Submitted to the Annals of Statistics

STRONG APPROXIMATIONS FOR EMPIRICAL PROCESSES INDEXED BY LIPSCHITZ FUNCTIONS

BY MATIAS D. CATTANEO[1,a], AND RUIQI (RAE) YU[1,b]

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

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 possibly Lipschitz functions, improving on previous results in the literature. In the setting considered by [29], and if the function class is Lipschitzian, our result improves the approximation rate $n^{-1/(2d)}$ to $n^{-1/\max\{d,2\}}$, up to a $\operatorname{polylog}(n)$ term, where n denotes the sample size. Remarkably, we establish a valid uniform Gaussian strong approximation at the 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 [23]. Second, a uniform Gaussian strong approximation is established for multiplicative separable empirical processes indexed by possibly Lipschitz functions, which addresses some outstanding problems in the literature [13, Section 3]. Finally, two other uniform Gaussian strong approximation results are presented when the function class is a sequence of Haar basis based on quasi-uniform partitions. Applications to nonparametric density and regression estimation are discussed.

1. Introduction.

Let $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$, $i = 1, \dots, n$, 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 in law as $n \to \infty$ to a Gaussian process in $\ell^\infty(\mathcal{H})$, the space of uniformly bounded real functions on $\mathcal{H}$. This weak convergence 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 $\mathbb{E}[Z(h_1)Z(h_2)] = \mathbb{E}[h_1(\mathbf{x}_i)h_2(\mathbf{x}_i)] - \mathbb{E}[h_1(\mathbf{x}_i)]\mathbb{E}[h_2(\mathbf{x}_i)]$ for all $h_1, h_2 \in \mathcal{H}$ when $\mathcal{H}$ is not n-varying, or its limit as $n \to \infty$ otherwise. See [33] and [20] for textbook overviews.

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

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

MSC2020 subject classifications: Primary 60F17; secondary 62E17, 62G20.

Keywords and phrases: empirical processes, strong approximation, Gaussian approximation, uniform inference, local empirical process, nonparametric regression.

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 established using non-asymptotic probability concentration inequalities, which 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 Gaussian process Z_n is "pre-asymptotic", it can offer a better finite sample approximation to the sampling distribution of X_n than the large sample approximation based on the limiting Gaussian process Z in (2).

There is a large literature on strong approximations for empirical processes, offering different levels of tightness for the bound ϱ_n in (3). In particular, the univariate case ($d = 1$) is mostly settled. A major breakthrough was accomplished by [23, 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: $\mathbf{x}_i \sim \mathsf{Uniform}(\mathcal{X})$, $\mathcal{X} = [0, 1]$, and $\mathcal{H} = \{\mathbb{1}(\cdot \leq x) : x \in [0, 1]\}$, where $\mathbb{1}(\cdot)$ denotes the indicator function. See [5] and [25] for more technical discussions on the Hungarian construction, and [14], [24] and [28] for textbook overviews. The KMT result was later extended by [18] and [19] to univariate empirical processes indexed by functions with uniformly bounded total variation: for $\mathbf{x}_i \sim \mathbb{P}_X$ supported on $\mathcal{X} = \mathbb{R}$ and continuously distributed, the authors obtained

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

in (3), with $\mathcal{H}$ satisfying a bounded variation condition. More recently, [8, Lemma SA26] gave a self-contained proof of a slightly generalized KMT result allowing for a larger class of distributions $\mathbb{P}_X$. See Remark 1 for details. As a statistical application, the authors considered univariate kernel density estimation [34], 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 more difficult, since the KMT approach does not easily generalize to multivariate data. Foundational results include [27], [22], and [29]. 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, [29] 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 [2], and demonstrates an unfortunate dimension penalty in the convergence rate of the d-variate uniform Gaussian strong approximation. As a statistical application, the author 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 [29]'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, [29, 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 $\mathcal{H}$.

2. As discussed by [13, Section 3], applying [29]'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, [29]'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 [13] to develop 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 [29]'s result.

This paper presents new uniform Gaussian strong approximation results for empirical processes that address the two aforementioned limitations. Section 3 studies the general empirical process (1), and establishes a uniform Gaussian strong approximation explicitly allowing for the possibility that $\mathcal{H}$ is Lipschitzian (Theorem 1). This result not only encompasses, but also generalizes previous results in the literature by allowing for $d \geq 1$ under more generic entropy conditions and weaker conditions on the underlying data generating process. For comparison, if we impose the regularity conditions in [29] 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 improving (5), in addition to matching (4) when $d = 1$; see Remark 1 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 $d \geq 3$. 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 allows for other entropy notions for $\mathcal{H}$ beyond the classical VC-type condition, and delivers improvements over [22]. See Remark 2 for details. Section 3 discusses how our improvements are achieved, and outstanding technical roadblocks.

Section 4 is motivated by the second aforementioned limitation in prior uniform Gaussian strong approximation results, and thus studies the *residual-based empirical process*:

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

for $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, $i = 1,\dots,n$, a random sample now also including an outcome variable $y_i \in \mathbb{R}$. 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 an n-varying local smoother based on kernel, series, or nearest-neighbor methods, while $r(\cdot)$ is some transformation such as $r(y) = y$ for conditional mean or $r(y) = \mathbb{1}(y \leq \cdot)$ for conditional distribution estimation. [13, 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 the sample $(\mathbf{z}_i : 1 \leq i \leq n)$, and thus available strong approximation results can be applied directly, including [22], [29], and our new Theorem 1. However, those off-the-shelf results require stringent assumptions and can deliver suboptimal approximation rates. First, available results require $\mathbf{z}_i$ to admit a bounded and positive Lebesgue density on $[0,1]^{d+1}$, possibly after some specific transformation, 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 $g(\cdot)$ denotes n-varying local smoothing weights based on a kernel function with bandwidth $b \to 0$ as $n \to \infty$, and $r(y) = y$, [29] 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 suboptimal because nb^d is the (pointwise) effective sample size for the kernel regression estimator.

A key observation underlying the potential suboptimality 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. Thus, Section 4 presents a novel uniform Gaussian strong approximation for the residual-based empirical process (Theorem 2), which explicitly exploits the multiplicative separability of $\mathcal{H} = \mathcal{G} \times \mathcal{R}$ and the possibly Lipschitz continuity of the function class, while also removing stringent assumptions imposed on the underlying data generating process. When applied to the local kernel regression empirical processes, our best result gives a uniform Gaussian strong approximation rate

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

thereby improving over both [29] leading to (5), and Theorem 1 leading to (8). The correct effective sample size nb^d is achieved, under weaker regularity conditions. As a statistical application, Section 4.1 leverages Theorem 2 to establish the best known uniform Gaussian strong approximation result for local polynomial regression estimators [17].

Following [29], the proof of Theorem 1 in Section 3 first approximates 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 [28, Chapter 10, for a textbook introduction] to construct a strong approximation. It thus requires balancing two approximation errors: a projection error ("bias") emerging from the mean square projection based on a Haar basis, and a coupling error ("variance") 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 to improve the projection error, but no coupling methods available in the literature for the resulting projected process. The proof of Theorem 2 in Section 4 employs a similar projection and coupling decomposition approach, but treats $\mathcal{G}$ and $\mathcal{R}$ separately in order to leverage the multiplicative separability of the residual process $(R_n(g,r) : (g,r) \in \mathcal{G} \times \mathcal{R})$. In particular, the proof designs cells for projection and coupling approximation that are asymmetric in the direction of $\mathbf{x}_i$ and y_i components to obtain the uniform Gaussian strong approximation. This distinct proof strategy relaxes some underlying assumptions (most notably, on the distribution of y_i), and delivers a better strong approximation rate for some local empirical processes than what would be obtained by directly applying Theorem 1.

In general, however, neither Theorem 1 nor Theorem 2 dominates each other, nor their underlying assumptions imply each other, and therefore both are of interest, depending on the statistical problem under consideration. Their proofs employ different strategies (most notably, in terms of the dyadic cells expansion used) to leverage the specific structure of X_n and R_n. It is an open question whether the uniform Gaussian strong approximation rates obtained from Theorems 1 and 2 are optimal under the assumptions imposed.

As a way to circumvent the technical limitations underlying the proof strategies of Theorem 1 and Theorem 2, Section 5 presents two other uniform Gaussian strong approximation results when $\mathcal{H}$ is spanned by a possibly increasing sequence of finite Haar functions on *quasi-uniform* partitions, for the general empirical process (Theorem 3) and for the residual-based empirical process (Theorem 4). These theorems shut down the projection error, and also rely on a generalized Tusnády's Lemma established in this paper, to establish valid couplings over more general partitioning schemes and under weaker regularity conditions. 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 $d \geq 1$, up to a $\operatorname{polylog}(n)$ term, where $\operatorname{polylog}(n) = \log^a(n)$ for some $a > 0$, and possibly an additional "bias" term induced exclusively by the cardinality of $\mathcal{R}$. As statistical applications, we establish uniform Gaussian strong approximations for the classical histogram density estimator, and for Haar partitioning-based regression estimators such as those arising in the context of certain regression tree and related nonparametric methods [4, 21, 7].

The supplemental appendix [11] contains all technical proofs, additional theoretical results of independent interest, and other omitted details.

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 introductions and overviews, see [14], [24], [16], [3], [26], [20], [28], [37], and references therein. See also [13, Section 3] for discussion and further references concerning local empirical processes and their role in nonparametric curve estimation.

The celebrated KMT construction [23], Yurinskii's coupling [35], and Zaitsev's coupling [36] are three well-known approaches that can be used to establish a uniform Gaussian strong approximation 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 [27], [22], [29], [18], and [19], among others. As summarized in the introduction, our first main result (Theorem 1) encompasses and improves on 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 literature [13, 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: for example, see [10] for results based on Yurinskii's coupling, and [32] for results based on Zaitsev's coupling. Finally, employing a different approach, [15] obtain a uniform Gaussian strong approximation for the multivariate empirical process indexed by half plane indicators with a dimension-independent approximation rate, up to $\operatorname{polylog}(n)$ terms.

2. Notation.

We employ standard notations from the empirical process literature, suitably modified and specialized to improve exposition. See, for example, [1], [33] and [20] for background definitions and more details.

The q-dimensional Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^q$ and symmetric positive semidefinite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{q \times q}$ is denoted by $\mathsf{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The binomial distribution with parameter $n \in \mathbb{N}$ and $p \in [0,1]$ is denoted by $\mathsf{Bin}(n,p)$. $|\mathcal{A}|$ denotes the cardinality of the set $\mathcal{A}$. For a vector $\mathbf{a} \in \mathbb{R}^q$, $\|\mathbf{a}\|$ denotes the Euclidean norm and $\|\mathbf{a}\|_\infty$ denotes the maximum norm of $\mathbf{a}$. For a matrix $\mathbf{A} \in \mathbb{R}^{q \times q}$, $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_d(\mathbf{A}) \geq 0$ denote the singular values of $\mathbf{A}$. For $1 \leq i_1 \leq j_2 \leq n$ and $1 \leq j_1 \leq j_2 \leq n$, $\mathbf{A}_{i_1:i_2,j_1:j_2}$ denotes the submatrix $(A_{ij})_{i_1 \leq i \leq i_2, j_1 \leq j \leq j_2}$ of $\mathbf{A}$, and $\mathbf{A}_{i_1,j_1:j_2}$, $\mathbf{A}_{i_1:i_2,j_1}$ are likewise defined. For sequences of real numbers, we write $a_n = o(b_n)$ if $\limsup_{n\to\infty} |\frac{a_n}{b_n}| = 0$, and write $a_n = O(b_n)$ if there exists some constant C and $N > 0$ such that $n > N$ implies $|a_n| \leq C|b_n|$. For sequences of random variables, we write $a_n = o_{\mathbb{P}}(b_n)$ if $\limsup_{n\to\infty} \mathbb{P}[|\frac{a_n}{b_n}| \geq \varepsilon] = 0$ for all $\varepsilon > 0$, and write $a_n = O_{\mathbb{P}}(b_n)$ if $\limsup_{M\to\infty} \limsup_{n\to\infty} \mathbb{P}[|\frac{a_n}{b_n}| \geq M] = 0$.

Let $\mathcal{U}, \mathcal{V} \subseteq \mathbb{R}^q$. We define $\mathcal{U} + \mathcal{V} = \{\mathbf{u} + \mathbf{v} : \mathbf{u} \in \mathcal{U}, \mathbf{v} \in \mathcal{V}\}$ and $\|\mathcal{U}\|_\infty = \sup\{\|\mathbf{u}_1 - \mathbf{u}_2\|_\infty : \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}\}$, and $\mathcal{B}(\mathcal{U})$ denotes the Borel σ-algebra generated by $\mathcal{U}$ and $\mathcal{B}(\mathcal{U}) \otimes \mathcal{B}(\mathcal{V})$ denotes the product σ-algebra. Let μ be a measure on $(\mathcal{U}, \mathcal{B}(\mathcal{U}))$, and $\phi : (\mathcal{V}, \mathcal{B}(\mathcal{V})) \mapsto (\mathcal{U}, \mathcal{B}(\mathcal{U}))$ be a measurable function. $\mu \circ \phi$ denotes the measure on $(\mathcal{V}, \mathcal{B}(\mathcal{V}))$ such that $\mu \circ \phi(V) = \mu(\phi(V))$ for any $V \in \mathcal{B}(\mathcal{V})$. For $R \in \mathcal{B}(\mathcal{U})$, let $\mu|_R$ be the restriction of μ on R, that is, $\mu|_R(U) = \mu(U \cap R)$ for all $U \in \mathcal{B}(\mathcal{U})$. Two measures μ and ν on the measure space $(\mathcal{U}, \mathcal{B}(\mathcal{U}))$ agree on $R \in \mathcal{B}(\mathcal{U})$ if $\mu|_R = \nu|_R$. The support of μ is $\mathrm{Supp}(\mu) = \mathrm{closure}(\cup\{U \in \mathcal{B}(\mathcal{U}) : \mu(U) \neq 0\})$. The Lebesgue measure is denoted by $\mathfrak{m}$. Let f be a real-valued function on the measure space $(\mathcal{U}, \mathcal{B}(\mathcal{U}), \mu)$. Define the L_p norms $\|f\|_{\mu,p} = (\int |f|^p d\mu)^{1/p}$ for $1 \leq p < \infty$ and $\|f\|_\infty = \sup_{\mathbf{x}\in\mathcal{U}} |f(\mathbf{x})|$, and let $\mathrm{Supp}(f) = \{\mathbf{u} \in \mathcal{U} : f(\mathbf{u}) > 0\}$ be the support of f. $L_p(\mu)$ is the class of all real-valued measurable functions f on $(\mathcal{U}, \mathcal{B}(\mathcal{U}))$ such that $\|f\|_{\mu,p} < \infty$, for $1 \leq p < \infty$. The semi-metric $\mathfrak{d}_\mu$ on $L_2(\mu)$ is defined by $\mathfrak{d}_\mu(f, g) = (\|f - g\|_{\mu,2}^2 - (\int f\, d\mu - \int g\, d\mu)^2)^{1/2}$, for $f, g \in L_2(\mu)$. Whenever it exits, $\nabla f(\mathbf{x})$ denotes the Jacobian matrix of f at $\mathbf{x}$. If $\mathcal{F}$ and $\mathcal{G}$ are two sets of functions from measure space $(\mathcal{U}, \mathcal{B}(\mathcal{U}))$ and $(\mathcal{V}, \mathcal{B}(\mathcal{V}))$ to $\mathbb{R}$, respectively, then $\mathcal{F} \times \mathcal{G}$ denotes the class of measurable functions $\{(f, g) : f \in \mathcal{F}, g \in \mathcal{G}\}$ from $(\mathcal{U} \times \mathcal{V}, \mathcal{B}(\mathcal{U}) \otimes \mathcal{B}(\mathcal{V}))$ to $\mathbb{R}$. For a measure μ on $(\mathcal{U} \times \mathcal{V}, \mathcal{B}(\mathcal{U}) \otimes \mathcal{B}(\mathcal{V}))$, the semi-metric $\mathfrak{d}_\mu$ on $\mathcal{G} \times \mathcal{R}$ is defined by $\mathfrak{d}_\mu((g_1, r_1), (g_2, r_2)) = (\|g_1 r_1 - g_2 r_2\|_{\mu,2}^2 - (\int g_1 r_1 d\mu - \int g_2 r_2 d\mu)^2)^{1/2}$. For a semi-metric space $(\mathcal{S}, d)$, the covering number $N(\mathcal{S}, d, \varepsilon)$ is the minimal number of balls $B_v(\varepsilon) = \{u : d(u, v) < \varepsilon\}$, $v \geq 1$, needed to cover $\mathcal{S}$.

2.1. *Main Definitions.* Let $\mathcal{F}$ be a class of measurable functions from a probability space $(\mathbb{R}^q, \mathcal{B}(\mathbb{R}^q), \mathbb{P})$ to $\mathbb{R}$. We introduce several definitions that capture 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 $\lim_{m\to\infty} g_m(\mathbf{u}) = f(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{R}^q$.

DEFINITION 2. Let $\mathrm{Supp}(\mathcal{F}) = \cup_{f\in\mathcal{F}} \mathrm{Supp}(f)$. A probability measure $\mathbb{Q}_{\mathcal{F}}$ on $(\mathbb{R}^q, \mathcal{B}(\mathbb{R}^q))$ is a surrogate measure for $\mathbb{P}$ with respect to $\mathcal{F}$ if

(i) $\mathbb{Q}_{\mathcal{F}}$ agrees with $\mathbb{P}$ on $\mathrm{Supp}(\mathbb{P}) \cap \mathrm{Supp}(\mathcal{F})$.
(ii) $\mathbb{Q}_{\mathcal{F}}(\mathrm{Supp}(\mathcal{F}) \setminus \mathrm{Supp}(\mathbb{P})) = 0$.

Let $\mathcal{Q}_{\mathcal{F}} = \mathrm{Supp}(\mathbb{Q}_{\mathcal{F}})$.

DEFINITION 3. For $q = 1$ and an interval $\mathcal{I} \subseteq \mathbb{R}$, the pointwise total variation of $\mathcal{F}$ over $\mathcal{I}$ is

$$\mathtt{pTV}_{\mathcal{F},\mathcal{I}} = \sup_{f\in\mathcal{F}} \sup_{P\geq 1} \sup_{\mathcal{P}_P\in\mathcal{I}} \sum_{i=1}^{P-1} |f(a_{i+1}) - f(a_i)|,$$

where $\mathcal{P}_P = \{(a_1, \ldots, a_P) : a_1 \leq \cdots \leq a_P\}$ denotes the collection of all partitions of $\mathcal{I}$.

DEFINITION 4. For a non-empty $\mathcal{C} \subseteq \mathbb{R}^q$, the total variation of $\mathcal{F}$ over $\mathcal{C}$ is

$$\mathtt{TV}_{\mathcal{F},\mathcal{C}} = \inf_{\mathcal{U}\in\mathcal{O}(\mathcal{C})} \sup_{f\in\mathcal{F}} \sup_{\phi\in\mathcal{D}_q(\mathcal{U})} \int_{\mathbb{R}^q} f(\mathbf{u})\, \mathrm{div}(\phi)(\mathbf{u}) d\mathbf{u} / \|\|\phi\|_2\|_\infty,$$

where $\mathcal{O}(\mathcal{C})$ denotes the collection of all open sets that contains $\mathcal{C}$, and $\mathcal{D}_q(\mathcal{U})$ denotes the space of infinitely differentiable functions from $\mathbb{R}^q$ to $\mathbb{R}^q$ with compact support contained in $\mathcal{U}$.

DEFINITION 5. For a non-empty $\mathcal{C} \subseteq \mathbb{R}^q$, the local total variation constant of $\mathcal{F}$ over $\mathcal{C}$, is a positive number $\mathtt{K}_{\mathcal{F},\mathcal{C}}$ such that for any cube $\mathcal{D} \subseteq \mathbb{R}^q$ with edges of length ℓ parallel to the coordinate axises,

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

DEFINITION 6. For a non-empty $\mathcal{C} \subseteq \mathbb{R}^q$, the envelopes of $\mathcal{F}$ over $\mathcal{C}$ are

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

DEFINITION 7. For a non-empty $\mathcal{C} \subseteq \mathbb{R}^q$, the Lipschitz constant of $\mathcal{F}$ over $\mathcal{C}$ is

$$\mathtt{L}_{\mathcal{F},\mathcal{C}} = \sup_{f\in\mathcal{F}} \sup_{\mathbf{u}_1,\mathbf{u}_2\in\mathcal{C}} \frac{|f(\mathbf{u}_1) - f(\mathbf{u}_2)|}{\|\mathbf{u}_1 - \mathbf{u}_2\|_\infty}.$$

DEFINITION 8. For a non-empty $\mathcal{C} \subseteq \mathbb{R}^q$, the L_1 bound of $\mathcal{F}$ over $\mathcal{C}$ is

$$\mathtt{E}_{\mathcal{F},\mathcal{C}} = \sup_{f\in\mathcal{F}} \int_{\mathcal{C}} |f| d\mathbb{P}.$$

DEFINITION 9. For a non-empty $\mathcal{C} \subseteq \mathbb{R}^q$, the uniform covering number of $\mathcal{F}$ with envelope $M_{\mathcal{F},\mathcal{C}}$ over $\mathcal{C}$ is

$$\mathtt{N}_{\mathcal{F},\mathcal{C}}(\delta, M_{\mathcal{F},\mathcal{C}}) = \sup_{\mu} N(\mathcal{F}, \|\cdot\|_{\mu,2}, \delta\|M_{\mathcal{F},\mathcal{C}}\|_{\mu,2}), \qquad \delta \in (0,\infty),$$

where the supremum is taken over all finite discrete measures on $(\mathcal{C}, \mathcal{B}(\mathcal{C}))$. We assume that $M_{\mathcal{F},\mathcal{C}}(\mathbf{u})$ is finite for every $\mathbf{u} \in \mathcal{C}$.

DEFINITION 10. For a non-empty $\mathcal{C} \subseteq \mathbb{R}^q$, the uniform entropy integral of $\mathcal{F}$ with envelope $M_{\mathcal{F},\mathcal{C}}$ over $\mathcal{C}$ is

$$J_{\mathcal{C}}(\delta, \mathcal{F}, M_{\mathcal{F},\mathcal{C}}) = \int_0^\delta \sqrt{1 + \log \mathtt{N}_{\mathcal{F},\mathcal{C}}(\varepsilon, M_{\mathcal{F},\mathcal{C}})} d\varepsilon,$$

where it is assumed that $M_{\mathcal{F},\mathcal{C}}(\mathbf{u})$ is finite for every $\mathbf{u} \in \mathcal{C}$.

DEFINITION 11. For a non-empty $\mathcal{C} \subseteq \mathbb{R}^q$, $\mathcal{F}$ is a VC-type class with envelope $M_{\mathcal{F},\mathcal{C}}$ over $\mathcal{C}$ if (i) $M_{\mathcal{F},\mathcal{C}}$ is measurable and $M_{\mathcal{F},\mathcal{C}}(\mathbf{u})$ is finite for every $\mathbf{u} \in \mathcal{C}$, and (ii) there exist $\mathtt{c}_{\mathcal{F},\mathcal{C}} > 0$ and $\mathtt{d}_{\mathcal{F},\mathcal{C}} > 0$ such that

$$\mathtt{N}_{\mathcal{F},\mathcal{C}}(\varepsilon, M_{\mathcal{F},\mathcal{C}}) \leq \mathtt{c}_{\mathcal{F},\mathcal{C}} \varepsilon^{-\mathtt{d}_{\mathcal{F},\mathcal{C}}}, \qquad \varepsilon \in (0,1).$$

DEFINITION 12. For a non-empty $\mathcal{C} \subseteq \mathbb{R}^q$, $\mathcal{F}$ is a polynomial-entropy class with envelope $M_{\mathcal{F},\mathcal{C}}$ over $\mathcal{C}$ if (i) $M_{\mathcal{F},\mathcal{C}}$ is measurable and $M_{\mathcal{F},\mathcal{C}}(\mathbf{u})$ is finite for every $\mathbf{u} \in \mathcal{C}$, and (ii) there exist $\mathtt{a}_{\mathcal{F},\mathcal{C}} > 0$ and $\mathtt{b}_{\mathcal{F},\mathcal{C}} > 0$ such that

$$\log \mathtt{N}_{\mathcal{F},\mathcal{C}}(\varepsilon, M_{\mathcal{F},\mathcal{C}}) \leq \mathtt{a}_{\mathcal{F},\mathcal{C}} \varepsilon^{-\mathtt{b}_{\mathcal{F},\mathcal{C}}}, \qquad \varepsilon \in (0,1).$$

If a surrogate measure $\mathbb{Q}_{\mathcal{F}}$ for $\mathbb{P}$ with respect to $\mathcal{F}$ has been assumed, and it is clear from the context, we drop the dependence on $\mathcal{C} = \mathcal{Q}_{\mathcal{F}}$ for all quantities in Definitions 4–12. That is, to save notation, we set $\mathtt{TV}_{\mathcal{F}} = \mathtt{TV}_{\mathcal{F},\mathcal{Q}_{\mathcal{F}}}$, $\mathtt{K}_{\mathcal{F}} = \mathtt{K}_{\mathcal{F},\mathcal{Q}_{\mathcal{F}}}$, $\mathtt{M}_{\mathcal{F}} = \mathtt{M}_{\mathcal{F},\mathcal{Q}_{\mathcal{F}}}$, $M_{\mathcal{F}}(\mathbf{u}) = M_{\mathcal{F},\mathcal{Q}_{\mathcal{F}}}(\mathbf{u})$, $\mathtt{L}_{\mathcal{F}} = \mathtt{L}_{\mathcal{F},\mathcal{Q}_{\mathcal{F}}}$, and so on, whenever there is no confusion.

3. General Empirical Process.

Let

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

and recall Section 2.1 and the notation conventions introduced there.

THEOREM 1. Suppose $(\mathbf{x}_i : 1 \leq i \leq n)$ are i.i.d. random vectors taking values in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with common law $\mathbb{P}_X$ supported on $\mathcal{X} \subseteq \mathbb{R}^d$, and the following conditions hold.

(i) $\mathcal{H}$ is a real-valued pointwise measurable class of functions on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P}_X)$.
(ii) There exists a surrogate measure $\mathbb{Q}_{\mathcal{H}}$ for $\mathbb{P}_X$ with respect to $\mathcal{H}$ such that $\mathbb{Q}_{\mathcal{H}} = \mathfrak{m} \circ \phi_{\mathcal{H}}$, where the *normalizing transformation* $\phi_{\mathcal{H}} : \mathcal{Q}_{\mathcal{H}} \mapsto [0,1]^d$ is a diffeomorphism.
(iii) $\mathtt{M}_{\mathcal{H}} < \infty$ 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 on $(\mathcal{H}, \mathfrak{d}_{\mathbb{P}_X})$ 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

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

where

$$\begin{aligned} \mathsf{A}_n(t,\delta) = {} & \min\big\{\mathsf{m}_{n,d}\sqrt{\mathtt{M}_{\mathcal{H}}}, \mathsf{l}_{n,d}\sqrt{\mathsf{c}_2 \mathsf{L}_{\mathcal{H}}}\big\}\sqrt{\mathsf{c}_1 \mathtt{TV}_{\mathcal{H}}}\sqrt{t + \log \mathtt{N}_{\mathcal{H}}(\delta, M_{\mathcal{H}})} \\ & + \sqrt{\frac{\mathtt{M}_{\mathcal{H}}}{n}} \min\big\{\sqrt{\log n}\sqrt{\mathtt{M}_{\mathcal{H}}}, \sqrt{\mathsf{c}_3 \mathsf{K}_{\mathcal{H}} + \mathtt{M}_{\mathcal{H}}}\big\}(t + \log \mathtt{N}_{\mathcal{H}}(\delta, M_{\mathcal{H}})) \end{aligned}$$

$$\mathsf{c}_1 = d \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} \prod_{j=1}^{d-1} \sigma_j(\nabla \phi_{\mathcal{H}}(\mathbf{x})), \qquad \mathsf{c}_2 = \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} \frac{1}{\sigma_d(\nabla \phi_{\mathcal{H}}(\mathbf{x}))}, \qquad \mathsf{c}_3 = 2^{d-1} d^{d/2-1} \mathsf{c}_1 \mathsf{c}_2^{d-1},$$

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 uniform Gaussian strong approximation theorem is given in full generality to accommodate different applications. Section 3.1 discusses the role of the surrogate measure and normalizing transformation, and Section 3.2 discusses leading special cases and compares our results to prior literature. The proof of Theorem 1 is in [11, Section SA-II], 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 oscillations) 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 $\mathsf{F}_n(t,\delta)$ emerging from Talagrand's inequality [20, Theorem 3.3.9] combined with a standard maximal inequality [13, Theorem 5.2].

Following [29], 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 denoting the L_2-projection onto piecewise constant functions on a carefully chosen partition of $\mathcal{X}$. We introduce a class of recursive *quasi-dyadic* cells expansion of $\mathcal{X}$, which we employ to generalize prior results in the literature, including 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 $\mathsf{Normal}(\frac{n}{2}, \frac{n}{4})$, due to the fact that the constant approximation within each recursive partitioning cell generates counts based on i.i.d. data. Building on the celebrated Tusnády's Lemma, [29, Theorem 2.1] established a remarkable coupling result for bounded functions L_2-projected on a dyadic cells expansion of $\mathcal{X}$. We build on his powerful ideas, and establish an analogous result for the case of Lipschitz functions L_2-projected on dyadic cells expansion of $\mathcal{X}$, thereby obtaining a tighter coupling error. A limitation of these results is that they only apply to a dyadic cells expansion due to the specifics of Tusnády's Lemma.

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 errors of the mean square projection onto a Haar basis based on *quasi-dyadic* cells expansion of $\mathcal{X}$. We handle this error using Bernstein inequality, while also taking into account explicitly the potential Lipschitz structure of the functions, and the more generic cell structure.

Balancing the coupling error and the two projection errors in (10) gives term $\mathsf{A}_n(t,\delta)$ in Theorem 1. Section SA-II of [11] provides all technical details, and additional results that may be of independent interest.

3.1. *Surrogate Measure and Normalizing Transformation.*

Theorem 1 assumes the existence of a surrogate measure $\mathbb{Q}_{\mathcal{H}}$, and a normalizing transformation $\phi_{\mathcal{H}}$, which together restrict $\mathbb{P}_X$ to be absolutely continuous with respect to $\mathfrak{m}$ on $\mathcal{X} \cap \operatorname{Supp}(\mathcal{H})$, while incorporating features of the support of $\mathcal{H}$. We provide examples of $\mathbb{Q}_{\mathcal{H}}$ and $\phi_{\mathcal{H}}$, discuss primitive sufficient conditions, and bound the constants c_1, c_2, and c_3 explicitly.

As a first simple example, suppose that $\mathbf{x}_i \sim \mathsf{Uniform}(\mathcal{X})$ with $\mathcal{X} = \times_{l=1}^d [\mathsf{a}_l, \mathsf{b}_l]$, where $-\infty < \mathsf{a}_l < \mathsf{b}_l < \infty$, $l = 1, 2, \ldots, d$. Setting $\mathbb{Q}_{\mathcal{H}} = \mathbb{P}_X$ and $\phi_{\mathcal{H}}(x_1, \cdots, x_d) = ((\mathsf{b}_1 - \mathsf{a}_1)^{-1}(x_1 - \mathsf{a}_1), \cdots, (\mathsf{b}_d - \mathsf{a}_d)^{-1}(x_d - \mathsf{a}_d))$ verifies assumption (ii) in Theorem 1. In this case, $\mathsf{c}_1 = d \max_{1 \leq l \leq d} |\mathsf{b}_l - \mathsf{a}_l| \prod_{l=1}^d |\mathsf{b}_l - \mathsf{a}_l|^{-1}$, $\mathsf{c}_2 = \max_{1 \leq l \leq d} |\mathsf{b}_l - \mathsf{a}_l|$ and $\mathsf{c}_3 = 2^{d-1} d^{d/2} \max_{1 \leq l \leq d} |\mathsf{b}_l - \mathsf{a}_l|^d \prod_{l=1}^d |\mathsf{b}_l - \mathsf{a}_l|^{-1}$.

When $\mathbb{P}_X$ is not the uniform distribution, or $\mathcal{X}$ is not isomorphic to the d-dimensional unit cube, a careful choice of $\mathbb{Q}_{\mathcal{H}}$ and $\phi_{\mathcal{H}}$ is needed. In many interesting cases, the *Rosenblatt transformation* can be used to exhibit a valid normalizing transformation, together with an appropriate choice of $\mathbb{Q}_{\mathcal{H}}$ taking into account $\mathcal{X}$ and $\operatorname{Supp}(\mathcal{H})$. For a random vector $\mathbf{V} = (V_1, \cdots, V_d) \in \mathbb{R}^d$ with distribution $\mathbb{P}_V$, the Rosenblatt transformation is

$$T_{\mathbb{P}_V}(v_1, \cdots, v_d) = \begin{bmatrix} \mathbb{P}_V(V_1 \leq v_1) \\ \mathbb{P}_V(V_2 \leq v_2 | V_1 = v_1) \\ \vdots \\ \mathbb{P}_V(V_d \leq v_d | V_1 = v_1, \cdots, V_{d-1} = v_{d-1}) \end{bmatrix}.$$

To discuss the role of the Rosenblatt transformation in constructing a valid normalizing transformation, we consider the following two cases.

Case 1: Rectangular $\mathcal{Q}_{\mathcal{H}}$. Suppose that $\mathbb{Q}_{\mathcal{H}}$ admits a Lebesgue density f_Q supported on $\mathcal{Q}_{\mathcal{H}} = \times_{l=1}^{d}[\mathsf{a}_l, \mathsf{b}_l]$, $-\infty \leq \mathsf{a}_l < \mathsf{b}_l \leq \infty$. Then, the Rosenblatt transformation $\phi_{\mathcal{H}} = T_{\mathbb{Q}_{\mathcal{H}}}$ is a normalizing transformation, and we obtain

$$\mathsf{c}_1 = d \sup_{\mathbf{u}\in\mathcal{Q}_{\mathcal{H}}} \frac{f_Q(\mathbf{u})}{\min\{f_{Q,1}(u_1), f_{Q,2|1}(u_2|u_1), \cdots, f_{Q,d|-d}(u_d|u_1, \cdots, u_{d-1})\}},$$

$$\mathsf{c}_2 = \sup_{\mathbf{u}\in\mathcal{Q}_{\mathcal{H}}} \frac{1}{\min\{f_{Q,1}(u_1), f_{Q,2|1}(u_2|u_1), \cdots, f_{Q,d|-d}(u_d|u_1, \cdots, u_{d-1})\}},$$

and $\mathsf{c}_3 = 2^{d-1}d^{d/2-1}\mathsf{c}_1\mathsf{c}_2^{d-1}$, where $f_{Q,j|-j}(\cdot|u_1, \cdots, u_{j-1})$ denotes the conditional density of $Q_j|Q_1 = u_1, \cdots, Q_{j-1} = u_{j-1}$ for $\mathbf{Q} = (Q_1, \cdots, Q_d) \sim \mathbb{Q}_{\mathcal{H}}$.

This case covers several examples of interest, which give primitive conditions for assumption (ii) in Theorem 1:

(a) Suppose $\mathcal{Q}_{\mathcal{H}} = \times_{l=1}^{d}[\mathsf{a}_l, \mathsf{b}_l]$ is bounded. Then, for f_Q bounded and bounded away from zero on $\mathcal{Q}_{\mathcal{H}}$,

$$\mathsf{c}_1 \leq d\frac{\overline{f}_Q^2}{\underline{f}_Q}\overline{\mathcal{Q}}_{\mathcal{H}} \qquad \text{and} \qquad \mathsf{c}_2 \leq \frac{\overline{f}_Q}{\underline{f}_Q}\overline{\mathcal{Q}}_{\mathcal{H}},$$

where $\underline{f}_Q = \inf_{\mathbf{x}\in\mathcal{Q}_{\mathcal{H}}} f_X(\mathbf{x})$, $\overline{f}_Q = \sup_{\mathbf{x}\in\mathcal{Q}_{\mathcal{H}}} f_Q(\mathbf{x})$, and $\overline{\mathcal{Q}}_{\mathcal{H}} = \max_{1\leq l\leq d}|\mathsf{b}_l - \mathsf{a}_l|$. If $\mathcal{X} = \times_{l=1}^{d}[\mathsf{a}_l, \mathsf{b}_l]$ is bounded and $\mathbb{P}_X$ admits a bounded Lebesgue density f_X on $\mathcal{X}$, then we can set $\mathbb{Q}_{\mathcal{H}} = \mathbb{P}_X$ and $\phi_{\mathcal{H}} = T_{\mathbb{P}_X}$. This case corresponds to [29, Theorem 1.1], and the bounds for c_1 and c_3 coincide with those in [29, Section 3, *Transformation of the r.v.'s*]. Alternatively, if $\mathcal{X}$ is unbounded but $\operatorname{Supp}(\mathcal{H})$ is bounded, we may still be able to find $\mathbb{Q}_{\mathcal{H}}$ supported on a bounded rectangle. We illustrate this case with Example 1 in Section 3.2.

(b) Suppose $\mathcal{Q}_{\mathcal{H}} = \times_{l=1}^{d}[\mathsf{a}_l, \mathsf{b}_l]$ is unbounded. This is often the case when $\mathcal{X}$ and $\operatorname{Supp}(\mathcal{H})$ are unbounded (but note that setting $\mathcal{X} \cap \operatorname{Supp}(\mathcal{H})$ could be bounded in some cases). To fix ideas, let $\mathbf{x}_i \sim \mathsf{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, we can set $\mathbb{Q}_{\mathcal{H}} = \mathbb{P}_X$ and $\phi_{\mathcal{H}} = T_{\mathbb{P}_X}$, and obtain

$$\begin{aligned} \mathsf{c}_1 &\leq d \sup_{\mathbf{x}\in\mathcal{Q}_{\mathcal{H}}} \max\{f_{X,1}(x_1), f_{X,2|1}(x_2|x_1), \cdots, f_{X,d|-d}(x_d|x_{-d})\}^{d-1} \\ &\leq d \min_{1\leq k\leq d}\{\boldsymbol{\Sigma}_{k,k} - \boldsymbol{\Sigma}_{k,1:k-1}\boldsymbol{\Sigma}_{1:k-1,1:k-1}^{-1}\boldsymbol{\Sigma}_{1:k-1,k}\}^{-(d-1)/2} \end{aligned} \tag{11}$$

bounded, but c_2 (and hence c_3) unbounded. This result shows that even when the support of $\mathbb{P}_X$ is unbounded, a valid uniform Gaussian strong approximation can be established in certain cases (albeit the Lipschitz property is not used).

Case 2: Non-Rectangular $\mathcal{Q}_{\mathcal{H}}$. Due to the irregularity of $\mathcal{X}$ and $\operatorname{Supp}(\mathcal{H})$, in some settings only a surrogate measure $\mathbb{Q}_{\mathcal{H}}$ with non-rectangular $\mathcal{Q}_{\mathcal{H}}$ may exist. Then, we can compose the Rosenblatt transformation with another mapping capturing the shape of $\mathcal{Q}_{\mathcal{H}}$ to exhibit a valid normalizing transformation. Suppose that $\mathbb{Q}_{\mathcal{H}}$ admits a Lebesgue density f_Q supported on $\mathcal{Q}_{\mathcal{H}}$, and there exists a diffeomorphism $\chi : \mathcal{Q}_{\mathcal{H}} \mapsto [0,1]^d$. Setting $\phi_{\mathcal{H}} = T_{\mathbb{Q}_{\mathcal{H}}\circ\chi^{-1}} \circ \chi$ gives a valid normalizing transformation, with

$$\mathsf{c}_1 \leq d\frac{\overline{f}_Q^2}{\underline{f}_Q}\mathsf{S}_\chi \qquad \text{and} \qquad \mathsf{c}_2 \leq \frac{\overline{f}_Q}{\underline{f}_Q}\mathsf{S}_\chi,$$

where $\mathsf{S}_\chi = \frac{\sup_{\mathbf{x}\in[0,1]^d}|\det(\nabla\chi^{-1}(\mathbf{x}))|}{\inf_{\mathbf{x}\in[0,1]^d}|\det(\nabla\chi^{-1}(\mathbf{x}))|}\|\|\nabla\chi^{-1}\|_2\|_\infty$. See also Example 1 in Section 3.2.

To recap, Theorem 1 requires the existence of a surrogate measure and a normalizing transformation, which restrict the probability law of the data and take advantage of specific features of the function class. In particular, assumption (ii) in Theorem 1 does not require $\mathcal{X}$ to be compact if either (11) is bounded (as it occurs when $\mathbb{P}_X$ is the Gaussian distribution) or $\mathrm{Supp}(\mathcal{H})$ is bounded (as we illustrate in Example 1 in Section 3.2). See Section SA-II.2 of [11] for details.

3.2. *Special Cases and Related Literature.* We introduce our first statistical example.

EXAMPLE 1 (Kernel Density Estimation). Suppose that $\mathbb{P}_X$ admits a continuous Lebesgue density f_X on its support $\mathcal{X}$. The classical kernel density estimator is

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

where $K : \mathcal{K} \to \mathbb{R}$ is a continuous function with $\mathcal{K} \subseteq \mathbb{R}^d$ compact, and $\int_{\mathcal{K}} K(\mathbf{w})d\mathbf{w} = 1$. In statistical applications, the bandwidth $b \to 0$ as $n \to \infty$ to enable nonparametric estimation [34]. Consider establishing a strong approximation for the localized empirical process $(\xi_n(\mathbf{w}) : \mathbf{w} \in \mathcal{W})$, $\mathcal{W} \subseteq \mathcal{X}$, where

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

with $\mathcal{H} = \{h_{\mathbf{w}}(\cdot) = b^{-d/2}K((\cdot - \mathbf{w})/b) : \mathbf{w} \in \mathcal{W}\}$. It follows that $\mathtt{M}_{\mathcal{H},\mathbb{R}^d} = O(b^{-d/2})$. ▲

Variants of Example 1 have been discussed extensively in prior literature on strong approximations 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, [18] and [19] established strong approximations for the univariate case ($d = 1$) under i.i.d. sampling with $\mathcal{X}$ unbounded, [9] established strong approximations for the univariate case ($d = 1$) under i.i.d. sampling with $\mathcal{X}$ compact, [29] established strong approximations for the multivariate case ($d > 1$) under i.i.d. sampling with $\mathcal{X}$ compact, [31] established strong approximations for the multivariate case ($d > 1$) under i.i.d. sampling with $\mathcal{X}$ unbounded, and [8] established strong approximations for the univariate case ($d = 1$) under non-i.i.d. dyadic data with $\mathcal{X}$ compact. [13, Remark 3.1] provides further discussion and references. See also [12] for an application of [29] to uniform inference for conditional density estimation.

We can use Example 1 to further illustrate the role of $\mathbb{Q}_{\mathcal{H}}$ and $\phi_{\mathcal{H}}$.

EXAMPLE 1 (continued). Recall that $\mathcal{X}$ is the support of $\mathbb{P}_X$, $\mathcal{W} \subseteq \mathcal{X}$ is the index set for the class $\mathcal{H}$, and $\mathcal{K}$ is the compact support of K. It follows that $\mathrm{Supp}(\mathcal{H}) = \mathcal{W} + b \cdot \mathcal{K}$. We illustrate two sets of primitive conditions implying assumption (ii) in Theorem 1.

- Suppose that $\mathcal{X} = \times_{l=1}^{d}[\mathsf{a}_l, \mathsf{b}_l]$, $-\infty \leq \mathsf{a}_l < \mathsf{b}_l \leq \infty$, and $\mathcal{W}$ is arbitrary. Then, we can set $\mathbb{Q}_{\mathcal{H}} = \mathbb{P}_X$ and $\phi_{\mathcal{H}} = T_{\mathbb{P}_X}$, and the discussion in parts (a) and (b) of Case 1 in Section 3.1 applies, which implies assumption (ii) in Theorem 1 under the assumptions imposed therein. Furthermore, when $\mathcal{X}$ is bounded, $\mathsf{c}_1 = O(1)$ and $\mathsf{c}_2 = O(1)$, and hence $\mathsf{c}_3 = O(1)$, because f_X is continuous and positive on $\mathcal{X}$. This is part (a) in Case 1 of Section 3.1, and also the example in [29, Section 4]. No information on $\mathrm{Supp}(\mathcal{H})$ is used.
- Suppose that $\mathcal{X}$ is arbitrary, and $\mathcal{W}$ is bounded. Then, it may be possible to find $\mathbb{Q}_{\mathcal{H}}$ supported on a bounded set, even if $\mathcal{X}$ is unbounded. For example, suppose that $\mathcal{X} = \mathbb{R}^d_+$,

$\mathcal{W} = \times_{l=1}^{d}[\mathsf{a}_l, \mathsf{b}_l]$, $0 \leq \mathsf{a}_l < \mathsf{b}_l < \infty$, and $\mathcal{K} = [-1,1]^d$. Then, for instance, we can take $\mathbb{Q}_{\mathcal{H}}$ with Lebesgue density

$$f_Q(\mathbf{x}) = \begin{cases} f_X(\mathbf{x}) & \text{if } \mathbf{x} \in \times_{l=1}^{d}[\overline{\mathsf{a}}_l, \overline{\mathsf{b}}_l], \\ (1 - \mathbb{P}_X(\times_{l=1}^{d}[\overline{\mathsf{a}}_l, \overline{\mathsf{b}}_l]))/\mathfrak{m}(\Upsilon) & \text{if } \mathbf{x} \in \Upsilon, \\ 0 & \text{otherwise,} \end{cases}$$

where $\overline{\mathsf{a}}_l = \max\{\mathsf{a}_l - b, 0\}$, $\overline{\mathsf{b}}_l = \mathsf{b}_l + b$, $\Upsilon = \times_{l=1}^{d}[\overline{\mathsf{a}}_l, \overline{\mathsf{b}}_l + 1] \setminus \times_{l=1}^{d}[\overline{\mathsf{a}}_l, \overline{\mathsf{b}}_l]$, and $\phi_{\mathcal{H}} = T_{\mathbb{Q}_{\mathcal{H}} \circ \chi^{-1}} \circ \chi$ with $\chi(x_1, \cdots, x_d) = ((\overline{\mathsf{b}}_1 - \overline{\mathsf{a}}_1)^{-1}(x_1 - \overline{\mathsf{a}}_1), \cdots, (\overline{\mathsf{b}}_d - \overline{\mathsf{a}}_d)^{-1}(x_d - \overline{\mathsf{a}}_d))$. It follows that assumption (ii) in Theorem 1 holds. A more general example is discussed in [11, Section SA-II.6].

Finally, the surrogate measure and normalizing transformation could be used to incorporate truncation arguments. We do not dive into this idea for brevity. ▲

We now specialize Theorem 1 to several cases of practical interest. We employ the definitions and notation conventions given in Section 2.1. To streamline the presentation, we also assume that $\mathtt{c}_1 < \infty$ and $\mathtt{c}_2 < \infty$ (hence $\mathtt{c}_3 < \infty$) in the remaining of Section 3. See [11, Section SA-II] for details.

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

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 $\mathtt{M}_{\mathcal{H}}$ over $\mathcal{Q}_{\mathcal{H}}$ with constants $\mathtt{c}_{\mathcal{H}} \geq e$ and $\mathtt{d}_{\mathcal{H}} \geq 1$. Then, (3) holds with

$$\varrho_n = \mathsf{m}_{n,d}\sqrt{\log n}\sqrt{\mathtt{c}_1 \mathtt{M}_{\mathcal{H}} \mathtt{TV}_{\mathcal{H}}} + \frac{\log n}{\sqrt{n}} \min\{\sqrt{\log n}\sqrt{\mathtt{M}_{\mathcal{H}}}, \sqrt{\mathtt{c}_3 \mathtt{K}_{\mathcal{H}} + \mathtt{M}_{\mathcal{H}}}\}\sqrt{\mathtt{M}_{\mathcal{H}}}.$$

This corollary recovers the main result in [29, 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 [18, 19, 8]. Thus, Corollary 1 contributes to the literature by covering all $d \geq 1$ cases simultaneously, allowing for possibly weaker regularity conditions on $\mathbb{P}_X$ through the surrogate measure and normalizing transformation, and making explicit the dependence on d, $\mathcal{X}$, and all other features of the underlying data generating process. This additional contribution can be useful for non-asymptotic probability concentration arguments, or for truncation arguments (see [31] for an example). Nonetheless, for $d \geq 2$, the main intellectual content of Corollary 1 is due to [29]; we present it here for completeness and as a prelude for our upcoming results.

For $d = 1$, Corollary 1 delivers the optimal univariate KMT approximation rate when $\mathtt{K}_{\mathcal{H}} = O(1)$, which employs a weaker notion of total variation relative to prior literature, but at the expense of requiring additional conditions, as the following remark explains.

REMARK 1 (Univariate Strong Approximation). In Section 2 of [18] and the proof of [19], the authors considered univariate ($d = 1$) i.i.d. continuously distributed random variables, and established the strong approximation:

$$\mathbb{P}\left(\|X_n - Z_n^X\|_{\mathcal{H}} > \mathtt{pTV}_{\mathcal{H},\mathbb{R}} \frac{t + C_1 \log n}{\sqrt{n}}\right) \leq C_2 \exp(-C_3 t), \qquad t > 0,$$

where C_1, C_2, C_3 are universal constants. [8, 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.

For any interval $\mathcal{I}$ in $\mathbb{R}$, $\mathtt{TV}_{\mathcal{H},\mathcal{I}} \leq \mathtt{pTV}_{\mathcal{H},\mathcal{I}}$ provided that $\mathtt{M}_{\mathcal{H},\mathcal{I}} < \infty$ [1, Theorem 3.27]. Therefore, Theorem 1 employs a weaker notation of total variation, but imposes complexity requirements on $\mathcal{H}$ and the existence of a normalizing transformation. In contrast, [18], [19] and [8] do not imposed those extra conditions, but their results only apply 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 $\mathtt{TV}_{\mathcal{H}} = O(b^{d/2-1})$ and $\mathtt{K}_{\mathcal{H}} = O(b^{-d/2})$. 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 $d = 1$ [18, 19, 8] and $d \geq 2$ [29]. 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 [13, 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 $d \in \{1, 2, 3\}$, 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 $\mathcal{H} = \{b^{-d/2}K((\cdot - \mathbf{w})/b) : \mathbf{w} \in \mathcal{W}\}$ forms a VC-type class, and the other conditions in Corollary 1. The Lipschitz assumption holds 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, with a slightly suboptimal strong approximation rate (an extra $\sqrt{\log n}$ term appears when $d \geq 2$).

3.2.2. *VC-type Lipschitz Functions.* It is known that the uniform Gaussian strong approximation rate in Corollary 1 is optimal under the assumptions imposed [2]. 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 ($\mathtt{L}_{\mathcal{H}} < \infty$).

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 envelope function $\mathtt{M}_{\mathcal{H}}$ over $\mathcal{Q}_{\mathcal{H}}$ with constants $\mathtt{c}_{\mathcal{H}} \geq e$ and $\mathtt{d}_{\mathcal{H}} \geq 1$. Then, (3) holds with

$$\begin{aligned}\varrho_n &= \min\{\mathsf{m}_{n,d}\sqrt{\mathtt{M}_{\mathcal{H}}}, \mathsf{l}_{n,d}\sqrt{\mathtt{c}_2\mathtt{L}_{\mathcal{H}}}\}\sqrt{\log n}\sqrt{\mathtt{c}_1\mathtt{TV}_{\mathcal{H}}} \\ &\quad + \frac{\log n}{\sqrt{n}}\min\{\sqrt{\log n}\sqrt{\mathtt{M}_{\mathcal{H}}}, \sqrt{\mathtt{c}_3\mathtt{K}_{\mathcal{H}} + \mathtt{M}_{\mathcal{H}}}\}\sqrt{\mathtt{M}_{\mathcal{H}}}.\end{aligned}$$

Putting aside $\mathtt{M}_{\mathcal{H}}$ and $\mathtt{TV}_{\mathcal{H}}$, this corollary shows that if $\mathtt{L}_{\mathcal{H}} < \infty$, then the rate of strong approximation can be 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 $\mathtt{K}_{\mathcal{H}} = O(b^{-d/2})$. Therefore, Corollary 2 establishes 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: the main roadblock underlying the proof strategy is related to the coupling approach based on the 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 $\mathcal{Q}_{\mathcal{H}}$. 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 [29, Theorem 2.1] can be improved. However, further improvements for smoother functions appear to necessitate an approximation approach that would not generate dyadic binomial counts, thereby rendering current coupling approaches inapplicable.

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

EXAMPLE 1 (continued). Under the conditions imposed, $\mathtt{L}_{\mathcal{H}} = O(b^{-d/2-1})$, 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 [13, Remark 3.1(ii)], Example 1 shows 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 $\|\xi_n\|_{\mathcal{H}} = \sup_{\mathbf{w}\in\mathcal{W}} |\xi_n(\mathbf{w})|$ 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 [29, Section 4] already assumed K is Lipschitizian for the verification of the conditions imposed by his strong approximation result (cf. Corollary 1).

3.2.3. *Polynomial-entropy Functions.* [22] 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 [29]. Furthermore, [22] employed a Haar approximation condition, which plays a similar role as the total variation and the Lipschitz conditions exploited in our paper. To enable a precise comparison to [22], the next corollary considers a class $\mathcal{H}$ satisfying a polynomial-entropy condition.

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

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

$$\varrho_n = \mathsf{m}_{n,d}\sqrt{\mathtt{c}_1\mathtt{M}_{\mathcal{H}}\mathtt{TV}_{\mathcal{H}}}\Big(\sqrt{\log n} + (\mathtt{c}_1\mathsf{m}_{n,d}^2\mathtt{M}_{\mathcal{H}}^{-1}\mathtt{TV}_{\mathcal{H}})^{-\frac{\mathtt{b}_{\mathcal{H}}}{4}}\Big)$$
$$+ \sqrt{\frac{\mathtt{M}_{\mathcal{H}}}{n}}\min\{\sqrt{\log n}\sqrt{\mathtt{M}_{\mathcal{H}}}, \sqrt{\mathtt{c}_3\mathtt{K}_{\mathcal{H}} + \mathtt{M}_{\mathcal{H}}}\}\Big(\log n + (\mathtt{c}_1\mathsf{m}_{n,d}^2\mathtt{M}_{\mathcal{H}}^{-1}\mathtt{TV}_{\mathcal{H}})^{-\frac{\mathtt{b}_{\mathcal{H}}}{2}}\Big),$$

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

$$\varrho_n = \mathsf{l}_{n,d}\sqrt{\mathtt{c}_1\mathtt{c}_2\mathtt{L}_{\mathcal{H}}\mathtt{TV}_{\mathcal{H}}}\Big(\sqrt{\log n} + (\mathtt{c}_1\mathtt{c}_2\mathsf{l}_{n,d}^2\mathtt{M}_{\mathcal{H}}^{-2}\mathtt{L}_{\mathcal{H}}\mathtt{TV}_{\mathcal{H}})^{-\frac{\mathtt{b}_{\mathcal{H}}}{4}}\Big)$$
$$+ \sqrt{\frac{\mathtt{M}_{\mathcal{H}}}{n}}\min\{\sqrt{\log n}\sqrt{\mathtt{M}_{\mathcal{H}}}, \sqrt{\mathtt{c}_3\mathtt{K}_{\mathcal{H}} + \mathtt{M}_{\mathcal{H}}}\}\Big(\log n + (\mathtt{c}_1\mathtt{c}_2\mathsf{l}_{n,d}^2\mathtt{M}_{\mathcal{H}}^{-2}\mathtt{L}_{\mathcal{H}}\mathtt{TV}_{\mathcal{H}})^{-\frac{\mathtt{b}_{\mathcal{H}}}{2}}\Big).$$

This corollary reports a simplified version of our result, which corresponds to the best possible bound for the discussion in this section. See [11, Section SA-II] for the general case. It is possible to apply Corollary 3 to Example 1, although the result is suboptimal relative to the previous results leveraging a VC-type condition.

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

Our running example shows that a uniform Gaussian strong approximation based on polynomial-entropy conditions can lead to suboptimal KMT approximation rates. However, for other (larger) function classes, those results may be useful. The following remark discusses an example studied in [22], and illustrates our contributions in that context.

REMARK 2 (Polynomial-entropy Condition). Suppose $\mathbb{P}_X$ is $\mathsf{Uniform}(\mathcal{X})$ with $\mathcal{X} = [0,1]^d$, and $\mathcal{H}$ a subclass of $C^q(\mathcal{X})$ with C^q-norm uniformly bounded by 1 and $2 \leq d < q$. [22, page 111] discusses this example after his Theorem 11.3, and reports the uniform Gaussian strong approximation rate $n^{-\frac{q-d}{2qd}} \operatorname{polylog}(n)$. See [22], or [11, Section SA-I], for the additional notation and definitions used in this example.

Corollary 3 is applicable to this case, upon setting $(\mathbb{Q}_{\mathcal{H}}, \phi_{\mathcal{H}}) = (\mathbb{P}_X, \mathrm{Id})$ with Id denoting the identity map from $[0,1]^d$ to $[0,1]^d$. It follows that $\mathtt{M}_{\mathcal{H}} = 1$, $\mathtt{TV}_{\mathcal{H}} = 1$, $\mathtt{L}_{\mathcal{H}} = 1$. [33, Theorem 2.7.1] shows that $\mathcal{H}$ is a polynomial-entropy class with constants $\mathtt{a}_{\mathcal{H}} = \mathtt{C}_{q,d}$ and $\mathtt{b}_{\mathcal{H}} = d/q$, where $\mathtt{C}_{q,d}$ is a constant depending on q and d only. 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 [22].

The improvement is explained by two differences between [22] 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 [22]. Second, using the uniform entropy condition approach, we get $\log N(\mathcal{H}, \|\cdot\|_{\mathbb{P}_X,2}, \varepsilon) = O(\varepsilon^{-d/q})$, while [22] started with the bracketing number condition $\log N_{[\,]}(\mathcal{H}, \|\cdot\|_{\mathbb{P}_X,1}, \varepsilon) = O(\varepsilon^{-d/q})$ and, with the help of his Lemma 8.4, applied Theorem 3.1 with $\alpha = \frac{d}{d+q}$ in his Equation (3.2). The proof of his Theorem 3.1 leverages the fact that his Equation (3.2) implies that $\log N(\mathcal{H}, \|\cdot\|_{\mathbb{P}_X,2}, \varepsilon) = O(\varepsilon^{-2d/q})$, and his approximation rate is looser by a power of two when compared to the uniform entropy condition underlying our Corollary 3. Setting $\mathtt{L}_{\mathcal{H}} = \infty$, $\mathtt{b}_{\mathcal{H}} = 2d/q$, and keeping the other constants, Corollary 3*(i)* would give $\varrho_n = n^{-\frac{q-d}{2qd}} \operatorname{polylog}(n)$, which is the same rate as in [22]. Finally, Theorem 3.2 in [22] allows for $\log N(\mathcal{H}, \|\cdot\|_{\mathbb{P}_X,2}, \varepsilon) = O(\varepsilon^{-2\rho})$ where ρ is not implied by his Equation (3.2), and his result would give the strong approximation rate $n^{-\frac{2q-d}{4qd}} \operatorname{polylog}(n)$. □

4. Residual-Based Empirical Process.

Consider the simple local empirical process discussed in [13, Section 3.1]:

$$S_n(\mathbf{w}) = \frac{1}{nb^d} \sum_{i=1}^{n} K\Big(\frac{\mathbf{x}_i - \mathbf{w}}{b}\Big) y_i, \qquad \mathbf{w} \in \mathcal{W}, \tag{12}$$

where $\mathbf{x}_i \sim \mathbb{P}_X$, $y_i \sim \mathbb{P}_Y$, and $b \to 0$ as $n \to \infty$. Using our notation, $\big(\sqrt{nb^d}(S_n(\mathbf{w}) - \mathbb{E}[S_n(\mathbf{w})|\mathbf{x}_1, \cdots, \mathbf{x}_n]) : \mathbf{w} \in \mathcal{W}\big) = (R_n(g,r) : g \in \mathcal{G}, r \in \mathcal{R})$ with $\mathcal{G} = \{b^{-d/2}K(\frac{\cdot - \mathbf{w}}{b}) : \mathbf{w} \in \mathcal{W}\}$ 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 [13, Remark 3.1(iii)], a direct application of [29], or of our Theorem 1, views $\mathbf{z}_i = (\mathbf{x}_i, y_i) \sim \mathbb{P}_Z$ as the underlying $(d+1)$-dimensional random vectors 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 $\mathbb{P}_Z$, [29]'s strong approximation result verifies (3) with rate (6), which is also verified via Corollary 1. Furthermore, imposing a Lipschitz property on $\mathcal{H} = \mathcal{G} \times \mathcal{R}$, Corollary 2 would give the improved strong approximation result (8), under regularity conditions.

The strong approximation results for S_n 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 suboptimal because only the d-dimensional covariate $\mathbf{x}_i$ are being smoothed out. Since the pointwise variance of the process is of order $n^{-1}b^{-d}$, the correct effective sample size should be nb^d, up to $\mathrm{polylog}(n)$ terms. Therefore, applying [29], or our improved Theorem 1, leads to a suboptimal uniform Gaussian strong approximation for S_n. Second, applying [29], or our improved Theorem 1, requires $\mathbb{P}_Z$ to be continuously distributed and supported on $[0,1]^{d+1}$, possibly after applying a normalizing transformation. This requirement imposes non-trivial restrictions on $\mathbb{P}_Z$ and, in particular, on $\mathbb{P}_Y$, limiting the applicability of the strong approximation results. See [13, Remark 3.1(iii)] for more discussion.

Motivated by the aforementioned limitations, the following theorem explicitly studies the residual-based empirical process 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 [11, Section SA-IV]. Recall Section 2.1 and the notation conventions introduced therein.

THEOREM 2. Suppose $(\mathbf{z}_i = (\mathbf{x}_i, y_i) : 1 \leq i \leq n)$ are i.i.d. random vectors taking values in $(\mathbb{R}^{d+1}, \mathcal{B}(\mathbb{R}^{d+1}))$ with common law $\mathbb{P}_Z$, where $\mathbf{x}_i$ has distribution $\mathbb{P}_X$ supported on $\mathcal{X} \subseteq \mathbb{R}^d$, y_i has distribution $\mathbb{P}_Y$ supported on $\mathcal{Y} \subseteq \mathbb{R}$, and the following conditions hold.

(i) $\mathcal{G}$ is a real-valued pointwise measurable class of functions on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P}_X)$.
(ii) There exists a surrogate measure $\mathbb{Q}_{\mathcal{G}}$ for $\mathbb{P}_X$ with respect to $\mathcal{G}$ such that $\mathbb{Q}_{\mathcal{G}} = \mathfrak{m} \circ \phi_{\mathcal{G}}$, where the *normalizing transformation* $\phi_{\mathcal{G}} : \mathcal{Q}_{\mathcal{G}} \mapsto [0,1]^d$ is a diffeomorphism.
(iii) $\mathcal{G}$ is a VC-type class with function $\mathtt{M}_{\mathcal{G}}$ over $\mathcal{Q}_{\mathcal{G}}$ with $\mathtt{c}_{\mathcal{G}} \geq e$ and $\mathtt{d}_{\mathcal{G}} \geq 1$.
(iv) $\mathcal{R}$ is a real-valued pointwise measurable class of functions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_Y)$.
(v) $\mathcal{R}$ is a VC-type class with envelope $M_{\mathcal{R},\mathcal{Y}}$ over $\mathcal{Y}$ with $\mathtt{c}_{\mathcal{R},\mathcal{Y}} \geq e$ and $\mathtt{d}_{\mathcal{R},\mathcal{Y}} \geq 1$, where $M_{\mathcal{R},\mathcal{Y}}(y) + \mathtt{pTV}_{\mathcal{R},(-|y|,|y|)} \leq \mathtt{v}(1+|y|^\alpha)$ for all $y \in \mathcal{Y}$, for some $\mathtt{v} > 0$, and for some $\alpha \geq 0$. Furthermore, if $\alpha > 0$, then $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(|y_i|)|\mathbf{x}_i = \mathbf{x}] \leq 2$.
(vi) There exists a constant $\mathtt{k}$ such that $|\log_2 \mathtt{E}_{\mathcal{G}}| + |\log_2 \mathtt{TV}| + |\log_2 \mathtt{M}_{\mathcal{G}}| \leq \mathtt{k} \log_2 n$, where $\mathtt{TV} = \max\{\mathtt{TV}_{\mathcal{G}}, \mathtt{TV}_{\mathcal{G} \times \mathcal{V}_{\mathcal{R}}, \mathcal{Q}_{\mathcal{G}}}\}$ with $\mathcal{V}_{\mathcal{R}} = \{\theta(\cdot, r) : r \in \mathcal{R}\}$, and $\theta(\mathbf{x}, r) = \mathbb{E}[r(y_i)|\mathbf{x}_i = \mathbf{x}]$.

Then, on a possibly enlarged probability space, there exists a sequence of mean-zero Gaussian processes $(Z_n^R(g,r) : (g,r) \in \mathcal{G} \times \mathcal{R})$ with almost sure continuous trajectories on $(\mathcal{G} \times \mathcal{R}, \mathfrak{d}_{\mathbb{P}_Z})$ 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
- $\mathbb{P}\big[\|R_n - Z_n^R\|_{\mathcal{G} \times \mathcal{R}} > C_1 C_{\mathtt{v},\alpha} \mathsf{T}_n(t)\big] \leq C_2 e^{-t}$ for all $t > 0$,

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

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

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

$$\mathtt{c}_1 = d\sup_{\mathbf{x}\in\mathcal{Q}_{\mathcal{G}}}\prod_{j=1}^{d-1}\sigma_j(\nabla\phi_{\mathcal{G}}(\mathbf{x})), \qquad \mathtt{c}_2 = \sup_{\mathbf{x}\in\mathcal{Q}_{\mathcal{G}}}\frac{1}{\sigma_d(\nabla\phi_{\mathcal{G}}(\mathbf{x}))},$$

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

This theorem establishes a uniform Gaussian strong approximation under regularity conditions specifically tailored to leverage the multiplicative separable structure of R_n defined in (7). Conditions (i)–(iii) in Theorem 2 are analogous to the conditions imposed in Corollaries 1 and 2 for the general empirical process. Conditions (iv)–(v) in Theorem 2 are new, mild restrictions on the portion of the stochastic process corresponding to the outcome y_i. Condition (v) either assumes $\mathcal{R}$ is uniformly bounded, or restricts the tail decay of the function class $\mathcal{R}$, without imposing restrictive assumptions on the distribution $\mathbb{P}_Y$. Finally, condition (vi) is imposed only to simplify the exposition; see [11] for the general result. We require a $\mathtt{pTV}$ condition on $\mathcal{R}$ in (v), but $\mathtt{TV}$ conditions on $\mathcal{G}$ and $\mathcal{G}\times\mathcal{V}_{\mathcal{R}}$ in (vi), because $\mathbb{P}_X$ admits a Lebesgue density, but $\mathbb{P}_Y$ may not.

The proof strategy of Theorem 2 is similar to the proof for the general empirical process (Theorem 1), and is given in [11, Section SA-IV]. First, we discretize to a δ-net to obtain

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

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 onto 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 the partitioning cells based on two key observations: (i) regularity conditions are often imposed on the conditional distribution of $y_i|\mathbf{x}_i$, as opposed to on 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 or atoms, while $\mathcal{G}$ is a VC-type class of n-varying functions with a possibly more regular $\mathbb{P}_X$ having compact support. Furthermore, 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 [11] for details on our proposed asymmetric dyadic cells expansion.

Given our dyadic expansion exploiting the structure of 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 $L_2(\mathbb{P}_Z)$ to piecewise constant functions that explicitly factorizes the product $g(\mathbf{x}_i)r(y_i)$. In fact, as we discuss in [11], 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 generalized Tusnády's lemma for more general binomial counts [11] allows for the Gaussian coupling of any piecewise-constant functions over our asymmetrically constructed dyadic cells. A generalization of [29, 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, the 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 and other local empirical process examples.

The following corollary summarizes the main result from Theorem 2.

COROLLARY 4 (VC-Type Lipschitz Functions). Suppose the conditions of Theorem 2 hold with constants $\mathtt{c}$ and $\mathtt{d}$. Then, $\|R_n - Z_n^R\|_{\mathcal{G}\times\mathcal{R}} = O(\varrho_n)$ a.s. with

$$\varrho_n = \min\left\{\frac{(\mathtt{c}_1^d \mathtt{M}_{\mathcal{G}}^{d+1}\mathtt{TV}^d\mathtt{E}_{\mathcal{G}})^{\frac{1}{2d+2}}}{n^{1/(2d+2)}}, \frac{(\mathtt{c}_1^{\frac{d}{2}}\mathtt{c}_2^{\frac{d}{2}}\mathtt{M}_{\mathcal{G}}\mathtt{TV}^{\frac{d}{2}}\mathtt{E}_{\mathcal{G}}\mathtt{L}^{\frac{d}{2}})^{\frac{1}{d+2}}}{n^{1/(d+2)}}\right\}(\log n)^{\alpha+3/2} + \frac{(\log n)^{\alpha+1}}{\sqrt{n}}\mathtt{M}_{\mathcal{G}}.$$

This corollary shows that our best attainable uniform Gaussian strong approximation rate for R_n is $n^{-1/(d+2)}\operatorname{polylog}(n)$, putting aside $\mathtt{c}_1$, $\mathtt{c}_2$, $\mathtt{M}_{\mathcal{G}}$, $\mathtt{TV}$, $\mathtt{E}_{\mathcal{G}}$, and $\mathtt{L}$. It is not possible to give a strict ranking between Corollary 2 and Corollary 4. 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 $n^{-\min\{1/(d+1),1/2\}}\operatorname{polylog}(n)$, putting aside the various constants and underlying assumptions. On the other hand, Corollary 4 can deliver a tighter strong approximation under weaker regularity conditions whenever $\mathcal{H} = \mathcal{G} \times \mathcal{R}$ and $\mathcal{G}$ varies with n, as in the case of the local empirical processes arising from nonparametric regression. The next section offers an application illustrating this point.

See [11, Section SA-IV] for proofs and other omitted details. In addition, Section SA-III in [11] present uniform Gaussian strong approximation results for a general multiplicative-separable empirical process, which may be of interest but is not discussed in the paper to conserve space.

4.1. *Example: Local Polynomial Regression.* Suppose that $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are i.i.d random vectors taking values in $(\mathbb{R}^{d+1}, \mathcal{B}(\mathbb{R}^{d+1}))$, with $\mathbf{x}_i \sim \mathbb{P}_X$ admitting a continuous Lebesgue density on its support $\mathcal{X} = [0,1]^d$. Consider the class of estimands

$$\theta(\mathbf{w}; r) = \mathbb{E}[r(y_i)|\mathbf{x}_i = \mathbf{w}], \qquad \mathbf{w} \in \mathcal{W} \subseteq \mathcal{X}, \quad r \in \mathcal{R}, \tag{13}$$

where we focus on two leading cases to streamline the discussion: $\mathcal{R}_1 = \{\mathrm{Id}\}$ corresponds to the conditional expectation $\mu(\mathbf{w}) = \mathbb{E}[y_i|\mathbf{x}_i = \mathbf{w}]$, and $\mathcal{R}_2 = \{\mathbb{1}(\cdot \leq y) : y \in \mathbb{R}\}$ corresponds to the conditional distribution function $F(y|\mathbf{w}) = \mathbb{E}[\mathbb{1}(y_i \leq y)|\mathbf{x}_i = \mathbf{w}]$. In the first case, $\mathcal{R}$ is a singleton but the identity 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 has compact support $\mathcal{K}$. 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 $0 \leq |\boldsymbol{\nu}| \leq \mathfrak{p}$, where $\mathbf{u}^{\boldsymbol{\nu}} = u_1^{\nu_1} \cdots u_d^{\nu_d}$, $\boldsymbol{\nu}! = \nu_1! \cdots \nu_d!$ and $|\boldsymbol{\nu}| = \nu_1 + \cdots + \nu_d$, for $\mathbf{u} = (u_1, \cdots, u_d)^\top$ and $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_d)^\top$. A local polynomial regression estimator of $\theta(\mathbf{w}; r)$ is

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

with $\mathbf{w} \in \mathcal{W} \subseteq \mathcal{X}$, $r \in \mathcal{R}_1$ or $r \in \mathcal{R}_2$, and $\mathbf{e}_1$ denoting the first standard basis vector. See [17] for a textbook review. The estimation error can be decomposed into three terms:

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

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

It follows that the linear term is

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

for $(g, r) \in \mathcal{G} \times \mathcal{R}_l$, $l = 1, 2$, and where $\mathcal{G} = \{b^{-d/2} \mathfrak{K}_{\mathbf{w}}(\frac{\cdot - \mathbf{w}}{b}) : \mathbf{w} \in \mathcal{W}\}$ with $\mathfrak{K}_{\mathbf{w}}(\mathbf{u}) = \mathbf{e}_1^\top \mathbf{H}_{\mathbf{w}}^{-1} \mathbf{p}(\mathbf{u}) K(\mathbf{u})$ the equivalent boundary-adaptive kernel function. Furthermore, under the regularity conditions given in [11, Section SA-IV.6], which relate to uniform smoothness and moment restrictions for the conditional distribution of $y_i | \mathbf{x}_i$,

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

$$\sup_{\mathbf{w} \in \mathcal{W}, r \in \mathcal{R}_2} \big| \mathbf{e}_1^\top (\widehat{\mathbf{H}}_{\mathbf{w}}^{-1} - \mathbf{H}_{\mathbf{w}}^{-1}) \mathbf{S}_{\mathbf{w},r} \big| = O((nb^d)^{-1} \log n) \quad \text{a.s.},$$

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

provided that $\log(n)/(nb^d) \to 0$. Therefore, the goal reduces to establishing a Gaussian strong approximation for the residual-based empirical process $(R_n(g, r) : (g, r) \in \mathcal{G} \times \mathcal{R}_l)$, $l = 1, 2$. We discuss different attempts to establish such approximation result, culminating with the application of our Theorem 2.

As discussed in [13, Remark 3.1], a first attempt is to deploy Theorem 1.1 in [29] (or, equivalently, Corollary 1). Viewing the empirical process as based on the random sample $\mathbf{z}_i = (\mathbf{x}_i, y_i) \sim \mathbb{P}_Z$, $i = 1, 2, \cdots, n$, Theorem 1.1 in [29] requires $\mathbb{P}_Z$ to be continuously distributed with positive Lebesgue density on its support $[0, 1]^{d+1}$. For this reason, [13, Remark 3.1] assumes that $(\mathbf{x}_i, y_i) = (\mathbf{x}_i, \varphi(\mathbf{x}_i, u_i))$ where the joint law $\mathbb{P}_B$ of $\mathbf{b}_i = (\mathbf{x}_i, u_i)$ admits a continuous Lebesgue density supported on $\mathcal{B} = [0, 1]^{d+1}$. If $\mathtt{M}_{\{\varphi\},\mathcal{B}} < \infty$, $\mathtt{K}_{\{\varphi\},\mathcal{B}} < \infty$, $\sup_{g \in \mathcal{G}} \mathtt{TV}_{\{\varphi\},\operatorname{supp}(g) \times [0,1]} < \infty$, and other regularity conditions hold, then it can be shown [11, Section SA-IV.6] that applying [29] to $(X_n(h) : h \in \mathcal{H}_l)$ based on $(\mathbf{b}_i : 1 \leq i \leq n)$ with $\mathcal{H}_l = \{g \cdot (r \circ \varphi) - g \cdot \theta(\cdot, r) : g \in \mathcal{G}, r \in \mathcal{R}_l\}$, $l = 1, 2$, gives a Gaussian strong approximation with rate (6). Without the local total variation condition $\mathtt{K}_{\{\varphi\},\mathcal{B}} < \infty$, an additional $\sqrt{\log n}$ multiplicative factor appears in the final rate.

The previous result does not exploit Lipschitz continuity, so a natural second attempt is to employ Corollary 2 to improve it. Retaining the same assumptions, but now also assuming that φ is Lipschitz, our Theorem 1 gives a Gaussian strong approximation for $(X_n(h) : h \in \mathcal{H}_1)$ with rate (8). Theorem 1 does not give an improvement for $\mathcal{R}_2$ because the Lipschitz condition is not satisfied. See [11, Section SA-IV.6].

The two attempts so far impose restrictive 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 2 addresses both problems: since $\operatorname{Supp}(\mathcal{H}) = \mathcal{W} + b\mathcal{K}$, and under standard regularity conditions, we can set $\mathbb{Q}_{\mathcal{H}}$ and $\phi_{\mathcal{H}}$ according to the discussion in Example 1, and thus we verify in [11, Section SA-IV.6] that $\mathtt{c}_1 = O(1)$, $\mathtt{c}_2 = O(1)$, $\mathtt{M}_{\mathcal{G}} = O(b^{-d/2})$, $\mathtt{E}_{\mathcal{G}} = O(b^{d/2})$, $\mathtt{K}_{\mathcal{G}} = O(b^{-d/2})$, $\mathtt{TV} = O(b^{d/2-1})$, and $\mathtt{L} = O(b^{-d/2-1})$. This 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 $\sup_{\mathbf{w}\in\mathcal{W}} \mathbb{E}[\exp(|y_i|)|\mathbf{x}_i = \mathbf{w}] < \infty$, 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 the following strong approximations hold:

- Let $\widehat{\mu}(\mathbf{w}) = \widehat{\theta}(\mathbf{w}; r)$ for $r \in \mathcal{R}_1$. Recall that $\mathcal{R}_1$ consists of the singleton of identity function Id. If $b^{\mathfrak{p}+1}(nb^d)^{\frac{d+4}{2d+4}}(\log n)^{-1/2} + (nb^d)^{-\frac{d+1}{d+2}}(\log n)^2 = O(1)$, then

  $$\sup_{\mathbf{w}\in\mathcal{W}} \left|\sqrt{nb^d}\big(\widehat{\mu}(\mathbf{w}) - \mu(\mathbf{w})\big) - Z_n^R(\mathbf{w})\right| = O(\mathsf{r}_n) \quad \text{a.s.}, \qquad \mathsf{r}_n = \left(\frac{(\log n)^{1+d/2}}{nb^d}\right)^{\frac{1}{d+2}},$$

  where $\mathbb{C}\mathrm{ov}(Z_n^R(\mathbf{w}_1), Z_n^R(\mathbf{w}_2)) = nb^d\mathbb{C}\mathrm{ov}(\mathbf{e}_1^\top\mathbf{H}_{\mathbf{w}_1}^{-1}\mathbf{S}_{\mathbf{w}_1,\mathrm{Id}}, \mathbf{e}_1^\top\mathbf{H}_{\mathbf{w}_2}^{-1}\mathbf{S}_{\mathbf{w}_2,\mathrm{Id}})$ for all $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$.
- Let $\widehat{F}(y|\mathbf{w}) = \widehat{\theta}(\mathbf{w}; r_y)$ for $r_y = \mathbb{1}(\cdot \leq y) \in \mathcal{R}_2$. If $b^{\mathfrak{p}+1}(nb^d)^{(d+4)/(2d+4)}(\log n)^{-1/2} = O(1)$ and $(nb^d)^{-1}\log n = o(1)$, then

  $$\sup_{\mathbf{w}\in\mathcal{W}, y\in\mathbb{R}} \left|\sqrt{nb^d}\big(\widehat{F}(y|\mathbf{w}) - F(y|\mathbf{w})\big) - Z_n^R(\mathbf{w}, y)\right| = O(\mathsf{r}_n) \quad \text{a.s.},$$

  where $\mathbb{C}\mathrm{ov}(Z_n^R(\mathbf{w}_1, u_1), Z_n^R(\mathbf{w}_2, u_2)) = nb^d\mathbb{C}\mathrm{ov}(\mathbf{e}_1^\top\mathbf{H}_{\mathbf{w}_1}^{-1}\mathbf{S}_{\mathbf{w}_1,r_{u_1}}, \mathbf{e}_1^\top\mathbf{H}_{\mathbf{w}_2}^{-1}\mathbf{S}_{\mathbf{w}_2,r_{u_2}})$ for all $(\mathbf{w}_1, u_1), (\mathbf{w}_2, u_2)$ in $\mathcal{W} \times \mathbb{R}$ and $r_{u_1}, r_{u_2} \in \mathcal{R}_2$.

This example gives a statistical application where Theorem 2 offers a strict improvement on the accuracy of the Gaussian strong approximation over [29], and the improved Theorem 1 upon incorporating a Lipschitz condition on the function class. See [11, Section SA-IV.6] for omitted details. It remains an open question whether the result in this section provides the best Gaussian strong approximation for local polynomial regression or, more generally, for a local empirical process. The results presented are the best in the literature, but we are unaware of lower bounds that would confirm the approximation rates are unimprovable.

5. Quasi-Uniform Haar Functions. Assuming the existence of a surrogate measure and a normalizing transformation, or otherwise restricting the data generating process, 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 $d \geq 3$, putting aside $\mathtt{c}_1$, $\mathtt{c}_2$, $\mathtt{c}_3$, $\mathtt{M}_{\mathcal{H}}$, $\mathtt{L}_{\mathcal{H}}$, $\mathtt{TV}_{\mathcal{H}}$, and $\mathtt{K}_{\mathcal{H}}$. The possibly suboptimal strong approximation rate 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 a cover of $\mathcal{X}$. Likewise, Theorem 2 established an improved uniform Gaussian strong approximation for the residual-based empirical process (7), but the result is also limited by the mean square projection error incurred by employing a Haar basis expansion based on a carefully chosen, asymmetric partitioning of the support of $\mathbf{z}_i = (\mathbf{x}_i, y_i)$.

Motivated by the limitations introduced by the mean square projection error underlying the proofs of Theorems 1 and 2, this section presents uniform Gaussian strong approximations for $(X_n(h) : h \in \mathcal{H})$ and $(R_n(g,r) : (g,r) \in \mathcal{G} \times \mathcal{R})$ when $\mathcal{H}$ and $\mathcal{G}$ belong to the span of a Haar basis based on a *quasi-uniform* partition with cardinality L, which can be viewed as an approximation based on $L \to \infty$ as $n \to \infty$. We do not require the existence of a normalizing transformation, allow for more general partitioning schemes than dyadic cells expansions, and impose minimal restrictions on the data generating process, while achieving the univariate KMT optimal strong approximation rate based on the effective sample size n/L for all $d \geq 1$. The strong approximation results presented in this section generalize two ideas from the regression Splines literature [21]: (i) the cells forming the Haar basis are assumed to be quasi-uniform with respect to a surrogate measure $\mathcal{Q}_{\mathcal{H}}$; and (ii) the number of active cells of the Haar basis affects the strong approximation. We apply the strong approximation results to histogram density estimation, and partitioning-based regression estimation based on Haar basis, which includes certain regression trees [4] and other related methods [7]. Proof and omitted technical details are given in [11, Section SA-V].

5.1. *General Empirical Process.* The following result is the analogue of Theorem 1.

THEOREM 3. Suppose $(\mathbf{x}_i : 1 \leq i \leq n)$ are i.i.d. random vectors taking values in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with common law $\mathbb{P}_X$ supported on $\mathcal{X} \subseteq \mathbb{R}^d$, and the following condition holds.

(i) $\mathcal{H} \subseteq \operatorname{Span}\{\mathbb{1}_{\Delta_l} : 0 \leq l < L\}$ is a class of Haar functions on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P}_X)$.
(ii) There exists a surrogate measure $\mathbb{Q}_{\mathcal{H}}$ for $\mathbb{P}_X$ with respect to $\mathcal{H}$ such that $\{\Delta_l : 0 \leq l < L\}$ forms a *quasi-uniform partition* of $\mathcal{Q}_{\mathcal{H}}$ with respect to $\mathbb{Q}_{\mathcal{H}}$:

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

(iii) $\mathtt{M}_{\mathcal{H}} < \infty$.

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 on $(\mathcal{H}, \mathfrak{d}_{\mathbb{P}_X})$ 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

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

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

This theorem shows that if $n^{-1}L\log(nL)\to 0$, then a valid strong approximation can be achieved with exponential probability concentration. The proof of Theorem 3 leverages the fact that the L_2-projection error is zero by construction, but recognizes that [29, Theorem 2.1] does not apply because the partitions are *quasi-dyadic*, preventing the use of the celebrated Tusnády's inequality. Instead, in [11], we present two technical results to circumvent that limitation: (i) we combine [6, Lemma 2] and [30, Lemma 2] to establish a version of Tusnády's inequality that allows for more general binomial random variables $\mathsf{Bin}(n,p)$ with $\underline{p}\le p\le\overline{p}$, the error bound holding uniformly in p, as required by the quasi-dyadic partitioning structure; and (ii) we generalize [29, Theorem 2.1] to the case of quasi-dyadic cells.

Assuming a VC-type condition on $\mathcal{H}$, and putting aside $\mathtt{M}_{\mathcal{H}}$, $\mathtt{E}_{\mathcal{H}}$, and $\mathtt{S}_{\mathcal{H}}$, it follows that (3) holds with $\varrho_n=\sqrt{\log(n)}/\sqrt{n/L}+\log(n)/\sqrt{n}$. More generally, we have the following.

COROLLARY 5 (VC-type Haar Functions). Suppose the conditions of Theorem 3 hold. In addition, assume that $\mathcal{H}$ is a VC-type class with function $\mathtt{M}_{\mathcal{H}}$ over $\mathcal{Q}_{\mathcal{H}}$ with constants $\mathtt{c}_{\mathcal{H}}\ge e$ and $\mathtt{d}_{\mathcal{H}}\ge 1$. Then, if $n^{-1}L\log(nL)\to 0$, (3) holds with

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

We offer a simple statistical application of Theorem 3 in the next example.

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

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

where $\{\Delta_l:0\le l<P\}$ are disjoint and satisfy $\max_{0\le l<P}\mathbb{P}_X(\Delta_l)\le\rho\min_{0\le l<P}\mathbb{P}_X(\Delta_l)$.

For L proportional to $\mathbb{P}_X(\Delta_l)^{-1}$, up to ρ, we establish a strong approximation for the localized empirical process $(\zeta_n(\mathbf{w}):\mathbf{w}\in\mathcal{W})$, $\mathcal{W}\subseteq\mathcal{X}$, where

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

with $\mathcal{H}=\{h_{\mathbf{w}}(\cdot)=L^{1/2}\sum_{l=0}^{P-1}\mathbb{1}(\mathbf{w}\in\Delta_l)\mathbb{1}(\cdot\in\Delta_l):\mathbf{w}\in\mathcal{W}\}$ a collection of Haar basis functions based on the partition $\{\Delta_l:0\le l<P\}$. It follows that $\mathtt{M}_{\mathcal{H},\mathbb{R}^d}=L^{1/2}$ and $\mathtt{S}_{\mathcal{H}}=1$.

If $\mathcal{W}=\mathcal{X}$, then we set $L=P$, $\mathbb{Q}_{\mathcal{H}}=\mathbb{P}_X$, $\mathcal{Q}_{\mathcal{H}}=\mathcal{X}$, and the conditions of Theorem 3 are satisfied with $\mathtt{E}_{\mathcal{H}}=L^{-1/2}$. Then, for $X_n=\zeta_n$, (3) holds with $\varrho_n=\log(nL)/\sqrt{n/L}$, assuming that $n^{-1}L\log(nL)\to 0$.

If $\mathcal{W}\subsetneq\mathcal{X}$, assume $\mathcal{W}\subseteq\sqcup_{0\le l<P}\Delta_l$. If $\mathbb{P}_X(\sqcup_{0\le l<P}\Delta_l)<1$, then $\{\Delta_l:0\le l<P\}$ is no longer a quasi-uniform partition of $\mathcal{X}$ with respect to $\mathbb{P}_X$. The surrogate measure can help in this setting: we may add or refine cells to handle the residual probability $\mathbb{P}_X[(\sqcup_{0\le l<P}\Delta_l)^c]$. For example, suppose that for some $\mathring{P}\in\mathbb{N}$ we have

$$\mathring{P}\le\frac{\mathbb{P}_X((\sqcup_{0\le l<P}\Delta_l)^c)}{\min_{0\le l<P}\mathbb{P}_X(\Delta_l)}<\mathring{P}+1.$$

Set $L=P+\mathring{P}$. For any collection of disjoint cells $\{\Delta_l:P\le l<L\}$ in $\mathcal{X}\cup\operatorname{Supp}(\mathcal{H})^c$, take $\mathbb{Q}_{\mathcal{H}}$ to agree with $\mathbb{P}_X$ on $\sqcup_{0\le l<P}\Delta_l$ and $\mathbb{Q}_{\mathcal{H}}(\Delta_l)=\mathring{P}^{-1}\mathbb{P}_X[(\sqcup_{0\le l<P}\Delta_l)^c]$ for $l=P,\dots,L-1$. Then, the enlarged class of cells $\{\Delta_l:0\le l<L+K\}$ and the probability measure $\mathbb{Q}_{\mathcal{H}}$ satisfy conditions (i) and (ii) in Theorem 3. It follows that $\mathtt{E}_{\mathcal{H}}=L^{-1/2}$ and hence, for $X_n=\zeta_n$, (3) holds with $\varrho_n=\log(nL)/\sqrt{n/L}$, assuming that $n^{-1}L\log(nL)\to 0$. In particular, the quasi-uniformity condition of $\mathbb{P}_X$ is required on a cover of $\mathcal{W}$, instead of

on a cover of $\mathcal{X}$, at the expense of possibly increasing the number of cells to account for the residual probability $\mathbb{P}_X[(\sqcup_{0\le l<P}\Delta_l)^c]$. ▲

Theorem 3, and in particular Example 2, showcases the existence of a class of stochastic processes for which a uniform Gaussian strong approximation can be 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 projection error ($\mathcal{H}$ is spanned by a Haar basis), and the coupling error is controlled via our generalized Tusnády's inequality. See [11] for details.

5.2. *Residual-Based Empirical Process.* The next result is the analogue of Theorem 2.

THEOREM 4. Suppose $(\mathbf{z}_i = (\mathbf{x}_i, y_i) : 1 \le i \le n)$ are i.i.d. random vectors taking values in $(\mathbb{R}^{d+1}, \mathcal{B}(\mathbb{R}^{d+1}))$ with common law $\mathbb{P}_Z$, where $\mathbf{x}_i$ has distribution $\mathbb{P}_X$ supported on $\mathcal{X} \subseteq \mathbb{R}^d$, y_i has distribution $\mathbb{P}_Y$ supported on $\mathcal{Y} \subseteq \mathbb{R}$, and the following conditions hold.

(i) $\mathcal{G} \subseteq \operatorname{Span}\{\mathbb{1}_{\Delta_l} : 0 \le l < L\}$ is a class of Haar functions on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P}_X)$.
(ii) There exists a surrogate measure $\mathbb{Q}_{\mathcal{G}}$ for $\mathbb{P}_X$ with respect to $\mathcal{G}$ such that $\{\Delta_l : 0 \le l < L\}$ forms a *quasi-uniform partition* of $\mathcal{Q}_{\mathcal{G}}$ with respect to $\mathbb{Q}_{\mathcal{G}}$:

$$\mathcal{Q}_{\mathcal{G}} \subseteq \sqcup_{0\le l<L}\Delta_l \qquad \text{and} \qquad \frac{\max_{0\le l<L}\mathbb{Q}_{\mathcal{G}}(\Delta_l)}{\min_{0\le l<L}\mathbb{Q}_{\mathcal{G}}(\Delta_l)} \le \rho < \infty.$$

(iii) $\mathcal{G}$ is a VC-type class with envelope function $\mathtt{M}_{\mathcal{G}}$ over $\mathcal{Q}_{\mathcal{G}}$ with $\mathtt{c}_{\mathcal{G}} \ge e$ and $\mathtt{d}_{\mathcal{G}} \ge 1$.
(iv) $\mathcal{R}$ is a real-valued pointwise measurable class of functions on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_Y)$.
(v) $\mathcal{R}$ is a VC-type class with envelope $M_{\mathcal{R},\mathcal{Y}}$ over $\mathcal{Y}$ with $\mathtt{c}_{\mathcal{R},\mathcal{Y}} \ge e$ and $\mathtt{d}_{\mathcal{R},\mathcal{Y}} \ge 1$, where $M_{\mathcal{R},\mathcal{Y}}(y) + \mathtt{pTV}_{\mathcal{R},(-|y|,|y|)} \le \mathtt{v}(1+|y|^\alpha)$ for all $y \in \mathcal{Y}$, for some $\mathtt{v} > 0$, and for some $\alpha \ge 0$. Furthermore, if $\alpha > 0$, then $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[\exp(|y_i|)|\mathbf{x}_i = \mathbf{x}] \le 2$.
(vi) There exists a constant $\mathtt{k}$ such that $|\log_2 \mathtt{E}_{\mathcal{G}}| + |\log_2 \mathtt{M}_{\mathcal{G}}| + |\log_2 L| \le \mathtt{k}\log_2 n$.

Then, on a possibly enlarged probability space, there exists a sequence of mean-zero Gaussian processes $(Z_n^R(g,r) : (g,r) \in \mathcal{G} \times \mathcal{R})$ with almost sure continuous trajectories on $(\mathcal{G} \times \mathcal{R}, \mathfrak{d}_{\mathbb{P}_Z})$ 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
- $\mathbb{P}[\|R_n - Z_n^R\|_{\mathcal{G}\times\mathcal{R}} > C_1 C_{\mathtt{v},\alpha}(C_\rho \mathsf{U}_n(t) + \mathsf{V}_n(t))] \le C_2 e^{-t} + L e^{-C_\rho n/L}$ for all $t > 0$,

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

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

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

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

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

The first term, $\mathsf{U}_n(t)$, can be interpreted as a "variance" contribution based on the effective sample size n/L, up to $\operatorname{polylog}(n)$ 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 $\mathbb{E}[r(y_i)|\mathbf{x}_i = \mathbf{x}]$, which may not necessarily lie in the span of Haar basis. In the special case when $\mathcal{R}$ 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, but such a construction is not possible when uniformity over $\mathcal{R}$ is desired.

Theorem 4 gives the following uniform Gaussian strong approximation result.

COROLLARY 6 (VC-type Haar Basis). Suppose the conditions of Theorem 4 hold with constants $\mathtt{c}$ and $\mathtt{d}$. Then, if $n^{-1}L\log(nL) \to 0$, $\|R_n - Z_n^R\|_{\mathcal{G}\times\mathcal{R}} = O(\varrho_n)$ a.s. with

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

Setting aside $\mathtt{M}_{\mathcal{G}}$ and $\mathtt{E}_{\mathcal{G}}$, an approximation rate is $(\log n)^{2\alpha+1}(n/L)^{-1/2} + \mathbb{1}(|\mathcal{R}| > 1)(\max_{0\le l<L}\|\Delta_l\|_\infty)\sqrt{\log n}$, which can achieve the optimal univariate KMT strong approximation rate based on the effective sample size n/L, up to a $\mathrm{polylog}(n)$ term, when $\mathcal{R}$ is a singleton function class. See [11, Section SA-V] for details.

The next section illustrates Theorem 4 with an example studying nonparametric regression estimation based on a Haar basis approximation.

5.3. *Example: Haar Partitioning-based Regression.* Suppose $(\mathbf{z}_i = (\mathbf{x}_i, y_i), 1 \le i \le n)$ are i.i.d. random vectors taking values in $(\mathcal{X}\times\mathbb{R}, \mathcal{B}(\mathcal{X}\times\mathbb{R}))$ with $\mathcal{X}\subseteq\mathbb{R}^d$. As in Section 4.1, consider the regression estimand (13), focusing again on the two examples $\mathcal{R}_1$ and $\mathcal{R}_2$. Instead of local polynomial regression, we study the Haar partitioning-based estimator:

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

where $\mathbf{p}(\mathbf{u}) = (\mathbb{1}(\mathbf{u}\in\Delta_l) : 0 \le l < L)$, and $\mathbf{w}\in\mathcal{W}\subseteq\mathcal{X}$. As in Example 2, either $\mathcal{W} = \mathcal{X}$ or $\mathcal{W}\subsetneq\mathcal{X}$, but for simplicity we discuss only the former case, and hence we assume that $\{\Delta_l : 0 \le l < L\}$ is a quasi-uniform partition of $\mathcal{Q}_{\mathcal{H}} = \mathcal{X}$ with respect to $\mathbb{Q}_{\mathcal{H}} = \mathbb{P}_X$.

The estimation error can again be decomposed into three terms:

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

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 linearization term takes the form

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

for $l = 1, 2$, where $\mathcal{G} = \{k_{\mathbf{w}}(\cdot) : \mathbf{w}\in\mathcal{W}\}$ with $k_{\mathbf{w}}(\mathbf{u}) = L^{-1/2}\sum_{0\le l<L}\mathbb{1}(\mathbf{w}\in\Delta_l)\mathbb{1}(\mathbf{u}\in\Delta_l)/\mathbb{P}_X(\Delta_l)$ the equivalent kernel. Under standard regularity conditions including smoothness and moment assumptions [11, Section SA-V.3],

$$\sup_{r\in\mathcal{R}_1}\big|\mathbf{e}_1^\top(\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1})\mathbf{T}_r\big| = O(\log(nL)L/n + (\log(nL)L/n)^{3/2}\log n) \qquad \text{a.s.},$$

$$\sup_{r\in\mathcal{R}_2}\big|\mathbf{e}_1^\top(\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1})\mathbf{T}_r\big| = O(\log(nL)L/n) \qquad \text{a.s.},$$

$$\sup_{\mathbf{w}\in\mathcal{W}, r\in\mathcal{R}_l}\big|\mathbb{E}[\check{\theta}(\mathbf{w}, r)|\mathbf{x}_1, \cdots, \mathbf{x}_n] - \theta(\mathbf{w}, r)\big| = O\big(\max_{0\le l<L}\|\Delta_l\|_\infty\big) \qquad \text{a.s.}, \quad l = 1, 2,$$

provided that $\log(nL)L/n \to 0$. 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 $\mathtt{d}_{\mathcal{R}_1} = 1$, $\alpha = 1$ if $\sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[\exp(|y_i|)|\mathbf{x}_i = \mathbf{x}] \leq 2$, and condition (v) 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 (v) in Theorem 4 with $\alpha = 0$, and we can take $\mathtt{c}_{\mathcal{R}_2}$ to be some universal constant and $\mathtt{d}_{\mathcal{R}_2} = 2$ by [33, Theorem 2.6.7], which implies that $\|R_n - Z_n^R\|_{\mathcal{G}\times\mathcal{R}_1} = O(\varrho_n)$ a.s. with

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

provided that $\log(n)L/n \to 0$. A uniform Gaussian strong approximation for the Haar partitioning-based regression processes $(\sqrt{n/L}(\check{\theta}(\mathbf{w},r) - \theta(\mathbf{w},r)) : (\mathbf{w},r) \in \mathcal{W} \times \mathcal{R}_l)$, $l = 1, 2$, follows directly from the results obtained above, as illustrated in Section 4.1.

This example showcases a statistical application of our strong approximation result (Theorem 4) where the optimal univariate KMT strong approximation rate based on the effective sample size n/L is achievable, up to $\operatorname{polylog}(n)$ terms and the complexity of $\mathcal{R}$. See [11, Section SA-V.3] for omitted details.

Acknowledgments. We specially thank Boris Hanin for many insightful discussions. We also thank Rajita Chandak, Jianqing Fan, Kengo Kato, Jason Klusowski, Xinwei Ma, Boris Shigida, Jennifer Sun, Rocio Titiunik, Will Underwood, and two reviewers for their comments and suggestions.

Funding. The first author was supported by the National Science Foundation through grants DMS-2210561 and SES-2241575.

SUPPLEMENTARY MATERIAL

Proofs and other technical results

The supplementary material [11] collects detailed proofs of our main results, and also provides other technical results that may be of independent interest.

REFERENCES

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

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

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

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

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

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

[7] Cattaneo, M. D., Farrell, M. H. and Feng, Y. (2020). Large Sample Properties of Partitioning-Based Series Estimators. *Annals of Statistics* **48** 1718–1741.

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

[9] Cattaneo, M. D., Jansson, M. and Ma, X. (2024). Local Regression Distribution Estimators. *Journal of Econometrics* **240** 105074.

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

[11] Cattaneo, M. D. and Yu, R. (2024). Supplement to 'Strong Approximations for Empirical Processes Indexed by Lipschitz Functions'.

[12] Cattaneo, M. D., Chandak, R., Jansson, M. and Ma, X. (2024). Local Polynomial Conditional Density Estimators. *Bernoulli* **30** 3193–3223.

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

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

[15] 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** 1 – 56.

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

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

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

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

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

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

[22] 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** 73–118.

[23] 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.

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

[25] 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* 415–428. Springer.

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

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

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

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

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

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

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

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

[34] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall/CRC.

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

[36] 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** 203–220.

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