left(\frac{t^2}{C_1^2L^2J_2(\delta)^2}\right)-1\right)^{-1}\vee 1\leq 2\exp\left(\frac{-t^2}{C_1^2L^2J_2(\delta)^2}\right).$$

Part 4: conclusion

The result follows by scaling t and η and enlarging constants if necessary. □

C.1.4 Applications to nonparametric regression

Proof (Proposition 4.4.1)

Proceed according to the decomposition in Section 4.4.1. By stationarity and Lemma SA-2.1 in Cattaneo et al. (2020), we have $\sup_w\|p(w)\|_1\lesssim 1$ and also $\|H\|_1\lesssim n/k$ and $\|H^{-1}\|_1\lesssim k/n$.

Part 1: bounding $\beta_{\infty,2}$ and $\beta_{\infty,3}$

Set $X_i = p(W_i)\varepsilon_i$ so $S = \sum_{i=1}^n X_i$, and set $\sigma_i^2 = \sigma^2(W_i)$ and $V_i = \mathrm{Var}[X_i \mid \mathcal{H}_{i-1}] = \sigma_i^2 p(W_i)p(W_i)^\mathsf{T}$. Recall from Corollary 4.2.2 that for $r \in \{2,3\}$,

$$\beta_{\infty,r} = \sum_{i=1}^n \mathbb{E}\left[\|X_i\|_2^r \|X_i\|_\infty + \|V_i^{1/2} Z_i\|_2^r \|V_i^{1/2} Z_i\|_\infty\right]$$

with $Z_i \sim \mathcal{N}(0,1)$ i.i.d. and independent of V_i. For the first term, we use $\sup_w \|p(w)\|_2 \lesssim 1$ and bounded third moments of ε_i:

$$\mathbb{E}\left[\|X_i\|_2^r \|X_i\|_\infty\right] \le \mathbb{E}\left[|\varepsilon_i|^3 \|p(W_i)\|_2^{r+1}\right] \lesssim 1.$$

For the second term, apply Lemma C.1.3 conditionally on $\mathcal{H}_n$ with $\sup_w \|p(w)\|_2 \lesssim 1$ to see

$$\begin{aligned}
\mathbb{E}\left[\|V_i^{1/2} Z_i\|_2^r \|V_i^{1/2} Z_i\|_\infty\right] &\lesssim \sqrt{\log 2k}\ \mathbb{E}\left[\max_{1\le j\le k} (V_i)_{jj}^{1/2} \left(\sum_{j=1}^k (V_i)_{jj}\right)^{r/2}\right] \\
&\lesssim \sqrt{\log 2k}\ \mathbb{E}\left[\sigma_i^{r+1} \max_{1\le j\le k} p(W_i)_j \left(\sum_{j=1}^k p(W_i)_j^2\right)^{r/2}\right] \lesssim \sqrt{\log 2k}\ \mathbb{E}\left[\sigma_i^{r+1}\right] \lesssim \sqrt{\log 2k}.
\end{aligned}$$

Putting these together yields $\beta_{\infty,2} \lesssim n\sqrt{\log 2k}$ and $\beta_{\infty,3} \lesssim n\sqrt{\log 2k}$.

Part 2: bounding Ω

Set $\Omega = \sum_{i=1}^n \left(V_i - \mathbb{E}[V_i]\right)$ so

$$\Omega = \sum_{i=1}^n \left(\sigma_i^2 p(W_i)p(W_i)^\mathsf{T} - \mathbb{E}\left[\sigma_i^2 p(W_i)p(W_i)^\mathsf{T}\right]\right).$$

Observe that Ω_{jl} is the sum of a zero-mean strictly stationary α-mixing sequence and so $\mathbb{E}[\Omega_{jl}^2] \lesssim n$ by Lemma C.1.6(i). Since the basis functions satisfy Assumption 3 in Cattaneo et al. (2020), Ω has a bounded number of non-zero entries in each row, so by Jensen's inequality

$$\mathbb{E}\left[\|\Omega\|_2\right] \le \mathbb{E}\left[\|\Omega\|_\mathrm{F}\right] \le \left(\sum_{j=1}^k \sum_{l=1}^k \mathbb{E}\left[\Omega_{jl}^2\right]\right)^{1/2} \lesssim \sqrt{nk}.$$

Part 3: strong approximation

By Corollary 4.2.2 and the previous parts, with any sequence $R_n \to \infty$,

$$\begin{aligned}\|S - T\|_\infty &\lesssim_{\mathbb{P}} \beta_{\infty,2}^{1/3}(\log 2k)^{1/3} R_n + \sqrt{\log 2k}\sqrt{\mathbb{E}[\|\Omega\|_2]} R_n \\ &\lesssim_{\mathbb{P}} n^{1/3}\sqrt{\log 2k} R_n + (nk)^{1/4}\sqrt{\log 2k} R_n.\end{aligned}$$

If further $\mathbb{E}\left[\varepsilon_i^3 \mid \mathcal{H}_{i-1}\right] = 0$ then the third-order version of Corollary 4.2.2 applies since

$$\pi_3 = \sum_{i=1}^n \sum_{|\kappa|=3} \mathbb{E}\Big[\big|\mathbb{E}[X_i^\kappa \mid \mathcal{H}_{i-1}]\big|\Big] = \sum_{i=1}^n \sum_{|\kappa|=3} \mathbb{E}\Big[\big|p(W_i)^\kappa\, \mathbb{E}[\varepsilon_i^3 \mid \mathcal{H}_{i-1}]\big|\Big] = 0,$$

giving

$$\|S - T\|_\infty \lesssim_{\mathbb{P}} \beta_{\infty,3}^{1/4}(\log 2k)^{3/8} R_n + \sqrt{\log 2k}\sqrt{\mathbb{E}[\|\Omega\|_2]} R_n \lesssim_{\mathbb{P}} (nk)^{1/4}\sqrt{\log 2k} R_n.$$

By Hölder's inequality and with $\|H^{-1}\|_1 \lesssim k/n$ we have

$$\sup_{w\in\mathcal{W}} \Big|p(w)^\mathsf{T} H^{-1} S - p(w)^\mathsf{T} H^{-1} T\Big| \le \sup_{w\in\mathcal{W}} \|p(w)\|_1 \|H^{-1}\|_1 \|S - T\|_\infty \lesssim n^{-1}k \|S - T\|_\infty.$$

Part 4: convergence of $\hat{H}$

We have $\hat{H} - H = \sum_{i=1}^n \big(p(W_i)p(W_i)^\mathsf{T} - \mathbb{E}\left[p(W_i)p(W_i)^\mathsf{T}\right]\big)$. Observe that $(\hat{H} - H)_{jl}$ is the sum of a zero-mean strictly stationary α-mixing sequence and so $\mathbb{E}[(\hat{H} - H)_{jl}^2] \lesssim n$ by Lemma C.1.6(i). Since the basis functions satisfy Assumption 3 in Cattaneo et al. (2020), $\hat{H} - H$ has a bounded number of non-zero entries in each row and so by Jensen's inequality

$$\mathbb{E}\Big[\|\hat{H} - H\|_1\Big] = \mathbb{E}\left[\max_{1\le i\le k} \sum_{j=1}^k \big|(\hat{H} - H)_{ij}\big|\right] \le \mathbb{E}\left[\sum_{1\le i\le k}\left(\sum_{j=1}^k |(\hat{H} - H)_{ij}|\right)^2\right]^{\frac{1}{2}} \lesssim \sqrt{nk}.$$

Part 5: bounding the matrix term

Note $\|\hat{H}^{-1}\|_1 \leq \|H^{-1}\|_1 + \|\hat{H}^{-1}\|_1\|\hat{H} - H\|_1\|H^{-1}\|_1$ so by the previous part, we deduce

$$\|\hat{H}^{-1}\|_1 \leq \frac{\|H^{-1}\|_1}{1 - \|\hat{H} - H\|_1\|H^{-1}\|_1} \lesssim_{\mathbb{P}} \frac{k/n}{1 - \sqrt{nk}\, k/n} \lesssim_{\mathbb{P}} \frac{k}{n}$$

as $k^3/n \to 0$. Note that by the martingale structure, since $p(W_i)$ is bounded and supported on a region with volume at most of the order $1/k$, and as W_i has a Lebesgue density,

$$\mathrm{Var}[T_j] = \mathrm{Var}[S_j] = \mathrm{Var}\left[\sum_{i=1}^n \varepsilon_i p(W_i)_j\right] = \sum_{i=1}^n \mathbb{E}\left[\sigma_i^2 p(W_i)_j^2\right] \lesssim \frac{n}{k}.$$

So by the Gaussian maximal inequality in Lemma C.1.4, $\|T\|_\infty \lesssim_{\mathbb{P}} \sqrt{\frac{n \log 2k}{k}}$. Since $k^3/n \to 0$,

$$\begin{aligned}
\sup_{w \in \mathcal{W}} \left|p(w)^{\mathsf{T}}(\hat{H}^{-1} - H^{-1})S\right| &\leq \sup_{w \in \mathcal{W}} \|p(w)^{\mathsf{T}}\|_1\|\hat{H}^{-1}\|_1\|\hat{H} - H\|_1\|H^{-1}\|_1\|S - T\|_\infty \\
&\quad + \sup_{w \in \mathcal{W}} \|p(w)^{\mathsf{T}}\|_1\|\hat{H}^{-1}\|_1\|\hat{H} - H\|_1\|H^{-1}\|_1\|T\|_\infty \\
&\lesssim_{\mathbb{P}} \frac{k^2}{n^2}\sqrt{nk}\Big(n^{1/3}\sqrt{\log 2k} + (nk)^{1/4}\sqrt{\log 2k}\Big) + \frac{k^2}{n^2}\sqrt{nk}\sqrt{\frac{n \log 2k}{k}} \\
&\lesssim_{\mathbb{P}} \frac{k^2}{n}\sqrt{\log 2k}.
\end{aligned}$$

Part 6: conclusion of the main result

By the previous parts, with $G(w) = p(w)^{\mathsf{T}}H^{-1}T$,

$$\begin{aligned}
&\sup_{w \in \mathcal{W}} \left|\hat{\mu}(w) - \mu(w) - p(w)^{\mathsf{T}}H^{-1}T\right| \\
&= \sup_{w \in \mathcal{W}} \left|p(w)^{\mathsf{T}}H^{-1}(S - T) + p(w)^{\mathsf{T}}(\hat{H}^{-1} - H^{-1})S + \mathrm{Bias}(w)\right| \\
&\lesssim_{\mathbb{P}} \frac{k}{n}\|S - T\|_\infty + \frac{k^2}{n}\sqrt{\log 2k} + \sup_{w \in \mathcal{W}} |\mathrm{Bias}(w)| \\
&\lesssim_{\mathbb{P}} \frac{k}{n}\left(n^{1/3}\sqrt{\log 2k} + (nk)^{1/4}\sqrt{\log 2k}\right) R_n + \frac{k^2}{n}\sqrt{\log 2k} + \sup_{w \in \mathcal{W}} |\mathrm{Bias}(w)| \\
&\lesssim_{\mathbb{P}} n^{-2/3}k\sqrt{\log 2k}R_n + n^{-3/4}k^{5/4}\sqrt{\log 2k}R_n + \frac{k^2}{n}\sqrt{\log 2k} + \sup_{w \in \mathcal{W}} |\mathrm{Bias}(w)| \\
&\lesssim_{\mathbb{P}} n^{-2/3}k\sqrt{\log 2k}R_n + \sup_{w \in \mathcal{W}} |\mathrm{Bias}(w)|
\end{aligned}$$

since $k^3/n \to 0$. If further $\mathbb{E}\left[\varepsilon_i^3 \mid \mathcal{H}_{i-1}\right] = 0$ then

$$\begin{aligned}\sup_{w\in\mathcal{W}}\left|\hat{\mu}(w) - \mu(w) - p(w)^\mathsf{T} H^{-1} T\right| &\lesssim_\mathbb{P} \frac{k}{n}\|S - T\|_\infty + \frac{k^2}{n}\sqrt{\log 2k} + \sup_{w\in\mathcal{W}} |\operatorname{Bias}(w)| \\ &\lesssim_\mathbb{P} n^{-3/4}k^{5/4}\sqrt{\log 2k}R_n + \sup_{w\in\mathcal{W}} |\operatorname{Bias}(w)|.\end{aligned}$$

Finally, we verify the variance bounds for the Gaussian process. With $\sigma^2(w)$ bounded above,

$$\begin{aligned}\operatorname{Var}[G(w)] &= p(w)^\mathsf{T} H^{-1} \operatorname{Var}\left[\sum_{i=1}^n p(W_i)\varepsilon_i\right] H^{-1} p(w) \\ &= p(w)^\mathsf{T} H^{-1} \mathbb{E}\left[\sum_{i=1}^n p(W_i)p(W_i)^\mathsf{T}\sigma^2(W_i)\right] H^{-1} p(w) \\ &\lesssim \|p(w)\|_2^2 \|H^{-1}\|_2^2 \|H\|_2 \lesssim k/n.\end{aligned}$$

Similarly, since $\sigma^2(w)$ is bounded away from zero,

$$\operatorname{Var}[G(w)] \gtrsim \|p(w)\|_2^2 \|H^{-1}\|_2^2 \|H^{-1}\|_2^{-1} \gtrsim k/n. \qquad \square$$

Part 7: bounding the bias

We delegate the task of carefully deriving bounds on the bias to Cattaneo et al. (2020), who provide a high-level assumption on the approximation error in Assumption 4 and then use it to derive bias bounds in Section 3 of the form $\sup_{w\in\mathcal{W}} |\operatorname{Bias}(w)| \lesssim_\mathbb{P} k^{-\gamma}$. This assumption is then verified for B-splines, wavelets, and piecewise polynomials in their supplemental appendix.

Proof (Proposition 4.4.2)

Part 1: infeasible supremum approximation

Provided that the bias is negligible, for all $s > 0$ we have

$$\begin{aligned}&\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\frac{\hat{\mu}(w) - \mu(w)}{\sqrt{\rho(w,w)}}\right| \le t\right) - \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\frac{G(w)}{\sqrt{\rho(w,w)}}\right| \le t\right)\right| \\ &\quad\le \sup_{t\in\mathbb{R}}\mathbb{P}\left(t \le \sup_{w\in\mathcal{W}}\left|\frac{G(w)}{\sqrt{\rho(w,w)}}\right| \le t + s\right) + \mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\frac{\hat{\mu}(w) - \mu(w) - G(w)}{\sqrt{\rho(w,w)}}\right| > s\right).\end{aligned}$$

By the Gaussian anti-concentration result given as Corollary 2.1 in Chernozhukov et al. (2014a) applied to a discretization of $\mathcal{W}$, the first term is at most $s\sqrt{\log n}$ up to a constant factor, and the second term converges to zero whenever $\frac{1}{s}\left(\frac{k^3(\log k)^3}{n}\right)^{1/6} \to 0$. Thus a suitable value of s exists whenever $\frac{k^3(\log n)^6}{n} \to 0$.

Part 2: feasible supremum approximation

By Chernozhukov et al. (2013a, Lemma 3.1) and discretization, with $\rho(w,w') = \mathbb{E}[\hat{\rho}(w,w')]$,

$$\begin{aligned}
&\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\frac{\hat{G}(w)}{\sqrt{\hat{\rho}(w,w)}}\right| \le t \;\middle|\; \mathbf{W},\mathbf{Y}\right) - \mathbb{P}\left(\left|\frac{G(w)}{\sqrt{\rho(w,w)}}\right| \le t\right)\right| \\
&\quad\lesssim_{\mathbb{P}} \sup_{w,w'\in\mathcal{W}}\left|\frac{\hat{\rho}(w,w')}{\sqrt{\hat{\rho}(w,w)\hat{\rho}(w',w')}} - \frac{\rho(w,w')}{\sqrt{\rho(w,w)\rho(w',w')}}\right|^{1/3}(\log n)^{2/3} \\
&\quad\lesssim_{\mathbb{P}} \left(\frac{n}{k}\right)^{1/3}\sup_{w,w'\in\mathcal{W}}|\hat{\rho}(w,w') - \rho(w,w')|^{1/3}(\log n)^{2/3} \\
&\quad\lesssim_{\mathbb{P}} \left(\frac{n(\log n)^2}{k}\right)^{1/3}\sup_{w,w'\in\mathcal{W}}\left|p(w)^{\mathsf{T}}\hat{H}^{-1}\left(\hat{V}[S] - \mathrm{Var}[S]\right)\hat{H}^{-1}p(w')\right|^{1/3} \\
&\quad\lesssim_{\mathbb{P}} \left(\frac{k(\log n)^2}{n}\right)^{1/3}\left\|\hat{V}[S] - \mathrm{Var}[S]\right\|_2^{1/3},
\end{aligned}$$

and vanishes in probability whenever $\frac{k(\log n)^2}{n}\left\|\hat{V}[S]-\mathrm{Var}[S]\right\|_2 \to_{\mathbb{P}} 0$. For the plug-in estimator,

$$\begin{aligned}
\left\|\hat{V}[S] - \mathrm{Var}[S]\right\|_2 &= \left\|\sum_{i=1}^n p(W_i)p(W_i^{\mathsf{T}})\hat{\sigma}^2(W_i) - n\mathbb{E}\left[p(W_i)p(W_i^{\mathsf{T}})\sigma^2(W_i)\right]\right\|_2 \\
&\lesssim_{\mathbb{P}} \sup_{w\in\mathcal{W}}|\hat{\sigma}^2(w) - \sigma^2(w)|\left\|\hat{H}\right\|_2 \\
&\quad + \left\|\sum_{i=1}^n p(W_i)p(W_i^{\mathsf{T}})\sigma^2(W_i) - n\mathbb{E}\left[p(W_i)p(W_i^{\mathsf{T}})\sigma^2(W_i)\right]\right\|_2 \\
&\lesssim_{\mathbb{P}} \frac{n}{k}\sup_{w\in\mathcal{W}}|\hat{\sigma}^2(w) - \sigma^2(w)| + \sqrt{nk},
\end{aligned}$$

where the second term is bounded by the same argument used to bound $\|\hat{H} - H\|_1$. Thus, the feasible approximation is valid whenever $(\log n)^2\sup_{w\in\mathcal{W}}|\hat{\sigma}^2(w) - \sigma^2(w)| \to_{\mathbb{P}} 0$ and $\frac{k^3(\log n)^4}{n} \to 0$. The validity of the uniform confidence band follows immediately. □

Proof (Proposition 4.4.3)

We apply Proposition 4.3.1 with the metric $d(f_w, f_{w'}) = \|w - w'\|_2$ and the function class

$$\mathcal{F} = \left\{ (W_i, \varepsilon_i) \mapsto e_1^\mathsf{T} H(w)^{-1} K_h(W_i - w) p_h(W_i - w) \varepsilon_i : \ w \in \mathcal{W} \right\},$$

with ψ chosen as a suitable Bernstein Orlicz function.

Part 1: bounding $H(w)^{-1}$

Recall that $H(w) = \sum_{i=1}^n \mathbb{E}[K_h(W_i - w) p_h(W_i - w) p_h(W_i - w)^\mathsf{T}]$ and let $a(w) \in \mathbb{R}^k$ with $\|a(w)\|_2 = 1$. Since the density of W_i is bounded away from zero on $\mathcal{W}$,

$$\begin{aligned} a(w)^\mathsf{T} H(w) a(w) &= n\mathbb{E}\left[\big(a(w)^\mathsf{T} p_h(W_i - w)\big)^2 K_h(W_i - w) \right] \\ &\gtrsim n \int_{\mathcal{W}} \big(a(w)^\mathsf{T} p_h(u - w)\big)^2 K_h(u - w) \, \mathrm{d}u \gtrsim n \int_{\frac{\mathcal{W}-w}{h}} \big(a(w)^\mathsf{T} p(u)\big)^2 K(u) \, \mathrm{d}u. \end{aligned}$$

This is continuous in $a(w)$ on the compact set $\|a(w)\|_2 = 1$ and $p(u)$ forms a polynomial basis so $a(w)^\mathsf{T} p(u)$ has finitely many zeroes. Since $K(u)$ is compactly supported and $h \to 0$, the above integral is eventually strictly positive for all $x \in \mathcal{W}$, and hence is bounded below uniformly in $w \in \mathcal{W}$ by a positive constant. Therefore $\sup_{w \in \mathcal{W}} \|H(w)^{-1}\|_2 \lesssim 1/n$.

Part 2: bounding β_δ

Let $\mathcal{F}_\delta$ be a δ-cover of $(\mathcal{F}, d)$ with cardinality $|\mathcal{F}_\delta| \asymp \delta^{-m}$ and let $\mathcal{F}_\delta(W_i, \varepsilon_i) = \big(f(W_i, \varepsilon_i) : f \in \mathcal{F}_\delta\big)$. Define the truncated errors $\tilde{\varepsilon}_i = \varepsilon_i \mathbb{I}\{-a \log n \le \varepsilon_i \le b \log n\}$ and note that $\mathbb{E}\big[e^{|\varepsilon_i|/C_\varepsilon}\big] < \infty$ implies that $\mathbb{P}(\exists i : \tilde{\varepsilon}_i \neq \varepsilon_i) \lesssim n^{1-(a \vee b)/C_\varepsilon}$. Hence, by choosing a and b large enough, with high probability, we can replace all ε_i by $\tilde{\varepsilon}_i$. Further, it is always possible to increase either a or b along with some randomization to ensure that $\mathbb{E}[\tilde{\varepsilon}_i] = 0$. Since K is

bounded and compactly supported, W_i has a bounded density and $|\tilde{\varepsilon}_i| \lesssim \log n$,

$$\begin{aligned}
\left|\!\left|\!\left| f(W_i, \tilde{\varepsilon}_i) \right|\!\right|\!\right|_2 &= \mathbb{E}\left[\left| e_1^{\mathsf{T}} H(w)^{-1} K_h(W_i - w) p_h(W_i - w) \tilde{\varepsilon}_i \right|^2 \right]^{1/2} \\
&\leq \mathbb{E}\left[\|H(w)^{-1}\|_2^2 K_h(W_i - w)^2 \|p_h(W_i - w)\|_2^2 \sigma^2(W_i) \right]^{1/2} \\
&\lesssim n^{-1} \mathbb{E}\left[K_h(W_i - w)^2 \right]^{1/2} \lesssim n^{-1} h^{-m/2}, \\
\left|\!\left|\!\left| f(W_i, \tilde{\varepsilon}_i) \right|\!\right|\!\right|_\infty &\leq \left|\!\left|\!\left| \|H(w)^{-1}\|_2 K_h(W_i - w) \|p_h(W_i - w)\|_2 |\tilde{\varepsilon}_i| \right|\!\right|\!\right|_\infty \\
&\lesssim n^{-1} \left|\!\left|\!\left| K_h(W_i - w) \right|\!\right|\!\right|_\infty \log n \lesssim n^{-1} h^{-m} \log n.
\end{aligned}$$

Therefore

$$\mathbb{E}\left[\|\mathcal{F}_\delta(W_i, \tilde{\varepsilon}_i)\|_2^2 \|\mathcal{F}_\delta(W_i, \tilde{\varepsilon}_i)\|_\infty \right] \leq \sum_{f \in \mathcal{F}_\delta} \left|\!\left|\!\left| f(W_i, \tilde{\varepsilon}_i) \right|\!\right|\!\right|_2^2 \max_{f \in \mathcal{F}_\delta} \left|\!\left|\!\left| f(W_i, \tilde{\varepsilon}_i) \right|\!\right|\!\right|_\infty \lesssim n^{-3} \delta^{-m} h^{-2m} \log n.$$

Let $V_i(\mathcal{F}_\delta) = \mathbb{E}\big[\mathcal{F}_\delta(W_i, \tilde{\varepsilon}_i) \mathcal{F}_\delta(W_i, \tilde{\varepsilon}_i)^{\mathsf{T}} \mid \mathcal{H}_{i-1} \big]$ and $Z_i \sim \mathcal{N}(0, I_d)$ be i.i.d. and independent of $\mathcal{H}_n$. Note that $V_i(f, f) = \mathbb{E}[f(W_i, \tilde{\varepsilon}_i)^2 \mid W_i] \lesssim n^{-2} h^{-2m}$ and $\mathbb{E}[V_i(f, f)] = \mathbb{E}[f(W_i, \tilde{\varepsilon}_i)^2] \lesssim n^{-2} h^{-m}$. Thus by Lemma C.1.3,

$$\begin{aligned}
\mathbb{E}\left[\left\| V_i(\mathcal{F}_\delta)^{1/2} Z_i \right\|_2^2 \left\| V_i(\mathcal{F}_\delta)^{1/2} Z_i \right\|_\infty \right] &= \mathbb{E}\left[\mathbb{E}\left[\left\| V_i(\mathcal{F}_\delta)^{1/2} Z_i \right\|_2^2 \left\| V_i(\mathcal{F}_\delta)^{1/2} Z_i \right\|_\infty \mid \mathcal{H}_n \right] \right] \\
&\leq 4\sqrt{\log 2|\mathcal{F}_\delta|}\, \mathbb{E}\left[\max_{f \in \mathcal{F}_\delta} \sqrt{V_i(f, f)} \sum_{f \in \mathcal{F}_\delta} V_i(f, f) \right] \\
&\lesssim n^{-3} h^{-2m} \delta^{-m} \sqrt{\log(1/\delta)}.
\end{aligned}$$

Thus since $\log(1/\delta) \asymp \log(1/h) \asymp \log n$,

$$\beta_\delta = \sum_{i=1}^n \mathbb{E}\left[\|\mathcal{F}_\delta(W_i, \tilde{\varepsilon}_i)\|_2^2 \|\mathcal{F}_\delta(W_i, \tilde{\varepsilon}_i)\|_\infty + \left\| V_i(\mathcal{F}_\delta)^{1/2} Z_i \right\|_2^2 \left\| V_i(\mathcal{F}_\delta)^{1/2} Z_i \right\|_\infty \right] \lesssim \frac{\log n}{n^2 h^{2m} \delta^m}.$$

Part 3: bounding Ω_δ

Let $C_K > 0$ be the radius of a ℓ^2-ball containing the support of K and note that

$$\begin{aligned}\big|V_i(f,f')\big| &= \Big|\mathbb{E}\Big[e_1^\mathsf{T} H(w)^{-1} p_h(W_i - w) e_1^\mathsf{T} H(w')^{-1} p_h(W_i - w') \\ &\qquad \times K_h(W_i - w) K_h(W_i - w') \tilde{\varepsilon}_i^2 \;\Big|\; \mathcal{H}_{i-1}\Big]\Big| \\ &\lesssim n^{-2} K_h(W_i - w) K_h(W_i - w') \\ &\lesssim n^{-2} h^{-m} K_h(W_i - w) \mathbb{I}\{\|w - w'\|_2 \le 2C_K h\}.\end{aligned}$$

Since W_i are α-mixing with $\alpha(j) < e^{-2j/C_\alpha}$, Lemma C.1.6(ii) with $r = 3$ gives

$$\begin{aligned}&\operatorname{Var}\left[\sum_{i=1}^n V_i(f,f')\right] \\ &\lesssim \sum_{i=1}^n \mathbb{E}\left[|V_i(f,f')|^3\right]^{2/3} \lesssim n^{-3} h^{-2m} \mathbb{E}\left[K_h(W_i - w)^3\right]^{2/3} \mathbb{I}\{\|w - w'\|_2 \le 2C_K h\} \\ &\lesssim n^{-3} h^{-2m} (h^{-2m})^{2/3} \mathbb{I}\{\|w - w'\|_2 \le 2C_K h\} \\ &\lesssim n^{-3} h^{-10m/3} \mathbb{I}\{\|w - w'\|_2 \le 2C_K h\}.\end{aligned}$$

Therefore, by Jensen's inequality,

$$\begin{aligned}\mathbb{E}\big[\|\Omega_\delta\|_2\big] \le \mathbb{E}\big[\|\Omega_\delta\|_\mathrm{F}\big] &\le \mathbb{E}\left[\sum_{f,f' \in \mathcal{F}_\delta} (\Omega_\delta)_{f,f'}^2\right]^{1/2} \le \left(\sum_{f,f' \in \mathcal{F}_\delta} \operatorname{Var}\left[\sum_{i=1}^n V_i(f,f')\right]\right)^{1/2} \\ &\lesssim n^{-3/2} h^{-5m/3} \left(\sum_{f,f' \in \mathcal{F}_\delta} \mathbb{I}\{\|w - w'\|_2 \le 2C_K h\}\right)^{1/2} \\ &\lesssim n^{-3/2} h^{-5m/3} \big(h^m \delta^{-2m}\big)^{1/2} \lesssim n^{-3/2} h^{-7m/6} \delta^{-m}.\end{aligned}$$

Note that we could have used $\|\cdot\|_1$ rather than $\|\cdot\|_\mathrm{F}$, but this term is negligible either way.

Part 4: regularity of the stochastic processes

For each $f, f' \in \mathcal{F}$, define the mean-zero and α-mixing random variables

$$u_i(f,f') = e_1^\mathsf{T}\big(H(w)^{-1} K_h(W_i - w) p_h(W_i - w) - H(w')^{-1} K_h(W_i - w') p_h(W_i - w')\big)\tilde{\varepsilon}_i.$$

Note that for all $1 \le j \le k$, by the Lipschitz property of the kernel and monomials,

$$\begin{aligned}
&\big|K_h(W_i - w) - K_h(W_i - w')\big| \\
&\quad \lesssim h^{-m-1}\|w - w'\|_2\big(\mathbb{I}\{\|W_i - w\| \le C_K h\} + \mathbb{I}\{\|W_i - w'\| \le C_K h\}\big), \\
&\big|p_h(W_i - w)_j - p_h(W_i - w')_j\big| \lesssim h^{-1}\|w - w'\|_2,
\end{aligned}$$

to deduce that for any $1 \le j, l \le k$,

$$\begin{aligned}
\big|H(w)_{jl} - H(w')_{jl}\big| &= \big|n\mathbb{E}\big[K_h(W_i - w)p_h(W_i - w)_j p_h(W_i - w)_l \\
&\qquad - K_h(W_i - w')p_h(W_i - w')_j p_h(W_i - w')_l\big]\big| \\
&\le n\mathbb{E}\left[\big|K_h(W_i - w) - K_h(W_i - w')\big|\,|p_h(W_i - w)_j p_h(W_i - w)_l|\right] \\
&\quad + n\mathbb{E}\left[\big|p_h(W_i - w)_j - p_h(W_i - w')_j\big|\,\big|K_h(W_i - w')p_h(W_i - w)_l\big|\right] \\
&\quad + n\mathbb{E}\left[\big|p_h(W_i - w)_l - p_h(W_i - w')_l\big|\,\big|K_h(W_i - w')p_h(W_i - w')_j\big|\right] \\
&\lesssim nh^{-1}\|w - w'\|_2.
\end{aligned}$$

Therefore, as the dimension of the matrix $H(w)$ is fixed,

$$\big\|H(w)^{-1} - H(w')^{-1}\big\|_2 \le \big\|H(w)^{-1}\big\|_2\big\|H(w')^{-1}\big\|_2\big\|H(w) - H(w')\big\|_2 \lesssim \frac{\|w - w'\|_2}{nh}.$$

Hence

$$\begin{aligned}
\big|u_i(f, f')\big| &\le \big\|H(w)^{-1}K_h(W_i - w)p_h(W_i - w) - H(w')^{-1}K_h(W_i - w')p_h(W_i - w')\tilde{\varepsilon}_i\big\|_2 \\
&\le \big\|H(w)^{-1} - H(w')^{-1}\big\|_2\big\|K_h(W_i - w)p_h(W_i - w)\tilde{\varepsilon}_i\big\|_2 \\
&\quad + \big|K_h(W_i - w) - K_h(W_i - w')\big|\big\|H(w')^{-1}p_h(W_i - w)\tilde{\varepsilon}_i\big\|_2 \\
&\quad + \big\|p_h(W_i - w) - p_h(W_i - w')\big\|_2\big\|H(w')^{-1}K_h(W_i - w')\tilde{\varepsilon}_i\big\|_2 \\
&\lesssim \frac{\|w - w'\|_2}{nh}\big|K_h(W_i - w)\tilde{\varepsilon}_i\big| + \frac{1}{n}\big|K_h(W_i - w) - K_h(W_i - w')\big|\,|\tilde{\varepsilon}_i| \\
&\lesssim \frac{\|w - w'\|_2 \log n}{nh^{m+1}},
\end{aligned}$$

and from the penultimate line, we also deduce that

$$\begin{aligned}\mathrm{Var}[u_i(f,f')] &\lesssim \frac{\|w-w'\|_2^2}{n^2h^2}\mathbb{E}\left[K_h(W_i-w)^2\sigma^2(X_i)\right] \\ &\quad + \frac{1}{n^2}\mathbb{E}\left[\big(K_h(W_i-w)-K_h(W_i-w')\big)^2\sigma^2(X_i)\right] \lesssim \frac{\|w-w'\|_2^2}{n^2h^{m+2}}.\end{aligned}$$

Further, $\mathbb{E}[u_i(f,f')u_j(f,f')] = 0$ for $i \neq j$ so by Lemma C.1.7(ii), for a constant $C_1 > 0$,

$$\mathbb{P}\left(\Big|\sum_{i=1}^n u_i(f,f')\Big| \geq \frac{C_1\|w-w'\|_2}{\sqrt{n}h^{m/2+1}}\left(\sqrt{t}+\sqrt{\frac{(\log n)^2}{nh^m}}\sqrt{t}+\sqrt{\frac{(\log n)^6}{nh^m}}t\right)\right) \leq C_1e^{-t}.$$

Therefore, adjusting the constant if necessary and since $nh^m \gtrsim (\log n)^7$,

$$\mathbb{P}\left(\Big|\sum_{i=1}^n u_i(f,f')\Big| \geq \frac{C_1\|w-w'\|_2}{\sqrt{n}h^{m/2+1}}\left(\sqrt{t}+\frac{t}{\sqrt{\log n}}\right)\right) \leq C_1e^{-t}.$$

Van de Geer and Lederer (2013, Lemma 2) with $\psi(x) = \exp\left(\left(\sqrt{1+2x/\sqrt{\log n}}-1\right)^2\log n\right)-1$ now shows that

$$\Big\|\sum_{i=1}^n u_i(f,f')\Big\|_\psi \lesssim \frac{\|w-w'\|_2}{\sqrt{n}h^{m/2+1}}$$

so we take $L = \frac{1}{\sqrt{n}h^{m/2+1}}$. Noting $\psi^{-1}(t) = \sqrt{\log(1+t)} + \frac{\log(1+t)}{2\sqrt{\log n}}$ and $N_\delta \lesssim \delta^{-m}$,

$$\begin{aligned}J_\psi(\delta) &= \int_0^\delta \psi^{-1}\big(N_\varepsilon\big)\,\mathrm{d}\varepsilon + \delta\psi^{-1}\big(N_\delta\big) \lesssim \frac{\delta\log(1/\delta)}{\sqrt{\log n}} + \delta\sqrt{\log(1/\delta)} \lesssim \delta\sqrt{\log n}, \\ J_2(\delta) &= \int_0^\delta \sqrt{\log N_\varepsilon}\,\mathrm{d}\varepsilon \lesssim \delta\sqrt{\log(1/\delta)} \lesssim \delta\sqrt{\log n}.\end{aligned}$$

Part 5: strong approximation

Recalling that $\tilde{\varepsilon}_i = \varepsilon_i$ for all i with high probability, by Proposition 4.3.1, for all $t, \eta > 0$ there exists a zero-mean Gaussian process $T(w)$ satisfying

$$\mathbb{E}\left[\left(\sum_{i=1}^n f_w(W_i,\varepsilon_i)\right)\left(\sum_{i=1}^n f_{w'}(W_i,\varepsilon_i)\right)\right] = \mathbb{E}\big[T(w)T(w')\big]$$

for all $w, w' \in \mathcal{W}$ and

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\sum_{i=1}^{n} f_w(W_i,\varepsilon_i) - T(w)\right| \geq C_\psi(t+\eta)\right) \\
&\quad\leq C_\psi \inf_{\delta>0}\inf_{\mathcal{F}_\delta}\left\{\frac{\beta_\delta^{1/3}(\log 2|\mathcal{F}_\delta|)^{1/3}}{\eta} + \left(\frac{\sqrt{\log 2|\mathcal{F}_\delta|}\sqrt{\mathbb{E}\left[\|\Omega_\delta\|_2\right]}}{\eta}\right)^{2/3}\right. \\
&\qquad\left. + \psi\left(\frac{t}{LJ_\psi(\delta)}\right)^{-1} + \exp\left(\frac{-t^2}{L^2 J_2(\delta)^2}\right)\right\} \\
&\quad\leq C_\psi\left\{\frac{\left(\frac{\log n}{n^2h^{2m}\delta^m}\right)^{1/3}(\log n)^{1/3}}{\eta} + \left(\frac{\sqrt{\log n}\sqrt{n^{-3/2}h^{-7m/6}\delta^{-m}}}{\eta}\right)^{2/3}\right. \\
&\qquad\left. + \psi\left(\frac{t}{\frac{1}{\sqrt{n}h^{m/2+1}}J_\psi(\delta)}\right)^{-1} + \exp\left(\frac{-t^2}{\left(\frac{1}{\sqrt{n}h^{m/2+1}}\right)^2 J_2(\delta)^2}\right)\right\} \\
&\quad\leq C_\psi\left\{\frac{(\log n)^{2/3}}{n^{2/3}h^{2m/3}\delta^{m/3}\eta} + \left(\frac{n^{-3/4}h^{-7m/12}\delta^{-m/2}\sqrt{\log n}}{\eta}\right)^{2/3}\right. \\
&\qquad\left. + \psi\left(\frac{t\sqrt{n}h^{m/2+1}}{\delta\sqrt{\log n}}\right)^{-1} + \exp\left(\frac{-t^2 nh^{m+2}}{\delta^2\log n}\right)\right\}.
\end{aligned}$$

Noting $\psi(x) \geq e^{x^2/4}$ for $x \leq 4\sqrt{\log n}$, any $R_n \to \infty$ gives the probability bound

$$\sup_{w\in\mathcal{W}}\left|\sum_{i=1}^{n} f_w(W_i,\varepsilon_i) - T(w)\right| \lesssim_{\mathbb{P}} \frac{(\log n)^{2/3}}{n^{2/3}h^{2m/3}\delta^{m/3}}R_n + \frac{\sqrt{\log n}}{n^{3/4}h^{7m/12}\delta^{m/2}}R_n + \frac{\delta\sqrt{\log n}}{\sqrt{n}h^{m/2+1}}.$$

Optimizing over δ gives $\delta \asymp \left(\frac{\log n}{nh^{m-6}}\right)^{\frac{1}{2m+6}} = h\left(\frac{\log n}{nh^{3m}}\right)^{\frac{1}{2m+6}}$ and so

$$\sup_{w\in\mathcal{W}}\left|\sum_{i=1}^{n} f_w(W_i,\varepsilon_i) - T(w)\right| \lesssim_{\mathbb{P}} \left(\frac{(\log n)^{m+4}}{n^{m+4}h^{m(m+6)}}\right)^{\frac{1}{2m+6}} R_n.$$

Part 6: convergence of $\hat{H}(w)$

For $1 \leq j, l \leq k$ define the zero-mean random variables

$$u_{ijl}(w) = K_h(W_i - w)p_h(W_i - w)_j p_h(W_i - w)_l - \mathbb{E}\big[K_h(W_i - w)p_h(W_i - w)_j p_h(W_i - w)_l\big]$$

and note that $|u_{ijl}(w)| \lesssim h^{-m}$. By Lemma C.1.7(i) for a constant $C_2 > 0$ and all $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} u_{ijl}(w)\right| > C_2 h^{-m}\big(\sqrt{nt} + (\log n)(\log\log n)t\big)\right) \leq C_2 e^{-t}.$$

Further, note that by Lipschitz properties,

$$\left|\sum_{i=1}^{n} u_{ijl}(w) - \sum_{i=1}^{n} u_{ijl}(w')\right| \lesssim h^{-m-1}\|w - w'\|_2$$

so there is a δ-cover of $(\mathcal{W}, \|\cdot\|_2)$ with size at most $n^a\delta^{-a}$ for some $a > 0$. Adjusting C_2,

$$\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\sum_{i=1}^{n} u_{ijl}(w)\right| > C_2 h^{-m}\big(\sqrt{nt} + (\log n)(\log\log n)t\big) + C_2 h^{-m-1}\delta\right) \leq C_2 n^a \delta^{-a} e^{-t}$$

and hence

$$\sup_{w\in\mathcal{W}}\left|\sum_{i=1}^{n} u_{ijl}(w)\right| \lesssim_{\mathbb{P}} h^{-m}\sqrt{n\log n} + h^{-m}(\log n)^3 \lesssim_{\mathbb{P}} \sqrt{\frac{n\log n}{h^{2m}}}.$$

Therefore

$$\sup_{w\in\mathcal{W}}\|\hat{H}(w) - H(w)\|_2 \lesssim_{\mathbb{P}} \sqrt{\frac{n\log n}{h^{2m}}}.$$

Part 7: bounding the matrix term

Firstly, note that since $\sqrt{\frac{\log n}{nh^{2m}}} \to 0$, we have that uniformly in $w \in \mathcal{W}$

$$\|\hat{H}(w)^{-1}\|_2 \leq \frac{\|H(w)^{-1}\|_2}{1 - \|\hat{H}(w) - H(w)\|_2\|H(w)^{-1}\|_2} \lesssim_{\mathbb{P}} \frac{1/n}{1 - \sqrt{\frac{n\log n}{h^{2m}}}\frac{1}{n}} \lesssim_{\mathbb{P}} \frac{1}{n}.$$

Therefore

$$\begin{aligned}\sup_{w\in\mathcal{W}}\left|e_1^{\mathsf{T}}\big(\hat{H}(w)^{-1} - H(w)^{-1}\big)S(w)\right| &\leq \sup_{w\in\mathcal{W}}\big\|\hat{H}(w)^{-1} - H(w)^{-1}\big\|_2\|S(w)\|_2 \\ &\leq \sup_{w\in\mathcal{W}}\big\|\hat{H}(w)^{-1}\big\|_2\big\|H(w)^{-1}\big\|_2\big\|\hat{H}(w) - H(w)\big\|_2\|S(w)\|_2 \lesssim_{\mathbb{P}} \sqrt{\frac{\log n}{n^3h^{2m}}}\sup_{w\in\mathcal{W}}\|S(w)\|_2.\end{aligned}$$

Now for $1 \le j \le k$ write $u_{ij}(w) = K_h(W_i - w)p_h(W_i - w)_j\tilde{\varepsilon}_i$ so that $S(w)_j = \sum_{i=1}^n u_{ij}(w)$ with high probability. Note that $u_{ij}(w)$ are zero-mean with $\mathrm{Cov}[u_{ij}(w), u_{i'j}(w)] = 0$ for $i \neq i'$. Also $|u_{ij}(w)| \lesssim h^{-m}\log n$ and $\mathrm{Var}[u_{ij}(w)] \lesssim h^{-m}$. By Lemma C.1.7(ii) for a constant $C_3 > 0$,

$$\mathbb{P}\left(\Big|\sum_{i=1}^n u_{ij}(w)\Big| \ge C_3\big((h^{-m/2}\sqrt{n} + h^{-m}\log n)\sqrt{t} + h^{-m}(\log n)^3 t\big)\right) \le C_3 e^{-t},$$
$$\mathbb{P}\left(\Big|\sum_{i=1}^n u_{ij}(w)\Big| > C_3\left(\sqrt{\frac{tn}{h^m}} + \frac{t(\log n)^3}{h^m}\right)\right) \le C_3 e^{-t},$$

where we used $nh^m \gtrsim (\log n)^2$ and adjusted the constant if necessary. As before, $u_{ij}(w)$ is Lipschitz in w with a constant which is at most polynomial in n, so for some $a > 0$

$$\mathbb{P}\left(\sup_{w\in\mathcal{W}}\Big|\sum_{i=1}^n u_{ij}(w)\Big| > C_3\left(\sqrt{\frac{tn}{h^m}} + \frac{t(\log n)^3}{h^m}\right)\right) \le C_3 n^a e^{-t},$$
$$\sup_{w\in\mathcal{W}}\|S(w)\|_2 \lesssim_{\mathbb{P}} \sqrt{\frac{n\log n}{h^m}} + \frac{(\log n)^4}{h^m} \lesssim_{\mathbb{P}} \sqrt{\frac{n\log n}{h^m}}$$

as $nh^m \gtrsim (\log n)^7$. Finally,

$$\sup_{w\in\mathcal{W}}\big|e_1^{\mathsf{T}}\big(\hat{H}(w)^{-1} - H(w)^{-1}\big)S(w)\big| \lesssim_{\mathbb{P}} \sqrt{\frac{\log n}{n^3h^{2m}}}\sqrt{\frac{n\log n}{h^m}} \lesssim_{\mathbb{P}} \frac{\log n}{\sqrt{n^2h^{3m}}}.$$

Part 8: bounding the bias

Since $\mu \in \mathcal{C}^\gamma$, we have, by the multivariate version of Taylor's theorem,

$$\mu(W_i) = \sum_{|\kappa|=0}^{\gamma-1}\frac{1}{\kappa!}\partial^\kappa\mu(w)(W_i - w)^\kappa + \sum_{|\kappa|=\gamma}\frac{1}{\kappa!}\partial^\kappa\mu(w')(W_i - w)^\kappa$$

for some w' on the line segment connecting w and W_i. Now since $p_h(W_i - w)_1 = 1$,

$$e_1^{\mathsf{T}}\hat{H}(w)^{-1}\sum_{i=1}^n K_h(W_i - w)p_h(W_i - w)\mu(w)$$
$$= e_1^{\mathsf{T}}\hat{H}(w)^{-1}\sum_{i=1}^n K_h(W_i - w)p_h(W_i - w)p_h(W_i - w)^{\mathsf{T}}e_1\mu(w) = e_1^{\mathsf{T}}e_1\mu(w) = \mu(w).$$

Therefore

$$\begin{aligned}
\operatorname{Bias}(w) &= e_1^{\mathsf{T}}\hat{H}(w)^{-1}\sum_{i=1}^{n} K_h(W_i-w)p_h(W_i-w)\mu(W_i)-\mu(w)\\
&= e_1^{\mathsf{T}}\hat{H}(w)^{-1}\sum_{i=1}^{n} K_h(W_i-w)p_h(W_i-w)\\
&\quad\times\left(\sum_{|\kappa|=0}^{\gamma-1}\frac{1}{\kappa!}\partial^\kappa\mu(w)(W_i-w)^\kappa+\sum_{|\kappa|=\gamma}\frac{1}{\kappa!}\partial^\kappa\mu(w')(W_i-w)^\kappa-\mu(w)\right)\\
&= \sum_{|\kappa|=1}^{\gamma-1}\frac{1}{\kappa!}\partial^\kappa\mu(w)e_1^{\mathsf{T}}\hat{H}(w)^{-1}\sum_{i=1}^{n} K_h(W_i-w)p_h(W_i-w)(W_i-w)^\kappa\\
&\quad+\sum_{|\kappa|=\gamma}\frac{1}{\kappa!}\partial^\kappa\mu(w')e_1^{\mathsf{T}}\hat{H}(w)^{-1}\sum_{i=1}^{n} K_h(W_i-w)p_h(W_i-w)(W_i-w)^\kappa\\
&= \sum_{|\kappa|=\gamma}\frac{1}{\kappa!}\partial^\kappa\mu(w')e_1^{\mathsf{T}}\hat{H}(w)^{-1}\sum_{i=1}^{n} K_h(W_i-w)p_h(W_i-w)(W_i-w)^\kappa,
\end{aligned}$$

where we used that $p_h(W_i-w)$ is a vector containing monomials in W_i-w of order up to γ, so $e_1^{\mathsf{T}}\hat{H}(w)^{-1}\sum_{i=1}^{n} K_h(W_i-w)p_h(W_i-w)(W_i-w)^\kappa=0$ whenever $1\le|\kappa|\le\gamma$. Finally,

$$\begin{aligned}
\sup_{w\in\mathcal{W}}|\operatorname{Bias}(w)| &= \sup_{w\in\mathcal{W}}\left|\sum_{|\kappa|=\gamma}\frac{1}{\kappa!}\partial^\kappa\mu(w')e_1^{\mathsf{T}}\hat{H}(w)^{-1}\sum_{i=1}^{n} K_h(W_i-w)p_h(W_i-w)(W_i-w)^\kappa\right|\\
&\lesssim_{\mathbb{P}} \sup_{w\in\mathcal{W}}\max_{|\kappa|=\gamma}\left|\partial^\kappa\mu(w')\right|\|\hat{H}(w)^{-1}\|_2\left\|\sum_{i=1}^{n} K_h(W_i-w)p_h(W_i-w)\right\|_2 h^\gamma\\
&\lesssim_{\mathbb{P}} \frac{h^\gamma}{n}\sup_{w\in\mathcal{W}}\left\|\sum_{i=1}^{n} K_h(W_i-w)p_h(W_i-w)\right\|_2.
\end{aligned}$$

Write $\tilde{u}_{ij}(w)=K_h(W_i-w)p_h(W_i-w)_j$ and note $|\tilde{u}_{ij}(w)|\lesssim h^{-m}$ and $\mathbb{E}[\tilde{u}_{ij}(w)]\lesssim 1$, so

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}\tilde{u}_{ij}(w)-\mathbb{E}\left[\sum_{i=1}^{n}\tilde{u}_{ij}(w)\right]\right|>C_4h^{-m}\big(\sqrt{nt}+(\log n)(\log\log n)t\big)\right)\le C_4e^{-t}$$

by Lemma C.1.7(i) for a constant C_4, By Lipschitz properties, this implies

$$\sup_{w\in\mathcal{W}}\left|\sum_{i=1}^{n}\tilde{u}_{ij}(w)\right|\lesssim_{\mathbb{P}} n\left(1+\sqrt{\frac{\log n}{nh^{2m}}}\right)\lesssim_{\mathbb{P}} n.$$

Therefore $\sup_{w\in\mathcal{W}}|\operatorname{Bias}(w)|\lesssim_{\mathbb{P}} nh^\gamma/n\lesssim_{\mathbb{P}} h^\gamma$.

Part 9: conclusion

By the previous parts,

$$\begin{aligned}\sup_{w\in\mathcal{W}}|\hat{\mu}(w)-\mu(w)-T(w)| &\leq \sup_{w\in\mathcal{W}}\left|e_1^{\mathsf{T}}H(w)^{-1}S(w)-T(w)\right| \\ &\quad+\sup_{w\in\mathcal{W}}\left|e_1^{\mathsf{T}}\big(\hat{H}(w)^{-1}-H(w)^{-1}\big)S(w)\right|+\sup_{w\in\mathcal{W}}|\operatorname{Bias}(w)| \\ &\lesssim_{\mathbb{P}}\left(\frac{(\log n)^{m+4}}{n^{m+4}h^{m(m+6)}}\right)^{\frac{1}{2m+6}}R_n+\frac{\log n}{\sqrt{n^2h^{3m}}}+h^{\gamma} \\ &\lesssim_{\mathbb{P}}\frac{R_n}{\sqrt{nh^m}}\left(\frac{(\log n)^{m+4}}{nh^{3m}}\right)^{\frac{1}{2m+6}}+h^{\gamma},\end{aligned}$$

where the last inequality follows because $nh^{3m}\to\infty$ and $\frac{1}{2m+6}\leq\frac{1}{2}$. Finally, we verify the upper and lower bounds on the variance of the Gaussian process. Since the spectrum of $H(w)^{-1}$ is bounded above and below by $1/n$,

$$\begin{aligned}\operatorname{Var}[T(w)] &= \operatorname{Var}\left[e_1^{\mathsf{T}}H(w)^{-1}\sum_{i=1}^{n}K_h(W_i-w)p_h(W_i-w)\varepsilon_i\right] \\ &= e_1^{\mathsf{T}}H(w)^{-1}\operatorname{Var}\left[\sum_{i=1}^{n}K_h(W_i-w)p_h(W_i-w)\varepsilon_i\right]H(w)^{-1}e_1^{\mathsf{T}} \\ &\lesssim \|H(w)^{-1}\|_2^2\max_{1\leq j\leq k}\sum_{i=1}^{n}\operatorname{Var}\left[K_h(W_i-w)p_h(W_i-w)_j\sigma(W_i)\right] \\ &\lesssim \frac{1}{n^2}n\frac{1}{h^m}\lesssim\frac{1}{nh^m}.\end{aligned}$$

Similarly, $\operatorname{Var}[T(w)]\gtrsim\frac{1}{nh^m}$ by the same argument used to bound eigenvalues of $H(w)^{-1}$. □

C.2 High-dimensional central limit theorems for martingales

We present an application of our main results to high-dimensional central limit theorems for martingales. Our main contribution here is the generality of our results, which are broadly applicable to martingale data and impose minimal extra assumptions. In exchange for the scope and breadth of our results, we naturally do not necessarily achieve state-of-the-art distributional approximation errors in certain special cases, such as with independent data or when restricting the class of sets over which the central limit theorem must hold. Extensions

of our high-dimensional central limit theorem results to mixingales and other approximate martingales, along with third-order refinements and Gaussian mixture target distributions, are possible through methods akin to those used to establish our main results in Section 4.2, but we omit these for succinctness.

Our approach to deriving a high-dimensional martingale central limit theorem proceeds as follows. Firstly, the upcoming Proposition C.2.1 uses our main result on martingale coupling (Corollary 4.2.2) to reduce the problem to that of providing anti-concentration results for high-dimensional Gaussian vectors. We then demonstrate the utility of this reduction by employing a few such anti-concentration methods from the existing literature. Proposition C.2.2 gives a feasible implementation via the Gaussian multiplier bootstrap, enabling valid resampling-based inference using the resulting conditional Gaussian distribution. Finally, in Section C.2.1 we provide an example application: distributional approximation for ℓ^p-norms of high-dimensional martingale vectors in Kolmogorov–Smirnov distance, relying on some recent results concerning Gaussian perimetric inequalities (Nazarov, 2003; Kozbur, 2021; Giessing, 2023; Chernozhukov, Chetverikov, and Kato, 2017b).

We begin this section with some notation. Assume the setup of Corollary 4.2.2 and suppose Σ is non-random. Let $\mathcal{A}$ be a class of measurable subsets of $\mathbb{R}^d$ and take $T \sim \mathcal{N}(0, \Sigma)$. For $\eta > 0$ and $p \in [1, \infty]$ define the Gaussian perimetric quantity

$$\Delta_p(\mathcal{A}, \eta) = \sup_{A \in \mathcal{A}} \left\{ \mathbb{P}(T \in A_p^\eta \setminus A) \vee \mathbb{P}(T \in A \setminus A_p^{-\eta}) \right\},$$

where $A_p^\eta = \{x \in \mathbb{R}^d : \|x - A\|_p \leq \eta\}$, $A_p^{-\eta} = \mathbb{R}^d \setminus (\mathbb{R}^d \setminus A)_p^\eta$, and $\|x - A\|_p = \inf_{x' \in A} \|x - x'\|_p$. Using this perimetric term allows us to convert coupling results to central limit theorems as follows. Denote by $\Gamma_p(\eta)$ the rate of strong approximation attained in Corollary 4.2.2:

$$\Gamma_p(\eta) = 24 \left(\frac{\beta_{p,2} \phi_p(d)^2}{\eta^3} \right)^{1/3} + 17 \left(\frac{\mathbb{E}\left[\|\Omega\|_2 \right] \phi_p(d)^2}{\eta^2} \right)^{1/3}.$$

Proposition C.2.1 (High-dimensional central limit theorem for martingales)

Take the setup of Corollary 4.2.2, and Σ non-random. For a class $\mathcal{A}$ of measurable sets in $\mathbb{R}^d$,

$$\sup_{A\in\mathcal{A}}\big|\mathbb{P}(S\in A)-\mathbb{P}(T\in A)\big|\le\inf_{p\in[1,\infty]}\inf_{\eta>0}\big\{\Gamma_p(\eta)+\Delta_p(\mathcal{A},\eta)\big\}. \tag{C.5}$$

Proof (Proposition C.2.1)

This follows from Strassen's theorem (Lemma C.1.1), but we provide a proof for completeness.

$$\mathbb{P}(S\in A)\le\mathbb{P}(T\in A)+\mathbb{P}(T\in A_p^\eta\setminus A)+\mathbb{P}(\|S-T\|>\eta)$$

and applying this to $\mathbb{R}^d\setminus A$ gives

$$\begin{aligned}\mathbb{P}(S\in A)&=1-\mathbb{P}(S\in\mathbb{R}^d\setminus A)\\&\ge 1-\mathbb{P}(T\in\mathbb{R}^d\setminus A)-\mathbb{P}(T\in(\mathbb{R}^d\setminus A)_p^\eta\setminus(\mathbb{R}^d\setminus A))-\mathbb{P}(\|S-T\|>\eta)\\&=\mathbb{P}(T\in A)-\mathbb{P}(T\in A\setminus A_p^{-\eta})-\mathbb{P}(\|S-T\|>\eta).\end{aligned}$$

Since this holds for all $p\in[1,\infty]$,

$$\begin{aligned}\sup_{A\in\mathcal{A}}\big|\mathbb{P}(S\in A)-\mathbb{P}(T\in A)\big|&\le\sup_{A\in\mathcal{A}}\big\{\mathbb{P}(T\in A_p^\eta\setminus A)\vee\mathbb{P}(T\in A\setminus A_p^{-\eta})\big\}+\mathbb{P}(\|S-T\|>\eta)\\&\le\inf_{p\in[1,\infty]}\inf_{\eta>0}\big\{\Gamma_p(\eta)+\Delta_p(\mathcal{A},\eta)\big\}.\end{aligned}$$

□

The term $\Delta_p(\mathcal{A},\eta)$ in (C.5) is a Gaussian anti-concentration quantity so it depends on the law of S only through the covariance matrix Σ. A few results are available in the literature for bounding this term. For instance, with $\mathcal{A}=\mathcal{C}=\{A\subseteq\mathbb{R}^d \text{ is convex}\}$, Nazarov (2003) showed

$$\Delta_2(\mathcal{C},\eta)\asymp\eta\sqrt{\|\Sigma^{-1}\|_{\mathrm{F}}}, \tag{C.6}$$

whenever Σ is invertible. Proposition C.2.1 with $p=2$ and (C.6) yield for convex sets

$$\sup_{A\in\mathcal{C}}\big|\mathbb{P}(S\in A)-\mathbb{P}(T\in A)\big|\lesssim\inf_{\eta>0}\left\{\left(\frac{\beta_{p,2}d}{\eta^3}\right)^{1/3}+\left(\frac{\mathbb{E}[\|\Omega\|_2]d}{\eta^2}\right)^{1/3}+\eta\sqrt{\|\Sigma^{-1}\|_{\mathrm{F}}}\right\}.$$

Alternatively, one can take $\mathcal{A} = \mathcal{R}$, the class of axis-aligned rectangles in $\mathbb{R}^d$. By Nazarov's Gaussian perimetric inequality (Nazarov, 2003; Chernozhukov et al., 2017a),

$$\Delta_\infty(\mathcal{R}, \eta) \leq \frac{\eta(\sqrt{2 \log d} + 2)}{\sigma_{\min}} \tag{C.7}$$

whenever $\min_j \Sigma_{jj} \geq \sigma_{\min}^2$ for some $\sigma_{\min} > 0$. Proposition C.2.1 with $p = \infty$ and (C.7) yields

$$\sup_{A \in \mathcal{R}} \left| \mathbb{P}(S \in A) - \mathbb{P}(T \in A) \right| \lesssim \inf_{\eta > 0} \left\{ \left(\frac{\beta_{\infty,2} \log 2d}{\eta^3} \right)^{1/3} + \left(\frac{\mathbb{E}[\|\Omega\|_2] \log 2d}{\eta^2} \right)^{1/3} + \frac{\eta \sqrt{\log 2d}}{\sigma_{\min}} \right\}.$$

In situations where $\liminf_n \min_j \Sigma_{jj} = 0$, it may be possible in certain cases to regularize the minimum variance away from zero and then apply a Gaussian–Gaussian rectangular approximation result such as Lemma 2.1 from Chernozhukov, Chetverikov, and Koike (2023).

Remark C.2.1 (Comparisons with the literature)
The literature on high-dimensional central limit theorems has developed rapidly in recent years (see Zhai, 2018; Koike, 2021; Buzun, Shvetsov, and Dylov, 2022; Lopes, 2022; Chernozhukov et al., 2023, and references therein), particularly for the special case of sums of independent random vectors on the rectangular sets $\mathcal{R}$. Our corresponding results are rather weaker in terms of dependence on the dimension than for example Chernozhukov et al. (2023, Theorem 2.1). This is an inherent issue due to our approach of first considering the class of all Borel sets and only afterwards specializing to the smaller class $\mathcal{R}$, where sharper results in the literature directly target the Kolmogorov–Smirnov distance via Stein's method and Slepian interpolation.

Next, we present a version of Proposition C.2.1 in which the covariance matrix Σ is replaced by an estimator $\hat{\Sigma}$. This ensures that the associated conditionally Gaussian vector is feasible and can be resampled, allowing Monte Carlo quantile estimation via a Gaussian multiplier bootstrap.

Proposition C.2.2 (Bootstrap central limit theorem for martingales)

Assume the setup of Corollary 4.2.2, with Σ non-random, and let $\hat{\Sigma}$ be an $\mathbf{X}$-measurable random $d \times d$ positive semi-definite matrix, where $\mathbf{X} = (X_1, \ldots, X_n)$. For a class $\mathcal{A}$ of measurable subsets of $\mathbb{R}^d$,

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}\big(S \in A\big) - \mathbb{P}\big(\hat{\Sigma}^{1/2} Z \in A \mid \mathbf{X}\big) \right| \\ \leq \inf_{p \in [1,\infty]} \inf_{\eta > 0} \left\{ \Gamma_p(\eta) + 2\Delta_p(\mathcal{A}, \eta) + 2d \exp\left(\frac{-\eta^2}{2d^{2/p} \left\| \hat{\Sigma}^{1/2} - \Sigma^{1/2} \right\|_2^2} \right) \right\},$$

where $Z \sim \mathcal{N}(0, I_d)$ is independent of $\mathbf{X}$.

Proof (Proposition C.2.2)

Since $T = \Sigma^{1/2} Z$ is independent of $\mathbf{X}$,

$$\left| \mathbb{P}\big(S \in A\big) - \mathbb{P}\left(\hat{\Sigma}^{1/2} Z \in A \mid \mathbf{X}\right) \right| \\ \leq \left| \mathbb{P}\big(S \in A\big) - \mathbb{P}\big(T \in A\big) \right| + \left| \mathbb{P}\big(\Sigma^{1/2} Z \in A\big) - \mathbb{P}\left(\hat{\Sigma}^{1/2} Z \in A \mid \mathbf{X}\right) \right|.$$

The first term is bounded by Proposition C.2.1; the second by Lemma C.1.5 conditional on $\mathbf{X}$.

$$\left| \mathbb{P}\big(S \in A\big) - \mathbb{P}\left(\hat{\Sigma}^{1/2} Z \in A \mid \mathbf{X}\right) \right| \\ \leq \Gamma_p(\eta) + \Delta_p(\mathcal{A}, \eta) + \Delta_{p'}(\mathcal{A}, \eta') + 2d \exp\left(\frac{-\eta'^2}{2d^{2/p'} \left\| \hat{\Sigma}^{1/2} - \Sigma^{1/2} \right\|_2^2} \right)$$

for all $A \in \mathcal{A}$ and any $p, p' \in [1, \infty]$ and $\eta, \eta' > 0$. Taking a supremum over A and infima over $p = p'$ and $\eta = \eta'$ yields the result. We do not need $p = p'$ and $\eta = \eta'$ in general. □

A natural choice for $\hat{\Sigma}$ in certain situations is the sample covariance matrix $\sum_{i=1}^n X_i X_i^\mathsf{T}$, or a correlation-corrected variant thereof. In general, whenever $\hat{\Sigma}$ does not depend on unknown quantities, one can sample from the law of $\hat{T} = \hat{\Sigma}^{1/2} Z$ conditional on $\mathbf{X}$ to approximate the distribution of S. Proposition C.2.2 verifies that this Gaussian multiplier bootstrap approach is valid whenever $\hat{\Sigma}$ and Σ are sufficiently close. To this end, Theorem X.1.1 in Bhatia (1997) gives $\left\| \hat{\Sigma}^{1/2} - \Sigma^{1/2} \right\|_2 \leq \left\| \hat{\Sigma} - \Sigma \right\|_2^{1/2}$ and Problem X.5.5 in the same gives

$\left\|\hat{\Sigma}^{1/2} - \Sigma^{1/2}\right\|_2 \leq \left\|\Sigma^{-1/2}\right\|_2 \left\|\hat{\Sigma} - \Sigma\right\|_2$ when Σ is invertible. The latter often gives a tighter bound when the minimum eigenvalue of Σ can be bounded away from zero, and consistency of $\hat{\Sigma}$ can be established using a range of matrix concentration inequalities.

In Section C.2.1 we apply Proposition C.2.1 to the special case of approximating the distribution of the ℓ^p-norm of a high-dimensional martingale. Proposition C.2.2 is then used to ensure that feasible distributional approximations are also available.

C.2.1 Application: distributional approximation of martingale ℓ^p-norms

In empirical applications, including nonparametric significance tests (Lopes, Lin, and Müller, 2020) and nearest neighbor search procedures (Biau and Mason, 2015), an estimator or test statistic can be expressed under the null hypothesis as the ℓ^p-norm of a zero-mean martingale for some $p \in [1, \infty]$. In the notation of Corollary 4.2.2, it is of interest to bound Kolmogorov–Smirnov quantities of the form $\sup_{t \geq 0} \left|\mathbb{P}(\|S\|_p \leq t) - \mathbb{P}(\|T\|_p \leq t)\right|$. Let $\mathcal{B}_p$ be the class of closed ℓ^p-balls in $\mathbb{R}^d$ centered at the origin and set $\Delta_p(\eta) := \Delta_p(\mathcal{B}_p, \eta) = \sup_{t \geq 0} \mathbb{P}(t < \|T\|_p \leq t + \eta)$.

Proposition C.2.3 (Distributional approximation of martingale ℓ^p-norms)
Assume the setup of Corollary 4.2.2, with Σ non-random. Then for $T \sim \mathcal{N}(0, \Sigma)$,

$$\sup_{t \geq 0} \left|\mathbb{P}(\|S\|_p \leq t) - \mathbb{P}\left(\|T\|_p \leq t\right)\right| \leq \inf_{\eta > 0} \left\{\Gamma_p(\eta) + \Delta_p(\eta)\right\}. \tag{C.8}$$

Proof (Proposition C.2.3)
Applying Proposition C.2.1 with $\mathcal{A} = \mathcal{B}_p$ gives

$$\begin{aligned}\sup_{t \geq 0} \left|\mathbb{P}(\|S\|_p \leq t) - \mathbb{P}\left(\|T\|_p \leq t\right)\right| &= \sup_{A \in \mathcal{B}_p} \left|\mathbb{P}(S \in A) - \mathbb{P}(T \in A)\right| \\ &\leq \inf_{\eta > 0} \left\{\Gamma_p(\eta) + \Delta_p(\mathcal{B}_p, \eta)\right\} \leq \inf_{\eta > 0} \left\{\Gamma_p(\eta) + \Delta_p(\eta)\right\}. \quad \square\end{aligned}$$

The right-hand side of (C.8) can be controlled in various ways. In the case of $p = \infty$, note that ℓ^∞-balls are rectangles so $\mathcal{B}_\infty \subseteq \mathcal{R}$ and (C.7) applies, giving $\Delta_\infty(\eta) \leq \eta(\sqrt{2 \log d} + 2)/\sigma_{\min}$ whenever $\min_j \Sigma_{jj} \geq \sigma^2_{\min}$. Alternatively, Giessing (2023, Theorem 1) provides $\Delta_\infty(\eta) \lesssim \eta / \sqrt{\mathrm{Var}[\|T\|_\infty] + \eta^2}$. By Hölder duality of ℓ^p-norms, we can write $\|T\|_p = \sup_{\|u\|_q \leq 1} u^\mathsf{T} T$

where $1/p + 1/q = 1$. Applying the Gaussian process anti-concentration result of Giessing (2023, Theorem 2) yields the more general $\Delta_p(\eta) \lesssim \eta/\sqrt{\mathrm{Var}[\|T\|_p] + \eta^2}$. Thus, the problem can be reduced to that of bounding $\mathrm{Var}\left[\|T\|_p\right]$, with techniques for doing so discussed in Giessing (2023, Section 4). Alongside the ℓ^p-norms, other functionals can be analyzed in this manner, including the maximum and other order statistics (Kozbur, 2021; Giessing, 2023).

To conduct inference in this setting, we must feasibly approximate the quantiles of $\|T\|_p$. To that end, take a significance level $\tau \in (0,1)$ and set $\hat{q}_p(\tau) = \inf\left\{t \in \mathbb{R} : \mathbb{P}(\|\hat{T}\|_p \leq t \mid \mathbf{X}) \geq \tau\right\}$ where $\hat{T} \mid \mathbf{X} \sim \mathcal{N}(0, \hat{\Sigma})$, with $\hat{\Sigma}$ any $\mathbf{X}$-measurable positive semi-definite estimator of Σ. Note that for the canonical estimator $\hat{\Sigma} = \sum_{i=1}^n X_i X_i^\mathsf{T}$ we can write $\hat{T} = \sum_{i=1}^n X_i Z_i$ with $Z_1, \ldots, Z_n$ i.i.d. standard Gaussian independent of $\mathbf{X}$, yielding the Gaussian multiplier bootstrap. Now assuming the law of $\|\hat{T}\|_p \mid \mathbf{X}$ has no atoms, we can apply Proposition C.2.2 to see

$$\begin{aligned}\sup_{\tau\in(0,1)} \left|\mathbb{P}\left(\|S\|_p \leq \hat{q}_p(\tau)\right) - \tau\right| &\leq \mathbb{E}\left[\sup_{t\geq 0}\left|\mathbb{P}(\|S\|_p \leq t) - \mathbb{P}(\|\hat{T}\|_p \leq t \mid \mathbf{X})\right|\right] \\ &\leq \inf_{\eta>0}\left\{\Gamma_p(\eta) + 2\Delta_p(\eta) + 2d\,\mathbb{E}\left[\exp\left(\frac{-\eta^2}{2d^{2/p}\left\|\hat{\Sigma}^{1/2} - \Sigma^{1/2}\right\|_2^2}\right)\right]\right\},\end{aligned}$$

and hence the bootstrap is valid whenever $\|\hat{\Sigma}^{1/2} - \Sigma^{1/2}\|_2^2$ is sufficiently small. See the preceding discussion regarding methods for bounding this object.

Remark C.2.2 (One-dimensional distributional approximations)

In our application to distributional approximation of ℓ^p-norms, the object of interest $\|S\|_p$ is a one-dimensional functional of the high-dimensional martingale; contrast this with the more general Proposition C.2.1 which directly considers the d-dimensional random vector S. As such, our coupling-based approach may be improved in certain settings by applying a more carefully tailored smoothing argument. For example, Belloni and Oliveira (2018) employ a "log sum exponential" bound (see also Chernozhukov et al., 2013a) for the maximum statistic $\max_{1\leq j\leq d} S_j$ along with a coupling due to Chernozhukov et al. (2014b) to attain an improved dependence on the dimension. Naturally, their approach does not permit the formulation of high-dimensional central limit theorems over arbitrary classes of Borel sets as in our Proposition C.2.1.

Bibliography

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598.

Anastasiou, A., Balasubramanian, K., and Erdogdu, M. A. (2019). Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. In *Conference on Learning Theory*, pp. 115–137. Proceedings of Machine Learning Research.

Arcones, M. A. (1995). A Bernstein-type inequality for U-statistics and U-processes. *Statistics & Probability Letters*, 22(3):239–247.

Arcones, M. A. and Giné, E. (1993). Limit theorems for U-processes. *Annals of Probability*, pp. 1494–1542.

Arnould, L., Boyer, C., and Scornet, E. (2023). Is interpolation benign for random forest regression? In *International Conference on Artificial Intelligence and Statistics*, pp. 5493–5548. Proceedings of Machine Learning Research.

Atchadé, Y. F. and Cattaneo, M. D. (2014). A martingale decomposition for quadratic forms of Markov chains (with applications). *Stochastic Processes and their Applications*, 124(1):646–677.

Baxter, B. J. C. (1994). Norm estimates for inverses of Toeplitz distance matrices. *Journal of Approximation Theory*, 79(2):222–242.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Fernández-Val, I. (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, 213(1):4–29.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Belloni, A. and Oliveira, R. I. (2018). A high dimensional central limit theorem for martingales, with applications to context tree models. *Preprint*. `arXiv:1809.02741`.

Berthet, P. and Mason, D. M. (2006). Revisiting two strong approximation results of Dudley and Philipp. *Lecture Notes–Monograph Series*, 51:155–172. High Dimensional Probability.

Bhatia, R. (1997). *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer, New York, NY.

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

Biau, G. and Mason, D. M. (2015). High-dimensional p-norms. In *Mathematical Statistics and Limit Theorems*, edited by M. Hallin, D. M. Mason, D. Pfeifer, and J. G. Steinebach, pp. 21–40. Springer.

Birgé, L. (2001). An alternative point of view on Lepski's method. *Lecture Notes–Monograph Series*, 36:113–133. State of the Art in Probability and Statistics.

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

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144.

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

Bureau of Meteorology, Australian Government (2017). Daily weather observations. `http://www.bom.gov.au/climate/data/`. Accessed October 2023.

Buzun, N., Shvetsov, N., and Dylov, D. V. (2022). Strong Gaussian approximation for the sum of random vectors. In *Conference on Learning Theory*, volume 178, pp. 1693–1715. Proceedings of Machine Learning Research.

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

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

Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A. (2004). Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning*, p. 18.

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

Cattaneo, M. D., Feng, Y., and Underwood, W. G. (2024). Uniform inference for kernel density estimators with dyadic data. *Journal of the American Statistical Association*, forthcoming.

Cattaneo, M. D., Klusowski, J. M., and Underwood, W. G. (2023). Inference with Mondrian random forests. *Preprint*. `arXiv:2310.09702`.

Cattaneo, M. D., Masini, R. P., and Underwood, W. G. (2022). Yurinskii's coupling for martingales. *Preprint*. `arXiv:2210.00362`.

Chatterjee, S. (2006). A generalization of the Lindeberg principle. *Annals of Probability*, 34(6):2061–2076.

Chen, X. and Kato, K. (2020). Jackknife multiplier bootstrap: finite sample approximations to the U-process supremum with applications. *Probability Theory and Related Fields*, 176(3):1097–1163.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013a). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics*, 41(6):2786–2819.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014a). Anti-concentration and honest, adaptive confidence bands. *Annals of Statistics*, 42(5):1787–1818.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014b). Gaussian approximation of suprema of empirical processes. *Annals of Statistics*, 42(4):1564–1597.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2016). Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related Gaussian couplings. *Stochastic Processes and their Applications*, 126(12):3632–3651.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017a). Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, 45(4):2309–2352.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017b). Detailed proof of Nazarov's inequality. *Preprint*. `arXiv:1711.10696`.

Chernozhukov, V., Chetverikov, D., and Koike, Y. (2023). Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *Annals of Applied Probability*, 33(3):2374–2425.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013b). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.

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

Chiang, H. D., Kato, K., and Sasaki, Y. (2023). Inference for high-dimensional exchangeable arrays. *Journal of the American Statistical Association*, 118(543):1595–1605.

Chiang, H. D. and Tan, B. Y. (2023). Empirical likelihood and uniform convergence rates for dyadic kernel density estimation. *Journal of Business and Economic Statistics*, 41(3):906–914.

Cuny, C. and Merlevède, F. (2014). On martingale approximations and the quenched weak invariance principle. *Annals of Probability*, 42(2):760–793.

Davezies, L., D'Haultfœuille, X., and Guyonvarch, Y. (2021). Empirical process results for exchangeable arrays. *Annals of Statistics*, 49(2):845–862.

de la Peña, V. H. and Montgomery-Smith, S. J. (1995). Decoupling inequalities for the tail probabilities of multivariate U-statistics. *Annals of Probability*, 23(2):806–816.

Dedecker, J., Merlevède, F., and Volnỳ, D. (2007). On the weak invariance principle for non-adapted sequences under projective criteria. *Journal of Theoretical Probability*, 20:971–1004.

Dehling, H. (1983). Limit theorems for sums of weakly dependent Banach space valued random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 63(3):393–432.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach. *Econometrica*, 64(5):1001–1004.

Dudley, R. and Philipp, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 62(4):509–552.

Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

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

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

Eggermont, P. P. B. and LaRiccia, V. N. (2009). *Maximum Penalized Likelihood Estimation: Volume II: Regression*. Springer Series in Statistics. Springer, New York, NY.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, New York, NY.

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020). *Statistical Foundations of Data Science*. Data Science Series. Chapman & Hall/CRC, New York, NY.

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

Gao, C. and Ma, Z. (2021). Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing. *Statistical Science*, 36(1):16–33.

Gao, W., Xu, F., and Zhou, Z.-H. (2022). Towards convergence rate analysis of random forests for classification. *Artificial Intelligence*, 313:103 788.

Giessing, A. (2023). Anti-concentration of suprema of Gaussian processes and Gaussian order statistics. *Preprint*. `arXiv:2310.12119`.

Giné, E., Koltchinskii, V., and Sakhanenko, L. (2004). Kernel density estimators: convergence in distribution for weighted sup-norms. *Probability Theory and Related Fields*, 130(2):167–198.

Giné, E., Latała, R., and Zinn, J. (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II*, edited by E. Giné, D. M. Mason, and J. A. Wellner, pp. 13–38. Birkhäuser, Boston, MA.

Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Annals of Statistics*, 38(2):1122–1170.

Giné, E. and Nickl, R. (2021). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Graham, B. S. (2020). Network data. In *Handbook of Econometrics*, edited by S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin, volume 7, pp. 111–218. Elsevier.

Graham, B. S., Niu, F., and Powell, J. L. (2021). Minimax risk and uniform convergence rates for nonparametric dyadic regression. Technical report, National Bureau of Economic Research.

Graham, B. S., Niu, F., and Powell, J. L. (2024). Kernel density estimation for undirected dyadic data. *Journal of Econometrics*, 240(2).

Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Annals of Statistics*, 20(2):675–694.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and its Application*. Academic Press, New York, NY.

Hall, P. and Kang, K.-H. (2001). Bootstrapping nonparametric density estimators with empirically chosen bandwidths. *Annals of Statistics*, 29(5):1443–1468.

Head, K. and Mayer, T. (2014). Gravity equations: Workhorse, toolkit, and cookbook. In *Handbook of International Economics*, edited by G. Gopinath, E. Helpman, and K. Rogoff, volume 4, pp. 131–195. Elsevier.

Hoover, D. N. (1979). Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ.*

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

Kenny, D. A., Kashy, D. A., and Cook, W. L. (2020). *Dyadic Data Analysis*. Methodology in the Social Sciences Series. Guilford Press.

Khasminskii, R. Z. (1978). A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory of Probability and its Applications*, 23(4):794–798.

Klusowski, J. M. (2021). Sharp analysis of a simple model for random forests. In *International Conference on Artificial Intelligence and Statistics*, pp. 757–765. Proceedings of Machine Learning Research.

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

Koike, Y. (2021). Notes on the dimension dependence in high-dimensional central limit theorems for hyperrectangles. *Japanese Journal of Statistics and Data Science*, 4:257–297.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Series in Statistics. Springer, New York, NY.

Komlós, J., Major, P., and Tusnády, G. (1975). An approximation of partial sums of independent RVs, and the sample DF. I. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32(1-2):111–131.

Kozbur, D. (2021). Dimension-free anticoncentration bounds for Gaussian order statistics with discussion of applications to multiple testing. *Preprint*. `arXiv:2107.10766`.

Kwapien, S. and Szulga, J. (1991). Hypercontraction methods in moment inequalities for series of independent random variables in normed spaces. *Annals of Probability*, 19(1):369–379.

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

Lakshminarayanan, B., Roy, D. M., and Teh, Y. W. (2016). Mondrian forests for large-scale regression when uncertainty matters. In *Artificial Intelligence and Statistics*, pp. 1478–1487. Proceedings of Machine Learning Research.

Laurent, M. and Rendl, F. (2005). Semidefinite programming and integer programming. In *Discrete Optimization*, edited by K. Aardal, G. L. Nemhauser, and R. Weismantel, volume 12 of *Handbooks in Operations Research and Management Science*, pp. 393–514. Elsevier.

Le Cam, L. (1988). On the Prokhorov distance between the empirical process and the associated Gaussian bridge. Technical report, University of California, Berkeley.

Le Gall, J.-F. (2016). *Brownian Motion, Martingales, and Stochastic Calculus*, volume 274 of *Graduate Texts in Mathematics*. Springer, Berlin, Heidelberg.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*. Classics in Mathematics. Springer, Berlin, Heidelberg.

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

Li, J. and Liao, Z. (2020). Uniform nonparametric inference for time series. *Journal of Econometrics*, 219(1):38–51.

Lopes, M. E. (2022). Central limit theorem and bootstrap approximation in high dimensions: Near $1/n$ rates via implicit smoothing. *Annals of Statistics*, 50(5):2492–2513.

Lopes, M. E., Lin, Z., and Müller, H.-G. (2020). Bootstrapping max statistics in high dimensions: Near-parametric rates under weak variance decay and application to functional and multinomial data. *Annals of Statistics*, 48(2):1214–1229.

Luke, D. A. and Harris, J. K. (2007). Network analysis in public health: history, methods, and applications. *Annual Review of Public Health*, 28:69–93.

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

Magda, P. and Zhang, N. (2018). Martingale approximations for random fields. *Electronic Communications in Probability*, 23(28):1–9.

Matsushita, Y. and Otsu, T. (2021). Jackknife empirical likelihood: small bandwidth, sparse network and high-dimensional asymptotics. *Biometrika*, 108(3):661–674.

McLeish, D. L. (1975). Invariance principles for dependent variables. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32(3):165–178.

Merlevède, F., Peligrad, M., and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High Dimensional Probability V, the Luminy volume*, edited by C. Houdré, V. Koltchinskii, D. M. Mason, and M. Peligrad, pp. 273–292. Institute of Mathematical Statistics.

Minsker, S. and Wei, X. (2019). Moment inequalities for matrix-valued U-statistics of order 2. *Electronic Journal of Probability*, 24(133):1–32.

MOSEK ApS (2021). *The MOSEK Optimizer API for C manual. Version 9.3.*

Mourtada, J., Gaïffas, S., and Scornet, E. (2017). Universal consistency and minimax rates for online Mondrian forests. *Advances in Neural Information Processing Systems*, 30.

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

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

Nazarov, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In *Geometric Aspects of Functional Analysis*, edited by V. D. Milman and G. Schechtman, pp. 169–187. Springer.

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

Peligrad, M. (2010). Conditional central limit theorem via martingale approximation. In *Dependence in Probability, Analysis and Number Theory, volume in memory of Walter Philipp*, edited by I. Berkes, R. C. Bradley, H. Dehling, M. Peligrad, and R. Tichy, pp. 295–311. Kendrick Press.

Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Rakhlin, A., Sridharan, K., and Tewari, A. (2015). Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1):111–153.

Ray, K. and van der Vaart, A. (2021). On the Bernstein–von Mises theorem for the Dirichlet process. *Electronic Journal of Statistics*, 15(1):2224–2246.

Rio, E. (2017). *Asymptotic Theory of Weakly Dependent Random Processes*, volume 80 of *Probability Theory and Stochastic Modelling*. Springer, Berlin, Heidelberg.

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

Royden, H. L. and Fitzpatrick, P. (1988). *Real Analysis*. Macmillan, New York, NY.

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

Scillitoe, A., Seshadri, P., and Girolami, M. (2021). Uncertainty quantification for data-driven turbulence modelling with Mondrian forests. *Journal of Computational Physics*, 430:110 116.

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

Settati, A. (2009). Gaussian approximation of the empirical process under random entropy conditions. *Stochastic Processes and their Applications*, 119(5):1541–1560.

Sheehy, A. and Wellner, J. A. (1992). Uniform Donsker classes of functions. *Annals of Probability*, 20(4):1983–2030.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer Series in Statistics. Springer Science, New York, NY.

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

van de Geer, S. and Lederer, J. (2013). The Bernstein–Orlicz norm and deviation inequalities. *Probability Theory and Related Fields*, 157(1):225–250.

van de Geer, S. A. (2000). *Empirical Processes in M-Estimation*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, New York, NY.

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

Wand, M. P. and Jones, M. C. (1994). *Kernel Smoothing*, volume 60 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, New York, NY.

Wu, W. B. and Woodroofe, M. (2004). Martingale approximations for sums of stationary processes. *Annals of Probability*, 32(2):1674–1690.

Yurinskii, V. V. (1978). On the error of the Gaussian approximation for convolutions. *Theory of Probability & its Applications*, 22(2):236–247.

Zaitsev, A. Y. (1987a). Estimates of the Lévy–Prokhorov distance in the multivariate central limit theorem for random variables with finite exponential moments. *Theory of Probability & Its Applications*, 31(2):203–220.

Zaitsev, A. Y. (1987b). On the Gaussian approximation of convolutions under multidimensional analogues of S. N. Bernstein's inequality conditions. *Probability Theory and Related Fields*, 74(4):535–566.

Zhai, A. (2018). A high-dimensional CLT in $\mathcal{W}_2$ distance with near optimal convergence rate. *Probability Theory and Related Fields*, 170(3):821–845.

Zhao, O. and Woodroofe, M. (2008). On martingale approximations. *Annals of Applied Probability*, 18(5):1831–1847.

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