Yurinskii's Coupling for Martingales

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

February 29, 2024

Abstract

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

Keywords: coupling, strong approximation, mixingales, martingales, dependent data, Gaussian mixture approximation, time series, empirical processes, uniform inference, series estimation, local polynomial estimation.

[1]Department of Operations Research and Financial Engineering, Princeton University.
[2]Department of Statistics, University of California, Davis.

Contents

1 Introduction

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

When compared to other coupling approaches, such as the celebrated Hungarian construction (Komlós et al., 1975) or Zaitsev's coupling (Zaitsev, 1987a,b), Yurinskii's approach stands out for its simplicity, robustness, and wider applicability, while also offering tighter couplings in some applications (see below for more discussion and examples). These features have led many scholars to use Yurinskii's coupling to study the distributional features of high-dimensional statistical procedures in a variety of settings, often with the end goal of developing uncertainty quantification or hypothesis testing methods. For example, in recent years, Yurinskii's coupling has been used to construct Gaussian approximations for the suprema of empirical processes (Chernozhukov et al., 2014b); to establish distribution theory for non-Donsker stochastic t-processes generated in nonparametric series regression (Belloni et al., 2015); to prove distributional approximations for high-dimensional ℓ^p-norms (Biau and Mason, 2015); to develop distribution theory for vector-valued martingales (Belloni and Oliveira, 2018; Li and Liao, 2020); to derive a law of the iterated logarithm for stochastic gradient descent optimization methods (Anastasiou et al., 2019); to establish uniform distributional results for nonparametric high-dimensional quantile processes (Belloni et al., 2019); to develop distribution theory for non-Donsker stochastic t-processes generated in partitioning-based series regression (Cattaneo et al., 2020); to deduce Bernstein–von Mises theorems in high-dimensions (Ray and van der Vaart, 2021); and to develop distribution theory for non-Donsker U-processes based on dyadic network data (Cattaneo et al., 2024). There are also many other early applications of Yurinskii's coupling: Dudley and Philipp (1983) and Dehling (1983) establish invariance principles for Banach space-valued random variables, and Le Cam (1988) and Sheehy and Wellner (1992) obtain uniform Donsker results for empirical processes, to name just a few.

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

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

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

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

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

Closest to our work are the unpublished manuscript by Belloni and Oliveira (2018) and the recent paper by Li and Liao (2020), which both investigated distribution theory for martingale data using Yurinskii's coupling and related methods. Specifically, Li and Liao (2020) established a Gaussian ℓ^2-norm Yurinskii coupling for mixingales and martingales under the assumption that the covariance structure has a minimum eigenvalue bounded away from zero. As formally demonstrated in this paper (Section 3.1), such eigenvalue assumptions can be prohibitively strong in practically relevant applications. In contrast, our Yurinskii coupling does not impose any restrictions on covariance degeneracy (iii), in addition to offering several other new features not present in Li and Liao (2020), including (i), (ii), (iv), and applicability to general ℓ^p-norms. In addition, we correct a slight technical inaccuracy in their proof relating to the derivation of bounds in probability (Remark 2.1). Belloni and Oliveira (2018) did not establish a Yurinskii coupling for martingales, but rather a central limit theorem for smooth functions of high-dimensional martingales using the celebrated second-order Lindeberg method (see Chatterjee, 2006, and references therein), explicitly accounting for covariance degeneracy. As a consequence, their result could be leveraged to deduce a Yurinskii coupling for martingales with additional, non-trivial technical work (see our Appendix A for details). Nevertheless, a Yurinskii coupling derived from Belloni and Oliveira (2018) would not feature (i), (ii), (iv), or general ℓ^p-norms, as our results do. We discuss further the connections between our work and the related literature in the upcoming sections, both when introducing our main theoretical results and when presenting the examples and statistical applications.

The most general coupling result of this paper (Theorem 2.1) is presented in Section 2, where we also specialize it to a slightly weaker yet more user-friendly formulation (Proposition 2.1). Our Yurinskii coupling for approximate martingales is a strict generalization of all previous Yurinskii couplings available in the literature, offering a Gaussian mixture strong approximation for approximate martingale vectors in ℓ^p-norm, with an improved rate of approximation when the third moments of the data are negligible, and with no assumptions on the spectrum of the data covariance matrix. A key technical innovation underlying the proof of Theorem 2.1 is that we explicitly account for the possibility that the minimum eigenvalue of the variance may be zero, or its lower bound may be unknown, with the argument proceeding using a carefully tailored regularization. Establishing a coupling to a Gaussian mixture distribution is achieved by an appropriate conditioning argument, leveraging a conditional version of Strassen's theorem established by Chen and Kato (2020), along with some related technical work detailed in Appendix A. A third-order coupling is obtained via a modification of a standard smoothing technique for Borel sets from classical versions of Yurinskii's coupling, enabling improved approximation errors whenever third moments are negligible.

In Proposition 2.1, we explicitly tune the parameters of the aforementioned regularization to obtain a simpler, parameter-free version of Yurinskii's coupling for approximate martingales, again offering Gaussian mixture coupling distributions and an improved third-order approximation error. This specialization of our main result takes an agnostic approach to potential singularities in the data covariance matrix and, as such, may be improved in specific applications where additional knowledge of the covariance structure is available. Section 2 also presents some further refinements when additional structure is imposed, deriving Yurinskii couplings for mixingales, martingales and independent data as Corollaries 2.1, 2.2 and 2.3, respectively. We take the opportunity to discuss and correct in Remark 2.1 a technical issue which is often neglected (Pollard, 2002; Li and Liao, 2020) when using Yurinskii's coupling to derive bounds in probability. Section 2.5 presents a stylized example portraying the relevance of our main technical results in the context of canonical factor models, illustrating the importance of each of our new Yurinskii coupling features (i)–(iv).

Section 3 considers a substantive application of our main results: strong approximation of martingale empirical processes. We begin with the motivating example of canonical kernel density estimation, demonstrating how Yurinskii's coupling can be applied, and showing in Lemma 3.1 why it is essential that we do not place any conditions on the minimum eigenvalue of the variance matrix (iii). We then present a general-purpose strong approximation for martingale empirical processes in Proposition 3.1, combining classical results in the empirical process literature (van der Vaart and Wellner, 1996) with our Corollary 2.2. This statement appears to be the first of its kind for martingale data, and when specialized to independent (and not necessarily identically distributed) data, it is shown to be superior to the best known comparable strong approximation result available in the literature (Berthet and Mason, 2006). Our improvement comes from using Yurinskii's coupling for the ℓ^∞-norm, where Berthet and Mason (2006) apply Zaitsev's coupling (Zaitsev, 1987a,b) with the larger ℓ^2-norm.

Section 4 further illustrates the applicability of our results through two examples in nonparametric regression estimation. Firstly, we deduce a strong approximation for partitioning-based least squares series estimators with time series data, applying Corollary 2.2 directly and additionally imposing only a mild mixing condition on the regressors. We show that our Yurinskii coupling for martingale vectors delivers the same distributional approximation rate as the best known result for independent data, and discuss how this can be leveraged to yield a feasible statistical inference procedure. We also show that if the residuals have vanishing conditional third moment, an improved rate of Gaussian approximation can be established. Secondly, we deduce a strong approximation for local polynomial estimators with time series data, using our result on martingale empirical processes (Proposition 3.1) and again imposing a mixing assumption. Appealing to empirical process theory is essential here as, in contrast with series estimators, local polynomials do not possess certain additive separability properties. The bandwidth restrictions we require are relatively mild, and, as far as we know, they have not been improved upon even with independent data.

Section 5 concludes the paper. All proofs are collected in Appendix A, which also includes other technical lemmas of potential independent interest. In Appendix B we present some further results on applications of our theory to deriving high-dimensional central limit theorems for martingales.

1.1 Notation

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

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

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

2 Main results

We begin with our most general result: an ℓ^p-norm Yurinskii coupling of a sum of vector-valued approximate martingale differences to a Gaussian mixture-distributed random vector. The general result is presented in Theorem 2.1, while Proposition 2.1 gives a simplified and slightly weaker version which is easier to use in applications. We then further specialize Proposition 2.1 to three scenarios with successively stronger assumptions, namely mixingales, martingales, and independent data in Corollaries 2.1, 2.2 and 2.3 respectively. In each case we allow for possibly random quadratic variations (cf. mixing convergence), thereby establishing a Gaussian mixture coupling in the general setting. In Remark 2.1 we comment on and correct an often overlooked technicality relating to the derivation of bounds in probability from Yurinskii's coupling. As a first illustration of the power of our generalized ℓ^p-norm Yurinskii coupling, we present in Section 2.5 a simple factor model example relating to all three of the aforementioned scenarios, discussing further how our contributions are related to the existing literature.

Theorem 2.1 (Strong approximation for vector-valued approximate martingales)
Take a complete probability space with a countably generated filtration $\mathcal{H}_0, \ldots, \mathcal{H}_n$ for some $n \geq 1$, supporting the $\mathbb{R}^d$-valued square-integrable random vectors $X_1, \ldots, X_n$. Let $S = \sum_{i=1}^n X_i$ and define

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

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

$$\begin{aligned} \mathbb{P}\big(\|S - T\|_p > 6\eta\big) \leq \inf_{t>0} &\left\{ 2\mathbb{P}\big(\|Z\|_p > t\big) + \min\left\{ \frac{\beta_{p,2} t^2}{\eta^3}, \frac{\beta_{p,3} t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3} \right\} \right\} \\ &+ \inf_{M \succeq 0} \Big\{ 2\mathbb{P}\big(\Omega \npreceq M\big) + \delta_p(M, \eta) + \varepsilon_p(M, \eta) \Big\} + \mathbb{P}\big(\|U\|_p > \eta\big), \end{aligned} \tag{1}$$

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

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

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

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

This theorem offers four novel contributions to the literature on coupling theory and strong approximation, as discussed in the introduction. Firstly (i), it allows for approximate vector-valued martingales, with the variables $\tilde{X}_i$ forming martingale differences with respect to $\mathcal{H}_i$ by construction, and U quantifying the associated martingale approximation error. Such martingale approximation techniques for sequences of dependent random vectors are well established and have been used in a range of scenarios: see, for example, Wu and Woodroofe (2004), Dedecker et al. (2007), Zhao and Woodroofe (2008), Peligrad (2010), Atchadé and Cattaneo (2014), Cuny and Merlevède (2014), Magda and Zhang (2018), and references therein. In Section 2.2 we demonstrate how this approximation can be established in practice by restricting our general theorem to the special case of mixingales, while the upcoming example in Section 2.5 provides an illustration in the context of auto-regressive factor models.

Secondly (ii), Theorem 2.1 allows for the resulting coupling variable T to follow a multivariate Gaussian distribution only conditionally, and thus we offer a useful analog of mixing convergence in the context of strong approximation. To be more precise, the random matrix $\sum_{i=1}^n V_i$ is the quadratic variation of the constructed martingale $\sum_{i=1}^n \tilde{X}_i$, and we approximate it using the $\mathcal{H}_0$-measurable random matrix Σ. This yields the coupling variable $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$, which can alternatively be written as $T = \Sigma^{1/2} Z$ with $Z \sim \mathcal{N}(0, I_d)$ independent of $\mathcal{H}_0$. The errors in this quadratic variation approximation are accounted for by the terms $\mathbb{P}(\Omega \npreceq M)$, $\delta_p(M, \eta)$ and $\varepsilon_p(M, \eta)$, utilizing a regularization argument through the free matrix parameter M. If a non-random Σ is used, then T is unconditionally Gaussian, and one can take $\mathcal{H}_0$ to be the trivial σ-algebra. As demonstrated in our proof, our approach to establishing a mixing approximation is different from naively taking an unconditional version of Yurinskii's coupling and applying it conditionally on $\mathcal{H}_0$, which will not deliver the same coupling as in Theorem 2.1 for a few reasons. To begin with, we explicitly indicate in the conditions of Theorem 2.1 where conditioning is required. Next, our error of approximation is given unconditionally, involving only marginal expectations and probabilities. Finally, we provide a rigorous account of the construction of the conditionally Gaussian coupling variable T via a conditional version of Strassen's theorem (Chen and Kato, 2020). Section 2.3 illustrates how a strong approximation akin to mixing convergence can arise when the data forms an exact martingale, and Section 2.5 gives a simple example relating to factor modeling in statistics and data science.

As a third contribution to the literature (iii), and of particular importance for applications, Theorem 2.1 makes no requirements on the minimum eigenvalue of the quadratic variation of the approximating martingale sequence. Instead, our proof technique employs a careful regularization scheme designed to account for any such exact or approximate rank degeneracy in Σ. This capability is fundamental in some applications, a fact which we illustrate in Section 3.1 by demonstrating the significant improvements in strong approximation errors delivered by Theorem 2.1 relative to those obtained using prior results in the literature.

Finally (iv), Theorem 2.1 gives a third-order strong approximation alongside the usual second-order version considered in all prior literature. More precisely, we observe that an analog of the term $\beta_{p,2}$ is present in the classical Yurinskii coupling and comes from a Lindeberg telescoping sum argument, replacing random variables by Gaussians with the same mean and variance to match the first and second moments. Whenever the third moments of $\tilde{X}_i$ are negligible (quantified by π_3), this moment-matching argument can be extended to third-order terms, giving a new term $\beta_{p,3}$. In certain settings, such as when the data is symmetrically distributed around zero, using $\beta_{p,3}$ rather than $\beta_{p,2}$ can give smaller approximation errors in the coupling given in (1). Such a refinement can be viewed as a strong approximation counterpart to classical Edgeworth expansion methods. We illustrate this phenomenon in our upcoming applications to nonparametric inference (Section 4).

2.1 User-friendly formulation of main result

The result in Theorem 2.1 is given in a somewhat implicit manner, involving infima over the free parameters $t > 0$ and $M \succeq 0$, and it is not clear how to compute these in general. In the upcoming Proposition 2.1, we set $M = \nu^2 I_d$ and approximately optimize over $t > 0$ and $\nu > 0$, resulting in a simplified and slightly weaker version of our main general result. In specific applications, where there is additional knowledge of the quadratic variation structure, other choices of regularization schemes may be more appropriate. Nonetheless, the choice $M = \nu^2 I_d$ leads to arguably the principal result of our work, due to its simplicity and utility in statistical applications. For convenience, define the functions $\phi_p : \mathbb{N} \to \mathbb{R}$ for $p \in [0, \infty]$,

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

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

Proposition 2.1 (Simplified strong approximation for vector-valued approximate martingales) *Assume the setup and notation of Theorem 2.1. For each $\eta > 0$ and $p \in [1, \infty]$, there exists a random vector $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$ satisfying*

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

If further $\pi_3 = 0$ then

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

Proposition 2.1 makes clear the potential benefit of a third-order coupling when $\pi_3 = 0$, as in this case the bound features $\beta_{p,3}^{1/4}$ rather than $\beta_{p,2}^{1/3}$. If π_3 is small but non-zero, an analogous result can easily be derived by adjusting the optimal choices of t and ν, but we omit this for clarity of notation. In applications (see Section 4.1), this reduction of the exponent can provide a significant improvement in terms of the dependence of the bound on the sample size n, the dimension d, and other problem-specific quantities. When using our results for strong approximation, it is usual to set $p = \infty$ to bound the maximum discrepancy over the entries of a vector (to construct uniform confidence sets, for example). In this setting, we have that $\phi_\infty(d) = \sqrt{2 \log 2d}$ has a sub-Gaussian slow-growing dependence on the dimension. The remaining term depends on $\mathbb{E}[\|\Omega\|_2]$ and requires that the matrix Σ be a good approximation of $\sum_{i=1}^n V_i$, while remaining $\mathcal{H}_0$-measurable. In some applications (such as factor modeling; see Section 2.5), it can be shown that the quadratic variation $\sum_{i=1}^n V_i$ remains random and $\mathcal{H}_0$-measurable even in large samples, giving a natural choice for Σ.

In the next few sections, we continue to refine Proposition 2.1, presenting a sequence of results with increasingly strict assumptions on the dependence structure of the data X_i. These allow us to demonstrate the broad applicability of our main results, providing more explicit bounds in settings which are likely to be of special interest. In particular, we consider mixingales, martingales, and independent data, comparing our derived results with those in the existing literature.

2.2 Mixingales

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

$$\mathbb{E}\left[\|\mathbb{E}\left[X_i \mid \mathcal{H}_{i-r}\right]\|_p\right] \leq c_i \zeta_r, \tag{2}$$

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

$$\mathbb{E}\left[\big\|X_i - \mathbb{E}\big[X_i \mid \mathcal{H}_{i+r}\big]\big\|_p\right] \leq c_i \zeta_{r+1}. \tag{3}$$

These conditions are satisfied, for example, if X_i are integrable strongly α-mixing random variables (McLeish, 1975), or if X_i are generated by an auto-regressive or auto-regressive moving average process (see Section 2.5), among many other possibilities (Bradley, 2005). Then, in the notation of Theorem 2.1, we have by Markov's inequality that

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

Combining Proposition 2.1 with this martingale error bound yields the following result for mixingales.

Corollary 2.1 (Strong approximation for vector-valued mixingales)
Assume the setup and notation of Theorem 2.1, and suppose that the mixingale conditions (2) and (3) hold. For each $\eta > 0$ and $p \in [1, \infty]$ there exists a random vector $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$ satisfying

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

If further $\pi_3 = 0$ then

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

The closest antecedent to Corollary 2.1 is found in Li and Liao (2020, Theorem 4), who also considered Yurinskii's coupling for mixingales. Our result improves on this work in the following manner: it removes any requirements on the minimum eigenvalue of the quadratic variation of the mixingale sequence; it allows for general ℓ^p-norms with $p \in [1, \infty]$; it establishes a coupling to a multivariate Gaussian mixture distribution in general; and it permits third-order couplings (when $\pi_3 = 0$). These improvements have important practical implications as demonstrated in Sections 2.5 and 4, where significantly better coupling approximation errors are demonstrated for a variety of statistical applications. On the technical side, our result is rigorously established using a conditional version of Strassen's theorem (Chen and Kato, 2020), a carefully crafted regularization argument, and a third-order Lindeberg method (see Chatterjee, 2006, and references therein, for more discussion on the standard second-order Lindeberg method). Furthermore, as explained in Remark 2.1, we clarify a technical issue in Li and Liao (2020) surrounding the derivation of valid probability bounds for $\|S - T\|_p$.

Corollary 2.1 focused on mixingales for simplicity, but, as previously discussed, any method for constructing a martingale approximation $\tilde{X}_i$ and bounding the resulting error U could be used instead in Proposition 2.1 to derive a similar result.

2.3 Martingales

For our second refinement, suppose that X_i form martingale differences with respect to $\mathcal{H}_i$. In this case, $\mathbb{E}[X_i \mid \mathcal{H}_n] = X_i$ and $\mathbb{E}[X_i \mid \mathcal{H}_0] = 0$, so $U = 0$, and the martingale approximation error term vanishes. Applying Proposition 2.1 in this setting directly yields the following result.

Corollary 2.2 (Strong approximation for vector-valued martingales)
With the setup and notation of Theorem 2.1, suppose X_i is $\mathcal{H}_i$-measurable with $\mathbb{E}[X_i \mid \mathcal{H}_{i-1}] = 0$ for $1 \leq i \leq n$. Then, for each $\eta > 0$ and $p \in [1, \infty]$, there is a random vector $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$ with

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

If further $\pi_3 = 0$ then

$$\mathbb{P}\big(\|S - T\|_p > \eta\big) \leq 24 \left(\frac{\beta_{p,3} \phi_p(d)^3}{\eta^4} \right)^{1/4} + 17 \left(\frac{\mathbb{E}\left[\|\Omega\|_2\right] \phi_p(d)^2}{\eta^2} \right)^{1/3}. \tag{5}$$

The closest antecedents to Corollary 2.2 are Belloni and Oliveira (2018) and Li and Liao (2020), who also implicitly or explicitly considered Yurinskii's coupling for martingales. More specifically, Li and Liao (2020, Theorem 1) established an explicit ℓ^2-norm Yurinskii coupling for martingales under a strong assumption on the minimum eigenvalue of the martingale quadratic variation, while Belloni and Oliveira (2018, Theorem 2.1) established a central limit theorem for vector-valued martingale sequences employing the standard second-order Lindeberg method, implying that their proof could be adapted to deduce a Yurinskii coupling for martingales with the help of a conditional version of Strassen's theorem (Chen and Kato, 2020) and some additional nontrivial technical work.

Corollary 2.2 improves over this prior work as follows. With respect to Li and Liao (2020), our result establishes an ℓ^p-norm Gaussian mixture Yurinskii coupling for martingales without any requirements on the minimum eigenvalue of the martingale quadratic variation, and permits a third-order coupling if $\pi_3 = 0$. The first probability bound (4) in Corollary 2.2 gives the same rate of strong approximation as that in Theorem 1 of Li and Liao (2020) when $p = 2$, with non-random Σ, and when the eigenvalues of a normalized version of Σ are bounded away from zero. In Section 3.1 we demonstrate the crucial importance of removing this eigenvalue lower bound restriction in applications involving nonparametric kernel estimators, while in Section 4.1 we demonstrate how the availability of a third-order coupling (5) can give improved approximation rates in applications involving nonparametric series estimators with conditionally symmetrically distributed residual errors. Finally, our technical work improves on Li and Liao (2020) in two respects: (i) we employ a conditional version of Strassen's theorem (see Lemma A.1 in the appendix) to appropriately handle the conditioning arguments; and (ii) we deduce valid probability bounds for $\|S - T\|_p$, as the following Remark 2.1 makes clear.

Remark 2.1 (Yurinskii's coupling and bounds in probability)
Given a sequence of random vectors S_n, Yurinskii's method provides a coupling in the following form: for each n and any $\eta > 0$, there exists a random vector T_n with $\mathbb{P}\big(\|S_n - T_n\| > \eta\big) < r_n(\eta)$, where $r_n(\eta)$ is the approximation error. Crucially, each coupling variable T_n is a function of the desired approximation level η and, as such, deducing bounds in probability on $\|S_n - T_n\|$ requires some extra care. One option is to select a sequence $R_n \to \infty$ and note that $\mathbb{P}\big(\|S_n - T_n\| > r_n^{-1}(1/R_n)\big) < 1/R_n \to 0$ and hence $\|S_n - T_n\| \lesssim_{\mathbb{P}} r_n^{-1}(1/R_n)$. In this case, T_n depends on the choice of R_n, which can in turn typically be chosen to diverge slowly enough to cause no issues in applications.

Technicalities akin to those outlined in Remark 2.1 have been both addressed and neglected alike in the prior literature. Pollard (2002, Chapter 10.4, Example 16) apparently misses this subtlety, providing an inaccurate bound in probability based on the Yurinskii coupling. Li and Liao (2020) seem to make the same mistake in the proof of their Lemma A2, which invalidates the conclusion of their Theorem 1. In contrast, Belloni et al. (2015) and Belloni et al. (2019) directly provide bounds in $o_{\mathbb{P}}$ instead of $O_{\mathbb{P}}$, circumventing these issues in a manner similar to our approach involving a diverging sequence R_n.

To see how this phenomenon applies to our main results, observe that the second-order martingale coupling given as (4) in Corollary 2.2 implies that for any $R_n \to \infty$,

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

This bound is comparable to that obtained by Li and Liao (2020, Theorem 1) with $p = 2$, albeit with their formulation missing the R_n correction terms. In Section 4.1 we discuss further their (amended) result, in the setting of nonparametric series estimation. Our approach using $p = \infty$ obtains superior distributional approximation rates, alongside exhibiting various other improvements such as the aforementioned third-order coupling.

Turning to the comparison with Belloni and Oliveira (2018), our Corollary 2.2 again offers the same improvements, with the only exception being that the authors did account for the implications of a possibly vanishing minimum eigenvalue. However, their results exclusively concern high-dimensional central limit theorems for vector-valued martingales (see our Section B), and therefore while their findings could in principle enable the derivation of a result similar to our Corollary 2.2, this would require additional technical work on their behalf in multiple ways: (i) a correct application of a conditional version of Strassen's theorem (see Lemma A.1 in the appendix); (ii) the development of a third-order Borel set smoothing technique and associated ℓ^p-norm moment control (see Lemmas A.2, A.3, and A.4); (iii) a careful truncation scheme to account for $\Omega \not\succeq 0$; and (iv) a valid third-order Lindeberg argument (see Lemma A.8), among others.

2.4 Independence

As a final refinement, suppose that X_i are independent and zero-mean conditionally on $\mathcal{H}_0$, and take $\mathcal{H}_i$ to be the filtration generated by $X_1, \ldots, X_i$ and $\mathcal{H}_0$ for $1 \leq i \leq n$. Then, taking $\Sigma = \sum_{i=1}^n V_i$ gives $\Omega = 0$, and hence Corollary 2.2 immediately yields the following result.

Corollary 2.3 (Strong approximation for sums of independent vectors)
Assume the setup of Theorem 2.1, and suppose X_i are independent given $\mathcal{H}_0$, with $\mathbb{E}[X_i \mid \mathcal{H}_0] = 0$. Then, for each $\eta > 0$ and $p \in [1, \infty]$, with $\Sigma = \sum_{i=1}^n V_i$, there exists $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$ satisfying

$$\mathbb{P}\big(\|S - T\|_p > \eta\big) \leq 24 \left(\frac{\beta_{p,2} \phi_p(d)^2}{\eta^3} \right)^{1/3}. \tag{6}$$

If further $\pi_3 = 0$ then

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

Taking $\mathcal{H}_0$ to be trivial, the first inequality (6) in Corollary 2.3 provides an ℓ^p-norm approximation analogous to that presented in Belloni et al. (2019). By further restricting to $p = 2$, we recover the original Yurinskii coupling as presented in Le Cam (1988, Theorem 1) and Pollard (2002, Theorem 10). Thus, in the independent data setting, our result improves on prior work as follows: (i) it establishes a coupling to a multivariate Gaussian mixture distribution; and (ii) it permits a third-order coupling if $\pi_3 = 0$.

2.5 Stylized example: factor modeling

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

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

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

We employ the above factor model to give a first illustration of the applicability of our main result Theorem 2.1, the user-friendly Proposition 2.1, and their specialized Corollaries 2.1–2.3. We consider three different sets of conditions to demonstrate the applicability of each of our corollaries for mixingales, martingales, and independent data, respectively. We assume throughout that $(\varepsilon_1, \ldots, \varepsilon_n)$ is zero-mean and finite variance, and that $(\varepsilon_1, \ldots, \varepsilon_n)$ is independent of L and $(f_1, \ldots, f_n)$. Let $\mathcal{H}_i$ be the σ-algebra generated by L, $(f_1, \ldots, f_i)$ and $(\varepsilon_1, \ldots, \varepsilon_i)$, with $\mathcal{H}_0$ the σ-algebra generated by L alone.

- **Independent data**. Suppose that the factors $(f_1, \ldots, f_n)$ are independent conditional on L and satisfy $\mathbb{E}[f_i \mid L] = 0$. Then, since X_i are independent conditional on $\mathcal{H}_0$ and with $\mathbb{E}[X_i \mid \mathcal{H}_0] = \mathbb{E}[L f_i + \varepsilon_i \mid L] = 0$, we can apply Corollary 2.3 to $\sum_{i=1}^n X_i$. In general, we will obtain a coupling variable which has the Gaussian mixture distribution $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = \sum_{i=1}^n (L \operatorname{Var}[f_i \mid L] L^\mathsf{T} + \operatorname{Var}[\varepsilon_i])$. In the special case where L is non-random and $\mathcal{H}_0$ is trivial, the coupling is Gaussian. Furthermore, if $f_i \mid L$ and ε_i are symmetric about zero and bounded almost surely, then $\pi_3 = 0$, and the coupling is improved.

- **Martingales**. Suppose instead that we assume only a martingale condition on the latent factor variables so that $\mathbb{E}[f_i \mid L, f_1, \ldots, f_{i-1}] = 0$. Then $\mathbb{E}[X_i \mid \mathcal{H}_{i-1}] = L\,\mathbb{E}[f_i \mid \mathcal{H}_{i-1}] = 0$ and Corollary 2.2 is applicable to $\sum_{i=1}^n X_i$. The preceding comments on Gaussian mixture distributions and third-order couplings continue to apply.

- **Mixingales**. Finally, assume that the factors follow the auto-regressive model $f_i = Af_{i-1} + u_i$ where $A \in \mathbb{R}^{m\times m}$ is non-random and $(u_1, \ldots, u_n)$ are zero-mean, independent, and independent of $(\varepsilon_1, \ldots, \varepsilon_n)$. Then $\mathbb{E}[f_i \mid f_0] = A^i f_0$, so taking $p \in [1, \infty]$ we see that $\mathbb{E}\big[\|\mathbb{E}[f_i \mid f_0]\|_p\big] = \mathbb{E}\big[\|A^i f_0\|_p\big] \leq \|A\|_p^i\,\mathbb{E}[\|f_0\|_p]$, and that clearly $f_i - \mathbb{E}[f_i \mid \mathcal{H}_n] = 0$. Thus, whenever $\|A\|_p < 1$, the geometric sum formula implies that we can apply the mixingale result from Corollary 2.1 to $\sum_{i=1}^n X_i$. The conclusions on Gaussian mixture distributions and third-order couplings parallel the previous cases.

This simple application to factor modeling gives a preliminary illustration of the power of our main results, encompassing settings which could not be handled by employing Yurinskii couplings available in the existing literature. Even with independent data, we offer new Yurinskii couplings to Gaussian mixture distributions (due to the presence of the common random factor loading L), which could be further improved whenever the factors and residuals possess symmetric (conditional) distributions. Furthermore, our results do not impose any restrictions on the minimum eigenvalue of Σ, thereby allowing for more general factor structures. These improvements are maintained in the martingale, mixingale, and weakly dependent stationary data settings.

3 Strong approximation for martingale empirical processes

In this section, we demonstrate how our main results can be applied to some more substantive problems in statistics. Having until this point studied only finite-dimensional (albeit potentially high-dimensional) random vectors, we now turn our attention to infinite-dimensional stochastic processes. Specifically, we consider empirical processes of the form

$$S(f) = \sum_{i=1}^n f(X_i), \qquad f \in \mathcal{F},$$

with $\mathcal{F}$ a problem-specific class of real-valued functions, where each $f(X_i)$ forms a martingale difference sequence with respect to an appropriate filtration. We construct (conditionally) Gaussian processes $T(f)$ for which an upper bound on the uniform coupling error $\sup_{f\in\mathcal{F}} |S(f) - T(f)|$ is precisely quantified. We control the complexity of $\mathcal{F}$ using metric entropy under Orlicz norms.

The novel strong approximation results which we present concern the entire martingale empirical process $(S(f) : f \in \mathcal{F})$, as opposed to just the scalar supremum of the empirical process, $\sup_{f\in\mathcal{F}} |S(f)|$. This distinction has been carefully noted by Chernozhukov et al. (2014b), who studied Gaussian approximation of empirical process suprema in the independent data setting and wrote (p. 1565): "A related but different problem is that of approximating *whole* empirical processes by a sequence of Gaussian processes in the sup-norm. This problem is more difficult than [approximating the supremum of the empirical process]." Indeed, the results we establish in this section are for a strong approximation for the entire empirical process by a sequence of Gaussian mixture processes in the supremum norm, when the data has a martingale difference structure (cf. Corollary 2.2). Our results can be further generalized to approximate martingale empirical processes (cf. Corollary 2.1), but we do not consider this extension to reduce notation and the technical burden.

3.1 Motivating example: kernel density estimation

We begin with a brief study of a canonical example of an empirical process which is non-Donsker (thus precluding the use of uniform central limit theorems) due to the presence of a function class whose complexity increases with the sample size: the kernel density estimator with i.i.d. scalar data. We give an overview of our general strategy for strong approximation of stochastic processes via discretization, and show explicitly in Lemma 3.1 how it is crucial that we do not impose lower bounds on the eigenvalues of the discretized covariance matrix. Detailed calculations for this section are relegated to Appendix A for conciseness.

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

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

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

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

Writing $S(\mathcal{X}_\delta)$ for $\big(S(x) : x \in \mathcal{X}_\delta\big) \in \mathbb{R}^N$, and noting that this is a sum of i.i.d. random vectors, we apply Corollary 2.3 as $\max_{x \in \mathcal{X}_\delta} |S(x) - T(x)| = \|S(\mathcal{X}_\delta) - T(\mathcal{X}_\delta)\|_\infty$. We thus obtain that, for each $\eta > 0$, there exists a Gaussian vector $T(\mathcal{X}_\delta)$ with the same covariance matrix as $S(\mathcal{X}_\delta)$ satisfying

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

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

Next, it is not difficult to show by chaining with the Bernstein–Orlicz and sub-Gaussian norms respectively (van der Vaart and Wellner, 1996, Section 2.2) that if $\log(N/h) \lesssim \log n$ and $nh \gtrsim \log n$,

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

Finally, for any sequence $R_n \to \infty$ (see Remark 2.1), the resulting bound on the coupling error is

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

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

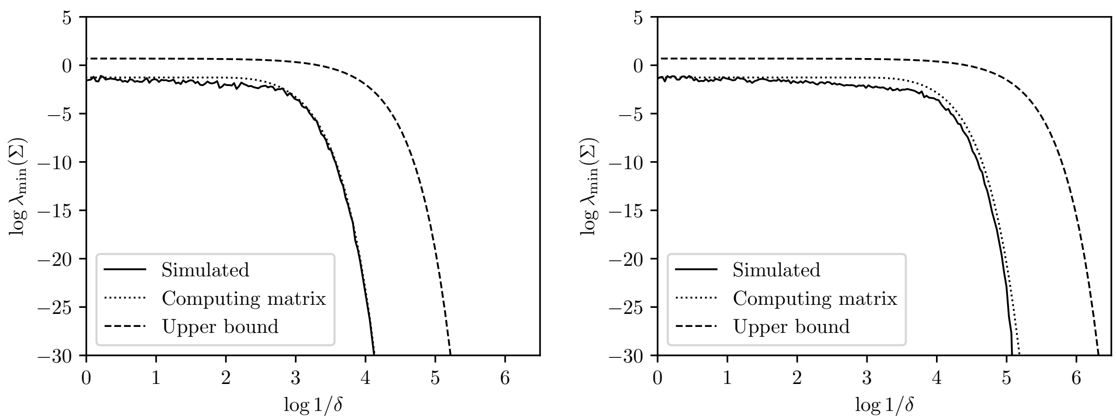
The discretization strategy outlined above is at the core of the proof strategy for our upcoming Proposition 3.1. Since we will consider martingale empirical processes, our proof will rely on

Corollary 2.2, which, unlike the martingale Yurinskii coupling established by Li and Liao (2020), does not require a lower bound on the minimum eigenvalue of Σ. Using the simple kernel density example just discussed, we now demonstrate precisely the crucial importance of removing such eigenvalue conditions. The following Lemma 3.1 shows that the discretized covariance matrix $\Sigma = nh\,\mathrm{Var}[S(\mathcal{X}_\delta)]$ has exponentially small eigenvalues, which in turn will negatively affect the strong approximation bound if the Li and Liao (2020) coupling were to be used instead of the results in this paper.

Lemma 3.1 (Minimum eigenvalue of a kernel density estimator covariance matrix)
The minimum eigenvalue of $\Sigma = nh\,\mathrm{Var}[S(\mathcal{X}_\delta)] \in \mathbb{R}^{N\times N}$ satisfies the upper bound

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

Figure 1 shows how the upper bound in Lemma 3.1 captures the behavior of the simulated minimum eigenvalue of Σ. In particular, the smallest eigenvalue decays exponentially fast in the discretization level δ and the bandwidth h. As seen in the calculations above, the coupling rate depends on δ/h, while the bias will generally depend on h, implying that both δ and h must converge to zero to ensure valid statistical inference. In general, this will lead to Σ possessing extremely small eigenvalues, rendering strong approximation approaches such as that of Li and Liao (2020) ineffective in such scenarios.



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

Figure 1: Upper bounds on the minimum eigenvalue of the discretized covariance matrix in kernel density estimation, with $n = 100$ and $a = 0.2$. Simulated: the kernel density estimator is simulated, resampling the data 100 times to estimate its covariance. Computing matrix: the minimum eigenvalue of the limiting covariance matrix Σ is computed explicitly. Upper bound: the bound derived in Lemma 3.1 is shown.

The discussion in this section focuses on the strong approximation of the centered process $\hat{g}(x) - \mathbb{E}[\hat{g}(x)]$. In practice, the goal is often rather to approximate the feasible process $\hat{g}(x) - g(x)$. The difference between these is captured by the smoothing bias $\mathbb{E}[\hat{g}(x)] - g(x)$, which is straightforward to control with $\sup_{x\in\mathcal{X}} \left|\mathbb{E}[\hat{g}(x)] - g(x)\right| \lesssim \frac{h}{a} e^{-a^2/(2h^2)}$. See Section 4 for further discussion.

3.2 General result for martingale empirical processes

We now give our general result on a strong approximation for martingale empirical processes, obtained by applying the first result (4) in Corollary 2.2 with $p = \infty$ to a discretization of the empirical process, as in Section 3.1. We then control the increments in the stochastic processes using chaining with Orlicz norms, but note that other tools are available, including generalized entropy with bracketing (van de Geer, 2000) and sequential symmetrization (Rakhlin 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 $\mathcal{F}$, write $\mathcal{F}(x) = \big(f(x) : f \in \mathcal{F}\big)$. Define the set of Orlicz functions

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

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

Proposition 3.1 (Strong approximation for martingale empirical processes)
Let X_i be random variables for $1 \leq i \leq n$ taking values in a measurable space $\mathcal{X}$, and $\mathcal{F}$ be a pointwise measurable class of functions from $\mathcal{X}$ to $\mathbb{R}$. Let $\mathcal{H}_0, \ldots, \mathcal{H}_n$ be a filtration such that each X_i is $\mathcal{H}_i$-measurable, with $\mathcal{H}_0$ the trivial σ-algebra, and suppose that $\mathbb{E}[f(X_i) \mid \mathcal{H}_{i-1}] = 0$ for all $f \in \mathcal{F}$. Define $S(f) = \sum_{i=1}^n f(X_i)$ for $f \in \mathcal{F}$ and let $\Sigma : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ be an almost surely positive semi-definite $\mathcal{H}_0$-measurable random function. Suppose that for a non-random metric d on $\mathcal{F}$, constant L and $\psi \in \Psi$,

$$\Sigma(f, f) - 2\Sigma(f, f') + \Sigma(f', f') + \lVert\!| S(f) - S(f') |\!\rVert_\psi^2 \leq L^2 d(f, f')^2 \quad a.s. \tag{7}$$

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

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

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

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

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

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

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

Proposition 3.1 is given in a rather general form to accommodate a range of different settings and applications. In particular, consider the following well known Orlicz functions.

Polynomial: $\psi(x) = x^a$ for $a \geq 2$ has $\lVert\!| X |\!\rVert_2 \leq \lVert\!| X |\!\rVert_\psi$ and $\sqrt{\log x} \leq \sqrt{a}\psi^{-1}(x)$.

Exponential: $\psi(x) = \exp(x^a) - 1$ for $a \in [1, 2]$ has $\lVert\!| X |\!\rVert_2 \leq 2\lVert\!| X |\!\rVert_\psi$ and $\sqrt{\log x} \leq \psi^{-1}(x)$.

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

For these Orlicz functions and when $\Sigma(f, f') = \mathrm{Cov}[S(f), S(f')]$ is non-random, the terms involving Σ in (7) can be controlled by the Orlicz ψ-norm term; similarly, J_2 is bounded by J_ψ. Further, C_ψ can be replaced by a universal constant C which does not depend on the parameter a. See Section 2.2 in van der Vaart and Wellner (1996) for details. If the conditional third moments of $f(X_i)$ given $\mathcal{H}_{i-1}$ are all zero (if f and X_i are appropriately symmetric, for example), then the second inequality in Corollary 2.2 can be applied to obtain a tighter coupling inequality; the details of this are omitted for brevity, and the proof would proceed in exactly the same manner.

In general, however, Proposition 3.1 allows for a random covariance function, yielding a coupling to a stochastic process that is Gaussian only conditionally. Such a process can equivalently be viewed as a mixture of Gaussian processes, writing $T = \Sigma^{1/2} Z$ with an operator square root and where Z is a Gaussian white noise on $\mathcal{F}$ independent of $\mathcal{H}_0$. This extension is in contrast with much of the existing strong approximation and empirical process literature, which tends to focus on couplings and weak convergence results with marginally Gaussian processes.

A similar approach was taken by Berthet and Mason (2006), who used a Gaussian coupling due to Zaitsev (1987a,b) along with a discretization method to obtain strong approximations for empirical processes with independent data. They handled fluctuations in the stochastic processes with uniform L^2 covering numbers and bracketing numbers where we opt instead for chaining with Orlicz norms. Our version using the (martingale) Yurinskii coupling can improve upon theirs in approximation rate even for independent data 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 $d_{\mathbb{Q}}(f, f')^2 = \mathbb{E}_{\mathbb{Q}}\bigl[(f - f')^2\bigr]$ 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 take $\Sigma(f, f') = \mathrm{Cov}[S(f), S(f')]$ so $\Omega_\delta = 0$. Let $\mathcal{F}_\delta$ be a minimal δ-cover of $(\mathcal{F}, d_{\mathbb{P}})$ with cardinality $N_\delta \lesssim \delta^{-\nu_0}$ where $\delta \to 0$. It is not difficult to show that $\beta_\delta \lesssim n\delta^{-\nu_0}\sqrt{\log(1/\delta)}$. Theorem 2.2.8 and Theorem 2.14.1 in van der Vaart and Wellner (1996) give

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

where we used the VC-type property to bound the entropy integral. So by our Proposition 3.1, for any sequence $R_n \to \infty$ (see Remark 2.1),

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

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

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

showing that our approach achieves a better approximation rate whenever $\nu_0 > 4/3$. In particular, our method is superior in richer function classes with larger VC-type dimension. For example, if $\mathcal{F}$ is smoothly 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, Equation 5.3) are compelled to use the inequality $\|\cdot\|_\infty \leq \|\cdot\|_2$ in the coupling step, a bound which is loose when the dimension of the vectors (here on the order of $\delta^{-\nu_0}$) is even moderately large. We use the fact that our version of Yurinskii's coupling applies directly to the supremum norm, giving sharper dependence on the dimension.

In Section 4.2 we apply Proposition 3.1 to obtain strong approximations for local polynomial estimators in the nonparametric regression setting. In contrast with the series estimators of the upcoming Section 4.1, local polynomial estimators are not linearly separable and hence cannot be analyzed directly using the finite-dimensional Corollary 2.2.

4 Applications to nonparametric regression

We illustrate the applicability of our previous strong approximation results with two substantial and classical examples in nonparametric regression estimation. Firstly, we present an analysis of partitioning-based series estimators, in which we can apply Corollary 2.2 directly due to an intrinsic linear separability property. Secondly, we consider local polynomial estimators, this time using Proposition 3.1 due to the presence of a non-linearly separable martingale empirical process.

4.1 Partitioning-based series estimators

Partitioning-based least squares methods are essential tools for estimation and inference in nonparametric regression, encompassing splines, piecewise polynomials, compactly supported wavelets and decision trees as special cases. See Cattaneo et al. (2020) for further details and references throughout this section. We illustrate the usefulness of Corollary 2.2 by deriving a Gaussian strong approximation for partitioning series estimators based on multivariate martingale data. Proposition 4.1 shows how we achieve the best known rate of strong approximation for independent data by imposing an additional mild α-mixing condition to control the time series dependence of the regressors.

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

Our goal is to approximate the law of the stochastic process $(\hat{\mu}(w) - \mu(w) : w \in \mathcal{W})$, which upon rescaling is typically not asymptotically tight as $k \to \infty$ and thus does not converge weakly. Nevertheless, exploiting the intrinsic linearity of the estimator $\hat{\mu}(w)$, we can apply Corollary 2.2 directly to construct a Gaussian strong approximation. Specifically, we write

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

where $H = \sum_{i=1}^n \mathbb{E}\left[p(W_i) p(W_i)^\mathsf{T}\right]$ is the expected outer product matrix, $S = \sum_{i=1}^n p(W_i)\varepsilon_i$ is the score vector, and $\mathrm{Bias}(w) = p(w)^\mathsf{T} \hat{H}^{-1} \sum_{i=1}^n p(W_i)\mu(W_i) - \mu(w)$. Imposing some mild time series restrictions and assuming stationarity for simplicity, it is not difficult to show (see Appendix A) that $\|\hat{H} - H\|_1 \lesssim_{\mathbb{P}} \sqrt{nk}$ 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 μ. Thus, it remains to study the k-dimensional mean-zero martingale S by applying Corollary 2.2 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 (Bradley, 2005).

Proposition 4.1 (Strong approximation for partitioning series estimators)
Consider the nonparametric regression setup described above and further assume the following:

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

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

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

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

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

Then, for any sequence $R_n \to \infty$, *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*

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

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

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

The core of the proof of Proposition 4.1 involves applying Corollary 2.2 with $S = \sum_{i=1}^n p(W_i)\varepsilon_i$ and $p = \infty$ to construct $T \sim \mathcal{N}\big(0, \mathrm{Var}[S]\big)$ such that $\|S - T\|_\infty$ is small, and then setting $G(w) = p(w)^{\mathsf{T}} H^{-1} T$. So long as the bias can be appropriately controlled, this result allows for uniform inference procedures such as uniform confidence bands or shape specification testing. The condition $k^3/n \to 0$ is the same (up to logs) as that imposed by Cattaneo et al. (2020) for i.i.d. data, which gives the best known strong approximation rate for this problem. Thus, Proposition 4.1 gives the same best approximation rate without requiring any extra restrictions for α-mixing time series data.

Our results improve substantially on Li and Liao (2020, Theorem 1): using the notation of our Corollary 2.2, and with any sequence $R_n \to \infty$, a valid (see Remark 2.1) version of their martingale Yurinskii coupling is

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

where $B_n = \sum_{i=1}^n \mathbb{E}[\|X_i\|_2^3]$ and r_n is a term controlling the convergence of the quadratic variation, playing a similar role to our term $\mathbb{E}[\|\Omega\|_2]$. Under the assumptions of our Proposition 4.1, applying this result with $S = \sum_{i=1}^n p(W_i)\varepsilon_i$ yields a rate no better than $\|S - T\|_2 \lesssim_{\mathbb{P}} (nk)^{1/3} R_n$. As such, they attain a rate of strong approximation no faster than

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

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-norm version of Yurinskii's coupling rather than the more recently established ℓ^∞-norm version. Further, our approach allows for an improved rate of distributional approximation whenever the residuals have zero conditional third moment.

To illustrate the statistical applicability of Proposition 4.1, consider constructing a feasible uniform confidence band for the regression function μ, using standardization and Studentization for statistical power improvements. We assume throughout that the bias is negligible. Proposition 4.1 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

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

where $\rho(w,w') = \mathbb{E}[G(w)G(w')]$. 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 4.1) that with $\mathbf{W} = (W_1, \ldots, W_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$,

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

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

Proposition 4.2 (Feasible uniform confidence bands for partitioning series estimators)
Assume the setup of the preceding section. Then

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

where

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

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

4.2 Local polynomial estimators

As a second example application we consider nonparametric regression estimation with martingale data employing local polynomial methods (Fan and Gijbels, 1996). In contrast with the partitioning-based series methods of Section 4.1, local polynomials induce stochastic processes which are not linearly separable, allowing us to showcase the empirical process result given in Proposition 3.1.

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

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

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

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

where $\hat{H}(w) = \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) p_h(W_i - w)^\mathsf{T}$, $H(w) = \mathbb{E}\big[\hat{H}(w)\big]$, $S(w) = \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) \varepsilon_i$ and $\mathrm{Bias}(w) = e_1^\mathsf{T} \hat{H}(w)^{-1} \sum_{i=1}^n K_h(W_i - w) p_h(W_i - w) \mu(W_i) - \mu(w)$. A key distinctive feature of local polynomial regression is that both $\hat{H}(w)$ and $S(w)$ are functions of the evaluation point $w \in \mathcal{W}$; contrast this with the partitioning-based series estimator discussed in Section 4.1, for which neither $\hat{H}$ nor S depend on w. Therefore we use Proposition 3.1 to obtain a Gaussian strong approximation for the martingale empirical process directly.

Under some mild regularity conditions, including stationarity for simplicity and an α-mixing assumption on the time-dependence of the data, we first show $\sup_{w \in \mathcal{W}} \|\hat{H}(w) - H(w)\|_2 \lesssim_\mathbb{P} \sqrt{n h^{-2m} \log n}$. 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 3.1 by setting

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

With this approach, we obtain the following result.

Proposition 4.3 (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 for any sequence $R_n \to \infty$, 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*

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

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

If the residuals further satisfy $\mathbb{E}\left[\varepsilon_i^3 \mid \mathcal{H}_{i-1}\right] = 0$, then a third-order Yurinskii coupling delivers an improved rate of strong approximation for Proposition 4.3; this is omitted here for brevity. For completeness, the proof of Proposition 4.3 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 4.1 for partitioning-based series estimators, and as far as we know, has not been improved upon for non-i.i.d. data.

Applying an anti-concentration result for Gaussian process suprema, such as Corollary 2.1 in Chernozhukov et al. (2014a), allows one to write a Kolmogorov–Smirnov bound comparing the law of $\sup_{w \in \mathcal{W}} |\hat{\mu}(w) - \mu(w)|$ to that of $\sup_{w \in \mathcal{W}} |T(w)|$. With an appropriate covariance estimator, we can further replace $T(w)$ by a feasible version $\hat{T}(w)$ or its Studentized counterpart, enabling procedures for uniform inference analogous to the confidence bands constructed in Section 4.1. We omit the details of this to conserve space but note that our assumptions on W_i and ε_i ensure that Studentization is possible even when the discretized covariance matrix has small eigenvalues (Section 3.1), as we normalize only by the diagonal entries. Chernozhukov et al. (2014b, Remark 3.1) achieve better rates for approximating the supremum of the t-process based on i.i.d. data in Kolmogorov–Smirnov distance by bypassing the step where we first approximate the entire stochastic process (see Section 3 for a discussion).

We finally remark that in this setting of kernel-based local empirical processes, it is essential that our initial strong approximation result (Corollary 2.2) does not impose a lower bound on the eigenvalues of the variance matrix Σ. This effect was demonstrated by Lemma 3.1 and its surrounding discussion in Section 3.1, and as such, the result of Li and Liao (2020) is unsuited for this application due to its strong minimum eigenvalue assumption.

5 Conclusion

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

Acknowledgments

We thank Jianqing Fan, Alexander Giessing, Michael Jansson, Jason Klusowski, and Boris Shigida for comments.

Funding

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).

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.

Atchadé, Y. F. and Cattaneo, M. D. (2014). A martingale decomposition for quadratic forms of Markov chains (with applications). *Stochastic Processes and their Applications*, 124(1):646–677.

Baxter, B. J. C. (1994). Norm estimates for inverses of Toeplitz distance matrices. *Journal of Approximation Theory*, 79(2):222–242.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Fernández-Val, I. (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, 213(1):4–29.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Belloni, A. and Oliveira, R. I. (2018). A high dimensional central limit theorem for martingales, with applications to context tree models. *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.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144.

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. *Annals of Statistics*, 48(3):1718–1741.

Cattaneo, M. D., Feng, Y., and Underwood, W. G. (2024). Uniform inference for kernel density estimators with dyadic data. *Journal of the American Statistical Association,* forthcoming.

Chatterjee, S. (2006). A generalization of the Lindeberg principle. *Annals of Probability*, 34(6):2061–2076.

Chen, X. and Kato, K. (2020). Jackknife multiplier bootstrap: finite sample approximations to the U-process supremum with applications. *Probability Theory and Related Fields*, 176(3):1097–1163.

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

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

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

Chernozhukov, V., Chetverikov, D., and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309 – 2352.

Chernozhukov, V., Chetverikov, D., and Koike, Y. (2023). Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *Annals of Applied Probability*, 33(3):2374–2425.

Cuny, C. and Merlevède, F. (2014). On martingale approximations and the quenched weak invariance principle. *The Annals of Probability*, 42(2):760–793.

Dedecker, J., Merlevède, F., and Volnỳ, D. (2007). On the weak invariance principle for non-adapted sequences under projective criteria. *Journal of Theoretical Probability*, 20:971–1004.

Dehling, H. (1983). Limit theorems for sums of weakly dependent Banach space valued random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 63(3):393–432.

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.

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020). *Statistical Foundations of Data Science*. CRC press.

Giessing, A. (2023). Anti-concentration of suprema of Gaussian processes and Gaussian order statistics. *arXiv preprint arXiv:2310.12119*.

Koike, Y. (2021). Notes on the dimension dependence in high-dimensional central limit theorems for hyperrectangles. *Japanese Journal of Statistics and Data Science*, 4:257–297.

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, Berkeley.

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

Lopes, M. E. (2022). Central limit theorem and bootstrap approximation in high dimensions: Near $1/n$ rates via implicit smoothing. *The Annals of Statistics*, 50(5):2492–2513.

Lopes, M. E., Lin, Z., and Müller, H.-G. (2020). Bootstrapping max statistics in high dimensions: Near-parametric rates under weak variance decay and application to functional and multinomial data. *Annals of Statistics*, 48(2):1214–1229.

Magda, P. and Zhang, N. (2018). Martingale approximations for random fields. *Electronic Communications in Probability*, 23(28):1–9.

McLeish, D. L. (1975). Invariance principles for dependent variables. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32(3):165–178.

Merlevède, F., Peligrad, M., and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, 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.

Peligrad, M. (2010). Conditional central limit theorem via martingale approximation. In *Dependence in Probability, Analysis and Number Theory, volume in memory of Walter Philipp*, pages 295–311. Kendrick Press I. Berkes, RC Bradley, H. Dehling, M. Peligrad, R. Tichy Editors.

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

Sheehy, A. and Wellner, J. A. (1992). Uniform Donsker classes of functions. *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.

Wu, W. B. and Woodroofe, M. (2004). Martingale approximations for sums of stationary processes. *Annals of Probability*, 32(2):1674–1690.

Yurinskii, V. V. (1978). On the error of the Gaussian approximation for convolutions. *Theory of Probability & its Applications*, 22(2):236–247.

Zaitsev, A. Y. (1987a). Estimates of the Lévy–Prokhorov distance in the multivariate central limit theorem for random variables with finite exponential moments. *Theory of Probability & Its Applications*, 31(2):203–220.

Zaitsev, A. Y. (1987b). On the Gaussian approximation of convolutions under multidimensional analogues of S. N. Bernstein's inequality conditions. *Probability Theory and Related Fields*, 74(4):535–566.

Zhai, A. (2018). A high-dimensional CLT in $\mathcal{W}_2$ distance with near optimal convergence rate. *Probability Theory and Related Fields*, 170(3):821–845.

Zhao, O. and Woodroofe, M. (2008). On martingale approximations. *Annals of Applied Probability*, 18(5):1831–1847.

A Proofs of main results

A.1 Preliminary lemmas

We give a sequence of preliminary lemmas which are useful for establishing our main results. Firstly, we present a conditional version of Strassen's theorem for the ℓ^p-norm (Chen and Kato, 2020, Theorem B.2), stated for completeness as Lemma A.1.

Lemma A.1 (A conditional Strassen theorem for the ℓ^p-norm)
Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space supporting the $\mathbb{R}^d$-valued random variable X for some $d \geq 1$. Let $\mathcal{H}'$ be a countably generated sub-σ-algebra of $\mathcal{H}$ and suppose there exists a $\mathrm{Unif}[0,1]$ *random variable on $(\Omega, \mathcal{H}, \mathbb{P})$ which is independent of the σ-algebra generated by X and $\mathcal{H}'$. Consider a regular conditional distribution $F(\cdot \mid \mathcal{H}')$ satisfying the following. Firstly, $F(A \mid \mathcal{H}')$ is an $\mathcal{H}'$-measurable random variable for all Borel sets $A \in \mathcal{B}(\mathbb{R}^d)$. Secondly, $F(\cdot \mid \mathcal{H}')(\omega)$ is a Borel probability measure on $\mathbb{R}^d$ for all $\omega \in \Omega$. Taking $\eta, \rho > 0$ and $p \in [1, \infty]$, with $\mathbb{E}^*$ the outer expectation, if*

$$\mathbb{E}^*\left[\sup_{A \in \mathcal{B}(\mathbb{R}^d)} \left\{\mathbb{P}\big(X \in A \mid \mathcal{H}'\big) - F\big(A_p^\eta \mid \mathcal{H}'\big)\right\}\right] \leq \rho,$$

where $A_p^\eta = \{x \in \mathbb{R}^d : \|x - A\|_p \leq \eta\}$ and $\|x - A\|_p = \inf_{x' \in A} \|x - x'\|_p$, then there exists an $\mathbb{R}^d$-valued random variable Y with $Y \mid \mathcal{H}' \sim F(\cdot \mid \mathcal{H}')$ and $\mathbb{P}(\|X - Y\|_p > \eta) \leq \rho$.

Proof (Lemma A.1)
By Theorem B.2 in Chen and Kato (2020), noting that the σ-algebra generated by Z is countably generated and using the metric induced by the ℓ^p-norm. □

Next, we present in Lemma A.2 an analytic result concerning the smooth approximation of Borel set indicator functions, similar to that given in Belloni et al. (2019, Lemma 39).

Lemma A.2 (Smooth approximation of Borel indicator functions)
Let $A \subseteq \mathbb{R}^d$ be a Borel set and $Z \sim \mathcal{N}(0, I_d)$. For $\sigma, \eta > 0$ and $p \in [1, \infty]$, define

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

Then f is infinitely differentiable and with $\varepsilon = \mathbb{P}(\|Z\|_p > \eta/\sigma)$, for all $k \geq 0$, any multi-index $\kappa = (\kappa_1, \ldots, \kappa_d) \in \mathbb{N}^d$, and all $x, y \in \mathbb{R}^d$, we have $|\partial^\kappa f_{A\eta\sigma}(x)| \leq \frac{\sqrt{\kappa!}}{\sigma^{|\kappa|}}$ and

$$\left|f_{A\eta\sigma}(x + y) - \sum_{|\kappa|=0}^{k} \frac{1}{\kappa!}\partial^\kappa f_{A\eta\sigma}(x) y^\kappa\right| \leq \frac{\|y\|_p \|y\|_2^k}{\sigma^k \eta \sqrt{k!}},$$
$$(1 - \varepsilon)\mathbb{I}\big\{x \in A\big\} \leq f_{A\eta\sigma}(x) \leq \varepsilon + (1 - \varepsilon)\mathbb{I}\big\{x \in A^{3\eta}\big\}.$$

Proof (Lemma A.2)
Drop the subscripts on $g_{A\eta}$ and $f_{A\eta\sigma}$. By Taylor's theorem with Lagrange remainder, for a $t \in [0, 1]$,

$$\left|f(x + y) - \sum_{|\kappa|=0}^{k} \frac{1}{\kappa!}\partial^\kappa f(x) y^\kappa\right| \leq \left|\sum_{|\kappa|=k} \frac{y^\kappa}{\kappa!}\big(\partial^\kappa f(x + ty) - \partial^\kappa f(x)\big)\right|.$$

Now with $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$,

$$f(x) = \mathbb{E}\big[g(x + \sigma W)\big] = \int_{\mathbb{R}^d} g(x + \sigma u) \prod_{j=1}^{d} \phi(u_j) \, \mathrm{d}u = \frac{1}{\sigma^d} \int_{\mathbb{R}^d} g(u) \prod_{j=1}^{d} \phi\left(\frac{u_j - x_j}{\sigma}\right) \mathrm{d}u$$

and since the integrand is bounded, we exchange differentiation and integration to compute

$$\begin{aligned}
\partial^\kappa f(x) &= \frac{1}{\sigma^{d+|\kappa|}} \int_{\mathbb{R}^d} g(u) \prod_{j=1}^{d} \partial^{\kappa_j} \phi\left(\frac{u_j - x_j}{\sigma}\right) \mathrm{d}u = \left(\frac{-1}{\sigma}\right)^{|\kappa|} \int_{\mathbb{R}^d} g(x + \sigma u) \prod_{j=1}^{d} \partial^{\kappa_j} \phi(u_j) \, \mathrm{d}u \\
&= \left(\frac{-1}{\sigma}\right)^{|\kappa|} \mathbb{E}\left[g(x + \sigma Z) \prod_{j=1}^{d} \frac{\partial^{\kappa_j} \phi(Z_j)}{\phi(Z_j)}\right], \qquad (8)
\end{aligned}$$

where $Z \sim \mathcal{N}(0, I_d)$. Recalling that $|g(x)| \leq 1$ and applying the Cauchy–Schwarz inequality,

$$|\partial^\kappa f(x)| \leq \frac{1}{\sigma^{|\kappa|}} \prod_{j=1}^{d} \mathbb{E}\left[\left(\frac{\partial^{\kappa_j} \phi(Z_j)}{\phi(Z_j)}\right)^2\right]^{1/2} \leq \frac{1}{\sigma^{|\kappa|}} \prod_{j=1}^{d} \sqrt{\kappa_j!} = \frac{\sqrt{\kappa!}}{\sigma^{|\kappa|}},$$

where we used the expected square of the Hermite polynomial of degree κ_j against the standard Gaussian measure is $\kappa_j!$. By the reverse triangle inequality, $|g(x + ty) - g(x)| \leq t\|y\|_p/\eta$, so by (8),

$$\begin{aligned}
\left|\sum_{|\kappa|=k} \frac{y^\kappa}{\kappa!} \big(\partial^\kappa f(x + ty) - \partial^\kappa f(x)\big)\right| &= \left|\sum_{|\kappa|=k} \frac{y^\kappa}{\kappa!} \frac{1}{\sigma^{|\kappa|}} \mathbb{E}\left[\big(g(x + ty + \sigma Z) - g(x + \sigma Z)\big) \prod_{j=1}^{d} \frac{\partial^{\kappa_j} \phi(Z_j)}{\phi(Z_j)}\right]\right| \\
&\leq \frac{t\|y\|_p}{\sigma^k \eta} \mathbb{E}\left[\left|\sum_{|\kappa|=k} \frac{y^\kappa}{\kappa!} \prod_{j=1}^{d} \frac{\partial^{\kappa_j} \phi(Z_j)}{\phi(Z_j)}\right|\right].
\end{aligned}$$

Therefore by the Cauchy–Schwarz inequality,

$$\begin{aligned}
\left(\sum_{|\kappa|=k} \frac{y^\kappa}{\kappa!} \big(\partial^\kappa f(x + ty) - \partial^\kappa f(x)\big)\right)^2 &\leq \frac{t^2\|y\|_p^2}{\sigma^{2k}\eta^2} \mathbb{E}\left[\left(\sum_{|\kappa|=k} \frac{y^\kappa}{\kappa!} \prod_{j=1}^{d} \frac{\partial^{\kappa_j} \phi(Z_j)}{\phi(Z_j)}\right)^2\right] \\
&= \frac{t^2\|y\|_p^2}{\sigma^{2k}\eta^2} \sum_{|\kappa|=k} \sum_{|\kappa'|=k} \frac{y^{\kappa+\kappa'}}{\kappa!\kappa'!} \prod_{j=1}^{d} \mathbb{E}\left[\frac{\partial^{\kappa_j} \phi(Z_j)}{\phi(Z_j)} \frac{\partial^{\kappa'_j} \phi(Z_j)}{\phi(Z_j)}\right].
\end{aligned}$$

Orthogonality of Hermite polynomials shows this is zero if $\kappa_j \neq \kappa'_j$. By the multinomial theorem,

$$\left|f(x + y) - \sum_{|\kappa|=0}^{k} \frac{1}{\kappa!} \partial^\kappa f(x) y^\kappa\right| \leq \frac{\|y\|_p}{\sigma^k \eta} \left(\sum_{|\kappa|=k} \frac{y^{2\kappa}}{\kappa!}\right)^{1/2} \leq \frac{\|y\|_p}{\sigma^k \eta \sqrt{k!}} \left(\sum_{|\kappa|=k} \frac{k!}{\kappa!} y^{2\kappa}\right)^{1/2} \leq \frac{\|y\|_p \|y\|_2^k}{\sigma^k \eta \sqrt{k!}}.$$

For the final result, since $f(x) = \mathbb{E}\left[g(x + \sigma Z)\right]$ and $\mathbb{I}\big\{x \in A^\eta\big\} \leq g(x) \leq \mathbb{I}\big\{x \in A^{2\eta}\big\}$, we have

$$\begin{aligned}
&f(x) \leq \mathbb{P}\left(x + \sigma Z \in A^{2\eta}\right) \leq \mathbb{P}\left(\|Z\|_p > \frac{\eta}{\sigma}\right) + \mathbb{I}\left\{x \in A^{3\eta}\right\} \mathbb{P}\left(\|Z\|_p \leq \frac{\eta}{\sigma}\right) = \varepsilon + (1 - \varepsilon)\mathbb{I}\left\{x \in A^{3\eta}\right\}, \\
&f(x) \geq \mathbb{P}\left(x + \sigma Z \in A^\eta\right) \leq \mathbb{I}\left\{x \in A\right\} \mathbb{P}\left(\|Z\|_p \leq \frac{\eta}{\sigma}\right) = (1 - \varepsilon)\mathbb{I}\left\{x \in A\right\}.
\end{aligned}$$

□

We provide a useful Gaussian inequality in Lemma A.3 which helps bound the $\beta_{\infty,k}$ moment terms appearing in several places throughout the paper.

Lemma A.3 (A useful Gaussian inequality)
Let $X \sim \mathcal{N}(0, \Sigma)$ where $\sigma_j^2 = \Sigma_{jj} \le \sigma^2$ for all $1 \le j \le d$. Then

$$\mathbb{E}\left[\|X\|_2^2 \|X\|_\infty\right] \le 4\sigma\sqrt{\log 2d} \sum_{j=1}^d \sigma_j^2 \qquad \textit{and} \qquad \mathbb{E}\left[\|X\|_2^3 \|X\|_\infty\right] \le 8\sigma\sqrt{\log 2d} \left(\sum_{j=1}^d \sigma_j^2\right)^{3/2}.$$

Proof (Lemma A.3)
By Cauchy–Schwarz, and with $k \in \{2, 3\}$, we have $\mathbb{E}\left[\|X\|_2^k \|X\|_\infty\right] \le \mathbb{E}\big[\|X\|_2^{2k}\big]^{1/2} \mathbb{E}\big[\|X\|_\infty^2\big]^{1/2}$. For the first term, by Hölder's inequality and the fourth and sixth moments of the normal distribution,

$$\begin{aligned}
\mathbb{E}\big[\|X\|_2^4\big] &= \mathbb{E}\left[\left(\sum_{j=1}^d X_j^2\right)^2\right] = \sum_{j=1}^d \sum_{k=1}^d \mathbb{E}\big[X_j^2 X_k^2\big] \le \left(\sum_{j=1}^d \mathbb{E}\big[X_j^4\big]^{1/2}\right)^2 = 3\left(\sum_{j=1}^d \sigma_j^2\right)^2, \\
\mathbb{E}\big[\|X\|_2^6\big] &= \mathbb{E}\left[\left(\sum_{j=1}^d X_j^2\right)^3\right] = \sum_{j=1}^d \sum_{k=1}^d \sum_{l=1}^d \mathbb{E}\big[X_j^2 X_k^2 X_l^2\big] \le \left(\sum_{j=1}^d \mathbb{E}\big[X_j^6\big]^{1/3}\right)^3 = 15\left(\sum_{j=1}^d \sigma_j^2\right)^3.
\end{aligned}$$

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

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

$\square$

We provide an ℓ^p-norm tail probability bound for Gaussian variables in Lemma A.4, motivating the definition of the term $\phi_p(d)$.

Lemma A.4 (Gaussian ℓ^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 A.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

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

by Jensen's inequality. For $p = \infty$, with $\sigma^2 = \max_j \Sigma_{jj}$, for $t > 0$,

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

again by Jensen's inequality. Setting $t = \frac{\sigma}{\sqrt{2 \log 2d}}$ gives $\mathbb{E}\big[\|X\|_\infty\big] \le \sigma\sqrt{2 \log 2d}$. $\square$

We give a Gaussian–Gaussian ℓ^p-norm approximation as Lemma A.5, which is useful for ensuring approximations remain valid upon substituting an estimator for the true variance matrix.

Lemma A.5 (Gaussian–Gaussian approximation in ℓ^p-norm)
Let $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$ be positive semi-definite and take $Z \sim \mathcal{N}(0, I_d)$. For $p \in [1,\infty]$ we have

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

Proof (Lemma A.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,

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

The same result holds for $p = \infty$ since

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

Now we apply this to 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

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

□

We also include, for completeness, a variance bound (Lemma A.6) and an exponential concentration inequality (Lemma A.7) for α-mixing random variables.

Lemma A.6 (Variance bounds for α-mixing random variables)
Let $X_1, \ldots, X_n$ be real-valued α-mixing random variables with mixing coefficients $\alpha(j)$. Then

(i) If for constants M_i we have $|X_i| \leq M_i$ a.s. then

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

(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

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

Proof (Lemma A.6)
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 $r > 2$,

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

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

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

□

Lemma A.7 (Exponential concentration 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 $\sum_{j=1}^n |\operatorname{Cov}[X_i, X_j]| \leq \sigma^2$. 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 A.7)
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 $M\sqrt{nt} + M(\log n)(\log\log n)t$.

(ii) By Theorem 2 in Merlevède et al. (2009),

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| > t\right) \leq \exp\left(-\frac{C_2 t^2}{n\sigma^2 + M^2 + Mt(\log n)^2}\right).$$

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

□

A.2 Main results

To establish Theorem 2.1, we first give the analogous result for martingales as Lemma A.8. Our approach is similar to that used in modern versions of Yurinskii's coupling for independent data, as in Theorem 1 in Le Cam (1988) and Theorem 10 in Chapter 10 of Pollard (2002). The proof of Lemma A.8 relies on constructing a "modified" martingale, which is close to the original martingale, but which has an $\mathcal{H}_0$-measurable terminal quadratic variation.

Lemma A.8 (Strong approximation for vector-valued martingales)
Let $X_1, \ldots, X_n$ be $\mathbb{R}^d$-valued square-integrable random vectors adapted to a countably generated filtration $\mathcal{H}_0, \ldots, \mathcal{H}_n$. Suppose that $\mathbb{E}[X_i \mid \mathcal{H}_{i-1}] = 0$ for all $1 \leq i \leq n$ and define the martingale $S = \sum_{i=1}^n X_i$. Let $V_i = \mathrm{Var}[X_i \mid \mathcal{H}_{i-1}]$ and $\Omega = \sum_{i=1}^n V_i - \Sigma$ where Σ is a positive semi-definite $\mathcal{H}_0$-measurable $d \times d$ random matrix. For each $\eta > 0$ and $p \in [1, \infty]$ there is $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$ with

$$\begin{aligned}\mathbb{P}\big(\|S - T\|_p > 5\eta\big) \leq \inf_{t>0} &\left\{2\mathbb{P}\big(\|Z\|_p > t\big) + \min\left\{\frac{\beta_{p,2} t^2}{\eta^3}, \frac{\beta_{p,3} t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3}\right\}\right\} \\ &+ \inf_{M \succeq 0} \big\{2\gamma(M) + \delta_p(M, \eta) + \varepsilon_p(M, \eta)\big\},\end{aligned}$$

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

$$\beta_{p,k} = \sum_{i=1}^n \mathbb{E}\left[\|X_i\|_2^k \|X_i\|_p + \|V_i^{1/2} Z_i\|_2^k \|V_i^{1/2} Z_i\|_p\right], \qquad \gamma(M) = \mathbb{P}\big(\Omega \npreceq M\big),$$

$$\delta_p(M, \eta) = \mathbb{P}\left(\left\|\big((\Sigma + M)^{1/2} - \Sigma^{1/2}\big) Z\right\|_p \geq \eta\right), \qquad \pi_3 = \sum_{i=1}^{n+m} \sum_{|\kappa|=3} \mathbb{E}\Big[\big|\mathbb{E}\left[X_i^\kappa \mid \mathcal{H}_{i-1}\right]\big|\Big],$$

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

for $k \in \{2, 3\}$, with $Z, Z_1, \ldots, Z_n$ i.i.d. standard Gaussian variables on $\mathbb{R}^d$ independent of $\mathcal{H}_n$.

Proof (Lemma A.8)

Part 1: constructing a modified martingale

Take $M \succeq 0$ a fixed positive semi-definite $d \times d$ matrix. We start by constructing a new martingale based on S whose quadratic variation is $\Sigma + M$. Take $m \geq 1$ and define

$$H_k = \Sigma + M - \sum_{i=1}^{k} V_i, \qquad \tau = \sup\big\{k \in \{0, 1, \ldots, n\} : H_k \succeq 0\big\},$$

$$\tilde{X}_i = X_i \mathbb{I}\{i \leq \tau\} + \frac{1}{\sqrt{m}} H_\tau^{1/2} Z_i \mathbb{I}\{n+1 \leq i \leq n+m\}, \quad \tilde{S} = \sum_{i=1}^{n+m} \tilde{X}_i,$$

where $Z_{n+1}, \ldots, Z_{n+m}$ is an i.i.d. sequence of standard Gaussian vectors in $\mathbb{R}^d$ independent of $\mathcal{H}_n$, noting that $H_0 = \Sigma + M \succeq 0$ a.s. Define the filtration $\tilde{\mathcal{H}}_0, \ldots, \tilde{\mathcal{H}}_{n+m}$, where $\tilde{\mathcal{H}}_i = \mathcal{H}_i$ for $0 \leq i \leq n$ and is the σ-algebra generated by $\mathcal{H}_n$ and $Z_{n+1}, \ldots, Z_i$ for $n+1 \leq i \leq n+m$. Observe that τ is a stopping time with respect to $\tilde{\mathcal{H}}_i$ because $H_{i+1} - H_i = -V_{i+1} \preceq 0$ almost surely, so $\{\tau \leq i\} = \{H_{i+1} \not\succeq 0\}$ for $0 \leq i < n$. This depends only on $V_1, \ldots, V_{i+1}$ and Σ which are $\tilde{\mathcal{H}}_i$-measurable. Similarly, $\{\tau = n\} = \{H_n \succeq 0\} \in \tilde{\mathcal{H}}_{n-1}$. Let $\tilde{V}_i = V_i \mathbb{I}\{i \leq \tau\}$ for $1 \leq i \leq n$ and $\tilde{V}_i = H_\tau / m$ for $n+1 \leq i \leq n+m$. Note that $\tilde{X}_i$ is $\tilde{\mathcal{H}}_i$-measurable and $\tilde{V}_i$ is $\tilde{\mathcal{H}}_{i-1}$-measurable. Further, $\mathbb{E}\Big[\tilde{X}_i \mid \tilde{\mathcal{H}}_{i-1}\Big] = 0$ and $\mathbb{E}\Big[\tilde{X}_i \tilde{X}_i^\mathsf{T} \mid \tilde{\mathcal{H}}_{i-1}\Big] = \tilde{V}_i$.

Part 2: bounding the difference between the original and modified martingales

By the triangle inequality,

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

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 $\mathbb{P}\big(\|H_\tau^{1/2} Z\|_p > \eta\big) \leq \mathbb{P}\big(\|H_n^{1/2} Z\|_p > \eta, \, \Omega \preceq M\big) + \mathbb{P}\big(\Omega \not\preceq M\big)$. Therefore

$$\mathbb{P}\big(\|S - \tilde{S}\|_p > \eta\big) \leq 2\mathbb{P}\big(\Omega \not\preceq M\big) + \mathbb{P}\big(\|(M - \Omega)^{1/2} Z\|_p > \eta, \, \Omega \preceq M\big) = 2\gamma(M) + \varepsilon_p(M, \eta). \tag{9}$$

Part 3: strong approximation of the modified martingale

Let $\tilde{Z}_1, \ldots, \tilde{Z}_{n+m}$ be i.i.d. $\mathcal{N}(0, I_d)$ and independent of $\tilde{\mathcal{H}}_{n+m}$. Define $\check{X}_i = \tilde{V}_i^{1/2} \tilde{Z}_i$ and $\check{S} = \sum_{i=1}^{n+m} \check{X}_i$. Fix a Borel set $A \subseteq \mathbb{R}^d$ and $\sigma, \eta > 0$ and let $f = f_{A\eta\sigma}$ be the function defined in Lemma A.2. By the Lindeberg method, write the telescoping sum

$$\mathbb{E}\Big[f\big(\tilde{S}\big) - f\big(\check{S}\big) \mid \mathcal{H}_0\Big] = \sum_{i=1}^{n+m} \mathbb{E}\Big[f\big(Y_i + \tilde{X}_i\big) - f\big(Y_i + \check{X}_i\big) \mid \mathcal{H}_0\Big]$$

where $Y_i = \sum_{j=1}^{i-1} \tilde{X}_i + \sum_{j=i+1}^{n+m} \check{X}_i$. By Lemma A.2 we have for $k \geq 0$

$$\left|\mathbb{E}\big[f(Y_i + \tilde{X}_i) - f(Y_i + \check{X}_i) \mid \mathcal{H}_0\big] - \sum_{|\kappa|=0}^{k} \frac{1}{\kappa!} \mathbb{E}\left[\partial^\kappa f(Y_i) \left(\tilde{X}_i^\kappa - \check{X}_i^\kappa\right) \mid \mathcal{H}_0\right]\right|$$
$$\leq \frac{1}{\sigma^k \eta \sqrt{k!}} \mathbb{E}\left[\|\tilde{X}_i\|_p \|\tilde{X}_i\|_2^k + \|\check{X}_i\|_p \|\check{X}_i\|_2^k \mid \mathcal{H}_0\right].$$

With $k \in \{2,3\}$, we bound each summand. With $|\kappa| = 0$ we have $\tilde{X}_i^\kappa = \check{X}_i^\kappa$, so consider $|\kappa| = 1$. Noting that $\sum_{i=1}^{n+m} \tilde{V}_i = \Sigma + M$, define

$$\tilde{Y}_i = \sum_{j=1}^{i-1} \tilde{X}_j + \tilde{Z}_i \left(\sum_{j=i+1}^{n+m} \tilde{V}_j \right)^{1/2} = \sum_{j=1}^{i-1} \tilde{X}_j + \tilde{Z}_i \left(\Sigma + M - \sum_{j=1}^{i} \tilde{V}_j \right)^{1/2}$$

and let $\check{\mathcal{H}}_i$ be the σ-algebra generated by $\tilde{\mathcal{H}}_{i-1}$ and $\tilde{Z}_i$. Note that $\tilde{Y}_i$ is $\check{\mathcal{H}}_i$-measurable and that Y_i and $\tilde{Y}_i$ have the same distribution conditional on $\tilde{\mathcal{H}}_{n+m}$. So

$$\begin{aligned}
\sum_{|\kappa|=1} \frac{1}{\kappa!} \mathbb{E}\left[\partial^\kappa f(Y_i)\big(\tilde{X}_i^\kappa - \check{X}_i^\kappa\big) \mid \mathcal{H}_0\right] &= \mathbb{E}\left[\nabla f(Y_i)^\mathsf{T}\big(\tilde{X}_i - \tilde{V}_i^{1/2}\tilde{Z}_i\big) \mid \mathcal{H}_0\right] \\
&= \mathbb{E}\left[\nabla f(\tilde{Y}_i)^\mathsf{T}\tilde{X}_i \mid \mathcal{H}_0\right] - \mathbb{E}\left[\nabla f(Y_i)^\mathsf{T}\tilde{V}_i^{1/2}\tilde{Z}_i \mid \mathcal{H}_0\right] \\
&= \mathbb{E}\left[\nabla f(\tilde{Y}_i)^\mathsf{T}\mathbb{E}\left[\tilde{X}_i \mid \check{\mathcal{H}}_i\right] \mid \mathcal{H}_0\right] - \mathbb{E}\left[\tilde{Z}_i\right]\mathbb{E}\left[\nabla f(Y_i)^\mathsf{T}\tilde{V}_i^{1/2} \mid \mathcal{H}_0\right] \\
&= \mathbb{E}\left[\nabla f(\tilde{Y}_i)^\mathsf{T}\mathbb{E}\left[\tilde{X}_i \mid \tilde{\mathcal{H}}_{i-1}\right] \mid \mathcal{H}_0\right] - 0 = 0.
\end{aligned}$$

Next, if $|\kappa| = 2$ then

$$\begin{aligned}
\sum_{|\kappa|=2} \frac{1}{\kappa!} \mathbb{E}\left[\partial^\kappa f(Y_i)\left(\tilde{X}_i^\kappa - \check{X}_i^\kappa\right) \mid \mathcal{H}_0\right] &= \frac{1}{2}\mathbb{E}\left[\tilde{X}_i^\mathsf{T}\nabla^2 f(Y_i)\tilde{X}_i - \tilde{Z}_i^\mathsf{T}\tilde{V}_i^{1/2}\nabla^2 f(Y_i)\tilde{V}_i^{1/2}\tilde{Z}_i \mid \mathcal{H}_0\right] \\
= \frac{1}{2}\mathbb{E}\left[\mathbb{E}\left[\operatorname{Tr}\nabla^2 f(\tilde{Y}_i)\tilde{X}_i\tilde{X}_i^\mathsf{T} \mid \check{\mathcal{H}}_i\right] \mid \mathcal{H}_0\right] &- \frac{1}{2}\mathbb{E}\left[\operatorname{Tr}\tilde{V}_i^{1/2}\nabla^2 f(Y_i)\tilde{V}_i^{1/2} \mid \mathcal{H}_0\right]\mathbb{E}\left[\tilde{Z}_i\tilde{Z}_i^\mathsf{T}\right] \\
= \frac{1}{2}\mathbb{E}\left[\operatorname{Tr}\nabla^2 f(Y_i)\mathbb{E}\left[\tilde{X}_i\tilde{X}_i^\mathsf{T} \mid \tilde{\mathcal{H}}_{i-1}\right] \mid \mathcal{H}_0\right] &- \frac{1}{2}\mathbb{E}\left[\operatorname{Tr}\nabla^2 f(Y_i)\tilde{V}_i \mid \mathcal{H}_0\right] = 0.
\end{aligned}$$

Finally if $|\kappa| = 3$, then since $\check{X}_i \sim \mathcal{N}(0, \tilde{V}_i)$ conditional on $\tilde{\mathcal{H}}_{n+m}$, we have by symmetry of the Gaussian distribution and Lemma A.2,

$$\begin{aligned}
&\left|\sum_{|\kappa|=3} \frac{1}{\kappa!}\mathbb{E}\left[\partial^\kappa f(Y_i)\left(\tilde{X}_i^\kappa - \check{X}_i^\kappa\right) \mid \mathcal{H}_0\right]\right| \\
&\quad= \left|\sum_{|\kappa|=3} \frac{1}{\kappa!}\left(\mathbb{E}\left[\partial^\kappa f(\tilde{Y}_i)\mathbb{E}\left[\tilde{X}_i^\kappa \mid \check{\mathcal{H}}_i\right] \mid \mathcal{H}_0\right] - \mathbb{E}\left[\partial^\kappa f(Y_i)\,\mathbb{E}\left[\check{X}_i^\kappa \mid \tilde{\mathcal{H}}_{n+m}\right] \mid \mathcal{H}_0\right]\right)\right| \\
&\quad= \left|\sum_{|\kappa|=3} \frac{1}{\kappa!}\mathbb{E}\left[\partial^\kappa f(Y_i)\,\mathbb{E}\left[\tilde{X}_i^\kappa \mid \tilde{\mathcal{H}}_{i-1}\right] \mid \mathcal{H}_0\right]\right| \leq \frac{1}{\sigma^3}\sum_{|\kappa|=3}\mathbb{E}\left[\left|\mathbb{E}\left[\tilde{X}_i^\kappa \mid \tilde{\mathcal{H}}_{i-1}\right]\right| \mid \mathcal{H}_0\right].
\end{aligned}$$

Combining these and summing over i with $k = 2$ shows

$$\mathbb{E}\left[f(\tilde{S}) - f(\check{S}) \mid \mathcal{H}_0\right] \leq \frac{1}{\sigma^2\eta\sqrt{2}}\sum_{i=1}^{n+m}\mathbb{E}\left[\|\tilde{X}_i\|_p\|\tilde{X}_i\|_2^2 + \|\check{X}_i\|_p\|\check{X}_i\|_2^2 \mid \mathcal{H}_0\right]$$

On the other hand, taking $k = 3$ gives

$$\begin{aligned}
\mathbb{E}\left[f(\tilde{S}) - f(\check{S}) \mid \mathcal{H}_0\right] &\leq \frac{1}{\sigma^3\eta\sqrt{6}}\sum_{i=1}^{n+m}\mathbb{E}\left[\|\tilde{X}_i\|_p\|\tilde{X}_i\|_2^3 + \|\check{X}_i\|_p\|\check{X}_i\|_2^3 \mid \mathcal{H}_0\right] \\
&\quad+ \frac{1}{\sigma^3}\sum_{i=1}^{n+m}\sum_{|\kappa|=3}\mathbb{E}\left[\left|\mathbb{E}\left[\tilde{X}_i^\kappa \mid \tilde{\mathcal{H}}_{i-1}\right]\right| \mid \mathcal{H}_0\right].
\end{aligned}$$

For $1 \leq i \leq n$ we have $\|\tilde{X}_i\| \leq \|X_i\|$ and $\|\check{X}_i\| \leq \|V_i^{1/2}\tilde{Z}_i\|$. For $n+1 \leq i \leq n+m$ we have $\tilde{X}_i = H_\tau^{1/2} Z_i/\sqrt{m}$ and $\check{X}_i = H_\tau^{1/2}\tilde{Z}_i/\sqrt{m}$ which are equal in distribution given $\mathcal{H}_0$. Therefore with

$$\tilde{\beta}_{p,k} = \sum_{i=1}^n \mathbb{E}\left[\|X_i\|_p \|X_i\|_2^k + \|V_i^{1/2} Z_i\|_p \|V_i^{1/2} Z_i\|_2^k \mid \mathcal{H}_0\right],$$

we have, since $k \in \{2, 3\}$,

$$\sum_{i=1}^{n+m} \mathbb{E}\left[\|\tilde{X}_i\|_p \|\tilde{X}_i\|_2^k + \|\check{X}_i\|_p \|\check{X}_i\|_2^k \mid \mathcal{H}_0\right] \leq \tilde{\beta}_{p,k} + \frac{2}{\sqrt{m}} \mathbb{E}\left[\|H_\tau^{1/2} Z\|_p \|H_\tau^{1/2} Z\|_2^k \mid \mathcal{H}_0\right].$$

Since H_i is weakly decreasing under the semi-definite partial order, we have $H_\tau \preceq H_0 = \Sigma + M$ implying that $|(H_\tau)_{jj}| \leq \|\Sigma + M\|_{\max}$ and $\mathbb{E}\big[|(H_\tau^{1/2} Z)_j|^3 \mid \mathcal{H}_0\big] \leq \sqrt{8/\pi}\, \|\Sigma + M\|_{\max}^{3/2}$. Hence as $p \geq 1$ and $k \in \{2, 3\}$,

$$\begin{aligned}
\mathbb{E}\left[\|H_\tau^{1/2} Z\|_p \|H_\tau^{1/2} Z\|_2^k \mid \mathcal{H}_0\right] &\leq \mathbb{E}\left[\|H_\tau^{1/2} Z\|_1^{k+1} \mid \mathcal{H}_0\right] \leq d^{k+1} \max_{1 \leq j \leq d} \mathbb{E}\left[|(H_\tau^{1/2} Z)_j|^{k+1} \mid \mathcal{H}_0\right] \\
&\leq 3d^4 \, \|\Sigma + M\|_{\max}^{(k+1)/2} \leq 6d^4 \, \|\Sigma\|_{\max}^{(k+1)/2} + 6d^4 \|M\|.
\end{aligned}$$

Assuming some X_i is not identically zero so the result is non-trivial, and supposing that Σ is bounded a.s. (replacing Σ by $\Sigma \cdot \mathbb{I}\{\|\Sigma\|_{\max} \leq C\}$ for an appropriately large C if necessary), take m large enough that

$$\frac{2}{\sqrt{m}} \mathbb{E}\left[\|H_\tau^{1/2} Z\|_p \|H_\tau^{1/2} Z\|_2^k \mid \mathcal{H}_0\right] \leq \frac{1}{4} \beta_{p,k}. \tag{10}$$

Further, if $|\kappa| = 3$ then $\left|\mathbb{E}\big[\tilde{X}_i^\kappa \mid \tilde{\mathcal{H}}_{i-1}\big]\right| \leq \left|\mathbb{E}\left[X_i^\kappa \mid \mathcal{H}_{i-1}\right]\right|$ for $1 \leq i \leq n$ while by symmetry of the Gaussian distribution $\mathbb{E}\left[\tilde{X}_i^\kappa \mid \tilde{\mathcal{H}}_{i-1}\right] = 0$ for $n+1 \leq i \leq n+m$. Hence with

$$\tilde{\pi}_3 = \sum_{i=1}^{n+m} \sum_{|\kappa|=3} \mathbb{E}\Big[\left|\mathbb{E}\left[X_i^\kappa \mid \mathcal{H}_{i-1}\right]\right| \mid \mathcal{H}_0\Big],$$

we have

$$\mathbb{E}\left[f(\tilde{S}) - f(\check{S}) \mid \mathcal{H}_0\right] \leq \min\left\{\frac{3\tilde{\beta}_{p,2}}{4\sigma^2\eta} + \frac{\beta_{p,2}}{4\sigma^2\eta}, \frac{3\tilde{\beta}_{p,3}}{4\sigma^3\eta} + \frac{\beta_{p,3}}{4\sigma^3\eta} + \frac{\tilde{\pi}_3}{\sigma^3}\right\}.$$

Along with Lemma A.2, and with $\sigma = \eta/t$ and $\varepsilon = \mathbb{P}(\|Z\|_p > t)$, we conclude that

$$\begin{aligned}
\mathbb{P}(\tilde{S} \in A \mid \mathcal{H}_0) &= \mathbb{E}\big[\mathbb{I}\{\tilde{S} \in A\} - f(\tilde{S}) \mid \mathcal{H}_0\big] + \mathbb{E}\big[f(\tilde{S}) - f(\check{S}) \mid \mathcal{H}_0\big] + \mathbb{E}\big[f(\check{S}) \mid \mathcal{H}_0\big] \\
&\leq \varepsilon\, \mathbb{P}(\tilde{S} \in A \mid \mathcal{H}_0) + \min\left\{\frac{3\tilde{\beta}_{p,2}}{4\sigma^2\eta} + \frac{\beta_{p,2}}{4\sigma^2\eta}, \frac{3\tilde{\beta}_{p,3}}{4\sigma^3\eta} + \frac{\beta_{p,3}}{4\sigma^3\eta} + \frac{\tilde{\pi}_3}{\sigma^3}\right\} + \varepsilon + (1-\varepsilon)\mathbb{P}\big(\check{S} \in A_p^{3\eta} \mid \mathcal{H}_0\big) \\
&\leq \mathbb{P}\big(\check{S} \in A_p^{3\eta} \mid \mathcal{H}_0\big) + 2\mathbb{P}(\|Z\|_p > t) + \min\left\{\frac{3\tilde{\beta}_{p,2} t^2}{4\eta^3} + \frac{\beta_{p,2} t^2}{4\eta^3}, \frac{3\tilde{\beta}_{p,3} t^3}{4\eta^4} + \frac{\beta_{p,3} t^3}{4\eta^4} + \frac{\tilde{\pi}_3 t^3}{\eta^3}\right\}.
\end{aligned}$$

Taking a supremum and an outer expectation yields with $\beta_{p,k} = \mathbb{E}\big[\tilde{\beta}_{p,k}\big]$ and $\pi_3 = \mathbb{E}[\tilde{\pi}_3]$,

$$\mathbb{E}^*\left[\sup_{A \in \mathcal{B}(\mathbb{R}^d)} \Big\{\mathbb{P}(\tilde{S} \in A \mid \mathcal{H}_0) - \mathbb{P}\big(\check{S} \in A_p^{3\eta} \mid \mathcal{H}_0\big)\Big\}\right] \leq 2\mathbb{P}(\|Z\|_p > t) + \min\left\{\frac{\beta_{p,2} t^2}{\eta^3}, \frac{\beta_{p,3} t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3}\right\}.$$

Finally, since $\check{S} = \sum_{i=1}^n \tilde{V}_i^{1/2} \tilde{Z}_i \sim \mathcal{N}(0, \Sigma + M)$ conditional on $\mathcal{H}_0$, the conditional Strassen theorem in Lemma A.1 ensures the existence of $\tilde{S}$ and $\tilde{T} \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma + M)$ such that

$$\mathbb{P}\left(\|\tilde{S} - \tilde{T}\|_p > 3\eta\right) \leq \inf_{t>0} \left\{2\mathbb{P}(\|Z\|_p > t) + \min\left\{\frac{\beta_{p,2} t^2}{\eta^3}, \frac{\beta_{p,3} t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3}\right\}\right\}, \tag{11}$$

since the infimum is attained by continuity of $\|Z\|_p$.

Part 4: conclusion

We show how to write $\tilde{T} = (\Sigma + M)^{1/2} W$ where $W \sim \mathcal{N}(0, I_d)$ and use this representation to construct $T \mid \mathcal{H}_0 \sim \mathcal{N}(0, \Sigma)$. By the spectral theorem, let $\Sigma + M = U \Lambda U^\mathsf{T}$ where U is a $d \times d$ orthogonal random matrix and Λ is a diagonal $d \times d$ random matrix with diagonal entries satisfying $\lambda_1 \geq \cdots \geq \lambda_r > 0$ and $\lambda_{r+1} = \cdots = \lambda_d = 0$ where $r = \mathrm{rank}(\Sigma + M)$. Let Λ^+ be the Moore–Penrose pseudo-inverse of Λ (obtained by inverting its non-zero elements) and define $W = U(\Lambda^+)^{1/2} U^\mathsf{T} \tilde{T} + U \tilde{W}$, where the first r elements of $\tilde{W}$ are zero and the last $d - r$ elements are i.i.d. $\mathcal{N}(0,1)$ independent from $\tilde{T}$. Then, it is easy to check that $W \sim \mathcal{N}(0, I_d)$ and that $\tilde{T} = (\Sigma + M)^{1/2} W$. Now define $T = \Sigma^{1/2} W$ so

$$\mathbb{P}\big(\|T - \tilde{T}\|_p > \eta\big) = \mathbb{P}\big(\big\|\big((\Sigma + M)^{1/2} - \Sigma^{1/2}\big) W\big\|_p > \eta\big) = \delta_p(M, \eta). \tag{12}$$

Finally (9), (11), (12), the triangle inequality and a union bound conclude the proof since by taking an infimum over $M \succeq 0$, and by possibly reducing the constant of $1/4$ in (10) to account for this infimum being potentially unattainable,

$$\begin{aligned}
\mathbb{P}\big(\|S - T\|_p > 5\eta\big) &\leq \mathbb{P}\big(\|\tilde{S} - \tilde{T}\|_p > 3\eta\big) + \mathbb{P}\big(\|S - \tilde{S}\|_p > \eta\big) + \mathbb{P}\big(\|T - \tilde{T}\|_p > \eta\big) \\
&\leq \inf_{t>0} \left\{2\mathbb{P}\big(\|Z\|_p > t\big) + \min\left\{\frac{\beta_{p,2} t^2}{\eta^3}, \frac{\beta_{p,3} t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3}\right\}\right\} \\
&\quad + \inf_{M \succeq 0} \big\{2\gamma(M) + \delta_p(M, \eta) + \varepsilon_p(M, \eta)\big\}.
\end{aligned}$$

□

Applying Lemma A.8 and the martingale approximation immediately yields Theorem 2.1.

Proof (Theorem 2.1)

Apply Lemma A.8 to the martingale $\sum_{i=1}^n \tilde{X}_i$, noting that $S - \sum_{i=1}^n \tilde{X}_i = U$. □

Providing bounds for quantities in Theorem 2.1 gives a user-friendly version as Proposition 2.1.

Proof (Proposition 2.1)

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

Part 1: bounding $\mathbb{P}(\|Z\|_p > t)$

By Markov's inequality and Lemma A.4, we have $\mathbb{P}(\|Z\|_p > t) \leq \mathbb{E}[\|Z\|_p]/t \leq \phi_p(d)/t$.

Part 2: bounding $\gamma(M)$

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

Part 3: bounding $\delta(M, \eta)$

By Markov's inequality and Lemma A.4, using $\max_j |M_{jj}| \leq \|M\|_2$ for $M \succeq 0$,

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

For semi-definite matrices the eigenvalue operator commutes with smooth matrix functions so

$$\|(\Sigma + M)^{1/2} - \Sigma^{1/2}\|_2 = \max_{1 \leq j \leq d} \left|\sqrt{\lambda_j(\Sigma) + \nu^2} - \sqrt{\lambda_j(\Sigma)}\right| \leq \nu$$

and hence $\delta_p(M, \eta) \leq \phi_p(d) \nu / \eta$.

Part 4: bounding $\varepsilon(M,\eta)$

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

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

Thus by Theorem 2.1 and the previous parts,

$$\begin{aligned}
\mathbb{P}\big(\|S-T\|_p > 6\eta\big) &\leq \inf_{t>0}\left\{2\mathbb{P}\big(\|Z\|_p > t\big) + \min\left\{\frac{\beta_{p,2}t^2}{\eta^3}, \frac{\beta_{p,3}t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3}\right\}\right\} \\
&\quad + \inf_{M\succeq 0}\left\{2\gamma(M) + \delta_p(M,\eta) + \varepsilon_p(M,\eta)\right\} + \mathbb{P}\big(\|U\|_p > \eta\big) \\
&\leq \inf_{t>0}\left\{\frac{2\phi_p(d)}{t} + \min\left\{\frac{\beta_{p,2}t^2}{\eta^3}, \frac{\beta_{p,3}t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3}\right\}\right\} \\
&\quad + \inf_{\nu>0}\left\{\frac{2\mathbb{E}\left[\|\Omega\|_2\right]}{\nu^2} + \frac{2\phi_p(d)\nu}{\eta}\right\} + \frac{\phi_p(d)\sqrt{\mathbb{E}\left[\|\Omega\|_2\right]}}{\eta} + \mathbb{P}\big(\|U\|_p > \eta\big).
\end{aligned}$$

In general, set $t = 2^{1/3}\phi_p(d)^{1/3}\beta_{p,2}^{-1/3}\eta$ and $\nu = \mathbb{E}[\|\Omega\|_2]^{1/3}\phi_p(d)^{-1/3}\eta^{1/3}$, replacing η with $\eta/6$ to see

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

Whenever $\pi_3 = 0$ we can set $t = 2^{1/4}\phi_p(d)^{1/4}\beta_{p,3}^{-1/4}\eta$, and with ν as above we obtain

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

□

After establishing Proposition 2.1, Corollaries 2.1, 2.2 and 2.3 follow easily, as in the main text.

Proof (Corollary 2.1)
This follows from Proposition 2.1 with $\mathbb{P}(\|U\|_p > \frac{\eta}{6}) \leq \frac{6}{\eta}\sum_{i=1}^n c_i(\zeta_i + \zeta_{n-i+1})$. □

Proof (Corollary 2.2)
This follows from Proposition 2.1 with $U = 0$ a.s. □

Proof (Corollary 2.3)
This follows from Corollary 2.2 with $\Omega = 0$ a.s. □

We conclude this section with a discussion expanding on the comments made in Remark 2.1 on deriving bounds in probability from Yurinskii's coupling. Consider for illustration the independent data result (6) given in Corollary 2.3: for each $\eta > 0$, there exists $T_n \mid \mathcal{H}_0 \sim \mathcal{N}(0,\Sigma)$ satisfying

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

where here we make explicit the dependence on the sample size n for clarity. The naive approach to converting this into a probability bound for $\|S_n - T_n\|_p$ is to select η to ensure the right-hand side is of order 1, arguing that the probability can then be made arbitrarily small by taking, in this case, η to be a large enough multiple of $\beta_{p,2}^{1/3}\phi_p(d)^{2/3}$. However, the somewhat subtle mistake is in neglecting the fact that the realization of the coupling variable T_n will in general depend on η, rendering the resulting bound invalid. As an explicit example of this phenomenon, take $\eta > 1$ and suppose $\|S_n - T_n(\eta)\| = \eta$ with probability $1 - 1/\eta$ and $\|S_n - T_n(\eta)\| = n$ with probability $1/\eta$. Then $\mathbb{P}\big(\|S_n - T_n(\eta)\| > \eta\big) = 1/\eta$ but it is not true for any η that $\|S_n - T_n(\eta)\| \lesssim_{\mathbb{P}} 1$.

We propose in Remark 2.1 the following fix. Instead of selecting η to ensure the right-hand side is of order 1, we instead choose it so the bound converges (slowly) to zero. This is easily achieved by taking the naive and incorrect bound and multiplying by some divergent sequence R_n. The resulting inequality reads, in the case of (6) with $\eta = \beta_{p,2}^{1/3}\phi_p(d)^{2/3}R_n$,

$$\mathbb{P}\Big(\|S_n - T_n\|_p > \beta_{p,2}^{1/3}\phi_p(d)^{2/3}R_n\Big) \leq \frac{24}{R_n} \to 0.$$

We thus recover, for the price of a rate which is slower by an arbitrarily small amount, a valid upper bound in probability, as we can immediately conclude that

$$\|S_n - T_n\|_p \lesssim_{\mathbb{P}} \beta_{p,2}^{1/3}\phi_p(d)^{2/3}R_n.$$

A.3 Strong approximation for martingale empirical processes

We begin by presenting some calculations omitted from the main text relating to the motivating example of kernel density estimation with i.i.d. data. First, the bias of this estimator is bounded as

$$\big|\mathbb{E}\big[\hat{g}(x)\big] - g(x)\big| = \left|\int_{\frac{-x}{h}}^{\frac{1-x}{h}} K(\xi)\,\mathrm{d}\xi - 1\right| \leq 2\int_{\frac{a}{h}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}}\,\mathrm{d}\xi \leq \frac{h}{a}\sqrt{\frac{2}{\pi}} e^{-\frac{a^2}{2h^2}}.$$

Next, we perform the calculations necessary to apply Corollary 2.3. Define $k_{ij} = \frac{1}{nh}K\left(\frac{X_i - x_j}{h}\right)$ and $k_i = (k_{ij} : 1 \leq j \leq N)$. Then $\|k_i\|_\infty \leq \frac{1}{nh\sqrt{2\pi}}$ a.s. and $\mathbb{E}[\|k_i\|_2^2] \leq \frac{N}{n^2 h}\int_{-\infty}^{\infty} K(\xi)^2\,\mathrm{d}\xi \leq \frac{N}{2n^2 h\sqrt{\pi}}$. Let $V = \mathrm{Var}[k_i] \in \mathbb{R}^{N\times N}$, so assuming that $1/h \geq \log 2N$, by Lemma A.3 we bound

$$\beta_{\infty,2} = n\mathbb{E}\left[\|k_i\|_2^2\|k_i\|_\infty\right] + n\mathbb{E}\left[\|V^{1/2}Z\|_2^2\|V^{1/2}Z\|_\infty\right] \leq \frac{N}{\sqrt{8}n^2h^2\pi} + \frac{4N\sqrt{\log 2N}}{\sqrt{8}n^2h^{3/2}\pi^{3/4}} \leq \frac{N}{n^2h^2}.$$

Finally, we verify the stochastic continuity bounds. By the Lipschitz property of K, it is easy to show that for $x, x' \in \mathcal{X}$ we have $\left|\frac{1}{h}K\left(\frac{X_i - x}{h}\right) - \frac{1}{h}K\left(\frac{X_i - x'}{h}\right)\right| \lesssim \frac{|x-x'|}{h^2}$ almost surely, and also that $\mathbb{E}\Big[\left|\frac{1}{h}K\left(\frac{X_i - x}{h}\right) - \frac{1}{h}K\left(\frac{X_i - x'}{h}\right)\right|^2\Big] \lesssim \frac{|x-x'|^2}{h^3}$. By chaining with the Bernstein–Orlicz norm and polynomial covering numbers,

$$\sup_{|x-x'|\leq\delta} \big\|S(x) - S(x')\big\|_\infty \lesssim_{\mathbb{P}} \delta\sqrt{\frac{\log n}{nh^3}}$$

whenever $\log(N/h) \lesssim \log n$ and $nh \gtrsim \log n$. By a Gaussian process maximal inequality (van der Vaart and Wellner, 1996, Corollary 2.2.8) the same bound holds for $T(x)$ with

$$\sup_{|x-x'|\leq\delta} \big\|T(x) - T(x')\big\|_\infty \lesssim_{\mathbb{P}} \delta\sqrt{\frac{\log n}{nh^3}}.$$

Proof (Lemma 3.1)
For $x, x' \in [a, 1-a]$, the scaled covariance function of this nonparametric estimator is

$$nh\,\mathrm{Cov}\left[\hat{g}(x), \hat{g}(x')\right] = \frac{1}{h}\mathbb{E}\left[K\left(\frac{X_i - x}{h}\right)K\left(\frac{X_i - x'}{h}\right)\right] - \frac{1}{h}\mathbb{E}\left[K\left(\frac{X_i - x}{h}\right)\right]\mathbb{E}\left[K\left(\frac{X_i - x'}{h}\right)\right]$$
$$= \frac{1}{2\pi}\int_{\frac{-x}{h}}^{\frac{1-x}{h}} \exp\left(-\frac{t^2}{2}\right)\exp\left(-\frac{1}{2}\left(t + \frac{x-x'}{h}\right)^2\right)\mathrm{d}t - hI(x)I(x')$$

where $I(x) = \frac{1}{\sqrt{2\pi}}\int_{-x/h}^{(1-x)/h} e^{-t^2/2}\,\mathrm{d}t$. Completing the square and integration by substitution gives

$$nh\,\mathrm{Cov}\left[\hat{g}(x), \hat{g}(x')\right] = \frac{1}{2\pi}\exp\left(-\frac{1}{4}\left(\frac{x-x'}{h}\right)^2\right)\int_{\frac{-x-x'}{2h}}^{\frac{2-x-x'}{2h}} \exp\left(-t^2\right)\,\mathrm{d}t - hI(x)I(x').$$

Now we show that since x, x' are not too close to the boundary of $[0,1]$, the limits in the above integral can be replaced by $\pm\infty$. Note that $\frac{-x-x'}{2h} \leq \frac{-a}{h}$ and $\frac{2-x-x'}{2h} \geq \frac{a}{h}$ so

$$\int_{-\infty}^{\infty}\exp\left(-t^2\right)\,\mathrm{d}t - \int_{\frac{-x-x'}{2h}}^{\frac{2-x-x'}{2h}}\exp\left(-t^2\right)\,\mathrm{d}t \leq 2\int_{a/h}^{\infty}\exp\left(-t^2\right)\,\mathrm{d}t \leq \frac{h}{a}\exp\left(-\frac{a^2}{h^2}\right).$$

Therefore since $\int_{-\infty}^{\infty} e^{-t^2}\,\mathrm{d}t = \sqrt{\pi}$,

$$\left|nh\,\mathrm{Cov}\left[\hat{g}(x), \hat{g}(x')\right] - \frac{1}{2\sqrt{\pi}}\exp\left(-\frac{1}{4}\left(\frac{x-x'}{h}\right)^2\right) + hI(x)I(x')\right| \leq \frac{h}{2\pi a}\exp\left(-\frac{a^2}{h^2}\right).$$

Define the $N \times N$ matrix $\tilde{\Sigma}_{ij} = \frac{1}{2\sqrt{\pi}}\exp\left(-\frac{1}{4}\left(\frac{x_i - x_j}{h}\right)^2\right)$. By Baxter (1994, Proposition 2.4, Proposition 2.5 and Equation 2.10), with $\mathcal{B}_k = \left\{b \in \mathbb{R}^{\mathbb{Z}} : \sum_{i\in\mathbb{Z}} \mathbb{I}\{b_i \neq 0\} \leq k\right\}$,

$$\inf_{k\in\mathbb{N}}\inf_{b\in\mathbb{R}^k}\frac{\sum_{i=1}^k\sum_{j=1}^k b_i b_j\, e^{-\lambda(i-j)^2}}{\sum_{i=1}^k b_i^2} = \sqrt{\frac{\pi}{\lambda}}\sum_{i=-\infty}^{\infty}\exp\left(-\frac{(\pi e + 2\pi i)^2}{4\lambda}\right).$$

We use Riemann sums, noting that $\pi e + 2\pi x = 0$ at $x = -e/2 \approx -1.359$. Consider the substitutions $\mathbb{Z} \cap (-\infty, -3] \mapsto (-\infty, -2]$, $\{-2, -1\} \mapsto \{-2, -1\}$ and $\mathbb{Z} \cap [0, \infty) \mapsto [-1, \infty)$.

$$\sum_{i\in\mathbb{Z}} e^{-(\pi e + 2\pi i)^2/4\lambda} \leq \int_{-\infty}^{-2} e^{-(\pi e + 2\pi x)^2/4\lambda}\,\mathrm{d}x + e^{-(\pi e - 4\pi)^2/4\lambda} + e^{-(\pi e - 2\pi)^2/4\lambda} + \int_{-1}^{\infty} e^{-(\pi e + 2\pi x)^2/4\lambda}\,\mathrm{d}x.$$

Now use the substitution $t = \frac{\pi e + 2\pi x}{2\sqrt{\lambda}}$ and suppose $\lambda < 1$, yielding

$$\sum_{i\in\mathbb{Z}} e^{-(\pi e + 2\pi i)^2/4\lambda} \leq \frac{\sqrt{\lambda}}{\pi}\int_{-\infty}^{\frac{\pi e - 4\pi}{2\sqrt{\lambda}}} e^{-t^2}\,\mathrm{d}t + e^{-(\pi e - 4\pi)^2/4\lambda} + e^{-(\pi e - 2\pi)^2/4\lambda} + \frac{\sqrt{\lambda}}{\pi}\int_{\frac{\pi e - 2\pi}{2\sqrt{\lambda}}}^{\infty} e^{-t^2}\,\mathrm{d}t$$
$$\leq \left(1 + \frac{1}{\pi}\frac{\lambda}{4\pi - \pi e}\right)e^{-(\pi e - 4\pi)^2/4\lambda} + \left(1 + \frac{1}{\pi}\frac{\lambda}{\pi e - 2\pi}\right)e^{-(\pi e - 2\pi)^2/4\lambda}$$
$$\leq \frac{13}{12}e^{-(\pi e - 4\pi)^2/4\lambda} + \frac{8}{7}e^{-(\pi e - 2\pi)^2/4\lambda} \leq \frac{9}{4}\exp\left(-\frac{5}{4\lambda}\right).$$

Therefore

$$\inf_{k\in\mathbb{N}}\inf_{b\in\mathcal{B}_k}\frac{\sum_{i\in\mathbb{Z}}\sum_{j\in\mathbb{Z}}b_ib_j\,e^{-\lambda(i-j)^2}}{\sum_{i\in\mathbb{Z}}b_i^2}<\frac{4}{\sqrt{\lambda}}\exp\left(-\frac{5}{4\lambda}\right)<4e^{-1/\lambda}.$$

From this and since $\tilde{\Sigma}_{ij}=\frac{1}{2\sqrt{\pi}}e^{-\lambda(i-j)^2}$ with $\lambda=\frac{1}{4(N-1)^2h^2}\leq\frac{\delta^2}{h^2}$, for each h and some $\delta\leq h$,

$$\lambda_{\min}(\tilde{\Sigma})\leq 2e^{-h^2/\delta^2}.$$

Recall that

$$\left|\Sigma_{ij}-\tilde{\Sigma}_{ij}+hI(x_i)I(x_j)\right|\leq\frac{h}{2\pi a}\exp\left(-\frac{a^2}{h^2}\right).$$

Now for any positive semi-definite $N\times N$ matrices A and B and vector v we have $\lambda_{\min}(A-vv^\mathsf{T})\leq\lambda_{\min}(A)$ and $\lambda_{\min}(B)\leq\lambda_{\min}(A)+\|B-A\|_2\leq\lambda_{\min}(A)+N\|B-A\|_{\max}$. Hence with $I_i=I(x_i)$,

$$\lambda_{\min}(\Sigma)\leq\lambda_{\min}(\tilde{\Sigma}-hII^\mathsf{T})+\frac{Nh}{2\pi a}\exp\left(-\frac{a^2}{h^2}\right)\leq 2e^{-h^2/\delta^2}+\frac{h}{\pi a\delta}e^{-a^2/h^2}.$$

□

Proof (Proposition 3.1)
Let $\mathcal{F}_\delta$ be a δ-cover of $(\mathcal{F},d)$. Using a union bound, we can write

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

Part 1: bounding the error on $\mathcal{F}_\delta$
We apply Corollary 2.2 with $p=\infty$ to the martingale difference sequence $\mathcal{F}_\delta(X_i)=\big(f(X_i):f\in\mathcal{F}_\delta\big)$ which takes values in $\mathbb{R}^{|\mathcal{F}_\delta|}$. Square integrability can be assumed otherwise $\beta_\delta=\infty$. Note $\sum_{i=1}^n\mathcal{F}_\delta(X_i)=S(\mathcal{F}_\delta)$ and $\phi_\infty(\mathcal{F}_\delta)\leq\sqrt{2\log 2|\mathcal{F}_\delta|}$. Therefore there exists a conditionally Gaussian vector $T(\mathcal{F}_\delta)$ with the same covariance structure as $S(\mathcal{F}_\delta)$ conditional on $\mathcal{H}_0$ satisfying

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

Part 2: bounding the fluctuations in $S(f)$
Since $\left\|S(f)-S(f')\right\|_\psi\leq Ld(f,f')$, by Theorem 2.2.4 in van der Vaart and Wellner (1996)

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

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

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

Part 3: bounding the fluctuations in $T(f)$
By the Vorob'ev–Berkes–Philipp theorem (Dudley, 1999), $T(\mathcal{F}_\delta)$ extends to a conditionally Gaussian process $T(f)$. Firstly since $\left\|\left|T(f) - T(f')\right|\right\|_2 \leq L d(f, f')$ conditionally on $\mathcal{H}_0$, and $T(f)$ is a conditional Gaussian process, we have $\left\|T(f) - T(f')\right\|_{\psi_2} \leq 2 L d(f, f')$ conditional on $\mathcal{H}_0$ by van der Vaart and Wellner (1996, Chapter 2.2, Complement 1), where $\psi_2(x) = \exp(x^2) - 1$. Thus again by Theorem 2.2.4 in van der Vaart and Wellner (1996), again conditioning on $\mathcal{H}_0$,

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

for some universal constant $C_1 > 0$, where we used $\psi_2^{-1}(x) = \sqrt{\log(1+x)}$ and monotonicity of covering numbers. Then by Markov's inequality and the definition of the Orlicz norm,

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

Part 4: conclusion
The result follows by combining the parts, scaling t and η and enlarging constants if necessary. □

A.4 Applications to nonparametric regression

Proof (Proposition 4.1)
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 $\beta_{\infty,2}$ and $\beta_{\infty,3}$
Set $X_i = p(W_i)\varepsilon_i$ so $S = \sum_{i=1}^n X_i$ and set $\sigma_i^2 = \sigma^2(W_i)$ and $V_i = \mathrm{Var}[X_i \mid \mathcal{H}_{i-1}] = \sigma_i^2 p(W_i) p(W_i)^\mathsf{T}$. Recall from Corollary 2.2 that for $r \in \{2, 3\}$,

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

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

$$\mathbb{E}\left[\|X_i\|_2^r \|X_i\|_\infty\right] \leq \mathbb{E}\left[|\varepsilon_i|^3 \|p(W_i)\|_2^{r+1}\right] \lesssim 1.$$

For the second term, apply Lemma A.3 conditionally on $\mathcal{H}_n$ and again use $\sup_w \|p(w)\|_2 \lesssim 1$ to see

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

Putting these together yields $\beta_{\infty,2} \lesssim n\sqrt{\log 2k}$ and $\beta_{\infty,3} \lesssim n\sqrt{\log 2k}$.

Part 2: bounding Ω
Set $\Omega = \sum_{i=1}^n \big(V_i - \mathbb{E}[V_i]\big)$ as in Lemma A.8 so $\Omega = \sum_{i=1}^n \big(\sigma_i^2 p(W_i)p(W_i)^\mathsf{T} - \mathbb{E}\left[\sigma_i^2 p(W_i)p(W_i)^\mathsf{T}\right]\big)$. 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 A.6(i). Since the basis functions satisfy Assumption 3 in Cattaneo et al. (2020), Ω has a bounded number of non-zero entries in each row, and so by Jensen's inequality

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

Part 3: strong approximation
By Corollary 2.2 and the previous parts, with any sequence $R_n \to \infty$,

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

If further $\mathbb{E}\left[\varepsilon_i^3 \mid \mathcal{H}_{i-1}\right] = 0$ then the third-order version of Corollary 2.2 applies since

$$\pi_3 = \sum_{i=1}^n \sum_{|\kappa|=3} \mathbb{E}\Big[\big|\mathbb{E}[X_i^\kappa \mid \mathcal{H}_{i-1}]\big|\Big] = \sum_{i=1}^n \sum_{|\kappa|=3} \mathbb{E}\Big[\big|p(W_i)^\kappa\, \mathbb{E}[\varepsilon_i^3 \mid \mathcal{H}_{i-1}]\big|\Big] = 0,$$

giving

$$\|S - T\|_\infty \lesssim_\mathbb{P} \beta_{\infty,3}^{1/4} (\log 2k)^{3/8} R_n + \sqrt{\log 2k}\sqrt{\mathbb{E}[\|\Omega\|_2]}R_n \lesssim_\mathbb{P} (nk)^{1/4}\sqrt{\log 2k}R_n.$$

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

$$\sup_{w \in \mathcal{W}} \left|p(w)^\mathsf{T} H^{-1} S - p(w)^\mathsf{T} H^{-1} T\right| \leq \sup_{w \in \mathcal{W}} \|p(w)\|_1 \|H^{-1}\|_1 \|S - T\|_\infty \lesssim n^{-1} k \|S - T\|_\infty.$$

Part 4: convergence of $\hat{H}$
We have $\hat{H} - H = \sum_{i=1}^n \big(p(W_i)p(W_i)^\mathsf{T} - \mathbb{E}\left[p(W_i)p(W_i)^\mathsf{T}\right]\big)$. Observe that $(\hat{H} - H)_{jl}$ is the sum of a zero-mean strictly stationary α-mixing sequence and so $\mathbb{E}[(\hat{H} - H)_{jl}^2] \lesssim n$ by Lemma A.6(i). Since the basis functions satisfy Assumption 3 in Cattaneo et al. (2020), $\hat{H} - H$ has a bounded number of non-zero entries in each row and so by Jensen's inequality

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

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

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

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

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

Part 6: conclusion of the main result

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

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

since $k^3/n \to 0$. If further $\mathbb{E}\left[\varepsilon_i^3 \mid \mathcal{H}_{i-1}\right] = 0$ then

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

Finally, we verify the variance bounds for 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: bounding the bias

We delegate the task of carefully deriving bounds on the bias to Cattaneo et al. (2020), who provide a high-level assumption on the approximation error in Assumption 4 and then use it to derive bias bounds in Section 3 of the form $\sup_{w \in \mathcal{W}} |\mathrm{Bias}(w)| \lesssim_{\mathbb{P}} k^{-\gamma}$. This assumption is then verified for B-splines, wavelets and piecewise polynomials in their supplemental appendix.

□

Proof (Proposition 4.2)
Part 1: infeasible supremum approximation
Provided that the bias is negligible, for all $s > 0$ we have

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\frac{\hat{\mu}(w)-\mu(w)}{\sqrt{\rho(w,w)}}\right|\le t\right)-\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\frac{G(w)}{\sqrt{\rho(w,w)}}\right|\le t\right)\right|$$
$$\le \sup_{t\in\mathbb{R}}\mathbb{P}\left(t\le\sup_{w\in\mathcal{W}}\left|\frac{G(w)}{\sqrt{\rho(w,w)}}\right|\le t+s\right)+\mathbb{P}\left(\sup_{w\in\mathcal{W}}\left|\frac{\hat{\mu}(w)-\mu(w)-G(w)}{\sqrt{\rho(w,w)}}\right|>s\right).$$

By the Gaussian anti-concentration result given as Corollary 2.1 in Chernozhukov et al. (2014a) applied to a discretization of $\mathcal{W}$, the first term is at most $s\sqrt{\log n}$ up to a constant factor, and the second term converges to zero whenever $\frac{1}{s}\left(\frac{k^3(\log k)^3}{n}\right)^{1/6}\to 0$. Thus a suitable value of s exists whenever $\frac{k^3(\log n)^6}{n}\to 0$.

Part 2: feasible supremum approximation
By Chernozhukov et al. (2013, Lemma 3.1) and discretization, with $\rho(w,w')=\mathbb{E}[\hat{\rho}(w,w')]$,

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

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

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

where the second term is bounded by the same argument used to bound $\|\hat{H}-H\|_1$. Thus, the feasible approximation is valid whenever $(\log n)^2\sup_{w\in\mathcal{W}}|\hat{\sigma}^2(w)-\sigma^2(w)|\to_{\mathbb{P}}0$ and $\frac{k^3(\log n)^4}{n}\to 0$. The validity of the uniform confidence band follows immediately. □

Proof (Proposition 4.3)
We apply Proposition 3.1 with the metric $d(f_w,f_{w'})=\|w-w'\|_2$ and the function class

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

with ψ chosen as a suitable Bernstein Orlicz function.

Part 1: bounding $H(w)^{-1}$
Recall that $H(w) = \sum_{i=1}^n \mathbb{E}[K_h(W_i - w) p_h(W_i - w) p_h(W_i - w)^\mathsf{T}]$ and let $a(w) \in \mathbb{R}^k$ with $\|a(w)\|_2 = 1$. Since the density of W_i is bounded away from zero on $\mathcal{W}$,

$$\begin{aligned} a(w)^\mathsf{T} H(w) a(w) &= n \mathbb{E}\left[\left(a(w)^\mathsf{T} p_h(W_i - w)\right)^2 K_h(W_i - w)\right] \\ &\gtrsim n \int_{\mathcal{W}} \left(a(w)^\mathsf{T} p_h(u - w)\right)^2 K_h(u - w) \,\mathrm{d}u \gtrsim n \int_{\frac{\mathcal{W}-w}{h}} \left(a(w)^\mathsf{T} p(u)\right)^2 K(u) \,\mathrm{d}u. \end{aligned}$$

This is continuous in $a(w)$ on the compact set $\|a(w)\|_2 = 1$ and $p(u)$ forms a polynomial basis so $a(w)^\mathsf{T} p(u)$ has finitely many zeroes. Since $K(u)$ is compactly supported and $h \to 0$, the above integral is eventually strictly positive for all $x \in \mathcal{W}$, and hence is bounded below uniformly in $w \in \mathcal{W}$ by a positive constant. Therefore $\sup_{w \in \mathcal{W}} \|H(w)^{-1}\|_2 \lesssim 1/n$.

Part 2: bounding β_δ
Let $\mathcal{F}_\delta$ be a δ-cover of $(\mathcal{F}, d)$ with cardinality $|\mathcal{F}_\delta| \asymp \delta^{-m}$ and let $\mathcal{F}_\delta(W_i, \varepsilon_i) = \big(f(W_i, \varepsilon_i) : f \in \mathcal{F}_\delta\big)$. Define the truncated errors $\tilde{\varepsilon}_i = \varepsilon_i \mathbb{I}\{-a \log n \leq \varepsilon_i \leq b \log n\}$ and note that $\mathbb{E}\big[e^{|\varepsilon_i|/C_\varepsilon}\big] < \infty$ implies that $\mathbb{P}(\exists i : \tilde{\varepsilon}_i \neq \varepsilon_i) \lesssim n^{1-(a \vee b)/C_\varepsilon}$. Hence, by choosing a and b large enough, with high probability, we can replace all ε_i by $\tilde{\varepsilon}_i$. Further, it is always possible to increase either a or b along with some randomization to ensure that $\mathbb{E}[\tilde{\varepsilon}_i] = 0$. Since K is bounded and compactly supported, W_i has a bounded density and $|\tilde{\varepsilon}_i| \lesssim \log n$,

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

Therefore

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

Let $V_i(\mathcal{F}_\delta) = \mathbb{E}\big[\mathcal{F}_\delta(W_i, \tilde{\varepsilon}_i) \mathcal{F}_\delta(W_i, \tilde{\varepsilon}_i)^\mathsf{T} \mid \mathcal{H}_{i-1}\big]$ and $Z_i \sim \mathcal{N}(0, I_d)$ be i.i.d. and independent of $\mathcal{H}_n$. Note that $V_i(f, f) = \mathbb{E}[f(W_i, \tilde{\varepsilon}_i)^2 \mid W_i] \lesssim n^{-2} h^{-2m}$ and $\mathbb{E}[V_i(f, f)] = \mathbb{E}[f(W_i, \tilde{\varepsilon}_i)^2] \lesssim n^{-2} h^{-m}$. Thus by Lemma A.3,

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

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

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

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

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

Since W_i are α-mixing with $\alpha(j) < e^{-2j/C_\alpha}$, Lemma A.6(ii) with $r=3$ gives

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

Therefore, by Jensen's inequality,

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

Note that we could have used $\|\cdot\|_1$ rather than $\|\cdot\|_\mathrm{F}$, but this term is negligible either way.

Part 4: regularity of the stochastic processes
For each $f, f' \in \mathcal{F}$, define the mean-zero and α-mixing random variables

$$
u_i(f,f') = \big(e_1^\mathsf{T} H(w)^{-1} K_h(W_i-w) p_h(W_i-w) - e_1^\mathsf{T} H(w')^{-1} K_h(W_i-w') p_h(W_i-w')\big)\tilde\varepsilon_i.
$$

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

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

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

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

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

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

Hence

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

and from the penultimate line, we also deduce that

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

Further, $\mathbb{E}[u_i(f,f')u_j(f,f')] = 0$ for $i \neq j$ so by Lemma A.7(ii), for a constant $C_1 > 0$,

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

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 $\tilde{\varepsilon}_i = \varepsilon_i$ for all i with high probability, by Proposition 3.1, for all $t, \eta > 0$ there exists a zero-mean Gaussian process $T(w)$ satisfying

$$
\mathbb{E}\left[\left(\sum_{i=1}^n f_w(W_i,\varepsilon_i)\right)\left(\sum_{i=1}^n f_{w'}(W_i,\varepsilon_i)\right)\right] = \mathbb{E}\big[T(w)T(w')\big]
$$

for all $w, w' \in \mathcal{W}$ and

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

Noting $\psi(x) \geq e^{x^2/4}$ for $x \leq 4\sqrt{\log n}$ and taking any sequence $R_n \to \infty$ gives the probability bound

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

Optimizing over δ gives $\delta \asymp \left(\frac{\log n}{n h^{m-6}}\right)^{\frac{1}{2m+6}} = h \left(\frac{\log n}{n h^{3m}}\right)^{\frac{1}{2m+6}}$ and so

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

Part 6: convergence of $\hat{H}(w)$
For $1 \leq j, l \leq k$ define the zero-mean random variables

$$
u_{ijl}(w) = K_h(W_i - w) p_h(W_i - w)_j p_h(W_i - w)_l - \mathbb{E}\big[K_h(W_i - w) p_h(W_i - w)_j p_h(W_i - w)_l\big]
$$

and note that $|u_{ijl}(w)| \lesssim h^{-m}$. By Lemma A.7(i) for a constant $C_2 > 0$ and all $t > 0$,

$$
\mathbb{P}\left(\left|\sum_{i=1}^n u_{ijl}(w)\right| > C_2 h^{-m} \big(\sqrt{nt} + (\log n)(\log \log n) t\big)\right) \leq C_2 e^{-t}.
$$

Further, note that by Lipschitz properties,

$$
\left|\sum_{i=1}^n u_{ijl}(w) - \sum_{i=1}^n u_{ijl}(w')\right| \lesssim h^{-m-1} \|w - w'\|_2
$$

so there is a δ-cover of $(\mathcal{W}, \|\cdot\|_2)$ with cardinality at most $n^a \delta^{-a}$ for some $a > 0$. Adjusting C_2,

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

and hence

$$\sup_{w \in \mathcal{W}} \left| \sum_{i=1}^n u_{ijl}(w) \right| \lesssim_{\mathbb{P}} h^{-m} \sqrt{n \log n} + h^{-m} (\log n)^3 \lesssim_{\mathbb{P}} \sqrt{\frac{n \log n}{h^{2m}}}.$$

Therefore

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

Part 7: bounding the matrix term

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

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

Therefore

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

Now for $1 \leq j \leq k$ write $u_{ij}(w) = K_h(W_i - w) p_h(W_i - w)_j \tilde{\varepsilon}_i$ so that $S(w)_j = \sum_{i=1}^n u_{ij}(w)$ with high probability. Note that $u_{ij}(w)$ are zero-mean with $\mathrm{Cov}[u_{ij}(w), u_{i'j}(w)] = 0$ for $i \neq i'$. Also $|u_{ij}(w)| \lesssim h^{-m} \log n$ and $\mathrm{Var}[u_{ij}(w)] \lesssim h^{-m}$. Thus by Lemma A.7(ii) for a constant $C_3 > 0$,

$$\mathbb{P}\left(\Big| \sum_{i=1}^n u_{ij}(w) \Big| \geq C_3 \big((h^{-m/2} \sqrt{n} + h^{-m} \log n) \sqrt{t} + h^{-m} (\log n)^3 t \big) \right) \leq C_3 e^{-t},$$

$$\mathbb{P}\left(\Big| \sum_{i=1}^n u_{ij}(w) \Big| > C_3 \left(\sqrt{\frac{tn}{h^m}} + \frac{t (\log n)^3}{h^m} \right) \right) \leq C_3 e^{-t},$$

where we used $n h^m \gtrsim (\log n)^2$ 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 $a > 0$

$$\mathbb{P}\left(\sup_{w \in \mathcal{W}} \Big| \sum_{i=1}^n u_{ij}(w) \Big| > C_3 \left(\sqrt{\frac{tn}{h^m}} + \frac{t (\log n)^3}{h^m} \right) \right) \leq C_3 n^a e^{-t},$$

$$\sup_{w \in \mathcal{W}} \|S(w)\|_2 \lesssim_{\mathbb{P}} \sqrt{\frac{n \log n}{h^m}} + \frac{(\log n)^4}{h^m} \lesssim_{\mathbb{P}} \sqrt{\frac{n \log n}{h^m}}$$

as $n h^m \gtrsim (\log n)^7$. Finally

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

Part 8: bounding the bias

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

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

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

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

Therefore

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

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

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

Now write $\tilde{u}_{ij}(w) = K_h(W_i - w) p_h(W_i - w)_j$ and note that $|\tilde{u}_{ij}(w)| \lesssim h^{-m}$ and $\mathbb{E}[\tilde{u}_{ij}(w)] \lesssim 1$. By Lemma A.7(i), for a constant C_4,

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

As in previous parts, by Lipschitz properties, this implies

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

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

Part 9: conclusion
By the previous parts,

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

where the last inequality follows because $nh^{3m}\to\infty$ and $\frac{1}{2m+6}\leq\frac{1}{2}$. Finally, we verify the upper and lower bounds on the variance of the Gaussian process. Since the spectrum of $H(w)^{-1}$ is bounded above and below by $1/n$,

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

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

B High-dimensional central limit theorems for martingales

We present an application of our main results to high-dimensional central limit theorems for martingales. Our main contribution here is the generality of our results, which are broadly applicable to martingale data and impose minimal extra assumptions. In exchange for the scope and breadth of our results, we naturally do not necessarily achieve state-of-the-art distributional approximation errors in certain special cases, such as with independent data or when restricting the class of sets over which the central limit theorem must hold. Extensions of our high-dimensional central limit theorem results to mixingales and other approximate martingales, along with third-order refinements and Gaussian mixture target distributions, are possible through methods akin to those used to establish our main results in Section 2, but we omit these for succinctness.

Our approach to deriving a high-dimensional martingale central limit theorem proceeds as follows. Firstly, the upcoming Proposition B.1 uses our main result on martingale coupling (Corollary 2.2) to reduce the problem to that of providing anti-concentration results for high-dimensional Gaussian vectors. We then demonstrate the utility of this reduction by employing a few such anti-concentration methods from the existing literature. Proposition B.2 gives a feasible implementation via the Gaussian multiplier bootstrap, enabling valid resampling-based inference using the resulting conditional Gaussian distribution. Finally in Section B.1 we provide an explicit example application: distributional approximation for ℓ^p-norms of high-dimensional martingale vectors in Kolmogorov–Smirnov distance, relying on some recent results concerning Gaussian perimetric inequalities (Nazarov, 2003; Kozbur, 2021; Giessing, 2023).

We begin this section with some notation. Assume the setup of Corollary 2.2 and suppose Σ is non-random. Let $\mathcal{A}$ be a class of measurable subsets of $\mathbb{R}^d$ and take $T \sim \mathcal{N}(0, \Sigma)$. For $\eta > 0$ and $p \in [1, \infty]$ define the Gaussian perimetric quantity

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

where $A_p^\eta = \{x \in \mathbb{R}^d : \|x - A\|_p \leq \eta\}$, $A_p^{-\eta} = \mathbb{R}^d \setminus (\mathbb{R}^d \setminus A)_p^\eta$ and $\|x - A\|_p = \inf_{x' \in A} \|x - x'\|_p$. Using this perimetric term allows us to convert coupling results to central limit theorems as follows. Denote by $\Gamma_p(\eta)$ the rate of strong approximation attained in Corollary 2.2:

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

Proposition B.1 (High-dimensional central limit theorem for martingales)
Assume the setup of Corollary 2.2, with Σ non-random. For a class $\mathcal{A}$ of measurable subsets of $\mathbb{R}^d$,

$$\sup_{A \in \mathcal{A}} \big|\mathbb{P}(S \in A) - \mathbb{P}(T \in A)\big| \leq \inf_{p \in [1,\infty]} \inf_{\eta > 0} \big\{\Gamma_p(\eta) + \Delta_p(\mathcal{A}, \eta)\big\}. \tag{13}$$

Proof (Proposition B.1)
This follows from Strassen's theorem (Lemma A.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 $p \in [1, \infty]$,

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

□

The term $\Delta_p(\mathcal{A}, \eta)$ in (13) 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 $\mathcal{A} = \mathcal{C} = \{A \subseteq \mathbb{R}^d \text{ is convex}\}$, Nazarov (2003) showed

$$\Delta_2(\mathcal{C}, \eta) \asymp \eta \sqrt{\|\Sigma^{-1}\|_{\mathrm{F}}}, \tag{14}$$

whenever Σ is invertible. Then Proposition B.1 with $p = 2$ combined with (14) yields for convex sets

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

Alternatively, one can take $\mathcal{A} = \mathcal{R}$, the class of axis-aligned rectangles in $\mathbb{R}^d$. By Nazarov's Gaussian perimetric inequality (Nazarov, 2003; Chernozhukov et al., 2017),

$$\Delta_\infty(\mathcal{R}, \eta) \leq \frac{\eta(\sqrt{2 \log d} + 2)}{\sigma_{\min}} \tag{15}$$

whenever $\min_j \Sigma_{jj} \geq \sigma_{\min}^2$ for some $\sigma_{\min} > 0$. Then Proposition B.1 with $p = \infty$ and (15) yields

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

In situations where $\liminf_n \min_j \Sigma_{jj} = 0$, it may be possible in certain cases to regularize the minimum variance away from zero and then apply a Gaussian–Gaussian rectangular approximation result such as Lemma 2.1 from Chernozhukov et al. (2023).

Remark B.1 (Comparisons with the literature)
The literature on high-dimensional central limit theorems has developed rapidly in recent years (see Zhai, 2018; Koike, 2021; Buzun et al., 2022; Lopes, 2022; Chernozhukov et al., 2023, and references therein), particularly for the special case of sums of independent random vectors on the rectangular sets $\mathcal{R}$. Our corresponding results are rather weaker in terms of dependence on the dimension than for example Chernozhukov et al. (2023, Theorem 2.1). This is an inherent issue due to our approach of first considering the class of all Borel sets and only afterwards specializing to the smaller class $\mathcal{R}$, where sharper results in the literature directly target the Kolmogorov–Smirnov distance via Stein's method and Slepian interpolation.

Next, we present a version of Proposition B.1 in which the covariance matrix Σ is replaced by an estimator $\hat{\Sigma}$. This ensures that the associated conditionally Gaussian vector is feasible and can be resampled, allowing Monte Carlo quantile estimation via a Gaussian multiplier bootstrap.

Proposition B.2 (Bootstrap central limit theorem for martingales)
Assume the setup of Corollary 2.2, with Σ non-random, and let $\hat{\Sigma}$ be an $\mathbf{X}$-measurable random $d \times d$ positive semi-definite matrix, where $\mathbf{X} = (X_1, \ldots, X_n)$. For a class $\mathcal{A}$ of measurable subsets of $\mathbb{R}^d$,

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

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

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

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

The first term is bounded by Proposition B.1; the second by applying Lemma A.5 conditional on $\mathbf{X}$.

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

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. $\square$

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

In Section B.1 we apply Proposition B.1 to the special case of approximating the distribution of the ℓ^p-norm of a high-dimensional martingale. Proposition B.2 is then used to ensure that feasible distributional approximations are also available.

B.1 Application: 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 (possibly high-dimensional) martingale for some $p \in [1, \infty]$. In the notation of Corollary 2.2, it is therefore of interest to bound Kolmogorov–Smirnov quantities of the form

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

Let $\mathcal{B}_p$ be the class of closed ℓ^p-balls in $\mathbb{R}^d$ centered at the origin and set

$$\Delta_p(\eta) := \Delta_p(\mathcal{B}_p, \eta) = \sup_{t \geq 0} \mathbb{P}(t < \|T\|_p \leq t + \eta).$$

Proposition B.3 (Distributional approximation of martingale ℓ^p-norms)
Assume the setup of Corollary 2.2, with Σ non-random. Then for $T \sim \mathcal{N}(0, \Sigma)$,

$$\sup_{t \geq 0} \left| \mathbb{P}(\|S\|_p \leq t) - \mathbb{P}\left(\|T\|_p \leq t\right) \right| \leq \inf_{\eta > 0} \left\{ \Gamma_p(\eta) + \Delta_p(\eta) \right\}. \tag{16}$$

Proof (Proposition B.3)
Applying Proposition B.1 with $\mathcal{A} = \mathcal{B}_p$ gives

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

$\square$

The right-hand side of (16) can be controlled in various ways. In the case of $p = \infty$, note that ℓ^∞-balls are rectangles so $\mathcal{B}_\infty \subseteq \mathcal{R}$ and (15) applies, giving $\Delta_\infty(\eta) \leq \eta(\sqrt{2 \log d} + 2)/\sigma_{\min}$ whenever $\min_j \Sigma_{jj} \geq \sigma_{\min}^2$. Alternatively, Giessing (2023, Theorem 1) provides $\Delta_\infty(\eta) \lesssim \eta / \sqrt{\operatorname{Var}[\|T\|_\infty] + \eta^2}$. In fact, by Hölder duality of ℓ^p-norms, we can write $\|T\|_p = \sup_{\|u\|_q \leq 1} u^{\mathsf{T}} T$ where $1/p + 1/q = 1$. Then, applying the Gaussian process anti-concentration result of Giessing (2023, Theorem 2) yields the more general $\Delta_p(\eta) \lesssim \eta/\sqrt{\mathrm{Var}[\|T\|_p] + \eta^2}$. Thus, the problem can be reduced to that of bounding $\mathrm{Var}\left[\|T\|_p\right]$, with techniques for doing so discussed, for example, in Giessing (2023, Section 4). 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; Giessing, 2023).

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

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

with $\hat{\Sigma}$ any $\mathbf{X}$-measurable positive semi-definite estimator of Σ. Note that for the canonical estimator $\hat{\Sigma} = \sum_{i=1}^n X_i X_i^\mathsf{T}$ we can write $\hat{T} = \sum_{i=1}^n X_i Z_i$ with $Z_1, \ldots, Z_n$ i.i.d. standard Gaussian independent of $\mathbf{X}$, yielding the Gaussian multiplier bootstrap. Now assuming the law of $\|\hat{T}\|_p \mid \mathbf{X}$ has no atoms, we can apply Proposition B.2 to see

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

and hence the bootstrap is valid whenever $\|\hat{\Sigma}^{1/2} - \Sigma^{1/2}\|_2^2$ is sufficiently small. See the preceding discussion regarding methods for bounding this object.

Remark B.2 (One-dimensional distributional approximations)
In our application to distributional approximation of ℓ^p-norms, the object of interest $\|S\|_p$ is a one-dimensional functional of the high-dimensional martingale; contrast this with the more general Proposition B.1 which directly considers the d-dimensional random vector S. As such, our coupling-based approach may be improved in certain settings by applying a more carefully tailored smoothing argument. For example, Belloni and Oliveira (2018) employ a "log sum exponential" bound (see also Chernozhukov et al., 2013) for the maximum statistic $\max_{1 \leq j \leq d} S_j$ along with a coupling due to Chernozhukov et al. (2014b) to attain an improved dependence on the dimension. Naturally their approach does not permit the formulation of high-dimensional central limit theorems over arbitrary classes of Borel sets as in our Proposition B.1.