Yurinskii's Coupling for Martingales

Matias D. Cattaneo[1] Ricardo P. Masini[2] William G. Underwood[1]

October 2, 2022

Abstract

Yurinskii's coupling is a popular tool for finite-sample distributional approximation in mathematical statistics and applied probability, offering a Gaussian strong approximation for sums of random vectors under easily verified conditions with an explicit rate of approximation. Originally stated for sums of independent random vectors in ℓ^2-norm, it has recently been extended to the ℓ^p-norm, where $1 \leq p \leq \infty$, and to vector-valued martingales in ℓ^2-norm under some rather strong conditions. We provide as our main result a generalization of all of the previous forms of Yurinskii's coupling, giving a Gaussian strong approximation for martingales in ℓ^p-norm under relatively weak conditions. We apply this result to some areas of statistical theory, including high-dimensional martingale central limit theorems and uniform strong approximations for martingale empirical processes. Finally we give a few illustrative examples in statistical methodology, applying our results to partitioning-based series estimators for nonparametric regression, distributional approximation of ℓ^p-norms of high-dimensional martingales, and local polynomial regression estimators. We address issues of feasibility, demonstrating implementable statistical inference procedures in each section.

Keywords: coupling, martingales, strong approximation, dependent data, time series, high-dimensional central limit theorems, empirical processes, uniform inference, series estimation, local polynomial estimation.

[1]Department of Operations Research and Financial Engineering, Princeton University.
[2]Center for Statistics and Machine Learning, Princeton University.

Contents

1 Introduction

Yurinskii's coupling (Yurinskii, 1978) has proven to be an important tool for developing non-asymptotic distributional approximations in high-dimensional 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 d-dimensional Gaussian vector T with the same covariance structure as S ($\mathrm{Var}[S] = \mathrm{Var}[T]$) and which is close to S in probability, bounding the discrepancy $\|S - T\|$ as a function of n, d, the norm used 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 coupling stands out for its simplicity, robustness and wide applicability, while also offering tighter couplings in some applications (see below for more discussion and examples). These features have led many authors 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 (i) to construct Gaussian approximations for the suprema of empirical processes (Chernozhukov et al., 2014b); (ii) to establish distribution theory for non-Donsker stochastic t-processes generated in nonparametric series regression (Belloni et al., 2015); (iii) to prove distributional approximations for high-dimensional ℓ^p-norms (Biau and Mason, 2015); (iv) to derive a law of the iterated logarithm for stochastic gradient descent optimization methods (Anastasiou et al., 2019); (v) to establish uniform distribution theory for nonparametric high-dimensional quantile processes (Belloni et al., 2019); (vi) to develop distribution theory for non-Donsker stochastic t-processes generated in partitioning-based series regression (Cattaneo et al., 2020); (vii) to deduce Bernstein–von Mises theorems in high-dimensions (Ray and van der Vaart, 2021); and (viii) to develop distribution theory for non-Donsker U-processes based on dyadic network data (Cattaneo et al., 2022). 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.

One limitation of Yurinskii's coupling, which also afflicts the other coupling methods mentioned above, is that it can only be applied to sums of independent random vectors. This paper addresses said shortcoming by presenting a new Yurinskii coupling for sums of vector-valued martingale differences (Hall and Heyde, 2014). We then harness this result to obtain general-purpose Gaussian strong approximations for high-dimensional random vectors and stochastic processes based on martingale data. A key feature of our Yurinskii coupling for martingales is that it does not impose any restriction on the eigenvalues of the variance of the vector-valued martingale. As such, our result improves upon that of Li and Liao (2020), who recently established a Yurinskii-type coupling for martingales under the assumption that the minimum eigenvalue of the variance is bounded away from zero. As discussed below, their possibly high-dimensional eigenvalue restriction can be difficult or impossible to verify in some important applications.

The main coupling result of this paper (Theorem 1) is presented in Section 2, where we also specialize it in a slightly weaker formulation, which is much easier to use (Proposition 1). Our Yurinskii coupling for martingales is a strict generalization of all of the previous Yurinskii couplings available in the literature, offering a Gaussian strong approximation for martingale vectors in ℓ^p-norm with no assumptions on the spectrum of the variance matrix. The key innovation underlying the proof of Theorem 1 is that we explicitly account for the possibility that the minimum eigenvalue of the variance may be zero or that its lower bound may be unknown, with the argument proceeding by means of a carefully tailored regularization. Proposition 1 then explicitly tunes the regularization parameter to obtain a simpler and regularization-free Yurinskii coupling for martingales. This specialization of our main result takes an agnostic approach to potential singularities in the variance of the martingale, and as such may be improved in specific applications where additional knowledge of the covariance structure is available.

Section 3 illustrates the broad applicability of our Yurinskii coupling for martingale vectors with two substantive applications to statistical theory. Firstly, we obtain a distributional Gaussian approximation for possibly high-dimensional martingale vectors (Proposition 2). This result complements a recent literature on probability and statistics studying the same problem but with independent data (see Buzun et al., 2022; Chernozhukov et al., 2022, and references therein). Our result also improves upon Belloni and Oliveira (2018) by offering a simpler central limit theorem for high-dimensional martingales which is easier to apply. Secondly, we present a general-purpose strong approximation for martingale empirical processes (Proposition 4), combining classical results in the empirical process literature (van der Vaart and Wellner, 1996) with our Proposition 1. This statement appears to be the first of its kind for martingale data, and when specialized to independent data it is shown to be superior to the best known 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 with the larger ℓ_2-norm.

Section 4 is dedicated to showcasing some more practical examples which build upon the preceding theory. Firstly, we apply our main result (Proposition 1) directly to deduce a strong approximation for partitioning-based least squares series estimators based on time series data, additionally imposing only a mild mixing condition on the regressors. We show that our Yurinskii coupling for martingale vectors delivers the same approximation rate as the best known rate for independent data, and then discuss how the strong approximation can be leveraged to yield a feasible statistical inference procedure. Secondly, we apply our martingale central limit theorem (Proposition 2) to deduce Gaussian-based approximations of martingale ℓ^p-norms in Kolmogorov–Smirnov distance, relying on recent results concerning Gaussian perimetric inequalities. Thirdly, we use our result on martingale empirical processes (Proposition 4) to deduce a strong approximation for local polynomial estimators (Fan and Gijbels, 1996) with time series data, again imposing a mild mixing assumption. The bandwidth restrictions we require are relatively mild, and as far as we know have not been improved upon even with independent data.

Finally, Section 5 concludes the paper. All proofs are collected in Appendix A, which also includes other technical lemmas of potential independent interest.

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 $\vert\kern-0.25ex\vert\kern-0.25ex\vert X \vert\kern-0.25ex\vert\kern-0.25ex\vert_\psi$ to denote the Orlicz ψ-norm (van der Vaart and Wellner, 1996, Section 2.2) and $\vert\kern-0.25ex\vert\kern-0.25ex\vert X \vert\kern-0.25ex\vert\kern-0.25ex\vert_p$ 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 $M \succeq 0$.

For scalar sequences x_n and y_n, 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. Write $x_n \asymp y_n$ to indicate both $x_n \lesssim y_n$ and $y_n \lesssim x_n$, and $x_n \to x$ for limits. Similarly, for random variables X_n and Y_n, 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 $X_n \to_{\mathbb{P}} X$ for limits in probability. For real numbers a and b we use $a \vee b = \max\{a, b\}$.

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).

2 Main results

We begin with our Yurinskii-type Gaussian strong approximation for vector-valued martingales in ℓ^p-norm. Our main result is presented in Theorem 1 while Proposition 1 gives a simplified and slightly weaker version which is easier to use in applications.

Theorem 1 (Strong approximation for martingale vectors)
Let $X_1, \ldots, X_n$ be $\mathbb{R}^d$-valued square-integrable random variables adapted to a filtration $\mathcal{H}_1, \ldots, \mathcal{H}_n$, with $\mathcal{H}_0$ the trivial σ-algebra. Suppose $\mathbb{E}[X_i \mid \mathcal{H}_{i-1}] = 0$ for all $1 \leq i \leq n$ and let $V_i = \mathrm{Var}[X_i \mid \mathcal{H}_{i-1}]$. Define the martingale $S = \sum_{i=1}^n X_i$ and let $\Sigma = \mathrm{Var}[S]$ and $\Omega = \sum_{i=1}^n \big(V_i - \mathbb{E}[V_i]\big)$. Then for each $\eta > 0$ and $p \in [1, \infty]$ there exists $T \sim \mathcal{N}(0, \Sigma)$ such that

$$\mathbb{P}\big(\|S - T\|_p > 6\eta\big) \leq 2 \inf_{t>0} \left\{ \alpha_p(t) + \frac{\beta_p}{\eta^3} t^2 \right\} + \inf_{M \succeq 0} \big\{ 2\gamma(M) + \delta_p(M, \eta) + \varepsilon_p(M, \eta) \big\}, \tag{1}$$

where the second infimum is taken over all positive semi-definite $d \times d$ non-random matrices, and

$$\alpha_p(t) = \mathbb{P}\big(\|Z\|_p > t\big), \qquad \beta_p = \sum_{i=1}^n \mathbb{E}\left[\|X_i\|_2^2 \|X_i\|_p + \|V_i^{1/2} Z_i\|_2^2 \|V_i^{1/2} Z_i\|_p \right],$$

$$\gamma(M) = \mathbb{P}\big(\Omega \npreceq M\big), \qquad \delta_p(M, \eta) = \mathbb{P}\left(\big\| \big((\Sigma + M)^{1/2} - \Sigma^{1/2}\big) Z \big\|_p \geq \eta \right),$$

$$\varepsilon_p(M, \eta) = \mathbb{P}\left(\big\| (M - \Omega)^{1/2} Z \big\|_p \geq \eta, \ \Omega \preceq M \right),$$

with $Z, Z_1, \ldots, Z_n$ i.i.d. standard Gaussian variables on $\mathbb{R}^d$ independent of $\mathcal{H}_n$.

The second term on the right-hand side of (1) controls for the randomness of the quadratic variation $\sum_{i=1}^n V_i$. If this quantity is almost surely constant then $\Omega = 0 \in \mathbb{R}^{d \times d}$ a.s. and we may take $M = 0 \in \mathbb{R}^{d \times d}$ so that (1) simplifies to

$$\mathbb{P}\big(\|S - T\|_p > 6\eta\big) \leq 2 \inf_{t>0} \left\{ \alpha_p(t) + \frac{\beta_p}{\eta^3} t^2 \right\}.$$

If further $V_i = \mathbb{E}[V_i]$ almost surely for every $1 \leq i \leq n$ then we recover the same bound as for independent random variables, namely Yurinskii's coupling for $p = 2$ as in Yurinskii (1978) and Pollard (2002, Theorem 10); and the more general version for $p \in [1, \infty]$ given by Belloni et al. (2019, Lemma 38).

More broadly, the second term on the right-hand side of (1) emerges from a regularization scheme designed to account for potential degeneracy of the variance Σ of the martingale. Setting $M = \nu^2 I_d$ in Theorem 1 and minimizing the right-hand side of (1) over $t, \nu > 0$ yields the following proposition, which is arguably the main result of our paper because of its simplicity and utility in statistical applications.

Proposition 1 (Simplified strong approximation for martingale vectors)
Assume the setup of Theorem 1. For each $\eta > 0$ and $p \in [1, \infty]$ there exists $T \sim \mathcal{N}(0, \Sigma)$ such that

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

where $\phi_p(d) = \sqrt{p d^{2/p}}$ for $p \in [1, \infty)$ and $\phi_\infty(d) = \sqrt{2 \log 2d}$, and

$$\beta_p = \sum_{i=1}^n \mathbb{E}\left[\|X_i\|_2^2 \|X_i\|_p + \|V_i^{1/2} Z_i\|_2^2 \|V_i^{1/2} Z_i\|_p \right].$$

In particular, this implies that

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

The probability bound in Proposition 1 gives the same rate of strong approximation as that in Theorem 1 of Li and Liao (2020) when $p = 2$. However we make no assumptions about the non-degeneracy of S, whereas Li and Liao (2020) impose a lower bound on the eigenvalues of Σ.

In Section 4.1 we use Proposition 1 to obtain strong approximations for partitioning-based series estimators in the nonparametric regression setting. Despite the fact that such estimators are function-valued rather than vector-valued, Proposition 1 still applies due to a certain linear separability property of the underlying structure of the estimators.

3 Applications to statistical theory

In Section 3.1 we present two substantive applications of our main result to high-dimensional central limit theorems for martingales. Proposition 2 reduces the problem to that of establishing anti-concentration results for Gaussian vectors and Proposition 3 demonstrates a feasible implementation via the Gaussian multiplier bootstrap. In Section 3.2 we deduce results for strong approximation of martingale empirical processes. These processes may be indexed by functions, and we state our results in terms of metric entropy under Orlicz norms.

3.1 High-dimensional central limit theorems for martingales

We begin this section with some notation. Let $\mathcal{A}$ be a class of measurable subsets of $\mathbb{R}^d$ and $T \sim \mathcal{N}(0,\Sigma)$ be as in Theorem 1. For $\eta > 0$ and $p \in [1,\infty]$ define the Gaussian perimetric quantity

$$\Delta_p(\mathcal{A},\eta) = \sup_{A\in\mathcal{A}} \big\{\mathbb{P}(T \in A_p^\eta \setminus A) \vee \mathbb{P}(T \in A \setminus A_p^{-\eta})\big\},$$

where $A_p^\eta = \{x \in \mathbb{R}^d : \|x - A\|_p \le \eta\}$ and $A_p^{-\eta} = \mathbb{R}^d \setminus (\mathbb{R}^d \setminus A)_p^\eta$ and $\|x-A\|_p = \inf_{x'\in A}\|x-x'\|_p$. Denote by $\Gamma_p(\eta)$ the rate of strong approximation attained in Proposition 1:

$$\Gamma_p(\eta) = 24\left(\frac{\beta_p\phi_p(d)^2}{\eta^3}\right)^{1/3} + 17\left(\frac{\phi_p(d)\sqrt{\mathbb{E}[\|\Omega\|_2]}}{\eta}\right)^{2/3}.$$

Proposition 2 (High-dimensional central limit theorem for martingales)
Assume the same setup as in Theorem 1. For any class $\mathcal{A}$ of measurable subsets of $\mathbb{R}^d$,

$$\sup_{A\in\mathcal{A}} \big|\mathbb{P}(S\in A) - \mathbb{P}(T\in A)\big| \le \inf_{p\in[1,\infty]}\inf_{\eta>0}\big\{\Gamma_p(\eta) + \Delta_p(\mathcal{A},\eta)\big\}. \tag{2}$$

Note that the term $\Delta_p(\mathcal{A},\eta)$ in (2) 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, in the case $\mathcal{A} = \mathcal{C} = \{A \subseteq \mathbb{R} \text{ is convex}\}$, Nazarov (2003) showed

$$\Delta_2(\mathcal{C},\eta) \asymp \eta\sqrt{\|\Sigma^{-1}\|_{\mathrm{F}}} \tag{3}$$

whenever Σ is invertible. Then Proposition 2 with $p = 2$ combined with (3) yields

$$\sup_{A\in\mathcal{C}} \big|\mathbb{P}(S\in A) - \mathbb{P}(T\in A)\big| \lesssim \inf_{\eta>0}\left\{\left(\frac{\beta_2 d}{\eta^3}\right)^{1/3} + \left(\frac{\sqrt{d\mathbb{E}[\|\Omega\|_2]}}{\eta}\right)^{2/3} + \eta\|\Sigma^{-1}\|_{\mathrm{F}}\right\}. \tag{4}$$

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; Buzun et al., 2022),

$$\Delta_\infty(\mathcal{R}, \eta) \leq \frac{\eta(\sqrt{2\log d} + 2)}{\sigma_{\min}} \tag{5}$$

if $\min_j \Sigma_{jj} \geq \sigma_{\min}^2$ for some $\sigma_{\min} > 0$. Then Proposition 2 with $p = \infty$ along with (5) yields

$$\sup_{A\in\mathcal{R}} \left|\mathbb{P}(S \in A) - \mathbb{P}(T \in A)\right| \lesssim \inf_{\eta>0} \left\{ \left(\frac{\beta_\infty \log 2d}{\eta^3}\right)^{1/3} + \left(\frac{\sqrt{\mathbb{E}[\|\Omega\|_2] \log 2d}}{\eta}\right)^{2/3} + \frac{\eta\sqrt{\log 2d}}{\sigma_{\min}} \right\}. \tag{6}$$

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 et al. (2022).

The literature on Gaussian approximations for high-dimensional sums of independent random vectors has developed rapidly in recent years (see Buzun et al., 2022; Chernozhukov et al., 2022, and references therein). For example, if X_i are i.i.d. and bounded a.s. and if $\mathrm{Var}[X_i]$ has minimum eigenvalue bounded away from zero, Chernozhukov et al. (2022, Theorem 2.1) recently showed

$$\sup_{A\in\mathcal{R}} \left|\mathbb{P}(S \in A) - \mathbb{P}(T \in A)\right| \lesssim \frac{(\log d)^{5/2}}{\sqrt{n}}.$$

For comparison, under the same conditions, we have $\Omega = 0$ by independence, $\beta_\infty \lesssim nd\sqrt{\log d}$ and $\sigma_{\min} \gtrsim \sqrt{n}$; our closest comparable result (6) yields

$$\sup_{A\in\mathcal{R}} \left|\mathbb{P}(S \in A) - \mathbb{P}(T \in A)\right| \lesssim \left(\frac{d(\log d)^3}{\sqrt{n}}\right)^{1/6}.$$

Thus our results are substantially weaker for i.i.d. data, an inherent issue due to our approach of first constructing a coupling for the high-dimensional vector S and only then specializing to a distributional approximation. In contrast, sharper results in the literature directly target the distribution via Stein's method and Slepian interpolation. The main contribution of this section is therefore to obtain a Gaussian distributional approximation for sums of high-dimensional martingale difference data, for which alternative high-dimensional central limit theorem proof strategies are not readily available.

We remark that it may be possible to somewhat improve our approach to high-dimensional central limit theorems by adjusting the proof of Theorem 1. If the family of sets under consideration (e.g., the class of rectangles $\mathcal{R}$) is substantially smaller than the family of Borel sets, one might be able to improve the smoothing argument (Lemma 2) on this smaller class and skip the Strassen argument entirely (Lemma 1).

Next, we present a version of Proposition 2 where the covariance matrix Σ is replaced by an estimator $\widehat{\Sigma}$. This ensures that the associated conditionally Gaussian vector is feasible and can be resampled, allowing Monte Carlo estimation with a Gaussian multiplier bootstrap.

Proposition 3 (Bootstrap central limit theorem for martingales)
Assume the same setup as in Theorem 1 and let $\widehat{\Sigma}$ be an $\mathbf{X}$-measurable random $d \times d$ matrix which is a.s. positive semi-definite, where $\mathbf{X} = (X_1, \ldots, X_n)$. For any class $\mathcal{A}$ of measurable subsets of $\mathbb{R}^d$,

$$\begin{aligned}
&\sup_{A\in\mathcal{A}} \left|\mathbb{P}(S \in A) - \mathbb{P}\big(\widehat{\Sigma}^{1/2} Z \in A \mid \mathbf{X}\big)\right| \\
&\quad \leq \inf_{p\in[1,\infty]} \inf_{\eta>0} \left\{ \Gamma_p(\eta) + 2\Delta_p(\mathcal{A}, \eta) + 2d \exp\left(\frac{-\eta^2}{2d^{2/p}\big\|\widehat{\Sigma}^{1/2} - \Sigma^{1/2}\big\|_2^2}\right) \right\}
\end{aligned}$$

where $Z \sim \mathcal{N}(0, I_d)$ is independent of $\mathbf{X}$.

A natural choice for $\widehat{\Sigma}$ is the sample covariance matrix $\sum_{i=1}^n X_i X_i^\mathsf{T}$. In general, whenever $\widehat{\Sigma}$ does not depend on unknown quantities, we can sample from the law of $\widehat{T} = \widehat{\Sigma}^{1/2} Z$ conditional on $\mathbf{X}$ to approximate the distribution of S. Proposition 3 verifies that this Gaussian multiplier bootstrap approach is valid whenever $\widehat{\Sigma}$ and Σ are sufficiently close. To this end, Theorem X.1.1 in Bhatia (1997) gives $\big\|\widehat{\Sigma}^{1/2} - \Sigma^{1/2}\big\|_2 \leq \big\|\widehat{\Sigma} - \Sigma\big\|_2^{1/2}$ and Problem X.5.5 in Bhatia (1997) gives $\big\|\widehat{\Sigma}^{1/2} - \Sigma^{1/2}\big\|_2 \leq \big\|\Sigma^{-1/2}\big\|_2 \big\|\widehat{\Sigma} - \Sigma\big\|_2$ when Σ is invertible. The latter often gives a tighter bound when the minimum eigenvalue of Σ can be bounded away from zero.

In Section 4.2 we apply Proposition 2 to the special case of approximating the distribution of the ℓ^p-norm of a high-dimensional martingale. Proposition 3 is then used to ensure that feasible distributional approximations are also available.

3.2 Strong approximation for martingale empirical processes

Our next result gives a strong approximation for martingale empirical processes, obtained by applying Proposition 1 with $p = \infty$ to a discretization of the empirical process. We 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 et al., 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 $\mathcal{F}$, we use the notation $\mathcal{F}(x) = \big(f(x) : f \in \mathcal{F}\big)$. Define the set of Orlicz functions

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

Proposition 4 (Strong approximation for martingale empirical processes)
Let X_i be random variables for $1 \leq i \leq n$ taking values in a measurable space $\mathcal{X}$. Let $\mathcal{F}$ be a pointwise measurable class of functions from $\mathcal{X}$ to $\mathbb{R}$. For each i let $\mathcal{H}_i$ be a σ-algebra such that $X_1, \ldots, X_i$ are $\mathcal{H}_i$-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 suppose that we have, for some non-random metric d on $\mathcal{F}$, constant L and $\psi \in \Psi$,

$$\big\|S(f) - S(f')\big\|_2 + \big\|S(f) - S(f')\big\|_\psi \leq L d(f, f'). \tag{7}$$

Then for all $t, \eta > 0$ there exists a zero-mean Gaussian process $T(f)$ indexed by $f \in \mathcal{F}$ satisfying $\mathbb{E}\big[S(f)S(f')\big] = \mathbb{E}\big[T(f)T(f')\big]$ for all $f, f' \in \mathcal{F}$ and

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

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

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

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

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

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

Proposition 4 is given in a rather general form to accommodate different settings and applications. In particular, as is well known in the literature, consider the following common 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: $\psi(x) = \exp\Big(\Big(\frac{\sqrt{1+2ax}-1}{a}\Big)^2\Big) - 1$ for $a > 0$ has $\|\!|X|\!\|_2 \leq (1+a)\|\!|X|\!\|_\psi$ and $\sqrt{\log x} \leq \psi^{-1}(x)$.

For these Orlicz functions the first term in (7) can be controlled by bounding the second 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.

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 handle fluctuations in the stochastic processes with uniform L^2 covering numbers and bracketing numbers where we opt for chaining with Orlicz norms. Our version using the (martingale) Yurinskii coupling can improve upon theirs in approximation rate for independent data under certain circumstances, as follows. Suppose the setup of Proposition 1 in Berthet and Mason (2006); that is, $X_1, \ldots, X_n$ 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. Then using uniform L^2 covering numbers rather than Orlicz norm chaining in our Proposition 4 gives the following. Firstly as X_i are i.i.d. we have $\Omega_\delta = 0$. Let $\mathcal{F}_\delta$ be a minimal δ-cover of $(\mathcal{F}, d_\mathbb{P})$ with cardinality N_δ where $\delta \to 0$. By Lemma 6 we have $\beta_\delta \lesssim n\delta^{-\nu_0}\sqrt{\log(1/\delta)}$. Theorem 2.14.1 and Theorem 2.2.8 in van der Vaart and Wellner (1996) give

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

where we used the VC-type property to bound the entropy integrals. Therefore

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

where we minimized over δ in the last step. In contrast, Berthet and Mason (2006) achieve

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

Comparing these shows 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 parametrized 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) are compelled to use the inequality $\|\cdot\|_\infty \leq \|\cdot\|_2$ in the coupling step, a bound which is loose when the dimension (here on the order of $\delta^{-\nu_0}$) is even moderately large. We exploit the fact that our version of Yurinskii's coupling applies directly to the supremum norm, yielding much sharper dependence on the dimension.

A Gaussian multiplier bootstrap analog of Proposition 4 could be given to parallel Proposition 3, but is omitted to conserve space; there is no fundamental innovation relative to Proposition 4 and the statement would be cumbersome.

In Section 4.3 we apply Proposition 4 to obtain strong approximations for local polynomial estimators in the nonparametric regression setting. In contrast with the series estimators of Section 4.1, local polynomial estimators are not linearly separable and hence cannot be analyzed directly using Proposition 1.

4 Illustrative examples

We illustrate the applicability of our previous results with three distinct examples. In the first, an analysis of partitioning-based series estimators for nonparametric regression, we are able to apply Proposition 1 directly due to the intrinsic linear separability of the estimator. The second relies on Propositions 2 and 3 and concerns the distributional approximation of ℓ^p-norms of high-dimensional martingale vectors. In the third and final example we consider local polynomial estimators for nonparametric regression, using Proposition 4 due to the presence of a non-linearly separable martingale empirical process.

4.1 Partitioning-based series estimators

Partitioning-based least squares methods are important 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 details and references throughout this section. We illustrate the usefulness of Proposition 1 by deriving a Gaussian strong approximation for partitioning series estimators based on multivariate martingale data. Proposition 5 shows that 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 $Y_i = \mu(W_i) + \varepsilon_i$ 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 $p(W_i)$ be a k-dimensional vector of bounded basis functions on $\mathcal{W}$ which are locally-supported on a quasi-uniform partition. Under minimal regularity conditions, the least-squares partitioning-based series estimator is $\widehat{\mu}(w) = p(w)^\mathsf{T} \widehat{H}^{-1} \sum_{i=1}^n p(W_i) Y_i$ with $\widehat{H} = \sum_{i=1}^n p(W_i) p(W_i)^\mathsf{T}$. The approximation power of the estimator $\widehat{\mu}(w)$ derives from letting $k \to \infty$ as $n \to \infty$. With decision trees, for example, $p(w)$ is comprised of indicator functions over k axis-aligned rectangles forming a partition of $\mathcal{W}$ (a Haar basis).

Our goal is to approximate the law of the stochastic process $(\widehat{\mu}(w) - \mu(w) : w \in \mathcal{W})$, which is not asymptotically tight and thus does not converge weakly. Nevertheless, exploiting the intrinsic linearity of the estimator $\widehat{\mu}(w)$, we can apply Proposition 1 directly to construct a Gaussian strong approximation. Specifically,

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

where $H = \sum_{i=1}^n \mathbb{E}[p(W_i)p(W_i)^\mathsf{T}]$ is the regressor variance matrix, $S = \sum_{i=1}^n p(W_i)\varepsilon_i$ is the score vector and $\mathrm{Bias}(w) = p(w)^\mathsf{T} \widehat{H}^{-1} \sum_{i=1}^n p(W_i)\mu(W_i) - \mu(w)$. Imposing some mild time series restrictions and assuming stationarity for simplicity, it is not difficult to show that $\|\widehat{H} - H\|_2 \lesssim_\mathbb{P} \sqrt{nk}$ and $\sup_{w \in \mathcal{W}} |\mathrm{Bias}(w)| \lesssim_\mathbb{P} k^{-\gamma}$ 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. Thus it remains to study the mean-zero martingale S by applying Proposition 1 with $X_i = p(W_i)\varepsilon_i$. Controlling the convergence of the quadratic variation term $\mathbb{E}[\|\Omega\|_2]$ also requires some time series dependence assumptions; we impose an α-mixing condition on $(W_1, \ldots, W_n)$ for illustration.

Proposition 5 (Strong approximation for partitioning series estimators)
Consider the nonparametric regression setup described above and further assume the following:

(i) $(W_i, \varepsilon_i)_{1 \leq i \leq n}$ *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) $p(w)$ *forms a basis with k features satisfying Assumptions 2 and 3 in Cattaneo et al. (2020).*

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

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

provided that the number of basis functions satisfies $k^3/n \to 0$.

The core of the proof of Proposition 5 involves applying Proposition 1 with $S = \sum_{i=1}^n p(W_i)\varepsilon_i$ to construct $T \sim \mathcal{N}\big(0, \mathrm{Var}[S]\big)$ such that $\|S - T\|_\infty \lesssim_\mathbb{P} \big(n^{1/3} + (nk)^{1/4}\big)\sqrt{\log k}$. 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); we do not require any extra restrictions for α-mixing time series compared with i.i.d. data. The assumptions made on $p(w)$ are mild enough to accommodate splines, wavelets, piecewise polynomials and decision trees constructed independently of the data. Furthermore, in the case of martingale data, our result improves on Li and Liao (2020, Theorem 2) by offering faster strong approximation rates under weaker sufficient conditions.

To illustrate the statistical applicability of Proposition 5, 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 5 and anti-concentration for Gaussian suprema (Chernozhukov et al., 2014a, Corollary 2.1) can be combined to obtain a distributional approximation for the supremum statistic whenever $k^3(\log n)^6/n \to 0$, giving

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

where $\rho(w, w') = \mathbb{E}[G(w)G(w')]$. Furthermore, using a Gaussian–Gaussian comparison result (Chernozhukov et al., 2013, Lemma 3.1) and anti-concentration again, it is not difficult to show (see the proof of Proposition 5) that with $\mathbf{W} = (W_1, \ldots, W_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$,

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

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

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

where

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

is the conditional quantile of the supremum of the Studentized Gaussian process. This quantile can be estimated by Monte Carlo resampling, drawing from the conditional law of $\widehat{G}(w) \mid \mathbf{W}, \mathbf{Y}$ with an appropriate discretization of $w \in \mathcal{W}$.

We take the opportunity here to compare our rate of strong approximation for series estimation with that of Li and Liao (2020). Using the notation of our Theorem 1, they derive the martingale Yurinskii coupling

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

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 Ω term. Under the assumptions of our Proposition 5, applying this result with $S = \sum_{i=1}^n p(W_i)\varepsilon_i$ yields a rate no better than $\|S - T\|_2 \lesssim_{\mathbb{P}} (nk)^{1/3}$. As such, they attain a rate of strong approximation no faster than

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

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 version of Yurinskii's coupling rather than the more recently established ℓ^∞ version.

4.2 Distributional approximation of martingale ℓ^p-norms

In some empirical applications, including nonparametric significance tests (Lopes et al., 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 $p \in [1, \infty]$, possibly in high dimension. In the notation of Theorem 1, it is therefore of interest to bound quantities of the form

$$\sup_{t \geq 0} \big| \mathbb{P}(\|S\|_p \leq t) - \mathbb{P}(\|T\|_p \leq t) \big|.$$

We use this setup to illustrate the applicability of Propositions 2 and 3. Using the notation of Section 3.1, let $\mathcal{B}_p$ be the class of closed ℓ^p-balls 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 6 (Distributional approximation of martingale ℓ^p-norms)
Assume the setup of Theorem 1 and define $\Gamma_p(\eta)$ as in Section 3.1. Then for $T \sim \mathcal{N}(0, \Sigma)$,

$$\sup_{t \geq 0} \left| \mathbb{P}(\|S\|_p \leq t) - \mathbb{P}\left(\|T\|_p \leq t\right) \right| \leq \inf_{\eta > 0} \left\{ \Gamma_p(\eta) + \Delta_p(\eta) \right\}. \tag{8}$$

The right-hand side of (8) can be controlled in various ways. For the case of $p = 2$, Götze et al. (2019) give $\Delta_2(\eta) \lesssim \eta \|\Sigma\|_{\mathrm{F}}^{-1/2}$. When $p = \infty$, note that ℓ^∞-balls are rectangles so $\mathcal{B}_\infty \subseteq \mathcal{R}$ and (5) applies so that $\Delta_\infty(\eta) \leq \eta(\sqrt{2 \log d} + 2)/\sigma_{\min}$ whenever $\min_j \Sigma_{jj} \geq \sigma_{\min}^2$. More generally $\Delta_p(\eta)$ can be bounded using anti-concentration of the ℓ^p-norm of a Gaussian random vector whenever such results are available. We note that alongside the ℓ^p-norms, other functionals can be analyzed in this manner, including the maximum statistic and other order statistics (Kozbur, 2021).

To conduct inference in this situation, we need to feasibly approximate the quantiles of $\|S\|_p$. To that end, take a significance level $\tau \in (0, 1)$ and define

$$\widehat{q}_p(\tau) = \inf\left\{t \in \mathbb{R} : \mathbb{P}(\|\widehat{T}\|_p \leq t \mid \mathbf{X}) \geq \tau\right\} \quad \text{where} \quad \widehat{T} \mid \mathbf{X} \sim \mathcal{N}(0, \widehat{\Sigma}),$$

with $\widehat{\Sigma}$ any $\mathbf{X}$-measurable positive semi-definite estimator of Σ. Note that for the canonical estimator $\widehat{\Sigma} = \sum_{i=1}^n X_i X_i^\mathsf{T}$ we can write $\widehat{T} = \sum_{i=1}^n X_i Z_i$ 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 $\|\widehat{T}\|_p \mid \mathbf{X}$ has no atoms, we can apply Proposition 3 to see

$$\begin{aligned}
\sup_{\tau \in (0,1)} \left| \mathbb{P}\left(\|S\|_p \leq \widehat{q}_p(\tau)\right) - \tau \right| &\leq \mathbb{E}\left[\sup_{t \geq 0} \left| \mathbb{P}(\|S\|_p \leq t) - \mathbb{P}(\|\widehat{T}\|_p \leq t \mid \mathbf{X}) \right| \right] \\
&\leq \inf_{\eta > 0} \left\{ \Gamma_p(\eta) + 2\Delta_p(\eta) + 2d\, \mathbb{E}\left[\exp\left(\frac{-\eta^2}{2d^{2/p} \left\| \widehat{\Sigma}^{1/2} - \Sigma^{1/2} \right\|_2^2} \right) \right] \right\}
\end{aligned}$$

and hence the bootstrap is valid whenever $\|\widehat{\Sigma}^{1/2} - \Sigma^{1/2}\|_2^2$ is sufficiently small. See the discussion at the end of Section 3.1 regarding methods for bounding this object.

Belloni and Oliveira (2018) obtained a central limit theorem for the maximum of a multivariate martingale using a coupling due to Chernozhukov et al. (2014b). Assuming the martingale differences X_i are bounded and the minimum eigenvalue of $\mathrm{Var}[T]/n$ is bounded away from zero, using Markov's inequality to bound $V_n - V$ (in their notation) and by Gaussian anti-concentration, their results establish that for all $\eta > 0$,

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\max_{1 \leq j \leq d} S_j \leq t \right) - \mathbb{P}\left(\max_{1 \leq j \leq d} T_j \leq t \right) \right| \lesssim \frac{\sqrt{\mathbb{E}[\|\Omega\|_2] \log 2d}}{\eta} + \frac{n (\log 2d)^{7/2}}{\eta^3} + \eta \sqrt{\frac{\log 2d}{n}},$$

requiring that $\mathbb{E}[\|\Omega\|_2]/n \to 0$ up to logs whenever d is at most polynomial in n. On the other hand, following the approach of our Proposition 6 gives

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\max_{1 \leq j \leq d} S_j \leq t \right) - \mathbb{P}\left(\max_{1 \leq j \leq d} T_j \leq t \right) \right| \lesssim \frac{(nd)^{1/3} \sqrt{\log 2d}}{\eta} + \frac{(\mathbb{E}[\|\Omega\|_2] \log 2d)^{1/3}}{\eta^{2/3}} + \eta \sqrt{\frac{\log 2d}{n}},$$

which requires both $\mathbb{E}[\|\Omega\|_2]/n \to 0$ and $d^2/n \to 0$ up to logs. While our assumptions are more restrictive than those of Belloni and Oliveira (2018), our approach is valid not only for the maximum statistic but also more generally for the ℓ^p-norm where $1 \leq p \leq \infty$.

4.3 Local polynomial estimators

As a third and final example we consider nonparametric regression estimation with martingale data again but now employing classical local polynomial methods (Fan and Gijbels, 1996). In contrast with the partitioning-based series methods of Section 4.1, local polynomial methods induce stochastic processes which are not linearly separable, allowing us to show the applicability of Proposition 4.

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. Using multi-index notation, let $p(w)$ be the k-dimensional vector collecting the monomials $w^\nu/\nu!$ for $0 \leq |\nu| \leq \gamma$, where $w^\nu = w_1^{\nu_1} \cdots w_m^{\nu_m}$ and $\nu! = \nu_1 \cdots \nu_m$ and $|\nu| = \nu_1 + \cdots + \nu_m$. Set $p_h(w) = p(w/h)$. The local polynomial regression estimator of $\mu(w)$ is

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

with $e_1 \in \mathbb{R}^k$ being the first standard unit vector.

The goal is again to approximate the distribution of the non-Donsker process $(\widehat{\mu}(w) - \mu(w) : w \in \mathcal{W})$, which can be decomposed as follows:

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

where $H(w) = \sum_{i=1}^n \mathbb{E}[K_h(W_i - w) p_h(W_i - w) p_h(W_i - w)^\mathsf{T}]$, $\widehat{H}(w) = \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) p_h(W_i - w)^\mathsf{T}$, $S(w) = \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) \varepsilon_i$ and $\mathrm{Bias}(w) = e_1^\mathsf{T} \widehat{H}(w)^{-1} \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) \mu(W_i) - \mu(w)$. A key distinctive feature of local polynomial regression is that both the $\widehat{H}(w)$ and $S(w)$ are functions of the evaluation point $w \in \mathcal{W}$; contrast this with the partitioning-based series estimator discussed in Section 4.1 for which neither $\widehat{H}$ nor S are functions of w. Therefore we need to use Proposition 4 to obtain a Gaussian strong approximation.

Under some mild regularity conditions, including stationarity for simplicity and an α-mixing assumption on the time-dependence of the data, we first show $\sup_{w \in \mathcal{W}} \|\widehat{H}(w) - H(w)\|_2 \lesssim_\mathbb{P} \sqrt{n h^{-2m} \log n}$. Further, $\sup_{w \in \mathcal{W}} |\mathrm{Bias}(w)| \lesssim_\mathbb{P} h^\gamma$ provided that the regression function is sufficiently smooth. Thus it remains to analyze the martingale empirical process $\big(e_1^\mathsf{T} H(w)^{-1} S(w) : w \in \mathcal{W}\big)$ via Proposition 4 by setting

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

With this approach, we obtain the following result.

Proposition 7 (Strong approximation for local polynomial estimators)
Under the nonparametric regression setup described above, assume further that

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

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

(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 some* $C_\varepsilon > 0$ *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 function satisfying* $\int K(w) \, \mathrm{d}w = 1$.

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

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

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

For completeness, the proof of Proposition 7 verifies that if the regression function $\mu(w)$ is γ times continuously differentiable on $\mathcal{W}$ then $\sup_w |\mathrm{Bias}(w)| \lesssim_{\mathbb{P}} h^\gamma$. 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 5 for partitioning-based series estimators, and as far as we know has not been improved upon for martingale data. With i.i.d. data in the multidimensional setting, a coupling due to Rio (1994) can offer improvements, while with i.i.d. data in the one-dimensional setting ($m = 1$) the Komlós–Major–Tusnády coupling (Komlós et al., 1975) may attain unimprovable strong approximation rates.

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 $\sup_{w \in \mathcal{W}} |\widehat{\mu}(w) - \mu(w)|$ to $\sup_{w \in \mathcal{W}} |T(w)|$. With an appropriate covariance estimator, we can further replace $T(w)$ by a feasible version $\widehat{T}(w)$ or its Studentized counterpart, enabling procedures for uniform inference analogous to the confidence bands constructed in Section 4.1. We omit the details to conserve space. Chernozhukov et al. (2014b, Remark 3.1) achieve better rates for i.i.d. data in Kolmogorov–Smirnov distance by bypassing the step where we first approximate the entire stochastic process. Jacod et al. (2021) take a similar approach to us in first providing a strong approximation for the t-statistic process and the deducing a coupling for the supremum t-statistic, in the context of volatility estimation for semimartingales in financial asset pricing.

We finally remark that in this setting of kernel-based local empirical processes it is essential that our initial strong approximation result (Theorem 1) does not impose a lower bound on the eigenvalues of the variance matrix Σ. This is because the proof of Proposition 7 applies the Gaussian coupling on a δ-cover $\mathcal{W}_\delta$ of $(\mathcal{W}, \|\cdot\|_2)$ where $\delta/h \to 0$, and it is not clear how to control the size of the smallest eigenvalue of the discretized covariance matrix $\mathrm{Cov}[\widehat{\mu}(w), \widehat{\mu}(w')]_{w,w' \in \mathcal{W}_\delta}$. As such, the result of Li and Liao (2020) is unsuited for this application due to its minimum eigenvalue assumption.

5 Conclusion

We introduced a new version of Yurinskii's coupling which strictly generalizes all previous forms of the result, giving a Gaussian strong approximation for martingale data in ℓ^p-norm where $1 \leq p \leq \infty$. We demonstrated the applicability of our main result to some areas of statistical theory, including high-dimensional martingale central limit theorems and uniform strong approximations for martingale empirical processes. We also gave illustrative examples in statistical methodology, applying our results to partitioning-based series estimators, distributional approximation of ℓ^p-norms of high-dimensional martingales, and local polynomial estimators. At each stage we addressed issues of feasibility and provided implementable inference procedures.

Acknowledgments

We thank Jianqing Fan and Alexander Giessing for comments. The authors gratefully acknowledge financial support from the National Science Foundation through grant DMS-2210561, and Cattaneo gratefully acknowledges financial support from the National Institute of Health (R01 GM072611-16).

A Proofs

A.1 Main results

The proof of Theorem 1 depends on several auxiliary results. First, we require Strassen's theorem for the ℓ^p-norm (Pollard, 2002, Theorem 8), stated for completeness as Lemma 1. Next, we present an analytic result about smooth approximation of indicator functions (Belloni et al., 2019, Lemma 39), given as Lemma 2. We then establish Lemma 3, a Yurinskii-type coupling result for martingales with non-random terminal quadratic variation. 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 Pollard (2002). The proof of Theorem 1 relies on constructing a "modified" martingale, which is close to the original martingale, but which has a non-random terminal quadratic variation. Lemma 3 is then applied to this modified martingale.

Lemma 1 (Strassen's theorem for the p-norm)
Let $\mathbb{P}_X$ and $\mathbb{P}_Y$ be Borel probability distributions on $\mathbb{R}^d$ and take $\eta, \rho > 0$ and $p \in [1, \infty]$. There are $X \sim \mathbb{P}_X$ and $Y \sim \mathbb{P}_Y$ with

$$\mathbb{P}\big(\|X - Y\|_p > \eta\big) \leq \rho \iff \mathbb{P}_X(A) \leq \mathbb{P}_Y(A^\eta) + \rho \text{ for all Borel sets } A.$$

Proof (Lemma 1)
By Theorem 8 in Pollard (2002), noting that every law on $\mathbb{R}^d$ is tight. □

Lemma 2 (Smooth approximation of 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

$$g_{A,\eta}(x) = \big(1 - \eta^{-1}\|x - A^\eta\|_p\big) \vee 0, \qquad f_{A,\sigma,\eta}(x) = \mathbb{E}\left[g_{A,\eta}(x + \sigma Z)\right].$$

Then for all $x, y \in \mathbb{R}^d$,

$$\left| f_{A,\sigma,\eta}(x+y) - f_{A,\sigma,\eta}(x) - y^\mathsf{T} \nabla f_{A,\sigma,\eta}(x) - \frac{1}{2} y^\mathsf{T} \nabla^2 f_{A,\sigma,\eta}(x) y \right| \leq \frac{\|y\|_2^2 \|y\|_p}{\sigma^2 \eta}, \tag{9}$$

$$(1 - \varepsilon)\mathbb{I}\big\{x \in A\big\} \leq f_{A,\sigma,\eta}(x) \leq \varepsilon + (1 - \varepsilon)\mathbb{I}\big\{x \in A^{3\eta}\big\} \tag{10}$$

where $\varepsilon = \mathbb{P}(\|Z\|_p > \eta/\sigma)$.

Proof (Lemma 2)
See Lemma 39 in Belloni et al. (2019). □

Lemma 3 (Strong approximation for martingale vectors with non-random quadratic variation)
Let $X_1, \ldots, X_n$ be $\mathbb{R}^d$-valued random variables adapted to a filtration $\mathcal{H}_1, \ldots, \mathcal{H}_n$, with $\mathcal{H}_0$ the trivial σ-algebra. Suppose that $\mathbb{E}[X_i \mid \mathcal{H}_{i-1}] = 0$ for each $1 \leq i \leq n$. Let $V_i = \operatorname{Var}[X_i \mid \mathcal{H}_{i-1}]$ and assume it exists and is finite. Suppose that $\sum_{i=1}^n V_i = \Sigma$ almost surely where $\Sigma \in \mathbb{R}^{d \times d}$ is non-random. Then for each $\delta > 0$ and $p \in [1, \infty]$ there exists $T \sim \mathcal{N}(0, \Sigma)$ such that

$$\mathbb{P}\big(\|S - T\|_p > 3\delta\big) \leq \inf_{t > 0} \left\{ 2\alpha_p(t) + \frac{\beta_p}{\delta^3} t^2 \right\}$$

where

$$\alpha_p(t) = \mathbb{P}\big(\|Z\|_p > t\big), \qquad\qquad \beta_p = \sum_{i=1}^n \mathbb{E}\left[\|X_i\|_2^2\|X_i\|_p + \|V_i^{1/2}Z_i\|_2^2\|V_i^{1/2}Z_i\|_p\right]$$

with $Z, Z_1, \ldots, Z_n$ i.i.d. standard Gaussian on $\mathbb{R}^d$ independent of $\mathcal{H}_n$.

Proof (Lemma 3)
Let $Z_1, \ldots, Z_n$ be i.i.d. $\mathcal{N}(0, I_d)$ and independent of $X_1, \ldots, X_n$. Define $\widetilde{X}_i = V_i^{1/2}Z_i$ and $\widetilde{S} = \sum_{i=1}^n \widetilde{X}_i$. Fix a Borel set $A \subseteq \mathbb{R}^d$ and $\sigma, \eta > 0$ and let $f = f_{A,\sigma,\eta}$ be the function defined in Lemma 2. By the Lindeberg method, write the telescoping sum

$$\mathbb{E}\big[f(S) - f(\widetilde{S})\big] = \sum_{i=1}^n \mathbb{E}\big[f(Y_i + X_i) - f(Y_i + \widetilde{X}_i)\big]$$

where $Y_i = X_1 + \cdots + X_{i-1} + \widetilde{X}_{i+1} + \cdots + \widetilde{X}_n$. Take second-order Taylor expansions to obtain

$$\begin{aligned}
\mathbb{E}\big[f(Y_i + X_i)\big] &= \mathbb{E}\big[f(Y_i)\big] + \mathbb{E}\big[X_i^\mathsf{T}\nabla f(Y_i)\big] + \frac{1}{2}\mathbb{E}\big[X_i^\mathsf{T}\nabla^2 f(Y_i)X_i\big] + R_1^i, \\
\mathbb{E}\big[f(Y_i + \widetilde{X}_i)\big] &= \mathbb{E}\big[f(Y_i)\big] + \mathbb{E}\big[\widetilde{X}_i^\mathsf{T}\nabla f(Y_i)\big] + \frac{1}{2}\mathbb{E}\big[\widetilde{X}_i^\mathsf{T}\nabla^2 f(Y_i)\widetilde{X}_i\big] + R_2^i
\end{aligned}$$

where R_1^i and R_2^i denote the Taylor approximation errors. Subtracting the expressions yields

$$\begin{aligned}
&\mathbb{E}\big[f(Y_i + X_i) - f(Y_i + \widetilde{X}_i)\big] \\
&\quad = \mathbb{E}\big[(X_i - \widetilde{X}_i)^\mathsf{T}\nabla f(Y_i)\big] + \frac{1}{2}\mathbb{E}\big[X_i^\mathsf{T}\nabla^2 f(Y_i)X_i - \widetilde{X}_i^\mathsf{T}\nabla^2 f(Y_i)\widetilde{X}_i\big] + R_1^i - R_2^i
\end{aligned} \tag{11}$$

and we now bound each term on the right-hand side. Define

$$\begin{aligned}
\widetilde{Y}_i &= X_1 + \cdots + X_{i-1} + (V_{i+1} + \cdots + V_n)^{1/2} Z_i \\
&= X_1 + \cdots + X_{i-1} + (\Sigma - (V_1 + \cdots + V_i))^{1/2} Z_i
\end{aligned}$$

and let $\mathcal{G}_i$ be the σ-algebra generated by $\mathcal{H}_{i-1}$ and Z_i. Note that $\widetilde{Y}_i$ is $\mathcal{G}_i$-measurable and that Y_i and $\widetilde{Y}_i$ have the same distribution conditional on $\mathcal{H}_n$. Also Z_i is zero-mean and independent of X_i, Y_i and V_i and X_i is a martingale difference sequence with respect to $\mathcal{G}_i$. So for the first term in (11),

$$\mathbb{E}\big[(X_i - \widetilde{X}_i)^\mathsf{T}\nabla f(Y_i)\big] = \mathbb{E}\big[X_i^\mathsf{T}\nabla f(\widetilde{Y}_i)\big] - \mathbb{E}[Z_i]^\mathsf{T}\mathbb{E}\big[V_i^{1/2}\nabla f(Y_i)\big] = \mathbb{E}\big[\mathbb{E}[X_i \mid \mathcal{G}_i]^\mathsf{T}\nabla f(\widetilde{Y}_i)\big] - 0 = 0.$$

For the second term in (11) by conditioning on $\mathcal{H}_{i-1}$ and Y_i we have

$$\mathbb{E}\big[\widetilde{X}_i^\mathsf{T}\nabla^2 f(Y_i)\widetilde{X}_i\big] = \mathbb{E}\big[\operatorname{Tr}\nabla^2 f(Y_i)\widetilde{X}_i\widetilde{X}_i^\mathsf{T}\big] = \mathbb{E}\big[\operatorname{Tr}\nabla^2 f(Y_i)\mathbb{E}[\widetilde{X}_i\widetilde{X}_i^\mathsf{T} \mid \mathcal{H}_{i-1}]\big] = \mathbb{E}\big[\operatorname{Tr}\nabla^2 f(Y_i)V_i\big]$$

and, by the same properties used above,

$$\mathbb{E}\big[X_i^\mathsf{T}\nabla^2 f(Y_i)X_i\big] = \mathbb{E}\big[\operatorname{Tr}\nabla^2 f(\widetilde{Y}_i)\mathbb{E}[X_iX_i^\mathsf{T} \mid \mathcal{G}_i]\big] = \mathbb{E}\big[\operatorname{Tr}\nabla^2 f(Y_i)V_i)\big].$$

Hence the second term also vanishes. For the third and fourth terms in (11), we use (9) to obtain

$$|R_1^i| + |R_2^i| \leq \frac{1}{\sigma^2\eta}\Big(\mathbb{E}\big[\|X_i\|_2^2\|X_i\|_p\big] + \mathbb{E}\big[\|\widetilde{X}_i\|_2^2\|\widetilde{X}_i\|_p\big]\Big).$$

Combining the three terms and summing over i yields

$$\mathbb{E}\big[f(S) - f(\widetilde{S})\big] \leq \frac{1}{\sigma^2 \eta} \sum_{i=1}^n \Big(\mathbb{E}\big[\|X_i\|_2^2 \|X_i\|_p\big] + \mathbb{E}\big[\|\widetilde{X}_i\|_2^2 \|\widetilde{X}_i\|_p\big]\Big) = \frac{\beta_p}{\sigma^2 \eta}.$$

Along with (10) we conclude that

$$\begin{aligned}
\mathbb{P}(S \in A) &= \mathbb{E}\big[\mathbb{I}\{S \in A\} - f(S)\big] + \mathbb{E}\big[f(S) - f(\widetilde{S})\big] + \mathbb{E}\big[f(\widetilde{S})\big] \\
&\leq \varepsilon \mathbb{P}(S \in A) + \frac{\beta_p}{\sigma^2 \eta} + \varepsilon + (1-\varepsilon)\mathbb{P}\big(\widetilde{S} \in A^{3\eta}\big) \leq \mathbb{P}\big(\widetilde{S} \in A^{3\eta}\big) + 2\varepsilon + \frac{\beta_p}{\sigma^2 \eta}.
\end{aligned}$$

Set $\sigma = \eta / t$ for $t > 0$ to obtain

$$\mathbb{P}(S \in A) \leq \mathbb{P}\big(\widetilde{S} \in A^{3\eta}\big) + 2\mathbb{P}\big(\|Z\|_p > t\big) + \frac{\beta_p}{\eta^3} t^2.$$

Finally, since $\widetilde{S} = \sum_{i=1}^n V_i^{1/2} Z_i \sim \mathcal{N}(0, \Sigma)$ by the non-random quadratic variation property, Strassen's theorem (Lemma 1) ensures the existence of S and $T \sim \mathcal{N}(0, \Sigma)$ such that

$$\mathbb{P}(\|S - T\|_p > 3\eta) \leq \inf_{t > 0} \left\{ 2\alpha_p(t) + \frac{\beta_p}{\eta^3} t^2 \right\}.$$

□

Proof (Theorem 1)
Part 1: constructing the 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 non-random and equals $\Sigma + M$. Take $m \geq 1$ and define

$$\begin{aligned}
H_k &= \sum_{i=1}^n \mathbb{E}[V_i] - \sum_{i=1}^k V_i + M, & \tau &= \sup\big\{k \in \{0, 1, \ldots, n\} : H_k \succeq 0\big\}, \\
\widetilde{X}_i &= X_i \mathbb{I}\{i \leq \tau\} + \frac{1}{\sqrt{m}} H_\tau^{1/2} Z_i \mathbb{I}\{n+1 \leq i \leq n+m\}, & \widetilde{S} &= \sum_{i=1}^{n+m} \widetilde{X}_i,
\end{aligned}$$

where $Z_{n+1}, \ldots Z_{n+m}$ is an i.i.d. sequence of standard Gaussian vectors in $\mathbb{R}^d$ independent of $\mathcal{H}_n$. Also define the filtration $(\mathcal{G}_i : 1 \leq i \leq n+m)$, where $\mathcal{G}_i = \mathcal{H}_i$ for $0 \leq i \leq n$ and $\mathcal{G}_i$ is the σ-algebra generated by $\mathcal{F}_n$ and $Z_{n+1}, \ldots, Z_i$ for $n+1 \leq i \leq n+m$.

We now show that $\widetilde{S}$ satisfies the conditions of Lemma 3 with terminal quadratic variation $\Sigma + M$. Firstly, τ is a stopping time with respect to $\mathcal{F}_i$. To see this note that $H_{i+1} - H_i = -V_{i+1} \preceq 0$ almost surely so that $\{\tau \leq i\} = \{H_{i+1} \not\succeq 0\}$ for $0 \leq i < n$. This depends only on $V_1, \ldots, V_{i+1}$ which are $\mathcal{H}_i$-measurable. Similarly, $\{\tau = n\} = \{H_n \succeq 0\} \in \mathcal{H}_{n-1}$.

Next, H_τ is $\mathcal{H}_n$-measurable so $\widetilde{X}_i$ is $\mathcal{G}_i$-measurable for every $1 \leq i \leq n+m$. Further, by the martingale difference property of X_i, we have $\mathbb{E}\big[\widetilde{X}_i \mid \mathcal{G}_{i-1}\big] = \mathbb{I}\{i \leq \tau\}\mathbb{E}[X_i \mid \mathcal{H}_{i-1}] = 0$ for $1 \leq i \leq n$ and $\mathbb{E}\big[\widetilde{X}_i \mid \mathcal{G}_{i-1}\big] = m^{-1/2} H_\tau^{1/2} \mathbb{E}[Z_i \mid \mathcal{G}_{i-1}] = 0$ for $n+1 \leq i \leq n+m$. So $\widetilde{X}_i$ form martingale differences with respect to $\mathcal{G}_i$.

Finally, let $\widetilde{V}_i = \mathbb{E}\big[\widetilde{X}_i \widetilde{X}_i^\mathsf{T} \mid \mathcal{G}_{i-1}\big]$ so that $\widetilde{V}_i = V_i \mathbb{I}\{i \leq \tau\}$ for $1 \leq i \leq n$ and $\widetilde{V}_i = H_\tau / m$ for $n+1 \leq i \leq n+m$. Thus the terminal quadratic variation is $\sum_{i=1}^{n+m} \widetilde{V}_i = \sum_{i=1}^\tau V_i + H_\tau = \sum_{i=1}^n \mathbb{E}[V_i] + M = \Sigma + M$ by definition of H_τ.

Part 2: Gaussian coupling for the modified martingale
By Lemma 3 with $p \in [1, \infty]$ and $\eta > 0$, there is a random vector $\widetilde{T} \sim \mathcal{N}(0, \Sigma + M)$ (on some probability space) such that

$$\mathbb{P}\big(\|\widetilde{S} - \widetilde{T}\|_p > 3\eta\big) \le \inf_{t<0} \left\{ 2\alpha_p(t) + \frac{\widetilde{\beta}_p}{\eta^3} t^2 \right\},$$

with $\alpha_p(t)$ as defined in Lemma 3 and

$$\widetilde{\beta}_p = \sum_{i=1}^{n} \mathbb{E}\left[\|\widetilde{X}_i\|_2^2 \|\widetilde{X}_i\|_p + \|\widetilde{V}_i^{1/2} Z_i\|_2^2 \|\widetilde{V}_i^{1/2} Z_i\|_p \right] + \frac{1}{m^{3/2}} \sum_{i=n+1}^{n+m} \mathbb{E}\left[\|H_\tau^{1/2} Z_i\|_2^2 \|H_\tau^{1/2} Z_i\|_p \right].$$

Clearly, for the first term ($1 \le i \le n$) we have $\|\widetilde{X}_i\|_p \le \|X_i\|_p$ and $\|\widetilde{V}^{1/2} Z_i\|_p \le \|V_i^{1/2} Z_i\|_p$, while the second term is equal to $m^{-1/2}\mathbb{E}\big[\|H_\tau^{1/2} Z\|_2^2 \|H_\tau^{1/2} Z\|_p\big]$ where Z is an independent standard Gaussian variable. Since H_i is weakly decreasing under the semi-definite partial order, we have $H_\tau \preceq H_0$ implying that $\max_j |(H_\tau)_{jj}| \le \|H_0\|_{\max}$. Write $\breve{Z} = H_\tau^{1/2} Z$ so $\max_{1 \le j \le d} \mathbb{E}\big[\breve{Z}_j^2 \mid \mathcal{H}_n\big] \le \|H_0\|_{\max}$ and $\mathbb{E}\big[|\breve{Z}_j|^3\big] \le \sqrt{8/\pi}\, \|H_0\|_{\max}^{3/2}$. Hence as $p \ge 1$,

$$\mathbb{E}\left[\|H_\tau^{1/2} Z\|_2^2 \|H_\tau^{1/2} Z\|_p \right] \le \mathbb{E}\left[\|\breve{Z}\|_1^3 \right] \le d^3 \max_{1 \le j \le d} \mathbb{E}\left[|\breve{Z}_j|^3 \right] \le d^3 \sqrt{8/\pi}\, \|H_0\|_{\max}^{3/2}.$$

We may assume that $\beta_p = \sum_{i=1}^n \mathbb{E}[\|X_i\|_2^2 \|X_i\|_p + \|V_i^{1/2} Z_i\|_2^2 \|V_i^{1/2} Z_i\|_p]$ is finite as otherwise the theorem is vacuous. Take $m \ge 8d^6 \|H_0\|_{\max}^3 / (\pi \beta_p^2)$ so that $\widetilde{\beta}_p \le 2\beta_p$ and

$$\mathbb{P}\big(\|\widetilde{S} - \widetilde{T}\|_p > 3\eta\big) \le 2 \inf_{t>0} \left\{ \alpha_p(t) + \frac{\beta_p}{\eta^3} t^2 \right\}. \tag{12}$$

Part 3: bounding the difference between the original and modified martingales
By the triangle inequality,

$$\|S - \widetilde{S}\|_p \le \left\| \sum_{i=\tau+1}^{n} X_i \right\|_p + \left\| \frac{1}{\sqrt{m}} \sum_{i=n+1}^{m} H_\tau^{1/2} Z_i \right\|_p.$$

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 $\mathbb{P}\big(\|H_\tau^{1/2} Z\|_p > \eta\big) \le \mathbb{P}\big(\|H_n^{1/2} Z\|_p > \eta,\, \Omega \preceq M\big) + \mathbb{P}\big(\Omega \npreceq M\big)$, where we used $\{\Omega \preceq M\} = \{\tau = n\}$. Therefore

$$\mathbb{P}\big(\|S - \widetilde{S}\|_p > 2\eta\big) \le 2\mathbb{P}\big(\Omega \npreceq M\big) + \mathbb{P}\big(\|(M - \Omega)^{1/2} Z\|_p > \eta,\, \Omega \preceq M\big). \tag{13}$$

Part 4: conclusion
We will show how to write $\widetilde{T} = (\Sigma + M)^{1/2} W$ where $W \sim \mathcal{N}(0, I_d)$ and use this representation to construct T. By the spectral theorem, let $\Sigma + M = U \Lambda U^\mathsf{T}$ where U is a $d \times d$ orthogonal matrix and Λ is a diagonal $d \times d$ matrix with diagonal entries satisfying $\lambda_1 \ge \cdots \ge \lambda_r > 0$ 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 simply inverting its non-zero elements) and define $W = U(\Lambda^+)^{1/2} U^\mathsf{T} \widetilde{T} + U\widetilde{W}$, where the first r elements of $\widetilde{W}$ are zero and the last $d - r$ elements are i.i.d. $\mathcal{N}(0, 1)$ independent

from $\widetilde{T}$. Then it is easy to check that $W \sim \mathcal{N}(0, I_d)$ and that $\widetilde{T} = (\Sigma + M)^{1/2} W$. Now define $T = \Sigma^{1/2} W \sim \mathcal{N}(0, \Sigma)$ so that

$$\mathbb{P}\big(\|T - \widetilde{T}\|_p > \eta\big) = \mathbb{P}\Big(\big\|\big((\Sigma + M)^{1/2} - \Sigma^{1/2}\big) W\big\|_p > \eta\Big). \tag{14}$$

Finally (12), (13), (14), the triangle inequality and a union bound conclude the proof since by taking an infimum over $M \succeq 0$,

$$\begin{aligned}\mathbb{P}\big(\|S - T\|_p > 6\eta\big) &\le \mathbb{P}\big(\|\widetilde{S} - \widetilde{T}\|_p > 3\eta\big) + \mathbb{P}\big(\|S - \widetilde{S}\|_p > 2\eta\big) + \mathbb{P}\big(\|T - \widetilde{T}\|_p > \eta\big) \\ &\le 2 \inf_{t > 0} \left\{ \alpha_p(t) + \frac{\beta_p}{\eta^3} t^2 \right\} + \inf_{M \succeq 0} \big\{ 2\gamma(M) + \delta(M, \eta) + \varepsilon(M, \eta) \big\}.\end{aligned}$$

□

Before proving Proposition 1 we provide an ℓ^p-norm bound for Gaussian variables in Lemma 4.

Lemma 4 (Gaussian p-norm bound)
Let $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{d \times d}$ is positive semi-definite. Then

$$\mathbb{E}\left[\|X\|_p\right] \le \phi_p(d) \max_{1 \le j \le d} \sqrt{\Sigma_{jj}}$$

where $\phi_p(d) = \sqrt{p d^{2/p}}$ for $p \in [1, \infty)$ and $\phi_\infty(d) = \sqrt{2 \log 2d}$.

Proof (Lemma 4)
For $p \in [1, \infty)$, since each X_j is Gaussian, we have $\big(\mathbb{E}\big[|X_j|^p\big]\big)^{1/p} \le \sqrt{p\, \mathbb{E}[X_j^2]} = \sqrt{p \Sigma_{jj}}$. Therefore

$$\mathbb{E}\big[\|X\|_p\big] \le \left(\sum_{j=1}^d \mathbb{E}\big[|X_j|^p\big] \right)^{1/p} \le \left(\sum_{j=1}^d p^{p/2} \Sigma_{jj}^{p/2} \right)^{1/p} \le \sqrt{p d^{2/p}} \max_{1 \le j \le d} \sqrt{\Sigma_{jj}}.$$

For $p = \infty$, with $\sigma^2 = \max_j \Sigma_{jj}$, for $t > 0$,

$$\mathbb{E}\big[\|X\|_\infty\big] \le t \log \sum_{j=1}^d \mathbb{E}\Big[e^{|X_j|/t}\Big] \le t \log \sum_{j=1}^d \mathbb{E}\Big[2 e^{X_j/t}\Big] \le t \log(2 d e^{\sigma^2/(2t^2)}) \le t \log 2d + \frac{\sigma^2}{2t}.$$

Setting $t = \frac{\sigma}{\sqrt{2 \log 2d}}$ gives $\mathbb{E}\big[\|X\|_\infty\big] \le \sigma \sqrt{2 \log 2d}$. □

Proof (Proposition 1)
We set $M = \nu^2 I_d$ and bound each term appearing on the right-hand side of (1).

Part 1: bounding $\alpha_p(t)$
By Markov's inequality and Lemma 4, we have $\alpha_p(t) = \mathbb{P}\big(\|Z\|_p > t\big) \le \mathbb{E}[\|Z\|_p]/t \le \phi_p(d)/t$.

Part 2: bounding $\gamma(M)$
With $M = \nu^2 I_d$ and by Markov's inequality, $\gamma(M) = \mathbb{P}\big(\Omega \npreceq M\big) = \mathbb{P}\big(\|\Omega\|_2 > \nu^2\big) \le \nu^{-2} \mathbb{E}[\|\Omega\|_2]$.

Part 3: bounding $\delta(M, \eta)$
By Markov's inequality and Lemma 4, using $\max_j |M_{jj}| \le \|M\|_2$ for $M \succeq 0$,

$$\delta_p(M, \eta) = \mathbb{P}\left(\big\| \big((\Sigma + M)^{1/2} - \Sigma^{1/2}\big) Z \big\|_p \ge \eta \right) \le \frac{\phi_p(d)}{\eta} \big\|(\Sigma + M)^{1/2} - \Sigma^{1/2}\big\|_2.$$

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 $\varepsilon(M, \eta)$
Since $(M - \Omega)^{1/2} Z$ is a centered Gaussian conditional on $\mathcal{H}_n$ and $\{\Omega \preceq M\}$, We have by Markov's inequality, Lemma 4 and Jensen's inequality that

$$\begin{aligned}\varepsilon_p(M, \eta) &= \mathbb{P}\left(\left\|(M - \Omega)^{1/2} Z\right\|_p \geq \eta,\ \Omega \preceq M\right) \leq \frac{1}{\eta}\mathbb{E}\left[\mathbb{I}\{\Omega \preceq M\}\mathbb{E}\left[\left\|(M - \Omega)^{1/2} Z\right\|_p \mid \mathcal{H}_n\right]\right] \\ &\leq \frac{\phi_p(d)}{\eta}\mathbb{E}\left[\mathbb{I}\{\Omega \preceq M\} \max_{1 \leq j \leq d} \sqrt{(M - \Omega)_{jj}}\right] \leq \frac{\phi_p(d)}{\eta}\mathbb{E}\left[\sqrt{\|M - \Omega\|_2}\right] \\ &\leq \frac{\phi_p(d)}{\eta}\mathbb{E}\left[\sqrt{\|\Omega\|_2} + \nu\right] \leq \frac{\phi_p(d)}{\eta}\left(\sqrt{\mathbb{E}[\|\Omega\|_2]} + \nu\right).\end{aligned}$$

Part 5: conclusion
Thus by Theorem 1 and the previous parts,

$$\begin{aligned}\mathbb{P}\big(\|S - T\|_p > 6\eta\big) &\leq 2 \inf_{t > 0}\left\{\alpha_p(t) + \frac{\beta_p}{\eta^3}t^2\right\} + \inf_{M \succeq 0}\big\{2\gamma(M) + \delta(M, \eta) + \varepsilon(M, \eta)\big\} \\ &\leq 2 \inf_{t > 0}\left\{\frac{\phi_p(d)}{t} + \frac{\beta_p}{\eta^3}t^2\right\} + \inf_{\nu > 0}\left\{\frac{2\mathbb{E}[\|\Omega\|_2]}{\nu^2} + \frac{2\phi_p(d)\nu}{\eta}\right\} + \frac{\phi_p(d)\sqrt{\mathbb{E}[\|\Omega\|_2]}}{\eta}.\end{aligned}$$

Balancing the terms by setting $t = \phi_p(d)^{1/3}\beta_p^{-1/3}\eta$ and $\nu = \mathbb{E}[\|\Omega\|_2]^{1/3}\phi_p(d)^{-1/3}\eta^{1/3}$ yields

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

The result follows by noting that either the last term is greater than one, in which case the result is vacuous, or it is less than one and is dominated by the second term. Finally replace η by $\eta/6$. □

A.2 Applications to statistical theory

Proof (Proposition 2)
This follows from Strassen's theorem (Lemma 1), but we provide a proof for completeness. Note

$$\mathbb{P}(S \in A) \leq \mathbb{P}(T \in A) + \mathbb{P}(T \in A_p^\eta \setminus A) + \mathbb{P}(\|S - T\| > \eta)$$

and applying this to $\mathbb{R}^d \setminus A$ gives

$$\begin{aligned}\mathbb{P}(S \in A) &= 1 - \mathbb{P}(S \in \mathbb{R}^d \setminus A) \\ &\geq 1 - \mathbb{P}(T \in \mathbb{R}^d \setminus A) - \mathbb{P}(T \in (\mathbb{R}^d \setminus A)_p^\eta \setminus (\mathbb{R}^d \setminus A)) - \mathbb{P}(\|S - T\| > \eta) \\ &= \mathbb{P}(T \in A) - \mathbb{P}(T \in A \setminus A_p^{-\eta}) - \mathbb{P}(\|S - T\| > \eta).\end{aligned}$$

Since this holds for all $p \in [1, \infty]$,

$$\begin{aligned}\sup_{A \in \mathcal{A}}\big|\mathbb{P}(S \in A) - \mathbb{P}(T \in A)\big| &\leq \sup_{A \in \mathcal{A}}\big\{\mathbb{P}(T \in A_p^\eta \setminus A) \vee \mathbb{P}(T \in A \setminus A_p^{-\eta})\big\} + \mathbb{P}(\|S - T\| > \eta) \\ &\leq \inf_{p \in [1, \infty]} \inf_{\eta > 0}\big\{\Gamma_p(\eta) + \Delta_p(\mathcal{A}, \eta)\big\}.\end{aligned}$$

□

Before proving Proposition 3 we give a Gaussian–Gaussian ℓ^p-norm approximation as Lemma 5.

Lemma 5 (Gaussian–Gaussian approximation in 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

$$\mathbb{P}\left(\left\|\left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right) Z\right\|_p > t\right) \leq 2d \exp\left(\frac{-t^2}{2d^{2/p}\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|_2^2}\right).$$

Proof (Lemma 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

$$\begin{aligned}\mathbb{P}\left(\left\|\Sigma^{1/2} Z\right\|_\infty > t\right) &= \mathbb{P}\left(\max_{1\leq j\leq d} \left|\left(\Sigma^{1/2} Z\right)_j\right| > t\right) \leq \sum_{j=1}^d \mathbb{P}\left(\left|\left(\Sigma^{1/2} Z\right)_j\right| > t\right) \\ &= \sum_{j=1}^d \mathbb{P}\left(|\sigma_j Z_j| > t\right) \leq 2\sum_{j=1}^d \exp\left(\frac{-t^2}{2\sigma_j^2}\right) \leq 2d \exp\left(\frac{-t^2}{2\|\sigma\|_\infty^2}\right).\end{aligned}$$

Now we apply this with the matrix $\Sigma = \left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right)^2$. For $p \in [1,\infty)$,

$$\begin{aligned}\|\sigma\|_p^p = \sum_{j=1}^d (\Sigma_{jj})^{p/2} &= \sum_{j=1}^d \left(\left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right)^2\right)_{jj}^{p/2} \leq d \max_{1\leq j\leq d} \left(\left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right)^2\right)_{jj}^{p/2} \\ &\leq d \left\|\left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right)^2\right\|_2^{p/2} = d\left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_2^p\end{aligned}$$

Similarly for $p = \infty$ we have

$$\|\sigma\|_\infty = \max_{1\leq j\leq d} (\Sigma_{jj})^{1/2} = \max_{1\leq j\leq d} \left(\left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right)^2\right)_{jj}^{1/2} \leq \left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_2.$$

Thus for all $p \in [1,\infty]$ we have $\|\sigma\|_p \leq d^{1/p}\left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_2$, with $d^{1/\infty} = 1$ by convention. Hence

$$\mathbb{P}\left(\left\|\left(\Sigma_1^{1/2} - \Sigma_2^{1/2}\right) Z\right\|_p > t\right) \leq 2d\exp\left(\frac{-t^2}{2\|\sigma\|_p^2}\right) \leq 2d \exp\left(\frac{-t^2}{2d^{2/p}\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|_2^2}\right).$$

□

Proof (Proposition 3)
Since $T = \Sigma^{1/2} Z$ is independent of $\mathbf{X}$,

$$\begin{aligned}&\left|\mathbb{P}(S \in A) - \mathbb{P}\left(\widehat{\Sigma}^{1/2} Z \in A \mid \mathbf{X}\right)\right| \\ &\quad \leq \left|\mathbb{P}(S \in A) - \mathbb{P}(T \in A)\right| + \left|\mathbb{P}(\Sigma^{1/2} Z \in A) - \mathbb{P}\left(\widehat{\Sigma}^{1/2} Z \in A \mid \mathbf{X}\right)\right|.\end{aligned}$$

The first term is bounded by Proposition 2 and the second by applying Lemma 5 conditional on $\mathbf{X}$.

$$\left|\mathbb{P}\big(S \in A\big) - \mathbb{P}\left(\widehat{\Sigma}^{1/2} Z \in A \mid \mathbf{X}\right)\right|$$
$$\leq \Gamma_p(\eta) + \Delta_p(\mathcal{A}, \eta) + \Delta_{p'}(\mathcal{A}, \eta') + 2d \exp\left(\frac{-\eta'^2}{2d^{2/p'} \big\|\widehat{\Sigma}^{1/2} - \Sigma^{1/2}\big\|_2^2}\right)$$

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. Note that we need not insist that $p = p'$ and $\eta = \eta'$ in general. □

Proof (Proposition 4)
Let $\mathcal{F}_\delta$ be a δ-cover of $(\mathcal{F}, d)$. Using a union bound, we can write

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \big|S(f) - T(f)\big| \geq 2t + \eta\right) \leq \mathbb{P}\left(\sup_{f \in \mathcal{F}_\delta} \big|S(f) - T(f)\big| \geq \eta\right)$$
$$+ \mathbb{P}\left(\sup_{d(f,f') \leq \delta} \big|S(f) - S(f')\big| \geq t\right) + \mathbb{P}\left(\sup_{d(f,f') \leq \delta} \big|T(f) - T(f')\big| \geq t\right).$$

Part 1: bounding the error on $\mathcal{F}_\delta$
We apply Proposition 1 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 as 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 Gaussian vector $T(\mathcal{F}_\delta)$ with the same covariance structure as $S(\mathcal{F}_\delta)$ satisfying

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}_\delta} \big|S(f) - T(f)\big| \geq \eta\right) \leq \frac{24\beta_\delta^{1/3} \sqrt{2 \log 2|\mathcal{F}_\delta|}^{2/3}}{\eta} + 17\left(\frac{\sqrt{2 \log 2|\mathcal{F}_\delta|}\sqrt{\mathbb{E}\left[\|\Omega_\delta\|_2\right]}}{\eta}\right)^{2/3}.$$

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)

$$\left\|\sup_{d(f,f') \leq \delta} \big|S(f) - S(f')\big|\right\|_\psi \leq C_\psi L \left(\int_0^\delta \psi^{-1}(N_\varepsilon)\,\mathrm{d}\varepsilon + \delta \psi^{-1}(N_\delta)^2\right) = C_\psi L J_\psi(\delta).$$

Then by Markov's inequality and the definition of the Orlicz norm,

$$\mathbb{P}\left(\sup_{d(f,f') \leq \delta} \big|S(f) - S(f')\big| \geq t\right) \leq \psi\left(\frac{t}{C_\psi L J_\psi(\delta)}\right)^{-1}.$$

Part 3: bounding the fluctuations in $T(f)$
By the Vorob'ev–Berkes–Philipp theorem (Dudley, 1999), $T(\mathcal{F}_\delta)$ extends to a Gaussian process $T(f)$. Firstly since $\big\|\big|T(f) - T(f')\big|\big\|_2 \leq L d(f, f')$ and $T(f)$ is a Gaussian process, we have $\big\|T(f) - T(f')\big\|_{\psi_2} \leq 2L d(f, f')$ 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),

$$\left\|\sup_{d(f,f') \leq \delta} \big|T(f) - T(f')\big|\right\|_{\psi_2} \leq C_1 L \int_0^\delta \sqrt{\log N_\varepsilon}\,\mathrm{d}\varepsilon = C_1 L J_2(\delta)$$

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,

$$\mathbb{P}\left(\sup_{d(f,f') \leq \delta} \big|T(f) - T(f')\big| \geq t\right) \leq \left(\exp\left(\frac{t^2}{C_1^2 L^2 J_2(\delta)^2}\right) - 1\right)^{-1} \vee 1 \leq 2 \exp\left(\frac{-t^2}{C_1^2 L^2 J_2(\delta)^2}\right).$$

Part 4: conclusion

The result follows by combining the parts, scaling t and η and enlarging constants if necessary. $\square$

A.3 Illustrative examples

Before proving the results of Section 4, we provide a useful result (Lemma 6) which helps bound the β_p term from Theorem 1 in applications. We also include for completeness some variance bounds (Lemma 7) and exponential inequalities (Lemma 8) for α-mixing random variables.

Lemma 6 (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

$$\mathbb{E}\left[\|X\|_2^2 \|X\|_\infty\right] \leq 4\sigma\sqrt{\log 2d} \sum_{j=1}^d \sigma_j^2.$$

Proof (Lemma 6)
By the Cauchy–Schwarz inequality, $\mathbb{E}\left[\|X\|_2^2 \|X\|_\infty\right] \leq \mathbb{E}\big[\|X\|_2^4\big]^{1/2} \mathbb{E}\big[\|X\|_\infty^2\big]^{1/2}$. For the first term, by Cauchy–Schwarz and the fourth moment of a normal distribution,

$$\mathbb{E}\big[\|X\|_2^4\big] = \mathbb{E}\left[\bigg(\sum_{j=1}^d X_j^2\bigg)^2\right] = \sum_{j=1}^d \sum_{k=1}^d \mathbb{E}\big[X_j^2 X_k^2\big] \leq \bigg(\sum_{j=1}^d \sqrt{\mathbb{E}\big[X_j^4\big]}\bigg)^2 = 3\bigg(\sum_{j=1}^d \sigma_j^2\bigg)^2.$$

For the second term, by Jensen's inequality and the χ^2 moment generating function,

$$\mathbb{E}\big[\|X\|_\infty^2\big] = \mathbb{E}\left[\max_{1 \leq j \leq d} X_j^2\right] \leq 4\sigma^2 \log \sum_{j=1}^d \mathbb{E}\Big[e^{X_j^2/(4\sigma^2)}\Big] \leq 4\sigma^2 \log \sum_{j=1}^d \sqrt{2} \leq 4\sigma^2 \log 2d.$$

$\square$

Lemma 7 (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 $|X_i| \leq M_i$ a.s. then

$$\mathrm{Var}\left[\sum_{i=1}^n X_i\right] \leq 4 \sum_{j=1}^\infty \alpha(j) \sum_{i=1}^n M_i^2.$$

(ii) If $\alpha(j) \leq e^{-2j/C_\alpha}$ then for any $r > 2$ there is a constant C_r depending only on r such that

$$\mathrm{Var}\left[\sum_{i=1}^n X_i\right] \leq C_r C_\alpha \sum_{i=1}^n \mathbb{E}\big[|X_i|^r\big]^{2/r}.$$

Proof (Lemma 7)
Define $\alpha^{-1}(t) = \inf\{j \in \mathbb{N} : \alpha(j) \leq t\}$ and $Q_i(t) = \inf\{s \in \mathbb{R} : \mathbb{P}(|X_i| > s) \leq t\}$. By Corollary 1.1 in Rio (2017) and Hölder's inequality for $r > 2$,

$$\mathrm{Var}\left[\sum_{i=1}^n X_i\right] \leq 4 \sum_{i=1}^n \int_0^1 \alpha^{-1}(t) Q_i(t)^2 \,\mathrm{d}t \quad \leq 4 \sum_{i=1}^n \left(\int_0^1 \alpha^{-1}(t)^{\frac{r}{r-2}} \,\mathrm{d}t\right)^{\frac{r-2}{r}} \left(\int_0^1 |Q_i(t)|^r \,\mathrm{d}t\right)^{\frac{2}{r}} \mathrm{d}t.$$

Now note that if $U \sim \mathcal{U}(0,1)$ then $Q_i(U)$ has the same distribution as X_i. Therefore

$$\operatorname{Var}\left[\sum_{i=1}^n X_i\right] \leq 4 \left(\int_0^1 \alpha^{-1}(t)^{\frac{r}{r-2}} \, \mathrm{d}t\right)^{\frac{r-2}{r}} \sum_{i=1}^n \mathbb{E}[|X_i|^r]^{\frac{2}{r}}.$$

If $\alpha(j) \leq e^{-2j/C_\alpha}$ then $\alpha^{-1}(t) \leq \frac{-C_\alpha \log t}{2}$ so for some constant C_r depending only on r,

$$\operatorname{Var}\left[\sum_{i=1}^n X_i\right] \leq 2C_\alpha \left(\int_0^1 (-\log t)^{\frac{r}{r-2}} \, \mathrm{d}t\right)^{\frac{r-2}{r}} \sum_{i=1}^n \mathbb{E}[|X_i|^r]^{\frac{2}{r}} \leq C_r C_\alpha \sum_{i=1}^n \mathbb{E}[|X_i|^r]^{\frac{2}{r}}.$$

Alternatively, if for constants M_i we have $|X_i| \leq M_i$ a.s. then

$$\operatorname{Var}\left[\sum_{i=1}^n X_i\right] \leq 4 \int_0^1 \alpha^{-1}(t) \, \mathrm{d}t \sum_{i=1}^n M_i^2 \leq 4 \sum_{j=1}^\infty \alpha(j) \sum_{i=1}^n M_i^2.$$

□

Proof (Proposition 5)
We proceed according to the decomposition given in Section 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 β_∞
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 $V_i = \operatorname{Var}[X_i \mid \mathcal{H}_{i-1}] = \sigma_i^2 p(W_i) p(W_i)^\mathsf{T}$. Recall from Theorem 1 that

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

with $Z_i \sim \mathcal{N}(0,1)$ i.i.d. and independent of V_i. For the first term we use Hölder's inequality, the fact that $\sup_w \|p(w)\|_2 \lesssim 1$ and bounded third moments of ε_i:

$$\begin{aligned}
\mathbb{E}\left[\|X_i\|_2^2 \|X_i\|_\infty\right] &\leq \mathbb{E}\left[\|X_i\|_2^3\right]^{2/3} \mathbb{E}\left[\|X_i\|_\infty^3\right]^{1/3} \leq \mathbb{E}\left[\|p(W_i)\varepsilon_i\|_2^3\right]^{2/3} \mathbb{E}\left[\|p(W_i)\varepsilon_i\|_\infty^3\right]^{1/3} \\
&\leq \mathbb{E}\left[|\varepsilon_i|^3\right] \lesssim 1.
\end{aligned}$$

For the second term, we apply Lemma 6 conditionally on $\mathcal{H}_n$ and again use $\sup_w \|p(w)\|_2 \lesssim 1$ to see

$$\begin{aligned}
\mathbb{E}\left[\|V_i^{1/2} Z_i\|_2^2 \|V_i^{1/2} Z_i\|_\infty\right] &\leq 4\sqrt{\log 2k} \, \mathbb{E}\left[\max_{1 \leq j \leq k} (V_i)_{jj}^{1/2} \sum_{j=1}^k (V_i)_{jj}\right] \\
&\leq 4\sqrt{\log 2k} \, \mathbb{E}\left[\sigma_i^3 \max_{1 \leq j \leq k} p(W_i)_j \sum_{j=1}^k p(W_i)_{jj}^2\right] \\
&\leq 4\sqrt{\log 2k} \, \mathbb{E}\left[\sigma_i^3\right] \lesssim \sqrt{\log 2k}.
\end{aligned}$$

Putting these together yields

$$\beta_\infty \lesssim n\sqrt{\log 2k}.$$

Part 2: bounding Ω
Set $\Omega = \sum_{i=1}^n \left(V_i - \mathbb{E}[V_i]\right)$ as in Theorem 1 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 7(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, and so by Jensen's inequality

$$\mathbb{E}\left[\|\Omega\|_2\right] \leq \mathbb{E}\left[\|\Omega\|_\mathrm{F}\right] \leq \left(\sum_{j=1}^k \sum_{l=1}^k \mathbb{E}\left[\Omega_{jl}^2\right]\right)^{1/2} \lesssim \sqrt{nk}.$$

Note that we could have controlled $\|\cdot\|_2$ more tightly by using $\|\cdot\|_1$ rather than $\|\cdot\|_\mathrm{F}$, but this term will be seen to be negligible either way.

Part 3: strong approximation
By Proposition 1 and the previous parts,

$$\|S - T\|_\infty \lesssim_\mathbb{P} \beta_\infty^{1/3}(\log 2k)^{1/3} + \sqrt{\log 2k}\sqrt{\mathbb{E}[\|\Omega\|_2]} \lesssim_\mathbb{P} n^{1/3}\sqrt{\log 2k} + (nk)^{1/4}\sqrt{\log 2k}.$$

by Hölder's inequality and with $\|H^{-1}\|_1 \lesssim k/n$ we have

$$\begin{aligned}
\sup_{w\in\mathcal{W}} \left|p(w)^\mathsf{T} H^{-1} S - p(w)^\mathsf{T} H^{-1} T\right| &\leq \sup_{w\in\mathcal{W}} \|p(w)\|_1 \|H^{-1}\|_1 \|S - T\|_\infty \\
&\lesssim n^{-1}k\left(n^{1/3}\sqrt{\log 2k} + (nk)^{1/4}\sqrt{\log 2k}\right) \\
&\lesssim n^{-2/3}k\sqrt{\log 2k} + n^{-3/4}k^{5/4}\sqrt{\log 2k}.
\end{aligned}$$

Part 4: convergence of $\widehat{H}$
We have $\widehat{H} - H = \sum_{i=1}^n \left(p(W_i)p(W_i)^\mathsf{T} - \mathbb{E}\left[p(W_i)p(W_i)^\mathsf{T}\right]\right)$. Observe that $(\widehat{H} - H)_{jl}$ is the sum of a zero-mean strictly stationary α-mixing sequence and so $\mathbb{E}[(\widehat{H} - H)_{jl}^2] \lesssim n$ by Lemma 7(i). Since the basis functions satisfy Assumption 3 in Cattaneo et al. (2020), $\widehat{H} - H$ has a bounded number of non-zero entries in each row and so by Jensen's inequality

$$\mathbb{E}\left[\|\widehat{H} - H\|_1\right] = \mathbb{E}\left[\max_{1\leq i\leq k}\sum_{j=1}^k \left|(\widehat{H} - H)_{ij}\right|\right] \leq \mathbb{E}\left[\sum_{1\leq i\leq k}\left(\sum_{j=1}^k |(\widehat{H} - H)_{ij}|\right)^2\right]^{1/2} \lesssim \sqrt{nk}.$$

Part 5: bounding the matrix term
Note $\|\widehat{H}^{-1}\|_1 \leq \|H^{-1}\|_1 + \|\widehat{H}^{-1}\|_1\|\widehat{H} - H\|_1\|H^{-1}\|_1$ so by the previous part, we deduce that

$$\|\widehat{H}^{-1}\|_1 \leq \frac{\|H^{-1}\|_1}{1 - \|\widehat{H} - H\|_1\|H^{-1}\|_1} \lesssim_\mathbb{P} \frac{k/n}{1 - \sqrt{nk}\,k/n} \lesssim_\mathbb{P} \frac{k}{n}$$

as $k^3/n \to 0$. Also, 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 4, $\|T\|_\infty \lesssim_\mathbb{P} \sqrt{\frac{n \log 2k}{k}}$. Therefore

$$
\begin{aligned}
\sup_{w\in\mathcal{W}} \left| p(w)^\mathsf{T} (\widehat{H}^{-1} - H^{-1}) S \right| &\leq \sup_{w\in\mathcal{W}} \|p(w)^\mathsf{T}\|_1 \|\widehat{H}^{-1}\|_1 \|\widehat{H} - H\|_1 \|H^{-1}\|_1 \|S - T\|_\infty \\
&\quad + \sup_{w\in\mathcal{W}} \|p(w)^\mathsf{T}\|_1 \|\widehat{H}^{-1}\|_1 \|\widehat{H} - H\|_1 \|H^{-1}\|_1 \|T\|_\infty \\
&\lesssim_\mathbb{P} \frac{k}{n}\sqrt{nk}\frac{k}{n}\left(n^{1/3}\sqrt{\log 2k} + (nk)^{1/4}\sqrt{\log 2k}\right) + \frac{k}{n}\sqrt{nk}\frac{k}{n}\sqrt{\frac{n\log 2k}{k}} \\
&\lesssim_\mathbb{P} \frac{k^2}{n}\sqrt{\log 2k}
\end{aligned}
$$

since $k^3/n \to 0$.

Part 6: conclusion of the main result

By the previous parts, with $G(w) = p(w)^\mathsf{T} H^{-1} T$,

$$
\begin{aligned}
&\sup_{w\in\mathcal{W}} \left| \widehat{\mu}(w) - \mu(w) - p(w)^\mathsf{T} H^{-1} T \right| \\
&\quad = \sup_{w\in\mathcal{W}} \left| p(w)^\mathsf{T} H^{-1} (S - T) + p(w)^\mathsf{T} (\widehat{H}^{-1} - H^{-1}) S + \mathrm{Bias}(w) \right| \\
&\quad \lesssim_\mathbb{P} n^{-2/3} k \sqrt{\log 2k} + n^{-3/4} k^{5/4} \sqrt{\log 2k} + \frac{k^2}{n}\sqrt{\log 2k} + \sup_{w\in\mathcal{W}} |\mathrm{Bias}(w)| \\
&\quad \lesssim_\mathbb{P} n^{-2/3} k \sqrt{\log 2k} + \sup_{w\in\mathcal{W}} |\mathrm{Bias}(w)|
\end{aligned}
$$

since $k^3/n \to 0$. Finally, we verify the upper and lower bounds on the variance of the Gaussian process. Since $\sigma^2(w)$ is bounded above,

$$
\begin{aligned}
\mathrm{Var}[G(w)] &= p(w)^\mathsf{T} H^{-1} \,\mathrm{Var}\left[\sum_{i=1}^n p(W_i)\varepsilon_i\right] H^{-1} p(w) \\
&= p(w)^\mathsf{T} H^{-1} \mathbb{E}\left[\sum_{i=1}^n p(W_i) p(W_i)^\mathsf{T} \sigma^2(W_i)\right] H^{-1} p(w) \\
&\lesssim \|p(w)\|_2^2 \|H^{-1}\|_2^2 \|H\|_2 \lesssim k/n.
\end{aligned}
$$

Similarly, since $\sigma^2(w)$ is bounded away from zero,

$$
\mathrm{Var}[G(w)] \gtrsim \|p(w)\|_2^2 \|H^{-1}\|_2^2 \|H^{-1}\|_2^{-1} \gtrsim k/n.
$$

Part 7: infeasible supremum approximation

Provided that the bias is negligible, for all $s > 0$ we have

$$
\begin{aligned}
&\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\sup_{w\in\mathcal{W}} \left| \frac{\widehat{\mu}(w) - \mu(w)}{\sqrt{\rho(w,w)}} \right| \leq t\right) - \mathbb{P}\left(\sup_{w\in\mathcal{W}} \left| \frac{G(w)}{\sqrt{\rho(w,w)}} \right| \leq t\right) \right| \\
&\quad \leq \sup_{t\in\mathbb{R}} \mathbb{P}\left(t \leq \sup_{w\in\mathcal{W}} \left| \frac{G(w)}{\sqrt{\rho(w,w)}} \right| \leq t + s\right) + \mathbb{P}\left(\sup_{w\in\mathcal{W}} \left| \frac{\widehat{\mu}(w) - \mu(w) - G(w)}{\sqrt{\rho(w,w)}} \right| > s\right).
\end{aligned}
$$

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 8: feasible supremum approximation
By Chernozhukov et al. (2013, Lemma 3.1) and discretization, with $\rho(w,w') = \mathbb{E}[\widehat{\rho}(w,w')]$,

$$\begin{aligned}
&\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\sup_{w\in\mathcal{W}} \left| \frac{\widehat{G}(w)}{\sqrt{\widehat{\rho}(w,w)}} \right| \leq t \;\middle|\; \mathbf{W}, \mathbf{Y} \right) - \mathbb{P}\left(\left| \frac{G(w)}{\sqrt{\rho(w,w)}} \right| \leq t \right) \right| \\
&\quad\lesssim_{\mathbb{P}} \sup_{w,w'\in\mathcal{W}} \left| \frac{\widehat{\rho}(w,w')}{\sqrt{\widehat{\rho}(w,w)\widehat{\rho}(w',w')}} - \frac{\rho(w,w')}{\sqrt{\rho(w,w)\rho(w',w')}} \right|^{1/3} (\log n)^{2/3} \\
&\quad\lesssim_{\mathbb{P}} \left(\frac{n}{k}\right)^{1/3} \sup_{w,w'\in\mathcal{W}} |\widehat{\rho}(w,w') - \rho(w,w')|^{1/3} (\log n)^{2/3} \\
&\quad\lesssim_{\mathbb{P}} \left(\frac{n(\log n)^2}{k}\right)^{1/3} \sup_{w,w'\in\mathcal{W}} \left| p(w)^\mathsf{T} \widehat{H}^{-1} \left(\widehat{\mathrm{Var}}[S] - \mathrm{Var}[S] \right) \widehat{H}^{-1} p(w') \right|^{1/3} \\
&\quad\lesssim_{\mathbb{P}} \left(\frac{k(\log n)^2}{n}\right)^{1/3} \left\| \widehat{\mathrm{Var}}[S] - \mathrm{Var}[S] \right\|_2^{1/3},
\end{aligned}$$

and goes to zero in probability whenever $\frac{k(\log n)^2}{n} \big\| \widehat{\mathrm{Var}}[S] - \mathrm{Var}[S] \big\|_2 \to_{\mathbb{P}} 0$. For the plug-in estimator,

$$\begin{aligned}
\left\| \widehat{\mathrm{Var}}[S] - \mathrm{Var}[S] \right\|_2 &= \left\| \sum_{i=1}^n p(W_i) p(W_i^\mathsf{T}) \widehat{\sigma}^2(W_i) - n\mathbb{E}\left[p(W_i) p(W_i^\mathsf{T}) \sigma^2(W_i) \right] \right\|_2 \\
&\lesssim_{\mathbb{P}} \sup_{w\in\mathcal{W}} |\widehat{\sigma}^2(w) - \sigma^2(w)| \, \|\widehat{H}\|_2 \\
&\quad + \left\| \sum_{i=1}^n p(W_i) p(W_i^\mathsf{T}) \sigma^2(W_i) - n\mathbb{E}\left[p(W_i) p(W_i^\mathsf{T}) \sigma^2(W_i) \right] \right\|_2 \\
&\lesssim_{\mathbb{P}} \frac{n}{k} \sup_{w\in\mathcal{W}} |\widehat{\sigma}^2(w) - \sigma^2(w)| + \sqrt{nk},
\end{aligned}$$

where the second term is bounded by the same argument used to bound $\|\widehat{H} - H\|_1$. Thus the feasible approximation is valid whenever $(\log n)^2 \sup_{w\in\mathcal{W}} |\widehat{\sigma}^2(w) - \sigma^2(w)| \to_{\mathbb{P}} 0$ and $\frac{k^3(\log n)^4}{n} \to 0$. Recall that the infeasible version requires that also $\frac{k^3(\log n)^6}{n} \to 0$. □

Proof (Proposition 6)
Applying Proposition 2 with $\mathcal{A} = \mathcal{B}_p$ gives

$$\begin{aligned}
\sup_{t\geq 0} \left| \mathbb{P}(\|S\|_p \leq t) - \mathbb{P}\left(\|T\|_p \leq t\right) \right| &= \sup_{A\in\mathcal{B}_p} \left| \mathbb{P}(S\in A) - \mathbb{P}(T\in A) \right| \\
&\leq \inf_{\eta>0} \left\{ \Gamma_p(\eta) + \Delta_p(\mathcal{B}_p, \eta) \right\} \leq \inf_{\eta>0} \left\{ \Gamma_p(\eta) + \Delta_p(\eta) \right\}.
\end{aligned}$$

□

Lemma 8 (Exponential inequalities for α-mixing random variables)
Let $X_1, \ldots, X_n$ be zero-mean real-valued random variables with α-mixing coefficients $\alpha(j) \leq e^{-2j/C_\alpha}$.

(i) Suppose $|X_i| \leq M$ a.s. for each $1 \leq i \leq n$. Then for all $t > 0$ there is a constant C_1 such that

$$\mathbb{P}\left(\left| \sum_{i=1}^n X_i \right| > C_1 M \big(\sqrt{nt} + (\log n)(\log\log n) t \big) \right) \leq C_1 e^{-t}.$$

(ii) Suppose further that $\sum_{j=1}^n |\mathrm{Cov}[X_i, X_j]| \leq \sigma^2$*. Then for all* $t > 0$ *there is a constant* C_2 *with*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq C_2\big((\sigma\sqrt{n} + M)\sqrt{t} + M(\log n)^2 t\big)\right) \leq C_2 e^{-t}.$$

Proof (Lemma 8)
We apply results from Merlevède et al. (2009), adjusting constants where necessary.

(i) By Theorem 1 in Merlevède et al. (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 $M\sqrt{n}\sqrt{t} + M(\log n)(\log\log n)t$.

(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 $\sigma\sqrt{n}\sqrt{t} + M\sqrt{t} + M(\log n)^2 t$.

□

Proof (Proposition 7)
We apply Proposition 4 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 $\widetilde{\varepsilon}_i = \varepsilon_i \mathbb{I}\{-a \log n \leq \varepsilon_i \leq b \log n\}$ and note that $\mathbb{E}\big[e^{|\varepsilon_i|/C_\varepsilon}\big] < \infty$ implies that $\mathbb{P}(\exists i : \widetilde{\varepsilon}_i \neq \varepsilon_i) \lesssim n^{1-(a \vee b)/C_\varepsilon}$. Hence, by choosing a and b large enough, with high probability, we can replace all ε_i by $\widetilde{\varepsilon}_i$. Further, it is always possible to increase either a or b along with some randomization to ensure that $\mathbb{E}[\widetilde{\varepsilon}_i] = 0$. Since K is bounded and compactly supported, W_i has a bounded density and $|\widetilde{\varepsilon}_i| \lesssim \log n$,

$$
\begin{aligned}
\big\|\!\big| f(W_i, \widetilde{\varepsilon}_i) \big|\!\big\|_2 &= \mathbb{E}\left[\left| e_1^\mathsf{T} H(w)^{-1} K_h(W_i - w) p_h(W_i - w) \widetilde{\varepsilon}_i \right|^2\right]^{1/2} \\
&\leq \mathbb{E}\left[\|H(w)^{-1}\|_2^2 K_h(W_i - w)^2 \|p_h(W_i - w)\|_2^2 \sigma^2(W_i)\right]^{1/2} \\
&\lesssim n^{-1} \mathbb{E}\left[K_h(W_i - w)^2\right]^{1/2} \lesssim n^{-1} h^{-m/2}, \\
\big\|\!\big| f(W_i, \widetilde{\varepsilon}_i) \big|\!\big\|_\infty &\leq \big\|\!\big| \|H(w)^{-1}\|_2 K_h(W_i - w) \|p_h(W_i - w)\|_2 |\widetilde{\varepsilon}_i| \big|\!\big\|_\infty \\
&\lesssim n^{-1} \big\|\!\big| K_h(W_i - w) \big|\!\big\|_\infty \log n \lesssim n^{-1} h^{-m} \log n.
\end{aligned}
$$

Therefore

$$
\mathbb{E}\left[\|\mathcal{F}_\delta(W_i, \widetilde{\varepsilon}_i)\|_2^2 \|\mathcal{F}_\delta(W_i, \widetilde{\varepsilon}_i)\|_\infty\right] \leq \sum_{f \in \mathcal{F}_\delta} \big\|\!\big| f(W_i, \widetilde{\varepsilon}_i) \big|\!\big\|_2^2 \max_{f \in \mathcal{F}_\delta} \big\|\!\big| f(W_i, \widetilde{\varepsilon}_i) \big|\!\big\|_\infty \lesssim n^{-3} \delta^{-m} h^{-2m} \log n.
$$

Let $V_i(\mathcal{F}_\delta) = \mathbb{E}\big[\mathcal{F}_\delta(W_i, \widetilde{\varepsilon}_i) \mathcal{F}_\delta(W_i, \widetilde{\varepsilon}_i)^\mathsf{T} \mid \mathcal{H}_{i-1}\big]$ and $Z_i \sim \mathcal{N}(0, I_d)$ be i.i.d. and independent of $\mathcal{H}_n$. Note that $V_i(f, f) = \mathbb{E}[f(W_i, \widetilde{\varepsilon}_i)^2 \mid W_i] \lesssim n^{-2} h^{-2m}$ and $\mathbb{E}[V_i(f, f)] = \mathbb{E}[f(W_i, \widetilde{\varepsilon}_i)^2] \lesssim n^{-2} h^{-m}$. Thus by Lemma 6,

$$
\begin{aligned}
\mathbb{E}\left[\big\|V_i(\mathcal{F}_\delta)^{1/2} Z_i\big\|_2^2 \big\|V_i(\mathcal{F}_\delta)^{1/2} Z_i\big\|_\infty\right] &= \mathbb{E}\left[\mathbb{E}\left[\big\|V_i(\mathcal{F}_\delta)^{1/2} Z_i\big\|_2^2 \big\|V_i(\mathcal{F}_\delta)^{1/2} Z_i\big\|_\infty \mid \mathcal{H}_n\right]\right] \\
&\leq 4 \sqrt{\log 2|\mathcal{F}_\delta|}\, \mathbb{E}\left[\max_{f \in \mathcal{F}_\delta} \sqrt{V_i(f, f)} \sum_{f \in \mathcal{F}_\delta} V_i(f, f)\right] \\
&\lesssim n^{-3} h^{-2m} \delta^{-m} \sqrt{\log(1/\delta)}.
\end{aligned}
$$

Thus since $\log(1/\delta) \asymp \log(1/h) \asymp \log n$,

$$
\beta_\delta = \sum_{i=1}^n \mathbb{E}\left[\|\mathcal{F}_\delta(W_i, \widetilde{\varepsilon}_i)\|_2^2 \|\mathcal{F}_\delta(W_i, \widetilde{\varepsilon}_i)\|_\infty + \big\|V_i(\mathcal{F}_\delta)^{1/2} Z_i\big\|_2^2 \big\|V_i(\mathcal{F}_\delta)^{1/2} Z_i\big\|_\infty\right] \lesssim \frac{\log n}{n^2 h^{2m} \delta^m}.
$$

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

$$
\begin{aligned}
\big|V_i(f, f')\big| &= \left|\mathbb{E}\left[e_1^\mathsf{T} H(w)^{-1} p_h(W_i - w) e_1^\mathsf{T} H(w')^{-1} p_h(W_i - w') K_h(W_i - w) K_h(W_i - w') \widetilde{\varepsilon}_i^2 \;\middle|\; \mathcal{H}_{i-1}\right]\right| \\
&\lesssim n^{-2} K_h(W_i - w) K_h(W_i - w') \lesssim n^{-2} h^{-m} K_h(W_i - w) \mathbb{I}\{\|w - w'\|_2 \leq 2 C_K h\}.
\end{aligned}
$$

Since W_i are α-mixing with $\alpha(j) < e^{-2j/C_\alpha}$, Lemma 7(ii) with $r = 3$ gives

$$
\begin{aligned}
\operatorname{Var}\left[\sum_{i=1}^n V_i(f, f')\right] &\lesssim \sum_{i=1}^n \mathbb{E}\left[|V_i(f, f')|^3\right]^{2/3} \\
&\lesssim n^{-3} h^{-2m} \mathbb{E}\left[K_h(W_i - w)^3\right]^{2/3} \mathbb{I}\{\|w - w'\|_2 \leq 2 C_K h\} \\
&\lesssim n^{-3} h^{-2m} (h^{-2m})^{2/3} \mathbb{I}\{\|w - w'\|_2 \leq 2 C_K h\} \\
&\lesssim n^{-3} h^{-10m/3} \mathbb{I}\{\|w - w'\|_2 \leq 2 C_K h\}.
\end{aligned}
$$

Therefore, by Jensen's inequality,

$$\begin{aligned}
\mathbb{E}\big[\|\Omega_\delta\|_2\big] \leq \mathbb{E}\big[\|\Omega_\delta\|_{\mathrm{F}}\big] \leq \mathbb{E}\Bigg[\sum_{f,f' \in \mathcal{F}_\delta} (\Omega_\delta)_{f,f'}^2\Bigg]^{1/2} &\leq \left(\sum_{f,f' \in \mathcal{F}_\delta} \mathrm{Var}\left[\sum_{i=1}^n V_i(f,f')\right]\right)^{1/2} \\
&\lesssim n^{-3/2} h^{-5m/3} \left(\sum_{f,f' \in \mathcal{F}_\delta} \mathbb{I}\{\|w - w'\|_2 \leq 2 C_K h\}\right)^{1/2} \\
&\lesssim n^{-3/2} h^{-5m/3} \big(h^m \delta^{-2m}\big)^{1/2} \lesssim n^{-3/2} h^{-7m/6} \delta^{-m}.
\end{aligned}$$

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

$$u_i(f,f') = \big(e_1^\top H(w)^{-1} K_h(W_i - w) p_h(W_i - w) - e_1^\top H(w')^{-1} K_h(W_i - w') p_h(W_i - w')\big) \widetilde{\varepsilon}_i.$$

To bound this we use that for all $1 \leq j \leq k$, by the Lipschitz property of the kernel and monomials,

$$\begin{aligned}
\big|K_h(W_i - w) - K_h(W_i - w')\big| &\lesssim h^{-m-1} \|w - w'\|_2 \big(\mathbb{I}\{\|W_i - w\| \leq C_K h\} + \mathbb{I}\{\|W_i - w'\| \leq C_K h\}\big), \\
\big|p_h(W_i - w)_j - p_h(W_i - w')_j\big| &\lesssim h^{-1} \|w - w'\|_2,
\end{aligned}$$

to deduce that for any $1 \leq j, l \leq k$,

$$\begin{aligned}
&\big|H(w)_{jl} - H(w')_{jl}\big| \\
&\quad = \big|n \mathbb{E}\left[K_h(W_i - w) p_h(W_i - w)_j p_h(W_i - w)_l - K_h(W_i - w') p_h(W_i - w')_j p_h(W_i - w')_l\right]\big| \\
&\quad \leq n \mathbb{E}\left[\big|K_h(W_i - w) - K_h(W_i - w')\big| \left|p_h(W_i - w)_j p_h(W_i - w)_l\right|\right] \\
&\qquad + n \mathbb{E}\left[\big|p_h(W_i - w)_j - p_h(W_i - w')_j\big| \big|K_h(W_i - w') p_h(W_i - w)_l\big|\right] \\
&\qquad + n \mathbb{E}\left[\big|p_h(W_i - w)_l - p_h(W_i - w')_l\big| \big|K_h(W_i - w') p_h(W_i - w')_j\big|\right] \\
&\quad \lesssim n h^{-1} \|w - w'\|_2.
\end{aligned}$$

Therefore as the dimension of the matrix $H(w)$ is fixed,

$$\big\|H(w)^{-1} - H(w')^{-1}\big\|_2 \leq \big\|H(w)^{-1}\big\|_2 \big\|H(w')^{-1}\big\|_2 \big\|H(w) - H(w')\big\|_2 \lesssim \frac{\|w - w'\|_2}{nh}.$$

Hence

$$\begin{aligned}
\big|u_i(f,f')\big| &\leq \big\|H(w)^{-1} K_h(W_i - w) p_h(W_i - w) - H(w')^{-1} K_h(W_i - w') p_h(W_i - w') \widetilde{\varepsilon}_i\big\|_2 \\
&\leq \big\|H(w)^{-1} - H(w')^{-1}\big\|_2 \big\|K_h(W_i - w) p_h(W_i - w) \widetilde{\varepsilon}_i\big\|_2 \\
&\quad + \big|K_h(W_i - w) - K_h(W_i - w')\big| \big\|H(w')^{-1} p_h(W_i - w) \widetilde{\varepsilon}_i\big\|_2 \\
&\quad + \big\|p_h(W_i - w) - p_h(W_i - w')\big\|_2 \big\|H(w')^{-1} K_h(W_i - w') \widetilde{\varepsilon}_i\big\|_2 \\
&\lesssim \frac{\|w - w'\|_2}{nh} \big|K_h(W_i - w) \widetilde{\varepsilon}_i\big| + \frac{1}{n} \big|K_h(W_i - w) - K_h(W_i - w')\big| \left|\widetilde{\varepsilon}_i\right| \\
&\lesssim \frac{\|w - w'\|_2 \log n}{n h^{m+1}},
\end{aligned}$$

and from the penultimate line, we also deduce that

$$\begin{aligned}\operatorname{Var}[u_i(f,f')] &\lesssim \frac{\|w-w'\|_2^2}{n^2h^2}\mathbb{E}\left[K_h(W_i-w)^2\sigma^2(X_i)\right] + \frac{1}{n^2}\mathbb{E}\left[\left(K_h(W_i-w)-K_h(W_i-w')\right)^2\sigma^2(X_i)\right] \\ &\lesssim \frac{\|w-w'\|_2^2}{n^2h^{m+2}}.\end{aligned}$$

Further, $\mathbb{E}[u_i(f,f')u_j(f,f')] = 0$ for $i \neq j$ so by Lemma 8(ii), for a constant $C_1 > 0$,

$$\mathbb{P}\left(\Big|\sum_{i=1}^n u_i(f,f')\Big| \geq \frac{C_1\|w-w'\|_2}{\sqrt{n}h^{m/2+1}}\left(\sqrt{t} + \sqrt{\frac{(\log n)^2}{nh^m}}\sqrt{t} + \sqrt{\frac{(\log n)^6}{nh^m}t}\right)\right) \leq C_1 e^{-t}.$$

Therefore, adjusting the constant if necessary and since $nh^m \gtrsim (\log n)^7$,

$$\mathbb{P}\left(\Big|\sum_{i=1}^n u_i(f,f')\Big| \geq \frac{C_1\|w-w'\|_2}{\sqrt{n}h^{m/2+1}}\left(\sqrt{t} + \frac{t}{\sqrt{\log n}}\right)\right) \leq C_1 e^{-t}.$$

By Lemma 2 in van de Geer and Lederer (2013) with $\psi(x) = \exp\left(\left(\sqrt{1+2x/\sqrt{\log n}}-1\right)^2 \log n\right) - 1$,

$$\Big\|\sum_{i=1}^n u_i(f,f')\Big\|_\psi \lesssim \frac{\|w-w'\|_2}{\sqrt{n}h^{m/2+1}}$$

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 $\widetilde{\varepsilon}_i = \varepsilon_i$ for all i with high probability, by Proposition 4, 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

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{w\in\mathcal{W}} \left|\sum_{i=1}^n f_w(W_i,\varepsilon_i) - T(w)\right| \geq C_\psi(t+\eta)\right) \\
&\leq C_\psi \inf_{\delta>0}\inf_{\mathcal{F}_\delta}\left\{\frac{\beta_\delta^{1/3}(\log 2|\mathcal{F}_\delta|)^{1/3}}{\eta} + \left(\frac{\sqrt{\log 2|\mathcal{F}_\delta|}\sqrt{\mathbb{E}\left[\|\Omega_\delta\|_2\right]}}{\eta}\right)^{2/3} \right.\\
&\quad \left. + \psi\left(\frac{t}{LJ_\psi(\delta)}\right)^{-1} + \exp\left(\frac{-t^2}{L^2 J_2(\delta)^2}\right)\right\} \\
&\leq C_\psi\left\{\frac{\left(\frac{\log n}{n^2h^{2m}\delta^m}\right)^{1/3}(\log n)^{1/3}}{\eta} + \left(\frac{\sqrt{\log n}\sqrt{n^{-3/2}h^{-7m/6}\delta^{-m}}}{\eta}\right)^{2/3}\right.\\
&\quad \left. + \psi\left(\frac{t}{\frac{1}{\sqrt{n}h^{m/2+1}}J_\psi(\delta)}\right)^{-1} + \exp\left(\frac{-t^2}{\left(\frac{1}{\sqrt{n}h^{m/2+1}}\right)^2 J_2(\delta)^2}\right)\right\} \\
&\leq C_\psi\left\{\frac{(\log n)^{2/3}}{n^{2/3}h^{2m/3}\delta^{m/3}\eta} + \left(\frac{n^{-3/4}h^{-7m/12}\delta^{-m/2}\sqrt{\log n}}{\eta}\right)^{2/3}\right.\\
&\quad \left. + \psi\left(\frac{t\sqrt{n}h^{m/2+1}}{\delta\sqrt{\log n}}\right)^{-1} + \exp\left(\frac{-t^2nh^{m+2}}{\delta^2\log n}\right)\right\}.
\end{aligned}
$$

Noting that $\psi(x) \geq e^{x^2/4}$ for $x \leq 4\sqrt{\log n}$ gives the probability bound

$$
\sup_{w\in\mathcal{W}}\left|\sum_{i=1}^n f_w(W_i,\varepsilon_i) - T(w)\right| \lesssim_\mathbb{P} \frac{(\log n)^{2/3}}{n^{2/3}h^{2m/3}\delta^{m/3}} + \frac{\sqrt{\log n}}{n^{3/4}h^{7m/12}\delta^{m/2}} + \frac{\delta\sqrt{\log n}}{\sqrt{n}h^{m/2+1}}.
$$

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

$$
\sup_{w\in\mathcal{W}}\left|\sum_{i=1}^n f_w(W_i,\varepsilon_i) - T(w)\right| \lesssim_\mathbb{P} \left(\frac{(\log n)^{m+4}}{n^{m+4}h^{m(m+6)}}\right)^{\frac{1}{2m+6}}.
$$

Part 6: convergence of $\widehat{H}(w)$

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 8(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 cardinality at most $n^a \delta^{-a}$ for some $a > 0$. Adjusting C_2,

$$\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\sum_{i=1}^n u_{ijl}(w)\right| > C_2 h^{-m}\big(\sqrt{nt} + (\log n)(\log\log n)t\big) + C_2 h^{-m-1}\delta\right) \le C_2 n^a \delta^{-a} e^{-t}$$

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

$$\sup_{w\in\mathcal{W}}\|\widehat{H}(w) - H(w)\|_2 \lesssim_{\mathbb{P}} \sqrt{\frac{n\log n}{h^{2m}}}.$$

Part 7: bounding the matrix term

Firstly note that, since $\sqrt{\frac{\log n}{nh^{2m}}} \to 0$, we have that uniformly in $w \in \mathcal{W}$

$$\|\widehat{H}(w)^{-1}\|_2 \le \frac{\|H(w)^{-1}\|_2}{1 - \|\widehat{H}(w) - H(w)\|_2\|H(w)^{-1}\|_2} \lesssim_{\mathbb{P}} \frac{1/n}{1 - \sqrt{\frac{n\log n}{h^{2m}}}\frac{1}{n}} \lesssim_{\mathbb{P}} \frac{1}{n}.$$

Therefore

$$\begin{aligned}\sup_{w\in\mathcal{W}}\left|e_1^\mathsf{T}\big(\widehat{H}(w)^{-1} - H(w)^{-1}\big)S(w)\right| &\le \sup_{w\in\mathcal{W}}\left\|\widehat{H}(w)^{-1} - H(w)^{-1}\right\|_2\|S(w)\|_2 \\ &\le \sup_{w\in\mathcal{W}}\left\|\widehat{H}(w)^{-1}\right\|_2\left\|H(w)^{-1}\right\|_2\left\|\widehat{H}(w) - H(w)\right\|_2\|S(w)\|_2 \\ &\lesssim_{\mathbb{P}} \sqrt{\frac{\log n}{n^3h^{2m}}}\sup_{w\in\mathcal{W}}\|S(w)\|_2.\end{aligned}$$

Now for $1 \le j \le k$ write $u_{ij}(w) = K_h(W_i - w)p_h(W_i - w)_j\widetilde{\varepsilon}_i$ 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 $i \ne i'$. Also $|u_{ij}(w)| \lesssim h^{-m}\log n$ and $\mathrm{Var}[u_{ij}(w)] \lesssim h^{-m}$. Thus by Lemma 8(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

$$\sup_{w\in\mathcal{W}}\left|e_1^\mathsf{T}\big(\widehat{H}(w)^{-1} - H(w)^{-1}\big)S(w)\right| \lesssim_{\mathbb{P}} \sqrt{\frac{\log n}{n^3h^{2m}}}\sqrt{\frac{n\log n}{h^m}} \lesssim_{\mathbb{P}} \frac{\log n}{\sqrt{n^2h^{3m}}}.$$

Part 8: bounding the bias

Since $\mu \in \mathcal{C}^\gamma$, we have by the multivariate version of Taylor's theorem,

$$\mu(W_i) = \sum_{|\nu|=0}^{\gamma-1} \frac{1}{\nu!} \frac{\partial^{|\nu|}\mu(w)}{\partial w^\nu}(W_i - w)^\nu + \sum_{|\nu|=\gamma} \frac{1}{\nu!} \frac{\partial^{\gamma}\mu(w')}{\partial w^\nu}(W_i - w)^\nu$$

for some w' on the line segment connecting w and W_i. Now since $p_h(W_i - w)_1 = 1$,

$$\begin{aligned}
&e_1^\mathsf{T} \widehat{H}(w)^{-1} \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) \mu(w) \\
&\quad = e_1^\mathsf{T} \widehat{H}(w)^{-1} \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) p_h(W_i - w)^\mathsf{T} e_1 \mu(w) = e_1^\mathsf{T} e_1 \mu(w) = \mu(w).
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathrm{Bias}(w) &= e_1^\mathsf{T} \widehat{H}(w)^{-1} \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) \mu(W_i) - \mu(w) \\
&= e_1^\mathsf{T} \widehat{H}(w)^{-1} \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) \\
&\quad \times \left(\sum_{|\nu|=0}^{\gamma-1} \frac{1}{\nu!} \frac{\partial^{|\nu|}\mu(w)}{\partial w^\nu}(W_i - w)^\nu + \sum_{|\nu|=\gamma} \frac{1}{\nu!} \frac{\partial^{\gamma}\mu(w')}{\partial w^\nu}(W_i - w)^\nu - \mu(w) \right) \\
&= \sum_{|\nu|=0}^{\gamma-1} \frac{1}{\nu!} \frac{\partial^{|\nu|}\mu(w)}{\partial w^\nu} e_1^\mathsf{T} \widehat{H}(w)^{-1} \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w)(W_i - w)^\nu \\
&\quad + \sum_{|\nu|=\gamma} \frac{1}{\nu!} \frac{\partial^{\gamma}\mu(w')}{\partial w^\nu} e_1^\mathsf{T} \widehat{H}(w)^{-1} \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w)(W_i - w)^\nu \\
&= \sum_{|\nu|=\gamma} \frac{1}{\nu!} \frac{\partial^{\gamma}\mu(w')}{\partial w^\nu} e_1^\mathsf{T} \widehat{H}(w)^{-1} \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w)(W_i - w)^\nu,
\end{aligned}$$

where we used that $p_h(W_i - w)$ is a vector containing all monomials in $W_i - w$ of order up to γ, so $e_1^\mathsf{T} \widehat{H}(w)^{-1} \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w)(W_i - w)^\nu = 0$ whenever $1 \leq |\nu| \leq \gamma$. Finally

$$\begin{aligned}
\sup_{w \in \mathcal{W}} |\mathrm{Bias}(w)| &= \sup_{w \in \mathcal{W}} \left| \sum_{|\nu|=\gamma} \frac{1}{\nu!} \frac{\partial^{\gamma}\mu(w')}{\partial w^\nu} e_1^\mathsf{T} \widehat{H}(w)^{-1} \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w)(W_i - w)^\nu \right| \\
&\lesssim_\mathbb{P} \sup_{w \in \mathcal{W}} \max_{|\nu|=\gamma} \left| \frac{\partial^{\gamma}\mu(w')}{\partial w^\nu} \right| \|\widehat{H}(w)^{-1}\|_2 \left\| \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) \right\|_2 h^\gamma \\
&\lesssim_\mathbb{P} \frac{h^\gamma}{n} \sup_{w \in \mathcal{W}} \left\| \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) \right\|_2 .
\end{aligned}$$

Now write $\widetilde{u}_{ij}(w) = K_h(W_i - w) p_h(W_i - w)_j$ and note that $|\widetilde{u}_{ij}(w)| \lesssim h^{-m}$ and $\mathbb{E}[\widetilde{u}_{ij}(w)] \lesssim 1$. By Lemma 8(i), for a constant C_4,

$$\mathbb{P}\left(\left| \sum_{i=1}^n \widetilde{u}_{ij}(w) - \mathbb{E}\left[\sum_{i=1}^n \widetilde{u}_{ij}(w) \right] \right| > C_4 h^{-m} \big(\sqrt{nt} + (\log n)(\log\log n) t \big) \right) \leq C_4 e^{-t}.$$

As in previous parts, by Lipschitz properties, this implies

$$\sup_{w \in \mathcal{W}} \left| \sum_{i=1}^n \widetilde{u}_{ij}(w) \right| \lesssim_\mathbb{P} n \left(1 + \sqrt{\frac{\log n}{n h^{2m}}} \right) \lesssim_\mathbb{P} n.$$

Therefore $\sup_{w \in \mathcal{W}} |\mathrm{Bias}(w)| \lesssim_\mathbb{P} n h^\gamma / n \lesssim_\mathbb{P} h^\gamma$.

Part 9: conclusion
By the previous parts,

$$\begin{aligned}
&\sup_{w \in \mathcal{W}} |\widehat{\mu}(w) - \mu(w) - T(w)| \\
&\quad \leq \sup_{w \in \mathcal{W}} \left| e_1^\mathsf{T} H(w)^{-1} S(w) - T(w) \right| + \sup_{w \in \mathcal{W}} \left| e_1^\mathsf{T} \big(\widehat{H}(w)^{-1} - H(w)^{-1} \big) S(w) \right| + \sup_{w \in \mathcal{W}} |\mathrm{Bias}(w)| \\
&\quad \lesssim_\mathbb{P} \left(\frac{(\log n)^{m+4}}{n^{m+4} h^{m(m+6)}} \right)^{\frac{1}{2m+6}} + \frac{\log n}{\sqrt{n^2 h^{3m}}} + h^\gamma \lesssim_\mathbb{P} \frac{1}{\sqrt{n h^m}} \left(\frac{(\log n)^{m+4}}{n h^{3m}} \right)^{\frac{1}{2m+6}} + h^\gamma,
\end{aligned}$$

where the last inequality follows because $n h^{3m} \to \infty$ 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 $1/n$,

$$\begin{aligned}
\mathrm{Var}[T(w)] &= \mathrm{Var}\left[e_1^\mathsf{T} H(w)^{-1} \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) \varepsilon_i \right] \\
&= e_1^\mathsf{T} H(w)^{-1} \, \mathrm{Var}\left[\sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) \varepsilon_i \right] H(w)^{-1} e_1^\mathsf{T} \\
&\lesssim \|H(w)^{-1}\|_2^2 \max_{1 \leq j \leq k} \sum_{i=1}^n \mathrm{Var}\left[K_h(W_i - w) p_h(W_i - w)_j \sigma(W_i) \right] \lesssim \frac{1}{n^2} n \frac{1}{h^m} \lesssim \frac{1}{n h^m}.
\end{aligned}$$

Similarly $\mathrm{Var}[T(w)] \gtrsim \frac{1}{n h^m}$ by the same argument given to bound the eigenvalues of $H(w)^{-1}$. □

References

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*, pages 115–137. PMLR.

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. *arXiv preprint arXiv:1809.02741*.

Berthet, P. and Mason, D. M. (2006). Revisiting two strong approximation results of Dudley and Philipp. *Lecture Notes–Monograph Series*, pages 155–172.

Bhatia, R. (1997). *Matrix Analysis*, volume 169. Springer, New York, NY.

Biau, G. and Mason, D. M. (2015). High-dimensional p-norms. In *Mathematical statistics and limit theorems*, pages 21–40. Springer.

Buzun, N., Shvetsov, N., and Dylov, D. V. (2022). Strong Gaussian approximation for the sum of random vectors. In *Conference on Learning Theory*, pages 1693–1715. PMLR.

Cattaneo, M. D., Farrell, M. H., and Feng, Y. (2020). Large sample properties of partitioning-based series estimators. *The Annals of Statistics*, 48(3):1718–1741.

Cattaneo, M. D., Feng, Y., and Underwood, W. G. (2022). Uniform inference for kernel density estimators with dyadic data. *arXiv preprint arXiv:2201.05967*.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014a). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014b). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597.

Chernozhukov, V., Chetverikov, D., and Koike, Y. (2022). Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *Annals of Applied Probability,* forthcoming.

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.

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.

Eggermont, P. P. B. and LaRiccia, V. N. (2009). *Maximum Penalized Likelihood Estimation: Volume II: Regression.* Springer.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman & Hall/CRC, New York.

Götze, F., Naumov, A., Spokoiny, V., and Ulyanov, V. (2019). Large ball probabilities, Gaussian comparison and anti-concentration. *Bernoulli*, 25(4A):2538–2563.

Hall, P. and Heyde, C. C. (2014). *Martingale Limit Theory and its Application.* Academic Press.

Jacod, J., Li, J., and Liao, Z. (2021). Volatility coupling. *The Annals of Statistics*, 49(4):1982–1998.

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. *arXiv preprint arXiv:2107.10766*.

Le Cam, L. (1988). On the Prokhorov distance between the empirical process and the associated Gaussian bridge. Technical report, University of California, Berkley.

Li, J. and Liao, Z. (2020). Uniform nonparametric inference for time series. *Journal of Econometrics*, 219(1):38–51.

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. *The Annals of Statistics*, 48(2):1214–1229.

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*, pages 273–292. Institute of Mathematical Statistics.

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*, pages 169–187. Springer.

Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability.* 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. (1994). Local invariance principles and their application to density estimation. *Probability Theory and Related Fields*, 98(1):21–45.

Rio, E. (2017). *Asymptotic theory of weakly dependent random processes*, volume 80. Springer.

Sheehy, A. and Wellner, J. A. (1992). Uniform Donsker classes of functions. *The Annals of Probability*, 20(4):1983–2030.

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. 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.

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.