Submitted to the Annals of Statistics

YURINSKII'S COUPLING FOR MARTINGALES

BY MATIAS D. CATTANEO[1,a], RICARDO P. MASINI[2,b] AND WILLIAM G. UNDERWOOD[3,c]

[1]*Department of Operations Research and Financial Engineering, Princeton University,* [a]*cattaneo@princeton.edu*

[2]*Department of Statistics, University of California, Davis,* [b]*rmasini@ucdavis.edu*

[3]*Statistical Laboratory, University of Cambridge,* [c]*wgu21@cam.ac.uk*

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 verifiable 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, alongside central limit theorems for high-dimensional martingale vectors.

1. Introduction Yurinskii's coupling [53] 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 which has 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 norm, and some features of the underlying distribution. See, for example, Pollard [44, Chapter 10] for a textbook introduction, and Csörgö and Révész [22] and Lindvall [37] for background references.

When compared to other coupling approaches, such as the celebrated Hungarian construction [34] or Zaitsev's coupling [54, 55], 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 properties 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 [18]; to establish distribution theory for non-Donsker stochastic t-processes generated in nonparametric series regression [4]; to prove distributional approximations for high-dimensional ℓ^p-norms [8]; to

MSC2020 subject classifications: Primary 62E20, 62G20 ; secondary 60G42.

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

develop distribution theory for vector-valued martingales [3, 36]; to derive a law of the iterated logarithm for stochastic gradient descent optimization methods [1]; to establish uniform distributional results for nonparametric high-dimensional quantile processes [5]; to develop distribution theory for non-Donsker stochastic t-processes generated in partitioning-based series regression [11]; to deduce Bernstein–von Mises theorems in high-dimensional settings [46]; and to develop distribution theory for non-Donsker U-processes based on dyadic network data [12]. There are also many other early applications of Yurinskii's coupling: Dudley and Philipp [27] and Dehling [25] establish invariance principles for Banach space-valued random variables, and Le Cam [35] and Sheehy and Wellner [48] 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 primary 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 recent paper by Li and Liao [36] and the unpublished manuscript by Belloni and Oliveira [3], which both investigated distribution theory for martingale data using Yurinskii's coupling and related methods. Specifically, Li and Liao [36] 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 (see 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 [36], 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 (see Remark 1).

Belloni and Oliveira [3] 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 15, 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 the supplementary material [13] for details). Nevertheless, a Yurinskii coupling derived from Belloni and Oliveira [3] 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 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, making 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 that 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 [16, Theorem B.2; 41, Theorem 4], along with some related

technical work detailed in the supplementary material [13]. A third-order coupling is obtained via a modification of a standard smoothing technique for Borel sets from classical versions of Yurinskii's coupling (see Lemma SA.2 in the supplementary material [13]), 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. 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 1 a technical issue which is often neglected [44, 36] 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 [50] with our coupling from 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 [6]. Our improvement comes from using Yurinskii's coupling for the ℓ^∞-norm, where Berthet and Mason [6] apply Zaitsev's coupling [54, 55] 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 strong approximations 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. Appendix A demonstrates how our coupling results can be used to derive distributional Gaussian approximations (central limit theorems) for possibly high-dimensional martingale vectors (Proposition A.1). This result complements a recent literature on probability and statistics studying the same problem but with independent data [see 10, 38, 21, 33, and references therein]. We also present a version of this result employing a covariance estimator (Proposition A.2), enabling the construction of valid high-dimensional confidence sets via a Gaussian multiplier bootstrap.

All proofs are collected in the supplementary material [13], where we also include other technical lemmas of potential independent interest, alongside some further results on distributional approximations for ℓ^p-norms of high-dimensional martingale vectors.

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 [50, 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 $|x_n| \le C|y_n|$ 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 $\mathbb{P}(|X_n| \ge C|Y_n|) \le \varepsilon$, 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 $|\kappa| = \sum_{j=1}^d \kappa_j$, along with $\kappa! = \prod_{j=1}^d \kappa_j!$, and $x^\kappa = \prod_{j=1}^d x_j^{\kappa_j}$ for $x \in \mathbb{R}^d$.

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 [26, Theorem 1.1.10].

2. Main results We begin with our most general result: an ℓ^p-norm Yurinskii coupling for 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 many 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 Gaussian mixture couplings in the general setting. In Remark 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 $n \ge 1$, 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*

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

Let $V_i = \mathrm{Var}[\tilde{X}_i \mid \mathcal{H}_{i-1}]$ 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

$$\mathbb{P}\big(\|S-T\|_p > 6\eta\big) \le \inf_{t>0}\left\{2\mathbb{P}\big(\|Z\|_p > t\big) + \min\left\{\frac{\beta_{p,2}t^2}{\eta^3}, \frac{\beta_{p,3}t^3}{\eta^4} + \frac{\pi_3 t^3}{\eta^3}\right\}\right\}$$
$$+ \inf_{M \succeq 0}\Big\{2\mathbb{P}\big(\Omega \not\preceq M\big) + \delta_p(M,\eta) + \varepsilon_p(M,\eta)\Big\} + \mathbb{P}\big(\|U\|_p > \eta\big), \tag{1}$$

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,

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

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

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

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

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 [52], Wu [51], Dedecker, Merlevède and Volnỳ [24], Zhao and Woodroofe [56], Peligrad [43], Atchadé and Cattaneo [2], Cuny and Merlevède [23], Magda and Zhang [39], 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 $\mathbb{P}(\Omega \not\preceq M)$, $\delta_p(M,\eta)$ and $\varepsilon_p(M,\eta)$, utilizing a regularization argument through the free matrix parameter M. If a non-random Σ is used, then T is unconditionally Gaussian, and one can take $\mathcal{H}_0$ to be the trivial σ-algebra. As demonstrated in our proof, our approach to establishing a mixing approximation is different from naively taking an unconditional version of Yurinskii's coupling and applying it conditionally on $\mathcal{H}_0$, which will not deliver the same coupling as in Theorem 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 [16, Theorem B.2; 41, Theorem 4]. 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 conditional 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 quantity $\beta_{p,3}$. At this level of generality, it is not possible to obtain explicit bounds on π_3 because we make no assumptions on the relationship between the data X_i and the σ-algebras $\mathcal{H}_i$ (and therefore the variables $\tilde{X}_i$ resulting from the martingale approximation). However, if $X_1, \dots, X_n$ form martingale differences with respect to $\mathcal{H}_0, \dots, \mathcal{H}_n$, then $\tilde{X}_i = X_i$ almost surely (see Section 2.3). In this setting, assuming that $\mathbb{E}\big[X_i^\kappa \mid \mathcal{H}_{i-1}\big] = 0$ for each multi-index κ with $|\kappa| = 3$ (e.g. if the data is conditionally symmetrically distributed around zero), then using $\beta_{p,3}$ rather than $\beta_{p,2}$ can give smaller coupling approximation errors in (1). Such a refinement can be viewed as a strong approximation counterpart to classical Edgeworth expansion methods, and we illustrate this phenomenon in our upcoming applications to nonparametric inference (Section 4).

2.1. *User-friendly formulation of the 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 $\phi_p : \{1, 2, \dots\} \to \mathbb{R}$, for $p \in [0, \infty]$, by

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

With $Z \sim \mathcal{N}(0, I_d)$ and $t > 0$, these functions satisfy $\mathbb{P}(\|Z\|_p > t) \leq \mathbb{E}[\|Z\|_p]/t \leq \phi_p(d)/t$ (see Lemma SA.4 in the supplementary material [13]).

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 also

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

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 $c_1, \dots, c_n$ and $\zeta_0, \dots, \zeta_n$ such that for all $1 \leq i \leq n$ and $0 \leq r \leq i$,

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

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

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

These conditions are satisfied, for example, if X_i are integrable strongly α-mixing random variables [40], or if X_i are generated by an auto-regressive or auto-regressive moving average process (see Section 2.5), among many other possibilities [9]. Then, in the notation of Theorem 2.1, we have by Markov's inequality that

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

with $\zeta = 6 \sum_{i=1}^n c_i(\zeta_i + \zeta_{n-i+1})$. Combining Proposition 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*

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

If further $\pi_3 = 0$ *then*

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

The closest antecedent to Corollary 2.1 is found in Li and Liao [36, 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 Section 2.5 and Section 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, a carefully crafted regularization argument, and a third-order Lindeberg method. Furthermore (Remark 1), we clarify a technical issue in Li and Liao [36] 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 [3] and Li and Liao [36], who also (implicitly or explicitly) considered Yurinskii's coupling for martingales. More specifically, Li and Liao [36, 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 [3, Theorem 2.1] established a central limit theorem for vector-valued martingale sequences employing the standard second-order Lindeberg method. As such, their proof could be adapted to deduce a Yurinskii coupling for martingales with the help of a conditional version of Strassen's theorem and some additional nontrivial technical work.

Corollary 2.2 improves over this prior work as follows. With respect to Li and Liao [36], 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 [36] 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 [36] in two respects: (i) we employ a conditional version of Strassen's theorem (see Lemma SA.1 in the supplementary material [13]) to appropriately handle the conditioning arguments; and (ii) we deduce valid probability bounds for $\|S - T\|_p$, as the following Remark 1 makes clear.

REMARK 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 1 have been both addressed and neglected alike in the prior literature. Pollard [44, Chapter 10.4, Example 16] apparently misses this subtlety, providing an inaccurate bound in probability based on the Yurinskii coupling. Li and Liao [36] 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. [4] and Belloni et al. [5] 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 [36, 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 [3], 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, 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 (see the supplementary material [13]): (i) a correct application of a conditional version of Strassen's theorem (Lemma SA.1 in the supplementary material [13]); (ii) the development of a third-order Borel set smoothing technique and associated ℓ^p-norm moment control (Lemmas SA.2, SA.3, and SA.4); (iii) a careful truncation scheme to account for $\Omega \not\asymp 0$; and (iv) a valid third-order Lindeberg argument (Lemma SA.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 $X_1, \dots, X_i$ 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 [5]. By further restricting to $p = 2$, we recover the original Yurinskii coupling as presented in Le Cam [35, Theorem 1] and Pollard [44, 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 independent (idiosyncratic) disturbances $(\varepsilon_1, \ldots, \varepsilon_n)$. See Fan et al. [30], 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 each ε_i 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.

(i) *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.

(ii) *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.

(iii) *Mixingales*. Finally, assume that the factors follow the auto-regressive model $f_i = A f_{i-1} + u_i$ 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 the mixingale result from Corollary 2.1 applies 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 for each $f \in \mathcal{F}$, the variables $f(X_1), \dots, f(X_n)$ form martingale differences with respect to an appropriate filtration. We construct (conditionally) Gaussian processes $T(f)$ for which upper bounds on the uniform coupling error $\sup_{f \in \mathcal{F}} |S(f) - T(f)|$ are 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, Chetverikov and Kato [18], 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 strong approximations of entire empirical processes by sequences of Gaussian mixture processes in 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 (including mixingale empirical processes; cf. Corollary 2.1), but to reduce notation and the technical burden we do not consider this extension.

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 the supplementary material [13] 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 process $(\hat{g}(x) - \mathbb{E}[\hat{g}(x)] : x \in \mathcal{X})$ which is, upon rescaling, non-Donsker whenever the bandwidth decreases to zero in large samples. To match notation with the upcoming general result for empirical processes, set $f_x(u) = \frac{1}{n}(K_h(u-x) - \mathbb{E}[K_h(X_i - x)])$ so $S(x) := S(f_x) = \hat{g}(x) - \mathbb{E}[\hat{g}(x)]$. The next step is standard: a mesh separates the local oscillations of the processes from the finite-dimensional coupling. For $\delta \in (0, 1/2)$, set $N = \lfloor 1 + \frac{1-2a}{\delta} \rfloor$ and $\mathcal{X}_\delta = (a + (j-1)\delta : 1 \le j \le N)$. Letting $T(x)$ be the approximating stochastic process to be constructed, consider the following decomposition:

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

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 [26, 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 [50, 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 $R_n \to \infty$ (Remark 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 is then optimized to obtain the tightest possible strong approximation. In particular, since $N \lesssim 1/\delta$, setting $\delta \asymp n^{-1/8} h^{5/8} (\log n)^{-1/8}$ yields

$$\sup_{x \in \mathcal{X}} \big|S(x) - T(x)\big| \lesssim_{\mathbb{P}} \left(\frac{(\log n)^3}{n^5 h^7}\right)^{1/8} R_n$$

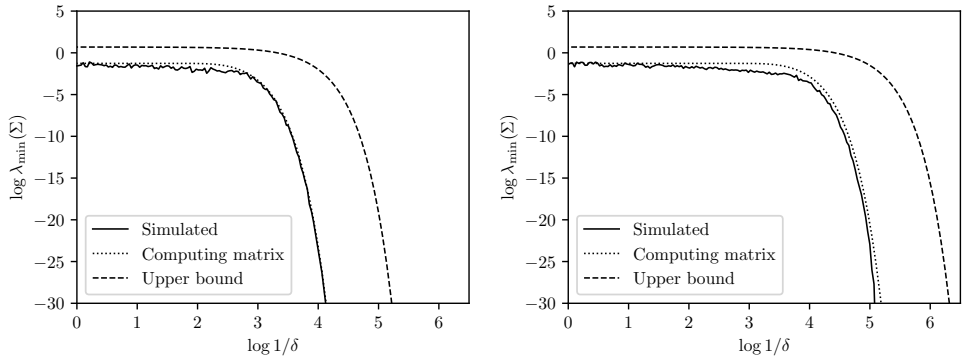
which, after standardization by $\sqrt{nh}$, vanishes whenever $R_n (\log n)^3/(nh^3) \to 0$. This is a more stringent assumption on the bandwidth h than $(\log n)/(nh) \to 0$ imposed by Giné, Koltchinskii and Sakhanenko [32] and Cattaneo and Yu [14] when employing a Hungarian construction [34], or $(\log n)^6/(nh) \to 0$ imposed by Chernozhukov, Chetverikov and Kato [18] when studying in particular the Kolmogorov–Smirnov distance between the scalar suprema. The difference in side restrictions is a result of the specific assumptions imposed and coupling approaches used; see Section 4.2 for related discussion.

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 [36], 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 $\Sigma = nh \operatorname{Var}[S(\mathcal{X}_\delta)]$ has exponentially small eigenvalues, which in turn will negatively affect the strong approximation bound if the Li and Liao [36] 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 [36] ineffective in such scenarios.



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

Fig 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 $\hat{g}(x) - g(x)$. 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 [49] and sequential symmetrization [45].

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

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

and, for real-valued Y, the Orlicz norm $\|\!|Y|\!\|_\psi = \inf\{C>0 : \mathbb{E}[\psi(|Y|/C)\le 1]\}$ as in van der Vaart and Wellner [50, Section 2.2].

PROPOSITION 3.1 (Strong approximation for martingale empirical processes). *Let X_i be random variables for $1\le i\le n$ taking values in a measurable space $\mathcal{X}$, and $\mathcal{F}$ be a pointwise measurable class of functions from $\mathcal{X}$ to $\mathbb{R}$. Let $\mathcal{H}_0,\dots,\mathcal{H}_n$ 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 $\psi\in\Psi$,*

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

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

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

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 $a\ge 2$ has $\|\!|X|\!\|_2\le\|\!|X|\!\|_\psi$ and $\sqrt{\log x}\le\sqrt{a}\psi^{-1}(x)$.
Exponential: $\psi(x)=\exp(x^a)-1$ for $a\in[1,2]$ has $\|\!|X|\!\|_2\le 2\|\!|X|\!\|_\psi$ and $\sqrt{\log x}\le\psi^{-1}(x)$.
Bernstein: $\psi(x)=\exp\Big(\Big(\frac{\sqrt{1+2ax}-1}{a}\Big)^2\Big)-1$ for $a>0$ has $\|\!|X|\!\|_2\le(1+a)\|\!|X|\!\|_\psi$ and $\sqrt{\log x}\le\psi^{-1}(x)$.

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 [50] for details. If the conditional third moments of $\mathcal{F}_\delta(X_i)$ 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 formally 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 [6], who used a Gaussian coupling due to Zaitsev [54, 55] 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 [6]; that is, $X_1, \dots, X_n$ are i.i.d. and $\sup_{\mathcal{F}} \|f\|_\infty \leq M$, with the VC-type assumption $\sup_{\mathbb{Q}} N(\varepsilon, \mathcal{F}, d_{\mathbb{Q}}) \leq c_0 \varepsilon^{-\nu_0}$ where $d_{\mathbb{Q}}(f, f')^2 = \mathbb{E}_{\mathbb{Q}}\big[(f - f')^2\big]$ for a measure $\mathbb{Q}$ on $\mathcal{X}$ and M, c_0, ν_0 are constants. Then, using uniform L^2 covering numbers rather than Orlicz norm chaining in our Proposition 3.1 gives the following. Firstly as X_i are i.i.d. we take $\Sigma(f, f') = \operatorname{Cov}[S(f), S(f')]$ so $\Omega_\delta = 0$. Let $\mathcal{F}_\delta$ be a minimal δ-cover of $(\mathcal{F}, d_{\mathbb{P}})$ with cardinality $N_\delta \lesssim \delta^{-\nu_0}$ where $\delta \to 0$. It is 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 [50] give

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

where we used the VC-type property to bound the entropy integral. So by our Proposition 3.1, for any sequence $R_n \to \infty$ (see Remark 1),

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

where we minimized over δ in the last step. Berthet and Mason [6, Proposition 1] achieved

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

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 [6, 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 the finite-dimensional result of Corollary 2.2 directly due to an intrinsic linear separability property. Secondly, we consider local polynomial estimators, this time using the stochastic process formulation in 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, Farrell and Feng [11] 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 $(W_1, \dots, W_{i+1}, \varepsilon_1, \dots, \varepsilon_i)$, $\mathbb{E}[\varepsilon_i \mid \mathcal{H}_{i-1}] = 0$ 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 [11, 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 the supplementary material [13]) 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 zero-mean 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 $(W_1, \dots, W_n)$ for illustration [9].

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

(i) $(W_i, \varepsilon_i)_{1 \leq i \leq n}$ *is strictly stationary.*
(ii) $W_1, \dots, W_n$ *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, Farrell and Feng [11].*

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

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

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

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

The core 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 $T \sim \mathcal{N}\big(0, \operatorname{Var}[S]\big)$ such that $\|S - T\|_\infty$ is small, and then setting $G(w) = p(w)^\mathsf{T} H^{-1} T$. So long as the bias can be appropriately controlled, this result allows for uniform inference procedures such as uniform confidence bands or shape specification testing. The condition $k^3/n \to 0$ is the same (up to logs) as that imposed by Cattaneo, Farrell and Feng [11] 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 [36, Theorem 1]: using the notation of our Corollary 2.2, and with any sequence $R_n \to \infty$, a valid (see Remark 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

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

Hence, for this approach to yield a valid strong approximation, the number of basis functions must satisfy $k^5/n \to 0$, a more restrictive assumption than our $k^3/n \to 0$ (up to logs). This difference is due to Li and Liao [36] 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 [19, Corollary 2.1] yield a distributional approximation for the supremum statistic whenever $k^3 (\log n)^6 / n \to 0$, giving

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

where $\rho(w,w') = \mathbb{E}[G(w)G(w')]$. Furthermore, using a Gaussian–Gaussian comparison result [17, Lemma 3.1] and anti-concentration again, it is not difficult to show (see the proof of Proposition 4.1) that with $\mathbf{W} = (W_1,\dots,W_n)$ and $\mathbf{Y} = (Y_1,\dots,Y_n)$,

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

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 as described above. 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

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

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

4.2. *Local polynomial estimators* As a second example application we consider nonparametric regression estimation with martingale data employing local polynomial methods [29]. 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 $1\le i\le n$ where W_i has compact connected support $\mathcal{W}\subseteq\mathbb{R}^m$, $\mathcal{H}_i$ is the σ-algebra generated by $(W_1,\dots,W_{i+1},\varepsilon_1,\dots,\varepsilon_i)$, $\mathbb{E}[\varepsilon_i \mid \mathcal{H}_{i-1}] = 0$, 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 $\gamma\ge 0$ 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 $0\le|\kappa|\le\gamma$, and set $p_h(w) = p(w/h)$. The local polynomial regression estimator of $\mu(w)$ is, with $e_1 = (1,0,\dots,0)^{\mathsf T}\in\mathbb{R}^k$,

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

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

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

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 $\sup_{w\in\mathcal{W}} \|\hat{H}(w) - H(w)\|_2 \lesssim_{\mathbb{P}} \sqrt{nh^{-2m}\log n}$. 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 $\left(e_1^{\mathsf{T}} H(w)^{-1} S(w) : w \in \mathcal{W}\right)$ via Proposition 3.1 by setting

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

With this approach, we obtain the following result.

PROPOSITION 4.3 (Strong approximation for local polynomial estimators). *Under the nonparametric regression setup described above, assume further that*

(i) $(W_i, \varepsilon_i)_{1\le i\le n}$ *is strictly stationary.*
(ii) $(W_i, \varepsilon_i)_{1\le i\le n}$ *is α-mixing with mixing coefficients $\alpha(j) \le e^{-2j/C_\alpha}$ for some $C_\alpha > 0$.*
(iii) W_i *has a Lebesgue density on $\mathcal{W}$ which is bounded above and away from zero.*
(iv) $\mathbb{E}\left[e^{|\varepsilon_i|/C_\varepsilon}\right] < \infty$ *for $C_\varepsilon > 0$ and $\mathbb{E}\left[\varepsilon_i^2 \mid \mathcal{H}_{i-1}\right] = \sigma^2(W_i)$ is bounded away from zero.*
(v) K *is a non-negative Lipschitz compactly supported kernel with $\int K(w)\,\mathrm{d}w < \infty$.*

Then for any $R_n \to \infty$, there is a zero-mean Gaussian process $T(w)$ 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 [28]. 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, Chetverikov and Kato [19], 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.

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 [36] is unsuited for this application due to its strong minimum eigenvalue assumption. Finally, for the special case of i.i.d. data, Chernozhukov, Chetverikov and Kato [18, Remark 3.1] achieve better rates for approximating the scalar supremum of the t-process in Kolmogorov–Smirnov distance by bypassing the step where we first approximate the entire stochastic process (see Section 3 for a discussion), while Cattaneo and Yu [14] obtain better strong approximations for the entire stochastic process under additional assumptions via a generalization of the celebrated Hungarian construction [34, 47].

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.

APPENDIX A: HIGH-DIMENSIONAL CENTRAL LIMIT THEOREMS FOR MARTINGALES

We present an application of our main results to central limit theorems for high-dimensional martingale vectors. Our main contribution in this section is found in 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 results to mixingales and other approximate martingales, along with third-order refinements and Gaussian mixture coupling 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 A.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 A.2 gives a feasible implementation via the Gaussian multiplier bootstrap, enabling valid resampling-based inference using the resulting conditional Gaussian distribution. In the supplementary material [13] we provide an example application: distributional approximation for ℓ^p-norms

of high-dimensional martingale vectors in Kolmogorov–Smirnov distance, relying on some recent results concerning Gaussian perimetric inequalities [see 42, 31, and references therein].

We begin 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 (anti-concentration) quantity

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

with $A_p^{\eta} = \{x \in \mathbb{R}^d : \|x - A\|_p \leq \eta\}$, $A_p^{-\eta} = \mathbb{R}^d \setminus (\mathbb{R}^d \setminus A)_p^{\eta}$ and $\|x - A\|_p = \inf_{x' \in A} \|x - x'\|_p$. This perimetric term allows one 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 A.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 $\mathbb{R}^d$,*

$$\sup_{A\in\mathcal{A}} \big| \mathbb{P}(S \in A) - \mathbb{P}(T \in A) \big| \leq \inf_{p\in[1,\infty]} \inf_{\eta>0} \big\{ \Gamma_p(\eta) + \Delta_p(\mathcal{A},\eta) \big\}. \tag{8}$$

PROOF (Proposition A.1). This follows directly from Strassen's theorem; see Lemma SA.1 in the supplementary materials [13]. □

The term $\Delta_p(\mathcal{A},\eta)$ in (8) depends on the law of S only through the covariance matrix Σ, and can be bounded using a selection of different results from the literature. For instance, with $\mathcal{A} = \mathcal{C} = \{A \subseteq \mathbb{R}^d \text{ is convex}\}$, Nazarov [42] showed

$$\Delta_2(\mathcal{C},\eta) \asymp \eta \sqrt{\|\Sigma^{-1}\|_{\mathrm{F}}}, \tag{9}$$

if Σ is invertible. Then Proposition A.1 with $p = 2$ combined with (9) 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, with $\mathcal{A} = \mathcal{R}$, the set of axis-aligned rectangles in $\mathbb{R}^d$, Nazarov [42, 20] gives

$$\Delta_\infty(\mathcal{R},\eta) \leq \frac{\eta(\sqrt{2\log d} + 2)}{\sigma_{\min}} \tag{10}$$

whenever $\min_j \Sigma_{jj} \geq \sigma_{\min}^2 > 0$. Proposition A.1 with $p = \infty$ and (10) then yields

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

In situations where $\liminf_n \min_j \Sigma_{jj} = 0$, it may be possible in certain cases to regularize the minimum variance away from zero and then apply a Gaussian–Gaussian rectangular approximation result such as Lemma 2.1 from Chernozhukov, Chetverikov and Koike [21]; we delegate this to future work.

REMARK 2 (Comparisons with the literature). The literature on high-dimensional central limit theorems has developed rapidly in recent years [see 10, 38, 21, 33, and references therein], particularly for the special case of sums of independent random vectors on rectangular sets

$\mathcal{R}$. As a consequence, the results in this appendix are weaker in terms of dependence on the dimension than those available in the literature. 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 $\mathcal{R}$. In contrast, sharper results in the literature, for example, directly target the Kolmogorov–Smirnov distance via Stein's method and Slepian interpolation. The main contribution of this section is therefore to obtain Gaussian distributional approximations for high-dimensional martingale vectors, a setting in which alternative proof strategies are not available.

As our final main result, we present a version of Proposition A.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 A.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 $\mathbf{X} = (X_1, \dots, X_n)$. For a class $\mathcal{A}$ of measurable subsets of $\mathbb{R}^d$,*

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

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

PROOF (Proposition A.2). Since $\Sigma^{1/2} Z$ is independent of $\mathbf{X}$, we have $\big|\mathbb{P}(S \in A) - \mathbb{P}\big(\hat{\Sigma}^{1/2} Z \in A \;\big|\; \mathbf{X}\big)\big| \leq \big|\mathbb{P}(S \in A) - \mathbb{P}\big(\Sigma^{1/2} Z \in A\big)\big| + \big|\mathbb{P}\big(\Sigma^{1/2} Z \in A\big) - \mathbb{P}\big(\hat{\Sigma}^{1/2} Z \in A \;\big|\; \mathbf{X}\big)\big|$. The first term is bounded by Proposition A.1; the second by Lemma SA.5 in [13] conditional on $\mathbf{X}$. Taking a supremum over A and infima over p and η yields the result. □

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 A.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 [7] gives $\big\|\hat{\Sigma}^{1/2} - \Sigma^{1/2}\big\|_2 \leq \big\|\hat{\Sigma} - \Sigma\big\|_2^{1/2}$ and Problem X.5.5 in the same gives $\big\|\hat{\Sigma}^{1/2} - \Sigma^{1/2}\big\|_2 \leq \big\|\Sigma^{-1/2}\big\|_2 \big\|\hat{\Sigma} - \Sigma\big\|_2$ when Σ is invertible. The latter often gives a tighter bound when the minimum eigenvalue of Σ can be bounded away from zero, and consistency of $\hat{\Sigma}$ can typically be established using a range of matrix concentration inequalities.

In the supplementary material [13] we apply Proposition A.1 to the special case of approximating the distribution of the ℓ^p-norm of a high-dimensional martingale. Proposition A.2 is then used to ensure that feasible distributional approximations are also available.

Acknowledgments We thank the Editor, Associate Editor, and several reviewers for their comments, which led to a much improved version of this paper. We also thank Jianqing Fan, Alexander Giessing, Boris Hanin, Michael Jansson, Jason Klusowski, Arun Kumar, Boris Shigida, and Rae Yu 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 Science Foundation through grant SES-2241575 and from the National Institute of Health through grant R01 GM072611-16.

SUPPLEMENTARY MATERIAL

Proofs of main results and additional technical material

The supplementary material [13] contains detailed proofs, along with results on martingale ℓ^p-norm approximations and technical lemmas which may be of independent interest.

REFERENCES

[1] 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* 115–137. PMLR.

[2] 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** 646–677.

[3] BELLONI, A. and OLIVEIRA, R. I. (2018). A high dimensional central limit theorem for martingales, with applications to context tree models. arXiv:1809.02741.

[4] 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** 345–366.

[5] BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and FERNÁNDEZ-VAL, I. (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics* **213** 4–29.

[6] BERTHET, P. and MASON, D. M. (2006). Revisiting two strong approximation results of Dudley and Philipp. *Lecture Notes–Monograph Series* 155–172.

[7] BHATIA, R. (1997). *Matrix Analysis* **169**. Springer, New York.

[8] BIAU, G. and MASON, D. M. (2015). High-dimensional p-norms. In *Mathematical Statistics and Limit Theorems* 21–40. Springer.

[9] BRADLEY, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys* **2** 107–144.

[10] BUZUN, N., SHVETSOV, N. and DYLOV, D. V. (2022). Strong Gaussian approximation for the sum of random vectors. In *Conference on Learning Theory* **178** 1693–1715. PMLR.

[11] CATTANEO, M. D., FARRELL, M. H. and FENG, Y. (2020). Large sample properties of partitioning-based series estimators. *Annals of Statistics* **48** 1718–1741.

[12] 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.

[13] CATTANEO, M. D., MASINI, R. P. and UNDERWOOD, W. G. (2024). Supplement to "Yurinskii's coupling for martingales".

[14] CATTANEO, M. D. and YU, R. R. (2024). Strong Approximations for Empirical Processes Indexed by Lipschitz Functions. arXiv:arXiv:2406.04191.

[15] CHATTERJEE, S. (2006). A generalization of the Lindeberg principle. *Annals of Probability* **34** 2061–2076.

[16] 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** 1097–1163.

[17] 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** 2786–2819.

[18] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014a). Gaussian approximation of suprema of empirical processes. *Annals of Statistics* **42** 1564–1597.

[19] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014b). Anti-concentration and honest, adaptive confidence bands. *Annals of Statistics* **42** 1787–1818.

[20] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2017). Central limit theorems and bootstrap in high dimensions. *Annals of Probability* **45** 2309–2352.

[21] CHERNOZHUKOV, V., CHETVERIKOV, D. and KOIKE, Y. (2023). Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *Annals of Applied Probability* **33** 2374–2425.

[22] CSÖRGÖ, M. and RÉVÉSZ, P. (1981). *Strong Approximations in Probability and Statistics*. *Probability and Mathematical Statistics: a series of monographs and textbooks*. Academic Press.

[23] CUNY, C. and MERLEVÈDE, F. (2014). On martingale approximations and the quenched weak invariance principle. *Annals of Probability* **42** 760–793.

[24] DEDECKER, J., MERLEVÈDE, F. and VOLNÝ, D. (2007). On the weak invariance principle for non-adapted sequences under projective criteria. *Journal of Theoretical Probability* **20** 971–1004.

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

[26] DUDLEY, R. M. (1999). *Uniform Central Limit Theorems*. *Cambridge Studies in Advanced Mathematics*. Cambridge University Press.

[27] DUDLEY, R. M. and PHILIPP, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **62** 509–552.

[28] EGGERMONT, P. P. B. and LARICCIA, V. N. (2009). *Maximum Penalized Likelihood Estimation: Volume II: Regression*. Springer.

[29] FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC.

[30] FAN, J., LI, R., ZHANG, C.-H. and ZOU, H. (2020). *Statistical Foundations of Data Science*. CRC Press.

[31] GIESSING, A. (2023). Anti-concentration of suprema of Gaussian processes and Gaussian order statistics. arXiv:2310.12119.

[32] GINÉ, E., KOLTCHINSKII, V. and SAKHANENKO, L. (2004). Kernel density estimators: convergence in distribution for weighted sup-norms. *Probability Theory and Related Fields* **130** 167–198.

[33] KOCK, A. B. and PREINERSTORFER, D. (2024). A remark on moment-dependent phase transitions in high-dimensional Gaussian approximations. *Statistics and Probability Letters* **211** 110149.

[34] 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** 111–131.

[35] LE CAM, L. (1988). On the Prokhorov distance between the empirical process and the associated Gaussian bridge. *Technical report*.

[36] LI, J. and LIAO, Z. (2020). Uniform nonparametric inference for time series. *Journal of Econometrics* **219** 38-51.

[37] LINDVALL, T. (1992). *Lectures on the Coupling Method*. Dover Publications, New York.

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

[39] MAGDA, P. and ZHANG, N. (2018). Martingale approximations for random fields. *Electronic Communications in Probability* **23** 1–9.

[40] MCLEISH, D. L. (1975). Invariance principles for dependent variables. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **32** 165–178.

[41] MONRAD, D. and PHILIPP, W. (1991). Nearby variables with nearby conditional laws and a strong approximation theorem for Hilbert space valued martingales. *Probability Theory and Related Fields* **88** 381–404.

[42] 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* 169–187. Springer.

[43] PELIGRAD, M. (2010). Conditional central limit theorem via martingale approximation. In *Dependence in Probability, Analysis and Number Theory, volume in memory of Walter Philipp* 295–311. Kendrick Press.

[44] POLLARD, D. (2002). *A User's Guide to Measure Theoretic Probability*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

[45] RAKHLIN, A., SRIDHARAN, K. and TEWARI, A. (2015). Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields* **161** 111–153.

[46] RAY, K. and VAN DER VAART, A. (2021). On the Bernstein–von Mises theorem for the Dirichlet process. *Electronic Journal of Statistics* **15** 2224–2246.

[47] RIO, E. (1994). Local invariance principles and their application to density estimation. *Probability Theory and Related Fields* **98** 21–45.

[48] SHEEHY, A. and WELLNER, J. A. (1992). Uniform Donsker classes of functions. *Annals of Probability* **20** 1983–2030.

[49] VAN DE GEER, S. A. (2000). *Empirical Processes in M-estimation* **6**. Cambridge University Press.

[50] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. *Springer Series in Statistics*. Springer, New York.

[51] WU, W. B. (2005). Nonlinear system theory: another look at dependence. *Proceedings of the National Academy of Sciences* **102** 14150–14154.

[52] WU, W. B. and WOODROOFE, M. (2004). Martingale approximations for sums of stationary processes. *Annals of Probability* **32** 1674–1690.

[53] YURINSKII, V. V. (1978). On the error of the Gaussian approximation for convolutions. *Theory of Probability & its Applications* **22** 236–247.

[54] 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** 203–220.

[55] 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** 535–566.

[56] ZHAO, O. and WOODROOFE, M. (2008). On martingale approximations. *Annals of Applied Probability* **18** 1831–1847.