Strong Approximations for Empirical Processes Indexed by Lipschitz Functions

Matias D. Cattaneo[1]   Ruiqi (Rae) Yu[1*]

June 6, 2024

Abstract

This paper presents new uniform Gaussian strong approximations for empirical processes indexed by classes of functions based on d-variate random vectors ($d \geq 1$). First, a uniform Gaussian strong approximation is established for general empirical processes indexed by Lipschitz functions, encompassing and improving on all previous results in the literature. When specialized to the setting considered by Rio (1994), and certain constraints on the function class hold, our result improves the approximation rate $n^{-1/(2d)}$ to $n^{-1/\max\{d,2\}}$, up to the same $\operatorname{polylog} n$ term, where n denotes the sample size. Remarkably, we establish a valid uniform Gaussian strong approximation at the optimal rate $n^{-1/2}\log n$ for $d = 2$, which was previously known to be valid only for univariate ($d = 1$) empirical processes via the celebrated Hungarian construction (Komlós *et al.*, 1975). Second, a uniform Gaussian strong approximation is established for a class of multiplicative separable empirical processes indexed by Lipschitz functions, which address some outstanding problems in the literature (Chernozhukov *et al.*, 2014, Section 3). In addition, two other uniform Gaussian strong approximation results are presented for settings where the function class takes the form of a sequence of Haar basis based on generalized quasi-uniform partitions. We demonstrate the improvements and usefulness of our new strong approximation results with several statistical applications to nonparametric density and regression estimation.

Keywords: empirical processes, coupling, Gaussian approximation, uniform inference, local empirical process, nonparametric regression.

[1] Department of Operations Research and Financial Engineering, Princeton University

[*] Corresponding author: rae.yu@princeton.edu

Contents

1 Introduction

Let $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$, $i = 1, 2, \dots, n$, be independent and identical distributed (i.i.d.) random vectors supported on a background probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The classical empirical process is

$$X_n(h) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(h(\mathbf{x}_i) - \mathbb{E}[h(\mathbf{x}_i)] \big), \qquad h \in \mathcal{H}, \tag{1}$$

where $\mathcal{H}$ is a (possibly n-varying) class of functions. Following the empirical process literature, and assuming $\mathcal{H}$ is "nice", the stochastic process $(X_n(h) : h \in \mathcal{H})$ is said to be Donsker if it converges (as $n \to \infty$) weakly to a Gaussian process in $\ell^\infty(\mathcal{H})$, the space uniformly bounded real functions on $\mathcal{H}$. This convergence in law result is typically denoted by

$$X_n \rightsquigarrow Z, \qquad \text{in } \ell^\infty(\mathcal{H}), \tag{2}$$

where $(Z(h) : h \in \mathcal{H})$ is a mean-zero Gaussian process with covariance function $\mathbb{E}[Z(h_1)Z(h_2)] = \mathbb{E}[h_1(\mathbf{x}_i)h_2(\mathbf{x}_i)] - \mathbb{E}[h_1(\mathbf{x}_i)]\mathbb{E}[h_2(\mathbf{x}_i)]$ for all $h_1, h_2 \in \mathcal{H}$ when $\mathcal{H}$ is not n-varying. See van der Vaart and Wellner (2013) and Giné and Nickl (2016) for textbook reviews.

A more challenging endeavour is to construct a uniform Gaussian strong approximation for the empirical process X_n. That is, if the background probability space is "rich" enough, or is otherwise properly enlarged, the goal is to construct a sequence of mean-zero Gaussian processes $(Z_n(h) : h \in \mathcal{H})$ with the same covariance structure as X_n (i.e., $\mathbb{E}[X_n(h_1)X_n(h_2)] = \mathbb{E}[Z_n(h_1)Z_n(h_2)]$ for all $h_1, h_2 \in \mathcal{H}$) such that

$$\|X_n - Z_n\|_{\mathcal{H}} := \sup_{h \in \mathcal{H}} \big| X_n(h) - Z_n(h) \big| = O(\varrho_n) \qquad \text{almost surely (a.s.)}, \tag{3}$$

for a non-random sequence $\varrho_n \to 0$ as $n \to \infty$. Such a refined approximation result is useful in a variety of contexts. For example, it gives a distributional approximation for non-Donsker empirical processes, for which (2) does not hold, and it also offers a precise quantification of the quality of the distributional approximation when (2) holds. In addition, (3) is typically obtained from precise probability concentration inequalities that can be used to construct statistical inference procedures requiring uniformity over $\mathcal{H}$ and/or the class of underlying data generating processes. Furthermore, because the sequence of Gaussian processes Z_n are "pre-asymptotic", they can offer better finite sample approximations to the sampling distribution of X_n when compared to the large sample approximation based on the limiting Gaussian process Z as in (2).

There is a large literature on strong approximations for empirical processes, offering different tightness levels for the bound ϱ_n in (3). In particular, the univariate case ($d = 1$) is mostly settled. A major breakthrough was accomplished by Komlós *et al.* (1975, KMT hereafter), who introduced the celebrated Hungarian construction to prove the optimal result $\varrho_n = n^{-1/2} \log n$ for the special case of the uniform empirical distribution process: $\mathcal{X} = [0, 1]$, $\mathbf{x}_i \sim \mathsf{Uniform}(\mathcal{X})$, and $\mathcal{H} = \{\mathbb{1}(\cdot \leq x) : x \in [0, 1]\}$, where $\mathbb{1}(\cdot)$ denotes the indicator function. See Bretagnolle and

Massart (1989) and Mason and Van Zwet (2011) for more technical discussions on the Hungarian construction, and Csörgó and Revész (1981) and Pollard (2002) for textbook introductions. The KMT result was later extended by Giné *et al.* (2004) and Giné and Nickl (2010) to univariate empirical processes indexed by functions with uniformly bounded total variation: for $\mathcal{X} = \mathbb{R}$ and $\mathbf{x}_i \sim \mathbb{P}_X$ continuously distributed, the authors obtained

$$\varrho_n = n^{-1/2} \log n, \tag{4}$$

in (3), with $\mathcal{H}$ satisfying a bounded variation condition (see Remark 2 below for details). More recently, Cattaneo *et al.* (2024b, Lemma SA26 in their supplemental appendix) gave a self-contained proof of a slightly generalized KMT result allowing for a larger class of distributions $\mathbb{P}_X$. As a statistical application, Giné *et al.* (2004) and Giné and Nickl (2010) considered univariate kernel density estimation with bandwidth $b \to 0$ as $n \to \infty$, and demonstrated that the optimal univariate KMT strong approximation rate $(nb)^{-1/2} \log n$ is achievable, where nb is the effective sample size.

Establishing strong approximations for general empirical processes with $d \geq 2$ is substantially more difficult, since the KMT approach does not easily generalize to multivariate data. Foundational results in the multidimensional context include Massart (1989), Koltchinskii (1994), and Rio (1994). In particular, assuming the function class $\mathcal{H}$ is uniformly bounded, has bounded total variation, and satisfies a VC-type condition, among other regularity conditions discussed precisely in the upcoming sections, Rio (1994) obtained

$$\varrho_n = n^{-1/(2d)} \sqrt{\log n}, \qquad d \geq 2, \tag{5}$$

in (3). This result is tight under the conditions imposed (Beck, 1985), and demonstrates an unfortunate dimension penalty in the convergence rate for d-variate uniform Gaussian strong approximation. As a statistical application, Rio (1994) also considered the kernel density estimator with bandwidth $b \to 0$ as $n \to \infty$, and established (3) with

$$\varrho_n = (nb^d)^{-1/(2d)} \sqrt{\log n}, \qquad d \geq 2,$$

where nb^d is the effective sample size.

While Rio (1994)'s KMT strong approximation result is unimprovable under the conditions he imposed, it has two limitations:

(1) The class of functions $\mathcal{H}$ may be too large, and further restrictions can open the door for improvements. For example, in his application to kernel density estimation, Rio (1994, Section 4) assumed that the class $\mathcal{H}$ is Lipschitzian to verify the sufficient conditions of his strong approximation theorem, but his theorem did not exploit the Lipschitz property in itself. (The Lipschitzian assumption is essentially without loss of generality in the kernel density estimation application.) It is an open question whether the optimal univariate KMT strong approximation rate (4) is achievable when $d \geq 2$, under additional restrictions on $\mathcal{H}$ (e.g., Lipschitz continuity).

(2) As discussed by Chernozhukov *et al.* (2014, Section 3), applying Rio (1994)'s strong approximation result directly to nonparametric local smoothing regression, a "local empirical process" in their terminology, leads to an even more suboptimal strong approximation rate in (3). For example, in the case of kernel regression estimation with d-dimensional covariates, Rio (1994)'s strong approximation would treat all $d+1$ variables (covariates and outcome) symmetrically, and thus it will give a strong approximation rate in (3) of the form

$$\varrho_n = (nb^{d+1})^{-1/(2d+2)}\sqrt{\log n}, \qquad d \geq 1, \tag{6}$$

where $b \to 0$ as $n \to \infty$, and under standard regular conditions. The main takeaway is that the resulting effective sample size is now nb^{d+1} when in reality it should be nb^d, since only the d-dimensional covariates are smoothed out for estimation of the conditional expectation. It is this unfortunate fact that prompted Chernozhukov *et al.* (2014) to developed strong approximation methods that target the scalar suprema of the stochastic process, $\sup_{h\in\mathcal{H}}|X_n(h)|$, instead of the stochastic process itself, $(X_n(h) : h \in \mathcal{H})$, as a way to circumvent the suboptimal strong approximation rates that would emerge from deploying directly results in the literature.

This paper presents new uniform Gaussian strong approximation results for empirical processes that address the two aforementioned limitations. To begin, Section 3 studies the general empirical process (1), and presents two main results. Theorem 1 establishes a uniform Gaussian strong approximation explicitly allowing for the possibility that $\mathcal{H}$ is Lipschitzian. This result not only encompasses, but also generalizes all previous results in the literature by allowing for $d \geq 1$ under more generic entropy conditions. For comparison, if we impose the regularity conditions in Rio (1994) and also assume $\mathcal{H}$ is Lipschitzian, then our result (Corollary 2) verifies (3) with

$$\varrho_n = n^{-1/d}\sqrt{\log n} + n^{-1/2}\log n, \qquad d \geq 1,$$

thereby substantially improving (5), in addition to matching (4) when $d = 1$; see Remark 2 for details. Remarkably, we demonstrate that the optimal univariate KMT strong approximation rate $n^{-1/2}\log n$ is achievable when $d = 2$, in addition to achieving the better approximation rate $n^{-1/d}\sqrt{\log n}$ when $d \geq 3$. For example, applying our result to the kernel density estimation example, we obtain the improved strong approximation rate $(nb^d)^{-1/d}\sqrt{\log n} + (nb^d)^{-1/2}\log n$, $d \geq 1$, under the same conditions imposed in prior literature. We thus show that the optimal univariate KMT uniform Gaussian strong approximation holds in (3) for bivariate kernel density estimation. Theorem 1 also considers other entropy notions for $\mathcal{H}$ beyond the classical VC-type condition, which allows us to demonstrate improvements over Koltchinskii (1994); see Remark 3 for details.

Section 3 also discusses how our rate improvements are achieved, and outlines the outstanding roadblocks in our proof strategy, which prevents us from achieving the univariate KMT uniform Gaussian strong approximation for the general empirical process (1) with $d \geq 3$. In essence, and following Rio (1994) and others, our proof first approximate in mean square the class of functions $\mathcal{H}$ using a Haar basis over carefully constructed disjoint dyadic cells, and then applies the celebrated

Tusnády's Lemma (Pollard, 2002, Chapter 10, for a textbook introduction) to construct a strong approximation. Thus, our proof requires balancing two approximation errors: (i) a "bias" error emerging from the mean square projection based on a Haar basis, and (ii) a "variance" error emerging from the coupling construction for the projected process. A key observation in our paper is that both errors can be improved by explicitly exploiting a Lipschitz assumption on $\mathcal{H}$. However, it appears that to achieve the univariate KMT uniform Gaussian strong approximation for the general empirical process (1) with $d \geq 3$, a mean square projection based on a higher-order function class would be needed, for which there are no coupling methods available in the literature.

As a way to circumvent the technical limitations underlying the proof strategy of Theorem 1, Section 3 also presents Theorem 2. This second main theorem establishes a uniform Gaussian strong approximation under the assumption that $\mathcal{H}$ is spanned by a possibly increasing sequence of finite Haar basis based on generic quasi-uniform cells. This theorem shuts down the projection error, and also relies on a generalized Tusnády's Lemma proven in the supplemental appendix, to establish a valid coupling over more general partitioning schemes. In this specialized setting, we demonstrate that a uniform Gaussian strong approximation at the optimal univariate KMT rate based on the corresponding effective sample size is possible for all $d \geq 1$ under certain regularity conditions. As a statistical application in this special setting, we consider the classical multivariate histogram density estimator. Furthermore, the ideas underlying Theorem 2 provide the basis for analyzing certain nonparametric regression estimation procedures based on tree or partitioning-based regression methods.

Section 4 is devoted to addressing the second aforementioned limitation in prior uniform Gaussian strong approximation results. Specifically, that section focuses on the following *residual-based empirical process*:

$$R_n(g,r) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(g(\mathbf{x}_i) r(y_i) - \mathbb{E}[g(\mathbf{x}_i) r(y_i) | \mathbf{x}_i] \big), \qquad (g,r) \in \mathcal{G} \times \mathcal{R}, \tag{7}$$

where our terminology reflects the fact that $g(\mathbf{x}_i)r(y_i) - \mathbb{E}[g(\mathbf{x}_i)r(y_i)|\mathbf{x}_i] = g(\mathbf{x}_i)\epsilon_i(r)$ with $\epsilon_i(r) := r(y_i) - \mathbb{E}[r(y_i)|\mathbf{x}_i]$, which can be interpreted as a residual in nonparametric local smoothing regression settings. In statistical applications, $g(\cdot)$ is typically a local smoother based on kernel, series, or nearest-neighbor methods, while $r(\cdot)$ is some transformation of interest such as $r(y) = y$ for conditional mean estimation or $r(y) = \mathbb{1}(y \leq \cdot)$ for conditional distribution estimation. Chernozhukov *et al.* (2014, Section 3.1) call these special cases of R_n a "local empirical process".

The residual-based empirical process $(R_n(g,r) : (g,r) \in \mathcal{G} \times \mathcal{R})$ may be viewed as a general empirical process (1) based on independent sample $(\mathbf{z}_i = (\mathbf{x}_i, y_i) : 1 \leq i \leq n)$, and thus available strong approximation results can be applied directly, including Rio (1994) and our new Theorem 1. However, those off-the-shelf results require over-stringent assumption and can deliver sub-optimal approximation rates. First, available results require $\mathbf{z}_i$ to admit a positive Lebesgue density on $[0,1]^{d+1}$, possibly after some transformation that is bounded with bounded total variation, thereby imposing strong restrictions on the marginal distribution of y_i. Second, available results can lead to the incorrect effective sample size for the strong approximation rate. For example, for a local empirical process where g denotes local smoothing weights such as a kernel function with bandwidth $b \to 0$ as $n \to \infty$, and $r(y) = y$, Rio (1994) gives the approximation rate (6), and our refined Theorem 1 for general empirical processes indexed by Lipschitz functions gives a uniform Gaussian strong approximation rate

$$\varrho_n = (nb^{d+1})^{-1/(d+1)} \sqrt{\log n} + (nb^d)^{-1/2} \log n, \tag{8}$$

where the effective sample size is still nb^{d+1}. This is necessarily suboptimal because the (pointwise) effective sample size for the local (kernel) regression estimator is nb^d.

A key observation underlying the potential sub-optimality of strong approximation results for local regression empirical processes is that all components of $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ are treated symmetrically. More precisely, as explained previously, the Gaussian strong approximation error balances a "bias" part, which captures the error made in project functions to piecewise constant on carefully chosen cells, and a "variance" part, which is the Gaussian strong approximation error for empirical process indexed by projected functions. Results for general empirical processes treat all coordinates of $\mathcal{H} = \mathcal{G} \times \mathcal{R}$ symmetrically, despite the fact that in certain statistical applications, such as nonparametric smoothing regression, $\mathcal{G}$ and $\mathcal{R}$ are distinctively different. For example, in the kernel regression case, $\mathcal{G}$ is an n-varying class of functions (via the bandwidth b) with envelope proportional to $b^{-d/2}$, a Lipschitz constant proportional to $b^{-d/2-1}$, and complexity measures depending on b and n as well, while $\mathcal{R}$ may be a singleton or otherwise have complexity independent of n. Therefore, a design of cells for projection and coupling that is asymmetric in the direction of $\mathbf{x}_i$ and y_i components may improve the uniform Gaussian strong approximation.

Theorem 3 in Section 4 presents a novel uniform Gaussian strong approximation for the residual-based empirical process $(R_n(g, r) : (g, r) \in \mathcal{G} \times \mathcal{R})$, which explicitly exploits the multiplicative separability of $\mathcal{H} = \mathcal{G} \times \mathcal{R}$ and the Lipschitz continuity of the function class $\mathcal{G}$, while also removing the over-stringent assumptions imposed on the distribution y_i. When applied to local regression smoothing empirical processes, our result gives a uniform Gaussian strong approximation rate of

$$\varrho_n = (nb^d)^{-1/(d+2)} \sqrt{\log n} + (nb^d)^{-1/2} \log n, \tag{9}$$

thereby improving over both Rio (1994) leading to (5), and Theorem 1 leading to (8). In Section 4.1, we leverage Theorem 3 and present a substantive statistical application establishing the best known uniform Gaussian strong approximation result for local polynomial regression estimators (Fan and Gijbels, 1996). It follows that our results offer a strong approximation rate with the correct effective sample size nb^d under substantially weaker conditions on the underlying data generating process and function index set $\mathcal{H} = \mathcal{G} \times \mathcal{R}$.

In general, however, neither Theorem 1 in Section 3 nor Theorem 3 in Section 4 dominates each other, and therefore both are of interest depending on the statistical problem under consideration. Furthermore, building on the ideas underlying Theorem 2, Section 4 also presents Theorem 4 where

$\mathcal{G}$ is further assumed to be spanned by a possibly increasing sequence of Haar basis based on generic quasi-uniform cells, while $\mathcal{R}$ is an arbitrary function class satisfying some mild regularity conditions. Remarkably, we are able to adapt our proof strategy to leverage the multiplicative structure of the residual-based empirical process $(R_n(g,r) : (g,r) \in \mathcal{G} \times \mathcal{R})$ in such a way that we establish a uniform Gaussian strong approximation at the optimal univariate KMT rate based on the effective sample size for all $d \geq 1$, up to a $\operatorname{polylog} n$ term, where $\operatorname{polylog} n := \log^{\kappa}(n)$ for some $\kappa > 0$, and an additional "bias" term reflecting exclusively the projection error associated with $\mathcal{R}$, which is zero when $\mathcal{R}$ is a singleton. As a substantive statistical application of our last main result Theorem 4, we establish a valid, optimal (up to a $\operatorname{polylog} n$ term) uniform Gaussian strong approximation for a large class of Haar partitioning-based regression estimators such as certain regression trees and related methods (Breiman *et al.*, 1984; Huang, 2003; Cattaneo *et al.*, 2020).

1.1 Related Literature

This paper contributes to the literature on strong approximations for empirical processes, and their applications to uniform inference for nonparametric smoothing methods. For foundational introductions and overviews, see Csörgó and Revész (1981), Einmahl and Mason (1998), Berthet and Mason (2006), Mason and Zhou (2012), Giné and Nickl (2016), Pollard (2002), Zaitsev (2013), and references therein. See also Chernozhukov *et al.* (2014, Section 3) for discussion and further references concerning local empirical processes and their role in nonparametric curve estimation.

The celebrated KMT construction (Komlós *et al.*, 1975), Yurinskii's coupling (Yurinskii, 1978), and Zaitsev's coupling (Zaitsev, 1987) are three well-known approaches that can be used for constructing uniform Gaussian strong approximations for empirical processes. Among them, the KMT approach often offers the tightest approximation rates when applicable, and is the focus of our paper: closely related literature includes Massart (1989), Koltchinskii (1994), Rio (1994), Giné *et al.* (2004), and Giné and Nickl (2010), among others. As summarized in the introduction, our main first result (Theorem 1) encompasses and substantially improves on all prior results in that literature. Furthermore, Theorems 2, 3, and 4 offer new results for more specific settings of interest in statistics, in particular addressing some outstanding problems in the statistical literature (Chernozhukov *et al.*, 2014, Section 3). We provide detailed comparisons to the prior literature in the upcoming sections.

We do not discuss the other coupling approaches because they deliver slower strong approximation rates under the assumptions imposed in this paper: see Cattaneo *et al.* (2024d) for results based on Yurinskii's coupling, and Settati (2009) for results based on Zaitsev's coupling. Finally, employing a different approach, Dedecker *et al.* (2014) obtain a uniform Gaussian strong approximations for the multivariate empirical process indexed by half plane indicators with a dimension-independent approximation rate, up to $\operatorname{polylog} n$ terms.

2 Notation and Main Definitions

We employ standard notations from the empirical process literature, suitably modified and specialized to improve exposition. See, for example, van der Vaart and Wellner (2013) and Giné and Nickl (2016) for background definitions and more details.

Sets. Suppose $\mathcal{U}$ and $\mathcal{V}$ are subsets of $\mathbb{R}^d$. $\mathfrak{m}(\mathcal{U})$ denotes the Lebesgue measure of $\mathcal{U}$, and $\mathcal{U}+\mathcal{V} := \{\mathbf{x}+\mathbf{y} : \mathbf{x} \in \mathcal{U}, \mathbf{y} \in \mathcal{V}\}$. Suppose $\mathcal{G}$ and $\mathcal{R}$ are sets of functions from measure space $(S, \mathcal{S})$ to $\mathbb{R}$ and $(T, \mathcal{T})$ to $\mathbb{R}$, respectively. Then $\mathcal{G} \times \mathcal{R}$ denotes $\{(g, r) : (S \times T, \mathcal{S} \otimes \mathcal{T}) \to \mathbb{R}, g \in \mathcal{G}, r \in \mathcal{R}\}$, where $\mathcal{S} \otimes \mathcal{R}$ denotes the product σ-algebra on $S \times T$. Denote $\|\mathcal{U}\|_\infty := \sup\{\|\mathbf{x}-\mathbf{y}\|_\infty : \mathbf{x}, \mathbf{y} \in \mathcal{U}\}$.

Norms. For vectors, $\|\cdot\|$ denotes the Euclidean norm and $\|\cdot\|_\infty$ denotes the supremum norm. For a real-valued random variable X, $\|X\|_p = \mathbb{E}[|X|^p]^{\frac{1}{p}}$ for $1 \leq p < \infty$. For $\alpha > 0$, $\|X\|_{\psi_\alpha} = \min\{\lambda > 0 : \mathbb{E}[\exp((|X|/\lambda)^\alpha)] \leq 2\}$. For a real-valued function g defined on a measure space $(S, \mathcal{S}, Q)$, define $Qg := \int g dQ$ and define $\|g\|_{Q,p} := (Q|g|^p)^{1/p}$ for $1 \leq p < \infty$, $\|g\|_\infty := \sup_{\mathbf{x} \in \mathcal{S}} |g(\mathbf{x})|$. In the case that $S \subseteq \mathbb{R}^l$ for some $l \in \mathbb{N}$, define $\|g\|_{\mathrm{Lip}} := \sup_{\mathbf{x},\mathbf{x}' \in \mathcal{S}} |g(\mathbf{x}) - g(\mathbf{x}')|/\|\mathbf{x}-\mathbf{x}'\|_\infty$. $\mathcal{L}^p(Q)$ is the class of all measurable functions g from S to $\mathbb{R}$ such that $\|g\|_{Q,p} < \infty$, $1 \leq p < \infty$. For $\alpha > 0$, define the C^α-norm of a real valued function on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ by $\|f\|_{C^\alpha} = \max_{|k| \leq \lfloor \alpha \rfloor} \sup_{\mathbf{x}} |D^k f(\mathbf{x})| + \max_{|k|=\alpha} \sup_{\mathbf{x} \neq \mathbf{y}} \frac{|D^k f(\mathbf{x}) - D^k f(\mathbf{y})|}{\|\mathbf{x}-\mathbf{y}\|_2^{\alpha - \lfloor \alpha \rfloor}}$. e_Q and ρ_Q are the semi-metrics on $\mathcal{L}^2(Q)$ such that $e_Q(f, g) = \|f - g\|_{Q,2}$ and $\rho_Q(f, g) = \sqrt{\|f-g\|_{Q,2}^2 - (Qf - Qg)^2}$. For a class of measurable functions $\mathcal{F} \subseteq \mathcal{L}^2(Q)$, $C(\mathcal{F}, \rho_{\mathbb{P}})$ is the class of all continuous functionals in $(\mathcal{F}, \rho_{\mathbb{P}})$.

Asymptotics. For reals sequences $|a_n| = o(|b_n|)$ if $\limsup \frac{a_n}{b_n} = 0$, $|a_n| \lesssim |b_n|$ if there exists some constant C and $N > 0$ such that $n > N$ implies $|a_n| \leq C|b_n|$. $|a_n| \lesssim_\alpha |b_n|$ if there exists some constant C_α and N_α only depending on α such that $|a_n| \leq C_\alpha b_n$ for all $n \geq N_\alpha$. For sequences of random variables $a_n = o_{\mathbb{P}}(b_n)$ if $\operatorname{plim}_{n\to\infty} \frac{a_n}{b_n} = 0$, $|a_n| \lesssim_{\mathbb{P}} |b_n|$ if $\limsup_{M\to\infty} \limsup_{n\to\infty} P[|\frac{a_n}{b_n}| \geq M] = 0$.

Empirical Processes. Let $(\mathcal{S}, d)$ be a semi-metric space. The covering number $N(\mathcal{S}, d, \varepsilon)$ is the minimal number of balls $B_s(\varepsilon) := \{t : d(t, s) < \varepsilon\}$ needed to cover $\mathcal{S}$. A $\mathbb{P}$-*Brownian bridge* is a centered Gaussian random function $W_n(f), f \in L_2(\mathcal{X}, \mathbb{P})$ with the covariance $\mathbb{E}[W_{\mathbb{P}}(f)W_{\mathbb{P}}(g)] = \mathbb{P}(fg) - \mathbb{P}(f)\mathbb{P}(g)$, for $f, g \in L_2(\mathcal{X}, \mathbb{P})$. A class $\mathcal{F} \subseteq L_2(\mathcal{X}, \mathbb{P})$ is $\mathbb{P}$-*pregaussian* if there is a version of $\mathbb{P}$-Brownian bridge $W_{\mathbb{P}}$ such that $W_{\mathbb{P}} \in C(\mathcal{F}; \rho_{\mathbb{P}})$ almost surely.

2.1 Main Definitions

Let $\mathcal{F}$ be a class of measurable functions from a measure space $(S, \mathcal{S}, \mu)$ to $\mathbb{R}$, $S \subseteq \mathbb{R}^q$ for some $q \in \mathbb{N}$. We first introduce several definitions that capture different properties of $\mathcal{F}$.

Definition 1. *$\mathcal{F}$ is* pointwise measurable *if it contains a countable subset $\mathcal{G}$ such that for any $f \in \mathcal{F}$, there exists a sequence $(g_m : m \geq 1) \subseteq \mathcal{G}$ such that $\lim_{m\to\infty} g_m(x) = f(x)$ for all $x \in S$.*

Definition 2. *For any $\mathcal{C} \in \mathcal{S}$ that is non-empty, the uniform total variation of $\mathcal{F}$ over $\mathcal{C}$ is*

$$\mathtt{TV}_{\mathcal{F},\mathcal{C}} = \sup_{f \in \mathcal{F}} \sup_{\phi \in \mathcal{D}_q(\mathcal{C})} \int f(\mathbf{x}) \operatorname{div}(\phi)(\mathbf{x}) d\mathbf{x} / \|\|\phi\|_2\|_\infty,$$

where $\mathcal{D}_q(\mathcal{C})$ denote the space of C^∞ functions from $\mathbb{R}^q$ to $\mathbb{R}^q$ with compact support in $\mathcal{C}$. To save notation, we set $\mathtt{TV}_{\mathcal{F}} = \mathtt{TV}_{\mathcal{F},\mathbb{R}^q}$.

Definition 3. *The local uniform total variation constant of $\mathcal{F}$ restricted to a subset of S, $\mathcal{D} \in \mathcal{S}$, is a positive number $\mathtt{K}_{\mathcal{F}}$ such that for any cube $\mathcal{C}$ that is a subset of $\mathcal{D}$ with edges of length ℓ parallel to the coordinate axises,*

$$\mathtt{TV}_{\mathcal{F},\mathcal{C}} \leq \mathtt{K}_{\mathcal{F},\mathcal{D}}\ell^{d-1}.$$

To save notation, we set $\mathtt{K}_{\mathcal{F}} = \mathtt{K}_{\mathcal{F},\mathbb{R}^q}$.

Definition 4. *The envelopes of the class $\mathcal{F}$ are*

$$\mathtt{M}_{\mathcal{F}} = \|M_{\mathcal{F}}\|_\infty, \qquad M_{\mathcal{F}}(\mathbf{x}) = \sup_{f\in\mathcal{F}} |f(\mathbf{x})|, \qquad \mathbf{x} \in \mathcal{S}.$$

Note that in the case that $\mathcal{F}$ is pointwise measurable, $M_{\mathcal{F}}$ is measurable.

Definition 5. *The Lipschitz constant for the class $\mathcal{F}$ is*

$$\mathtt{L}_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \sup_{\mathbf{x},\mathbf{x}'\in S} \frac{|f(\mathbf{x}) - f(\mathbf{x}')|}{\|\mathbf{x} - \mathbf{x}'\|_\infty} = \sup_{f\in\mathcal{F}} \|f\|_{\mathrm{Lip}},$$

Definition 6. *The uniform entropy integral for the class $\mathcal{F}$ is*

$$J(\delta, \mathcal{F}, M_{\mathcal{F}}) = \int_0^\delta \sup_Q \sqrt{1 + \log N(\mathcal{F}, e_Q, \varepsilon\|M_{\mathcal{F}}\|_{Q,2})} d\varepsilon,$$

where the supremum is taken over all finite discrete measures on $(S, \mathcal{S})$. Here we assume that $M_{\mathcal{F}}(\mathbf{x})$ is finite for every $\mathbf{x} \in S$.

Definition 7. *The uniform covering number of the class $\mathcal{F}$ is*

$$\mathtt{N}_{\mathcal{F}}(\delta) := \sup_{\mathcal{Q}} N(\mathcal{F}, e_Q, \delta\|M_{\mathcal{F}}\|_{\mathcal{Q},2}), \quad \delta \in (0, \infty),$$

where the supremum is taken over all finite discrete measures on $(S, \mathcal{S})$. Here we assume that $M_{\mathcal{F}}(\mathbf{x})$ is finite for every $\mathbf{x} \in S$.

Definition 8. *$\mathcal{F}$ is a VC-type class with envelope $M_{\mathcal{F}}$ if (i) $M_{\mathcal{F}}$ is measurable and $M_{\mathcal{F}}(\mathbf{x})$ is finite for every $\mathbf{x} \in S$, and (ii) there exists some positive constants $\mathtt{c}_{\mathcal{F}}$ and $\mathtt{d}_{\mathcal{F}}$ such that for all $0 < \varepsilon < 1$*

$$\sup_Q N(\mathcal{F}, e_Q, \varepsilon\|M_{\mathcal{F}}\|_{Q,2}) \leq \mathtt{c}_{\mathcal{F}}\varepsilon^{-\mathtt{d}_{\mathcal{F}}},$$

where the supremum is taken over all finite discrete measures on $(S, \mathcal{S})$.

Definition 9. *$\mathcal{F}$ is a Polynomial-entropy class with envelope $M_{\mathcal{F}}$ if (i) $M_{\mathcal{F}}$ is measurable and $M_{\mathcal{F}}(\mathbf{x})$ is finite for every $\mathbf{x} \in S$, and (ii) there exists some positive constants $\mathtt{a}_{\mathcal{F}}$ and $\mathtt{b}_{\mathcal{F}} < 2$ such that for all $0 < \varepsilon < 1$*

$$\log \sup_{Q} N(\mathcal{F}, e_Q, \varepsilon \|M_{\mathcal{F}}\|_{Q,2}) \leq \mathtt{a}_{\mathcal{F}} \varepsilon^{-\mathtt{b}_{\mathcal{F}}},$$

where the supremum is taken over all finite discrete measures on $(S, \mathcal{S})$.

Definition 10. *The uniform L_1 bound for the class $\mathcal{F}$ is*

$$\mathtt{E}_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \int_S |f| d\mu.$$

3 General Empirical Process

This section presents improved, in some cases optimal, strong approximations for the general empirical process $(X_n(h) : h \in \mathcal{H})$ defined in (1). We impose the following assumption on the underlying data generation.

Assumption A. *$(\mathbf{x}_i : 1 \leq i \leq n)$ are i.i.d. random vectors taking values in $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with $\mathcal{X}$ compact, and their common law $\mathbb{P}_X$ admits a Lebesgue density f_X continuous and positive on $\mathcal{X}$.*

The next theorem gives our first main strong approximation result. Let

$$\mathsf{c}_1 = \frac{\overline{f}_X^2}{\underline{f}_X}, \qquad \mathsf{c}_2 = \frac{\overline{f}_X}{\underline{f}_X} \qquad \text{and} \qquad \mathsf{c}_3 = (2\sqrt{d})^{d-1} \frac{\overline{f}_X^{d+1}}{\underline{f}_X^d}.$$

where $\overline{f}_X := \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x})$ and $\underline{f}_X := \inf_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x})$, and

$$\mathsf{m}_{n,d} := \begin{cases} n^{-1/2}\sqrt{\log n} & \text{if } d = 1 \\ n^{-1/(2d)} & \text{if } d \geq 2 \end{cases} \qquad \text{and} \qquad \mathsf{l}_{n,d} := \begin{cases} 1 & \text{if } d = 1 \\ n^{-1/2}\sqrt{\log n} & \text{if } d = 2 \\ n^{-1/d} & \text{if } d \geq 3 \end{cases}.$$

Theorem 1. *Suppose Assumption A holds with $\mathcal{X} = [0,1]^d$, and $\mathcal{H}$ is a class of real-valued pointwise measurable functions on $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P}_X)$ such that $\mathtt{M}_{\mathcal{H}} < \infty$ and $J(1, \mathcal{H}, \mathtt{M}_{\mathcal{H}}) < \infty$. Then, on a possibly enlarged probability space, there exists a sequence of mean-zero Gaussian processes $(Z_n^X(h) : h \in \mathcal{H})$ with almost sure continuous trajectories such that:*

- *$\mathbb{E}[X_n(h_1)X_n(h_2)] = \mathbb{E}[Z_n^X(h_1)Z_n^X(h_2)]$ for all $h_1, h_2 \in \mathcal{H}$, and*
- *$\mathbb{P}\big[\|X_n - Z_n^X\|_{\mathcal{H}} > C_1 \mathsf{S}_n(t)\big] \leq C_2 e^{-t}$ for all $t > 0$,*

where C_1 and C_2 are universal constants, and

$$\mathsf{S}_n(t) = \min_{\delta \in (0,1)} \{\mathsf{A}_n(t, \delta) + \mathsf{F}_n(t, \delta)\},$$

with

$$\begin{aligned}\mathsf{A}_n(t,\delta) &:= \min\left\{\mathsf{m}_{n,d}\sqrt{\mathtt{M}_{\mathcal{H}}}, \mathsf{l}_{n,d}\sqrt{\mathsf{c}_2\mathtt{L}_{\mathcal{H}}}\right\}\sqrt{d\mathsf{c}_1\mathtt{TV}_{\mathcal{H}}}\sqrt{t+\log \mathtt{N}_{\mathcal{H}}(\delta)} \\ &\quad + n^{-1/2}\min\left\{\sqrt{\log n}\sqrt{\mathtt{M}_{\mathcal{H}}}, \sqrt{d^3\mathsf{c}_3\mathtt{K}_{\mathcal{H}}}\right\}\sqrt{\mathtt{M}_{\mathcal{H}}}(t+\log \mathtt{N}_{\mathcal{H}}(\delta))\end{aligned}$$

and

$$\mathsf{F}_n(t,\delta) := J(\delta,\mathcal{H},\mathtt{M}_{\mathcal{H}})\mathtt{M}_{\mathcal{H}} + \frac{\mathtt{M}_{\mathcal{H}}J^2(\delta,\mathcal{H},\mathtt{M}_{\mathcal{H}})}{\delta^2\sqrt{n}} + \delta\mathtt{M}_{\mathcal{H}}\sqrt{t} + \frac{\mathtt{M}_{\mathcal{H}}}{\sqrt{n}}t.$$

This theorem on uniform Gaussian strong approximation is given in full generality to accommodate different applications. Section 3.1 below discusses leading special cases, and compares our results to prior literature. The proof of Theorem 1 is in Section SA-II of the supplemental appendix, but we briefly outline the general proof strategy here to highlight our improvements on prior literature and some open questions. The proof begins with the standard "discretization" or "meshing" decomposition:

$$\|X_n - Z_n^X\|_{\mathcal{H}} \leq \|X_n - X_n \circ \pi_{\mathcal{H}_\delta}\|_{\mathcal{H}} + \|X_n - Z_n^X\|_{\mathcal{H}_\delta} + \|Z_n^X \circ \pi_{\mathcal{H}_\delta} - Z_n^X\|_{\mathcal{H}},$$

where $\|X_n - Z_n^X\|_{\mathcal{H}_\delta}$ captures the coupling between the empirical process and the Gaussian process on a δ-net of $\mathcal{H}$, which is denoted by $\mathcal{H}_\delta$, while the terms $\|X_n - X_n \circ \pi_{\mathcal{H}_\delta}\|_{\mathcal{H}}$ and $\|Z_n^X \circ \pi_{\mathcal{H}_\delta} - Z_n^X\|_{\mathcal{H}}$ capture the "fluctuations" or "ocillation" relative to the meshing for each of the stochastic processes. The latter two errors are handled using standard empirical process results, which give the contribution $\mathsf{F}(\delta)$ emerging from Talagrand's inequality (Giné and Nickl, 2016, Theorem 3.3.9) combined with a standard maximal inequality (Chernozhukov *et al.*, 2014, Theorem 5.2). See Section SA-II.3 of the supplemental appendix for details.

Following Rio (1994), the "coupling" term $\|X_n - Z_n^X\|_{\mathcal{H}_\delta}$ is further decomposed using a mean square projection onto a Haar function space:

$$\|X_n - Z_n^X\|_{\mathcal{H}_\delta} \leq \|X_n - \Pi_0 X_n\|_{\mathcal{H}_\delta} + \|\Pi_0 X_n - \Pi_0 Z_n^X\|_{\mathcal{H}_\delta} + \|\Pi_0 Z_n^X - Z_n^X\|_{\mathcal{H}_\delta}, \tag{10}$$

where $\Pi_0 X_n(h) = X_n \circ \Pi_0 h$ with Π_0 the L_2 projection from $L_2([0,1]^d)$ to piecewise constant functions on a carefully chosen partition of $\mathcal{X}$. Section SA-II.1 introduces a class of recursive *quasi-dyadic* cells expansions of $\mathcal{X}$, which we employ to generalize prior results in the literature. Section SA-II.2 then describes the properties of the L_2 projection onto a Haar basis based on quasi-dyadic cells.

The term $\|\Pi_0 X_n - \Pi_0 Z_n^X\|_{\mathcal{H}_\delta}$ in (10) represents the strong approximation error for the projected process over a recursive dyadic collection of cells partitioning $\mathcal{X}$. Handling this error boils down to the coupling of $\mathsf{Bin}(n, \frac{1}{2})$ with $\mathsf{N}(\frac{n}{2}, \frac{n}{4})$, due to the fact that the constant approximation within each recursive partitioning cell generates count data. Building on the celebrated Tusnády's Lemma, Rio (1994, Theorem 2.1) established a remarkable coupling result for bounded functions L_2-projected on a dyadic cells expansion of $\mathcal{X}$. Our Lemma SA.10 builds on his powerful ideas, and establishes an analogous result for the case of Lipschitz functions L_2-projected on dyadic cells expansions of $\mathcal{X}$, thereby obtaining a tighter coupling error. A limitation of these results is that they only apply to a dyadic cell expansion due to the specifics of Tusnády's Lemma. Section 3.2 below discusses this limitation further, and presents some generalized results, which are further exploited in Section 4.

The terms $\|X_n - \Pi_0 X_n\|_{\mathcal{H}_\delta}$ and $\|\Pi_0 Z_n^X - Z_n^X\|_{\mathcal{H}_\delta}$ in (10) represent the L_2 projection errors onto a Haar basis based on *quasi-dyadic* cells expansion of $\mathcal{X}$. Lemma SA.9 handles this error using Bernstein inequality, taking into account explicitly the potential Lipschitz structure of the functions and the generic cell structure. Balancing these approximation errors with that of $\|\Pi_0 X_n - \Pi_0 Z_n^X\|_{\mathcal{H}_\delta}$ gives term $\mathsf{A}_n(t,\delta)$ in Theorem 1. Section SA-II of the supplemental appendix provides all technical details, and some additional results that may be of independent theoretical interest.

Theorem 1 restricts the data to be continuously distributed on the d-dimensional unit cube, a normalized tensor product of compact intervals. This restriction simplifies our proof because we employ the Rosenblatt transform (Lemma SA.12) to account for general distributions supported on $\mathcal{X} = [0,1]^d$. However, as the next remark discusses, the support restriction and the other assumptions in Theorem 1 can be weakened in certain cases.

Remark 1. Theorem 1 imposes Assumption A with $\mathcal{X} = [0,1]^d$, but these restrictions can be relaxed as follows.

Univariate case. When $d = 1$, we can remove all the restrictions on the distribution of $\mathbf{x}_i$ in Assumption A and allow for $\mathcal{X} = \mathbb{R}$, by directly applying the Rosenblatt transform so that $u_i = F_X(x_i) \sim \mathsf{Uniform}[0,1]$ i.i.d., $i = 1, 2, \ldots, n$, where $F_X(x) := \mathbb{P}_X[x_i \le x]$. It follows that $X_n(h) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (h \circ F_X^{-1})(u_i) - \mathbb{E}[(h \circ F_X^{-1})(u_i)]$. Then, $\widetilde{\mathcal{H}} = \{h \circ F_X^{-1} : h \in \mathcal{H}\}$ is pointwise measurable because $\mathcal{H}$ is assumed to be so, $\mathtt{M}_{\widetilde{\mathcal{H}}} = \mathtt{M}_{\mathcal{H}}$, $\mathtt{TV}_{\widetilde{\mathcal{H}}} = \mathtt{TV}_{\mathcal{H}}$, $J(\widetilde{\mathcal{H}}, H, \delta) = J(\mathcal{H}, H, \delta)$, and Theorem 1 holds with $\mathtt{L}_{\mathcal{H}} = \infty$ and $\mathtt{c}_1 = \mathtt{c}_2 = \mathtt{c}_3 = 1$. A similar argument can be found in Giné *et al.* (2004, Section 2) and in Cattaneo *et al.* (2024b, Lemma SA20). See Remark 2 below for related discussion.

Multivariate case. When $d > 1$, the support restriction $\mathcal{X} = [0,1]^d$ in Assumption A can be relaxed by assuming that there exists a diffeomorphism $\chi : \mathcal{X} \mapsto [0,1]^d$. In this case our results continue to hold with $\mathtt{c}_1$, $\mathtt{c}_2$ and $\mathtt{c}_3$ replaced by, respectively,

$$\mathtt{c}_1 = \frac{\overline{f}_X^2}{\underline{f}_X} \mathtt{S}_\chi, \qquad \mathtt{c}_2 = \frac{\overline{f}_X}{\underline{f}_X} \mathtt{S}_\chi, \qquad \text{and} \qquad \mathtt{c}_3 = (2\sqrt{d})^{d-1} \frac{\overline{f}_X^{d+1}}{\underline{f}_X^d} \mathtt{S}_\chi^d,$$

where $\mathtt{S}_\chi = \frac{\sup_{\mathbf{x} \in [0,1]^d} |\det(\nabla \chi^{-1}(\mathbf{x}))|}{\inf_{\mathbf{x} \in [0,1]^d} |\det(\nabla \chi^{-1}(\mathbf{x}))|} \| \|\nabla \chi^{-1}\|_2 \|_\infty$ with $\nabla \chi^{-1}(\mathbf{x})$ denoting the Jacobian of $\chi^{-1}(\mathbf{x})$, the inverse function of $\chi(\mathbf{x})$, and $\det(\cdot)$ denoting the determinant of its argument. □

The previous remark can be illustrated as follows. Suppose $(\mathbf{x}_i : 1 \le i \le n)$ are i.i.d. $\mathsf{Uniform}(\mathcal{X})$ with $\mathcal{X} = \times_{l=1}^d [a_l, b_l]$. Then, the Rosenblatt transform (Lemma SA.12) gives $\chi(x_1, \cdots, x_d) = ((b_1 - a_1)^{-1}(x_1 - a_1), \cdots, (b_d - a_d)^{-1}(x_d - a_d))$, $\mathtt{S}_\chi = \max_{1 \le l \le d} |b_l - a_l|$, $\mathtt{c}_1 = \max_{1 \le l \le d} |b_l - a_l| \prod_{l=1}^d |b_l - a_l|^{-1}$, $\mathtt{c}_2 = \max_{1 \le l \le d} |b_l - a_l|$ and $\mathtt{c}_3 = (2\sqrt{d})^{d-1} \max_{1 \le l \le d} |b_l - a_l|^d \prod_{l=1}^d |b_l - a_l|^{-1}$. Then, when $d = 1$, we have $\mathtt{TV}_{\widetilde{\mathcal{H}}} = \mathtt{TV}_{\mathcal{H}}$. However, when $d > 1$, $\mathtt{TV}_{\widetilde{\mathcal{H}}}$ is strictly greater than $\mathtt{TV}_{\mathcal{H}}$. This example illustrates the dimension penalty implied by the Rosenblatt transform when $d > 1$.

3.1 Special Cases and Related Literature

Theorem 1 can be specialized to several useful particular cases, which can be employed to compare our main results with prior literature. To this end, we introduce our first statistical example.

Example 1 (Kernel Density Estimation)**.** The classical kernel density estimator of $f_X(\mathbf{x})$ is

$$\widehat{f}_X(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{b^d} K\Big(\frac{\mathbf{x}_i - \mathbf{x}}{b}\Big),$$

where $K : \mathbb{R}^d \to \mathbb{R}$ be a compact supported continuous function such that $\int_{\mathbb{R}^d} K(\mathbf{x})d\mathbf{x} = 1$. In statistical applications, the bandwidth $b \to 0$ as $n \to \infty$ to enable nonparametric estimation (Wand and Jones, 1995). Consider establishing a strong approximation for the "localized" empirical process $(\xi_n(\mathbf{x}) : \mathbf{x} \in \mathcal{X})$, where

$$\xi_n(\mathbf{x}) := \sqrt{nb^d}\big(\widehat{f}_X(\mathbf{x}) - \mathbb{E}[\widehat{f}_X(\mathbf{x})]\big) = X_n(h), \qquad h \in \mathcal{H},$$

with $\mathcal{H} = \{b^{-d/2}K((\cdot - \mathbf{x})/b) : \mathbf{x} \in \mathcal{X}\}$. It follows that $\mathtt{M}_{\mathcal{H}} \lesssim b^{-d/2}$. ▲

Variants of Example 1 have been discussed extensively in prior literature because the process ξ_n is non-Donsker whenever $b \to 0$, and hence standard weak convergence results for empirical processes can not be used. For example, Giné *et al.* (2004) and Giné and Nickl (2010) established strong approximations for the univariate case ($d = 1$) under i.i.d. sampling with $\mathcal{X}$ unbounded, Cattaneo *et al.* (2024c) established strong approximations for the univariate case ($d = 1$) under i.i.d. sampling with $\mathcal{X}$ compact, Rio (1994) established strong approximations for the multivariate case ($d > 1$) under i.i.d. sampling with $\mathcal{X}$ compact, Sakhanenko (2015) established strong approximations for the multivariate case ($d > 1$) under i.i.d. sampling with $\mathcal{X}$ unbounded, and Cattaneo *et al.* (2024b) established strong approximations for the univariate case ($d = 1$) under non-i.i.d. dyadic data with $\mathcal{X}$ compact. Chernozhukov *et al.* (2014, Remark 3.1) provides further discussion and references. See also Cattaneo *et al.* (2024a) for an application of Rio (1994) to uniform inference for conditional density estimation.

3.1.1 VC-type Bounded Functions

Our first corollary considers a VC-type class $\mathcal{H}$ (Definition 8) of uniformly bounded functions ($\mathtt{M}_{\mathcal{H}} < \infty$), but without assuming they are Lipschitz functions ($\mathtt{L}_{\mathcal{H}} = \infty$).

Corollary 1 (VC-type Bounded Functions)**.** *Suppose the conditions of Theorem 1 hold. In addition, assume that $\mathcal{H}$ is a VC-type class with respect to envelope function $\mathtt{M}_{\mathcal{H}}$ with constant $\mathtt{c}_{\mathcal{H}} \geq e$ and exponent $\mathtt{d}_{\mathcal{H}} \geq 1$. Then,* (3) *holds with*

$$\varrho_n = \mathsf{m}_{n,d}\sqrt{\log n}\sqrt{\mathtt{M}_{\mathcal{H}}\mathtt{TV}_{\mathcal{H}}} + \frac{\log n}{\sqrt{n}}\min\{\sqrt{\log n}\sqrt{\mathtt{M}_{\mathcal{H}}}, \sqrt{\mathtt{K}_{\mathcal{H}} + \mathtt{M}_{\mathcal{H}}}\}\sqrt{\mathtt{M}_{\mathcal{H}}}.$$

This corollary recovers the main result in Rio (1994, Theorem 1.1) when $d \geq 2$, where $\mathsf{m}_{n,d} = n^{-1/(2d)}$. It also covers $d = 1$, where $\mathsf{m}_{n,1} = n^{-1/2}\sqrt{\log n}$, thereby allowing for a precise comparison with prior KMT strong approximation results in the univariate case (Giné *et al.*, 2004; Giné and Nickl, 2010; Cattaneo *et al.*, 2024b). Thus, Corollary 1 contributes to the literature by covering all $d \geq 1$ cases simultaneously. While not presented here to streamline the exposition, the proof of Corollary 1 further contributes to the literature by making explicit the dependence on d, $\mathcal{X}$, and other features of the underlying data generating process. This additional contribution can be useful for non-asymptotic probability concentration arguments, or for truncation arguments in cases where the random variables have low Lebesgue density (e.g., random variables with unbounded support); see Sakhanenko (2015) for an example. Nonetheless, for $d \geq 2$, the main intellectual content of Corollary 1 is due to Rio (1994); we present it here for completeness and as a prelude for the discussion of our upcoming results.

For $d = 1$, Corollary 1 delivers an optimal KMT result when $\mathtt{K}_{\mathcal{H}} \lesssim 1$, which employs a weaker notion of total variation relative to prior literature, but at the expense of requiring an additional VC-type condition, as the following remark explains.

Remark 2. In Section 2 of Giné *et al.* (2004) and the proof of Giné and Nickl (2010), the authors considered univariate ($d = 1$) i.i.d. continuously distributed random variables, and established the strong approximation:

$$\mathbb{P}\left(\|X_n - Z_n^X\|_{\mathcal{H}} > \frac{\mathtt{pTV}_{\mathcal{H}}(t + C_1 \log n)}{\sqrt{n}}\right) \leq C_2 \exp(-C_3 t),$$

where C_1, C_2, C_3 are absolute constants, and $\mathtt{pTV}_{\mathcal{H}}$ is the pointwise total variation

$$\mathtt{pTV}_{\mathcal{H}} := \sup_{h \in \mathcal{H}} \sup_{n \geq 1} \sup_{x_1 \leq \cdots \leq x_n} \sum_{i=1}^{n-1} |h(x_{i+1}) - h(x_i)|.$$

Cattaneo *et al.* (2020, Lemma SA20) slightly generalized the result (e.g., $\mathbb{P}_X$ is not required to be absolutely continuous with respect to the Lebesgue measure), and provided a self-contained proof.

The notion of total variation used in Theorem 1 is related to, but different than, $\mathtt{pTV}_{\mathcal{H}}$. From Ambrosio *et al.* (2000, Theorem 3.27), for any h that is locally integrable with respect to the Lebesgue measure, denoted by $h \in \mathcal{L}^1_{loc}(\mathbb{R})$, then

$$\mathtt{TV}_{\{g\}} = \inf\left\{\mathtt{pTV}_{\{g\}} : g = h, \text{Lebesgue-}a.e. \text{ in } \mathbb{R}\right\},$$

and the infimum is achieved. Because $\mathtt{M}_{\mathcal{H}} < \infty$, then $\mathcal{H} \subseteq \mathcal{L}^1_{loc}(\mathbb{R})$, and hence $\mathtt{TV}_{\mathcal{H}} \leq \mathtt{pTV}_{\mathcal{H}}$. Thus, our result employs a weaker notation of total variation but imposes additional entropy conditions. In contrast, the results in Giné *et al.* (2004), Giné and Nickl (2010), and Cattaneo *et al.* (2024b) do not have additional complexity requirements on $\mathcal{H}$ and allow for $\mathbb{P}_X$ not be dominated by the Lebesgue measure, but their proof strategy is only applicable when $d = 1$. □

We illustrate the usefulness of Corollary 1 with Example 1.

Example 1 (continued)**.** Let the conditions of Theorem 1 hold, and $nb^d/\log n \to \infty$. Prior literature further assumed K is Lipschitz to verify the conditions of Corollary 1 with $\mathtt{TV}_{\mathcal{H}} \lesssim b^{d/2-1}$ and $\mathtt{K}_{\mathcal{H}} \lesssim 1$. Then, for $X_n = \xi_n$, (3) holds with $\varrho_n = (nb^d)^{-1/(2d)}\sqrt{\log n} + (nb^d)^{-1/2}\log n$. ▲

The resulting uniform Gaussian approximation convergence rate in Example 1 matches prior literature for $d = 1$ (Giné *et al.*, 2004; Giné and Nickl, 2010; Cattaneo *et al.*, 2024b) and $d \geq 2$ (Rio, 1994). This result concerns the uniform Gaussian strong approximation of the *entire* stochastic process, which can then be specialized to deduce a strong approximation for the scalar suprema of the empirical process $\|\xi_n\|_{\mathcal{H}}$. As noted by Chernozhukov *et al.* (2014, Remark 3.1(ii)), the (almost sure) strong approximation rate in Example 1 is better than their strong approximation rate (in probability) for $\|\xi_n\|_{\mathcal{H}}$ when $d = 1, 2, 3$, but their approach specifically tailored to the scalar suprema delivers better strong approximation rates when $d \geq 4$.

Following prior literature, Example 1 imposed the additional condition that K is Lipschitz to verify that $\mathcal{H} = \{b^{-d/2}K((\cdot - \mathbf{x})/b) : \mathbf{x} \in \mathcal{X}\}$ forms a VC-type class, as well as other conditions in Corollary 1. The Lipschitz restriction is easily verified for most kernel functions used in practice. One notable exception is the uniform kernel, which is nonetheless covered by Corollary 1, and prior results in the literature, but with slightly sub-optimal strong approximation rates (an extra $\sqrt{\log n}$ term appears when $d \geq 2$).

3.1.2 VC-type Lipschitz Functions

It is known that the uniform Gaussian strong approximation rate in Corollary 1 is optimal under the assumptions imposed (Beck, 1985). However,the class of functions $\mathcal{H}$ often has additional structure in statistical applications that can be exploited to improve on Corollary 1. In Example 1, for instance, prior literature further assumed K is Lipschitz to verify the sufficient conditions. Therefore, our next corollary considers a VC-type class $\mathcal{H}$ now allowing for the possibility of Lipschitz functions ($\mathtt{L}_{\mathcal{H}} < \infty$). This is one of the main contributions of our paper.

Corollary 2 (VC-type Lipschitz Functions)**.** *Suppose the conditions of Theorem 1 hold. In addition, assume that $\mathcal{H}$ is a VC-type class with respect to envelope function $\mathtt{M}_{\mathcal{H}}$ with constant $\mathtt{c}_{\mathcal{H}} \geq e$ and exponent $\mathtt{d}_{\mathcal{H}} \geq 1$. Then,* (3) *holds with*

$$\varrho_n = \min\{\mathsf{m}_{n,d}\sqrt{\mathtt{M}_{\mathcal{H}}}, \mathsf{l}_{n,d}\sqrt{\mathtt{L}_{\mathcal{H}}}\}\sqrt{\log n}\sqrt{\mathtt{TV}_{\mathcal{H}}} + \frac{\log n}{\sqrt{n}} \min\{\sqrt{\log n}\sqrt{\mathtt{M}_{\mathcal{H}}}, \sqrt{\mathtt{K}_{\mathcal{H}} + \mathtt{M}_{\mathcal{H}}}\}\sqrt{\mathtt{M}_{\mathcal{H}}}.$$

Temporarily putting aside the potential contributions of $\mathtt{M}_{\mathcal{H}}$ and $\mathtt{TV}_{\mathcal{H}}$, this corollary shows that if $\mathtt{L}_{\mathcal{H}} < \infty$ then the rate of strong approximation can be substantially improved. In particular, for $d = 2$, $\mathsf{m}_{n,2} = n^{-1/4}$ but $\mathsf{l}_{n,2} = n^{-1/2}\sqrt{\log n}$, implying that $\varrho_n = n^{-1/2}\log n$ whenever $\mathtt{K}_{\mathcal{H}} \lesssim 1$. Therefore, to the best of our knowledge, Corollary 2 is the first result in the literature establishing a uniform Gaussian strong approximation for general empirical processes based on bivariate data that can achieve the optimal univariate KMT approximation rate. (An additional $\sqrt{\log n}$ penalty would appear if $\mathtt{K}_{\mathcal{H}} = \infty$.)

For $d \geq 3$, Corollary 2 also provides improvements relative to prior literature, but falls short of achieving the optimal univariate KMT approximation rate. Specifically, $\mathsf{m}_{n,d} = n^{-1/(2d)}$ but $\mathsf{l}_{n,d} = n^{-1/d}$ for $d \geq 3$, implying that $\varrho_n = n^{-1/d}\sqrt{\log n}$. It remains an open question whether further improvements are possible at this level of generality (cf. Section 3.2 below): the main roadblock underlying the proof strategy is related to the coupling approach based on the celebrated Tusnády's inequality for binomial counts, which in turn are generated by the aforementioned mean square approximation of the functions $h \in \mathcal{H}$ by local constant functions on carefully chosen partitions of $\mathcal{X}$. Our key observation underlying Corollary 2, and hence the limitation, is that for Lipschitz functions ($\mathtt{L}_{\mathcal{H}} < \infty$) both the projection error arising from the mean square approximation and the KMT coupling error by Rio (1994, Theorem 2.1) can be improved. However, further improvements for smoother functions appears to necessitate an approximation approach that would not generate dyadic binomial counts, thereby rendering current coupling approaches inapplicable. Section 3.2 discusses an extension based on a generalization of Tusnády's inequality for a special case of interest in statistics, and we also apply those ideas to other cases of interest in Section 4.

We revisit the kernel density estimation example to illustrate the power of Corollary 2.

Example 1 (continued)**.** Under the conditions already imposed, $\mathtt{L}_{\mathcal{H}} \lesssim b^{-d/2-1}$, and Corollary 2 implies that, for $X_n = \xi_n$, (3) holds with $\varrho_n = (nb^d)^{-1/d}\sqrt{\log n} + (nb^d)^{-1/2}\log n$. ▲

Returning to the discussion of Chernozhukov *et al.* (2014, Remark 3.1(ii)), Example 1 illustrates that our almost sure strong approximation rate for the entire empirical process is now better than their strong approximation (in probability) rate for the scalar suprema $\|\xi_n\|_{\mathcal{H}}$ when $d \leq 6$. On the other hand, their approach delivers a better strong approximation rate in probability for $\|\xi_n\|_{\mathcal{H}}$ when $d \geq 7$. Our improvement is obtained without imposing additional assumptions because Rio (1994, Section 4) already assumed K is Lipschitizian for the verification of the conditions imposed by his strong approximation result (cf. Corollary 1).

3.1.3 Polynomial-Entropy Functions

Koltchinskii (1994) also considered uniform Gaussian strong approximations for the general empirical process under other notions of entropy for $\mathcal{H}$, thereby allowing for more complex classes of functions when compared to Rio (1994). Furthermore, Koltchinskii (1994) employed a Haar approximation condition, which plays a similar role as to the total variation and the Lipschitz conditions exploited in our paper. Thanks to the generality of our Theorem 1, and to enable a precise comparison to Koltchinskii (1994), the next corollary considers a class $\mathcal{H}$ satisfying a polynomial entropy condition (Definition 9).

Corollary 3 (Polynomial-Entropy Functions)**.** *Suppose the conditions of Theorem 1 hold, and that $\mathcal{H}$ is a polynomial-entropy class with respect to envelope function $\mathtt{M}_{\mathcal{H}}$ with constant $\mathtt{a}_{\mathcal{H}} > 0$ and exponent $0 < \mathtt{b}_{\mathcal{H}} < 2$. Then,* (3) *holds as follows:*

(i) *If* $\mathtt{L}_{\mathcal{H}} \leq \infty$, *then*

$$\varrho_n = \mathsf{m}_{n,d}\sqrt{\mathtt{M}_{\mathcal{H}}\mathtt{TV}_{\mathcal{H}}}(\sqrt{\log n} + (\mathsf{m}_{n,d}^2\mathtt{M}_{\mathcal{H}}^{-1}\mathtt{TV}_{\mathcal{H}})^{-\frac{\mathtt{b}_{\mathcal{H}}}{4}})$$
$$+ \sqrt{\frac{\mathtt{M}_{\mathcal{H}}}{n}} \min\{\sqrt{\log n}\sqrt{\mathtt{M}_{\mathcal{H}}}, \sqrt{\mathtt{K}_{\mathcal{H}} + \mathtt{M}_{\mathcal{H}}}\}(\log n + (\mathsf{m}_{n,d}^2\mathtt{M}_{\mathcal{H}}^{-1}\mathtt{TV}_{\mathcal{H}})^{-\frac{\mathtt{b}_{\mathcal{H}}}{2}}),$$

(ii) *If* $\mathtt{L}_{\mathcal{H}} < \infty$, *then*

$$\varrho_n = \mathsf{l}_{n,d}\sqrt{\mathtt{L}_{\mathcal{H}}\mathtt{TV}_{\mathcal{H}}}(\sqrt{\log n} + (\mathsf{l}_{n,d}^2\mathtt{M}_{\mathcal{H}}^{-2}\mathtt{L}_{\mathcal{H}}\mathtt{TV}_{\mathcal{H}})^{-\frac{\mathtt{b}_{\mathcal{H}}}{4}})$$
$$+ \sqrt{\frac{\mathtt{M}_{\mathcal{H}}}{n}} \min\{\sqrt{\log n}\sqrt{\mathtt{M}_{\mathcal{H}}}, \sqrt{\mathtt{K}_{\mathcal{H}} + \mathtt{M}_{\mathcal{H}}}\}(\log n + (\mathsf{l}_{n,d}^2\mathtt{M}_{\mathcal{H}}^{-2}\mathtt{L}_{\mathcal{H}}\mathtt{TV}_{\mathcal{H}})^{-\frac{\mathtt{b}_{\mathcal{H}}}{2}}).$$

This corollary reports a simplified version of our result, which is the best possible bound for the discussion in this section. See Corollary SA.3 in the supplemental appendix for the general case. It is possible to apply Corollary 3 to Example 1, although the result is sub-optimal relative to the previous results leveraging a VC-type condition.

Example 1 (continued)**.** Under the conditions already imposed, for any $0 < \mathtt{b}_{\mathcal{H}} < 2$, we can take $\mathsf{a}_{\mathcal{H}} = \log(d+1) + d\mathtt{b}_{\mathcal{H}}^{-1}$ so that $\mathcal{H}$ is a polynomial-entropy class with constants $(\mathsf{a}_{\mathcal{H}}, \mathtt{b}_{\mathcal{H}})$. Then, Corollary 3*(ii)* implies that, for $X_n = \xi_n$, (3) holds with $\varrho_n = \mathsf{a}_{\mathcal{H}}^2(nb^d)^{-\frac{1}{d}(1-\frac{\mathtt{b}_{\mathcal{H}}}{2})}b^{-d\mathtt{b}_{\mathcal{H}}} + \mathsf{a}_{\mathcal{H}}^2(nb^d)^{-\frac{1}{2}+\frac{\mathtt{b}_{\mathcal{H}}}{d}}b^{-\frac{d\mathtt{b}_{\mathcal{H}}}{2}}$. ▲

Our running example shows that a uniform Gaussian strong approximation based on polynomial entropy conditions can lead to sub-optimal KMT approximation rates. However, for other (larger) classes of functions, those results are useful. The following remark discusses an example studied in Koltchinskii (1994), and further compares our contributions to his work.

Remark 3. Suppose Assumption A holds with $\mathbb{P}_X$ the uniform distribution on $\mathcal{X} = [0,1]^d$, and $\mathcal{H}$ a subclass of $C^q(\mathcal{X})$ with C^q-norm uniformly bounded by 1 and $2 \leq d < q$. Koltchinskii (1994, page 111) discusses this example after his Theorem 11.3, and reports a uniform Gaussian strong approximation $n^{-\frac{q-d}{2qd}} \operatorname{polylog} n$.

Corollary 3 is applicable to this case. More precisely, $\mathtt{M}_{\mathcal{H}} = 1$, $\mathtt{TV}_{\mathcal{H}} = 1$, $\mathtt{L}_{\mathcal{H}} = 1$, and van der Vaart and Wellner (2013, Theorem 2.7.1) shows that $\mathcal{H}$ is a polynomial-entropy class with constants $\mathsf{a}_{\mathcal{H}} = K$ and $\mathtt{b}_{\mathcal{H}} = d/q$, where K is a constant only depending on q and d. Then, Corollary 3*(ii)* implies that, for $X_n = \xi_n$, (3) holds with

$$\varrho_n = \begin{cases} n^{-\frac{1}{2}+\frac{1}{q}} \operatorname{polylog} n & \text{if } d = 2 \\ n^{-\frac{2q-d}{2dq}} \operatorname{polylog} n & \text{if } d > 2 \end{cases},$$

which gives a faster convergence rate than the one obtained by Koltchinskii (1994).

The improvement is explained by two differences between Koltchinskii (1994) and our approach. First, we explicitly incorporate the Lipschitz condition, and hence we can take $\beta = \frac{2}{d}$ instead of

$\beta = \frac{1}{d}$ in Equation (3.1) of Koltchinskii (1994). Second, using the uniform entropy condition approach, we get $\log N(\mathcal{H}, e_{\mathbb{P}_X}, \varepsilon) \leq K\varepsilon^{-d/q}$, while Koltchinskii (1994) started with the bracketing number condition $\log N_{[\,]}(\mathcal{F}, L_1(\mathbb{P}), \varepsilon) = O(\varepsilon^{-d/q})$ and, with the help of his Lemma 8.4, applied Theorem 3.1 with $\alpha = \frac{d}{d+q}$ in his Equation (3.2). As a result, because the proof of his Theorem 3.1 leverages the fact that Equation (3.2) implies that $\log N(\mathcal{H}, e_{\mathbb{P}_X}, \varepsilon) = O(\varepsilon^{-2d/q})$, and his approximation rate is looser by a power of two when compared to the uniform entropy condition underlying our Corollary 3.

Setting $\mathtt{L}_{\mathcal{H}} = \infty$, $\mathtt{b}_{\mathcal{H}} = \frac{2d}{q}$, and keeping the other constants the same, Corollary 3*(i)* would give $\varrho_n = n^{-\frac{q-d}{2qd}} \operatorname{polylog} n$, which is the same rate as in Koltchinskii (1994). Finally, Theorem 3.2 in Koltchinskii (1994) allows for $\log N(\mathcal{H}, e_{\mathbb{P}_X}, \varepsilon) = O(\varepsilon^{-2\rho})$ where ρ is not implied by his Equation (3.2), in which case his result would give the strong approximation rate $n^{-\frac{2q-d}{4qd}} \operatorname{polylog} n$. □

3.2 Quasi-Uniform Haar Basis

Theorem 1 established that the general empirical process (1) indexed by VC-type Lipschitz functions can admit a strong approximation (3) at the optimal univariate KMT rate $\varrho_n = n^{-1/2} \log n$ when $d \in \{1, 2\}$, and at the improved (but possibly suboptimal) rate $\varrho_n = n^{-1/d}\sqrt{\log n}$ when $d \geq 3$, in both cases putting aside the potential additional contributions controlled by $\mathtt{M}_{\mathcal{H}}$, $\mathtt{L}_{\mathcal{H}}$, $\mathtt{TV}_{\mathcal{H}}$, and $\mathtt{K}_{\mathcal{H}}$. When applied to kernel density estimation (Example 1), our results showed that $\varrho_n = (nb^d)^{-1/2} \log n$ when $d = 1, 2$, and $\varrho_n = (nb^d)^{-1/d}\sqrt{\log n}$ when $d \geq 3$, where nb^d is the "effective sample" size.

The possibly suboptimal strong approximation rate $\varrho_n = n^{-1/d}\sqrt{\log n}$ for $d \geq 3$ arises from the L_2 approximation of the functions $h \in \mathcal{H}$ by a Haar basis expansion based on a carefully chosen *dyadic* partition of $\mathcal{X}$. In this section, we demonstrate that the general empirical process (1) can admit a univariate KMT optimal strong approximation when $\mathcal{H}$ belongs to the span of Haar basis based on a *quasi-uniform* partition of $\mathcal{X}$ with cardinality L, which can be viewed as an approximation based on $L \to \infty$ as $n \to \infty$. More precisely, the following theorem showcases a setting where the univariate KMT optimal approximation rate based on the "effective sample" size n/L is achieved for all $d \geq 1$. Our formulation leverages and generalizes two ideas from the regression Splines literature (Huang, 2003): (i) the cells forming the Haar basis are assumed to be quasi-uniform with respect to $\mathbb{P}_X$; and (ii) the number of active cells of the Haar basis affect the strong approximation.

Theorem 2. *Suppose* $(\mathbf{x}_i : 1 \leq i \leq n)$ *are i.i.d. random vectors taking values in* $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ *with common law* $\mathbb{P}_X$, $\mathcal{X} \subseteq \mathbb{R}^d$, *and* $\mathcal{H}$ *is a class of functions on* $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P}_X)$ *such that* $\mathtt{M}_{\mathcal{H}} < \infty$ *and* $\mathcal{H} \subseteq \operatorname{Span}\{\mathbb{1}_{\Delta_l} : 0 \leq l < L\}$, *where* $\{\Delta_l : 0 \leq l < L\}$ *forms a quasi-uniform partition of* $\mathcal{X}$ *in the sense that*

$$\mathcal{X} \subseteq \sqcup_{0 \leq l \leq L} \Delta_l \qquad \textit{and} \qquad \frac{\max_{0 \leq l < L} \mathbb{P}_X(\Delta_l)}{\min_{0 \leq l < L} \mathbb{P}_X(\Delta_l)} \leq \rho < \infty.$$

Then, on a possibly enlarged probability space, there exists a sequence of mean-zero Gaussian processes $(Z_n^X(h) : h \in \mathcal{H})$ with almost sure continuous trajectories such that:

- $\mathbb{E}[X_n(h_1)X_n(h_2)] = \mathbb{E}[Z_n^X(h_1)Z_n^X(h_2)]$ *for all* $h_1, h_2 \in \mathcal{H}$, *and*
- $\mathbb{P}\big[\|X_n - Z_n^X\|_{\mathcal{H}} > C_1 C_\rho \mathsf{P}_n(t)\big] \leq C_2 e^{-t} + L e^{-C_\rho n/L}$ *for all* $t > 0$,

where C_1 and C_2 are universal constants, C_ρ is a constant that only depends on ρ, and

$$\mathsf{P}_n(t) = \min_{\delta \in (0,1)} \Big\{ \mathsf{H}_n(t,\delta) + \mathsf{F}_n(t,\delta) \Big\},$$

with

$$\mathsf{H}_n(t,\delta) := \sqrt{\frac{\mathtt{M}_{\mathcal{H}}\mathtt{E}_{\mathcal{H}}}{n/L}} \sqrt{t + \log \mathtt{N}_{\mathcal{H}}(\delta)} + \sqrt{\frac{\min\{\log_2(L), \mathtt{S}_{\mathcal{H}}^2\}}{n}} \mathtt{M}_{\mathcal{H}}(t + \log \mathtt{N}_{\mathcal{H}}(\delta)),$$

where $\mathtt{S}_{\mathcal{H}} = \sup_{h \in \mathcal{H}} \sum_{l=1}^{L} \mathbb{1}(\mathrm{Supp}(h) \cap \Delta_l \neq \emptyset)$.

This theorem shows that if $n^{-1} L \log L \to 0$, then a valid strong approximation can be achieved with exponential probability concentration. The proof of Theorem 2 leverages the fact that the L_2 projection error is zero by assumption, but recognizes that Rio (1994, Theorem 2.1) does not apply because the partitions are *quasi-dyadic*, preventing the use of the celebrated Tusnády's inequality. Instead, in Section SA-II of the supplemental appendix, we present two technical results to circumvent that limitation: (i) Lemma SA.6 combines Brown *et al.* (2010, Lemma 2) and Sakhanenko (1996, Lemma 2) to establish a new version of Tusnády's inequality that allows for more general binomial random variables $\mathsf{Bin}(n,p)$ with $\underline{p} \leq p \leq \overline{p}$, the error bound holding uniformly in p, as required by the quasi-dyadic partitioning structure; and (ii) Lemma SA.7 presents a generalization of Rio (1994, Theorem 2.1) to the case of quasi-dyadic partitions of $\mathcal{X}$.

Assuming a VC-type condition on $\mathcal{H}$, and putting aside the potential contributions of $\mathtt{M}_{\mathcal{H}}$, $\mathtt{E}_{\mathcal{H}}$, and $\mathtt{S}_{\mathcal{H}}$, it follows that (3) holds with $\varrho_n = \log(L)/(n/L)$, thereby achieving the optimal univariate KMT approximation rate for all $d \geq 1$ with "effective sample" size n/L. More precisely, we have the following corollary.

Corollary 4 (VC-type Haar Basis)**.** *Suppose the conditions of Theorem 2 hold. In addition, assume that $\mathcal{H}$ is a VC-type class with respect to envelope function $\mathtt{M}_{\mathcal{H}}$ with constant $\mathtt{c}_{\mathcal{H}} \geq e$ and exponent $\mathtt{d}_{\mathcal{H}} \geq 1$. Then, (3) holds with*

$$\varrho_n = \sqrt{\frac{\mathtt{M}_{\mathcal{H}}\mathtt{E}_{\mathcal{H}}}{n/L}} \sqrt{\log n} + \sqrt{\frac{\min\{\log_2(L), \mathtt{S}_{\mathcal{H}}^2\}}{n}} \mathtt{M}_{\mathcal{H}} \log n.$$

To provide a simple illustration of Theorem 2 to statistics, we consider the classical histogram density estimator.

Example 2 (Histogram Density Estimation)**.** The histogram density estimator of f_X is

$$\check{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=0}^{L-1}\mathbb{1}(\mathbf{x}_i \in \Delta_l)\mathbb{1}(\mathbf{x} \in \Delta_l),$$

where $\{\Delta_l : 0 \leq l < L\}$ forms a quasi-uniform partition of $\mathcal{X}$, where the partition size $L \to \infty$ as $n \to \infty$ in statistical applications. We consider establishing a strong approximation for the "localized" empirical process $(\zeta_n(\mathbf{x}) : \mathbf{x} \in \mathcal{X})$, where

$$\zeta_n(\mathbf{x}) := \sqrt{nL}\big(\check{f}(\mathbf{x}) - \mathbb{E}[\check{f}(\mathbf{x})]\big) = X_n(h), \qquad h \in \mathcal{H},$$

with $\mathcal{H}$ the collection of Haar basis functions based on the partition $\{\Delta_l : 0 \leq l < L\}$.

The conditions of Theorem 2 are satisfied with $\mathtt{M}_{\mathcal{H}} = L^{1/2}$, $\mathtt{E}_{\mathcal{H}} = L^{-1/2}$, and $\mathtt{S}_{\mathcal{H}} = 1$. It follows that, for $X_n = \zeta_n$, (3) holds with

$$\varrho_n = \frac{\log(nL)}{\sqrt{n/L}},$$

provided that $\log(nL)L/n \to 0$. ▲

Theorem 2, and in particular Example 2, showcases the existence of a class of stochastic processes for which a valid uniform Gaussian strong approximation is established with optimal univariate KMT rate in terms of the effective sample size n/L for all $d \geq 1$. This result is achieved because there is no error arising from the mean square approximation ($\mathcal{H}$ is assumed to be spanned by a Haar space), and with the help of our generalized Tusnády's inequality (Lemma SA.6).

Because the setup of Theorem 2 is rather special, the finding in this subsection is mostly of theoretical interest. However, our key ideas will be leveraged in the next section when studying regression estimation problems, where the quasi-uniform partitioning arises naturally in setting like regression trees (Breiman *et al.*, 1984) or nonparametric partitioning-based estimation (Cattaneo *et al.*, 2020).

4 Residual-Based Empirical Process

This section establishes improved uniform Gaussian strong approximation for the residual empirical process $(R_n(g,r) : (g,r) \in \mathcal{G} \times \mathcal{R})$ defined in (7). We impose the following assumption.

Assumption B. $(\mathbf{z}_i = (\mathbf{x}_i, y_i) : 1 \leq i \leq n)$ *are i.i.d. random vectors taking values in* $(\mathcal{X} \times \mathbb{R}, \mathcal{B}(\mathcal{X} \times \mathbb{R}))$ *with* $\mathcal{X}$ *compact, and* $\mathbf{x}_i \sim \mathbb{P}_X$ *admits a Lebesgue density* f_X *continuous and positive on* $\mathcal{X}$*.*

This assumption incorporates the presence of random variables $y_i \sim \mathbb{P}_Y$, but otherwise imposes the same regularity conditions as Assumption A for the marginal distribution $\mathbb{P}_X$ of $\mathbf{x}_i$. In particular, it does not restrict the support of $\mathbb{P}_Y$ nor requires $\mathbb{P}_Y$ to be dominated by the Lebesgue measure, which is important for some statistical applications.

To motivate this section, consider first the simple local empirical process discussed in Chernozhukov *et al.* (2014, Section 3.1):

$$S_n(\mathbf{x}) = \frac{1}{nb^d}\sum_{i=1}^{n} K\Big(\frac{\mathbf{x}_i - \mathbf{x}}{b}\Big) y_i, \qquad \mathbf{x} \in \mathcal{X}. \tag{11}$$

Using our notation for residual empirical process, $\big(\sqrt{nb^d}(S_n(\mathbf{x}) - \mathbb{E}[S_n(\mathbf{x})|\mathbf{x}_1, \cdots, \mathbf{x}_n]) : \mathbf{x} \in \mathcal{X}\big) = (R_n(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$ with $\mathcal{G} = \{b^{-d/2}K(\frac{\cdot - \mathbf{x}}{b}) : \mathbf{x} \in \mathcal{X}\}$ and $\mathcal{R} = \{\mathrm{Id}\}$, where Id denotes the identity map from $\mathbb{R}$ to $\mathbb{R}$. This setting corresponds to kernel regression estimation with K interpreted as the equivalent kernel; see Section 4.1 for details. As noted in Chernozhukov *et al.* (2014, Remark 3.1(iii)), a direct application of Rio (1994), or of our Theorem 1, views $\mathbf{z}_i$ as the underlying $(d + 1)$-dimensional vector of random variables entering the general empirical process X_n defined in (1). Specifically, under some regularity conditions on K and non-trivial restrictions on the joint distribution $\mathbb{P}_Z$, Rio (1994)'s strong approximation result verifies (3) with (6), which is also verified via Corollary 1. Furthermore, employing a Lipschitz property of $\mathcal{G} \times \mathcal{R}$, Corollary 2 would give the improved strong approximation result (8), under regularity conditions.

The strong approximation results for $S_n(\mathbf{x})$ illustrate two fundamental limitations because all the elements in $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ are treated symmetrically. First, the effective sample size emerging in the strong approximation rate is nb^{d+1}, which is necessarily suboptimal because only the d-dimensional covariate $\mathbf{x}_i$ are being smoothed out. In other words, since the pointwise variance of the process is of order $n^{-1}b^{-d}$, the correct effective sample size should be nb^d, and therefore applying Rio (1994), or our improved Theorem 1, leads to a suboptimal uniform Gaussian strong approximation for $S_n(\mathbf{x})$. Second, applying Rio (1994), or our improved Theorem 1, requires $\mathbf{z}_i = (\mathbf{x}_i, y_i) \sim \mathbb{P}_Z$ to be continuously distributed and supported on $[0, 1]^{d+1}$, possibly after applying the Rosenblatt transform (Lemma SA.12), as discussed in Remark 1. This requirement imposes non-trivial restrictions on the joint distribution $\mathbb{P}_Z$, and in particular on the marginal distribution of the outcome y_i, which limit the applicability of the resulting strong approximation results. For example, it could be assumed that $(\mathbf{x}_i, y_i) = (\mathbf{x}_i, \varphi(\mathbf{x}_i, u_i))$ where $(\mathbf{x}_i, u_i)$ satisfies Assumption A and φ is bounded with bounded uniform variation and local uniform variation; see Chernozhukov *et al.* (2014, Remark 3.1(iii)) for more discussion.

Motivated by the aforementioned limitations, the following theorem explicitly studies the residual empirical process $(R_n(g, r) : (g, r) \in \mathcal{G} \times \mathcal{R})$ defined in (7), leveraging its intrinsic multiplicative separable structure. We present our result under a VC-type condition on $\mathcal{G} \times \mathcal{R}$ to streamline the discussion, but a result at the same level of generality as Theorem 1 is given in the supplemental appendix (Section SA-I.2 and SA-I.3).

Theorem 3. *Suppose Assumption B holds with $\mathcal{X} = [0, 1]^d$, and the following conditions hold.*

(i) *$\mathcal{G}$ is a real-valued pointwise measurable class of functions on $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P}_X)$, and a VC-type class with respect to envelope function $\mathtt{M}_{\mathcal{G}}$ with constant $\mathtt{c}_{\mathcal{G}} \geq e$ and exponent $\mathtt{d}_{\mathcal{G}} \geq 1$.*

(ii) *$\mathcal{R}$ is a real-valued pointwise measurable class of functions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_Y)$, and a VC-type* *class with respect to* $\mathtt{M}_{\mathcal{R}}$ *with constant* $\mathtt{c}_{\mathcal{R}} \geq e$ *and exponent* $\mathtt{d}_{\mathcal{R}} \geq 1$. *Furthermore, one of the following holds:*

(a) $\mathtt{M}_{\mathcal{R}} \lesssim 1$ *and* $\mathtt{pTV}_{\mathcal{R}} \lesssim 1$, *and set* $\alpha = 0$, *or*

(b) $M_{\mathcal{R}}(y) \lesssim 1 + |y|^{\alpha}$ *and* $\mathtt{pTV}_{\mathcal{R},(-|y|,|y|)} \lesssim 1 + |y|^{\alpha}$ *for all* $y \in \mathbb{R}$ *and for some* $\alpha > 0$, *and* $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(y_i)|\mathbf{x}_i = \mathbf{x}] \leq 2$.

(iii) *There exists a constant* $\mathtt{c}_4$ *such that* $|\log_2 \mathtt{E}_{\mathcal{G}}| + |\log_2 \mathtt{TV}| + |\log_2 \mathtt{M}_{\mathcal{G}}| \leq \mathtt{c}_4 \log_2 n$, *where* $\mathtt{TV} = \max\{\mathtt{TV}_{\mathcal{G}}, \mathtt{TV}_{\mathcal{G} \times \mathcal{V}_{\mathcal{R}}}\}$ *with* $\mathcal{V}_{\mathcal{R}} := \{\theta(\cdot, r), r \in \mathcal{R}\}$, *and* $\theta(\cdot, r) : \mathcal{X} \to \mathbb{R}$ *is the function defined by* $\theta(\mathbf{x}, r) = \mathbb{E}[r(y_i)|\mathbf{x}_i = \mathbf{x}], \mathbf{x} \in \mathcal{X}$.

Then, on a possibly enlarged probability space, there exists a sequence of mean-zero Gaussian processes $(Z_n^R(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$ *with almost sure continuous trajectories such that:*

- $\mathbb{E}[R_n(g_1, r_1) R_n(g_2, r_2)] = \mathbb{E}[Z_n^R(g_1, r_1) Z_n^R(g_2, r_2)]$ *for all* $(g_1, r_1), (g_2, r_2) \in \mathcal{G} \times \mathcal{R}$, *and*
- $\mathbb{P}\left[\|R_n - Z_n^R\|_{\mathcal{G} \times \mathcal{R}} > C_1 C_\alpha \mathsf{T}_n(t)\right] \leq C_2 e^{-t}$ *for all* $t > 0$,

where C_1 *and* C_2 *are universal constants,* $C_\alpha = \max\{1 + (2\alpha)^{\frac{\alpha}{2}}, 1 + (4\alpha)^{\alpha}\}$, *and*

$$\mathsf{T}_n(t) := \mathsf{A}_n (t + \mathtt{c}_4 \log_2 n + \mathtt{d} \log(\mathtt{c} n))^{\alpha + \frac{3}{2}} \sqrt{d} + \frac{\mathtt{M}_{\mathcal{G}}}{\sqrt{n}} (t + \mathtt{c}_4 \log_2 n + \mathtt{d} \log(\mathtt{c} n))^{\alpha + 1},$$

$$\mathsf{A}_n := \min\left\{ \left(\frac{\mathtt{c}_1^d \mathtt{M}_{\mathcal{G}}^{d+1} \mathtt{TV}^d \mathtt{E}_{\mathcal{G}}}{n} \right)^{\frac{1}{2d+2}}, \left(\frac{\mathtt{c}_1^{\frac{d}{2}} \mathtt{c}_2^{\frac{d}{2}} \mathtt{M}_{\mathcal{G}} \mathtt{E}_{\mathcal{G}} \mathtt{TV}^{\frac{d}{2}} \mathtt{L}^{\frac{d}{2}}}{n} \right)^{\frac{1}{d+2}} \right\},$$

and $\mathtt{c} = \mathtt{c}_{\mathcal{G}} \mathtt{c}_{\mathcal{R}}$, $\mathtt{d} = \mathtt{d}_{\mathcal{G}} + \mathtt{d}_{\mathcal{R}}$, $\mathtt{L} = \max\{\mathtt{L}_{\mathcal{G}}, \mathtt{L}_{\mathcal{G} \times \mathcal{V}_{\mathcal{R}}}\}$.

This theorem establishes a uniform Gaussian strong approximation for the residual stochastic process $(R_n(g, r) : (g, r) \in \mathcal{G} \times \mathcal{R})$ defined in (7) under regularity conditions specifically tailored to leverage its multiplicative separable structure. Condition (i) in Theorem 3 is analogous to the conditions imposed in Corollaries 1 and 2 for the general empirical process. This is a mild, standard restriction on the portion of the stochastic process corresponding to the covariates $\mathbf{x}_i$. Condition (ii) in Theorem 3 is a new, mild condition on the portion of the stochastic process corresponding to the outcome y_i. This condition either assume $r(y_i)$ to be uniformly bounded, or restricts the tail decay of the function class $\mathcal{R}$ without requiring specific strong assumptions on the distribution $\mathbb{P}_Y$ and hence the joint distribution $\mathbb{P}_Z$ (cf. Chernozhukov *et al.* (2014, Remark 3.1(iii))). Finally, Condition (iii) is weak and imposed only to simplify the exposition; see Section SA-I.2 and SA-I.3 in the supplemental appendix for the general result. We require $\mathtt{pTV}$ conditions on $\mathcal{R}$ in (ii), and $\mathtt{TV}$ conditions on $\mathcal{G}$ and $\mathcal{G} \times \mathcal{V}_{\mathcal{R}}$ in (iii), because $\mathbf{x}_i$ has a Lebesgue density but y_i may not have one, which means values of $\mathcal{R}$ at a Lebesgue measure-zero set can affect the value of $R_n(g, r)$, but values of $\mathcal{G}$ and $\mathcal{G} \times \mathcal{V}_{\mathcal{R}}$ at a Lebesgue measure-zero set do not.

The proof strategy of Theorem 3 is the same as for the general empirical process (Theorem 1). First, we discretize to a δ-net to obtain

$$\|R_n - Z_n^R\|_{\mathcal{G} \times \mathcal{R}} \leq \|R_n - R_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta}\|_{\mathcal{G} \times \mathcal{R}} + \|R_n - Z_n^R\|_{(\mathcal{G} \times \mathcal{R})_\delta} + \|Z_n^R \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta} - Z_n^R\|_{\mathcal{G} \times \mathcal{R}},$$

where the terms capturing fluctuation off-the-net, $\|R_n - R_n \circ \pi_{(\mathcal{G}\times\mathcal{R})_\delta}\|_{\mathcal{G}\times\mathcal{R}}$ and $\|Z_n^R \circ \pi_{(\mathcal{G}\times\mathcal{R})_\delta} - Z_n^R\|_{\mathcal{G}\times\mathcal{R}}$, are handled via standard empirical process methods. Second, the remaining term $\|R_n - Z_n^R\|_{(\mathcal{G}\times\mathcal{R})_\delta}$, which captures the finite-class Gaussian approximation error, is once again decomposed via a suitable mean square "projection" from $L_2(\mathbb{R}^d \times \mathbb{R})$ to the class of piecewise constant Haar functions on a carefully chosen collection of cells partitioning the support of $\mathbb{P}_Z$. This is our point of departure from prior literature.

We design of cells based on two key observations: (i) regularity conditions are often imposed on the conditional distribution $y_i|\mathbf{x}_i$ (as opposed to their joint distribution); and (ii) $\mathcal{G}$ and $\mathcal{R}$ often require different regularity conditions. For example, in the classical regression case discussed previously, $\mathcal{R}$ is just the singleton identity function but $\mathbb{P}_Y$ may have unbounded support, while $\mathcal{G}$ is a VC-type class of n-varying functions with $\mathbb{P}_X$ compact supported. Thus, the dimension of y_i is a nuisance for the strong approximation, making results like Theorem 1 suboptimal in general. These observations suggest choosing dyadic cells by an asymmetric iterative splitting construction, where first the support of each dimension of $\mathbf{x}_i$ is partitioned, and only after the support of y_i is partitioned based on the conditional distribution of $y_i|\mathbf{x}_i$. See Section SA-III.1 in the supplemental appendix for details of our proposed dyadic cells expansion.

Given our dyadic expansion exploiting the structure of the residual empirical process R_n, we decompose the term $\|R_n - Z_n^R\|_{(\mathcal{G}\times\mathcal{R})_\delta}$ similarly to (10), leading to a "projected" piecewise constant process and the corresponding two projection errors. However, instead of employing the L_2-projection Π_0 as in (10), we now use another mapping Π_2 from $L_2(\mathbb{R}^d \times \mathbb{R})$ to piecewise constant functions that explicitly factorizes the product $g(\mathbf{x}_i)r(y_i)$. In fact, as we discuss in the supplemental appendix (Section SA-III.2), each base level cell $\mathcal{C}$ produced by our asymmetric dyadic splitting scheme can be written as a product of the form $\mathcal{X}_l \times \mathcal{Y}_m$, where $\mathcal{X}_l$ denotes the l-th cell for $\mathbf{x}_i$ and $\mathcal{Y}_m$ denotes the m-th cell for y_i. Thus, Π_2 is carefully chosen so that once we know $\mathbf{x} \in \mathcal{X}_l$ for some l, $\Pi_2[g,r](\mathbf{x},y) = \sum_{m=0}^{2^N-1} \mathbb{1}(y \in \mathcal{Y}_m)\mathbb{E}[r(y_i)|y_i \in \mathcal{Y}_m, \mathbf{x}_i \in \mathcal{X}_l]\mathbb{E}[g(\mathbf{x}_i)|\mathbf{x}_i \in \mathcal{X}_l]$, which only depends on y, and has envelope and total variation no greater than those for r.

Finally, our Tusnády's lemma for more general binomial counts (Lemma SA.6) allows for the Gaussian coupling of any piecewise-constant functions over our asymmetrically constructed dyadic cells. A generalization of Rio (1994, Theorem 2.1) enables upper bounding the Gaussian approximation error for processes indexed by piecewise constant functions by summing up a quadratic variation from all layers in the cell expansion. By the above choice of cells and projections, the contribution from the last layers corresponding to splitting y_i amounts to a sum of one-dimensional KMT coupling error from all possible $\mathcal{X}_l$ cells. In fact, we know one-dimensional KMT coupling is optimal and, as a consequence, requiring a vanishing contribution of y_i layers to the approximation error does not add extra requirements besides conditions on envelope functions and an L_1 bound for $\mathcal{G}$. This explains why we can obtain strong approximation rates reflecting the correct effective sample size underlying the empirical process for the kernel regression (or "local empirical process") example. The supplemental appendix contains all the technical details.

The following corollary summarizes the main result from Theorem 3.

Corollary 5 (Strong Approximation Residual Empirical Process)**.** *Suppose the conditions of Theorem 3 hold. Then,* $\|R_n - Z_n^R\|_{\mathcal{G}\times\mathcal{R}} = O(\varrho_n)$ *a.s. with*

$$\varrho_n = \min\left\{\frac{(\mathtt{M}_{\mathcal{G}}^{d+1}\mathtt{TV}^d\mathtt{E}_{\mathcal{G}})^{\frac{1}{2d+2}}}{n^{1/(2d+2)}}, \frac{(\mathtt{M}_{\mathcal{G}}\mathtt{TV}^{\frac{d}{2}}\mathtt{E}_{\mathcal{G}}\mathtt{L}^{\frac{d}{2}})^{\frac{1}{d+2}}}{n^{1/(d+2)}}\right\}(\log n)^{\alpha+3/2} + \frac{(\log n)^{\alpha+1}}{\sqrt{n}}\mathtt{M}_{\mathcal{G}}.$$

This corollary shows that our best attainable uniform Gaussian strong approximation rate for the residual empirical process R_n is $n^{-1/(d+2)}$ polylog n, putting aside the contributions from $\mathtt{M}_{\mathcal{G}}$, $\mathtt{TV} = \max\{\mathtt{TV}_{\mathcal{G}}, \mathtt{TV}_{\mathcal{G}\times\mathcal{V}_{\mathcal{R}}}\}$, $\mathtt{E}_{\mathcal{G}}$, and $\mathtt{L} = \max\{\mathtt{L}_{\mathcal{G}}, \mathtt{L}_{\mathcal{G}\times\mathcal{V}_{\mathcal{R}}}\}$. It is not possible to provide a strict ranking between Corollary 2 and Corollary 5. On the one hand, Corollary 2 treats all components in $\mathbf{z}_i$ symmetrically, and thus imposes stronger regularity conditions on $\mathbb{P}_Z$, but leads to the better approximation rate $n^{-\min\{1/(d+1),1/2\}}$ polylog n, putting aside the potential contributions of $\mathtt{M}_{\mathcal{G}\times\mathcal{R}}$, $\mathtt{TV}_{\mathcal{G}\times\mathcal{R}}$, $\mathtt{L}_{\mathcal{G}\times\mathcal{R}}$. On the other hand, as discussed previously, Corollary 5 can deliver a tighter strong approximation under much weaker regularity conditions whenever $\mathcal{H} = \mathcal{G}\times\mathcal{R}$ and $\mathcal{G}$ varies with n, as it is the case of the local empirical processes arising from nonparametric statistics. The following section offers a substantive application illustrating this point.

4.1 Example: Local Polynomial Regression

We demonstrate the applicability and improvements of Theorem 3 and Corollary 5 with a substantive application to nonparametric local polynomial regression (Fan and Gijbels, 1996). Assume $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ satisfy Assumption B, and consider the estimand

$$\theta(\mathbf{x}; r) = \mathbb{E}[r(y_i)|\mathbf{x}_i = \mathbf{x}], \qquad \mathbf{x} \in \mathcal{X}, \quad r \in \mathcal{R}, \tag{12}$$

where we focus on two leading cases to streamline the discussion: (i) $\mathcal{R}_1 := \{\mathrm{Id}\}$ corresponds to the conditional expectation $\mu(\mathbf{x}) := \mathbb{E}[y_i|\mathbf{x}_i = \mathbf{x}]$, and (ii) $\mathcal{R}_2 := \{\mathbb{1}(y_i \le y) : y \in \mathbb{R}\}$ corresponds to the conditional distribution function $F(y|\mathbf{x}) := \mathbb{E}[\mathbb{1}(y_i \le y)|\mathbf{x}_i = \mathbf{x}]$. In the first case, $\mathcal{R}$ is a singleton but the identify function calls for the possibility of $\mathbb{P}_Y$ not being dominated by the Lebesgue measure or perhaps being continuously distributed with unbounded support. In the second case, $\mathcal{R}$ is a VC-type class of indicator functions, and hence $r(y_i)$ is uniformly bounded, but establishing uniformity over $\mathcal{R}$ is of statistical interest (e.g., to construct specification hypothesis tests based on conditional distribution functions).

Suppose the kernel function $K : \mathbb{R}^d \to \mathbb{R}$ is non-negative, Lipschitz, and compact supported. Using standard multi-index notation, $\mathbf{p}(\mathbf{u})$ denotes the $\frac{(d+\mathfrak{p})!}{d!\mathfrak{p}!}$-dimensional vector collecting the ordered elements $\mathbf{u}^{\boldsymbol{\nu}}/\boldsymbol{\nu}!$ for $0 \le |\boldsymbol{\nu}| \le \mathfrak{p}$, where $\mathbf{u}^{\boldsymbol{\nu}} = u_1^{\nu_1}u_2^{\nu_2}\cdots u_d^{\nu_d}$, $\boldsymbol{\nu}! = \nu_1!\nu_2!\cdots\nu_d!$ and $|\boldsymbol{\nu}| = \nu_1+\nu_2+\cdots+\nu_d$, for $\mathbf{u} = (u_1, u_2, \cdots, u_d)^\top$ and $\boldsymbol{\nu} = (\nu_1, \nu_2, \cdots, \nu_d)^\top$. A local polynomial regression estimator of $\theta(\mathbf{x}; r)$ is

$$\widehat{\theta}(\mathbf{x}; r) := \mathbf{e}_1^\top\widehat{\boldsymbol{\beta}}(\mathbf{x}, r), \qquad \widehat{\boldsymbol{\beta}}(\mathbf{x}, r) := \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^n \big(r(y_i) - \mathbf{p}(\mathbf{x}_i - \mathbf{x})^\top\boldsymbol{\beta}\big)^2 K\Big(\frac{\mathbf{x}_i - \mathbf{x}}{b}\Big),$$

with $\mathbf{x} \in \mathcal{X}$, $r \in \mathcal{R}_1$ or $r \in \mathcal{R}_2$, and $\mathbf{e}_1$ denoting the first standard basis vector. The estimation error can be decomposed into three terms (linearization, non-linearity error, and smoothing bias):

$$\widehat{\theta}(\mathbf{x}, r) - \theta(\mathbf{x}, r) = \underbrace{\mathbf{e}_1^\top \mathbf{H}_\mathbf{x}^{-1} \mathbf{S}_{\mathbf{x},r}}_{\text{linearization}} + \underbrace{\mathbf{e}_1^\top (\widehat{\mathbf{H}}_\mathbf{x}^{-1} - \mathbf{H}_\mathbf{x}^{-1}) \mathbf{S}_{\mathbf{x},r}}_{\text{non-linearity error}} + \underbrace{\mathbb{E}[\widehat{\theta}(\mathbf{x}, r)|\mathbf{x}_1, \cdots, \mathbf{x}_n] - \theta(\mathbf{x}, r)}_{\text{smoothing bias}},$$

where $\widehat{\mathbf{H}}_\mathbf{x} = \frac{1}{n}\sum_{i=1}^n \mathbf{p}(\frac{\mathbf{x}_i - \mathbf{x}}{b})\mathbf{p}(\frac{\mathbf{x}_i - \mathbf{x}}{b})^\top b^{-d} K(\frac{\mathbf{x}_i - \mathbf{x}}{b})$, $\mathbf{H}_\mathbf{x} = \mathbb{E}[\mathbf{p}(\frac{\mathbf{x}_i - \mathbf{x}}{b})\mathbf{p}(\frac{\mathbf{x}_i - \mathbf{x}}{b})^\top b^{-d} K(\frac{\mathbf{x}_i - \mathbf{x}}{b})]$, and $\mathbf{S}_{\mathbf{x},r} = \frac{1}{n}\sum_{i=1}^n \mathbf{p}(\frac{\mathbf{x}_i - \mathbf{x}}{b}) b^{-d} K(\frac{\mathbf{x}_i - \mathbf{x}}{b})(r(y_i) - \mathbb{E}[r(y_i)|\mathbf{x}_i])$.

It follows immediately that the linear term is

$$\sqrt{nb^d}\mathbf{e}_1^\top \mathbf{H}_\mathbf{x}^{-1} \mathbf{S}_{\mathbf{x},r} = \frac{1}{\sqrt{nb^d}} \sum_{i=1}^n \mathcal{K}_\mathbf{x}\Big(\frac{\mathbf{x}_i - \mathbf{x}}{b}\Big)(r(y_i) - \mathbb{E}[r(y_i)|\mathbf{x}_i]) = R_n(g, r), \qquad g \in \mathcal{G}, r \in \mathcal{R}_l,$$

for $l = 1, 2$, and where $\mathcal{G} = \{b^{-d/2}\mathcal{K}_\mathbf{x}(\frac{\cdot - \mathbf{x}}{b}) : \mathbf{x} \in \mathcal{X}\}$ with $\mathcal{K}_\mathbf{x}(\mathbf{u}) = \mathbf{e}_1^\top \mathbf{H}_\mathbf{x}^{-1}\mathbf{p}(\mathbf{u})K(\mathbf{u})$ the equivalent boundary-adaptive kernel function. Furthermore, under the regularity conditions given in the supplemental appendix (Lemma SA.1), which relate to uniform smoothness and moment restrictions for the conditional distribution of $y_i|\mathbf{x}_i$, we have that

$$\sup_{\mathbf{x} \in \mathcal{X}, r \in \mathcal{R}_1} \left|\mathbf{e}_1^\top (\widehat{\mathbf{H}}_\mathbf{x}^{-1} - \mathbf{H}_\mathbf{x}^{-1}) \mathbf{S}_{\mathbf{x},r}\right| = O((nb^d)^{-1}\log n + (nb^d)^{-3/2}(\log n)^{5/2}) \qquad \text{a.s.},$$

$$\sup_{\mathbf{x} \in \mathcal{X}, r \in \mathcal{R}_2} \left|\mathbf{e}_1^\top (\widehat{\mathbf{H}}_\mathbf{x}^{-1} - \mathbf{H}_\mathbf{x}^{-1}) \mathbf{S}_{\mathbf{x},r}\right| = O((nb^d)^{-1}\log n) \qquad \text{a.s.},$$

$$\sup_{\mathbf{x} \in \mathcal{X}, r \in \mathcal{R}_l} \left|\mathbb{E}[\widehat{\theta}(\mathbf{x}, r)|\mathbf{x}_1, \cdots, \mathbf{x}_n] - \theta(\mathbf{x}, r)\right| = O(b^{1+\mathfrak{p}}) \qquad \text{a.s.}, \quad l = 1, 2.$$

Therefore, the goal reduces to establishing a Gaussian strong approximation for the residual-based empirical process $(R_n(g, r) : g \in \mathcal{G}, r \in \mathcal{R}_l)$, $l = 1, 2$. In the remaining of this subsection we discuss different attempts to establish such approximation result, culminating with the application of our Theorem 3.

As discussed in Chernozhukov *et al.* (2014, Remark 3.1), a first attempt is to deploy Theorem 1.1 in Rio (1994) (or, equivalently, Corollary 1). Viewing the empirical process as based on the random sample $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, $i = 1, 2, \cdots, n$, the theorem requires $\mathbb{P}_Z$ be continuously distributed with positive Lebesgue density on its support $\mathcal{X} = [0, 1]^{d+1}$ (using the notation of Assumption A). For this reason, Chernozhukov *et al.* (2014, Remark 3.1) assumes that $(\mathbf{x}_i, y_i) = (\mathbf{x}_i, \varphi(\mathbf{x}_i, u_i))$ where $(\mathbf{x}_i, u_i)$ has continuous and positive Lebesgue density supported on $\mathcal{X}$. Thus, if $\mathtt{M}_{\{\varphi\}} < \infty$, $\sup_{g \in \mathcal{G}} \mathtt{TV}_{\{\varphi\},\mathrm{supp}(g)} \lesssim \sup_{g \in \mathcal{G}} \mathfrak{m}(\mathrm{Supp}(g)) < \infty$, $\mathtt{K}_{\{\varphi\}} < \infty$, and other regularity conditions hold, then we show in the supplemental appendix (Example SA.1) that applying Rio (1994) to $(X_n(h) : h \in \mathcal{H})$ with $\mathcal{H} = \{(g \cdot \varphi) \circ \phi_Z^{-1}\}$, where ϕ_Z is the Rosenblatt transformation (see Lemma SA.12), gives a Gaussian strong approximation for $(R_n(g, r) : g \in \mathcal{G}, r \in \mathcal{R}_l)$, $l = 1, 2$, with rate (6). Without the condition on *local* uniform variation $\mathtt{K}_{\{\varphi\}} < \infty$, an additional $\sqrt{\log n}$ multiplicative factor appears.

The previous result does not exploit Lipschitz continuity, so a natural second attempt is to employ Corollary 2 to improve it. Retaining the same setup and assumptions, but now also assuming that φ is Lipschitz, our Theorem 1 gives a Gaussian strong approximation for $(R_n(g,r) : g \in \mathcal{G}, r \in \mathcal{R}_1)$ with rate (8). See Example SA.2 in the supplemental appendix. Importantly, Theorem 1 does not give an improvement for $\mathcal{R}_2$ because the Lipschitz condition is not satisfied.

The two attempts so far impose strong assumptions on the joint distribution of the data, and deliver approximation rates based on the incorrect effective sample size (and thus require $nb^{d+1} \to \infty$). Our Theorem 3 addresses both problems: suppose Assumption B holds and $K : \mathbb{R}^d \to \mathbb{R}$ is a compact supported Lipschitz continuous function, then we verify in the supplemental appendix (Example SA.3) that $\mathtt{M}_{\mathcal{G}} \lesssim b^{-d/2}$, $\mathtt{E}_{\mathcal{G}} \lesssim b^{d/2}$, $\mathtt{TV} \lesssim b^{d/2-1}$, and $\mathtt{L} \lesssim b^{-d/2-1}$, which gives $\|R_n - Z_n^R\|_{\mathcal{G}\times\mathcal{R}_2} = O(\varrho_n)$ a.s. with

$$\varrho_n = (nb^d)^{-1/(d+2)}\sqrt{\log n} + (nb^d)^{-1/2}\log n.$$

If, in addition, we assume $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[\exp(y_i)|\mathbf{x}_i = \mathbf{x}] < \infty$, then $\|R_n - Z_n^R\|_{\mathcal{G}\times\mathcal{R}_1} = O(\varrho_n)$ a.s. with

$$\varrho_n = (nb^d)^{-1/(d+2)}\sqrt{\log n} + (nb^d)^{-1/2}(\log n)^2.$$

As a consequence, our results verify that there exist valid uniform Gaussian approximations as follows:

- Let $\widehat{\mu}(\mathbf{x}) := \widehat{\theta}(\mathbf{x}; r)$ for $r \in \mathcal{R}_1$. If $b^{\mathfrak{p}+1}(nb^d)^{(d+4)/(2d+4)}(\log n)^{-1/2} + (nb^d)^{-(d+1)/(d+2)}(\log n)^2 = O(1)$, then

$$\sup_{\mathbf{x}\in\mathcal{X}} \left|\sqrt{nb^d}\big(\widehat{\mu}(\mathbf{x}) - \mu(\mathbf{x})\big) - Z_n^R(\mathbf{x})\right| \lesssim \left(\frac{(\log n)^{1+d/2}}{nb^d}\right)^{\frac{1}{d+2}} \qquad \text{a.s.},$$

  where $\mathbb{C}\mathrm{ov}(Z_n^R(\mathbf{x}), Z_n^R(\mathbf{x}')) = nb^d\mathbb{C}\mathrm{ov}(\mathbf{e}_1^\top \mathbf{H}_{\mathbf{x}}^{-1}\mathbf{S}_{\mathbf{x},r}, \mathbf{e}_1^\top \mathbf{H}_{\mathbf{x}'}^{-1}\mathbf{S}_{\mathbf{x}',r})$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and $r \in \mathcal{R}_1$.

- Let $\widehat{F}(r_y|\mathbf{x}) := \widehat{\theta}(\mathbf{x}; r_y)$ for $r_y \in \mathcal{R}_2$. If $b^{\mathfrak{p}+1}(nb^d)^{(d+4)/(2d+4)}(\log n)^{-1/2} = O(1)$, and also $(nb^d)^{-1}\log n = o(1)$, then

$$\sup_{\mathbf{x}\in\mathcal{X}, y\in\mathbb{R}} \left|\sqrt{nb^d}\big(\widehat{F}(y|\mathbf{x}) - F(y|\mathbf{x})\big) - Z_n^R(y,\mathbf{x})\right| \lesssim \left(\frac{(\log n)^{1+d/2}}{nb^d}\right)^{\frac{1}{d+2}} \qquad \text{a.s.},$$

  where $\mathbb{C}\mathrm{ov}(Z_n^R(\mathbf{x}), Z_n^R(\mathbf{x}')) = nb^d\mathbb{C}\mathrm{ov}(\mathbf{e}_1^\top \mathbf{H}_{\mathbf{x}}^{-1}\mathbf{S}_{\mathbf{x},r_y}, \mathbf{e}_1^\top \mathbf{H}_{\mathbf{x}'}^{-1}\mathbf{S}_{\mathbf{x}',r_{y'}})$ for all $(\mathbf{x},y), (\mathbf{x}',y') \in \mathcal{X}\times\mathbb{R}$ and $r_y, r_{y'} \in \mathcal{R}_2$.

This example gives a substantive statistical application where Theorem 3 offers a strict improvement on the accuracy of the Gaussian strong approximation over Rio (1994), and over Theorem 1 after incorporating the additional Lipschitz condition on the class of functions when applicable. It remains an open question whether the result in this section provides the best Gaussian strong approximation for local empirical processes or, in particular, for the local polynomial regression estimator. The results obtained are the best known in the literature to our knowledge, but we are unaware of lower bounds that would confirm the approximation rates are unimprovable.

4.2 Quasi-Uniform Haar Basis

In Section 3.2, we showed that when $\mathcal{H}$ lies in the span of a Haar basis, the Gaussian strong approximation rate can be optimal in the sense of achieving the univariate KMT approximation rate as a function of the effective sample size. This was a consequence of having no L_2-projection error in the construction of the strong approximation. In this section, we leverage the same idea to show that when $\mathcal{G}$ lies in the span of a Haar basis, it is possible to achieve nearly optimal Gaussian strong approximation rates for local empirical processes. This result has direct applicability to regression estimators based on Haar basis, including certain regression trees (Breiman *et al.*, 1984) and nonparametric partitioning-based estimators (Cattaneo *et al.*, 2020).

The following theorem gives our main result, which does not require that $\mathcal{R}$ lies in a Haar space, thereby highlighting once again the asymmetric roles that $\mathcal{G}$ and $\mathcal{R}$ play.

Theorem 4. *Suppose* $(\mathbf{z}_i = (\mathbf{x}_i, y_i), 1 \leq i \leq n)$ *are i.i.d. random variables taking values in* $(\mathcal{X} \times \mathbb{R}, \mathcal{B}(\mathcal{X} \times \mathbb{R}))$ *with* $\mathcal{X} \subseteq \mathbb{R}^d$*, and the following conditions hold.*

(i) $\mathcal{G}$ *is a class of functions on* $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P}_X)$ *such that* $\mathtt{M}_\mathcal{G} < \infty$ *and* $\mathcal{G} \subseteq \operatorname{Span}\{\mathbb{1}_{\Delta_l} : 0 \leq l < L\}$*, where* $\{\Delta_l : 0 \leq l < L\}$ *forms a quasi-uniform partition of* $\mathcal{X}$ *in the sense that*

$$\mathcal{X} \subseteq \sqcup_{0 \leq l < L} \Delta_l \qquad \text{and} \qquad \frac{\max_{0 \leq l < L} \mathbb{P}_X(\Delta_l)}{\min_{0 \leq l < L} \mathbb{P}_X(\Delta_l)} \leq \rho < \infty.$$

In addition, $\mathcal{G}$ *is a VC-type class with respect to envelope function* $\mathtt{M}_\mathcal{G}$ *with constant* $\mathtt{c}_\mathcal{G} \geq e$ *and exponent* $\mathtt{d}_\mathcal{G} \geq 1$.

(ii) $\mathcal{R}$ *is a real-valued pointwise measurable class of functions on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_Y)$*, and a VC-type class with respect to* $M_\mathcal{R}$ *with constant* $\mathtt{c}_\mathcal{R} \geq e$ *and exponent* $\mathtt{d}_\mathcal{R} \geq 1$*. Furthermore, one of the following holds:*

  (a) $\mathtt{M}_\mathcal{R} \lesssim 1$ *and* $\mathtt{pTV}_\mathcal{R} \lesssim 1$*, and set* $\alpha = 0$*, or*

  (b) $M_\mathcal{R}(y) \lesssim 1 + |y|^\alpha$*,* $\mathtt{pTV}_{\mathcal{R},(-|y|,|y|)} \lesssim 1 + |y|^\alpha$ *for all* $y \in \mathbb{R}$ *and for some* $\alpha > 0$*, and* $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(y_i) | \mathbf{x}_i = \mathbf{x}] \leq 2$.

(iii) *There exists a constant* $\mathtt{c}_5$ *such that* $|\log_2 \mathtt{E}_\mathcal{G}| + |\log_2 \mathtt{M}_\mathcal{G}| + |\log_2 L| \leq \mathtt{c}_5 \log_2 n$.

Then, on a possibly enlarged probability space, there exists mean-zero Gaussian processes $(Z_n^R(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$ *with almost sure continuous trajectory such that:*

- $\mathbb{E}[R_n(g_1, r_1) R_n(g_2, r_2)] = \mathbb{E}[Z_n^R(g_1, r_1) Z_n^R(g_2, r_2)]$ *for all* $(g_1, r_1), (g_2, r_2) \in \mathcal{G} \times \mathcal{R}$*, and*
- $\mathbb{P}[\|R_n - Z_n^R\|_{\mathcal{G} \times \mathcal{R}} > C_1 C_\alpha (C_\rho \mathsf{U}_n(t) + \mathsf{V}_n(t))] \leq C_2 e^{-t} + L e^{-C_\rho n/L}$ *for all* $t > 0$,

where C_1 *and* C_2 *are universal constants,* $C_\alpha = \max\{1 + (2\alpha)^{\frac{\alpha}{2}}, 1 + (4\alpha)^\alpha\}$*,* C_ρ *is a constant that only depends on* ρ*,*

$$\mathsf{U}_n(t) := \sqrt{\frac{d \mathtt{M}_\mathcal{G} \mathtt{E}_\mathcal{G}}{n/L}} (t + \mathtt{c}_5 \log_2(n) + \mathtt{d} \log(\mathtt{c} n))^{\alpha+1} + \frac{\mathtt{M}_\mathcal{G}}{\sqrt{n}} (\log n)^\alpha (t + \mathtt{c}_5 \log_2(n) + \mathtt{d} \log(\mathtt{c} n))^{\alpha+1}$$

with $\mathtt{c} = \mathtt{c}_{\mathcal{G}}\mathtt{c}_{\mathcal{R}}$, $\mathtt{d} = \mathtt{d}_{\mathcal{G}} + \mathtt{d}_{\mathcal{R}}$, *and*

$$\mathsf{V}_n(t) := \mathbb{1}(\operatorname{card}(\mathcal{R}) > 1)\sqrt{\mathtt{M}_{\mathcal{G}}\mathtt{E}_{\mathcal{G}}}\Big(\max_{0\leq l<L}\|\Delta_l\|_\infty\Big)\mathtt{L}_{\mathcal{V}_{\mathcal{R}}}\sqrt{t + \mathtt{c}_5\log_2(n) + \mathtt{d}\log(\mathtt{c}n)},$$

with $\mathcal{V}_{\mathcal{R}} := \{v_r : \mathbf{x} \mapsto \mathbb{E}[r(y_i)|\mathbf{x}_i = \mathbf{x}], \mathbf{x} \in \mathcal{X}, r \in \mathcal{R}\}$.

The first term ($\mathsf{U}_n(t)$) can be interpreted as a "variance" contribution based on "effective sample size" n/L, up to polylog(n) terms, while the second term ($\mathsf{V}_n(t)$) can be interpreted as a "bias" term that arises from the projection error for the conditional mean function $\theta(\cdot, r)$, which may not necessarily lie in the span of Haar basis. In the special case when $\mathcal{R} = \{r\}$ is a singleton we can construct the cells based on the condition distribution of $r(y_i) - \mathbb{E}[r(y_i)|\mathbf{x}_i]$, thereby making the conditional mean function (and hence the "bias" term) zero, while that is not possible when uniformity over $\mathcal{R}$ is desired.

Theorem 4 gives the following uniform Gaussian strong approximation result.

Corollary 6 (Haar Basis Residual Empirical Process)**.** *Suppose the conditions of Theorem 4 hold. Then,* $\|R_n - Z_n^R\|_{\mathcal{G}\times\mathcal{R}} = O(\varrho_n)$ *a.s. with*

$$\varrho_n = \sqrt{\frac{\mathtt{M}_{\mathcal{G}}\mathtt{E}_{\mathcal{G}}}{n/L}}(\log n)^{\alpha+1} + \frac{\mathtt{M}_{\mathcal{G}}}{\sqrt{n}}(\log n)^{2\alpha+1} + \mathbb{1}(\operatorname{card}(\mathcal{R}) > 1)\sqrt{\mathtt{M}_{\mathcal{G}}\mathtt{E}_{\mathcal{G}}}\Big(\max_{0\leq l<L}\|\Delta_l\|_\infty\Big)\sqrt{\log n}$$

Setting aside the roles of $\mathtt{M}_{\mathcal{G}}$ and $\mathtt{E}_{\mathcal{G}}$, the approximation rate is effectively $(\log n)^{\alpha+1}(n/L)^{-1/2} + \mathbb{1}(\operatorname{card}(\mathcal{R}) > 1)\max_{0\leq l<L}\|\Delta_l\|_\infty\sqrt{\log n}$, which can achieve the optimal univariate KMT strong approximation rate based on the effective sample size n/L, up to a polylog(n) term, when $\mathcal{R}$ is a singleton function class.

We illustrate the applicability to statistics of Theorem 4 with the following example considering nonparametric regression based on Haar basis approximation.

Example 3 (Haar Basis Regression Estimators)**.** Suppose ($\mathbf{z}_i = (\mathbf{x}_i, y_i), 1 \leq i \leq n$) are i.i.d. random variables taking values in $(\mathcal{X} \times \mathbb{R}, \mathcal{B}(\mathcal{X} \times \mathbb{R}))$ with $\mathcal{X} \subseteq \mathbb{R}^d$. As in Section 4.1, consider the regression estimand (12), focusing once again on the two leading examples $\mathcal{R}_1$ and $\mathcal{R}_2$. However, instead of local polynomial regression, now consider the Haar partitioning-based estimator:

$$\check{\theta}(\mathbf{x}, r) = \mathbf{p}(\mathbf{x})^\top\widehat{\boldsymbol{\gamma}}(r), \qquad \widehat{\boldsymbol{\gamma}}(r) = \operatorname*{argmin}_{\mathbf{g}\in\mathbb{R}^L}\sum_{i=1}^n\big(r(y_i) - \mathbf{p}(\mathbf{x}_i)^\top\mathbf{g}\big)^2,$$

where $\mathbf{p}(\mathbf{u}) = (\mathbb{1}(\mathbf{u} \in \Delta_l) : 0 \leq l < L)$ and $\{\Delta_l : 0 \leq l < L\}$ forms a quasi-uniform partition of $\mathcal{X}$ as defined in Theorem 4. The estimation error can again be decomposed into three terms (linearization, non-linearity error, and smoothing bias)

$$\check{\theta}(\mathbf{x}, r) - \theta(\mathbf{x}, r) = \underbrace{\mathbf{p}(\mathbf{x})^\top\mathbf{Q}^{-1}\mathbf{T}_r}_{\text{linearization}} + \underbrace{\mathbf{p}(\mathbf{x})^\top(\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1})\mathbf{T}_r}_{\text{non-linearity error}} + \underbrace{\mathbb{E}[\check{\theta}(\mathbf{x}, r)|\mathbf{x}_1, \cdots, \mathbf{x}_n] - \theta(\mathbf{x}, r)}_{\text{smoothing bias}},$$

where $\mathbf{Q} = \mathbb{E}[\mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\top]$, $\widehat{\mathbf{Q}} = \frac{1}{n}\sum_{i=1}^n \mathbf{p}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^\top$, and $\mathbf{T}_r = \frac{1}{n}\sum_{i=1}^n \mathbf{p}(\mathbf{x}_i)(r(y_i) - \mathbb{E}[r(y_i)|\mathbf{x}_i])$. In this example, the linear term takes the form

$$\sqrt{n/L}\mathbf{p}(\mathbf{x})^\top \mathbf{Q}^{-1}\mathbf{T}_r = \frac{1}{\sqrt{n}}\sum_{i=1}^n k_{\mathbf{x}}(\mathbf{x}_i)(r(y_i) - \mathbb{E}[r(y_i)|\mathbf{x}_i]) = R_n(g,r), \qquad g \in \mathcal{G}, r \in \mathcal{R}_l,$$

for $l = 1, 2$, where $\mathcal{G} = \{k_{\mathbf{x}}(\cdot) : \mathbf{x} \in \mathcal{X}\}$ with $k_{\mathbf{x}}(\mathbf{u}) = L^{-1/2}\sum_{0 \le l < L} \mathbb{1}(\mathbf{x} \in \Delta_l)\mathbb{1}(\mathbf{u} \in \Delta_l)/\mathbb{P}_X(\Delta_l)$ the equivalent kernel. Under standard regularity conditions including smoothness and moment assumptions (Lemma SA.2 in the supplemental appendix), we verify that

$$\begin{aligned}
\sup_{r \in \mathcal{R}_1} \left|\mathbf{e}_1^\top(\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1})\mathbf{T}_r\right| &= O(\log(nL)L/n + (\log(nL)L/n)^{3/2}\log n) \qquad \text{a.s.},\\
\sup_{r \in \mathcal{R}_2} \left|\mathbf{e}_1^\top(\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1})\mathbf{T}_r\right| &= O(\log(nL)L/n) \qquad \text{a.s.},\\
\sup_{\mathbf{x} \in \mathcal{X}, r \in \mathcal{R}_l} \left|\mathbb{E}[\check{\theta}(\mathbf{x},r)|\mathbf{x}_1, \cdots, \mathbf{x}_n] - \theta(\mathbf{x},r)\right| &= O\big(\max_{0 \le l < L}\|\Delta_l\|_\infty\big) \qquad \text{a.s.}, \quad l = 1, 2.
\end{aligned}$$

Finally, for the residual-based empirical process $(R_n(g,r) : g \in \mathcal{G}, r \in \mathcal{R}_l)$, $l = 1, 2$, we apply Theorem 4. First, $\mathtt{M}_{\mathcal{G}} = L^{1/2}$ and $\mathtt{E}_{\mathcal{G}} = L^{-1/2}$, and we can take $\mathtt{c}_{\mathcal{G}} = L$ and $\mathtt{d}_{\mathcal{G}} = 1$ because $\mathcal{G}$ has finite cardinality L. For the singleton case $\mathcal{R}_1$, we can take $\mathtt{c}_{\mathcal{R}_1} = 1$ and $\mathtt{d}_{\mathcal{R}_1} = 1$, and Condition (ii)(a) in Theorem 4 holds, which implies that $\|R_n - Z_n^R\|_{\mathcal{G}\times\mathcal{R}_1} = O(\varrho_n)$ a.s. with

$$\varrho_n = \frac{(\log(nL))^2}{\sqrt{n/L}},$$

provided that $(\log(nL)L/n \to 0$. For the VC-Type class $\mathcal{R}_2$, we can verify Condition (ii)(b) in Theorem 4 with $\alpha = 1$ if $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[\exp(y_i)|\mathbf{x}_i = \mathbf{x}] \le 2$, and we can take $\mathtt{c}_{\mathcal{R}_2}$ to be some absolute constant and $\mathtt{d}_{\mathcal{R}_2} = 2$ by van der Vaart and Wellner (2013, Theorem 2.6.7), which implies that $\|R_n - Z_n^R\|_{\mathcal{G}\times\mathcal{R}_1} = O(\varrho_n)$ a.s. with

$$\varrho_n = \frac{\log(nL)}{\sqrt{n/L}} + \max_{0 \le l < L}\|\Delta_l\|_\infty,$$

provided that $(\log(nL)L/n \to 0$.

A uniform Gaussian strong approximation for $(\sqrt{n/L}(\check{\theta}(\mathbf{x},r) - \theta(\mathbf{x},r)) : (\mathbf{x},r) \in \mathcal{X}\times\mathcal{R}_l)$, $l = 1, 2$, follows directly from the results obtained above, as previously discussed in Section 4.1. ▲

This example illustrates a substantive statistical application where the optimal univariate KMT strong approximation rate based on the effective sample size n/L, up to polylog(n) terms and the complexity of $\mathcal{R}$.

5 Acknowledgments

We thank Rajita Chandak, Jianqing Fan, Kengo Kato, Jason Klusowski, Xinwei Ma, Boris Shigida, Rocio Titiunik, and Will Underwood for comments. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grant DMS-2210561.

References

Ambrosio, L., Fusco, N., and Pallara, D. (2000). *Functions of bounded variation and free discontinuity problems*: Oxford university press.

Beck, J. (1985). "Lower bounds on the approximation of the multivariate empirical process," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *70*, 289–306.

Berthet, P. and Mason, D. M. (2006). "Revisiting two strong approximation results of Dudley and Philipp," *Lecture Notes–Monograph Series*, *51*, 155–172.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. J. (1984). *Classification and Regression Trees*: Chapman and Hall/CRC.

Bretagnolle, J. and Massart, P. (1989). "Hungarian Constructions from the Nonasymptotic Viewpoint," *Annals of Probability*, *17*(1), 239–256.

Brown, L. D., Cai, T. T., and Zhou, H. H. (2010). "Nonparametric regression in exponential families," *Annals of Statistics*, *38*(4), 2005–2046.

Cattaneo, M. D., Chandak, R., Jansson, M., and Ma, X. (2024a). "Local Polynomial Conditional Density Estimators," *Bernoulli*.

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

Cattaneo, M. D., Feng, Y., and Underwood, W. G. (2024b). "Uniform Inference for Kernel Density Estimators with Dyadic Data," *Journal of the American Statistical Association*.

Cattaneo, M. D., Jansson, M., and Ma, X. (2024c). "Local Regression Distribution Estimators," *Journal of Econometrics*.

Cattaneo, M. D., Masini, R. P., and Underwood, W. G. (2024d). "Yurinskii's Coupling for Martingales," *arXiv preprint arXiv:2210.00362*.

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

Csörgó, M. and Revész, P. (1981). *Strong Approximations in Probability and Statistics*, Probability and Mathematical Statistics : a series of monographs and textbooks: Academic Press.

Dedecker, J., Rio, E., and Merlevède, F. (2014). "Strong approximation of the empirical distribution function for absolutely regular sequences in $\mathbb{R}^d$," *Electronic Journal of Probability*, *19*(9), 1 – 56.

Einmahl, U. and Mason, D. M. (1998). "Strong Approximations to the Local Empirical Process," In *High Dimensional Probability*: Springer, 75–92.

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

Giné, E., Koltchinskii, V., and Sakhanenko, L. (2004). "Kernel Density Estimators: Convergence in Distribution for Weighted Sup-Norms," *Probability Theory and Related Fields*, *130*(2), 167–198.

Giné, E. and Nickl, R. (2010). "Confidence Bands in Density Estimation," *Annals of Statistics*, *38*(2), 1122–1170.

Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-dimensional Statistical Models*: Cambridge University Press.

Huang, J. (2003). "Local Asymptotics for Polynomial Spline Regression," *Annals of Statistics*, *31*(5), 1600–1635.

Koltchinskii, V. I. (1994). "Komlós-Major-Tusnády approximation for the general empirical process and Haar expansions of classes of functions," *Journal of Theoretical Probability*, *7*(1), 73–118.

Komlós, J., Major, P., and Tusnády, G. (1975). "An approximation of partial sums of independent RV'-s, and the sample DF. I," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *32*, 111–131.

Mason, D. M. and Van Zwet, W. R. (2011). "A Refinement of the KMT Inequality for the Uniform Empirical Process," In *Selected Works of Willem van Zwet*: Springer, 415–428.

Mason, D. M. and Zhou, H. H. (2012). "Quantile Coupling Inequalities and Their Applications," *Probability Surveys*, 39–479.

Massart, P. (1989). "Strong approximation for multivariate empirical and related processes, via KMT constructions," *Annals of probability*, 266–291.

Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*: Cambridge University Press.

Rio, E. (1994). "Local Invariance Principles and Their Application to Density Estimation," *Probability Theory and Related Fields*, *98*(1), 21–45.

Sakhanenko, A. (1996). "Estimates for the accuracy of coupling in the central limit theorem," *Siberian Mathematical Journal*, *37*(4), 811–823.

Sakhanenko, L. (2015). "Asymptotics of Suprema of Weighted Gaussian Fields with Applications to Kernel Density Estimators," *Theory of Probability & Its Applications*, *59*(3), 415–451.

Settati, A. (2009). "Gaussian approximation of the empirical process under random entropy conditions," *Stochastic processes and their Applications*, *119*(5), 1541–1560.

van der Vaart, A. and Wellner, J. (2013). *Weak convergence and empirical processes: with applications to statistics*: Springer Science & Business Media.

Wand, M. and Jones, M. (1995). *Kernel Smoothing*: Chapman & Hall/CRC.

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

Zaitsev, A. Y. (1987). "Estimates for the Lévy-Prokhorov distance in the multidimensional central limit theorem for random vectors with finite exponential moments," *Theory of Probability & its Applications*, *31*(2), 203–220.

Zaitsev, A. Y. (2013). "The Accuracy of Strong Gaussian Approximation for Sums of Independent Random Vectors," *Russian Mathematical Surveys*, *68*(4), 721–761.