

# STRONG APPROXIMATIONS FOR EMPIRICAL PROCESSES INDEXED BY LIPSCHITZ FUNCTIONS

BY MATIAS D. CATTANEO<sup>1,a</sup>, AND RUIQI (RAE) YU<sup>1,b</sup>

<sup>1</sup>*Department of Operations Research and Financial Engineering, Princeton University,*  
<sup>a</sup>*cattaneo@princeton.edu;* <sup>b</sup>*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 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

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

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  $n \rightarrow \infty$  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

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

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  $n \rightarrow \infty$  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

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

---

*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  $\varrho_n \rightarrow 0$  as  $n \rightarrow \infty$ . 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  $\varrho_n$  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:  $\mathbf{x}_i \sim \text{Uniform}(\mathcal{X})$ ,  $\mathcal{X} = [0, 1]$ , 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

$$(4) \quad \varrho_n = n^{-1/2} \log n,$$

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  $b \rightarrow 0$  as  $n \rightarrow \infty$ , 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

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

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  $b \rightarrow 0$  as  $n \rightarrow \infty$ , and established (3) with

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

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

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

where  $b \rightarrow 0$  as  $n \rightarrow \infty$ , 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

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

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*:

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

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  $b \rightarrow 0$  as  $n \rightarrow \infty$ , 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

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

where the effective sample size is still  $nb^{d+1}$ . This is suboptimal because  $nb^d$  is the (point-wise) 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

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

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  $\text{polylog}(n)$  term, where  $\text{polylog}(n) = \log^a(n)$  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  $\text{polylog}(n)$  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  $\text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . The binomial distribution with parameter  $n \in \mathbb{N}$  and  $p \in [0, 1]$  is denoted by  $\text{Bin}(n, p)$ .  $|\mathcal{A}|$  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  $\mathbf{A} \in \mathbb{R}^{q \times q}$ ,  $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \dots \geq \sigma_d(\mathbf{A}) \geq 0$  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  $\limsup_{n \rightarrow \infty} |a_n/b_n| = 0$ , 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  $\limsup_{n \rightarrow \infty} \mathbb{P}[|a_n/b_n| \geq \varepsilon] = 0$  for all  $\varepsilon > 0$ , and write  $a_n = O_{\mathbb{P}}(b_n)$  if  $\limsup_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} \mathbb{P}[|a_n/b_n| \geq M] = 0$ .

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  $\sigma$ -algebra generated by  $\mathcal{U}$  and  $\mathcal{B}(\mathcal{U}) \otimes \mathcal{B}(\mathcal{V})$  denotes the product  $\sigma$ -algebra. Let  $\mu$  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  $\mu$  on  $R$ , that is,  $\mu|_R(U) = \mu(U \cap R)$  for all  $U \in \mathcal{B}(\mathcal{U})$ . Two measures  $\mu$  and  $\nu$  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  $\mu$  is  $\text{Supp}(\mu) = \text{closure}(\cup\{U \in \mathcal{B}(\mathcal{U}) : \mu(U) \neq 0\})$ . The Lebesgue measure is denoted by  $\mathbf{m}$ . 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  $\text{Supp}(f) = \{\mathbf{u} \in \mathcal{U} : f(\mathbf{u}) > 0\}$  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 exists,  $\nabla f(\mathbf{x})$  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  $\mu$  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  $\lim_{m \rightarrow \infty} g_m(\mathbf{u}) = f(\mathbf{u})$  for all  $\mathbf{u} \in \mathbb{R}^q$ .

**DEFINITION 2.** Let  $\text{Supp}(\mathcal{F}) = \cup_{f \in \mathcal{F}} \text{Supp}(f)$ . 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  $\text{Supp}(\mathbb{P}) \cap \text{Supp}(\mathcal{F})$ .
- (ii)  $\mathbb{Q}_{\mathcal{F}}(\text{Supp}(\mathcal{F}) \setminus \text{Supp}(\mathbb{P})) = 0$ .

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

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

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

where  $\mathcal{P}_P = \{(a_1, \dots, a_P) : a_1 \leq \dots \leq a_P\}$  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

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



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  $K_{\mathcal{F},\mathcal{C}}$  such that for any cube  $\mathcal{D} \subseteq \mathbb{R}^q$  with edges of length  $\ell$  parallel to the coordinate axes,

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

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

$$M_{\mathcal{F},\mathcal{C}} = \sup_{\mathbf{u} \in \mathcal{C}} M_{\mathcal{F},\mathcal{C}}(\mathbf{u}), \quad M_{\mathcal{F},\mathcal{C}}(\mathbf{u}) = \sup_{f \in \mathcal{F}} |f(\mathbf{u})|, \quad \mathbf{u} \in \mathcal{C}.$$

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

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

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

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

**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

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

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

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

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  $c_{\mathcal{F},\mathcal{C}} > 0$  and  $d_{\mathcal{F},\mathcal{C}} > 0$  such that

$$N_{\mathcal{F},\mathcal{C}}(\varepsilon, M_{\mathcal{F},\mathcal{C}}) \leq c_{\mathcal{F},\mathcal{C}} \varepsilon^{-d_{\mathcal{F},\mathcal{C}}}, \quad \varepsilon \in (0, 1).$$

**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  $a_{\mathcal{F},\mathcal{C}} > 0$  and  $b_{\mathcal{F},\mathcal{C}} > 0$  such that

$$\log N_{\mathcal{F},\mathcal{C}}(\varepsilon, M_{\mathcal{F},\mathcal{C}}) \leq a_{\mathcal{F},\mathcal{C}} \varepsilon^{-b_{\mathcal{F},\mathcal{C}}}, \quad \varepsilon \in (0, 1).$$

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  $\text{TV}_{\mathcal{F}} = \text{TV}_{\mathcal{F},\mathcal{Q}_{\mathcal{F}}}$ ,  $K_{\mathcal{F}} = K_{\mathcal{F},\mathcal{Q}_{\mathcal{F}}}$ ,  $M_{\mathcal{F}} = M_{\mathcal{F},\mathcal{Q}_{\mathcal{F}}}$ ,  $M_{\mathcal{F}}(\mathbf{u}) = M_{\mathcal{F},\mathcal{Q}_{\mathcal{F}}}(\mathbf{u})$ ,  $L_{\mathcal{F}} = L_{\mathcal{F},\mathcal{Q}_{\mathcal{F}}}$ , and so on, whenever there is no confusion.

### 3. General Empirical Process. Let

$$m_{n,d} = \begin{cases} n^{-1/2} \sqrt{\log n} & \text{if } d = 1 \\ n^{-1/(2d)} & \text{if } d \geq 2 \end{cases} \quad \text{and} \quad l_{n,d} = \begin{cases} 1 & \text{if } d = 1 \\ n^{-1/2} \sqrt{\log n} & \text{if } d = 2, \\ n^{-1/d} & \text{if } d \geq 3 \end{cases}$$

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)  $M_{\mathcal{H}} < \infty$  and  $J(1, \mathcal{H}, M_{\mathcal{H}}) < \infty$ .

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
- $\mathbb{P}[\|X_n - Z_n^X\|_{\mathcal{H}} > C_1 S_n(t)] \leq C_2 e^{-t}$  for all  $t > 0$ ,

where  $C_1$  and  $C_2$  are universal constants, and

$$S_n(t) = \min_{\delta \in (0,1)} \{A_n(t, \delta) + F_n(t, \delta)\},$$

where

$$\begin{aligned} A_n(t, \delta) &= \min \{m_{n,d} \sqrt{M_{\mathcal{H}}}, l_{n,d} \sqrt{c_2 L_{\mathcal{H}}}\} \sqrt{c_1 \text{TV}_{\mathcal{H}}} \sqrt{t + \log N_{\mathcal{H}}(\delta, M_{\mathcal{H}})} \\ &\quad + \sqrt{\frac{M_{\mathcal{H}}}{n}} \min \{ \sqrt{\log n} \sqrt{M_{\mathcal{H}}}, \sqrt{c_3 K_{\mathcal{H}} + M_{\mathcal{H}}}\} (t + \log N_{\mathcal{H}}(\delta, M_{\mathcal{H}})) \end{aligned}$$

$$c_1 = d \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} \prod_{j=1}^{d-1} \sigma_j(\nabla \phi_{\mathcal{H}}(\mathbf{x})), \quad c_2 = \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} \frac{1}{\sigma_d(\nabla \phi_{\mathcal{H}}(\mathbf{x}))}, \quad c_3 = 2^{d-1} d^{d/2-1} c_1 c_2^{d-1},$$

and

$$F_n(t, \delta) = J(\delta, \mathcal{H}, M_{\mathcal{H}}) M_{\mathcal{H}} + \frac{M_{\mathcal{H}} J^2(\delta, \mathcal{H}, M_{\mathcal{H}})}{\delta^2 \sqrt{n}} + \delta M_{\mathcal{H}} \sqrt{t} + \frac{M_{\mathcal{H}}}{\sqrt{n}} t.$$

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  $\delta$ -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  $F_n(t, \delta)$  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:

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

where  $\Pi_0 X_n(h) = X_n \circ \Pi_0 h$  with  $\Pi_0$  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  $\text{Bin}(n, \frac{1}{2})$  with  $\text{Normal}(\frac{n}{2}, \frac{n}{4})$ , 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  $A_n(t, \delta)$  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  $\mathcal{X} \cap \text{Supp}(\mathcal{H})$ , 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  $\mathbf{x}_i \sim \text{Uniform}(\mathcal{X})$  with  $\mathcal{X} = \times_{l=1}^d [a_l, b_l]$ , where  $-\infty < a_l < b_l < \infty$ ,  $l = 1, 2, \dots, d$ . Setting  $\mathbb{Q}_{\mathcal{H}} = \mathbb{P}_X$  and  $\phi_{\mathcal{H}}(x_1, \dots, x_d) = ((b_1 - a_1)^{-1}(x_1 - a_1), \dots, (b_d - a_d)^{-1}(x_d - a_d))$  verifies assumption (ii) in Theorem 1. In this case,  $c_1 = d \max_{1 \leq l \leq d} |b_l - a_l| \prod_{l=1}^d |b_l - a_l|^{-1}$ ,  $c_2 = \max_{1 \leq l \leq d} |b_l - a_l|$  and  $c_3 = 2^{d-1} d^{d/2} \max_{1 \leq l \leq d} |b_l - a_l|^d \prod_{l=1}^d |b_l - a_l|^{-1}$ .

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  $\text{Supp}(\mathcal{H})$ . For a random vector  $\mathbf{V} = (V_1, \dots, V_d) \in \mathbb{R}^d$  with distribution  $\mathbb{P}_V$ , the Rosenblatt transformation is

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

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  $\mathcal{Q}_{\mathcal{H}} = \times_{l=1}^d [a_l, b_l]$ ,  $-\infty \leq a_l < b_l \leq \infty$ . Then, the Rosenblatt transformation  $\phi_{\mathcal{H}} = T_{\mathbb{Q}_{\mathcal{H}}}$  is a normalizing transformation, and we obtain

$$c_1 = d \sup_{\mathbf{u} \in \mathbb{Q}_{\mathcal{H}}} \frac{f_Q(\mathbf{u})}{\min\{f_{Q,1}(u_1), f_{Q,2|1}(u_2|u_1), \dots, f_{Q,d|d-1}(u_d|u_1, \dots, u_{d-1})\}},$$

$$c_2 = \sup_{\mathbf{u} \in \mathbb{Q}_{\mathcal{H}}} \frac{1}{\min\{f_{Q,1}(u_1), f_{Q,2|1}(u_2|u_1), \dots, f_{Q,d|d-1}(u_d|u_1, \dots, u_{d-1})\}},$$

and  $c_3 = 2^{d-1} d^{d/2-1} c_1 c_2^{d-1}$ , where  $f_{Q,j|j-1}(\cdot|u_1, \dots, u_{j-1})$  denotes the conditional density of  $Q_j|Q_1 = u_1, \dots, Q_{j-1} = u_{j-1}$  for  $\mathbf{Q} = (Q_1, \dots, Q_d) \sim \mathbb{Q}_{\mathcal{H}}$ .

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

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

$$c_1 \leq d \frac{\bar{f}_Q^2}{\underline{f}_Q} \bar{\mathcal{Q}}_{\mathcal{H}} \quad \text{and} \quad c_2 \leq \frac{\bar{f}_Q}{\underline{f}_Q} \bar{\mathcal{Q}}_{\mathcal{H}},$$

where  $\underline{f}_Q = \inf_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} f_Q(\mathbf{x})$ ,  $\bar{f}_Q = \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} f_Q(\mathbf{x})$ , and  $\bar{\mathcal{Q}}_{\mathcal{H}} = \max_{1 \leq l \leq d} |b_l - a_l|$ .

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

$$(11) \quad c_1 \leq d \sup_{\mathbf{x} \in \mathbb{Q}_{\mathcal{H}}} \max\{f_{X,1}(x_1), f_{X,2|1}(x_2|x_1), \dots, f_{X,d|d-1}(x_d|x_{-d})\}^{d-1}$$

$$\leq d \min_{1 \leq k \leq d} \{\boldsymbol{\Sigma}_{k,k} - \boldsymbol{\Sigma}_{k,1:k-1} \boldsymbol{\Sigma}_{1:k-1,1:k-1}^{-1} \boldsymbol{\Sigma}_{1:k-1,k}\}^{-(d-1)/2}$$

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  $\text{Supp}(\mathcal{H})$ , 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

$$c_1 \leq d \frac{\bar{f}_Q^2}{\underline{f}_Q} S_{\chi} \quad \text{and} \quad c_2 \leq \frac{\bar{f}_Q}{\underline{f}_Q} S_{\chi},$$

where  $S_{\chi} = \frac{\sup_{\mathbf{x} \in [0,1]^d} |\det(\nabla \chi^{-1}(\mathbf{x}))|}{\inf_{\mathbf{x} \in [0,1]^d} |\det(\nabla \chi^{-1}(\mathbf{x}))|} \|\|\nabla \chi^{-1}\|_2\|_{\infty}$ . 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  $\text{Supp}(\mathcal{H})$  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

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

where  $K : \mathcal{K} \rightarrow \mathbb{R}$  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  $b \rightarrow 0$  as  $n \rightarrow \infty$  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

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

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

Variants of Example 1 have been discussed extensively in prior literature on strong approximations because the process  $\xi_n$  is non-Donsker whenever  $b \rightarrow 0$ , 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  $\text{Supp}(\mathcal{H}) = \mathcal{W} + b \cdot \mathcal{K}$ . We illustrate two sets of primitive conditions implying assumption (ii) in Theorem 1.

- Suppose that  $\mathcal{X} = \times_{l=1}^d [a_l, b_l]$ ,  $-\infty \leq a_l < b_l \leq \infty$ , 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,  $c_1 = O(1)$  and  $c_2 = O(1)$ , and hence  $c_3 = O(1)$ , 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  $\text{Supp}(\mathcal{H})$  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  $\mathcal{X} = \mathbb{R}_+^d$ ,

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

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

where  $\bar{a}_l = \max\{a_l - b, 0\}$ ,  $\bar{b}_l = b_l + b$ ,  $\Upsilon = \times_{l=1}^d [\bar{a}_l, \bar{b}_l + 1] \setminus \times_{l=1}^d [\bar{a}_l, \bar{b}_l]$ , and  $\phi_{\mathcal{H}} = T_{\mathbb{Q}_{\mathcal{H}} \circ \chi^{-1}} \circ \chi$  with  $\chi(x_1, \dots, x_d) = ((\bar{b}_1 - \bar{a}_1)^{-1}(x_1 - \bar{a}_1), \dots, (\bar{b}_d - \bar{a}_d)^{-1}(x_d - \bar{a}_d))$ . 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.  $\blacktriangle$

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  $c_1 < \infty$  and  $c_2 < \infty$  (hence  $c_3 < \infty$ ) 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 ( $M_{\mathcal{H}} < \infty$ ), but without assuming they are Lipschitz ( $L_{\mathcal{H}} = \infty$ ).

**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  $M_{\mathcal{H}}$  over  $\mathcal{Q}_{\mathcal{H}}$  with constants  $c_{\mathcal{H}} \geq e$  and  $d_{\mathcal{H}} \geq 1$ . Then, (3) holds with

$$\varrho_n = m_{n,d} \sqrt{\log n} \sqrt{c_1 M_{\mathcal{H}} \text{TV}_{\mathcal{H}}} + \frac{\log n}{\sqrt{n}} \min\{\sqrt{\log n} \sqrt{M_{\mathcal{H}}}, \sqrt{c_3 K_{\mathcal{H}} + M_{\mathcal{H}}}\} \sqrt{M_{\mathcal{H}}}.$$

This corollary recovers the main result in [29, Theorem 1.1] when  $d \geq 2$ , where  $m_{n,d} = n^{-1/(2d)}$ . It also covers  $d = 1$ , where  $m_{n,1} = n^{-1/2} \sqrt{\log n}$ , 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  $K_{\mathcal{H}} = O(1)$ , 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:

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

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  $\mathbb{R}$ ,  $\text{TV}_{\mathcal{H}, \mathcal{I}} \leq \text{pTV}_{\mathcal{H}, \mathcal{I}}$  provided that  $M_{\mathcal{H}, \mathcal{I}} < \infty$  [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  $d = 1$ .  $\square$

We illustrate the usefulness of Corollary 1 with Example 1.

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

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 ( $L_{\mathcal{H}} < \infty$ ).

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  $M_{\mathcal{H}}$  over  $\mathcal{Q}_{\mathcal{H}}$  with constants  $c_{\mathcal{H}} \geq e$  and  $d_{\mathcal{H}} \geq 1$ . Then, (3) holds with

$$\begin{aligned} \varrho_n &= \min\{m_{n,d}\sqrt{M_{\mathcal{H}}}, l_{n,d}\sqrt{c_2L_{\mathcal{H}}}\}\sqrt{\log n}\sqrt{c_1\text{TV}_{\mathcal{H}}} \\ &\quad + \frac{\log n}{\sqrt{n}} \min\{\sqrt{\log n}\sqrt{M_{\mathcal{H}}}, \sqrt{c_3K_{\mathcal{H}} + M_{\mathcal{H}}}\}\sqrt{M_{\mathcal{H}}}. \end{aligned}$$

Putting aside  $M_{\mathcal{H}}$  and  $\text{TV}_{\mathcal{H}}$ , this corollary shows that if  $L_{\mathcal{H}} < \infty$ , then the rate of strong approximation can be improved. In particular, for  $d = 2$ ,  $m_{n,2} = n^{-1/4}$  but  $l_{n,2} = n^{-1/2}\sqrt{\log n}$ , implying that  $\varrho_n = n^{-1/2}\log n$  whenever  $K_{\mathcal{H}} = O(b^{-d/2})$ . 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  $K_{\mathcal{H}} = \infty$ .)

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,  $m_{n,d} = n^{-1/(2d)}$  but  $l_{n,d} = n^{-1/d}$  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 ( $L_{\mathcal{H}} < \infty$ ) 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,  $L_{\mathcal{H}} = O(b^{-d/2-1})$ , 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 Lipschitzian 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  $M_{\mathcal{H}}$  over  $\mathcal{Q}_{\mathcal{H}}$  with constants  $a_{\mathcal{H}} > 0$  and  $0 < b_{\mathcal{H}} < 2$ . Then, (3) holds as follows:

(i) If  $L_{\mathcal{H}} \leq \infty$ , then

$$\begin{aligned} \varrho_n &= m_{n,d} \sqrt{c_1 M_{\mathcal{H}} \text{TV}_{\mathcal{H}}} (\sqrt{\log n} + (c_1 m_{n,d}^2 M_{\mathcal{H}}^{-1} \text{TV}_{\mathcal{H}})^{-\frac{b_{\mathcal{H}}}{4}}) \\ &\quad + \sqrt{\frac{M_{\mathcal{H}}}{n}} \min\{\sqrt{\log n} \sqrt{M_{\mathcal{H}}}, \sqrt{c_3 K_{\mathcal{H}} + M_{\mathcal{H}}}\} (\log n + (c_1 m_{n,d}^2 M_{\mathcal{H}}^{-1} \text{TV}_{\mathcal{H}})^{-\frac{b_{\mathcal{H}}}{2}}), \end{aligned}$$

(ii) If  $L_{\mathcal{H}} < \infty$ , then

$$\begin{aligned} \varrho_n &= l_{n,d} \sqrt{c_1 c_2 L_{\mathcal{H}} \text{TV}_{\mathcal{H}}} (\sqrt{\log n} + (c_1 c_2 l_{n,d}^2 M_{\mathcal{H}}^{-2} L_{\mathcal{H}} \text{TV}_{\mathcal{H}})^{-\frac{b_{\mathcal{H}}}{4}}) \\ &\quad + \sqrt{\frac{M_{\mathcal{H}}}{n}} \min\{\sqrt{\log n} \sqrt{M_{\mathcal{H}}}, \sqrt{c_3 K_{\mathcal{H}} + M_{\mathcal{H}}}\} (\log n + (c_1 c_2 l_{n,d}^2 M_{\mathcal{H}}^{-2} L_{\mathcal{H}} \text{TV}_{\mathcal{H}})^{-\frac{b_{\mathcal{H}}}{2}}). \end{aligned}$$



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  $0 < b_{\mathcal{H}} < 2$ , we can take  $a_{\mathcal{H}} = \log(d + 1) + db_{\mathcal{H}}^{-1}$  so that  $\mathcal{H}$  is a polynomial-entropy class with constants  $(a_{\mathcal{H}}, b_{\mathcal{H}})$ . Then, Corollary 3(ii) implies that, for  $X_n = \xi_n$ , (3) holds with  $\varrho_n = a_{\mathcal{H}}^2 (nb^d)^{-\frac{1}{d}(1-\frac{b_{\mathcal{H}}}{2})} b^{-db_{\mathcal{H}}} + a_{\mathcal{H}}^2 (nb^d)^{-\frac{1}{2}+\frac{b_{\mathcal{H}}}{d}} b^{-\frac{db_{\mathcal{H}}}{2}}$ .  $\blacktriangle$

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 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  $n^{-\frac{q-d}{2dq}}$  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  $(\mathbb{Q}_{\mathcal{H}}, \phi_{\mathcal{H}}) = (\mathbb{P}_X, \text{Id})$  with Id denoting the identity map from  $[0, 1]^d$  to  $[0, 1]^d$ . It follows that  $M_{\mathcal{H}} = 1$ ,  $\text{TV}_{\mathcal{H}} = 1$ ,  $L_{\mathcal{H}} = 1$ . [33, Theorem 2.7.1] shows that  $\mathcal{H}$  is a polynomial-entropy class with constants  $a_{\mathcal{H}} = C_{q,d}$  and  $b_{\mathcal{H}} = d/q$ , where  $C_{q,d}$  is a constant depending on  $q$  and  $d$  only. Then, Corollary 3(ii) implies that, for  $X_n = \xi_n$ , (3) holds with

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

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  $L_{\mathcal{H}} = \infty$ ,  $b_{\mathcal{H}} = 2d/q$ , and keeping the other constants, Corollary 3(i) would give  $\varrho_n = n^{-\frac{q-d}{2dq}} \text{polylog}(n)$ , 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  $\rho$  is not implied by his Equation (3.2), and his result would give the strong approximation rate  $n^{-\frac{2q-d}{4qd}} \text{polylog}(n)$ .  $\square$

**4. Residual-Based Empirical Process.** Consider the simple local empirical process discussed in [13, Section 3.1]:

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

where  $\mathbf{x}_i \sim \mathbb{P}_X$ ,  $y_i \sim \mathbb{P}_Y$ , and  $b \rightarrow 0$  as  $n \rightarrow \infty$ . Using our notation,  $(\sqrt{nb^d}(S_n(\mathbf{w}) - \mathbb{E}[S_n(\mathbf{w})|\mathbf{x}_1, \dots, \mathbf{x}_n]) : \mathbf{w} \in \mathcal{W}) = (R_n(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  with  $\mathcal{G} = \{b^{-d/2}K(\frac{\cdot - \mathbf{w}}{b}) : \mathbf{w} \in \mathcal{W}\}$  and  $\mathcal{R} = \{\text{Id}\}$ , where  $\text{Id}$  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  $\text{polylog}(n)$  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  $\mathbb{Q}_\mathcal{G} = \mathbf{m} \circ \phi_\mathcal{G}$ , 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  $M_\mathcal{G}$  over  $\mathcal{Q}_\mathcal{G}$  with  $c_\mathcal{G} \geq e$  and  $d_\mathcal{G} \geq 1$ .
- (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  $c_{\mathcal{R}, \mathcal{Y}} \geq e$  and  $d_{\mathcal{R}, \mathcal{Y}} \geq 1$ , where  $M_{\mathcal{R}, \mathcal{Y}}(y) + \mathbf{pTV}_{\mathcal{R}, (-|y|, |y|)} \leq \mathbf{v}(1 + |y|^\alpha)$  for all  $y \in \mathcal{Y}$ , for some  $\mathbf{v} > 0$ , 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  $\mathbf{k}$  such that  $|\log_2 E_\mathcal{G}| + |\log_2 \text{TV}| + |\log_2 M_\mathcal{G}| \leq \mathbf{k} \log_2 n$ , where  $\text{TV} = \max\{\text{TV}_\mathcal{G}, \text{TV}_{\mathcal{G} \times \mathcal{V}_\mathcal{R}, \mathcal{Q}_\mathcal{G}}\}$  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
- $\mathbb{P}[\|R_n - Z_n^R\|_{\mathcal{G} \times \mathcal{R}} > C_1 C_{\mathbf{v}, \alpha} \mathbf{T}_n(t)] \leq C_2 e^{-t}$  for all  $t > 0$ ,

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

$$T_n(t) = A_n(t + k \log_2 n + d \log(cn))^{\alpha + \frac{3}{2}} \sqrt{d} + \frac{M_{\mathcal{G}}}{\sqrt{n}}(t + k \log_2 n + d \log(cn))^{\alpha + 1},$$

$$A_n = \min \left\{ \left( \frac{c_1^d M_{\mathcal{G}}^{d+1} \text{TV}^d \mathbf{E}_{\mathcal{G}}}{n} \right)^{\frac{1}{2d+2}}, \left( \frac{c_1^{\frac{d}{2}} c_2^{\frac{d}{2}} M_{\mathcal{G}} \mathbf{E}_{\mathcal{G}} \text{TV}^{\frac{d}{2}} L^{\frac{d}{2}}}{n} \right)^{\frac{1}{d+2}} \right\},$$

$$c_1 = d \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{G}}} \prod_{j=1}^{d-1} \sigma_j(\nabla \phi_{\mathcal{G}}(\mathbf{x})), \quad c_2 = \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{G}}} \frac{1}{\sigma_d(\nabla \phi_{\mathcal{G}}(\mathbf{x}))},$$

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

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 pTV condition on  $\mathcal{R}$  in (v), but 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  $\delta$ -net to obtain

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

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  $\Pi_0$  as in (10), we now use another mapping  $\Pi_2$  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,  $\Pi_2$  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  $c$  and  $d$ . Then,  $\|R_n - Z_n^R\|_{\mathcal{G} \times \mathcal{R}} = O(\varrho_n)$  a.s. with

$$\varrho_n = \min \left\{ \frac{(c_1^d M_{\mathcal{G}}^{d+1} \text{TV}^d \mathbf{E}_{\mathcal{G}})^{\frac{1}{2d+2}}}{n^{1/(2d+2)}}, \frac{(c_1^{\frac{d}{2}} c_2^{\frac{d}{2}} M_{\mathcal{G}} \text{TV}^{\frac{d}{2}} \mathbf{E}_{\mathcal{G}} L^{\frac{d}{2}})^{\frac{1}{d+2}}}{n^{1/(d+2)}} \right\} (\log n)^{\alpha+3/2} + \frac{(\log n)^{\alpha+1}}{\sqrt{n}} M_{\mathcal{G}}.$$

This corollary shows that our best attainable uniform Gaussian strong approximation rate for  $R_n$  is  $n^{-1/(d+2)}$  polylog( $n$ ), putting aside  $c_1$ ,  $c_2$ ,  $M_{\mathcal{G}}$ ,  $\text{TV}$ ,  $\mathbf{E}_{\mathcal{G}}$ , and  $L$ . 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  $\mathbb{P}_{\mathcal{Z}}$ , but leads to the better approximation rate  $n^{-\min\{1/(d+1), 1/2\}}$  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  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$  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

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

where we focus on two leading cases to streamline the discussion:  $\mathcal{R}_1 = \{\text{Id}\}$  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  $K : \mathbb{R}^d \rightarrow \mathbb{R}$  is non-negative, Lipschitz, and has compact support  $\mathcal{K}$ . Using standard multi-index notation,  $\mathbf{p}(\mathbf{u})$  denotes the  $\frac{(d+\mathbf{p})!}{d!\mathbf{p}!}$ -dimensional vector collecting the ordered elements  $\mathbf{u}^\nu/\nu!$  for  $0 \leq |\nu| \leq \mathbf{p}$ , where  $\mathbf{u}^\nu = u_1^{\nu_1} \cdots u_d^{\nu_d}$ ,  $\nu! = \nu_1! \cdots \nu_d!$  and  $|\nu| = \nu_1 + \cdots + \nu_d$ , for  $\mathbf{u} = (u_1, \dots, u_d)^\top$  and  $\nu = (\nu_1, \dots, \nu_d)^\top$ . A local polynomial regression estimator of  $\theta(\mathbf{w}; r)$  is

$$\widehat{\theta}(\mathbf{w}; r) = \mathbf{e}_1^\top \widehat{\boldsymbol{\beta}}(\mathbf{w}, r), \quad \widehat{\boldsymbol{\beta}}(\mathbf{w}, r) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^n (r(y_i) - \mathbf{p}(\mathbf{x}_i - \mathbf{w})^\top \boldsymbol{\beta})^2 K\left(\frac{\mathbf{x}_i - \mathbf{w}}{b}\right),$$

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:

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

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

It follows that the linear term is

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

for  $(g, r) \in \mathcal{G} \times \mathcal{R}_l$ ,  $l = 1, 2$ , and where  $\mathcal{G} = \{b^{-d/2} \mathfrak{K}_{\mathbf{w}}(\cdot - \frac{\mathbf{w}}{b}) : \mathbf{w} \in \mathcal{W}\}$  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  $y_i | \mathbf{x}_i$ ,

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

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

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

provided that  $\log(n)/(nb^d) \rightarrow 0$ . 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  $\mathbf{z}_i = (\mathbf{x}_i, y_i) \sim \mathbb{P}_Z$ ,  $i = 1, 2, \dots, n$ , 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  $M_{\{\varphi\}, \mathcal{B}} < \infty$ ,  $K_{\{\varphi\}, \mathcal{B}} < \infty$ ,  $\sup_{g \in \mathcal{G}} \text{TV}_{\{\varphi\}, \text{supp}(g) \times [0, 1]} < \infty$ , 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  $K_{\{\varphi\}, \mathcal{B}} < \infty$ , 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  $\varphi$  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  $nb^{d+1} \rightarrow \infty$ ). Our Theorem 2 addresses both problems: since  $\text{Supp}(\mathcal{H}) = \mathcal{W} + b\mathcal{K}$ , 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  $\mathbf{c}_1 = O(1)$ ,  $\mathbf{c}_2 = O(1)$ ,  $\mathbf{M}_{\mathcal{G}} = O(b^{-d/2})$ ,  $\mathbf{E}_{\mathcal{G}} = O(b^{d/2})$ ,  $\mathbf{K}_{\mathcal{G}} = O(b^{-d/2})$ ,  $\mathbf{TV} = O(b^{d/2-1})$ , and  $\mathbf{L} = O(b^{-d/2-1})$ . 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  $\hat{\mu}(\mathbf{w}) = \hat{\theta}(\mathbf{w}; r)$  for  $r \in \mathcal{R}_1$ . Recall that  $\mathcal{R}_1$  consists of the singleton of identity function Id. If  $b^{p+1}(nb^d)^{\frac{d+4}{2d+4}} (\log n)^{-1/2} + (nb^d)^{-\frac{d+1}{d+2}} (\log n)^2 = O(1)$ , then

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

where  $\text{Cov}(Z_n^R(\mathbf{w}_1), Z_n^R(\mathbf{w}_2)) = nb^d \text{Cov}(\mathbf{e}_1^\top \mathbf{H}_{\mathbf{w}_1}^{-1} \mathbf{S}_{\mathbf{w}_1, \text{Id}}, \mathbf{e}_1^\top \mathbf{H}_{\mathbf{w}_2}^{-1} \mathbf{S}_{\mathbf{w}_2, \text{Id}})$  for all  $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$ .

- Let  $\hat{F}(y|\mathbf{w}) = \hat{\theta}(\mathbf{w}; r_y)$  for  $r_y = \mathbb{1}(\cdot \leq y) \in \mathcal{R}_2$ . If  $b^{p+1}(nb^d)^{(d+4)/(2d+4)} (\log n)^{-1/2} = O(1)$  and  $(nb^d)^{-1} \log n = o(1)$ , then

$$\sup_{\mathbf{w} \in \mathcal{W}, y \in \mathbb{R}} |\sqrt{nb^d}(\hat{F}(y|\mathbf{w}) - F(y|\mathbf{w})) - Z_n^R(\mathbf{w}, y)| = O(r_n) \quad \text{a.s.},$$

where  $\text{Cov}(Z_n^R(\mathbf{w}_1, u_1), Z_n^R(\mathbf{w}_2, u_2)) = nb^d \text{Cov}(\mathbf{e}_1^\top \mathbf{H}_{\mathbf{w}_1}^{-1} \mathbf{S}_{\mathbf{w}_1, r_{u_1}}, \mathbf{e}_1^\top \mathbf{H}_{\mathbf{w}_2}^{-1} \mathbf{S}_{\mathbf{w}_2, r_{u_2}})$  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  $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{M}_{\mathcal{H}}, \mathbf{L}_{\mathcal{H}}, \mathbf{TV}_{\mathcal{H}}$ , and  $\mathbf{K}_{\mathcal{H}}$ . 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  $L \rightarrow \infty$  as  $n \rightarrow \infty$ . 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)  $\mathcal{H} \subseteq \text{Span}\{\mathbb{1}_{\Delta_l} : 0 \leq l < L\}$  is a class of Haar functions on  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P}_X)$ .
- (ii) There exists a surrogate measure  $\mathcal{Q}_{\mathcal{H}}$  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  $\mathcal{Q}_{\mathcal{H}}$ :

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

- (iii)  $M_{\mathcal{H}} < \infty$ .

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
- $\mathbb{P}[\|X_n - Z_n^X\|_{\mathcal{H}} > C_1 C_\rho P_n(t)] \leq C_2 e^{-t} + L e^{-C_\rho n/L}$  for all  $t > 0$ ,

where  $C_1$  and  $C_2$  are universal constants,  $C_\rho$  is a constant that only depends on  $\rho$ , and

$$P_n(t) = \min_{\delta \in (0,1)} \left\{ H_n(t, \delta) + F_n(t, \delta) \right\},$$

with

$$\begin{aligned} H_n(t, \delta) &= \sqrt{\frac{M_{\mathcal{H}} \mathbf{E}_{\mathcal{H}}}{n/L}} \sqrt{t + \log N_{\mathcal{H}}(\delta, M_{\mathcal{H}})} \\ &\quad + \sqrt{\frac{\min\{\log_2 L, \mathbf{S}_{\mathcal{H}}^2\}}{n}} M_{\mathcal{H}}(t + \log N_{\mathcal{H}}(\delta, M_{\mathcal{H}})), \end{aligned}$$

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

This theorem shows that if  $n^{-1}L \log(nL) \rightarrow 0$ , 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  $\text{Bin}(n, p)$  with  $\underline{p} \leq p \leq \bar{p}$ , 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  $M_{\mathcal{H}}$ ,  $E_{\mathcal{H}}$ , and  $S_{\mathcal{H}}$ , 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  $M_{\mathcal{H}}$  over  $\mathcal{Q}_{\mathcal{H}}$  with constants  $c_{\mathcal{H}} \geq e$  and  $d_{\mathcal{H}} \geq 1$ . Then, if  $n^{-1}L \log(nL) \rightarrow 0$ , (3) holds with

$$\varrho_n = \sqrt{\frac{M_{\mathcal{H}} E_{\mathcal{H}}}{n/L}} \sqrt{\log n} + \sqrt{\frac{\min\{\log_2 L, S_{\mathcal{H}}^2\}}{n}} M_{\mathcal{H}} \log n.$$

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

For  $L$  proportional to  $\mathbb{P}_X(\Delta_l)^{-1}$ , up to  $\rho$ , 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

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

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  $\{\Delta_l : 0 \leq l < P\}$ . It follows that  $M_{\mathcal{H}, \mathbb{R}^d} = L^{1/2}$  and  $S_{\mathcal{H}} = 1$ .

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

If  $\mathcal{W} \subsetneq \mathcal{X}$ , assume  $\mathcal{W} \subseteq \sqcup_{0 \leq l < P} \Delta_l$ . If  $\mathbb{P}_X(\sqcup_{0 \leq l < P} \Delta_l) < 1$ , then  $\{\Delta_l : 0 \leq l < P\}$  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  $\mathbb{P}_X[(\sqcup_{0 \leq l < P} \Delta_l)^c]$ . For example, suppose that for some  $\mathring{P} \in \mathbb{N}$  we have

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

Set  $L = P + \mathring{P}$ . For any collection of disjoint cells  $\{\Delta_l : P \leq l < L\}$  in  $\mathcal{X} \cup \text{Supp}(\mathcal{H})^c$ , take  $\mathcal{Q}_{\mathcal{H}}$  to agree with  $\mathbb{P}_X$  on  $\sqcup_{0 \leq l < P} \Delta_l$  and  $\mathcal{Q}_{\mathcal{H}}(\Delta_l) = \mathring{P}^{-1} \mathbb{P}_X[(\sqcup_{0 \leq l < P} \Delta_l)^c]$  for  $l = P, \dots, L-1$ . Then, the enlarged class of cells  $\{\Delta_l : 0 \leq l < L + K\}$  and the probability measure  $\mathcal{Q}_{\mathcal{H}}$  satisfy conditions (i) and (ii) in Theorem 3. It follows that  $E_{\mathcal{H}} = L^{-1/2}$  and hence, for  $X_n = \zeta_n$ , (3) holds with  $\varrho_n = \log(nL)/\sqrt{n/L}$ , assuming that  $n^{-1}L \log(nL) \rightarrow 0$ . 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  $\mathbb{P}_X[(\sqcup_{0 \leq l < P} \Delta_l)^c]$ .  $\blacktriangle$

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

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

- (iii)  $\mathcal{G}$  is a VC-type class with envelope function  $M_\mathcal{G}$  over  $\mathcal{Q}_\mathcal{G}$  with  $c_\mathcal{G} \geq e$  and  $d_\mathcal{G} \geq 1$ .
- (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  $c_{\mathcal{R}, \mathcal{Y}} \geq e$  and  $d_{\mathcal{R}, \mathcal{Y}} \geq 1$ , where  $M_{\mathcal{R}, \mathcal{Y}}(y) + \text{pTV}_{\mathcal{R}, (-|y|, |y|)} \leq v(1 + |y|^\alpha)$  for all  $y \in \mathcal{Y}$ , for some  $v > 0$ , 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  $k$  such that  $|\log_2 \mathbb{E}_\mathcal{G}| + |\log_2 M_\mathcal{G}| + |\log_2 L| \leq k \log_2 n$ .

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
- $\mathbb{P}[\|R_n - Z_n^R\|_{\mathcal{G} \times \mathcal{R}} > C_1 C_{v, \alpha} (C_\rho U_n(t) + V_n(t))] \leq C_2 e^{-t} + L e^{-C_\rho n/L}$  for all  $t > 0$ ,

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

$$U_n(t) = \left( \sqrt{\frac{dM_\mathcal{G} \mathbb{E}_\mathcal{G}}{n/L}} + \frac{M_\mathcal{G}}{\sqrt{n}} (\log n)^\alpha \right) (t + k \log_2 n + d \log(cn))^{\alpha+1}$$

with  $c = c_\mathcal{G} c_{\mathcal{R}, \mathcal{Y}}$ ,  $d = d_\mathcal{G} + d_{\mathcal{R}, \mathcal{Y}}$ , and

$$V_n(t) = \mathbb{1}(|\mathcal{R}| > 1) \sqrt{M_\mathcal{G} \mathbb{E}_\mathcal{G}} \left( \max_{0 \leq l < L} \|\Delta_l\|_\infty \right) L_{\mathcal{V}_\mathcal{R}} \sqrt{t + k \log_2 n + d \log(cn)},$$

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,  $U_n(t)$ , can be interpreted as a ‘‘variance’’ contribution based on the effective sample size  $n/L$ , up to polylog( $n$ ) terms, while the second term,  $V_n(t)$ , 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  $c$  and  $d$ . Then, if  $n^{-1}L \log(nL) \rightarrow 0$ ,  $\|R_n - Z_n^R\|_{\mathcal{G} \times \mathcal{R}} = O(\varrho_n)$  a.s. with

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

Setting aside  $M_{\mathcal{G}}$  and  $E_{\mathcal{G}}$ , an approximation rate is  $(\log n)^{2\alpha+1} (n/L)^{-1/2} + \mathbb{1}(|\mathcal{R}| > 1) (\max_{0 \leq l < L} \|\Delta_l\|_{\infty}) \sqrt{\log n}$ , which can achieve the optimal univariate KMT strong approximation rate based on the effective sample size  $n/L$ , up to a  $\text{polylog}(n)$  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  $(\mathbf{z}_i = (\mathbf{x}_i, y_i), 1 \leq i \leq n)$  are i.i.d. random vectors taking values in  $(\mathcal{X} \times \mathbb{R}, \mathcal{B}(\mathcal{X} \times \mathbb{R}))$  with  $\mathcal{X} \subseteq \mathbb{R}^d$ . 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:

$$\check{\theta}(\mathbf{w}, r) = \mathbf{p}(\mathbf{w})^{\top} \hat{\gamma}(r), \quad \hat{\gamma}(r) = \underset{\gamma \in \mathbb{R}^L}{\text{argmin}} \sum_{i=1}^n (r(y_i) - \mathbf{p}(\mathbf{x}_i)^{\top} \gamma)^2,$$

where  $\mathbf{p}(\mathbf{u}) = (\mathbb{1}(\mathbf{u} \in \Delta_l) : 0 \leq l < L)$ , 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  $\{\Delta_l : 0 \leq l < L\}$  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:

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

where  $\mathbf{Q} = \mathbb{E}[\mathbf{p}(\mathbf{x}_i) \mathbf{p}(\mathbf{x}_i)^{\top}]$ ,  $\hat{\mathbf{Q}} = \frac{1}{n} \sum_{i=1}^n \mathbf{p}(\mathbf{x}_i) \mathbf{p}(\mathbf{x}_i)^{\top}$ , 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  $k_{\mathbf{w}}(\mathbf{u}) = L^{-1/2} \sum_{0 \leq l < L} \mathbb{1}(\mathbf{w} \in \Delta_l) \mathbb{1}(\mathbf{u} \in \Delta_l) / \mathbb{P}_X(\Delta_l)$  the equivalent kernel. Under standard regularity conditions including smoothness and moment assumptions [11, Section SA-V.3],

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

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

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

provided that  $\log(nL)L/n \rightarrow 0$ . 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,  $M_{\mathcal{G}} = L^{1/2}$  and  $E_{\mathcal{G}} = L^{-1/2}$ , and we can take  $c_{\mathcal{G}} = L$  and  $d_{\mathcal{G}} = 1$  because  $\mathcal{G}$  has finite cardinality  $L$ . For the singleton case  $\mathcal{R}_1$ , we

can take  $c_{\mathcal{R}_1} = 1$  and  $d_{\mathcal{R}_1} = 1$ ,  $\alpha = 1$  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  $\log(nL)L/n \rightarrow 0$ . For the VC-Type class  $\mathcal{R}_2$ , we can verify condition (v) in Theorem 4 with  $\alpha = 0$ , and we can take  $c_{\mathcal{R}_2}$  to be some universal constant and  $d_{\mathcal{R}_2} = 2$  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  $\log(n)L/n \rightarrow 0$ . 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 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  $RV^1$ -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.



# Supplemental Appendix to “Strong Approximations for Empirical Processes Indexed by Lipschitz Functions”\*

Matias D. Cattaneo<sup>†</sup>      Ruiqi (Rae) Yu<sup>‡</sup>

November 14, 2024

## Abstract

This supplement appendix reports additional theoretical results not discussed in the paper to conserve space, and provides all the technical proofs. Section [SA-I](#) introduces additional notation and definitions used in the proofs. Section [SA-II](#) studies the general empirical process (Section 3 in the paper). Section [SA-III](#) studies the multiplicative-separable empirical process (not discussed in the paper but of independent interest). Section [SA-IV](#) studies the residual-based empirical process (Section 4 in the paper). Section [SA-V](#) studies the three empirical processes in the context of quasi-uniform Haar basis (Section 5 in the paper).

---

\*Corresponding author: [rae.yu@princeton.edu](mailto:rae.yu@princeton.edu). Support from the National Science Foundation through grants DMS-2210561 and SES-2241575 is gratefully acknowledged.

<sup>†</sup>Department of Operations Research and Financial Engineering, Princeton University.

<sup>‡</sup>Department of Operations Research and Financial Engineering, Princeton University.

# Contents

<b>SA-I</b>	<b>Additional Notation</b>	<b>2</b>
SA-I.1	Additional Main Definitions . . . . .	2
<b>SA-II</b>	<b>General Empirical Process</b>	<b>3</b>
SA-II.1	Preliminary Technical Results . . . . .	5
SA-II.1.1	Cells Expansions . . . . .	5
SA-II.1.2	Projection onto Piecewise Constant Functions . . . . .	6
SA-II.1.3	Strong Approximation Constructions . . . . .	8
SA-II.1.4	Meshing Error . . . . .	17
SA-II.1.5	Strong Approximation Errors . . . . .	18
SA-II.1.6	Projection Error . . . . .	21
SA-II.2	Surrogate Measure and Normalizing Transformation . . . . .	22
SA-II.3	General Result: Proof of Theorem 1 . . . . .	27
SA-II.4	Additional Results . . . . .	31
SA-II.5	Proofs of Corollaries 1, 2, and 3 . . . . .	32
SA-II.6	Example 1: Kernel Density Estimation . . . . .	32
SA-II.6.1	Surrogate Measure and Normalizing Transformation . . . . .	33
SA-II.6.2	Class $\mathcal{H}$ and Its Corresponding Constants . . . . .	33
<b>SA-III</b>	<b>Multiplicative-Separable Empirical Process</b>	<b>34</b>
SA-III.1	Preliminary Technical Results . . . . .	35
SA-III.1.1	Cells Expansions . . . . .	36
SA-III.1.2	Projection onto Piecewise Constant Functions . . . . .	37
SA-III.1.3	Strong Approximation Construction . . . . .	38
SA-III.1.4	Meshing Error . . . . .	42
SA-III.1.5	Strong Approximation Errors . . . . .	43
SA-III.1.6	Projection Error . . . . .	46
SA-III.2	General Result . . . . .	51
SA-III.3	Additional Results . . . . .	56
<b>SA-IV</b>	<b>Residual-Based Empirical Processes</b>	<b>57</b>
SA-IV.1	Preliminary Technical Results . . . . .	58
SA-IV.1.1	Projection onto Piecewise Constant Functions . . . . .	58
SA-IV.1.2	Strong Approximation Constructions . . . . .	59
SA-IV.1.3	Meshing Error . . . . .	61
SA-IV.1.4	Strong Approximation Errors . . . . .	62
SA-IV.1.5	Projection Error . . . . .	64
SA-IV.2	General Result . . . . .	65
SA-IV.3	Proof of Theorem 2 . . . . .	68
SA-IV.4	Proof of Corollary 4 . . . . .	68
SA-IV.5	Example: Local Polynomial Estimators . . . . .	68
<b>SA-V</b>	<b>Quasi-Uniform Haar Basis</b>	<b>76</b>
SA-V.1	General Empirical Process . . . . .	76
SA-V.1.1	Proof of Theorem 3 . . . . .	76
SA-V.1.2	Proof of Corollary 5 . . . . .	78
SA-V.1.3	Example 2: Histogram Density Estimation . . . . .	78
SA-V.2	Residual-Based (and Multiplicative Separable) Empirical Process . . . . .	79
SA-V.2.1	Proof of Theorem 4 . . . . .	84
SA-V.2.2	Proof of Corollary 6 . . . . .	85
SA-V.3	Example: Haar Partitioning-based Regression . . . . .	85

## SA-I Additional Notation

We introduce additional notation and definitions complementing those given in Section 2 of the paper. See [Ambrosio \*et al.\* \(2000\)](#), [van der Vaart and Wellner \(2013\)](#), [Giné and Nickl \(2016\)](#), and references therein, for background definitions and more details.

Let  $\mathcal{U}, \mathcal{V} \subseteq \mathbb{R}^q$ . We define  $\mathcal{U} - \mathcal{V} = \{\mathbf{u} - \mathbf{v} : \mathbf{u} \in \mathcal{U}, \mathbf{v} \in \mathcal{V}\}$ . We define  $\mathcal{U} \Delta \mathcal{V} = (\mathcal{U} \setminus \mathcal{V}) \cup (\mathcal{V} \setminus \mathcal{U})$ . Let  $\det(\mathbf{A})$  be the determinant of the matrix  $\mathbf{A}$ . Let  $\Phi(z)$  be the distribution function of  $\text{Normal}(0, 1)$ , and  $\text{Bern}(p)$  denote the Bernoulli distribution with parameter  $p \in (0, 1)$ . For a real-valued random variable  $X$ , the  $L_p$ -norm is defined as  $\|X\|_p = \mathbb{E}[|X|^p]^{1/p}$  for  $1 \leq p < \infty$ . The  $\sigma$ -algebra generated by  $X$  is denoted by  $\sigma(X)$ . For  $\alpha > 0$ , the  $\psi_\alpha$ -norm of  $X$  is given by  $\|X\|_{\psi_\alpha} = \min\{\lambda > 0 : \mathbb{E}[\exp((\frac{|X|}{\lambda})^\alpha)] \leq 2\}$ . For  $\mathbf{x} \in \mathbb{R}^q$  and  $r > 0$ , let  $B(\mathbf{x}, r)$  denote the Euclidean ball with radius  $r$  centered at  $\mathbf{x}$ . For a matrix  $\mathbf{A} \in \mathbb{R}^{q \times q}$ ,  $\|\mathbf{A}\|$  denotes its operator norm. Using standard empirical process notation,  $\mathbb{E}_n[f(\mathbf{x}_i)]$  denotes the empirical average  $n^{-1} \sum_{i=1}^n [f(\mathbf{x}_i) - \mathbb{E}[f(\mathbf{x}_i)]]$  based on random sample  $(\mathbf{x}_i : 1 \leq i \leq n)$ . For sequences of real numbers, we write  $a_n = \Omega(b_n)$  if there exists some constant  $C$  and  $N > 0$  such that  $n > N$  implies  $|a_n| \geq C|b_n|$ .

Let  $\mathcal{S} \subseteq \mathbb{R}^q$  and  $Q$  be a measure on  $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ . The semi-metric  $\mathfrak{d}_Q$  on  $L_2(Q)$  is defined by  $\mathfrak{d}_Q(f, g) = (\|f - g\|_{Q,2}^2 - (\int f dQ - \int g dQ)^2)^{1/2}$ , for  $f, g \in L_2(Q)$ . For a class  $\mathcal{F} \subseteq L_2(Q)$ , let  $C(\mathcal{F}, \mathfrak{d}_Q)$  denote the class of all continuous functionals on the space  $(\mathcal{F}, \mathfrak{d}_Q)$ . For  $\alpha > 0$ , the  $C^\alpha$ -norm of a real-valued measurable function  $f$  on  $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$  is given by  $\|f\|_{C^\alpha} = \max_{|k| \leq \lfloor \alpha \rfloor} \sup_{\mathbf{x} \in \mathcal{S}} |D^k f(\mathbf{x})| + \max_{|k| = \alpha} \sup_{\mathbf{x} \neq \mathbf{y} \in \mathcal{S}} \frac{|D^k f(\mathbf{x}) - D^k f(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|_2^{\alpha - \lfloor \alpha \rfloor}}$ . The space  $C^\alpha(\mathcal{S})$  denotes the collection of all real-valued measurable functions on  $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$  with  $C^\alpha$ -norm bounded by 1. For real-valued functions  $f, g$  on  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ , the convolution of  $f$  and  $g$  is the function  $f * g$  such that  $f * g(x) = \int_{-\infty}^{\infty} f(y)g(x - y)dy, x \in \mathbb{R}$ . If  $\mathcal{F}$  and  $\mathcal{G}$  are two sets of functions from measure space  $(\mathcal{U}, \mathcal{B}(\mathcal{U}))$  to  $\mathbb{R}$  and  $(\mathcal{V}, \mathcal{B}(\mathcal{V}))$  to  $\mathbb{R}$ , respectively, then  $\mathcal{F} \cdot \mathcal{G}$  denotes the class of measurable functions  $\{f \cdot g : f \in \mathcal{F}, g \in \mathcal{G}\}$  from  $(\mathcal{U} \times \mathcal{V}, \mathcal{B}(\mathcal{U}) \otimes \mathcal{B}(\mathcal{V}))$  to  $\mathbb{R}$ . For a semi-metric space  $(\mathcal{F}, \mathfrak{d})$  of real-valued measurable functions on  $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ ,  $N_{[]}(\varepsilon, \mathcal{F}, \mathfrak{d})$  denotes the bracketing number.

For a probability measure  $P$  on  $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ , a  $P$ -Brownian bridge is a centered Gaussian random function  $(W_P(f) : f \in L_2(P))$  with covariance given by  $\mathbb{E}[W_P(f)W_P(g)] = P(fg) - P(f)P(g)$  for  $f, g \in L_2(P)$ . A class  $\mathcal{F} \subseteq L_2(P)$  is said to be  $P$ -pregaussian if there exists a version of the  $P$ -Brownian bridge  $W_P$  such that  $W_P \in C(\mathcal{F}; \mathfrak{d}_P)$  almost surely.

Finally, we use  $a_n \lesssim b_n$  to denote that  $a_n = O(b_n)$  with only a universal constant, not a function of the data generating process or related parameters. For  $K \in \mathbb{N}$ , we repeatedly employ the index sets  $\mathcal{I}_K = \{(j, k) \in \mathbb{N} \times \mathbb{N} : 1 \leq j \leq K, 0 \leq k < 2^{K-j}\}$  and  $\mathcal{J}_K = \{(j, k) \in \mathbb{N} \times \mathbb{N} : 0 \leq j \leq K, 0 \leq k < 2^{K-j}\}$ .

### SA-I.1 Additional 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 additional definitions that capture properties of  $\mathcal{F}$ , complementing those in Section 2.1.

**Definition SA.1.** For a non-empty  $\mathcal{C} \subseteq \mathbb{R}^q$ , the smoothed uniform total variation of  $\mathcal{F}$  over  $\mathcal{C}$  is

$$\text{TV}_{\mathcal{F}, \mathcal{C}}^* = \sup_{f \in \mathcal{F}} \inf_{(f_\ell)_{\ell \in \mathbb{N}}} \limsup_{\ell \rightarrow \infty} \text{TV}_{\{f_\ell\}, \mathcal{C}},$$

where the infimum is taken over all sequences of functions  $(f_\ell)_{\ell \in \mathbb{N}}$  such that  $f_\ell \rightarrow f \in \mathcal{F}$  a.s.-m on  $(\mathcal{C}, \mathcal{B}(\mathcal{C}))$ , and  $f_\ell$  is differentiable and bounded by  $2M_{\mathcal{F}, \mathcal{C}}$  on  $\mathcal{C}$  for all  $\ell \geq 1$ .

**Definition SA.2.** For a non-empty  $\mathcal{C} \subseteq \mathbb{R}^q$ , the smoothed uniform local total variation of  $\mathcal{F}$  over  $\mathcal{C}$  is a positive number  $K_{\mathcal{F},\mathcal{C}}^*$  such that for any cube  $\mathcal{D} \subseteq \mathbb{R}^q$  with edges of length  $\ell$  parallel to the coordinate axes,

$$\text{TV}_{\mathcal{F},\mathcal{D} \cap \mathcal{C}}^* \leq K_{\mathcal{F},\mathcal{C}}^* \ell^{d-1}.$$

Suppose  $\mathcal{S}$  is also a class of measurable functions from the probability space  $(\mathbb{R}^q, \mathcal{B}(\mathbb{R}^q), \mathbb{P})$  to  $\mathbb{R}$ . We generalize the definition of the uniform covering number to  $\mathcal{F} \times \mathcal{S}$ .

**Definition SA.3.** For a non-empty  $\mathcal{C} \subseteq \mathbb{R}^q$ , the uniform covering number of  $\mathcal{F} \times \mathcal{S}$  with envelope  $M_{\mathcal{F},\mathcal{C}} M_{\mathcal{S},\mathcal{C}}$  over  $\mathcal{C}$  is

$$N_{\mathcal{F} \times \mathcal{S},\mathcal{C}}(\delta, M_{\mathcal{F},\mathcal{C}} M_{\mathcal{S},\mathcal{C}}) = \sup_{\mu} N(\mathcal{F} \times \mathcal{S}, \lambda_{\mu}, \delta \|M_{\mathcal{F},\mathcal{C}} M_{\mathcal{S},\mathcal{C}}\|_{\mu,2}), \quad \delta \in (0, \infty),$$

where the supremum is taken over all finite discrete measures on  $(\mathcal{C}, \mathcal{B}(\mathcal{C}))$ , and  $\lambda_{\mu}$  is the semi-metric on  $\mathcal{F} \times \mathcal{S}$  defined by

$$\lambda_{\mu}((f_1, g_1), (f_2, g_2))^2 = \int_{\mathcal{C}} (f_1(\mathbf{x})g_1(\mathbf{x}) - f_2(\mathbf{x})g_2(\mathbf{x}))^2 d\mu(\mathbf{x}).$$

We assume that  $M_{\mathcal{F},\mathcal{C}}(\mathbf{u})$  and  $M_{\mathcal{S},\mathcal{C}}(\mathbf{u})$  are finite for every  $\mathbf{u} \in \mathcal{C}$ .

**Definition SA.4.** For a non-empty  $\mathcal{C} \subseteq \mathbb{R}^q$ , the uniform entropy integral of  $\mathcal{F} \times \mathcal{S}$  with envelope  $M_{\mathcal{F},\mathcal{C}} M_{\mathcal{S},\mathcal{C}}$  over  $\mathcal{C}$  is

$$J_{\mathcal{C}}(\delta, \mathcal{F} \times \mathcal{S}, M_{\mathcal{F},\mathcal{C}} M_{\mathcal{S},\mathcal{C}}) = \int_0^{\delta} \sqrt{1 + \log N_{\mathcal{F} \times \mathcal{S},\mathcal{C}}(\varepsilon, M_{\mathcal{F},\mathcal{C}} M_{\mathcal{S},\mathcal{C}})} d\varepsilon,$$

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

## SA-II General Empirical Process

Recall that  $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ ,  $i = 1, \dots, n$ , are i.i.d. random vectors supported on a background probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , and the general empirical process is

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

where  $\mathcal{H}$  is a possibly  $n$ -varying class of functions. As briefly explained after Theorem 1 is presented in the paper, its proof relies on the following decomposition:

$$\begin{aligned} & \|X_n - Z_n^X\|_{\mathcal{H}} \\ & \leq \|X_n - X_n \circ \pi_{\mathcal{H}_{\delta}}\|_{\mathcal{H}} + \|X_n - Z_n^X\|_{\mathcal{H}_{\delta}} + \|Z_n^X \circ \pi_{\mathcal{H}_{\delta}} - Z_n^X\|_{\mathcal{H}} \\ & \leq \|X_n - X_n \circ \pi_{\mathcal{H}_{\delta}}\|_{\mathcal{H}} + \|X_n - \Pi_0 X_n\|_{\mathcal{H}_{\delta}} + \|\Pi_0 X_n - \Pi_0 Z_n^X\|_{\mathcal{H}_{\delta}} + \|\Pi_0 Z_n^X - Z_n^X\|_{\mathcal{H}_{\delta}} + \|Z_n^X \circ \pi_{\mathcal{H}_{\delta}} - Z_n^X\|_{\mathcal{H}}, \end{aligned}$$

where  $\mathcal{H}_{\delta}$  denotes a discretization (or meshing) of  $\mathcal{H}$  (i.e.,  $\delta$ -net of  $\mathcal{H}$ ), and 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) of  $X_n$  and  $Z_n^X$  relative to the meshing for each of the stochastic processes. These terms are handled using standard arguments for empirical processes. Then,

following [Rio \(1994\)](#), the term  $\|X_n - Z_n^X\|_{\mathcal{H}_\delta}$  is further decomposed into three terms:  $\|\Pi_0 X_n - \Pi_0 Z_n^X\|_{\mathcal{H}_\delta}$  and  $\|\Pi_0 Z_n^X - Z_n^X\|_{\mathcal{H}_\delta}$  represent a mean square projection onto a Haar function space, where  $\Pi_0 X_n(h) = X_n \circ \Pi_0 h$  with  $\Pi_0$  the  $L_2$  projection onto piecewise constant functions on a carefully chosen partition of  $\mathcal{X}$ , while the final term  $\|\Pi_0 X_n - \Pi_0 Z_n^X\|_{\mathcal{H}_\delta}$  captures the coupling between the projected empirical process and the projected Gaussian process (on a  $\delta$ -net of  $\mathcal{H}$ , after the  $L_2$  projection).

The proof of Theorem 1 first constructs the Gaussian process  $(Z_n^X(h) : h \in \mathcal{H})$  on a possibly enlarged probability space supporting the empirical process  $(X_n(h) : h \in \mathcal{H})$ , and then bounds each of the five error terms described above. The proof is given in Section [SA-II.3](#), and it exploits the existence of a surrogate measure and normalizing transformation (Section [SA-II.2](#)), along with a collection preliminary technical results (Section [SA-II.1](#)) that may be of independent interest. More specifically, our preliminary technical results are organized as follows:

- Section [SA-II.1.1](#) introduces a class of recursive quasi-dyadic cells expansion of  $\mathcal{X}$ , which we employ to generalize prior dyadic cell results in the literature.
- Section [SA-II.1.2](#) introduces the  $L_2$  projection onto piecewise constant functions, which can be written as a linear combination of the Haar basis based on the cells. As a consequence, the empirical process indexed by  $L_2$ -projected functions can be written as linear combinations of counts of i.i.d. data.
- Section [SA-II.1.3](#) constructs the Gaussian process  $(Z_n^X(h) : h \in \mathcal{H})$ . Since the constant approximation within each recursive partitioning cell generates counts based on i.i.d. data, the construction boils down to coupling binomial random variables with Gaussian random variables. The celebrated Tusnády’s inequality couples  $\text{Bin}(n, \frac{1}{2})$  with  $\text{Normal}(\frac{n}{2}, \frac{n}{4})$ , and gives an almost sure bound on the coupling error. In particular, the Gaussian random variable is given by a quantile transformation of the binomial random variable. Building on the quantile transformation idea, our Lemma [SA.4](#) studies the coupling between  $\text{Bin}(n, p)$  and  $\text{Normal}(np, np(1-p))$ , with the error bound given on a high probability set. Due to the dyadic correlation structure, a conditional quantile transformation is used to generate the Binomial–Gaussian pairs down the dyadic cells. Since the constructed Gaussian random variables have a joint distribution that coincides with the Brownian bridge integrated on cells, the Skorohod embedding lemma ([Dudley, 2014](#), Lemma 3.35) is then used to construct the Brownian bridge  $(Z_n^X(h) : h \in \mathcal{H})$  on a possibly enriched probability space supporting the data distribution.
- Section [SA-II.1.4](#) handles the meshing errors  $\|X_n - X_n \circ \pi_{\mathcal{H}_\delta}\|_{\mathcal{H}}$  and  $\|Z_n^X \circ \pi_{\mathcal{H}_\delta} - Z_n^X\|_{\mathcal{H}}$  using standard empirical process results, which give the contribution  $F(\delta)$  emerging from Talagrand’s inequality ([Giné and Nickl, 2016](#), Theorem 3.3.9) combined with a standard maximal inequality ([Chernozhukov et al., 2014](#), Theorem 5.2). This allows us to focus on the error on the  $\delta$ -net to study  $\|X_n - Z_n^X\|_{\mathcal{H}_\delta}$ .
- Section [SA-II.1.5](#) handles the strong approximation error  $\|\Pi_0 X_n - \Pi_0 Z_n^X\|_{\mathcal{H}_\delta}$ . Building on the Tusnády’s Lemma, [Rio \(1994, Theorem 2.1\)](#) established a remarkable coupling result for bounded functions  $L_2$ -projected on a dyadic cells expansion of  $\mathcal{X}$ . Our Lemma [SA.7](#) builds on his powerful ideas, and establishes an analogous result for the case of Lipschitz functions  $L_2$ -projected on dyadic cells expansions of  $\mathcal{X}$ , thereby obtaining a tighter coupling error. A limitation of these results is that they only apply to a dyadic cell expansion due to the specifics of Tusnády’s Lemma. Leveraging the coupling between  $\text{Bin}(n, p)$  and  $\text{Normal}(np, np(1-p))$ , our Lemma [SA.8](#) established a coupling result for bounded functions  $L_2$ -projected on a quasi-dyadic cells, although the result is restricted to a high probability event.

- Section SA-II.1.6 handles the  $L_2$ -projection errors  $\|X_n - \Pi_0 X_n\|_{\mathcal{H}_\delta}$  and  $\|\Pi_0 Z_n^X - Z_n^X\|_{\mathcal{H}_\delta}$  using Bernstein inequality, and taking into account explicitly the potential Lipschitz structure of the functions as well as the generic cell structure.

Section SA-II.2 introduces a reduction argument via the surrogate measure and the normalizing transformation in order to apply the preliminary technical results from Section SA-II.1 to prove Theorem 1. Specifically, the surrogate measure and normalizing transformation reduce the problem to the case where  $\mathbf{x}_i \sim \text{Uniform}([0, 1]^d)$ . Section SA-II.3 gives the proof of Theorem 1. Section SA-II.4 presents additional results of independent interest, which are used in Section SA-II.5 to prove the results discussed in Section 3.2 of the paper. Finally, Section SA-II.6 provides technical details underlying Example 1 in the paper.

## SA-II.1 Preliminary Technical Results

This section presents preliminary technical results that are used to prove Theorem 1. Whenever possible, these results are presented at a higher level of generality, and therefore may be of independent theoretical interest. Throughout this section, we employ the following assumption.

**Assumption SA.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)  $M_{\mathcal{H}, \mathcal{X}} < \infty$  and  $J_{\mathcal{X}}(1, \mathcal{H}, M_{\mathcal{H}, \mathcal{X}}) < \infty$ .

Compared to the assumptions in Theorem 1, this assumption does not require the existence of a surrogate measure and normalizing transformation. It will be applied in the analysis of each term in the error decomposition, where we work with the  $\mathbb{P}_X$  distribution. Section SA-II.2 illustrates how the normalizing transformation enables the use of the surrogate measure  $\mathbb{Q}_{\mathcal{H}}$ , providing greater flexibility in the data generating process. This reduction through the normalizing transformation is a crucial step in the proof of Theorem 1 (Section SA-II.3).

### SA-II.1.1 Cells Expansions

We introduce two definitions of quasi-dyadic cells expansions. Recall that  $\mathcal{I}_K = \{(j, k) \in \mathbb{N} \times \mathbb{N} : 1 \leq j \leq K, 0 \leq k < 2^{K-j}\}$  and  $\mathcal{J}_K = \{(j, k) \in \mathbb{N} \times \mathbb{N} : 0 \leq j \leq K, 0 \leq k < 2^{K-j}\}$ .

**Definition SA.5** (Quasi-Dyadic Expansion). *A collection of Borel measurable sets in  $\mathbb{R}^d$ ,  $\mathcal{C}_K(\mathbb{P}, \rho) = \{\mathcal{C}_{j,k} : (j, k) \in \mathcal{J}_K\}$ , is called a quasi-dyadic expansion of depth  $K$  with respect to probability measure  $\mathbb{P}$  if the following three conditions hold:*

- (i)  $\mathbb{P}(\mathcal{C}_{K,0}) = 1$ .
- (ii)  $\mathcal{C}_{j,k} = \mathcal{C}_{j-1,2k} \sqcup \mathcal{C}_{j-1,2k+1}$ , for all  $(j, k) \in \mathcal{J}_K$ .
- (iii)  $\max_{0 \leq k < 2^K} \mathbb{P}(\mathcal{C}_{0,k}) / \min_{0 \leq k < 2^K} \mathbb{P}(\mathcal{C}_{0,k}) \leq \rho < \infty$ .

When  $\rho = 1$ ,  $\mathcal{C}_K(\mathbb{P}, 1)$  is called a dyadic expansion of depth  $K$  with respect to  $\mathbb{P}$ .



This definition implies  $\frac{1}{2} \frac{2}{1+\rho} \leq \mathbb{P}(\mathcal{C}_{j-1,2k})/\mathbb{P}(\mathcal{C}_{j,k}) \leq \frac{1}{2} \frac{2\rho}{1+\rho}$  for all  $(j,k) \in \mathcal{I}_K$ , since each  $\mathcal{C}_{j-1,l}$  is a disjoint union of  $2^{j-1}$  cells of the form  $\mathcal{C}_{0,k}$ , which implies the third condition in Definition SA.5. Furthermore,  $\mathbb{P}(\mathcal{C}_{j-1,2k}) = \mathbb{P}(\mathcal{C}_{j-1,2k+1}) = \frac{1}{2} \mathbb{P}(\mathcal{C}_{j,k})$  in the special case  $\rho = 1$ , that is, the child level cells are obtained by splitting the parent level cells dyadically in probability.

The next definition specializes the dyadic expansion scheme to axis-aligned splits.

**Definition SA.6** (Axis-Aligned Quasi-Dyadic Expansion). *A collection of Borel measurable sets in  $\mathbb{R}^d$ ,  $\mathcal{A}_K(\mathbb{P}, \rho) = \{\mathcal{C}_{j,k} : (j,k) \in \mathcal{I}_K\}$ , is an axis-aligned quasi-dyadic expansion of depth  $K$  with respect to probability measure  $\mathbb{P}$  if it can be constructed via the following procedure:*

- (i) Initialization ( $q = 0$ ): Take  $\mathcal{C}_{K-q,0} = \text{Supp}(\mathbb{P})$ .
- (ii) Iteration ( $q = 1, \dots, K$ ): Given  $\mathcal{C}_{K-l,k}$  for  $0 \leq l \leq q-1, 0 \leq k < 2^l$ , take  $s = (q \bmod d) + 1$ , and construct  $\mathcal{C}_{K-q,2k} = \mathcal{C}_{K-q+1,k} \cap \{\mathbf{x} \in \mathbb{R}^d : \mathbf{e}_s^\top \mathbf{x} \leq c_{K-q+1,k}\}$  and  $\mathcal{C}_{K-q,2k+1} = \mathcal{C}_{K-q+1,k} \cap \{\mathbf{x} \in \mathbb{R}^d : \mathbf{e}_s^\top \mathbf{x} > c_{K-q+1,k}\}$  where  $c_{K-q+1,k}$  is a number chosen so that  $\mathbb{P}(\mathcal{C}_{K-q,2k})/\mathbb{P}(\mathcal{C}_{K-q+1,k}) \in [\frac{1}{1+\rho}, \frac{\rho}{1+\rho}]$  for all  $0 \leq k < 2^{q-1}$ . Continue until the collection  $(\mathcal{C}_{0,k} : 0 \leq k < 2^K)$  has been constructed.

If  $\rho = 1$  and  $\mathbb{P}$  is continuous, then  $\mathcal{A}_K(\mathbb{P}, \rho)$  is unique.

### SA-II.1.2 Projection onto Piecewise Constant Functions

For a quasi-dyadic expansion  $\mathcal{C}_K(\mathbb{P}, \rho)$ , the span of the Haar basis based on the terminal cells is

$$\mathcal{E}_K = \text{Span}\{\mathbb{1}_{\mathcal{C}_{0,k}} : 0 \leq k < 2^K\}.$$

For  $h \in L_2(\mathbb{P})$ , the mean square projection of  $h$  onto  $\mathcal{E}_K$  is

$$\Pi_0(\mathcal{C}_K(\mathbb{P}, \rho))[h] = \sum_{0 \leq k < 2^K} \frac{\mathbb{1}_{\mathcal{C}_{0,k}}}{\mathbb{P}(\mathcal{C}_{0,k})} \int_{\mathcal{C}_{0,k}} h(\mathbf{u}) d\mathbb{P}(\mathbf{u}).$$

Because  $\Pi_0(\mathcal{C}_K(\mathbb{P}, \rho))[h]$  is a linear combination of Haar functions, we obtain the following orthogonal decomposition.

**Lemma SA.1.** *For a quasi-dyadic expansion  $\mathcal{C}_K(\mathbb{P}, \rho)$  and any  $h \in L_2(\mathbb{P})$ , the mean square projection  $\Pi_0(\mathcal{C}_K(\mathbb{P}, \rho))[h]$  satisfies*

$$\Pi_0(\mathcal{C}_K(\mathbb{P}, \rho))[h] = \beta_{K,0}(h)e_{K,0} + \sum_{1 \leq j \leq K} \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(h)\tilde{e}_{j,k},$$

where

$$\beta_{j,k}(h) = \frac{1}{\mathbb{P}(\mathcal{C}_{j,k})} \int_{\mathcal{C}_{j,k}} h(\mathbf{u}) d\mathbb{P}(\mathbf{u}), \quad \tilde{\beta}_{j,k}(h) = \beta_{j-1,2k}(h) - \beta_{j-1,2k+1}(h),$$

$$e_{j,k} = \mathbb{1}_{\mathcal{C}_{j,k}}, \quad \tilde{e}_{j,k} = \frac{\mathbb{P}(\mathcal{C}_{j-1,2k+1})}{\mathbb{P}(\mathcal{C}_{j,k})} e_{j-1,2k} - \frac{\mathbb{P}(\mathcal{C}_{j-1,2k})}{\mathbb{P}(\mathcal{C}_{j,k})} e_{j-1,2k+1},$$

for all  $(j,k) \in \mathcal{I}_K = \{(j,k) \in \mathbb{N} \times \mathbb{N} : 1 \leq j \leq K, 0 \leq k < 2^{K-j}\}$ .

**Proof of Lemma SA.1.** First, we show that  $\{e_{K,0}\} \cup \{\tilde{e}_{j,k} : (j,k) \in \mathcal{I}_K\}$  is an orthogonal basis. For each  $(j,k) \in \mathcal{I}_K$ ,

$$\begin{aligned} \langle e_{K,0}, \tilde{e}_{j,k} \rangle &= \int_{\mathbb{R}^d} \frac{\mathbb{P}(\mathcal{C}_{j-1,2k+1})}{\mathbb{P}(\mathcal{C}_{j,k})} e_{j-1,2k}(\mathbf{u}) d\mathbb{P}(\mathbf{u}) - \int_{\mathbb{R}^d} \frac{\mathbb{P}(\mathcal{C}_{j-1,2k})}{\mathbb{P}(\mathcal{C}_{j,k})} e_{j-1,2k+1}(\mathbf{u}) d\mathbb{P}(\mathbf{u}) \\ &= \frac{\mathbb{P}(\mathcal{C}_{j-1,2k+1})\mathbb{P}(\mathcal{C}_{j-1,2k})}{\mathbb{P}(\mathcal{C}_{j,k})} - \frac{\mathbb{P}(\mathcal{C}_{j-1,2k})\mathbb{P}(\mathcal{C}_{j-1,2k+1})}{\mathbb{P}(\mathcal{C}_{j,k})} = 0, \end{aligned}$$

where  $\langle \cdot, \cdot \rangle$  denotes the inner product on  $L_2(\mathbb{P})$  given by  $\langle f, g \rangle = \int_{\mathbb{R}^d} f(\mathbf{u})g(\mathbf{u})d\mathbb{P}(\mathbf{u})$ ,  $f, g \in L_2(\mathbb{P})$ . Let  $(j_1, k_1), (j_2, k_2) \in \mathcal{I}_K$  such that  $(j_1, k_1) \neq (j_2, k_2)$ . We show  $\langle e_{j_1, k_1}, e_{j_2, k_2} \rangle = 0$  by considering two cases.

- *Case 1:*  $j_1 = j_2$  and  $k_1 \neq k_2$ , then  $\tilde{e}_{j_1, k_1}$  and  $\tilde{e}_{j_2, k_2}$  have different support, hence  $\langle \tilde{e}_{j_1, k_1}, \tilde{e}_{j_2, k_2} \rangle = 0$ .
- *Case 2:*  $j_1 \neq j_2$  and, without loss of generality, we assume  $j_1 < j_2$ . By (1) in Definition SA.5, either  $\mathcal{C}_{j_1, k_1} \cap \mathcal{C}_{j_2, k_2} = \emptyset$  or  $\mathcal{C}_{j_1, k_1} \subset \mathcal{C}_{j_2, k_2}$ .

In the first case, we also have  $\langle \tilde{e}_{j_1, k_1}, \tilde{e}_{j_2, k_2} \rangle = 0$ . In the second case, using (1) in Definition SA.5 again, either  $\mathcal{C}_{j_1, k_1} \subseteq \mathcal{C}_{j_2-1, 2k_2}$  or  $\mathcal{C}_{j_1, k_1} \subseteq \mathcal{C}_{j_2-1, 2k_2+1}$ . Assume, without loss of generality, that  $\mathcal{C}_{j_1, k_1} \subseteq \mathcal{C}_{j_2-1, 2k_2}$ . Then, for any  $(j_1, k_1), (j_2, k_2) \in \mathcal{I}_K$ ,

$$\begin{aligned} &\langle \tilde{e}_{j_1, k_1}, \tilde{e}_{j_2, k_2} \rangle \\ &= \langle \tilde{e}_{j_1, k_1}, \frac{\mathbb{P}(\mathcal{C}_{j_2-1, 2k_2})}{\mathbb{P}(\mathcal{C}_{j_2, k_2})} e_{j_2-1, 2k_2} \rangle \\ &= \frac{\mathbb{P}(\mathcal{C}_{j_2-1, 2k_2})}{\mathbb{P}(\mathcal{C}_{j_2, k_2})} \left[ \int_{\mathbb{R}^d} \frac{\mathbb{P}(\mathcal{C}_{j_1-1, 2k_1+1})}{\mathbb{P}(\mathcal{C}_{j_1, k_1})} e_{j_1-1, 2k_1}(\mathbf{u}) d\mathbb{P}(\mathbf{u}) - \int_{\mathbb{R}^d} \frac{\mathbb{P}(\mathcal{C}_{j_1-1, 2k_1})}{\mathbb{P}(\mathcal{C}_{j_1, k_1})} e_{j_1-1, 2k_1+1}(\mathbf{u}) d\mathbb{P}(\mathbf{u}) \right] \\ &= 0. \end{aligned}$$

Thus,  $\{e_{K,0}\} \cup \{\tilde{e}_{j,k} : (j,k) \in \mathcal{I}_K\}$  is an orthogonal basis for  $\mathcal{E}_K$ , and the  $L_2$  projection for all  $h \in L_2(\mathbb{P})$  is

$$\Pi_0(\mathcal{C}_K(\mathbb{P}, \rho))[h] = \frac{\langle h, e_{K,0} \rangle}{\langle e_{K,0}, e_{K,0} \rangle} e_{K,0} + \sum_{1 \leq j \leq K} \sum_{0 \leq k < 2^{K-j}} \frac{\langle h, \tilde{e}_{j,k} \rangle}{\langle \tilde{e}_{j,k}, \tilde{e}_{j,k} \rangle} \tilde{e}_{j,k}.$$

For all  $(j,k) \in \mathcal{I}_K$ , the coefficients are given by

$$\begin{aligned} \frac{\langle h, \tilde{e}_{j,k} \rangle}{\langle \tilde{e}_{j,k}, \tilde{e}_{j,k} \rangle} &= \frac{\int_{\mathbb{R}^d} h(\mathbf{u}) \tilde{e}_{j,k}(\mathbf{u}) d\mathbb{P}(\mathbf{u})}{\int_{\mathbb{R}^d} \tilde{e}_{j,k}(\mathbf{u}) \tilde{e}_{j,k}(\mathbf{u}) d\mathbb{P}(\mathbf{u})} \\ &= \frac{\mathbb{P}(\mathcal{C}_{j-1, 2k+1})\mathbb{P}(\mathcal{C}_{j-1, 2k})\mathbb{P}(\mathcal{C}_{j,k})^{-1}\beta_{j-1, 2k}(h) - \mathbb{P}(\mathcal{C}_{j-1, 2k})\mathbb{P}(\mathcal{C}_{j-1, 2k+1})\mathbb{P}(\mathcal{C}_{j,k})^{-1}\beta_{j-1, 2k+1}(h)}{\mathbb{P}(\mathcal{C}_{j-1, 2k+1})^2\mathbb{P}(\mathcal{C}_{j-1, 2k})\mathbb{P}(\mathcal{C}_{j,k})^{-2} + \mathbb{P}(\mathcal{C}_{j-1, 2k})^2\mathbb{P}(\mathcal{C}_{j-1, 2k+1})\mathbb{P}(\mathcal{C}_{j,k})^{-2}} \\ &= \frac{\mathbb{P}(\mathcal{C}_{j-1, 2k+1})\mathbb{P}(\mathcal{C}_{j-1, 2k})\mathbb{P}(\mathcal{C}_{j,k})^{-1}\beta_{j-1, 2k}(h) - \mathbb{P}(\mathcal{C}_{j-1, 2k})\mathbb{P}(\mathcal{C}_{j-1, 2k+1})\mathbb{P}(\mathcal{C}_{j,k})^{-1}\beta_{j-1, 2k+1}(h)}{\mathbb{P}(\mathcal{C}_{j-1, 2k+1})\mathbb{P}(\mathcal{C}_{j-1, 2k})\mathbb{P}(\mathcal{C}_{j,k})^{-1}} \\ &= \beta_{j-1, 2k}(h) - \beta_{j-1, 2k+1}(h) = \tilde{\beta}_{j,k}(h). \end{aligned}$$

Moreover,

$$\frac{\langle h, e_{K,0} \rangle}{\langle e_{K,0}, e_{K,0} \rangle} = \mathbb{P}(\mathcal{C}_{K,0})^{-1} \int_{\mathcal{C}_{K,0}} h(\mathbf{u}) d\mathbb{P}(\mathbf{u}) = \beta_{K,0}(h).$$

This concludes the proof.  $\square$

To save notation, we will write  $\Pi_0$  for  $\Pi_0(\mathcal{C}_K(\mathbb{P}, \rho))$  whenever the underlying cells expansion is clear from the context. For a class of functions  $\mathcal{H}$  on  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P})$  such that  $\mathcal{H} \subseteq L_2(\mathbb{P})$ , denote  $\Pi_0\mathcal{H} = \{\Pi_0 h : h \in \mathcal{H}\}$ .

### SA-II.1.3 Strong Approximation Constructions

This section employs the notations and conventions introduced in Sections SA-II.1.1 and SA-II.1.2. Unless explicitly stated otherwise, we assume a quasi-dyadic expansion  $\mathcal{C}_K(\mathbb{P}_X, \rho)$  is given. Let  $(\tilde{\xi}_{j,k} : (j,k) \in \mathcal{I}_K)$  be i.i.d. standard Gaussian random variables. Let  $F_{(j,k),m}$  be the cumulative distribution function of  $(S_{j,k} - mp_{j,k})/\sqrt{mp_{j,k}(1-p_{j,k})}$ , where  $S_{j,k}$  is a Bin( $m, p_{j,k}$ ) random variable with  $p_{j,k} = \mathbb{P}_X(\mathcal{C}_{j-1,2k})/\mathbb{P}_X(\mathcal{C}_{j,k})$ , and  $G_{(j,k),m}(t) = \inf\{x : F_{(j,k),m}(x) \geq t\}$ .

We define the collection of random variables  $(U_{j,k} : (j,k) \in \mathcal{J}_K)$  and  $(\tilde{U}_{j,k} : (j,k) \in \mathcal{I}_K)$  via the following iterative scheme:

1. *Initialization* ( $j = K$ ):  $U_{K,0} = n$ .
2. *Iteration* ( $j = K, K-1, \dots, 1$ ): For each  $1 \leq j \leq K$ , and given  $(U_{l,k} : j < l \leq K, 0 \leq k < 2^{K-l})$ , solve for  $(U_{j,k} : 0 \leq k < 2^{K-j})$  such that

$$\begin{aligned} \tilde{U}_{j,k} &= \sqrt{U_{j,k}p_{j,k}(1-p_{j,k})}G_{(j,k),U_{j,k}} \circ \Phi(\tilde{\xi}_{j,k}), \\ \tilde{U}_{j,k} &= (1-p_{j,k})U_{j-1,2k} - p_{j,k}U_{j-1,2k+1} = U_{j-1,2k} - p_{j,k}U_{j,k}, \\ U_{j-1,2k} + U_{j-1,2k+1} &= U_{j,k}, \end{aligned} \tag{SA-1}$$

where  $0 \leq k < 2^{K-j}$ . Continue till  $(U_{0,k} : 0 \leq k < 2^K)$  are defined.

Then,  $(U_{j,k} : (j,k) \in \mathcal{J}_K)$  has the same joint distribution as  $(\sum_{i=1}^n e_{j,k}(\mathbf{x}_i) : (j,k) \in \mathcal{J}_K)$  from Lemma SA.1. By the Vorob'ev–Berkes–Philipp theorem (Dudley, 2014, Theorem 1.31),  $(\tilde{\xi}_{j,k} : (j,k) \in \mathcal{I}_K)$  can be constructed on a possibly enlarged probability space such that the previously constructed  $U_{j,k}$  satisfies  $U_{j,k} = \sum_{i=1}^n e_{j,k}(\mathbf{x}_i)$  almost surely for all  $(j,k) \in \mathcal{J}_K$ . We will show that the  $\tilde{\xi}_{j,k}$ 's can be given as a Brownian bridge indexed by  $\tilde{e}_{j,k}$ 's from Lemma SA.1. Recall the definitions given in Section SA-II.1.2.

**Lemma SA.2.** *Suppose Assumption SA.1 holds, and a quasi-dyadic expansion  $\mathcal{C}_K(\mathbb{P}_X, \rho)$  is given. Then,  $\mathcal{H} \cup \Pi_0\mathcal{H} \subseteq L_2(\mathbb{P}_X)$  and is  $\mathbb{P}_X$ -pregaussian.*

**Proof of Lemma SA.2.** To simplify notation, the parameters of  $\mathcal{H}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{X}$ , and the index  $\mathcal{C}$  is omitted. Since  $M_{\mathcal{H}} < \infty$ ,  $\mathcal{H} \cup \Pi_0\mathcal{H} \subseteq L_2(\mathbb{P}_X)$ . Definition of  $\Pi_0$  from Section SA-II.1.2 implies that  $M_{\mathcal{H}} \cup \Pi_0\mathcal{H}$  is an envelope for  $\Pi_0\mathcal{H}$ .

Claim: For all  $0 < \delta < 1$ ,  $J(\Pi_0\mathcal{H}, M_{\mathcal{H}}, \delta) \leq J(\mathcal{H}, M_{\mathcal{H}}, \delta)$ .

Proof of Claim: Let  $Q$  be a finite discrete measure on  $\mathcal{X}$ . Let  $f, g \in \mathcal{H}$ . Then, by the definition of  $\Pi_0$  and Jensen's inequality,

$$\|\Pi_0 f - \Pi_0 g\|_{Q,2}^2 \leq \sum_{0 \leq k < 2^K} Q(\mathcal{C}_{0,k})2^K \int_{\mathcal{C}_{0,k}} (f - g)^2 d\mathbb{P}_X.$$

Define a measure  $\tilde{Q}$  such that for any  $A \in \mathcal{B}(\mathbb{R}^d)$ ,  $\tilde{Q}(A) = \sum_{0 \leq k < 2^K} Q(\mathcal{C}_{0,k})2^K \mathbb{P}_X(A \cap \mathcal{C}_{0,k})$ , then

$$\|\Pi_0 f - \Pi_0 g\|_{Q,2}^2 \leq \|f - g\|_{\tilde{Q},2}^2.$$

Take  $\mathcal{L}$  to be a  $\delta\mathbf{M}_{\mathcal{H}\mathcal{C}}$ -net of  $\mathcal{H}$  over  $\mathcal{X}$  with respect to  $\|\cdot\|_{\tilde{Q},2}$  with cardinality no greater than  $\mathbf{N}_{\mathcal{H}\mathcal{C}}(\delta, \mathbf{M}_{\mathcal{H}\mathcal{C}})$ . Let  $\Pi_0 f$  be in an arbitrary function in  $\Pi_0\mathcal{H}$ , there exists  $g \in \mathcal{L}$  such that  $\|\Pi_0 f - \Pi_0 g\|_{\tilde{Q},2}^2 \leq \|f - g\|_{\tilde{Q},2}^2 \leq \delta^2 \mathbf{M}_{\mathcal{H}\mathcal{C}}^2$ . The claim then follows.

It follows from the claim and (ii) from Assumption SA.1 that  $J(1, \mathcal{H} \cup \Pi_0\mathcal{H}, \mathbf{M}_{\mathcal{H}\mathcal{C}}) < \infty$ . By Dominated Convergence Theorem,  $\lim_{\delta \downarrow 0} J(\delta, \mathcal{H} \cup \Pi_0\mathcal{H}) < \infty$ . Since  $\mathbf{M}_{\mathcal{H}\mathcal{C}} < \infty$ ,  $\mathcal{H} \cup \Pi_0\mathcal{H}$  is totally bounded with respect to  $\|\cdot\|_{\mathbb{P}_X,2}$ . By separability of  $\mathcal{H}$  and van der Vaart and Wellner (2013, Corollary 2.2.9),  $\mathcal{H} \cup \Pi_0\mathcal{H}$  is  $\mathbb{P}_X$ -pregaussian.  $\square$

Under the conditions of Lemma SA.2, take  $(Z_n^X(h) : h \in \mathcal{H} \cup \Pi_0\mathcal{H})$  to be a  $\mathbb{P}_X$ -Brownian bridge such that  $Z_n^X(\cdot) \in C(\mathcal{H} \cup \Pi_0\mathcal{H}, \mathfrak{D}_{\mathbb{P}_X})$  almost surely. Since  $(Z_n^X(\tilde{e}_{j,k}) : (j,k) \in \mathcal{I}_K)$  are independent random variables with distribution  $\text{Normal}(0, \mathbb{P}_X(\mathcal{C}_{j-1,2k})\mathbb{P}_X(\mathcal{C}_{j-1,2k+1})\mathbb{P}_X(\mathcal{C}_{j,k})^{-1})$  for  $(j,k) \in \mathcal{I}_K$ , by Skorohod Embedding lemma (Dudley, 2014, Lemma 3.35), on a possibly enlarged probability space, the Brownian bridge  $(Z_n^X(h) : h \in \mathcal{H} \cup \Pi_0\mathcal{H})$  can be constructed such that it satisfies

$$\tilde{\xi}_{j,k} = \sqrt{\frac{\mathbb{P}_X(\mathcal{C}_{j,k})}{\mathbb{P}_X(\mathcal{C}_{j-1,2k})\mathbb{P}_X(\mathcal{C}_{j-1,2k+1})}} Z_n^X(\tilde{e}_{j,k}), \quad (\text{SA-2})$$

for all  $(j,k) \in \mathcal{I}_K$ . Moreover, for all  $g \in \Pi_0\mathcal{H}$ ,

$$\sqrt{n}X_n(g) = \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(g) \tilde{U}_{j,k} \quad \text{and} \quad \sqrt{n}Z_n^X(g) = \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(g) \tilde{V}_{j,k},$$

where  $\tilde{V}_{j,k} = \sqrt{n}Z_n^X(\tilde{e}_{j,k})$  for all  $(j,k) \in \mathcal{I}_K$ . The difference between  $X_n(g)$  and  $Z_n^X(g)$ , for all  $g \in \Pi_0\mathcal{H}$ , will rely on the coefficient  $(\tilde{\beta}_{j,k}(g) : (j,k) \in \mathcal{I}_K, g \in \Pi_0\mathcal{H})$  and the coupling between  $\tilde{U}_{j,k}$  and  $\tilde{V}_{j,k}$ , which is the essence of Theorem 2.1 in Rio (1994). Although Rio (1994, Theorem 2.1) is stated for i.i.d. Uniform( $[0,1]$ ) random variables, the underlying process only depends through the counts of the random variables taking values in each interval of the form  $[k2^{-j}, (k+1)2^{-j}]$  for  $(j,k) \in \mathcal{J}_K$ , which have the same distribution as the counts  $(\sum_{i=1}^n \mathbb{1}(\mathbf{x}_i \in \mathcal{C}_{j,k}) : (j,k) \in \mathcal{J}_K)$ . Therefore, we have the following corollary of Rio (1994, Theorem 2.1) under Assumption SA.1. Recall the definitions given in Section SA-II.1.2.

**Lemma SA.3.** *Suppose Assumption SA.1 holds, a dyadic expansion  $\mathcal{C}_K(\mathbb{P}_X, 1)$  is given, and  $(Z_n^X(h) : h \in \mathcal{H} \cup \Pi_0\mathcal{H})$  is the Gaussian process constructed as in (SA-2) on a possibly enlarged probability space. Then, for any  $g \in \Pi_0\mathcal{H}$  and any  $t > 0$ ,*

$$\mathbb{P}\left(\sqrt{n}|X_n(g) - Z_n^X(g)| \geq 24\sqrt{\|g\|_{\mathcal{E}_K}^2} t + 4\sqrt{\mathbf{C}_{\{g\},K}t}\right) \leq 2\exp(-t),$$

where

$$\|g\|_{\mathcal{E}_K}^2 = \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} |\tilde{\beta}_{j,k}(g)|^2$$

using the definitions in Lemma SA.1, and

$$\mathbf{C}_{\mathcal{F},K} = \sup_{f \in \mathcal{F}} \min \left\{ \sup_{(j,k) \in \mathcal{I}_K} \left[ \sum_{1 \leq l < j} (j-l)(j-l+1)2^{l-j} \sum_{0 \leq m < 2^{K-l}, \mathcal{C}_{l,m} \subseteq \mathcal{C}_{j,k}} \tilde{\beta}_{l,m}^2(f) \right] + \mathbf{M}_{\{f\},\mathcal{C}_{K,0}}^2, K\mathbf{M}_{\{f\},\mathcal{C}_{K,0}}^2 \right\},$$

for any  $\mathcal{F} \subseteq \mathcal{H} \cup \Pi_0\mathcal{H}$ .

**Proof of Lemma SA.3.** Let  $(w_i : 1 \leq i \leq n)$  be i.i.d.  $\text{Uniform}([0, 1])$ , and  $I_{j,k} = [k2^{-j}, (k+1)2^{-j})$  for  $(j, k) \in \mathcal{J}_K$ . Take  $B$  to be a Brownian bridge on  $[0, 1]$ , that is, there exists a standard Wiener process  $W$  such that  $B(t) = W(t) - tW(1)$  for all  $t \in [0, 1]$ . Take

$$\begin{aligned} v_{j,k} &= \sqrt{n} \int_0^1 \mathbb{1}(t \in I_{j,k}) dB(t), & (j, k) \in \mathcal{J}_K, \\ \tilde{v}_{j,k} &= v_{j-1,2k} - v_{j-1,2k+1}, & (j, k) \in \mathcal{I}_K. \end{aligned}$$

Take  $F_m$  to be the cumulative distribution function of  $(S_m - \frac{1}{2}m)/\sqrt{m/4}$ , where  $S_m$  is a  $\text{Bin}(m, 1/2)$  random variable, and  $G_m(t) = \inf\{x : F_m(x) \geq t\}$ . Define  $u_{j,k}$ 's and  $\tilde{u}_{j,k}$ 's via the iterative quantile transformation:

1. *Initialization:*  $u_{K,0} = n$ .
2. *Iteration:* For each  $0 \leq j \leq K-1$ , and given  $(u_{l,k} : 0 \leq k < 2^{K-l}, j < l \leq K)$ , then solve for  $(u_{j,k} : 0 \leq k < 2^{K-j})$  such that

$$\begin{aligned} \tilde{u}_{j,k} &= \frac{1}{2} \sqrt{u_{j,k}} G_{u_{j,k}} \circ \Phi(\tilde{\xi}_{j,k}), \\ \tilde{u}_{j,k} &= \frac{1}{2} u_{j-1,2k} - \frac{1}{2} u_{j-1,2k+1} = u_{j-1,2k} - \frac{1}{2} u_{j,k}, \\ u_{j-1,2k} + u_{j-1,2k+1} &= u_{j,k}, \end{aligned}$$

for  $0 \leq k < 2^{K-j}$ . Continue till  $(u_{0,k} : 0 \leq k < 2^K)$  are defined.

Then  $u_{j,k}$ 's have the same joint distribution as  $\sum_{i=1}^n \mathbb{1}(w_i \in I_{j,k})$ 's. Hence, by Skorohod Embedding lemma (Dudley, 2014, Lemma 3.35), on a rich enough probability space, we can take  $(B(t) : 0 \leq t \leq 1)$  such that  $u_{j,k} = \sum_{i=1}^n \mathbb{1}(w_i \in I_{j,k})$  for all  $(j, k) \in \mathcal{J}_K$ , almost surely.

Observe  $\{(\tilde{u}_{j,k}, \tilde{v}_{j,k}) : (j, k) \in \mathcal{I}_K\}$  and  $\{(\tilde{U}_{j,k}, \tilde{V}_{j,k}) : (j, k) \in \mathcal{I}_K\}$  have the same joint distribution, and

$$(X_n(g), Z_n^X(g)) = \left( \frac{1}{\sqrt{n}} \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(g) \tilde{U}_{j,k}, \frac{1}{\sqrt{n}} \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(g) \tilde{V}_{j,k} \right),$$

for all  $g \in \Pi_0 \mathcal{H}$ . Thus the distribution of the process  $\{(X_n(g), Z_n^X(g)) : g \in \Pi_0 \mathcal{H}\}$  is the same as distribution of

$$((x_n(g), z_n(g)) : g \in \Pi_0 \mathcal{H}) = \left( \left( \frac{1}{\sqrt{n}} \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(g) \tilde{u}_{j,k}, \frac{1}{\sqrt{n}} \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(g) \tilde{v}_{j,k} \right) : g \in \Pi_0 \mathcal{H} \right),$$

We can then apply Rio (1994, Theorem 2.1) on  $((x_n(g), z_n(g)) : g \in \Pi_0 \mathcal{H})$  and use its equi-distribution as  $((X_n(g), Z_n^X(g)) : g \in \Pi_0 \mathcal{H})$  to get for any  $\mathbf{p} = (p_1, \dots, p_K)$  with positive components such that  $\sum_{i=1}^K p_i \leq 1$ , if we take  $q_i = (2^i p_i)^{-1}$  and

$$M(\mathbf{p}, g) = 4 \sup_{(j,k) \in \mathcal{I}_K} \left[ \sum_{1 \leq l < j} q_{j-l} \sum_{0 \leq m < 2^{K-l} : \mathcal{C}_{l,m} \subseteq \mathcal{C}_{j,k}} \tilde{\beta}_{l,m}^2(g) \right], \quad g \in \Pi_0 \mathcal{H},$$

then for any  $t > 0$  and  $g \in \Pi_0\mathcal{H}$ ,

$$\mathbb{P}\left(\sqrt{n}|X_n(g) - Z_n^X(g)| \geq (\sqrt{M(\mathbf{p}, g)} + \mathbf{M}_{\{g\}, \mathcal{C}_{K,0}})t + \left(\left(\sum_{i=1}^K q_i/2\right)^{1/2} + 3\right)\|g\|_{\varepsilon_K}\sqrt{t}\right) \leq 2\exp(-t).$$

Following [Rio \(1994, Section 3\)](#), we choose either  $p_i = \frac{1}{2}\left(\frac{1}{K} + \frac{1}{i(i+1)}\right)$  to get

$$M(\mathbf{p}, g) \leq 8K\mathbf{M}_{\{g\}, \mathcal{C}_{K,0}} \quad \text{and} \quad \sum_{i=1}^K \frac{q_i}{2} < 8,$$

or  $p_i = \frac{1}{i(i+1)}$  to get

$$M(\mathbf{p}, g) \leq \sup_{(j,k) \in \mathcal{I}_K} \left[ \sum_{1 \leq l < j} (j-l)(j-l+1)2^{l-j} \sum_{0 \leq m < 2^{K-l}: \mathcal{C}_{l,m} \subseteq \mathcal{C}_{j,k}} \tilde{\beta}_{l,m}^2(g) \right] \text{ and } \sum_{i=1}^K \frac{q_i}{2} < 4.$$

The conclusion then follows.  $\square$

Lemma [SA.3](#) relies on a coupling of  $\text{Bin}(m, 1/2)$  random variables with Gaussian random variables. A weaker coupling also holds for  $\text{Bin}(m, p)$  with the error term only depending on how far away  $p$  is bounded away from 0 and 1, as the following lemma establishes.

**Lemma SA.4.** *Suppose  $X \sim \text{Bin}(m, p)$  where  $0 < \underline{p} < p < \bar{p} < 1$ . Then, there exists a random variable  $Z \sim \text{Normal}(0, 1)$ , and constants  $c_0, c_1, c_2, c_3 > 0$  only depending on  $\underline{p}$  and  $\bar{p}$ , such that whenever the event  $A = \{|X - mp| \leq c_1 m\}$  occurs and  $c_0 \sqrt{m} \geq 1$ , we have*

$$\left|X - mp - \sqrt{mp(1-p)}Z\right| \leq c_2 Z^2 + c_3 \quad \text{and} \quad |X - mp| \leq \frac{1}{c_0} + 2\sqrt{mp(1-p)}|Z|.$$

In particular, we can take  $c_0 > 0$  to be the solution of

$$60c_0\bar{p} \left(\sqrt{\frac{1-\underline{p}}{\underline{p}}}\right)^3 \exp\left(2\sqrt{\frac{1-\underline{p}}{\underline{p}}}c_0\right) + 60c_0(1-\underline{p}) \left(\sqrt{\frac{\bar{p}}{1-\bar{p}}}\right)^3 \exp\left(2\sqrt{\frac{\bar{p}}{1-\bar{p}}}c_0\right) = 1,$$

and  $c_1 = 15c_0\sqrt{\underline{p}(1-\bar{p})}$ ,  $c_2 = 1/(15c_0)$ ,  $c_3 = 1/c_0$ , and then set

$$Z = \Phi^{-1} \circ F\left((X - mp)/\sqrt{mp(1-p)}\right).$$

That is,  $Z$  can be taken via the quantile transformation based on  $F(x) = \mathbb{P}(X - mp < \sqrt{mp(1-p)}x)$ .

**Proof of Lemma SA.4.** Let  $(X_j : 1 \leq j \leq m)$  be i.i.d.  $\text{Bern}(p)$  with  $0 < \underline{p} < p < \bar{p} < 1$ . Take  $\xi_j = (X_j - p)/\sqrt{mp(1-p)}$  and  $S_m = \sum_{j=1}^m \xi_j$ . Then, for any  $a \in \mathbb{R}$ ,

$$\begin{aligned} L(a) &= \sum_{j=1}^m \mathbb{E} \left[ |\xi_j|^3 \exp(|a\xi_j|) \right] = \sum_{j=1}^m \mathbb{E} \left[ \left| \frac{X_j - p}{\sqrt{mp(1-p)}} \right|^3 \exp\left(a \left| \frac{X_j - p}{\sqrt{mp(1-p)}} \right| \right) \right] \\ &= mp \left( \frac{1-p}{\sqrt{mp(1-p)}} \right)^3 \exp\left(a \frac{1-p}{\sqrt{mp(1-p)}}\right) + m(1-p) \left( \frac{p}{\sqrt{mp(1-p)}} \right)^3 \exp\left(a \frac{p}{\sqrt{mp(1-p)}}\right). \end{aligned}$$

Take  $c_0 > 0$  such that

$$60c_0\bar{p} \left( \sqrt{\frac{1-p}{p}} \right)^3 \exp \left( 2\sqrt{\frac{1-p}{p}}c_0 \right) + 60c_0(1-p) \left( \sqrt{\frac{p}{1-p}} \right)^3 \exp \left( 2\sqrt{\frac{p}{1-p}}c_0 \right) = 1.$$

Then, for any  $m \in \mathbb{N}$  and  $\lambda = c_0\sqrt{m}$ , we have  $60\lambda L(2\lambda) \leq 1$ . [Sakhanenko \(1996, Lemma 2\)](#) implies that, whenever  $c_0\sqrt{m} \geq 1$  and the event  $\{|S_m| < c_0\sqrt{m}\}$  occurs,

$$|S_m - Z| \leq \frac{1}{c_0\sqrt{m}} + \frac{S_n^2}{60c_0\sqrt{m}}.$$

Moreover,  $Z$  can be taken such that  $Z = \Phi^{-1} \circ F(S_m)$ .

We then proceed as in the proof for Lemma 2 in [Brown et al. \(2010\)](#), where they show for each  $0 < p < 1$ , the coupling exits with  $c_0$  to  $c_3$  not depending on  $m$ , though they did not give explicit dependency of  $c_0$  to  $c_3$  on  $p$ . Take  $c_1$  such that  $c_1/(60c_0) < 1/2$ . In particular, we can take  $c_1 = 15c_0$ . Then, on the event  $\{|S_m| < c_1\sqrt{m}\}$ ,

$$|S_m - Z| \leq \frac{1}{c_0\sqrt{m}} + |S_m| \frac{c_1\sqrt{m}}{60c_0\sqrt{m}} \leq \frac{1}{c_0\sqrt{m}} + \frac{1}{2}|S_m|.$$

Hence, by triangle inequality,  $|S_m| \leq \frac{2}{c_0\sqrt{m}} + 2|Z|$ , and

$$|S_m - Z| \leq \frac{1}{c_0\sqrt{m}} + \frac{1}{60c_0\sqrt{m}} \left( \frac{2}{c_0\sqrt{m}} + 2|Z| \right)^2 \leq \frac{2}{c_0\sqrt{m}} + \frac{2}{15c_0\sqrt{m}}|Z|^2.$$

Since  $X = \sum_{j=1}^m X_j \sim \text{Bin}(m, p)$ , whenever the event  $\{|X - mp| < c_1m\sqrt{p(1-p)}\}$  occurs and  $c_0\sqrt{m} \geq 1$ ,

$$\left| X - mp - \sqrt{mp(1-p)}Z \right| \leq \frac{2}{c_0}\sqrt{p(1-p)} + \frac{2}{15c_0}\sqrt{p(1-p)}|Z|^2 \leq \frac{1}{c_0} + \frac{Z^2}{15c_0}.$$

Moreover,  $|S_m| \leq \frac{2}{c_0\sqrt{m}} + 2|Z|$  implies  $|X - mp| \leq \frac{1}{c_0} + 2\sqrt{mp(1-p)}|Z|$ .  $\square$

This generalization of Tusnády's Lemma enables the following strong approximation for the case of a quasi-dyadic cells expansion.

**Lemma SA.5.** *Suppose Assumption SA.1 holds, a quasi-dyadic expansion  $\mathcal{C}_K(\mathbb{P}_X, \rho)$  is given with  $\rho > 1$ , and  $(Z_n^X(h) : h \in \mathcal{H} \cup \Pi_0\mathcal{H})$  is the Gaussian process constructed as in (SA-2) on a possibly enlarged probability space. Then, for any  $g \in \Pi_0\mathcal{H}$  and for any  $t > 0$ ,*

$$\mathbb{P} \left( \sqrt{n} |X_n(g) - Z_n^X(g)| \geq c_\rho \sqrt{\|g\|_{\mathcal{E}_K}^2 t} + c_\rho \sqrt{\mathcal{C}_{\{g\}, K} t} \right) \leq 2 \exp(-t) + 2^{K+2} \exp(-c_\rho n 2^{-K}),$$

where  $c_\rho$  is a constant that only depends on  $\rho$ , and  $\|g\|_{\mathcal{E}_K}^2$  and  $\mathcal{C}_{\{g\}, K}$  are defined in Lemma SA.3.

**Proof of Lemma SA.5.** We adopt the coupling method from Section 2 of [Rio \(1994\)](#), extending it to accommodate quasi-dyadic cells. Instead of applying the well-known Tusnády inequality as in [Rio \(1994\)](#), which states that for  $X \sim \text{Bin}(m, \frac{1}{2})$ , there exists  $Z \sim \text{Normal}(0, 1)$  such that almost surely:

$$\left| X - \frac{m}{2} - \left( \frac{\sqrt{m}}{2} \right) Z \right| \leq 1 + \frac{Z^2}{8}, \quad \text{and} \quad \left| X - \frac{m}{2} \right| \leq 1 + \frac{\sqrt{m}}{2}|Z|,$$



we rely on Lemma SA.4, which allows for coupling in the case of  $\text{Bin}(m, p)$  with  $p \neq \frac{1}{2}$ , though restricted to a high-probability set. The proof proceeds in two parts: Part 1 establishes an upper bound for the small-probability event where the coupling inequalities from Lemma SA.4 fail to hold; Part 2 decomposes the error  $X_n(g) - Z_n^X(g)$  into the coupling errors corresponding to each pair of cells  $(\mathcal{C}_{j-1,2k}, \mathcal{C}_{j-1,2k+1})$ , following the strategy in Rio (1994), while accounting for the restriction to the high-probability set.

*Part 1: Strong Approximation Set-up.* By the construction at Equation (2), condition on  $U_{j,k}$ ,  $\tilde{U}_{j,k}$  has the same distribution as  $2\text{Bin}(U_{j,k}, p_{j,k}) - U_{j,k}$ , and the conditional quantile transformation relation  $\tilde{U}_{j,k} = \sqrt{U_{j,k}p_{j,k}(1-p_{j,k})}G_{(j,k),U_{j,k}} \circ \Phi(\tilde{\xi}_{j,k})$  holds. This allows for application of Lemma SA.4. Let  $\bar{p} = \rho$ ,  $\underline{p} = \rho^{-1}$ ,  $c_0$  to be the positive solution of

$$60c_0\bar{p} \left( \sqrt{\frac{1-\underline{p}}{\underline{p}}} \right)^3 \exp \left( 2\sqrt{\frac{1-\underline{p}}{\underline{p}}}c_0 \right) + 60c_0(1-\underline{p}) \left( \sqrt{\frac{\bar{p}}{1-\bar{p}}} \right)^3 \exp \left( 2\sqrt{\frac{\bar{p}}{1-\bar{p}}}c_0 \right) = 1,$$

$c_1 = 15c_0\sqrt{\underline{p}(1-\bar{p})}$ ,  $c_2 = 1/(15c_0)$ , and  $c_3 = 1/c_0$ . Consider the small probability set  $\mathcal{A}$  where the coupling inequalities from Lemma SA.4 are not guaranteed to hold,

$$\mathcal{A} = \left\{ |\tilde{U}_{j,k}| \leq c_1 U_{j,k} : (j, k) \in \mathcal{I}_K \right\},$$

and notice that we can always take  $c_1 \leq 1$  because  $|\tilde{U}_{j,k}| \leq U_{j,k}$  almost surely. Using Lemma SA.4 conditional on  $U_{j,k}$ , whenever  $\mathcal{A}$  occurs,

$$\begin{aligned} \left| \tilde{U}_{j,k} - \sqrt{U_{j,k}p_{j,k}(1-p_{j,k})}\tilde{\xi}_{j,k} \right| &< c_2\tilde{\xi}_{j,k}^2 + c_3, \\ \left| \tilde{U}_{j,k} \right| &\leq 1/c_0 + 2\sqrt{p_{j,k}(1-p_{j,k})}|\tilde{\xi}_{j,k}|, \end{aligned} \tag{SA-3}$$

for all  $(j, k) \in \mathcal{I}_K$ .

To bound  $\mathbb{P}(\mathcal{A}^c)$ , first notice that by Chernoff's inequality for Binomial distribution,  $\mathbb{P}(U_{j,k} \leq \mathbb{E}[U_{j,k}]/2) \leq \exp(-\mathbb{E}[U_{j,k}]/8)$  for all  $(j, k) \in \mathcal{I}_K$ , and  $\mathbb{P}(U_{j,k} \leq 2^{-1}\rho^{-1}n2^{j-K}) \leq \exp(-8^{-1}\rho^{-1}n2^{j-K})$  for all  $(j, k) \in \mathcal{I}_K$  because  $\rho^{-1}n2^{j-K} \leq \mathbb{E}[U_{j,k}] \leq \rho n2^{j-K}$ . Furthermore, using Hoeffding's inequality and the fact that  $\tilde{U}_{j,k} = U_{j-1,2k} - p_{j,k}U_{j,k} = U_{j-1,2k} - \mathbb{E}[U_{j-1,2k}|U_{j,k}]$ ,

$$\mathbb{P}\left(|\tilde{U}_{j,k}| \geq c_1 U_{j,k} \mid U_{j,k} \geq \frac{1}{2}\rho^{-1}n2^{-K+j}\right) \leq 2 \exp\left(-\frac{c_1^2 n 2^{-K+j}}{3\rho}\right).$$

Putting these together, and using the union bound,

$$\begin{aligned} \mathbb{P}(\mathcal{A}^c) &\leq \sum_{(j,k) \in \mathcal{I}_K} \mathbb{P}(|\tilde{U}_{j,k}| > c_1 U_{j,k}) \\ &\leq \sum_{(j,k) \in \mathcal{I}_K} \mathbb{P}\left(U_{j,k} \leq \frac{1}{2}\rho^{-1}n2^{-K+j}\right) + \mathbb{P}\left(|\tilde{U}_{j,k}| \geq c_1 U_{j,k} \mid U_{j,k} \geq \frac{1}{2}\rho^{-1}n2^{-K+j}\right) \\ &\leq \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \left\{ \exp(-8^{-1}\rho^{-1}n2^{j-K}) + 2 \exp(-c_1^2 \rho^{-1}n2^{j-K}/3) \right\} \\ &\leq 3 \cdot 2^K \exp(-\min\{c_1^2/3, 1/8\}\rho^{-1}n2^{-K}). \end{aligned} \tag{SA-4}$$

*Part 2: Bounding Strong Approximation Error.* We show that the proof of Theorem 2.1 in Rio (1994) still goes through for an approximate dyadic scheme. In other words, we show that the approximate dyadic scheme gives essentially the same Gaussian coupling rates as the dyadic scheme (Section SA-II.1.1). We employ the same notation as in Rio (1994), and for  $g \in \Pi_0\mathcal{H}$ , define

$$\begin{aligned}\Delta(g) &= (X - Z)(g), & X(g) &= \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(g) \tilde{U}_{j,k}, & Z(g) &= \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(g) \tilde{V}_{j,k}, \\ \Delta_1(g) &= (X - Y)(g), & Y(g) &= \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(g) \sqrt{U_{j,k} \tilde{p}_{j,k} (1 - \tilde{p}_{j,k})} \tilde{\xi}_{j,k}, \\ \Delta_2(g) &= (Y - Z)(g)\end{aligned}$$

It suffices to verify the following two claims.

Claim 1:  $\mathbb{E}[\exp(t\Delta_1(g))\mathbb{1}(\mathcal{A})] \leq \prod_{j=1}^K \prod_{0 \leq k < 2^{K-j}} \mathbb{E}[\cosh(t\tilde{\beta}_{j,k}(g)(2 + \tilde{\xi}_{j,k}^2/4))]$  for all  $g \in \Pi_0\mathcal{H}$ . Then, it follows from the proof of Lemma 2.2 in Rio (1994) that

$$\log \mathbb{E}[\exp(4t\Delta_1(g))\mathbb{1}(\mathcal{A})] \leq -\frac{83}{3}c_\rho^2 \left( \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}^2(g) \right) \log(1 - t^2),$$

for all  $|t| < 1$ .

Claim 2:  $\mathbb{E}[\exp(t\Delta_2)\mathbb{1}(\mathcal{A})] \leq \mathbb{E}[\exp(tc_\rho\Delta_3)]$  for all  $t > 0$ , where

$$\Delta_3(g) = \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(g) \tilde{\xi}_{j,k} \left( 1 + \sum_{l=j}^K \sum_{0 \leq q < 2^{K-l}} 2^{-|j-l|/2} |\tilde{\xi}_{l,q}| \mathbb{1}(\mathcal{C}_{l,q} \supseteq \mathcal{C}_{j,k}) \right),$$

for all  $g \in \Pi_0\mathcal{H}$ , and  $c_\rho$  a constant that only depends on  $\rho$ .

Proof of Claim 1: Let  $\mathcal{F}_j = \sigma(\{\tilde{\xi}_{l,k} : j < l \leq K, 0 \leq k < 2^{K-l}\})$  for all  $1 \leq j < K$ . In particular,  $\sigma(\{U_{l,k} : j \leq l \leq K, 0 \leq k < 2^{K-l}\}) \subseteq \mathcal{F}_j$ . Then, by Equation SA-3, for all  $t \in \mathbb{R}$ ,

$$\begin{aligned}\mathbb{E} \left[ \exp \left( t \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(g) \left( \tilde{U}_{j,k} - \sqrt{U_{j,k} \tilde{p}_{j,k} (1 - \tilde{p}_{j,k})} \tilde{\xi}_{j,k} \right) \right) \mathbb{1}(\mathcal{A}) \middle| \mathcal{F}_j \right] \\ \leq \mathbb{E} \left[ \prod_{0 \leq k < 2^{K-j}} \cosh \left( t \tilde{\beta}_{j,k}(g) (c_2 \tilde{\xi}_{j,k}^2 + c_3) \right) \mathbb{1}(\mathcal{A}) \middle| \mathcal{F}_j \right].\end{aligned}$$

Then, we will use the same induction argument as in the proof of Lemma 2.2 in Rio (1994): let

$$S_j(t) = \exp \left( t \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(g) \left( \tilde{U}_{j,k} - \sqrt{U_{j,k} \tilde{p}_{j,k} (1 - \tilde{p}_{j,k})} \tilde{\xi}_{j,k} \right) \right),$$

so that  $\mathbb{E}[\exp(t\Delta_1)\mathbb{1}(\mathcal{A})] = \mathbb{E}[\prod_{j=1}^K S_j(t)\mathbb{1}(\mathcal{A})]$ , and

$$T_j(t) = \prod_{0 \leq k < 2^{K-j}} \cosh\left(t\tilde{\beta}_{j,k}(g)(c_2\tilde{\xi}_{j,k}^2 + c_3)\right),$$

so that  $\prod_{j=1}^K \prod_{0 \leq k < 2^{K-j}} \mathbb{E}[\cosh(t\tilde{\beta}_{j,k}(2 + \tilde{\xi}_{j,k}^2/4))] = \mathbb{E}[\prod_{j=1}^K T_j(t)]$ . By Equation SA-3, for all  $1 \leq j \leq K$ ,

$$\mathbb{E}\left[S_j(t) \prod_{l=1}^{j-1} T_l(t)\mathbb{1}(\mathcal{A}) \middle| \mathcal{F}_j\right] \leq \mathbb{E}\left[\prod_{l=1}^j T_l(t)\mathbb{1}(\mathcal{A}) \middle| \mathcal{F}_j\right].$$

It follows that

$$\begin{aligned} \mathbb{E}[\exp(t\Delta_1)\mathbb{1}(\mathcal{A})] &= \mathbb{E}\left[\prod_{j=1}^K S_j(t)\mathbb{1}(\mathcal{A})\right] = \mathbb{E}\left[\mathbb{E}[S_1(t)\mathbb{1}(\mathcal{A})|\mathcal{F}_1] \prod_{j=2}^K S_j(t)\right] \leq \mathbb{E}\left[\mathbb{E}[T_1(t)\mathbb{1}(\mathcal{A})|\mathcal{F}_1] \prod_{j=2}^K S_j(t)\right] \\ &= \mathbb{E}\left[\mathbb{E}[T_1(t)S_2(t)\mathbb{1}(\mathcal{A})|\mathcal{F}_2] \prod_{j=3}^K S_j(t)\right] \leq \mathbb{E}\left[\mathbb{E}[T_1(t)T_2(t)\mathbb{1}(\mathcal{A})|\mathcal{F}_2] \prod_{j=3}^K S_j(t)\right] \\ &\leq \mathbb{E}\left[\prod_{j=1}^K T_j(t)\mathbb{1}(\mathcal{A})\right] \leq \mathbb{E}\left[\prod_{j=1}^K T_j(t)\right] = \prod_{j=1}^K \prod_{0 \leq k < 2^{K-j}} \mathbb{E}[\cosh(t\tilde{\beta}_{j,k}(h)(c_2\tilde{\xi}_{j,k}^2 + c_3))] \\ &\leq \prod_{j=1}^K \prod_{0 \leq k < 2^{K-j}} \mathbb{E}[\cosh(tc_\rho\tilde{\beta}_{j,k}(h)(\tilde{\xi}_{j,k}^2/4 + 2))] \end{aligned}$$

where in the last line, we have used independence of  $(\tilde{\xi}_{j,k} : 1 \leq j \leq K, 0 \leq k < 2^{K-j})$ . Without loss of generality, we assume that  $c_\rho \sup_{\mathbf{x} \in \mathcal{C}_{K,0}} |g(\mathbf{x})| \leq 1$ . Since we know  $(\tilde{\xi}_{j,k}, 1 \leq j \leq K, 0 \leq k < 2^{K-j})$  are i.i.d. standard Gaussian, the same upper bound established in Rio (1994) for the right hand side of the last display holds: for all  $g \in \Pi_0\mathcal{H}$ ,  $|t| < 1$ ,

$$\log \mathbb{E}[\exp(4t\Delta_1(g))\mathbb{1}(\mathcal{A})] \leq -\frac{83}{3}c_{\rho^2} \left( \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}^2(h) \right) \log(1 - t^2) = h_{\Delta_1}(t), \quad (\text{SA-5})$$

which concludes the verification of the first claim.

Proof of Claim 2: Denote  $q_{j,k} = \mathbb{P}_X(\mathcal{C}_{j,k})$  for  $(j,k) \in \mathcal{J}_K$ . By Equation (2), for any  $g \in \Pi_0\mathcal{H}$ , we have

$$\Delta_2(g) = \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(g) \left( \sqrt{U_{j,k}} - \sqrt{\mathbb{E}[U_{j,k}]} \right) \sqrt{\frac{q_{j-1,2k}q_{j-1,2k+1}}{q_{j,k}^2}} \tilde{\xi}_{j,k}.$$

We will use the same strategy as in Rio (1994) adapted to the quasi-dyadic case. Fix  $0 \leq l < 2^{K-j}$  and  $0 \leq j \leq K$ , and let  $k_l$  be the unique integer in  $[0, 2^{K-l})$  such that  $\mathcal{C}_{l,k_l} \supseteq \mathcal{C}_{j,k}$ . Then,

$$\begin{aligned} \sqrt{U_{j,k}} - \sqrt{\mathbb{E}[U_{j,k}]} &= \sum_{l=j}^{K-1} \sqrt{U_{l,k_l} \frac{q_{j,k}}{q_{l,k_l}}} - \sqrt{U_{l+1,k_{l+1}} \frac{q_{j,k}}{q_{l+1,k_{l+1}}}} \\ &= \sum_{l=j}^{K-1} \sqrt{\frac{q_{j,k}}{q_{l+1,k_{l+1}}}} \left( \sqrt{\frac{q_{l+1,k_{l+1}}}{q_{l,k_l}} U_{l,k_l}} - \sqrt{U_{l+1,k_{l+1}}} \right). \end{aligned}$$

By Equation SA-3, when the event  $\mathcal{A}$  holds,

$$\begin{aligned}
\left| \sqrt{\frac{q_{l+1,k_{l+1}}}{q_{l,k_l}} U_{l,k_l}} - \sqrt{U_{l+1,k_{l+1}}} \right| &\leq \frac{|\tilde{U}_{l,k_l}|}{\sqrt{\frac{q_{l+1,k_{l+1}}}{q_{l,k_l}} U_{l,k_l} + \sqrt{U_{l+1,k_{l+1}}}}} \\
&\leq \frac{2\sqrt{\frac{q_{l+1,2k_l}}{q_{l,k_l}} \frac{q_{l+1,2k_{l+1}}}{q_{l,k_l}} U_{l,k_l}} |\tilde{\xi}_{l,k_l}| + \min\{c_0^{-1}, \tilde{U}_{l,k_l}\}}{\sqrt{\frac{q_{l+1,k_{l+1}}}{q_{l,k_l}} U_{l,k_l} + \sqrt{U_{l+1,k_{l+1}}}}} \\
&\leq 2\sqrt{\frac{q_{l+1,2k_l+1}}{q_{l,k_l}} |\tilde{\xi}_{l,k_l}|} + \frac{\min\{c_0^{-1}, |\tilde{U}_{l,k_l}|\}}{\sqrt{\frac{q_{l+1,k_{l+1}}}{q_{l,k_l}} U_{l,k_l} + \sqrt{U_{l+1,k_{l+1}}}}}.
\end{aligned}$$

For the first summand,

$$\sum_{l=j}^{K-1} \sqrt{\frac{q_{j,k}}{q_{l+1,k_{l+1}}}} 2\sqrt{\frac{q_{l+1,2k_{l+1}}}{q_{l,k_l}}} |\tilde{\xi}_{l,k_l}| = \sum_{l=j}^{K-1} \sqrt{\prod_{j < s \leq l} p_{s,k_s}} 2\sqrt{p_{l,k_l}} |\tilde{\xi}_{l,k_l}| \leq c_\rho \sum_{l=j}^{K-1} 2^{-(l-j)/2} |\tilde{\xi}_{l,k_l}|.$$

For the second summand, we separate it into two terms as in Rio (1994). For  $\mathbb{1}(\tilde{U}_{l,k_l} \leq 0)$ , we have

$$\begin{aligned}
&\sum_{l=j}^{K-1} \sqrt{\frac{q_{j,k}}{q_{l+1,k_{l+1}}}} \frac{\min\{c_0^{-1}, -\tilde{U}_{l,k_l}\}}{\sqrt{\frac{q_{l+1,k_{l+1}}}{q_{l,k_l}} U_{l,k_l} + \sqrt{U_{l+1,k_{l+1}}}}} \mathbb{1}(\tilde{U}_{l,k_l} \leq 0) \\
&= \sum_{l=j}^{K-1} \sqrt{\frac{q_{j,k}}{q_{l+1,k_{l+1}}}} \frac{\min\{c_0^{-1}, -\tilde{U}_{l,k_l}\}}{\sqrt{U_{l+1,k_{l+1}} - \tilde{U}_{l,k_l} + \sqrt{U_{l+1,k_{l+1}}}}} \mathbb{1}(\tilde{U}_{l,k_l} \leq 0) \leq c_\rho,
\end{aligned}$$

since  $\sup_{0 \leq x \leq u} \min\{c_0^{-1}, x\} / (\sqrt{u} + \sqrt{u+x}) \lesssim 1$ . For  $\mathbb{1}(\tilde{U}_{l,k_l} > 0)$ , we have

$$\begin{aligned}
&\sum_{l=j}^{K-1} \sqrt{\frac{q_{j,k}}{q_{l+1,k_{l+1}}}} \frac{\min\{c_0^{-1}, \tilde{U}_{l,k_l}\}}{\sqrt{\frac{q_{l+1,k_{l+1}}}{q_{l,k_l}} U_{l,k_l} + \sqrt{U_{l+1,k_{l+1}}}}} \mathbb{1}(\tilde{U}_{l,k_l} > 0) \\
&\leq \sum_{l=j}^{K-1} \sqrt{\frac{q_{j,k}}{q_{l+1,k_{l+1}}}} \left( \sqrt{U_{l+1,k_{l+1}}} - \sqrt{\frac{q_{l+1,k_{l+1}}}{q_{l,k_l}} U_{l,k_l}} \right) \mathbb{1}\left( \frac{q_{l+1,k_{l+1}}}{q_{l,k_l}} U_{l,k_l} \leq U_{l+1,k_{l+1}} \leq \frac{q_{l+1,k_{l+1}}}{q_{l,k_l}} U_{l,k_l} + c_0^{-1} \right) \\
&\leq \sum_{l=j}^{K-1} \sqrt{\frac{q_{j,k}}{q_{l+1,k_{l+1}}}} \sqrt{c_0^{-1}} = \sum_{l=j}^{K-1} \sqrt{\prod_{j < s \leq l} p_{s,k_s}} \sqrt{c_0^{-1}} \leq c_\rho.
\end{aligned}$$

It follows that when the event  $\mathcal{A}$  holds,

$$\left| \sqrt{U_{j,k}} - \sqrt{\mathbb{E}[U_{j,k}]} \right| \leq c_\rho \left( 1 + \sum_{l=j}^{K-1} 2^{-(l-j)/2} \sum_{0 \leq q < 2^{K-l}} |\tilde{\xi}_{l,q}| \mathbb{1}(\mathcal{C}_{l,q} \supseteq \mathcal{C}_{j,k}) \right).$$

Using an induction argument, for all  $g \in \Pi_0 \mathcal{H}$ ,  $t > 0$ ,

$$\mathbb{E}[\exp(t\Delta_2(g)) \mathbb{1}(\mathcal{A})] \leq \mathbb{E}[\exp(tc_\rho \Delta_3(g))]. \tag{SA-6}$$

For any random variable  $W$ , define  $\gamma_W(t) = \log(\mathbb{E}[\exp(tc_\rho W)])$  for all  $t > 0$ , and  $h_W(u) = \sup_{t > 0} (tu -$

$\gamma_W(u)$ . Combining Equation (SA-5), for any  $g \in \Pi_0\mathcal{H}$ ,  $t > 0$ ,

$$\begin{aligned} \mathbb{P}(\Delta_1(g) \geq t \text{ and } \mathcal{A}) &\leq \inf_{u>0} \mathbb{P}(\exp(\Delta_1(g)u) \geq \exp(tu) \text{ and } \mathcal{A}) \leq \inf_{u>0} \exp(-tu) \mathbb{E}[\exp(\Delta_1(g)u) \mathbb{1}(\mathcal{A})] \\ &\leq \exp(-h_{\Delta_1(g)}(t)) = \exp\left(-\sup_{u>0} \left(tu + \frac{83}{3} c_\rho^2 \|g\|_{\varepsilon_K}^2 \log(1 - u^2/16)\right)\right), \end{aligned}$$

hence for any  $t > 0$ ,

$$\mathbb{P}(|\Delta_1(g)| \geq Cc_\rho \|g\|_{\varepsilon_K} \sqrt{t} + Ct \text{ and } \mathcal{A}) = \mathbb{P}(\Delta_1(g) \geq h_{\Delta_1(g)}^{-1}(t) \text{ and } \mathcal{A}) \leq 2 \exp(-t). \quad (\text{SA-7})$$

By Equation (SA-6), for any  $t > 0$ ,

$$\mathbb{P}(\Delta_2(g) \geq t \text{ and } \mathcal{A}) \leq \inf_{u>0} \exp(-tu) \mathbb{E}[\exp(\Delta_2(g)u) \mathbb{1}(\mathcal{A})] \leq \exp(-h_{\Delta_3(g)}(t)). \quad (\text{SA-8})$$

Since  $\Delta_3(g)$  only depends on  $((\tilde{\xi}_{j,k}, \tilde{\beta}_{j,k}(g)) : (j,k) \in \mathcal{I}_K)$ , the rest of the proof follows from Lemma 2.4 in Rio (1994). In particular, define

$$\Delta_4(g) = \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(g) \tilde{\xi}_{j,k}, \quad \Delta_5(g) = \Delta_3(g) - \Delta_4(g),$$

then identifying that  $\Delta_4(g)$  is Gaussian and applying Rio (1994, Lemma 2.4) with two choices of  $p_i$ -sequence,  $p_i = \frac{1}{2}(\frac{1}{K} + \frac{1}{i(i+1)})$  and  $p_i = \frac{1}{i(i+1)}$  separately on  $\Delta_5(g)$ , we get for any  $t > 0$ , and  $g \in \Pi_0\mathcal{H}$ ,

$$\mathbb{P}\left(|\Delta_2(g)| \geq c_\rho \|g\|_{\varepsilon_K} \sqrt{t} + c_\rho \sqrt{\mathcal{C}_{\{g\},K} t} \text{ and } \mathcal{A}\right) \leq \mathbb{P}\left(|\Delta_3(g)| \geq c_\rho \|g\|_{\varepsilon_K} \sqrt{t} + c_\rho \sqrt{\mathcal{C}_{\{g\},K} t}\right) \leq 2 \exp(-t).$$

Combining Equation (SA-4), (SA-7) and (SA-8), we get the stated result.  $\square$

#### SA-II.1.4 Meshing Error

For  $0 < \delta \leq 1$ , consider the  $(\delta M_{\mathcal{H},\mathcal{X}})$ -net of  $(\mathcal{H}, \|\cdot\|_{\mathbb{P}_{X,2}})$  over  $\mathcal{X}$ ,  $\mathcal{H}_\delta$ , with cardinality no larger than  $N_{\mathcal{H},\mathcal{X}}(\delta, M_{\mathcal{H},\mathcal{X}})$ . Define  $\pi_{\mathcal{H}_\delta} : \mathcal{H} \mapsto \mathcal{H}_\delta$  such that  $\|\pi_{\mathcal{H}_\delta}(h) - h\|_{\mathbb{P}_{X,2}} \leq \delta M_{\mathcal{H},\mathcal{X}}$  for all  $h \in \mathcal{H}$ . To simplify notation, in this section the parameters of  $\mathcal{H}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{X}$ , and the index  $\mathcal{C}$  is omitted whenever there is no confusion.

**Lemma SA.6.** *Suppose Assumption SA.1 holds, a quasi-dyadic expansion  $\mathbb{C}_K(\mathbb{P}_X, \rho)$  is given,  $(Z_n^X(h) : h \in \mathcal{H} \cup \Pi_0\mathcal{H})$  is the Gaussian process constructed as in (SA-2) on a possibly enlarged probability space, and  $\mathcal{H}_\delta$  is chosen in Section SA-II.1.4. Then, for all  $t > 0$  and  $0 < \delta < 1$ ,*

$$\begin{aligned} \mathbb{P}[\|X_n - X_n \circ \pi_{\mathcal{H}_\delta}\|_{\mathcal{H}} \gtrsim F_n(t, \delta)] &\leq \exp(-t), \\ \mathbb{P}[\|Z_n^X \circ \pi_{\mathcal{H}_\delta} - Z_n^X\|_{\mathcal{H}} \gtrsim M_{\mathcal{H}} J(\delta, \mathcal{H}, M_{\mathcal{H}}) + \delta M_{\mathcal{H}} \sqrt{t}] &\leq \exp(-t). \end{aligned}$$

**Proof of Lemma SA.6.** Take  $\mathcal{L} = \{h - \pi_{\mathcal{H}_\delta}(h) : h \in \mathcal{H}\}$  on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P}_X)$ . Then,  $\sup_{l \in \mathcal{L}} \|l\|_{\mathbb{P}_{X,2}} \leq \delta M_{\mathcal{H}}$  and, for all  $0 < \varepsilon < \delta$ ,

$$N_{\mathcal{L}}(\varepsilon, M_{\mathcal{H}}) \leq N_{\mathcal{L}}(\varepsilon, M_{\mathcal{H}}) N_{\mathcal{L}}(\delta, M_{\mathcal{H}}) \leq N_{\mathcal{L}}(\varepsilon, M_{\mathcal{H}})^2,$$

Hence  $J(u, \mathcal{L}, \mathbf{M}_{\mathcal{H}}) \leq 2J(u, \mathcal{H}, \mathbf{M}_{\mathcal{H}})$  for all  $0 < u < \delta$ . By Chernozhukov *et al.* (2014, Theorem 5.2), we have

$$\mathbb{E}[\|X_n - X_n \circ \pi_{\mathcal{H}_\delta}\|_{\mathcal{H}}] \lesssim J(\delta, \mathcal{H}, \mathbf{M}_{\mathcal{H}})\mathbf{M}_{\mathcal{H}} + \frac{\mathbf{M}_{\mathcal{H}}J^2(\delta, \mathcal{H}, \mathbf{M}_{\mathcal{H}})}{\delta^2\sqrt{n}}.$$

By Talagrand's inequality (Giné and Nickl, 2016, Theorem 3.3.9), for all  $t > 0$ ,

$$\mathbb{P}\left(\|X_n - X_n \circ \pi_{\mathcal{H}_\delta}\|_{\mathcal{H}} \gtrsim J(\delta, \mathcal{H}, \mathbf{M}_{\mathcal{H}})\mathbf{M}_{\mathcal{H}} + \frac{\mathbf{M}_{\mathcal{H}}J^2(\delta, \mathcal{H}, \mathbf{M}_{\mathcal{H}})}{\delta^2\sqrt{n}} + \delta\mathbf{M}_{\mathcal{H}}\sqrt{t} + \frac{\mathbf{M}_{\mathcal{H}}}{\sqrt{n}}t\right) \leq \exp(-t).$$

By van der Vaart and Wellner (2013, Corollary 2.2.9),

$$\mathbb{E}[\|Z_n - Z_n \circ \pi_{\mathcal{H}_\delta}\|_{\mathcal{H}}] \lesssim J(\delta, \mathcal{H}, \mathbf{M}_{\mathcal{H}})\mathbf{M}_{\mathcal{H}_\delta}.$$

By pointwise separability and a concentration inequality for Gaussian suprema, for all  $t > 0$ ,

$$\mathbb{P}\left(\|Z_n - Z_n \circ \pi_{\mathcal{H}_\delta}\|_{\mathcal{H}} \gtrsim J(\delta, \mathcal{H}, \mathbf{M}_{\mathcal{H}})\mathbf{M}_{\mathcal{H}} + \delta\mathbf{M}_{\mathcal{H}}\sqrt{t}\right) \leq \exp(-t),$$

which concludes the proof.  $\square$

### SA-II.1.5 Strong Approximation Errors

To simplify notation, in this section the parameters of  $\mathcal{H}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{X}$ , and the index  $\mathcal{C}$  is omitted whenever there is no confusion. The next lemma controls the strong approximation error for projected processes.

**Lemma SA.7.** *Suppose Assumption SA.1 holds, a dyadic expansion  $\mathcal{C}_K(\mathbb{P}_X, 1)$  is given,  $(Z_n^X(h) : h \in \mathcal{H} \cup \mathcal{E}_{K, \mathbf{M}_{\mathcal{H}}})$  is the Gaussian process constructed as in (SA-2) on a possibly enlarged probability space, and  $\mathcal{H}_\delta$  is chosen as in Section SA-II.1.4. For each  $1 \leq j \leq K$ , define the  $j$ -th level difference set*

$$\mathcal{U}_j = \cup_{0 \leq k < 2^{K-j}} (\mathcal{C}_{j-1, 2k+1} - \mathcal{C}_{j-1, 2k}).$$

Then, for all  $t > 0$ ,

$$\mathbb{P}\left[\|X_n \circ \Pi_0 - Z_n^X \circ \Pi_0\|_{\mathcal{H}_\delta} > 48\sqrt{\frac{\mathcal{R}_K(\mathcal{H}_\delta)}{n}}t + 4\sqrt{\frac{\mathbf{C}_{\mathcal{H}_\delta, K}}{n}}t\right] \leq 2\mathbf{N}_{\mathcal{H}_\delta}(\delta, \mathbf{M}_{\mathcal{H}_\delta})e^{-t},$$

where

$$\mathcal{R}_K(\mathcal{H}_\delta) = \sum_{j=1}^K \min\{\mathbf{M}_{\mathcal{H}_\delta}, \|\mathcal{U}_j\|_{\infty} \mathbf{L}_{\mathcal{H}_\delta}\} 2^{K-j} \min\left\{\sqrt{d} \sup_{\mathbf{x} \in \mathcal{X}} f_X^2(\mathbf{x}) 2^{2(K-j)} \|\mathcal{U}_j\|_{\infty} \mathbf{m}(\mathcal{U}_j) \mathbf{TV}_{\mathcal{H}_\delta}^*, \|\mathcal{U}_j\|_{\infty} \mathbf{L}_{\mathcal{H}_\delta}, \mathbf{E}_{\mathcal{H}_\delta}\right\},$$

and  $\mathbf{C}_{\mathcal{H}_\delta, K}$  is defined in Lemma SA.3. In the above display,  $f_X$  denotes the Lebesgue density of  $\mathbb{P}_X$ : if it does not exist, the term  $\sqrt{d} \sup_{\mathbf{x} \in \mathcal{X}} f_X^2(\mathbf{x}) 2^{2(K-j)} \|\mathcal{U}_j\|_{\infty} \mathbf{m}(\mathcal{U}_j) \mathbf{TV}_{\mathcal{H}_\delta}^*$  is taken to be infinity.

**Proof of Lemma SA.7.** We employ the same strategy as in the proof of Theorem 1.1 from Rio (1994), noting that incorporating the Lipschitz condition can lead to a tighter bound for strong approximation error.

A very first bound that we can obtain for is

$$\begin{aligned} \sum_{0 \leq k < 2^{K-j}} \left| \tilde{\beta}_{j,k}(h) \right| &\leq \sum_{\sum_{0 \leq k < 2^{K-j}} 2^{K-j} \int_{\mathcal{C}_{j,k}} |h(\mathbf{x})| d\mathbb{P}_X(\mathbf{x})} \\ &\leq 2^{K-j} \int_{\sqcup_{0 \leq k < 2^{K-(j-1)}} \mathcal{C}_{j-1,k}} |h(\mathbf{x})| d\mathbb{P}_X(\mathbf{x}) \leq 2^{K-j} \mathbf{E}_{\{h\}}. \end{aligned}$$

If we further assume  $\mathbb{P}_X$  admits a Lebesgue density  $f_X$ , then an analysis based on total variation of  $h$  can be done as follows. For each  $1 \leq j \leq K$ , there exists unique integers  $j_1, \dots, j_d$  such that  $0 \leq j_1 \leq \dots \leq j_d \leq j_1 + 1$  and  $\sum_{i=1}^d j_i = j$ . In particular, there exists a unique  $l = l(j) \in \{1, 2, \dots, d\}$  such that either  $l \leq d-1$  and  $j_l < j_{l+1}$  or  $l = d$  and  $j_d < j_1 + 1$ .

$$\begin{aligned} \tilde{\beta}_{j,k}(h) &= 2^{K-j} \int_{\mathcal{C}_{j-1,2k}} h(\mathbf{x}) f_X(\mathbf{x}) d\mathbf{x} - 2^{K-j} \int_{\mathcal{C}_{j-1,2k+1}} h(\mathbf{y}) f_X(\mathbf{y}) d\mathbf{y} \\ &= 2^{K-j} \int_{\mathcal{C}_{j-1,2k}} \left( h(\mathbf{x}) - \left( 2^{K-j} \int_{\mathcal{C}_{j-1,2k+1}} h(\mathbf{y}) f_X(\mathbf{y}) d\mathbf{y} \right) \right) f_X(\mathbf{x}) d\mathbf{x} \\ &= 2^{2(K-j)} \int_{\mathcal{C}_{j-1,2k}} \int_{\mathcal{C}_{j-1,2k+1}} (h(\mathbf{x}) - h(\mathbf{y})) f_X(\mathbf{x}) f_X(\mathbf{y}) d\mathbf{y} d\mathbf{x} \\ &= 2^{2(K-j)} \int_{\mathcal{C}_{j-1,2k}} \int_{\mathcal{C}_{j-1,2k+1} - \{\mathbf{x}\}} (h(\mathbf{x}) - h(\mathbf{x} + \mathbf{s})) f_X(\mathbf{x}) f_X(\mathbf{x} + \mathbf{s}) \mathbb{1}_{\mathcal{C}_{j-1,2k+1}}(\mathbf{x} + \mathbf{s}) d\mathbf{s} d\mathbf{x}. \end{aligned}$$

Since we have assumed  $f$  is bounded from above on  $\mathcal{X}$  and hence on  $\mathcal{C}_{K,0}$ , and  $\mathcal{C}_{j-1,2k+1} - \{\mathbf{x}\} \subseteq \mathcal{C}_{j-1,2k+1} - \mathcal{C}_{j-1,2k}$ ,

$$\left| \tilde{\beta}_{j,k}(h) \right| \leq 2^{2(K-j)} \int_{\mathcal{C}_{j-1,2k+1} - \mathcal{C}_{j-1,2k}} \int_{\mathcal{C}_{j-1,2k}} |h(\mathbf{x}) - h(\mathbf{x} + \mathbf{s})| f_X(\mathbf{x}) f_X(\mathbf{x} + \mathbf{s}) d\mathbf{x} d\mathbf{s}.$$

and therefore

$$\sum_{0 \leq k < 2^{K-j}} \left| \tilde{\beta}_{j,k}(h) \right| \leq 2^{2(K-j)} \int_{\mathcal{U}_j} \int_{\sqcup_{0 \leq k < 2^{K-j}} \mathcal{C}_{j-1,2k}} |h(\mathbf{x}) - h(\mathbf{x} + \mathbf{s})| f_X(\mathbf{x}) f_X(\mathbf{x} + \mathbf{s}) d\mathbf{x} d\mathbf{s},$$

where  $\mathcal{U}_j = \cup_{0 \leq k < 2^{K-j}} (\mathcal{C}_{j-1,2k+1} - \mathcal{C}_{j-1,2k})$ . Let  $(h_\ell)_{\ell \in \mathbb{N}}$  be any sequence of real-valued functions on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  such that  $h_\ell \rightarrow h$   $\mathbf{m}$ -almost surely, and are bounded by  $2M_{\mathcal{H}}$  on  $\mathcal{X}$ . Since we assumed  $M_{\mathcal{H}} < \infty$ , and  $h_\ell$  and  $h$  are bounded by  $2M_{\mathcal{H}}$  with  $\int_{\mathbb{R}^d} 2M_{\mathcal{H}} f_X(\mathbf{x}) d\mathbf{x} \leq 2M_{\mathcal{H}} < \infty$ , Dominated Convergence Theorem implies for any  $\mathbf{x} \in \mathcal{U}_j$ ,

$$\begin{aligned} \int_{\sqcup_{0 \leq k < 2^{K-j}} \mathcal{C}_{j-1,2k}} |h(\mathbf{x}) - h(\mathbf{x} + \mathbf{s})| f_X(\mathbf{x}) d\mathbf{x} &= \lim_{\ell \rightarrow \infty} \int_{\sqcup_{0 \leq k < 2^{K-j}} \mathcal{C}_{j-1,2k}} |h_\ell(\mathbf{x}) - h_\ell(\mathbf{x} + \mathbf{s})| f_X(\mathbf{x}) d\mathbf{x} \\ &= \lim_{\ell \rightarrow \infty} \int_{\sqcup_{0 \leq k < 2^{K-j}} \mathcal{C}_{j-1,2k}} \int_0^{\|\mathbf{s}\|} \|\nabla h_\ell(\mathbf{x} + t\mathbf{s}/\|\mathbf{s}\|)\| f_X(\mathbf{x}) dt d\mathbf{x} \\ &= \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \|\mathbf{s}\| \limsup_{\ell \rightarrow \infty} \text{TV}_{\{h_\ell\}}^*. \end{aligned}$$

Since the above inequality holds for all sequences  $(h_\ell)_{\ell \in \mathbb{N}}$  such that  $h_\ell \rightarrow h$   $\mathbf{m}$ -almost surely, and are bounded



by  $2\mathbf{M}_{\mathcal{H}}$  on  $\mathcal{X}$ , Definition SA.1 implies

$$\begin{aligned} \int_{\sqcup_{0 \leq k < 2^{K-j}} \mathcal{C}_{j-1, 2^k}} |h(\mathbf{x}) - h(\mathbf{x} + \mathbf{s})| f_X(\mathbf{x}) d\mathbf{x} &\leq \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \|\mathbf{s}\| \mathbf{TV}_{\{h\}}^* \\ &\leq \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \sqrt{d} \|\mathcal{U}_j\|_{\infty} \mathbf{TV}_{\{h\}}^*. \end{aligned}$$

It follows that

$$\sum_{0 \leq k < 2^{K-j}} \left| \tilde{\beta}_{j,k}(h) \right| \leq \sqrt{d} \left( \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \right)^2 2^{2(K-j)} \|\mathcal{U}_j\|_{\infty} \mathbf{m}(\mathcal{U}_j) \mathbf{TV}_{\{h\}}^*.$$

Moreover,  $|\tilde{\beta}_{j,k}(h)| \leq \min\{\mathbf{M}_{\{h\}}, \|\mathcal{U}_j\|_{\infty} \mathbf{L}_{\{h\}}\}$ , hence

$$\sup_{h \in \mathcal{H}_{\delta}} \|h\|_{\tilde{\mathcal{E}}_K}^2 = \sup_{h \in \mathcal{H}_{\delta}} \sum_{j=1}^K \sum_{0 \leq k < 2^{K-j}} |\tilde{\beta}_{j,k}(h)|^2 \leq \sup_{h \in \mathcal{H}_{\delta}} \sum_{j=1}^K \min\{\mathbf{M}_{\mathcal{H}_{\delta}}, \|\mathcal{U}_j\|_{\infty} \mathbf{L}_{\mathcal{H}_{\delta}}\} \sum_{0 \leq k < 2^{K-j}} |\tilde{\beta}_{j,k}(h)| \leq \mathcal{R}_K(\mathcal{H}_{\delta}),$$

where  $\mathcal{R}_K(\mathcal{H}_{\delta})$  is defined to be

$$\sum_{j=1}^K \min\{\mathbf{M}_{\mathcal{H}_{\delta}}, \|\mathcal{U}_j\|_{\infty} \mathbf{L}_{\mathcal{H}_{\delta}}\} 2^{K-j} \min \left\{ \sqrt{d} \left( \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \right)^2 2^{2(K-j)} \|\mathcal{U}_j\|_{\infty} \mathbf{m}(\mathcal{U}_j) \mathbf{TV}_{\mathcal{H}_{\delta}}^*, \|\mathcal{U}_j\|_{\infty} \mathbf{L}_{\mathcal{H}_{\delta}}, \mathbf{E}_{\mathcal{H}_{\delta}} \right\}.$$

Applying Lemma SA.3, for any  $h \in \mathcal{H}_{\delta}$ , for any  $t > 0$ , with probability at least  $1 - 2 \exp(-t)$ ,

$$|X_n \circ \Pi_0(h) - Z_n^X \circ \Pi_0(h)| \leq 48 \sqrt{\frac{\mathcal{R}_K(\mathcal{H}_{\delta})}{n} t} + \sqrt{\frac{\mathbf{C}_{\mathcal{H}_{\delta}, K}}{n} t}.$$

The result then follows from the fact that  $|\mathcal{H}_{\delta}| \leq \mathbf{N}_{\mathcal{H}}(\delta, \mathbf{M}_{\mathcal{H}})$  and a union bound argument.  $\square$

**Lemma SA.8.** *Suppose Assumption SA.1 holds, a quasi-dyadic expansion  $\mathcal{C}_K(\mathbb{P}_X, \rho)$  is given with  $\rho > 1$ ,  $(Z_n^X(h) : h \in \mathcal{H} \cup \Pi_0 \mathcal{H})$  is the Gaussian process constructed at Equation (SA-2) on a possibly enlarged probability space, and  $\mathcal{H}_{\delta}$  is chosen in Section SA-II.1.4. Then, for all  $t > 0$ ,*

$$\mathbb{P} \left[ \|X_n \circ \Pi_0 - Z_n^X \circ \Pi_0\|_{\mathcal{H}_{\delta}} > C_{\rho} \sqrt{\frac{\mathcal{R}_K(\mathcal{H}_{\delta})}{n} t} + C_{\rho} \sqrt{\frac{\mathbf{C}_{\mathcal{H}_{\delta}, K}}{n} t} \right] \leq 2\mathbf{N}_{\mathcal{H}}(\delta, \mathbf{M}_{\mathcal{H}}) e^{-t} + 2^K \exp(-C_{\rho} n 2^{-K}),$$

where  $C_{\rho}$  is a constant only depending on  $\rho$ ,  $\mathcal{R}_K(\mathcal{H}_{\delta})$  is defined in Lemma SA.7, and  $\mathbf{C}_{\mathcal{H}_{\delta}, K}$  is defined in Lemma SA.3.

**Proof of Lemma SA.8.** This follows from Lemma SA.5 and the fact that

$$\sup_{h \in \mathcal{H}_{\delta}} \|h\|_{\tilde{\mathcal{E}}_K}^2 \leq \mathcal{R}_K(\mathcal{H}_{\delta}), \quad h \in \mathcal{H}_{\delta},$$

from the proof of Lemma SA.7.  $\square$

### SA-II.1.6 Projection Error

To simplify notation, in this section the parameters of  $\mathcal{H}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{X}$ , and the index  $\mathcal{C}$  is omitted whenever there is no confusion. The following lemma controls the mean square projection onto piecewise constant functions.

**Lemma SA.9.** *Suppose Assumption SA.1 holds, a dyadic expansion  $\mathcal{C}_K(\mathbb{P}_X, 1)$  is given,  $(Z_n^X(h) : h \in \mathcal{H} \cup \Pi_0\mathcal{H})$  is the Gaussian process constructed as in (SA-2) on a possibly enlarged probability space, and  $\mathcal{H}_\delta$  is chosen in Section SA-II.1.4. In addition, assume  $\mathbb{P}_X$  admits a Lebesgue density  $f_X$  supported on  $\mathcal{X} \subseteq \mathbb{R}^d$ . Define quasi-dyadic variation set  $\mathcal{V} = \cup_{0 \leq k < 2^K} (\mathcal{C}_{0,k} - \mathcal{C}_{0,k})$ . Then, for all  $t > 0$ ,*

$$\begin{aligned} \mathbb{P} \left[ \|X_n - X_n \circ \Pi_0\|_{\mathcal{H}_\delta} > \sqrt{4\mathbf{V}_{\mathcal{H}_\delta} t} + \frac{4\mathbf{B}_{\mathcal{H}_\delta}}{3\sqrt{n}} t \right] &\leq 2\mathbf{N}_{\mathcal{H}_\delta}(\delta, \mathbf{M}_{\mathcal{H}_\delta}) e^{-t}, \\ \mathbb{P} \left[ \|Z_n^X - Z_n^X \circ \Pi_0\|_{\mathcal{H}_\delta} > \sqrt{4\mathbf{V}_{\mathcal{H}_\delta} t} \right] &\leq 2\mathbf{N}_{\mathcal{H}_\delta}(\delta, \mathbf{M}_{\mathcal{H}_\delta}) e^{-t}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{V}_{\mathcal{H}_\delta} &= \min\{2\mathbf{M}_{\mathcal{H}_\delta}, \mathbf{L}_{\mathcal{H}_\delta} \|\mathcal{V}\|_\infty\} \left( \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \right)^2 2^K \mathbf{m}(\mathcal{V}) \|\mathcal{V}\|_\infty \mathbf{TV}_{\mathcal{H}_\delta}^*, \\ \mathbf{B}_{\mathcal{H}_\delta} &= \min\{2\mathbf{M}_{\mathcal{H}_\delta}, \mathbf{L}_{\mathcal{H}_\delta} \|\mathcal{V}\|_\infty\}. \end{aligned}$$

In particular, if  $\mathbb{P}_X = \text{Uniform}([0, 1]^d)$  and  $\mathcal{C}_K(\mathbb{P}_X, 1) = \mathcal{A}_K(\mathbb{P}_X, 1)$ , then for all  $t > 0$ ,

$$\begin{aligned} \mathbb{P} \left[ \|X_n - X_n \circ \Pi_0\|_{\mathcal{H}_\delta} > \sqrt{4d \min\{2\mathbf{M}_{\mathcal{H}_\delta}, \mathbf{L}_{\mathcal{H}_\delta} 2^{-K}\} 2^{-K} \mathbf{TV}_{\mathcal{H}_\delta}^* t} + \frac{4 \min\{2\mathbf{M}_{\mathcal{H}_\delta}, \mathbf{L}_{\mathcal{H}_\delta} 2^{-K}\}}{3\sqrt{n}} t \right] &\leq 2\mathbf{N}_{\mathcal{H}_\delta}(\delta, \mathbf{M}_{\mathcal{H}_\delta}) e^{-t}, \\ \mathbb{P} \left[ \|Z_n^X - Z_n^X \circ \Pi_0\|_{\mathcal{H}_\delta} > \sqrt{4d \min\{2\mathbf{M}_{\mathcal{H}_\delta}, \mathbf{L}_{\mathcal{H}_\delta} 2^{-K}\} 2^{-K} \mathbf{TV}_{\mathcal{H}_\delta}^* t} \right] &\leq 2\mathbf{N}_{\mathcal{H}_\delta}(\delta, \mathbf{M}_{\mathcal{H}_\delta}) e^{-t}. \end{aligned}$$

**Proof of Lemma SA.9.** Let  $h \in \mathcal{H}$ . Then,  $|h(\mathbf{x}_i) - \Pi_0 h(\mathbf{x}_i)| \leq \min\{2\mathbf{M}_{\mathcal{H}_\delta}, \mathbf{L}_{\mathcal{H}_\delta} \|\mathcal{V}\|_\infty\} = \mathbf{B}_{\mathcal{H}_\delta}$ ,

$$\begin{aligned} \mathbb{E} [|h(\mathbf{x}_i) - \Pi_0 h(\mathbf{x}_i)|] &= \sum_{0 \leq k < 2^K} \int_{\mathcal{C}_{0,k}} \left| h(\mathbf{x}) - 2^K \int_{\mathcal{C}_{0,k}} h(\mathbf{y}) f_X(\mathbf{y}) d\mathbf{y} \right| f_X(\mathbf{x}) d\mathbf{x} \\ &\leq \sum_{0 \leq k < 2^K} 2^K \int_{\mathcal{C}_{0,k}} \int_{\mathcal{C}_{0,k}} |h(\mathbf{x}) - h(\mathbf{y})| f_X(\mathbf{y}) f_X(\mathbf{x}) d\mathbf{y} d\mathbf{x}. \end{aligned}$$

Using a change of variables  $\mathbf{s} = \mathbf{y} - \mathbf{x}$  and the fact that  $f_X$  is bounded above, we have

$$\begin{aligned} &\mathbb{E} [|h(\mathbf{x}_i) - \Pi_0 h(\mathbf{x}_i)|] \\ &\leq \sum_{0 \leq k < 2^K} 2^K \int_{\mathcal{C}_{0,k} - \mathcal{C}_{0,k}} \int_{\mathcal{C}_{0,k}} |h(\mathbf{x}) - h(\mathbf{x} + \mathbf{s})| f_X(\mathbf{x} + \mathbf{s}) f_X(\mathbf{x}) \mathbb{1}_{\mathcal{C}_{0,k}}(\mathbf{x} + \mathbf{s}) d\mathbf{x} d\mathbf{s} \\ &\leq 2^K \int_{\mathcal{V}} \int_{\mathcal{C}_{K,0}} |h(\mathbf{x}) - h(\mathbf{x} + \mathbf{s})| f_X(\mathbf{x} + \mathbf{s}) f_X(\mathbf{x}) d\mathbf{x} d\mathbf{s}. \end{aligned}$$

Let  $(h_\ell)_{\ell \in \mathbb{N}}$  be any sequence of real-valued functions on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  such that  $h_\ell \rightarrow h$   $\mathbf{m}$ -almost surely, and are bounded by  $2\mathbf{M}_{\mathcal{H}}$  on  $\mathcal{X}$ . Since we assumed  $\mathbf{M}_{\mathcal{H}} < \infty$ , and  $h_\ell$  and  $h$  are bounded by  $2\mathbf{M}_{\mathcal{H}}$ , by Dominated

Convergence Theorem we have that

$$\begin{aligned}
\int_{\mathcal{C}_{K,0}} |h(\mathbf{x}) - h(\mathbf{x} + \mathbf{s})| f_X(\mathbf{x}) d\mathbf{x} &= \lim_{\ell \rightarrow \infty} \int_{\mathcal{C}_{K,0}} |h_\ell(\mathbf{x}) - h_\ell(\mathbf{x} + \mathbf{s})| f_X(\mathbf{x}) d\mathbf{x} \\
&\leq \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \cdot \lim_{\ell \rightarrow \infty} \int_{\mathcal{X}} \int_0^{\|\mathbf{s}\|} \|\nabla h_\ell(\mathbf{x} + t\mathbf{s}/\|\mathbf{s}\|)\| dt d\mathbf{x} \\
&\leq \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \cdot \int_0^{\|\mathbf{s}\|} \lim_{\ell \rightarrow \infty} \int_{\mathcal{X}} \|\nabla h_\ell(\mathbf{x} + t\mathbf{s}/\|\mathbf{s}\|)\| d\mathbf{x} dt \\
&\leq \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \cdot \|\mathbf{s}\| \limsup_{\ell \rightarrow \infty} \text{TV}_{\{h_\ell\}}.
\end{aligned}$$

Since this holds for any sequence  $(h_\ell)_{\ell \in \mathbb{N}}$   $h_\ell \rightarrow h$   $\mathbf{m}$ -almost surely, and are bounded by  $2M_{\mathcal{H}}$  on  $\mathcal{X}$ , hence  $\int_{\mathcal{X}} |h(\mathbf{x}) - h(\mathbf{x} + \mathbf{s})| d\mathbf{x} \leq \|\mathbf{s}\| \text{TV}_{\{h\}}^*$ . It follows that

$$\mathbb{E}[|h(\mathbf{x}_i) - \Pi_0 h(\mathbf{x}_i)|] \leq \left( \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \right)^2 2^K \mathbf{m}(\mathcal{V}) \|\mathcal{V}\|_\infty \text{TV}_{\{h\}}^*,$$

and

$$\mathbb{V}[h(\mathbf{x}_i) - \Pi_0 h(\mathbf{x}_i)] \leq \min\{2M_{\mathcal{H}_\delta}, L_{\mathcal{H}_\delta} \|\mathcal{V}\|_\infty\} \left( \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \right)^2 2^K \mathbf{m}(\mathcal{V}) \|\mathcal{V}\|_\infty \text{TV}_{\mathcal{H}_\delta}^* = V_{\mathcal{H}_\delta},$$

for all  $h \in \mathcal{H}_\delta$ . Then, by Bernstein inequality, for any  $t > 0$ ,

$$\mathbb{P}(|X_n(h) - X_n(\Pi_0 h)| \geq t) \leq 2 \exp\left(-\frac{\frac{1}{2}t^2 n}{nV_{\mathcal{H}_\delta} + \frac{1}{3}B_{\mathcal{H}_\delta} t \sqrt{n}}\right) \leq 2 \exp\left(-\frac{1}{2} \min\left\{\frac{\frac{1}{2}t^2 n}{nV_{\mathcal{H}_\delta}}, \frac{\frac{1}{2}t^2 n}{\frac{1}{3}B_{\mathcal{H}_\delta} t \sqrt{n}}\right\}\right).$$

Set  $u = \frac{1}{2} \min\left\{\frac{\frac{1}{2}t^2 n}{nV_{\mathcal{H}_\delta}}, \frac{\frac{1}{2}t^2 n}{\frac{1}{3}B_{\mathcal{H}_\delta} t \sqrt{n}}\right\} > 0$ , then either  $t = 2\sqrt{V_{\mathcal{H}_\delta}}\sqrt{u}$  or  $t = \frac{4}{3}\frac{B_{\mathcal{H}_\delta}}{\sqrt{n}}u$ . Hence  $t \leq 2\sqrt{V_{\mathcal{H}_\delta}}\sqrt{u} + \frac{4}{3}\frac{B_{\mathcal{H}_\delta}}{\sqrt{n}}u$ . For any  $u > 0$ ,  $\mathbb{P}(|X_n(h) - X_n(\Pi_0 h)| \geq 2\sqrt{V_{\mathcal{H}_\delta}}\sqrt{u} + \frac{4}{3}\frac{B_{\mathcal{H}_\delta}}{\sqrt{n}}u) \leq 2 \exp(-u)$ . The result for  $\|X_n - X_n \circ \Pi_0\|_{\mathcal{H}_\delta}$  then follows from a union bound. The result for  $\|Z_n - Z_n \circ \Pi_0\|_{\mathcal{H}_\delta}$  follows from the fact that  $Z_n(h) - Z_n(\Pi_0 h)$  is a mean-zero Gaussian with variance  $\mathbb{V}[X_n(h) - X_n(\Pi_0 h)]$  and a union bound argument.  $\square$

## SA-II.2 Surrogate Measure and Normalizing Transformation

This section studies the properties of the surrogate measure  $\mathbb{Q}_{\mathcal{H}}$  and normalizing transformation  $\phi_{\mathcal{H}}$  introduced in condition (ii) of Theorem 1. The following lemma characterizes the connections between the original and the transformed parameters of  $\mathcal{H}$  (Definitions 4 to 12) when deploying  $\mathbb{Q}_{\mathcal{H}}$  and  $\phi_{\mathcal{H}}$ .

**Lemma SA.10.** *Suppose 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  $\mathbb{Q}_{\mathcal{H}} = \mathbf{m} \circ \phi_{\mathcal{H}}$ , where the normalizing transformation  $\phi_{\mathcal{H}} : \mathcal{Q}_{\mathcal{H}} \mapsto [0, 1]^d$  is a diffeomorphism.

Let  $\tilde{\mathcal{H}} = \{h \circ \phi_{\mathcal{H}}^{-1} : h \in \mathcal{H}\}$ . Then,

$$\begin{aligned}
\mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d} &= \mathbf{M}_{\mathcal{H}, \mathcal{Q}_{\mathcal{H}}}, & \mathbf{E}_{\tilde{\mathcal{H}}, [0,1]^d} &= \mathbf{E}_{\mathcal{H}, \mathcal{Q}_{\mathcal{H}}}, \\
\mathbf{N}_{\tilde{\mathcal{H}}, [0,1]^d}(\varepsilon, \mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d}) &= \mathbf{N}_{\mathcal{H}, \mathcal{Q}_{\mathcal{H}}}(\varepsilon, \mathbf{M}_{\mathcal{H}, \mathcal{Q}_{\mathcal{H}}}), & \varepsilon &\in (0, 1), \\
\mathbf{L}_{\tilde{\mathcal{H}}, [0,1]^d} &\leq \mathbf{c}_2 \mathbf{L}_{\mathcal{H}, \mathcal{Q}_{\mathcal{H}}}, & \mathbf{c}_2 &= \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} \frac{1}{\sigma_d(\nabla \phi_{\mathcal{H}}(\mathbf{x}))}, \\
\mathbf{TV}_{\tilde{\mathcal{H}}, [0,1]^d}^* &\leq d^{-1} \mathbf{c}_1 \mathbf{TV}_{\mathcal{H}, \mathcal{Q}_{\mathcal{H}}}, & \mathbf{c}_1 &= d \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} \prod_{j=1}^{d-1} \sigma_j(\nabla \phi_{\mathcal{H}}(\mathbf{x})), \\
\mathbf{K}_{\tilde{\mathcal{H}}, [0,1]^d}^* &\leq d^{-1/2} \mathbf{c}_3 \mathbf{K}_{\mathcal{H}, \mathcal{Q}_{\mathcal{H}}}, & \mathbf{c}_3 &= 2^{d-1} d^{d/2-1} \mathbf{c}_1 \mathbf{c}_2^{d-1}.
\end{aligned}$$

**Proof of Lemma SA.10.** The first three identities are self-evident. Consider next the relation between  $\mathbf{L}_{\tilde{\mathcal{H}}, [0,1]^d}$  and  $\mathbf{L}_{\mathcal{H}, \mathcal{Q}_{\mathcal{H}}}$ : for any  $h \in \mathcal{H}$ , using a change of variables and the differentiability of  $\phi_{\mathcal{H}}$ ,

$$\begin{aligned}
\mathbf{L}_{\{h \circ \phi_{\mathcal{H}}^{-1}\}, [0,1]^d} &= \sup_{\mathbf{u}, \mathbf{u}' \in [0,1]^d} \frac{|h \circ \phi_{\mathcal{H}}^{-1}(\mathbf{u}) - h \circ \phi_{\mathcal{H}}^{-1}(\mathbf{u}')|}{\|\mathbf{u} - \mathbf{u}'\|} \\
&\leq \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{Q}_{\mathcal{H}}} \frac{|h(\mathbf{x}) - h(\mathbf{x}')|}{\|\mathbf{x} - \mathbf{x}'\|} \frac{\|\mathbf{x} - \mathbf{x}'\|}{\|\phi_{\mathcal{H}}(\mathbf{x}) - \phi_{\mathcal{H}}(\mathbf{x}')\|} \\
&\leq \mathbf{L}_{\{h\}, \mathcal{Q}_{\mathcal{H}}} \sup_{\mathbf{u}, \mathbf{u}' \in [0,1]^d} \frac{|\phi_{\mathcal{H}}^{-1}(\mathbf{u}) - \phi_{\mathcal{H}}^{-1}(\mathbf{u}')|}{\|\mathbf{u} - \mathbf{u}'\|} \\
&\leq \mathbf{L}_{\{h\}, \mathcal{Q}_{\mathcal{H}}} \sup_{\mathbf{z} \in [0,1]^d} \sigma_1(\nabla \phi_{\mathcal{H}}^{-1}(\mathbf{z})) \\
&= \mathbf{L}_{\{h\}, \mathcal{Q}_{\mathcal{H}}} \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} \sigma_d(\nabla \phi_{\mathcal{H}}(\mathbf{x}))^{-1},
\end{aligned}$$

and the result follows.

Now consider the relation between  $\mathbf{TV}_{\tilde{\mathcal{H}}, [0,1]^d}$  and  $\mathbf{TV}_{\mathcal{H}, \mathcal{Q}_{\mathcal{H}}}$ . First suppose all functions in  $\mathcal{H}$  are differentiable, an integration by parts based on the definition of uniform total variation (Definition 5) and a change of variables calculation gives

$$\begin{aligned}
\mathbf{TV}_{\{h \circ \phi_{\mathcal{H}}^{-1}\}, [0,1]^d} &= \sup_{\varphi \in \mathcal{D}_d([0,1]^d)} \int_{[0,1]^d} h \circ \phi_{\mathcal{H}}^{-1}(\mathbf{x}) \operatorname{div}(\varphi)(\mathbf{x}) d\mathbf{x} / \|\varphi\|_2 \|\infty \\
&= \int_{\mathbf{u} \in [0,1]^d} \|\nabla(h \circ \phi_{\mathcal{H}}^{-1})(\mathbf{u})\| d\mathbf{u} \\
&= \int_{\mathbf{u} \in [0,1]^d} \|\nabla \phi_{\mathcal{H}}^{-1}(\mathbf{u})^\top \nabla h(\phi_{\mathcal{H}}^{-1}(\mathbf{u}))\| d\mathbf{u} \\
&= \int_{\mathcal{Q}_{\mathcal{H}}} \|\nabla \phi_{\mathcal{H}}^{-1}(\phi_{\mathcal{H}}(\mathbf{x}))^\top \nabla h(\mathbf{x})\| \cdot |\det(\nabla \phi_{\mathcal{H}}(\mathbf{x}))| d\mathbf{x} \\
&\leq \int_{\mathcal{Q}_{\mathcal{H}}} \|\nabla h(\mathbf{x})\| d\mathbf{x} \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} |\det(\nabla \phi_{\mathcal{H}}(\mathbf{x}))| \cdot \|\nabla \phi_{\mathcal{H}}^{-1}(\phi_{\mathcal{H}}(\mathbf{x}))\| \\
&\leq \mathbf{c}_1 \mathbf{TV}_{\{h\}, \mathcal{Q}_{\mathcal{H}}},
\end{aligned}$$

where in the last line we have used  $|\det(\nabla \phi_{\mathcal{H}}(\mathbf{x}))| = \prod_{j=1}^d \sigma_j(\nabla \phi_{\mathcal{H}}(\mathbf{x}))$ , and since  $\phi_{\mathcal{H}}$  is a diffeomorphism,  $\|\nabla \phi_{\mathcal{H}}^{-1}(\phi_{\mathcal{H}}(\mathbf{x}))\| = \sigma_1(\nabla \phi_{\mathcal{H}}^{-1}(\phi_{\mathcal{H}}(\mathbf{x}))) = \sigma_d(\nabla \phi_{\mathcal{H}}(\mathbf{x}))^{-1}$ . Now consider  $\mathcal{H}$  which contains possibly non-differentiable functions. Take  $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$  to be any smooth function with compact support such that  $\int_{\mathbb{R}^d} \psi(\mathbf{z}) d\mathbf{z} = 1$ , and take  $\psi_\varepsilon(\cdot) = \varepsilon^{-d} \psi(\cdot/\varepsilon)$ . For each  $\ell \in \mathbb{N}$ , define  $h_\ell = h * \psi_{\varepsilon_\ell}$ , where  $(\varepsilon_\ell)_{\ell \in \mathbb{N}}$  is a sequence

of non-increasing real positive numbers converging to zero with  $\varepsilon_1$  small enough. Then

$$\mathbf{TV}_{\tilde{\mathcal{H}}, [0,1]^d}^* \leq \sup_{h \in \mathcal{H}} \limsup_{\ell \rightarrow \infty} \mathbf{TV}_{\{h_\ell \circ \phi_{\mathcal{H}}^{-1}\}, [0,1]^d} = \sup_{h \in \mathcal{H}} \limsup_{\ell \rightarrow \infty} c_1 \mathbf{TV}_{\{h_\ell\}, \mathcal{Q}_{\mathcal{H}}} \leq c_1 \mathbf{TV}_{\mathcal{H}, \mathcal{Q}_{\mathcal{H}}},$$

where the first inequality is due to  $(h_\ell)_{\ell \in \mathbb{N}}$  being a particular sequence satisfying Definition SA.1, the second inequality from Lemma SA.10, the third inequality due to  $\mathbf{TV}_{\{h^* \psi\}, \mathcal{Q}_{\mathcal{H}}} \leq \mathbf{TV}_{\{h\}, \mathcal{Q}_{\mathcal{H}}}$  for any smooth  $\psi$ .

Moreover, let  $\mathcal{C} \subseteq \mathbb{R}^d$  be a cube with edges of length  $\mathbf{a}$  parallel to the coordinate axes. Then,  $\phi_{\mathcal{H}}^{-1}(\mathcal{C})$  is contained in another cube  $\mathcal{C}'$  with edges of length at most  $2\sqrt{d} \sup_{\mathbf{x} \in [0,1]^d} \|\nabla \phi_{\mathcal{H}}^{-1}(\mathbf{x})\| \mathbf{a}$ . Again, we first assume that each  $h \in \mathcal{H}$  is differentiable. Using a change of variables for the total variation calculation and the definition of  $K_{\{h\}, \mathcal{Q}_{\mathcal{H}}}$  (Definition 5), for any  $h \in \mathcal{H}$ ,

$$\begin{aligned} \mathbf{TV}_{\{h \circ \phi_{\mathcal{H}}^{-1}\}, \mathcal{C}} &= \int_{\mathcal{C}} \|\nabla(h \circ \phi_{\mathcal{H}}^{-1})(\mathbf{u})\| d\mathbf{u} \\ &\leq \int_{\mathcal{C}'} \|\nabla(h \circ \phi_{\mathcal{H}}^{-1})(\phi_{\mathcal{H}}(\mathbf{x}))\| \det(\nabla \phi_{\mathcal{H}}(\mathbf{x})) d\mathbf{x} \\ &\leq \int_{\mathcal{C}'} \|\nabla h(\mathbf{x})\| d\mathbf{x} \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} |\det(\nabla \phi_{\mathcal{H}}(\mathbf{x}))| \|\nabla \phi_{\mathcal{H}}^{-1}(\phi_{\mathcal{H}}(\mathbf{x}))\| \\ &\leq K_{\{h\}, \mathcal{Q}_{\mathcal{H}}} \left(2\sqrt{d} \sup_{\mathbf{x} \in [0,1]^d} \|\nabla \phi_{\mathcal{H}}^{-1}(\mathbf{x})\| \mathbf{a}\right)^{d-1} \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} |\det(\nabla \phi_{\mathcal{H}}(\mathbf{x}))| \|\nabla \phi_{\mathcal{H}}^{-1}(\phi_{\mathcal{H}}(\mathbf{x}))\| \\ &= d^{-1} (2\sqrt{d})^{d-1} c_1 c_2^{d-1} K_{\{h\}, \mathcal{Q}_{\mathcal{H}}} \mathbf{a}^{d-1} \\ &= d^{-1/2} c_3 K_{\{h\}, \mathcal{Q}_{\mathcal{H}}} \mathbf{a}^{d-1}, \end{aligned}$$

which implies

$$K_{\{\tilde{h}\}, [0,1]^d} \leq d^{-1/2} c_3 K_{\{h\}, \mathcal{Q}_{\mathcal{H}}}.$$

By similar smoothing arguments as for the TV terms, we can also show that  $K_{\tilde{\mathcal{H}}, [0,1]^d}^* \leq d^{-1/2} c_3 K_{\mathcal{H}, [0,1]^d}$  even when  $\mathcal{H}$  contains possibly non-differentiable functions.  $\square$

**Lemma SA.11.** *We recap the statements in Section 3.1 and present their proofs.*

- **Case 1: Uniform on Rectangle.** Suppose that  $\mathbf{x}_i \sim \text{Uniform}(\mathcal{X})$  with  $\mathcal{X} = \times_{l=1}^d [\mathbf{a}_l, \mathbf{b}_l]$ , where  $-\infty < \mathbf{a}_l < \mathbf{b}_l < \infty$ ,  $l = 1, 2, \dots, d$ . Setting  $\mathcal{Q}_{\mathcal{H}} = \mathbb{P}_X$ , a valid normalizing transformation is  $\phi_{\mathcal{H}}(x_1, \dots, x_d) = ((\mathbf{b}_1 - \mathbf{a}_1)^{-1}(x_1 - \mathbf{a}_1), \dots, (\mathbf{b}_d - \mathbf{a}_d)^{-1}(x_d - \mathbf{a}_d))$ , which verifies assumption (ii) in Theorem 1. In this case,  $c_1 = d \max_{1 \leq l \leq d} |\mathbf{b}_l - \mathbf{a}_l| \prod_{l=1}^d |\mathbf{b}_l - \mathbf{a}_l|^{-1}$ ,  $c_2 = \max_{1 \leq l \leq d} |\mathbf{b}_l - \mathbf{a}_l|$  and  $c_3 = 2^{d-1} d^{d/2} \max_{1 \leq l \leq d} |\mathbf{b}_l - \mathbf{a}_l|^d \prod_{l=1}^d |\mathbf{b}_l - \mathbf{a}_l|^{-1}$ .
- **Case 2: Rectangular  $\mathcal{Q}_{\mathcal{H}}$ .** Suppose that  $\mathcal{Q}_{\mathcal{H}}$  admits a Lebesgue density  $f_Q$  supported on  $\mathcal{Q}_{\mathcal{H}} = \times_{l=1}^d [\mathbf{a}_l, \mathbf{b}_l]$ ,  $-\infty \leq \mathbf{a}_l < \mathbf{b}_l \leq \infty$ . Then, the Rosenblatt transformation  $\phi_{\mathcal{H}} = T_{\mathcal{Q}_{\mathcal{H}}}$  is a normalizing transformation, and we obtain

$$\begin{aligned} c_1 &= d \sup_{\mathbf{u} \in \mathcal{Q}_{\mathcal{H}}} \frac{f_Q(\mathbf{u})}{\min\{f_{Q,1}(u_1), f_{Q,2|1}(u_2|u_1), \dots, f_{Q,d|d-1}(u_d|u_1, \dots, u_{d-1})\}}, \\ c_2 &= \sup_{\mathbf{u} \in \mathcal{Q}_{\mathcal{H}}} \frac{1}{\min\{f_{Q,1}(u_1), f_{Q,2|1}(u_2|u_1), \dots, f_{Q,d|d-1}(u_d|u_1, \dots, u_{d-1})\}}, \end{aligned}$$

and  $c_3 = 2^{d-1} d^{d/2-1} c_1 c_2^{d-1}$ .

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

(a) Suppose  $\mathcal{Q}_{\mathcal{H}} = \times_{l=1}^d [a_l, b_l]$  is bounded. Then,

$$c_1 \leq d \frac{\bar{f}_Q^2}{\underline{f}_Q} \bar{\mathcal{Q}}_{\mathcal{H}} \quad \text{and} \quad c_2 \leq \frac{\bar{f}_Q}{\underline{f}_Q} \bar{\mathcal{Q}}_{\mathcal{H}}.$$

(b) Suppose  $\mathcal{Q}_{\mathcal{H}} = \times_{l=1}^d [a_l, b_l]$  is unbounded. To fix ideas, let  $\mathbf{x}_i \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Then, setting  $\mathcal{Q}_{\mathcal{H}} = \mathbb{P}_X$  and  $\phi_{\mathcal{H}} = T_{\mathbb{P}_X}$  also gives a valid normalizing transformation, with

$$\begin{aligned} c_1 &\leq d \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} \max\{f_{X,1}(x_1), f_{X,2|1}(x_2|x_1), \dots, f_{X,d|1\dots d-1}(x_d|x_{-d})\}^{d-1} \\ &\leq d \min_{1 \leq k \leq d} \{\boldsymbol{\Sigma}_{k,k} - \boldsymbol{\Sigma}_{k,1:k-1} \boldsymbol{\Sigma}_{1:k-1,1:k-1}^{-1} \boldsymbol{\Sigma}_{1:k-1,k}\}^{-(d-1)/2} \end{aligned}$$

bounded, but  $c_2$  (and hence  $c_3$ ) unbounded.

- **Case 3: Non-Rectangular  $\mathcal{Q}_{\mathcal{H}}$ .** Suppose that  $\mathcal{Q}_{\mathcal{H}}$  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  $\phi_{\mathcal{H}} = T_{\mathcal{Q}_{\mathcal{H}} \circ \chi^{-1}} \circ \chi$  gives a valid normalizing transformation, with

$$c_1 \leq d \frac{\bar{f}_Q^2}{\underline{f}_Q} \mathbf{S}_{\chi} \quad \text{and} \quad c_2 \leq \frac{\bar{f}_Q}{\underline{f}_Q} \mathbf{S}_{\chi},$$

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

**Proof of Lemma SA.11.** We consider the three cases separately.

**Case 1: Uniform on Rectangle.** For every  $\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}$ , the singular values of  $\nabla \phi_{\mathcal{H}}(\mathbf{x})$  are  $(b_1 - a_1)^{-1}, \dots, (b_d - a_d)^{-1}$ . The values of  $c_1$  and  $c_2$  (and hence  $c_3$ ) then follow.

**Case 2: Rectangular  $\mathcal{Q}_{\mathcal{H}}$ .** We start with a proof for a general result for  $c_1, c_2, c_3$ , and then prove upper bounds for (a) and (b).

1. The General Case. Since  $\mathcal{Q}$  has a Lebesgue density  $f_Q$ ,

$$\nabla T_{\mathcal{Q}}(\mathbf{x}) = \begin{bmatrix} f_{Q,1}(x_1) & 0 & \cdots & 0 \\ * & f_{Q,2|1}(x_2|x_1) & \cdots & 0 \\ * & * & \vdots & 0 \\ * & * & \cdots & f_{Q,d|1,\dots,d-1}(x_d|x_1, \dots, x_{d-1}) \end{bmatrix}, \quad \mathbf{x} \in \mathcal{Q}_{\mathcal{H}}.$$

Because the singular values of  $\nabla\phi_{\mathcal{H}}(\mathbf{x}) = \nabla T_{\mathbb{Q}}(\mathbf{x})$  are the values on the diagonal,

$$\begin{aligned} \mathbf{c}_1 &= d \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} \frac{f_{\mathcal{Q}}(\mathbf{x})}{\min\{f_{Q,1}(x_1), f_{Q,2|1}(x_2|x_1), \dots, f_{Q,d|d-d}(x_d|x_{d-1})\}}, \\ \mathbf{c}_2 &= \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} \max\{f_{Q,1}(x_1)^{-1}, f_{Q,2|1}(x_2|x_1)^{-1}, \dots, f_{Q,d|d-d}(x_d|x_{d-1})^{-1}\}. \end{aligned}$$

2. Case (a):  $\mathcal{Q}_{\mathcal{H}} = \times_{l=1}^d [a_l, b_l]$  is bounded. Since we assumed the existence of an  $\mathcal{Q}_{\mathcal{H}}$  that is compact and  $\inf_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} f_{\mathcal{Q}}(\mathbf{x}) > 0$ , integrating on the rectangle gives

$$\frac{\bar{f}_{\mathcal{Q}}}{\underline{f}_{\mathcal{Q}}} \frac{1}{\bar{L}} \leq f_{Q,k|1, \dots, k-1}(x_k|x_1, \dots, x_{k-1}) = \frac{\int_{\prod_{l=k+1}^d [a_l, b_l]} f_{\mathcal{Q}}(x_1, \dots, x_k, \mathbf{z}) d\mathbf{z}}{\int_{\prod_{l=k}^d [a_l, b_l]} f_{\mathcal{Q}}(x_1, \dots, x_{k-1}, \mathbf{u}) d\mathbf{u}} \leq \frac{\bar{f}_{\mathcal{Q}}}{\underline{f}_{\mathcal{Q}}} \frac{1}{\underline{L}},$$

where  $\bar{L} = \max_{1 \leq l \leq d} (b_l - a_l)$  and  $\underline{L} = \min_{1 \leq l \leq d} (b_l - a_l)$ . Plugging in the generic bounds for  $\mathbf{c}_1$  and  $\mathbf{c}_2$ ,

$$\mathbf{c}_1 \leq d \left( \frac{\bar{f}_{\mathcal{Q}}}{\underline{f}_{\mathcal{Q}}} \max_{1 \leq k \leq d} \frac{1}{b_k - a_k} \right)^{d-1} \quad \text{and} \quad \mathbf{c}_2 \leq \frac{\bar{f}_{\mathcal{Q}}}{\underline{f}_{\mathcal{Q}}} \max_{1 \leq k \leq d} |b_k - a_k|.$$

3. Case (b):  $\mathbf{x}_i \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . The bound on  $\mathbf{c}_1$  follows from properties of the conditional distribution of multivariate Gaussian distribution. Since  $\inf_{\mathbf{x} \in \mathbb{R}^k} f_{k|1, \dots, k-1}(x_k|x_1, \dots, x_{k-1}) = 0$  for  $1 \leq k \leq d$ ,  $\mathbf{c}_2$  (and hence  $\mathbf{c}_3$ ) are unbounded.

**Case 3: Non-rectangular  $\mathcal{Q}_{\mathcal{H}}$ .** Since both  $T_{\mathbb{Q}_{\mathcal{H}}}$  and  $\chi$  are diffeomorphisms, we can use chain rule to get,

$$\begin{aligned} \mathbf{c}_1 &= \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} \prod_{j=1}^{d-1} \sigma_j(\nabla\phi_{\mathcal{H}}(\mathbf{x})) \\ &= \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} \det(\nabla\phi_{\mathcal{H}}(\mathbf{x})) \|\nabla\phi_{\mathcal{H}}^{-1}(\phi_{\mathcal{H}}(\mathbf{x}))\| \\ &\leq \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} \det(\nabla T_{\mathbb{Q}_{\mathcal{H}}}(\chi(\mathbf{x}))) \det(\nabla\chi(\mathbf{x})) \|\nabla\chi^{-1}(\chi(\mathbf{x}))\|_2 \|\nabla T_{\mathbb{Q}_{\mathcal{H}}}^{-1}(\phi_{\mathcal{H}}(\mathbf{x}))\|_2. \end{aligned}$$

Take  $\mathbf{w}_i = \chi(\mathbf{x}_i)$ , and denote by  $f_W$  the density of  $\mathbf{w}_i$ . Then

$$\nabla T_{\mathbb{Q}_{\mathcal{H}}}(x_1, \dots, x_d) = \begin{bmatrix} f_{W_1}(x_1) & 0 & \dots & 0 \\ * & f_{W_2|W_1}(x_2|x_1) & \dots & 0 \\ \vdots & \vdots & & \vdots \\ * & * & \dots & f_{W_d|W_1, \dots, W_{d-1}}(x_d|x_1, \dots, x_{d-1}) \end{bmatrix},$$

where \* denotes values that won't affect determinant or operator norm of the matrix  $\nabla T_{\mathbb{Q}_{\mathcal{H}}}$ . Hence,

$$\det(\nabla T_{\mathbb{Q}_{\mathcal{H}}}(\chi(\mathbf{x}))) = f_W(\chi(\mathbf{x})) = f_X(\mathbf{x}) |\det(\nabla\chi(\mathbf{x}))|^{-1}$$



and

$$\begin{aligned} \|\nabla T_{\mathcal{Q}_{\mathcal{H}}}^{-1}(\phi_{\mathcal{H}}(\mathbf{x}))\|_2 &= \sigma_d(\nabla T_{\mathcal{Q}_{\mathcal{H}}}(\chi(\mathbf{x})))^{-1} \leq \frac{\sup_{\mathbf{w} \in [0,1]^d} f_W(\mathbf{w})}{\inf_{\mathbf{w} \in [0,1]^d} f_W(\mathbf{w})} \\ &\leq \frac{\sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} f_X(\mathbf{x})}{\inf_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} f_X(\mathbf{x})} \cdot \frac{\sup_{\mathbf{x} \in [0,1]^d} |\det(\nabla \chi^{-1}(\mathbf{x}))|}{\inf_{\mathbf{x} \in [0,1]^d} |\det(\nabla \chi^{-1}(\mathbf{x}))|}. \end{aligned}$$

Putting together, we have

$$\mathbf{c}_1 \leq \frac{\bar{f}_X^2}{\underline{f}_X} \mathbf{S}_\chi,$$

with  $\bar{f}_X = \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} f_X(\mathbf{x})$ ,  $\underline{f}_X = \inf_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} f_X(\mathbf{x})$ ,  $\mathbf{S}_\chi = \frac{\sup_{\mathbf{x} \in [0,1]^d} |\det(\nabla \chi^{-1}(\mathbf{x}))|}{\inf_{\mathbf{x} \in [0,1]^d} |\det(\nabla \chi^{-1}(\mathbf{x}))|} \|\nabla \chi^{-1}\|_2 \|\cdot\|_\infty$ . Also,

$$\begin{aligned} \mathbf{c}_2 &= \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} \|\nabla \phi_{\mathcal{H}}^{-1}(\phi_{\mathcal{H}}(\mathbf{x}))\|_2 \leq \sup_{\mathbf{u} \in [0,1]^d} \|\nabla \chi^{-1}(\mathbf{u})\|_2 \sup_{\mathbf{u} \in [0,1]^d} \|\nabla T_{\mathcal{Q}_{\mathcal{H}}}^{-1}(\mathbf{u})\|_2 \\ &\leq \sup_{\mathbf{u} \in [0,1]^d} \|\nabla \chi^{-1}(\mathbf{u})\|_2 \frac{\sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} f_X(\mathbf{x})}{\inf_{\mathbf{x} \in \mathcal{Q}_{\mathcal{H}}} f_X(\mathbf{x})} \cdot \frac{\sup_{\mathbf{x} \in [0,1]^d} |\det(\nabla \chi^{-1}(\mathbf{x}))|}{\inf_{\mathbf{x} \in [0,1]^d} |\det(\nabla \chi^{-1}(\mathbf{x}))|} \leq \frac{\bar{f}_X}{\underline{f}_X} \mathbf{S}_\chi. \end{aligned}$$

This completes the proof.  $\square$

### SA-II.3 General Result: Proof of Theorem 1

First, we make a reduction through the surrogate measure and normalizing transformation. We want to show that under assumption (ii) in Theorem 1, the empirical process  $(X_n(h) : h \in \mathcal{H})$  can be written as an empirical process based on i.i.d Uniform( $[0,1]^d$ ) random variables. Let  $\mathcal{Z}_{\mathcal{H}} = \mathcal{X} \cap \text{Supp}(\mathcal{H})$ . Since  $\mathcal{Q}_{\mathcal{H}} = \mathbf{m} \circ \phi_{\mathcal{H}}$  by Assumption (ii) in Theorem 1, and  $\mathcal{Q}_{\mathcal{H}}|_{\mathcal{Z}_{\mathcal{H}}} = \mathbb{P}_X|_{\mathcal{Z}_{\mathcal{H}}}$ ,

$$\mathbb{P}_X|_{\mathcal{Z}_{\mathcal{H}}} = \mathbf{m} \circ \phi_{\mathcal{H}}|_{\mathcal{Z}_{\mathcal{H}}}.$$

To define the Uniform( $[0,1]^d$ ) random variables on the probability space that  $\mathbf{x}_i$ 's live in, we define a joint probability measure  $\mathbb{O}$  on  $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^{2d}))$  such that for all  $A \in \mathcal{B}(\mathbb{R}^{2d})$ :

$$\begin{aligned} \mathbb{O}(A \cap (\mathcal{Z}_{\mathcal{H}} \times \mathcal{Z}_{\mathcal{H}})) &= \mathbb{P}_X(\Pi_{1,d}(A \cap \{(\mathbf{x}, \mathbf{x}) : \mathbf{x} \in \mathcal{Z}_{\mathcal{H}}\})), \\ \mathbb{O}(A \cap (\mathcal{Z}_{\mathcal{H}} \times \mathcal{Z}_{\mathcal{H}}^c)) &= \mathbb{O}(A \cap (\mathcal{Z}_{\mathcal{H}}^c \times \mathcal{Z}_{\mathcal{H}})) = 0, \\ \mathbb{O}(A \cap (\mathcal{Z}_{\mathcal{H}}^c \times \mathcal{Z}_{\mathcal{H}}^c)) &= \int_{\mathcal{Z}_{\mathcal{H}}^c \cap \Pi_{d+1,2d}(A)} \frac{\mathbb{P}_X(A^{\mathbf{u}} \cap \mathcal{Z}_{\mathcal{H}}^c)}{\mathbb{P}_X(\mathcal{Z}_{\mathcal{H}}^c)} d(\mathbf{m} \circ \phi_{\mathcal{H}})(\mathbf{u}), \end{aligned}$$

where  $\Pi_{1,d}(A) = \{\mathbf{x} \in \mathbb{R}^d : (\mathbf{x}, \mathbf{u}) \in A \text{ for some } \mathbf{u} \in \mathbb{R}^d\}$ ,  $\Pi_{d+1,2d}(A) = \{\mathbf{u} \in \mathbb{R}^d : (\mathbf{x}, \mathbf{u}) \in A \text{ for some } \mathbf{x} \in \mathbb{R}^d\}$ , and  $A^{\mathbf{u}} = \{\mathbf{x} \in \mathbb{R}^d : (\mathbf{x}, \mathbf{u}) \in A\}$ . See Figure 1 for a graphical illustration.

Then we can check that (i) the marginals of  $\mathbb{O}$  are  $\mathbb{P}_X$  and  $\mathbf{m} \circ \phi_{\mathcal{H}}$ , respectively; (ii)  $\mathbb{O}|_{\mathcal{Z}_{\mathcal{H}} \times \mathbb{R}^d \cup \mathbb{R}^d \times \mathcal{Z}_{\mathcal{H}}}$  is supported on  $\{(\mathbf{x}, \mathbf{x}) : \mathbf{x} \in \mathcal{Z}_{\mathcal{H}}\}$ . By Skorohod embedding (Dudley, 2014, Lemma 3.35), on a possibly enlarged probability space, there exists a  $\mathbf{u}_i, 1 \leq i \leq n$  i.i.d. Uniform( $[0,1]^d$ ) such that  $(\mathbf{x}_i, \phi_{\mathcal{H}}^{-1}(\mathbf{u}_i))$  has joint law  $\mathbb{O}$ . In particular, if  $\mathbf{x}_i \in \mathcal{Z}_{\mathcal{H}}$ , then  $\mathbf{x}_i = \phi_{\mathcal{H}}^{-1}(\mathbf{u}_i)$ ; if  $\mathbf{x}_i \in \mathcal{Z}_{\mathcal{H}}^c$ , then  $\phi_{\mathcal{H}}^{-1}(\mathbf{u}_i) \in \mathcal{Z}_{\mathcal{H}}^c$ , and since  $\mathcal{Q}_{\mathcal{H}} \subseteq \mathcal{X} \cup (\cap_{h \in \mathcal{H}} \text{Supp}(h)^c)$ ,  $\phi_{\mathcal{H}}^{-1}(\mathbf{u}_i) \in \cap_{h \in \mathcal{H}} \text{Supp}(h)^c$ .

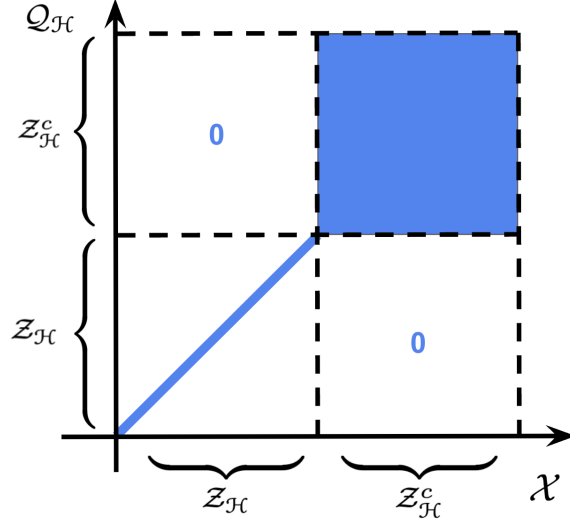


Figure 1: Illustration of  $\mathbb{O}$ .  $\mathbb{O}$  concentrates on  $\{(\mathbf{x}, \mathbf{x}) : \mathbf{x} \in \mathcal{Z}_{\mathcal{H}}\}$  in  $\mathcal{Z}_{\mathcal{H}} \times \mathcal{Z}_{\mathcal{H}}$ , agrees with the zero measure on  $\mathcal{Z}_{\mathcal{H}} \times \mathcal{Z}_{\mathcal{H}}^c$  and  $\mathcal{Z}_{\mathcal{H}}^c \times \mathcal{Z}_{\mathcal{H}}$ , and agrees with the product measure of  $\mathbb{P}_X \otimes (\mathbf{m} \circ \phi_{\mathcal{H}})$  on  $\mathcal{Z}_{\mathcal{H}}^c \times \mathcal{Z}_{\mathcal{H}}^c$ .

Thus, we take  $\tilde{h} = h \circ \phi_{\mathcal{H}}^{-1}$ , and consider the new class of functions  $\tilde{\mathcal{H}} = \{\tilde{h} : h \in \mathcal{H}\}$ . For any  $h \in \mathcal{H}$ ,

$$\begin{aligned} X_n(h) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n [h(\mathbf{x}_i) - \mathbb{E}[h(\mathbf{x}_i)]] \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n [h(\phi_{\mathcal{H}}^{-1}(\mathbf{u}_i)) - \mathbb{E}[h(\phi_{\mathcal{H}}^{-1}(\mathbf{u}_i))]] \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n [\tilde{h}(\mathbf{u}_i) - \mathbb{E}[\tilde{h}(\mathbf{u}_i)]], \end{aligned}$$

where the second equality follows because  $\mathbf{x}_i = \phi_{\mathcal{H}}^{-1}(\mathbf{u}_i)$  on the event  $\{\mathbf{x}_i \in \mathcal{Z}_{\mathcal{H}}\}$ , and  $h(\mathbf{x}_i) = h(\phi_{\mathcal{H}}^{-1}(\mathbf{u}_i)) = 0$  (a.s.) on the event  $\{\mathbf{x}_i \in \mathcal{Z}_{\mathcal{H}}^c\}$ . Hence, we work with an equivalent empirical process

$$\tilde{X}_n(\tilde{h}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [\tilde{h}(\mathbf{u}_i) - \mathbb{E}[\tilde{h}(\mathbf{u}_i)]], \quad \tilde{h} \in \tilde{\mathcal{H}}.$$

In particular,  $\mathbf{u}_i$  has the uniform distribution  $\mathbb{P}_U$  which has a Lebesgue density  $f_U$  that is bounded from above and below on its support  $[0, 1]^d$ ,

$$(X_n(h) : h \in \mathcal{H}) = (\tilde{X}_n(\tilde{h}) : \tilde{h} \in \tilde{\mathcal{H}}) \quad \text{almost surely,}$$

and Assumption SA.1 is satisfied with the random sample  $(\mathbf{u}_i : 1 \leq i \leq n)$  with  $\mathbf{u}_i \sim \mathbb{P}_U$  and the class of functions  $\tilde{\mathcal{H}}$ . We thus consider  $\mathcal{A}_K(\mathbb{P}_U, 1)$ , an axis aligned dyadic expansion of depth  $K$  with respect to probability measure  $\mathbb{P}_U = \text{Uniform}([0, 1]^d)$ . Suppose  $\mathcal{E}_K$  ( $\mathbb{M}_{\mathcal{H}, \mathcal{Q}_{\mathcal{H}}} = \mathbb{M}_{\tilde{\mathcal{H}}, [0, 1]^d}$ ) and  $\Pi_0 = \Pi_0[\mathcal{A}_K(\mathbb{P}_U, 1)]$  are defined based on  $\mathcal{A}_K(\mathbb{P}_U, 1)$  as in Section SA-II.1.2. By Lemma SA.2 and Lemma SA.10,  $\tilde{\mathcal{H}} \cup \Pi_0 \tilde{\mathcal{H}}$  is  $\mathbb{P}_U$ -pregaussian, hence by the same construction given in Section SA-II.1.3, on a possibly enlarged probability space, there exists a mean-zero Gaussian process  $\tilde{Z}_n^X$  indexed by  $\tilde{\mathcal{H}} \cup \Pi_0 \tilde{\mathcal{H}}$  such that with almost sure

continuous sample path such that

$$\mathbb{E}[\tilde{Z}_n^X(g)\tilde{Z}_n^X(f)] = \mathbb{E}[\tilde{X}_n(g)\tilde{X}_n(f)], \quad \forall g, f \in \tilde{\mathcal{H}} \cup \Pi_0\tilde{\mathcal{H}},$$

and  $U_{j,k} = \sum_{i=1}^n e_{j,k}(\mathbf{u}_i)$  for all  $(j,k) \in \mathcal{I}_K$ . Let  $\tilde{\mathcal{H}}_\delta$  be a  $\delta\mathbf{M}_{\tilde{\mathcal{H}},[0,1]^d} = \delta\mathbf{M}_{\mathcal{H},\mathcal{Q}_{\mathcal{H}}}$ -net of  $\tilde{\mathcal{H}}$  with cardinality no greater than  $N_{\tilde{\mathcal{H}},[0,1]^d}(\delta, \mathbf{M}_{\tilde{\mathcal{H}},[0,1]^d})$ .

The proof proceeds by bounding each of the terms in the decomposition

$$\begin{aligned} \|\tilde{X}_n - \tilde{Z}_n^X\|_{\tilde{\mathcal{H}}} &\leq \underbrace{\|\tilde{X}_n - \tilde{X}_n \circ \pi_{\tilde{\mathcal{H}}_\delta}\|_{\tilde{\mathcal{H}}}}_{\text{meshing error}} + \underbrace{\|\tilde{Z}_n^X \circ \pi_{\tilde{\mathcal{H}}_\delta} - \tilde{Z}_n^X\|_{\tilde{\mathcal{H}}}}_{\text{error on net}} + \|\tilde{X}_n - \tilde{Z}_n^X\|_{\tilde{\mathcal{H}}_\delta}, \\ \|\tilde{X}_n - \tilde{Z}_n^X\|_{\tilde{\mathcal{H}}_\delta} &\leq \underbrace{\|\Pi_0\tilde{X}_n - \Pi_0\tilde{Z}_n^X\|_{\tilde{\mathcal{H}}_\delta}}_{\text{approximation error}} + \underbrace{\|\tilde{X}_n - \Pi_0\tilde{X}_n\|_{\tilde{\mathcal{H}}_\delta} + \|\Pi_0\tilde{Z}_n^X - \tilde{Z}_n^X\|_{\tilde{\mathcal{H}}_\delta}}_{\text{projection error}}, \end{aligned}$$

and then balancing their contributions.

Given the cells  $\mathcal{A}_K(\mathbb{P}_U, 1)$ , we have  $\mathcal{U}_j \subseteq [-2^{-\frac{K-j}{d}+1}, 2^{-\frac{K-j}{d}+1}]^d$ . Then, by Lemma SA.7, for all  $t > 0$ ,

$$\mathbb{P}\left[\|\tilde{X}_n \circ \Pi_0 - \tilde{Z}_n^X \circ \Pi_0\|_{\tilde{\mathcal{H}}_\delta} > 48\sqrt{\frac{\mathcal{R}_K(\tilde{\mathcal{H}}_\delta)}{n}}t + 4\sqrt{\frac{\mathbf{C}_{\tilde{\mathcal{H}}_\delta, K}}{n}}t\right] \leq 2N_{\tilde{\mathcal{H}},[0,1]^d}(\delta, \mathbf{M}_{\tilde{\mathcal{H}},[0,1]^d})e^{-t},$$

where

$$\mathcal{R}_K(\tilde{\mathcal{H}}_\delta) \leq \begin{cases} \min\{\mathbf{TV}_{\tilde{\mathcal{H}}_\delta, [0,1]^d}^* \mathbf{M}_{\tilde{\mathcal{H}}_\delta, [0,1]^d}, \mathbf{TV}_{\tilde{\mathcal{H}}_\delta, [0,1]^d}^* \mathbf{L}_{\tilde{\mathcal{H}}_\delta, [0,1]^d}\}, & \text{if } d = 1, \\ \min\{2^K \mathbf{TV}_{\tilde{\mathcal{H}}_\delta, [0,1]^d}^* \mathbf{M}_{\tilde{\mathcal{H}}_\delta, [0,1]^d}, K \mathbf{TV}_{\tilde{\mathcal{H}}_\delta, [0,1]^d}^* \mathbf{L}_{\tilde{\mathcal{H}}_\delta, [0,1]^d}\}, & \text{if } d = 2, \\ \min\{2^{K(d-1)} \mathbf{TV}_{\tilde{\mathcal{H}}_\delta, [0,1]^d}^* \mathbf{M}_{\tilde{\mathcal{H}}_\delta, [0,1]^d}, 2^{K(d-2)} \mathbf{TV}_{\tilde{\mathcal{H}}_\delta, [0,1]^d}^* \mathbf{L}_{\tilde{\mathcal{H}}_\delta, [0,1]^d}\} & \text{if } d \geq 3. \end{cases}$$

Now we calculate the  $\mathbf{C}_{\tilde{\mathcal{H}}_\delta, K}$  term. Let  $\tilde{h} \in \tilde{\mathcal{H}}$  and take  $(\tilde{h}_\ell)_{\ell \in \mathbb{N}}$  be any sequence of real-valued functions on  $([0, 1]^d, \mathcal{B}([0, 1]^d))$  such that  $\tilde{h}_\ell \rightarrow \tilde{h}$   $\mathbf{m}$ -almost surely, and are bounded by  $2\mathbf{M}_{\tilde{\mathcal{H}}}$  on  $\mathcal{X}$ . Moreover, by Dominated Convergence Theorem, the definition of  $\mathbf{K}_{\tilde{\mathcal{H}}, [0,1]^d}^*$  and similar arguments as in the proof of Lemma SA.7, for each  $(j, k) \in \mathcal{I}_K$ ,

$$\begin{aligned} \sum_{m: \mathcal{C}_{l,m} \subseteq \mathcal{C}_{j,k}} |\tilde{\beta}_{j,k}(\tilde{h})| &\leq 2^{2(K-l)} \int_{\mathcal{U}_l} \int_{\mathcal{C}_{j,k}} |\tilde{h}(\mathbf{x}) - \tilde{h}(\mathbf{x} + \mathbf{s})| d\mathbf{x} d\mathbf{s} \\ &= \lim_{\ell \rightarrow \infty} 2^{2(K-l)} \int_{\mathcal{U}_l} \int_{\mathcal{C}_{j,k}} |\tilde{h}_\ell(\mathbf{x}) - \tilde{h}_\ell(\mathbf{x} + \mathbf{s})| d\mathbf{x} d\mathbf{s} \\ &\leq \limsup_{\ell \rightarrow \infty} 2^{2(K-l)} \int_{\mathcal{U}_l} \|\mathbf{s}\| \mathbf{TV}_{\{\tilde{h}_\ell\}, \mathcal{C}_{j,k}} d\mathbf{s}. \end{aligned}$$

Since the above inequality holds for all sequences  $(\tilde{h}_\ell)_{\ell \in \mathbb{N}}$  such that  $\tilde{h}_\ell \rightarrow \tilde{h}$   $\mathbf{m}$ -almost surely, and are bounded

by  $2\mathbf{M}_{\mathcal{H}}$  on  $[0, 1]^d$ , Definitions SA.1 and SA.2 implies

$$\begin{aligned}
\sum_{m:\mathcal{C}_{l,m} \subseteq \mathcal{C}_{j,k}} \left| \tilde{\beta}_{j,k}(\tilde{h}) \right| &\leq 2^{2(K-l)} \int_{\mathcal{U}_l} \|\mathbf{s}\| \mathbf{TV}_{\mathcal{H}, \mathcal{C}_{j,k}}^* d\mathbf{s} \\
&\leq 2^{2(K-l)} \int_{\mathcal{U}_l} \|\mathbf{s}\| \mathbf{K}_{\mathcal{H}, [0,1]^d}^* \|\mathcal{C}_{j,k}\|_{\infty}^{d-1} d\mathbf{s} \\
&\leq \sqrt{d} 2^{2(K-l)} \mathbf{m}(\mathcal{U}_l) \|\mathcal{U}_l\|_{\infty} \|\mathcal{C}_{j,k}\|_{\infty}^{d-1} \mathbf{K}_{\mathcal{H}, [0,1]^d}^* \\
&\leq \sqrt{d} 2^{\frac{d-1}{d}(j-l)} \mathbf{K}_{\mathcal{H}, [0,1]^d}^*.
\end{aligned}$$

It follows from the definition of  $\mathcal{C}_{\tilde{\mathcal{H}}, K}$  in Lemma SA.3 that

$$\mathcal{C}_{\tilde{\mathcal{H}}, K} \leq \min \left\{ \sqrt{K \mathbf{M}_{\mathcal{H}, [0,1]^d}^2}, \sqrt{\sqrt{d} \mathbf{M}_{\mathcal{H}, [0,1]^d} \mathbf{K}_{\mathcal{H}, [0,1]^d}^* + \mathbf{M}_{\mathcal{H}, [0,1]^d}^2} \right\}.$$

For projection error, by Lemma SA.9, for all  $t > 0$ , with probability at least  $1 - 2\mathbf{N}_{\tilde{\mathcal{H}}, [0,1]^d}(\delta, \mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d}) e^{-t}$ ,

$$\begin{aligned}
\|\tilde{X}_n - \tilde{X}_n \circ \Pi_0\|_{\tilde{\mathcal{H}}_{\delta}} &\leq \sqrt{4d \min\{2\mathbf{M}_{\tilde{\mathcal{H}}_{\delta}, [0,1]^d}, \mathbf{L}_{\tilde{\mathcal{H}}_{\delta}, [0,1]^d} 2^{-K}\} 2^{-K} \mathbf{TV}_{\tilde{\mathcal{H}}_{\delta}, [0,1]^d}^* t} \\
&\quad + \frac{4 \min\{2\mathbf{M}_{\tilde{\mathcal{H}}_{\delta}, [0,1]^d}, \mathbf{L}_{\tilde{\mathcal{H}}_{\delta}, [0,1]^d} 2^{-K}\}}{3\sqrt{n}} t, \\
\|\tilde{Z}_n^X - \tilde{Z}_n^X \circ \Pi_0\|_{\tilde{\mathcal{H}}_{\delta}} &\leq \sqrt{4d \min\{2\mathbf{M}_{\tilde{\mathcal{H}}_{\delta}, [0,1]^d}, \mathbf{L}_{\tilde{\mathcal{H}}_{\delta}, [0,1]^d} 2^{-K}\} 2^{-K} \mathbf{TV}_{\tilde{\mathcal{H}}_{\delta}, [0,1]^d}^* t}.
\end{aligned}$$

We balance the previous two errors by choosing  $K = \lfloor d^{-1} \log_2 n \rfloor$  and get for all  $t > 0$ , with probability at least  $1 - 2 \exp(-t)$ ,

$$\|\tilde{X}_n - \tilde{Z}_n^X\|_{\tilde{\mathcal{H}}_{\delta}} \leq \tilde{\mathbf{A}}_n(t, \delta),$$

where

$$\begin{aligned}
\tilde{\mathbf{A}}_n(t, \delta) &= \min \left\{ \mathbf{m}_{n,d} \sqrt{\mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d}}, \mathbf{1}_{n,d} \sqrt{\mathbf{L}_{\tilde{\mathcal{H}}, [0,1]^d}} \right\} \sqrt{d \mathbf{TV}_{\tilde{\mathcal{H}}, [0,1]^d}^*} \sqrt{(t + \log \tilde{\mathbf{N}}_{\tilde{\mathcal{H}}, [0,1]^d}(\delta, \mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d}))} \\
&\quad + \sqrt{\frac{\mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d}}{n}} \min \left\{ \sqrt{\log n} \sqrt{\mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d}}, \sqrt{\sqrt{d} \mathbf{K}_{\tilde{\mathcal{H}}, [0,1]^d}^* + \mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d}} \right\} (t + \log \tilde{\mathbf{N}}_{\tilde{\mathcal{H}}, [0,1]^d}(\delta, \mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d})).
\end{aligned}$$

By Lemma SA.6 we bound the meshing error by, for all  $t > 0$ ,

$$\begin{aligned}
\mathbb{P}[\|\tilde{X}_n - \tilde{X}_n \circ \pi_{\tilde{\mathcal{H}}_{\delta}}\|_{\tilde{\mathcal{H}}} > C \mathbf{F}_n(t, \delta)] &\leq \exp(-t), \\
\mathbb{P}[\|\tilde{Z}_n^X \circ \pi_{\tilde{\mathcal{H}}_{\delta}} - \tilde{Z}_n^X\|_{\tilde{\mathcal{H}}} > C(\mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d} J(\delta, \tilde{\mathcal{H}}, \mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d}) + \delta \mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d} \sqrt{t})] &\leq \exp(-t),
\end{aligned}$$

where

$$\mathbf{F}_n(t, \delta) = J(\delta, \tilde{\mathcal{H}}, \mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d}) \mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d} + \frac{\mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d} J^2(\delta, \tilde{\mathcal{H}}, \mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d})}{\delta^2 \sqrt{n}} + \delta \mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d} \sqrt{t} + \frac{\mathbf{M}_{\tilde{\mathcal{H}}, [0,1]^d}}{\sqrt{n}} t.$$

Take the Gaussian process  $(Z_n(h) : h \in \mathcal{H})$  such that, almost surely,  $Z_n(h) = \tilde{Z}_n(\tilde{h})$  for all  $h \in \mathcal{H}$ . The

result then follows from the decomposition that

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

and Lemma SA.10 to establish the relationships between the parameters of  $\mathcal{H}$  over  $\mathcal{Q}_{\mathcal{H}}$  and those of  $\tilde{\mathcal{H}}$  over  $[0, 1]^d$ .  $\square$

## SA-II.4 Additional Results

In what follows, 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  $\text{TV}_{\mathcal{F}} = \text{TV}_{\mathcal{F}, \mathcal{Q}_{\mathcal{F}}}$ ,  $\text{K}_{\mathcal{F}} = \text{K}_{\mathcal{F}, \mathcal{Q}_{\mathcal{F}}}$ ,  $\text{M}_{\mathcal{F}, \mathcal{X}} = \text{M}_{\mathcal{F}, \mathcal{Q}_{\mathcal{F}}}$ ,  $M_{\mathcal{F}, \mathcal{X}}(\mathbf{u}) = M_{\mathcal{F}, \mathcal{Q}_{\mathcal{F}}}(\mathbf{u})$ ,  $\text{L}_{\mathcal{F}} = \text{L}_{\mathcal{F}, \mathcal{Q}_{\mathcal{F}}}$ , and so on, whenever there is no confusion.

**Corollary SA.1** (VC-Type Bounded Functions). *Suppose the conditions of Corollary 1 hold. Then,*

$$\text{S}_n(t) = m_{n,d} \sqrt{c_1 \text{M}_{\mathcal{H}} \text{TV}_{\mathcal{H}}} \sqrt{t + \mathbf{d}_{\mathcal{H}} \log(c_{\mathcal{H}} n)} + \sqrt{\frac{\text{M}_{\mathcal{H}}}{n}} \min\{\sqrt{\log n} \sqrt{\text{M}_{\mathcal{H}}}, \sqrt{c_3 \text{K}_{\mathcal{H}} + \text{M}_{\mathcal{H}}}\} (t + \mathbf{d}_{\mathcal{H}} \log(c_{\mathcal{H}} n))$$

in Theorem 1.

*Proof of Corollary SA.1.* Take  $\delta = n^{-1/2}$ . Under the VC-type class condition,  $\log \text{N}_{\mathcal{H}}(n^{-1}, \text{M}_{\mathcal{H}}) \leq \log(c_{\mathcal{H}}) + \mathbf{d}_{\mathcal{H}} \log(n) \leq \mathbf{d}_{\mathcal{H}} \log(c_{\mathcal{H}} n)$ , where the last inequality holds since  $c_{\mathcal{H}} \geq e$  and  $\mathbf{d}_{\mathcal{H}} > 0$ . This gives

$$\begin{aligned} \text{A}_n(t, n^{-1/2}) &\leq m_{n,d} \sqrt{c_1 (t + \mathbf{d}_{\mathcal{H}} \log(c_{\mathcal{H}} n)) \text{M}_{\mathcal{H}} \text{TV}_{\mathcal{H}}} \\ &\quad + \min\{\sqrt{\log(n) \text{M}_{\mathcal{H}}}, \sqrt{c_3 \text{K}_{\mathcal{H}} + \text{M}_{\mathcal{H}}}\} \sqrt{\frac{\text{M}_{\mathcal{H}}}{n}} (t + \mathbf{d}_{\mathcal{H}} \log(c_{\mathcal{H}} n)). \end{aligned}$$

Moreover,  $J(\delta, \mathcal{H}, \text{M}_{\mathcal{H}}) \leq \int_0^\delta \sqrt{1 + \mathbf{d}_{\mathcal{H}} \log(c_{\mathcal{H}} \varepsilon^{-1})} d\varepsilon \leq 3\delta \sqrt{\mathbf{d}_{\mathcal{H}} \log(c_{\mathcal{H}}/\delta)}$ . It follows that

$$\text{F}_n(t, n^{-1/2}) \leq \frac{3\text{M}_{\mathcal{H}}}{\sqrt{n}} \mathbf{d}_{\mathcal{H}} \log(c_{\mathcal{H}} n) + \frac{\text{M}_{\mathcal{H}}}{\sqrt{n}} (\sqrt{t} + t).$$

The result then follows from Theorem 1.  $\square$

**Corollary SA.2** (VC-Type Lipschitz Functions). *Suppose the conditions of Corollary 2 hold. Then,*

$$\begin{aligned} \text{S}_n(t) &= \min\{m_{n,d} \sqrt{\text{M}_{\mathcal{H}}}, l_{n,d} \sqrt{c_2 \text{L}_{\mathcal{H}}}\} \sqrt{\text{TV}_{\mathcal{H}}} \sqrt{t + \mathbf{d}_{\mathcal{H}} \log(c_{\mathcal{H}} n)} \\ &\quad + \sqrt{\frac{\text{M}_{\mathcal{H}}}{n}} \min\{\sqrt{\log n} \sqrt{\text{M}_{\mathcal{H}}}, \sqrt{c_3 \text{K}_{\mathcal{H}} + \text{M}_{\mathcal{H}}}\} (t + \mathbf{d}_{\mathcal{H}} \log(c_{\mathcal{H}} n)) \end{aligned}$$

in Theorem 1.

*Proof of Corollary SA.2.* The result follows by taking  $\delta = n^{-1/2}$  and apply Theorem 1, with calculations similar to Corollary SA.1.  $\square$

**Corollary SA.3** (Polynomial-Entropy Functions). *Suppose the conditions of Corollary 2 hold. Then,*

$$\text{S}_n(t) = \mathbf{a}_{\mathcal{H}} (2 - \mathbf{b}_{\mathcal{H}})^{-2} \min\{\text{S}_n^{\text{bdd}}(t), \text{S}_n^{\text{lip}}(t), \text{S}_n^{\text{err}}(t)\}$$

in Theorem 1, where

$$\begin{aligned}
S_n^{bdd}(t) &= m_{n,d} \sqrt{c_1 M_{\mathcal{H}} \text{TV}_{\mathcal{H}}} (\sqrt{t} + (m_{n,d}^2 M_{\mathcal{H}}^{-1} \text{TV}_{\mathcal{H}})^{-\frac{b_{\mathcal{H}}}{4}}) \\
&\quad + \sqrt{\frac{M_{\mathcal{H}}}{n}} \min\{\sqrt{\log n} \sqrt{M_{\mathcal{H}}}, \sqrt{c_3 K_{\mathcal{H}} + M_{\mathcal{H}}}\} (t + (m_{n,d}^2 M_{\mathcal{H}}^{-1} \text{TV}_{\mathcal{H}})^{-\frac{b_{\mathcal{H}}}{2}}), \\
S_n^{lip}(t) &= l_{n,d} \sqrt{c_1 c_2 L_{\mathcal{H}} \text{TV}_{\mathcal{H}}} (\sqrt{t} + (l_{n,d}^2 M_{\mathcal{H}}^{-2} L_{\mathcal{H}} \text{TV}_{\mathcal{H}})^{-\frac{b_{\mathcal{H}}}{4}}) \\
&\quad + \sqrt{\frac{M_{\mathcal{H}}}{n}} \min\{\sqrt{\log n} \sqrt{M_{\mathcal{H}}}, \sqrt{c_3 K_{\mathcal{H}} + M_{\mathcal{H}}}\} (t + (l_{n,d}^2 M_{\mathcal{H}}^{-2} L_{\mathcal{H}} \text{TV}_{\mathcal{H}})^{-\frac{b_{\mathcal{H}}}{2}}), \\
S_n^{err}(t) &= \min\{m_{n,d} \sqrt{M_{\mathcal{H}}}, l_{n,d} \sqrt{c_2 L_{\mathcal{H}}}\} \sqrt{c_1 \text{TV}_{\mathcal{H}}} (\sqrt{t} + n^{\frac{b_{\mathcal{H}}}{2(b_{\mathcal{H}}+2)}}) \\
&\quad + \sqrt{\frac{M_{\mathcal{H}}}{n}} \min\{\sqrt{\log n} \sqrt{M_{\mathcal{H}}}, \sqrt{c_3 K_{\mathcal{H}} + M_{\mathcal{H}}}\} (t + n^{\frac{b_{\mathcal{H}}}{b_{\mathcal{H}}+2}}) + n^{-\frac{1}{b_{\mathcal{H}}+2}} M_{\mathcal{H}} \sqrt{t}.
\end{aligned}$$

*Proof of Corollary SA.3.* Under the polynomial entropy condition,  $\log N_{\mathcal{H}}(\delta) \leq a_{\mathcal{H}} \delta^{-b_{\mathcal{H}}}$ ,  $J(\delta, \mathcal{H}, M_{\mathcal{H}}) \leq \sqrt{a_{\mathcal{H}}}(2 - b_{\mathcal{H}})^{-1} \delta^{-b_{\mathcal{H}}/2+1}$ ,

$$\begin{aligned}
A_n(t, \delta) &\leq \min\{m_{n,d} \sqrt{M_{\mathcal{H}}}, l_{n,d} \sqrt{c_2 L_{\mathcal{H}}}\} \sqrt{\text{TV}_{\mathcal{H}}(t + a_{\mathcal{H}} \delta^{-b_{\mathcal{H}}})} \\
&\quad + \sqrt{\frac{M_{\mathcal{H}}}{n}} \min\{\sqrt{\log n} \sqrt{M_{\mathcal{H}}}, \sqrt{c_3 K_{\mathcal{H}} + M_{\mathcal{H}}}\} (t + a_{\mathcal{H}} \delta^{-b_{\mathcal{H}}}), \\
F_n(t, \delta) &\leq a_{\mathcal{H}} (2 - b_{\mathcal{H}})^{-2} \left( M_{\mathcal{H}} \delta^{-b_{\mathcal{H}}/2+1} + \frac{M_{\mathcal{H}}}{\sqrt{n}} \delta^{-b_{\mathcal{H}}} + \delta M_{\mathcal{H}} \sqrt{t} + \frac{M_{\mathcal{H}}}{\sqrt{n}} t \right).
\end{aligned}$$

Notice that the two terms  $\frac{M_{\mathcal{H}}}{\sqrt{n}} \delta^{-b_{\mathcal{H}}}$  and  $\frac{M_{\mathcal{H}}}{\sqrt{n}} t$  in  $F_n(t, \delta)$  are dominated by terms in  $A_n(t, \delta)$ . And when  $\delta \leq n^{-1/2}$ , the third term  $\delta M_{\mathcal{H}} \sqrt{t}$  is also dominated by terms in  $A_n(t, \delta)$ . To choose  $\delta$  that balance  $A_n$  and  $F_n$ , we consider the following three cases:

**Case 1:** Choosing  $\delta$  such that  $m_{n,d} \sqrt{M_{\mathcal{H}} \text{TV}_{\mathcal{H}} \delta^{-b_{\mathcal{H}}}} = M_{\mathcal{H}} \delta^{-b_{\mathcal{H}}/2+1}$ , gives  $\delta_* = m_{n,d} \sqrt{\text{TV}_{\mathcal{H}}/M_{\mathcal{H}}}$ . Notice that this choice also makes  $\delta M_{\mathcal{H}} \sqrt{t} \leq \sqrt{\frac{M_{\mathcal{H}}}{n}} \min\{\sqrt{\log n} \sqrt{M_{\mathcal{H}}}, \sqrt{c_3 K_{\mathcal{H}} + M_{\mathcal{H}}}\} (t + a_{\mathcal{H}} \delta^{-b_{\mathcal{H}}})$ . Thus, we get  $A_n(t, \delta_*) + F_n(t, \delta_*) \leq S_n^{bdd}(t)$ .

**Case 2:** Choosing  $\delta$  such that  $l_{n,d} \sqrt{L_{\mathcal{H}} \text{TV}_{\mathcal{H}} \delta^{-b_{\mathcal{H}}}} = M_{\mathcal{H}} \delta^{-b_{\mathcal{H}}/2+1}$ , gives  $\delta_* = l_{n,d} \sqrt{L_{\mathcal{H}} \text{TV}_{\mathcal{H}}/M_{\mathcal{H}}^2}$ . Again, this choice of  $\delta$  makes  $\delta M_{\mathcal{H}} \sqrt{t} \leq \sqrt{\frac{M_{\mathcal{H}}}{n}} \min\{\sqrt{\log n} \sqrt{M_{\mathcal{H}}}, \sqrt{c_3 K_{\mathcal{H}} + M_{\mathcal{H}}}\} (t + a_{\mathcal{H}} \delta^{-b_{\mathcal{H}}})$ . Thus, we get  $A_n(t, \delta_*) + F_n(t, \delta_*) \leq S_n^{lip}(t)$ .

**Case 3:** Choosing  $\delta$  such that  $M_{\mathcal{H}} n^{-1/2} \delta^{-b_{\mathcal{H}}} = M_{\mathcal{H}} \delta^{-b_{\mathcal{H}}/2+1}$ , gives  $\delta_* = n^{-1/(b_{\mathcal{H}}+2)}$ . Thus, we get  $A_n(t, \delta_*) + F_n(t, \delta_*) \leq S_n^{err}(t)$ .  $\square$

## SA-II.5 Proofs of Corollaries 1, 2, and 3

*Proof of Corollary 1.* Take  $t = C \log n$  with  $C > 1$  in Corollary SA.1.  $\square$

*Proof of Corollary 2.* Take  $t = C \log n$  with  $C > 1$  in Corollary SA.2.  $\square$

*Proof of Corollary 3.* Take  $t = C \log n$  with  $C > 1$  in Corollary SA.3.  $\square$

## SA-II.6 Example 1: Kernel Density Estimation

To simplify notation, in this section the parameters of  $\mathcal{H}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{Q}_{\mathcal{H}}$ , and the index  $\mathcal{C}$  is omitted whenever there is no confusion.

### SA-II.6.1 Surrogate Measure and Normalizing Transformation

We show that the two sets of primitive conditions discussed in the paper imply condition (ii) in Theorem 1.

First, consider the case  $\mathcal{X} = \times_{l=1}^d [\mathbf{a}_l, \mathbf{b}_l]$ ,  $-\infty \leq \mathbf{a}_l < \mathbf{b}_l \leq \infty$  and  $\mathcal{W}$  is arbitrary. Observe that  $\mathbb{Q}_{\mathcal{H}} = \mathbb{P}_X$  is always a valid surrogate measure for  $\mathbb{P}_X$  with respect to  $\mathcal{H}$ , according to Definition 2. The conclusion then follows from Case 1 from Section 3.1 with  $f_Q = f_X$ .

Second, consider the case when  $\mathcal{X}$  may be unbounded. We present a general construction, and then specialize it to the example discussed in the paper. Suppose we can find  $\mathcal{Q}_{\mathcal{H}}$  diffeomorphic to  $[0, 1]^d$  such that  $\mathcal{X} \cap \text{Supp}(\mathcal{H}) \subseteq \mathcal{Q}_{\mathcal{H}} \subseteq \mathcal{X} \cup \text{Supp}(\mathcal{H})^c$ , with  $\mathbb{P}_X(\mathcal{X} \cap \text{Supp}(\mathcal{H})) < 1$  and  $\mathbf{m}(\mathcal{Q}_{\mathcal{H}} \setminus (\mathcal{X} \cap \text{Supp}(\mathcal{H}))) > 0$ . Setting  $\mathbb{Q}_{\mathcal{H}}$  to be the probability measure with Lebesgue density  $f_Q$  such that

$$f_Q(\mathbf{x}) = \begin{cases} f_X(\mathbf{x}), & \text{if } \mathbf{x} \in \mathcal{X} \cap \text{Supp}(\mathcal{H}), \\ (1 - \mathbb{P}_X(\mathcal{X} \cap \text{Supp}(\mathcal{H}))) / \mathbf{m}(\mathcal{Q}_{\mathcal{H}} \setminus (\mathcal{X} \cap \text{Supp}(\mathcal{H}))), & \text{if } \mathbf{x} \in \mathcal{Q}_{\mathcal{H}} \setminus (\mathcal{X} \cap \text{Supp}(\mathcal{H})), \\ 0, & \text{otherwise.} \end{cases}$$

then  $\mathbb{Q}_{\mathcal{H}}$  is a surrogate measure of  $\mathbb{P}_X$  with respect to  $\mathcal{H}$ . Suppose  $\chi$  is a diffeomorphism from  $\mathcal{Q}_{\mathcal{H}}$  to  $[0, 1]^d$ . Since we assumed  $\mathcal{X} \cap \text{Supp}(\mathcal{H}) \subseteq \mathcal{Q}_{\mathcal{H}} \subseteq \mathcal{X} \cup \text{Supp}(\mathcal{H})^c$ , with  $\mathbb{P}_X(\mathcal{X} \cap \text{Supp}(\mathcal{H})) < 1$  and  $\mathbf{m}(\mathcal{Q}_{\mathcal{H}} \setminus (\mathcal{X} \cap \text{Supp}(\mathcal{H}))) > 0$ , we can check that (1)  $f_Q$  is supported and positive on  $\mathcal{Q}_{\mathcal{H}}$ , (2)  $f_Q$  agrees with  $f_X$  on  $\mathcal{X} \cap \text{Supp}(\mathcal{H})$ . Then Case 2 in Section 3.1 implies  $\phi_{\mathcal{H}} = T_{\mathcal{Q}_{\mathcal{H}} \circ \chi^{-1}} \circ \chi$  is a valid normalizing transformation, and condition (ii) in Theorem 1 holds. Suppose  $0 < \inf_{\mathbf{x} \in \mathcal{X} \cap \text{Supp}(\mathcal{H})} f_X(\mathbf{x}) < \sup_{\mathbf{x} \in \mathcal{X} \cap \text{Supp}(\mathcal{H})} f_X(\mathbf{x}) < \infty$  and  $\frac{\sup_{\mathbf{x} \in [0, 1]^d} |\det(\nabla \chi^{-1}(\mathbf{x}))|}{\inf_{\mathbf{x} \in [0, 1]^d} |\det(\nabla \chi^{-1}(\mathbf{x}))|} \|\|\nabla \chi^{-1}\|_2\|_{\infty} < \infty$ , then we have  $\mathbf{c}_1 = O(1)$  and  $\mathbf{c}_2 = O(1)$  (and hence  $\mathbf{c}_3 = O(1)$ ).

For a concrete example, consider the case  $\mathcal{X} = \mathbb{R}_+^d$ ,  $\mathcal{W} = \times_{l=1}^d [\mathbf{a}_l, \mathbf{b}_l]$ ,  $0 \leq \mathbf{a}_l < \mathbf{b}_l < \infty$ , and  $\mathcal{K} = [-1, 1]^d$ . Observe that  $\text{Supp}(\mathcal{H}) \cap \mathcal{X} = \times_{l=1}^d [(\mathbf{a}_l - b)_+, \mathbf{b}_l + b] = \times_{l=1}^d [\bar{\mathbf{a}}_l, \bar{\mathbf{b}}_l]$ . Since  $\mathcal{X} = \mathbb{R}_+^d$ ,  $\mathbb{P}_X(\times_{l=1}^d [\bar{\mathbf{a}}_l, \bar{\mathbf{b}}_l]) < 1$ . Moreover, we can check that  $\mathcal{X} \cap \text{Supp}(\mathcal{H}) \subseteq \mathcal{Q}_{\mathcal{H}} \subseteq \mathcal{X} \cup \text{Supp}(\mathcal{H})^c$  and  $\mathbb{Q}_{\mathcal{H}}$  agrees with  $\mathbb{P}_X$  on  $\mathcal{X} \cap \text{Supp}(\mathcal{H})$ . The rest then follows from the general construction above.

### SA-II.6.2 Class $\mathcal{H}$ and Its Corresponding Constants

Let  $\mathcal{H} = \{h_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\}$  with  $h_{\mathbf{w}}(\cdot) = b^{-d/2} K(b^{-1}(\mathbf{w} - \cdot))$ . Since  $K$  is compactly supported and Lipschitz,  $\mathbf{M}_{\{K\}} < \infty$ . Hence,  $\mathbf{M}_{\mathcal{H}} = b^{-d/2} \mathbf{M}_{\{K\}} \leq C_K b^{-d/2}$  and  $\mathbf{L}_{\mathcal{H}} \leq b^{-\frac{d}{2}-1} \mathbf{L}_{\{K\}} \leq C_K b^{-d/2-1}$ , where  $C_K$  is a constant that only depends on the kernel function  $K$ . Since  $\sup_{\mathbf{w} \in \mathcal{W}} \mathbf{m}(\text{Supp}(h_{\mathbf{w}})) \leq C_K b^d$  and each  $h_{\mathbf{w}}$  is differentiable,

$$\mathbf{TV}_{\mathcal{H}} = \sup_{\mathbf{w} \in \mathcal{W}} \int \|\nabla h_{\mathbf{w}}(\mathbf{u})\| d\mathbf{u} \leq \sup_{\mathbf{w} \in \mathcal{W}} \mathbf{m}(\text{Supp}(h_{\mathbf{w}})) \mathbf{L}_{\mathcal{H}} \leq C_K b^{d/2-1}.$$

To upper bound  $\mathbf{K}_{\mathcal{H}}$ , let  $\mathcal{D} \subseteq \mathcal{Q}_{\mathcal{H}}$  be a cube with edges of length  $\mathbf{a}$  parallel to the coordinate axes. Consider the following two cases: (i) if  $\mathbf{a} < b$ , then  $\mathbf{TV}_{\mathcal{H}, \mathcal{D}} \leq C_K b^{-d/2-1} \mathbf{a}^d \leq C_K b^{-d/2} \mathbf{a}^{d-1}$ ; (ii) if  $\mathbf{a} > b$ , then  $\mathbf{TV}_{\mathcal{H}, \mathcal{D}} \leq C_K \sup_{\mathbf{w} \in \mathcal{W}} \mathbf{m}(\text{Supp}(h_{\mathbf{w}})) \mathbf{L}_{\mathcal{H}} \leq C_K b^d b^{-d/2-1} \leq C_K b^{-d/2} b^{d-1} \leq C_K b^{-d/2} \mathbf{a}^{d-1}$ . This shows

$$\mathbf{K}_{\mathcal{H}} \leq C_K b^{-d/2}.$$

Next, by a change of variables,

$$\mathbf{E}_{\mathcal{H}} = \sup_{\mathbf{w} \in \mathcal{W}} \int b^{-\frac{d}{2}} |K(b^{-1}(\mathbf{w} - \mathbf{u}))| f_X(\mathbf{u}) d\mathbf{u} = \sup_{\mathbf{w} \in \mathcal{W}} \int b^{-\frac{d}{2}} |K(\mathbf{z})| f_X(\mathbf{w} - b\mathbf{z}) b^d d\mathbf{z} \leq C_K b^{d/2}.$$



Finally, we check that  $\mathcal{H}$  is a VC-type class. We will apply Lemma 7 from [Cattaneo \*et al.\* \(2024\)](#) on the class  $\mathbb{M}_{\mathcal{H}}^{-1}\mathcal{H}$ . To check the conditions in this lemma, define  $g_{\mathbf{w}}(\cdot) = b^{-\frac{d}{2}}\mathbb{M}_{\mathcal{H}}^{-1}K(\cdot)$  for all  $\mathbf{w} \in \mathcal{W}$ . Note that  $g_{\mathbf{w}}$  is the same function for all  $\mathbf{w} \in \mathcal{W}$  in this setting (but, more generally, our results allow for functions varying with the evaluation point such as in the case of boundary adaptive kernels). Then  $\mathbb{M}_{\mathcal{H}}^{-1}\mathcal{H} = \{g_{\mathbf{w}}(\frac{\cdot}{b}) : \mathbf{w} \in \mathcal{W}\}$ , and there exists a constant  $c_K$ , only depending on  $\mathbb{M}_{\{K\}}$  and  $\mathbb{L}_{\{K\}}$ , such that

$$\sup_{\mathbf{w} \in \mathcal{W}} \|g_{\mathbf{w}}\|_{\infty} \leq c_K, \quad \sup_{\mathbf{w} \in \mathcal{W}} \sup_{\mathbf{u}, \mathbf{v} \in \mathcal{Q}_{\mathcal{H}}} \frac{|g_{\mathbf{w}}(\mathbf{u}) - g_{\mathbf{w}}(\mathbf{v})|}{\|\mathbf{u} - \mathbf{v}\|_{\infty}} \leq c_K, \quad \sup_{\mathbf{w}, \mathbf{w}' \in \mathcal{W}} \sup_{\mathbf{u} \in \mathcal{Q}_{\mathcal{H}}} \frac{|g_{\mathbf{w}}(\mathbf{u}) - g_{\mathbf{w}'}(\mathbf{u})|}{\|\mathbf{w} - \mathbf{w}'\|_{\infty}} \leq c_K.$$

We can apply Lemma 7 from [Cattaneo \*et al.\* \(2024\)](#), which is modified upon Lemma 4.1 from [Rio \(1994\)](#), to show that for all  $0 < \varepsilon < 1$ ,  $N_{\mathbb{M}_{\mathcal{H}}^{-1}\mathcal{H}}(\varepsilon, 1) \leq c_K \varepsilon^{-d-1} + 1$ , and hence

$$N_{\mathcal{H}}(\varepsilon, \mathbb{M}_{\mathcal{H}}) \leq c_K \varepsilon^{-2d-2} + 1,$$

The conclusions on uniform Gaussian strong approximation rates then follow from Corollaries 1–3.

### SA-III Multiplicative-Separable Empirical Process

Let  $\mathbf{z}_i = (\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$ ,  $i = 1, \dots, n$ , be i.i.d. random vectors supported on a background probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . The multiplicative-separable empirical process is

$$G_n(g, r) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (g(\mathbf{x}_i)r(y_i) - \mathbb{E}[g(\mathbf{x}_i)r(y_i)]), \quad g \in \mathcal{G}, r \in \mathcal{R},$$

where  $\mathcal{G}$  and  $\mathcal{R}$  are possibly  $n$ -varying classes of functions. Notably, if we take  $\mathcal{H} = \mathcal{G} \cdot \mathcal{R} = \{g \cdot r : g \in \mathcal{G}, r \in \mathcal{R}\}$ , then the above process can also be written as a generic empirical process based on  $(\mathbf{z}_i : 1 \leq i \leq n)$  because

$$X_n(h) = X_n(g \cdot r) = \frac{1}{\sqrt{n}} \sum_{i=1}^n ((g \cdot r)(\mathbf{z}_i) - \mathbb{E}[(g \cdot r)(\mathbf{z}_i)]), \quad h = g \cdot r \in \mathcal{H} = \mathcal{G} \cdot \mathcal{R}.$$

Hence, the same decomposition for the  $X_n$  process also applies for the  $G_n$  process:

$$\begin{aligned} \|G_n - Z_n^G\|_{\mathcal{G} \times \mathcal{R}} &\leq \|G_n - Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_{\delta}} + \|G_n - G_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}}\|_{\mathcal{G} \times \mathcal{R}} + \|Z_n^G \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}} - Z_n^G\|_{\mathcal{G} \times \mathcal{R}} \\ &\leq \|\Pi_1 Z_n^G - Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_{\delta}} + \|G_n - \Pi_1 G_n\|_{(\mathcal{G} \times \mathcal{R})_{\delta}} + \|\Pi_1 G_n - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_{\delta}} \\ &\quad + \|G_n - G_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}}\|_{\mathcal{G} \times \mathcal{R}} + \|Z_n^G \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}} - Z_n^G\|_{\mathcal{G} \times \mathcal{R}}, \end{aligned}$$

where  $(\mathcal{G} \times \mathcal{R})_{\delta}$  denotes a discretization (or meshing) of  $\mathcal{G} \times \mathcal{R}$  (i.e.,  $\delta$ -net of  $\mathcal{G} \times \mathcal{R}$ ), and the terms  $\|G_n - G_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}}\|_{\mathcal{G} \times \mathcal{R}}$  and  $\|Z_n^G \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}} - Z_n^G\|_{\mathcal{G} \times \mathcal{R}}$  capture the fluctuations (or oscillations) of  $G_n$  and  $Z_n^G$  relative to the meshing for each of the stochastic processes.  $\|\Pi_1 G_n - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_{\delta}}$  and  $\|\Pi_1 Z_n^G - Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_{\delta}}$  represent projections onto a Haar function space, where  $\Pi_1 G_n(h) = G_n \circ \Pi_1 h$ . The operator  $\Pi_1$  is a projection onto piecewise constant functions that respects the multiplicative structure of the  $G_n$  process. The final term  $\|\Pi_1 G_n - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_{\delta}}$  captures the coupling between the empirical process and the Gaussian process (on a  $\delta$ -net of  $\mathcal{G} \times \mathcal{R}$ , after the projection  $\Pi_1$ ).

A general result under uniform entropy integral conditions is presented in Section [SA-III.2](#), and a corollary under a VC-type condition is presented in Section [SA-III.3](#). The proofs exploit the existence of a

surrogate measure and normalizing transformation of  $\mathcal{G}$  with respect to  $\mathbb{P}_X$ , the law of  $\mathbf{x}_1$ , as developed in Section SA-II.2. The preliminary technical results differ from those in Section SA-II by explicitly leveraging the multiplicative structure of the empirical process, and are organized as follows.

- Section SA-III.1.1 introduces the class of *cylindered quasi-dyadic cell expansions* based on  $\mathbb{P}_Z$ , which can be viewed as a special case of the *quasi-dyadic cell expansions* from Definition SA.5 that leverages the multiplicative structure. This cell expansion is tailored to the multiplicative structure, with the upper layers corresponding to splits in the  $\mathbf{x}_i$ -direction and the lower layers handling divisions along the  $y_i$ -direction.
- Section SA-III.1.2 introduces an alternative to the  $L_2$  projection onto piecewise constant functions on the chosen cells: the *product-factorized projection*,  $\Pi_1$ . This projection exploits the multiplicative structure of  $G_n$ , allowing the empirical process to treat  $\mathbf{x}_i$  and  $y_i$  as independent in layers where cells divide along  $\mathcal{Y}$ , thereby isolating contributions from  $\mathcal{G}$  and  $\mathcal{R}$ . To analyze the projection errors  $\|G_n - \Pi_1 G_n\|_{(\mathcal{G} \times \mathcal{R})_\delta}$  and  $\|Z_n^G - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta}$ , we also define the  $L_2$  projection onto piecewise constant functions on the chosen cells,  $\Pi_0$ .
- Section SA-III.1.3 constructs the Gaussian process  $(Z_n^G(g, r) : (g, r) \in \mathcal{G} \times \mathcal{R})$ . These constructions are essentially the same as those in Section SA-II.1.3, relying on coupling binomial random variables with Gaussian random variables.
- Section SA-III.1.4 handles the meshing errors  $\|G_n - G_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta}\|_{\mathcal{G} \times \mathcal{R}}$  and  $\|Z_n^G \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta} - Z_n^G\|_{\mathcal{G} \times \mathcal{R}}$  using standard empirical process results, which give the contribution  $F(\delta)$  emerging from Talagrand’s inequality (Giné and Nickl, 2016, Theorem 3.3.9) combined with a standard maximal inequality (Chernozhukov *et al.*, 2014, Theorem 5.2). This allows us to focus on the error on the  $\delta$ -net to simply study  $\|G_n - Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta}$ .
- Section SA-III.1.5 addresses the strong approximation error  $\|\Pi_1 G_n - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta}$ . The multiplicative structure of  $G_n$  and the pre-factorization of coefficients in  $\Pi_1 G_n$  and  $\Pi_1 Z_n^G$  enable a new bound on the strong approximation error for the empirical process indexed by piecewise constant functions. Specifically, we establish a bound on  $\mathbb{E}[\|\Pi_1 G_n - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta}^2]$  that is polynomial in the number of splits along the  $y_i$ -direction and exponential in the number of splits along the  $\mathbf{x}_i$ -direction. This is a key step in achieving a Gaussian strong approximation rate that treats splits along the  $y_i$ -dimension as residual contributions.
- Section SA-III.1.6 addresses the projection errors  $\|G_n - \Pi_1 G_n\|_{(\mathcal{G} \times \mathcal{R})_\delta}$  and  $\|Z_n^G - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta}$ . We begin by comparing the two projections, bounding the differences  $\|\Pi_1 G_n - \Pi_0 G_n\|_{(\mathcal{G} \times \mathcal{R})_\delta}$  and  $\|\Pi_1 Z_n^G - \Pi_0 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta}$ . Next, we control the  $L_2$  projection errors  $\|G_n - \Pi_0 G_n\|_{(\mathcal{G} \times \mathcal{R})_\delta}$  and  $\|Z_n^G - \Pi_0 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta}$  using Bernstein inequality and similar arguments as in Section SA-II.1.6.

### SA-III.1 Preliminary Technical Results

This section presents preliminary technical results that are used to prove Theorem SA.1. Whenever possible, these results are presented at a higher level of generality, and therefore may be of independent theoretical interest. Throughout this section, we employ the following assumption.

**Assumption SA.2.** *Suppose  $(\mathbf{z}_i = (\mathbf{x}_i, y_i) : 1 \leq i \leq n)$  are i.i.d. random vectors taking values in  $(\mathbb{R}^{d+1}, \mathcal{B}(\mathbb{R}^{d+1}))$ , where  $(\mathbf{x}_1, y_1)$  has joint distribution  $\mathbb{P}_Z$ . Suppose  $\mathbf{x}_1$  has distribution  $\mathbb{P}_X$  supported on  $\mathcal{X} \subseteq \mathbb{R}^d$ ,  $y_1$  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  $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P}_X)$ .
- (ii)  $M_{\mathcal{G}, \mathcal{X}} < \infty$  and  $J_{\mathcal{X}}(1, \mathcal{G}, M_{\mathcal{G}, \mathcal{X}}) < \infty$ .
- (iii)  $\mathcal{R}$  is a real-valued pointwise measurable class of functions on  $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), \mathbb{P}_Y)$ .
- (iv)  $M_{\mathcal{R}, \mathcal{Y}}(y) + \mathbf{pTV}_{\mathcal{R}, (-|y|, |y|)} \leq \mathbf{v}(1 + |y|^\alpha)$  for all  $y \in \mathcal{Y}$ , for some  $\mathbf{v} > 0$ , and for some  $\alpha \geq 0$ . Furthermore, if  $\alpha > 0$ , then  $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(|y_i|) | \mathbf{x}_i = \mathbf{x}] \leq 2$ .
- (v)  $J_{\mathcal{Y}}(\mathcal{R}, M_{\mathcal{R}, \mathcal{Y}}, 1) < \infty$ .

Compared to the assumptions in Theorem 2, this assumption does not require the existence of a surrogate measure or a normalizing transformation. It will be applied in the analysis of terms in the error decomposition, where we work with the distribution  $\mathbb{P}_{\mathcal{Z}}$ , and an extra condition on the existence of Lebesgue density of  $\mathbb{P}_X$  is assumed whenever necessary (Section SA-III.1.6). The surrogate measure and the normalizing transformation will be used in the proof of Theorem SA.1 with the help of Section SA-II.2, providing greater flexibility in the data generating process.

### SA-III.1.1 Cells Expansions

**Definition SA.7** (Cylindered Quasi-Dyadic Expansion of  $\mathbb{R}^d$ ). *Let  $\mathbb{P}$  denote the joint distribution of  $(\mathbf{X}, Y)$ , a random vector taking values in  $(\mathbb{R}^d \times \mathbb{R}, \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}))$ , and let  $\mathbb{P}_X$  be the marginal distribution of  $\mathbf{X}$ . For a given  $\rho \geq 1$ , a collection of Borel measurable sets in  $\mathbb{R}^{d+1}$ ,  $\mathcal{C}_{M,N}(\mathbb{P}, \rho) = \{\mathcal{C}_{j,k} : 0 \leq k < 2^{M+N-j}, 0 \leq j \leq M+N\}$ , is called a cylindered quasi-dyadic expansion of  $\mathbb{R}^{d+1}$  of depth  $M$  for the main subspace  $\mathbb{R}^d$  and depth  $N$  for the multiplier subspace  $\mathbb{R}$  with respect to  $\mathbb{P}$  if the following conditions hold:*

1. For all  $N \leq j \leq M+N$ ,  $0 \leq k < 2^{M+N-j}$ , there exists a set  $\mathcal{X}_{j-N,k} \subseteq \mathbb{R}^d$  such that  $\mathcal{C}_{j,k} = \mathcal{X}_{j-N,k} \times \mathcal{Y}_{*,N,0}$ , with  $\mathcal{Y}_{*,N,0}$  a subset of  $\mathbb{R}$  and  $\mathbb{P}(\mathcal{C}_{M+N,0}) = 1$ . The collection  $\mathcal{C}_M(\mathbb{P}_X, \rho) = \{\mathcal{X}_{l,k} : 0 \leq l \leq M, 0 \leq k < 2^{M-l}\}$  forms a quasi-dyadic expansion of depth  $M$  with respect to  $\mathbb{P}_X$ .
2. For all  $0 \leq j < N$  and  $0 \leq k < 2^{M+N-j}$ , let  $l$  and  $m$  be the unique non-negative integers such that  $k = 2^{N-j}l + m$ . Then there exists a set  $\mathcal{Y}_{l,j,m} \subseteq \mathbb{R}$  such that  $\mathcal{C}_{j,k} = \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}$ . Moreover, for each  $0 \leq l < 2^M$ , the collection  $\{\mathcal{Y}_{l,j,m} : 0 \leq j < N, 0 \leq m < 2^{N-j}\}$  forms a dyadic expansion of depth  $N$  with respect to the conditional distribution  $\mathbb{P}(Y \in \cdot | \mathbf{X} \in \mathcal{X}_{0,l})$ , and  $\mathcal{Y}_{l,N,0} = \mathcal{Y}_{*,N,0}$ .

When  $\rho = 1$ ,  $\mathcal{C}_{M,N}(\mathbb{P}, 1)$  is called a cylindered dyadic expansion. For notational simplicity, we write  $\mathbf{p}_X[\mathcal{C}_{M,N}(\mathbb{P}, \rho)] = \{\mathcal{X}_{l,k} : 0 \leq l \leq M, 0 \leq k < 2^{M-l}\}$ .

**Definition SA.8** (Axis-Aligned Quasi-Dyadic Expansion of  $\mathbb{R}^d$ ). *Let  $\mathbb{P}$  denote the joint distribution of  $(\mathbf{X}, Y)$ , a random vector taking values in  $(\mathbb{R}^d \times \mathbb{R}, \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}))$ , and let  $\mathbb{P}_X$  be the marginal distribution of  $\mathbf{X}$ . For a given  $\rho \geq 1$ , a collection of Borel measurable sets in  $\mathbb{R}^{d+1}$ ,  $\mathcal{A}_{M,N}(\mathbb{P}, \rho) = \{\mathcal{C}_{j,k} : 0 \leq k < 2^{M+N-j}, 0 \leq j \leq M+N\}$ , is called an axis-aligned cylindered quasi-dyadic expansion of  $\mathbb{R}^{d+1}$  of depth  $M$  in the main subspace  $\mathbb{R}^d$  and depth  $N$  in the multiplier subspace  $\mathbb{R}$  with respect to  $\mathbb{P}$  if the following conditions hold:*

1.  $\mathcal{A}_{M,N}(\mathbb{P}, \rho)$  is a cylindered quasi-dyadic expansion of  $\mathbb{R}^{d+1}$ , of depth  $M$  for the main subspace  $\mathbb{R}^d$  and depth  $N$  for the multiplier subspace  $\mathbb{R}$ , with respect to  $\mathbb{P}$ .
2.  $\mathbf{p}_X[\mathcal{A}_{M,N}(\mathbb{P}, \rho)] = \{\mathcal{X}_{l,k} : 0 \leq l \leq M, 0 \leq k < 2^{M-l}\}$  forms an axis-aligned quasi-dyadic expansion of depth  $M$  with respect to  $\mathbb{P}_X$ .

When  $\rho = 1$ ,  $\mathcal{A}_{M,N}(\mathbb{P}, 1)$  is called an axis-aligned cylindered dyadic expansion.

### SA-III.1.2 Projection onto Piecewise Constant Functions

Consider a cylindered quasi-dyadic expansion  $\mathcal{C}_{M,N}(\mathbb{P}, \rho)$  where  $\mathbb{P}$  is the joint distribution of a random vector  $(\mathbf{X}, Y)$  taking values in  $(\mathbb{R}^d \times \mathbb{R}, \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}))$ . Define the span of the Haar basis over the terminal cells as described in Section SA-II.1.2, specifically

$$\mathcal{E}_{M+N} = \text{Span}\{\mathbb{1}_{\mathcal{C}_{0,k}} : 0 \leq k < 2^{M+N}\}.$$

For  $h \in L_2(\mathbb{P})$ , recall that the mean square projection of  $h$  onto  $\mathcal{E}_{M+N}$  is given by

$$\Pi_0(\mathcal{C}_{M,N}(\mathbb{P}, \rho))[h] = \sum_{0 \leq k < 2^{M+N}} \frac{\mathbb{1}_{\mathcal{C}_{0,k}}}{\mathbb{P}(\mathcal{C}_{0,k})} \int_{\mathcal{C}_{0,k}} h(\mathbf{u}) d\mathbb{P}(\mathbf{u}).$$

and the  $\beta$ -coefficients are defined by

$$\beta_{j,k}(h) = \frac{1}{\mathbb{P}(\mathcal{C}_{j,k})} \int_{\mathcal{C}_{j,k}} h(\mathbf{u}) d\mathbb{P}(\mathbf{u}), \quad \tilde{\beta}_{j,k}(h) = \beta_{j-1,2k}(h) - \beta_{j-1,2k+1}(h).$$

Then we still have

$$\Pi_0(\mathcal{C}_{M,N}(\mathbb{P}, \rho))[h] = \beta_{K,0}(h)e_{K,0} + \sum_{1 \leq j \leq K} \sum_{0 \leq k < 2^{K-j}} \tilde{\beta}_{j,k}(h)\tilde{e}_{j,k},$$

where

$$e_{j,k} = \mathbb{1}_{\mathcal{C}_{j,k}}, \quad \tilde{e}_{j,k} = \frac{\mathbb{P}(\mathcal{C}_{j-1,2k+1})}{\mathbb{P}(\mathcal{C}_{j,k})} e_{j-1,2k} - \frac{\mathbb{P}(\mathcal{C}_{j-1,2k})}{\mathbb{P}(\mathcal{C}_{j,k})} e_{j-1,2k+1},$$

for all  $(j, k) \in \mathcal{I}_{M+N} = \{(j, k) \in \mathbb{N} \times \mathbb{N} : 1 \leq j \leq M+N, 0 \leq k < 2^{M+N-j}\}$ . We refer to  $\Pi_0(\mathcal{C}_{M,N}(\mathbb{P}, \rho))$  as  $\Pi_0$  for simplicity.

To address the separable structure of  $g(\mathbf{X})r(Y)$ , we define the *product-factorized projection* from  $L_2(\mathbb{P})$  to  $\mathcal{E}_{M+N} = \text{Span}\{\mathcal{C}_{0,k} = \mathcal{X}_{0,l} \times \mathcal{Y}_{l,0,m} : 0 \leq l < 2^M, 0 \leq m < 2^N, k = 2^N l + m\}$ , defined as

$$\Pi_1(\mathcal{C}_{M,N}(\mathbb{P}, \rho))[g, r] = \gamma_{M+N,0}(g, r)e_{M+N,0} + \sum_{1 \leq j \leq M+N} \sum_{0 \leq k < 2^{M+N-j}} \tilde{\gamma}_{j,k}(g, r)\tilde{e}_{j,k}, \quad (\text{SA-9})$$

and

$$\gamma_{j,k}(g, r) = \begin{cases} \mathbb{E}[g(\mathbf{X})r(Y)|\mathbf{X} \in \mathcal{X}_{j-N,k}], & \text{if } N \leq j \leq M+N, \\ \mathbb{E}[g(\mathbf{X})|\mathbf{X} \in \mathcal{X}_{0,l}] \cdot \mathbb{E}[r(Y)|\mathbf{X} \in \mathcal{X}_{0,l}, Y \in \mathcal{Y}_{l,0,m}], & \text{if } j < N, k = 2^{N-j}l + m, \end{cases}$$

and  $\tilde{\gamma}_{j,k}(g, r) = \gamma_{j-1,2k}(g, r) - \gamma_{j-1,2k+1}(g, r)$ . We refer to  $\Pi_1(\mathcal{C}_{M,N}(\mathbb{P}, \rho))$  as  $\Pi_1$  for simplicity.

The Haar basis representation in Equation (SA-9) decomposes the function into layers of increasingly localized fluctuations. However, at lower layers ( $1 \leq j \leq N$ ), the local fluctuation is characterized by the *product-factorized projection*  $\mathbb{E}[g(\mathbf{X})|\mathbf{X} \in \mathcal{X}_{0,l}] \cdot \mathbb{E}[r(Y)|\mathbf{X} \in \mathcal{X}_{0,l}, Y \in \mathcal{Y}_{l,0,m}]$ , rather than  $\mathbb{E}[g(\mathbf{X})r(Y)|\mathbf{X} \in \mathcal{X}_{0,l} \times \mathcal{Y}_{l,0,m}]$ . This distinction makes  $\Pi_1(\mathcal{C}_{M,N}(\mathbb{P}, \rho))[g, r]$  generally different from  $\Pi_0(\mathcal{C}_{M,N}(\mathbb{P}, \rho))[g \cdot r]$ .

Now, we define the empirical processes indexed by projected functions. For any real valued functions  $g$

on  $\mathbb{R}^d$  and  $r$  on  $\mathbb{R}$  such that  $\int_{\mathbb{R}^d} \int_{\mathbb{R}} g(\mathbf{x})^2 \mathbb{P}(dyd\mathbf{x}) < \infty$  and  $\int_{\mathbb{R}^d} \int_{\mathbb{R}} r(y)^2 \mathbb{P}(dyd\mathbf{x}) < \infty$ , we define

$$\begin{aligned}\Pi_1 G_n(g, r) &= X_n \circ \Pi_1[\mathcal{C}_{M,N}(\mathbb{P}, \rho)](g, r), \\ \Pi_0 G_n(g, r) &= X_n \circ \Pi_0[\mathcal{C}_{M,N}(\mathbb{P}, \rho)](gr),\end{aligned}\tag{SA-10}$$

recalling  $(X_n(f) : f \in \mathcal{F})$  is the empirical process based on a random sample  $(\mathbf{z}_i = (\mathbf{x}_i, y_i) : 1 \leq i \leq n)$  with

$$X_n(f) = n^{-1/2} \sum_{i=1}^n (f(\mathbf{x}_i, y_i) - \mathbb{E}[f(\mathbf{x}_i, y_i)]).$$

### SA-III.1.3 Strong Approximation Construction

In this section, we construct the Gaussian process  $Z_n^G$  (along with some auxiliary Gaussian processes) on a possibly enlarged probability space to couple with the empirical process  $G_n$ .

**Lemma SA.12.** *Suppose Assumption SA.2 holds, and a cylindered quasi-dyadic expansion  $\mathcal{C}_K(\mathbb{P}_Z, \rho)$  is given. Then,  $(\mathcal{G} \cdot \mathcal{R}) \cup \Pi_0(\mathcal{G} \times \mathcal{R}) \cup \Pi_1(\mathcal{G} \times \mathcal{R})$  is  $\mathbb{P}_Z$ -pregaussian.*

**Proof of Lemma SA.12.** By the entropy integral conditions on  $\mathcal{G}$  and  $\mathcal{R}$  and Definitions 10 and SA.4,

$$\begin{aligned}J_{\mathcal{X} \times \mathcal{Y}}(\mathcal{G} \cdot \mathcal{R}, M_{\mathcal{G}, \mathcal{X}} M_{\mathcal{R}, \mathcal{Y}}, \delta) &= J_{\mathcal{X} \times \mathcal{Y}}(\mathcal{G} \times \mathcal{R}, M_{\mathcal{G}, \mathcal{X}} M_{\mathcal{R}, \mathcal{Y}}, \delta) \\ &\leq \sqrt{2} J_{\mathcal{X} \times \mathcal{Y}}(\bar{\mathcal{G}}, M_{\mathcal{G}, \mathcal{X}}, \delta/\sqrt{2}) + \sqrt{2} J_{\mathcal{X} \times \mathcal{Y}}(\bar{\mathcal{R}}, M_{\mathcal{R}, \mathcal{Y}}, \delta/\sqrt{2}) \\ &\leq \sqrt{2} J_{\mathcal{X}}(\mathcal{G}, M_{\mathcal{G}, \mathcal{X}}, \delta/\sqrt{2}) + \sqrt{2} J_{\mathcal{Y}}(\mathcal{R}, M_{\mathcal{R}, \mathcal{Y}}, \delta/\sqrt{2})\end{aligned}$$

where  $\bar{\mathcal{G}} = \{(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \mapsto g(\mathbf{x}) : g \in \mathcal{G}\}$  and  $\bar{\mathcal{R}} = \{(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \mapsto r(y) : r \in \mathcal{R}\}$ .

Claim 1: For all  $0 < \delta < 1$ ,

$$J_{\mathcal{X} \times \mathcal{Y}}(\Pi_0(\mathcal{G} \times \mathcal{R}), c_{\mathbf{v}, \alpha} M_{\mathcal{G}, \mathcal{X}} N^\alpha, \delta) \leq J_{\mathcal{X} \times \mathcal{Y}}(\mathcal{G} \times \mathcal{R}, M_{\mathcal{G}, \mathcal{X}} M_{\mathcal{R}, \mathcal{Y}}, \delta),$$

where  $c_{\mathbf{v}, \alpha} = \mathbf{v} \max\{1 + (2\alpha)^{\frac{\alpha}{2}}, 1 + (4\alpha)^\alpha\}$ .

Proof of Claim 1: We consider the two cases of whether  $\alpha > 0$  in Assumption SA.2 (iv) separately.

If  $\alpha > 0$ , by Step 2 in Definition SA.7,  $\max_{0 \leq l < 2^{M+N}} \mathbb{E}[\exp(y_i/(N \log 2)) | (\mathbf{x}_i, y_i) \in \mathcal{C}_{0,l}] \leq 2$ . Hence

$$\begin{aligned}\max_{0 \leq l < 2^{M+N}} \sup_{r \in \mathcal{R}} \mathbb{E}[|r(y_i)| | (\mathbf{x}_i, y_i) \in \mathcal{C}_{0,l}] &\leq \mathbf{v} (1 + \max_{0 \leq l < 2^{M+N}} \mathbb{E}[|y_i|^\alpha | (\mathbf{x}_i, y_i) \in \mathcal{C}_{0,l}]) \\ &\leq \mathbf{v} (1 + (2N\sqrt{\alpha})^\alpha).\end{aligned}\tag{SA-11}$$

Definition of  $\Pi_0$  then implies

$$\sup_{g \in \mathcal{G}} \sup_{r \in \mathcal{R}} \sup_{(\mathbf{x}, y) \in \mathcal{C}_{M+N, 0}} |\Pi_0(gr)(\mathbf{x}, y)| \leq c_{\mathbf{v}, \alpha} M_{\mathcal{G}, \mathcal{X}} N^\alpha.\tag{SA-12}$$

Moreover, if  $\alpha = 0$ , Assumption SA.2 (iv) implies  $M_{\mathcal{R}, \mathcal{Y}} \leq 1$ , Equations (SA-11), (SA-12) hold with  $\alpha = 0$ .

Let  $Q$  be a finite discrete measure on  $\mathcal{X} \times \mathcal{Y}$ . Definition of  $\Pi_0$  and Jensen's inequality implies

$$\begin{aligned} \|\Pi_0 f - \Pi_0 g\|_{\tilde{Q},2}^2 &\leq \sum_{0 \leq k < 2^{M+N}} Q(C_{0,k}) (2^{M+N} \int_{C_{0,k}} f - g d\mathbb{P}_Z)^2 \\ &\leq \sum_{0 \leq k < 2^{M+N}} Q(C_{0,k}) 2^{M+N} \int_{C_{0,k}} (f - g)^2 d\mathbb{P}_Z, \quad \forall f, g \in \mathcal{G} \cdot \mathcal{R}. \end{aligned}$$

Define a measure  $\tilde{Q}$  such that for any  $A \in \mathcal{B}(\mathbb{R}^d \times \mathbb{R})$ ,  $\tilde{Q}(A) = \sum_{0 \leq k < 2^{M+N}} Q(C_{0,k}) 2^{M+N} \mathbb{P}_Z(A \cap C_{0,k})$ , then

$$\|\Pi_0 f - \Pi_0 g\|_{\tilde{Q},2}^2 \leq \|f - g\|_{\tilde{Q},2}^2, \quad \forall f, g \in \mathcal{G} \cdot \mathcal{R}.$$

Lemma SA.15 implies that there exists an  $\delta c_{v,\alpha} M_{\mathcal{G},\mathcal{X}} N^\alpha$ -net  $\mathcal{L}$  of  $\mathcal{G} \times \mathcal{R}$  with cardinality no greater than  $N_{\mathcal{G} \times \mathcal{R}, \mathcal{X} \times \mathcal{Y}}(\delta, \|M_{\mathcal{G},\mathcal{X}} M_{\mathcal{R},\mathcal{Y}}\|_{\tilde{Q},2})$  such that for all  $f \in \Pi_0(\mathcal{G} \times \mathcal{R})$ , there exists  $g \in \mathcal{L}$  such that

$$\|f - g\|_{\tilde{Q},2}^2 \leq \delta^2 \|M_{\mathcal{G},\mathcal{X}} M_{\mathcal{R},\mathcal{Y}}\|_{\tilde{Q},2}^2 \leq \delta^2 (c_{v,\alpha} M_{\mathcal{G},\mathcal{X}} N^\alpha)^2.$$

The claim then follows.

Claim 2: For all  $0 < \delta < 1$ ,

$$J_{\mathcal{X} \times \mathcal{Y}}(\Pi_1(\mathcal{G} \times \mathcal{R}), c_{v,\alpha} M_{\mathcal{G},\mathcal{X}} N^\alpha, \delta) \lesssim J_{\mathcal{X} \times \mathcal{Y}}(\mathcal{G} \times \mathcal{R}, M_{\mathcal{G},\mathcal{X}} M_{\mathcal{R},\mathcal{Y}}, \delta/3).$$

Proof of Claim 2: Definition SA.5 and the definition of product factorized projection imply that for the upper layers with  $N \leq j \leq M + N$ ,

$$\gamma_{j,k}(g, r) = \mathbb{E}[g(\mathbf{x}_i) r(y_i) | \mathbf{x}_i \in \mathcal{X}_{j-N,k}] = \mathbb{E}[g(\mathbf{x}_i) r(y_i) | (\mathbf{x}_i, y_i) \in \mathcal{C}_{j-N,k}],$$

that is, the coefficients coincide with those from the mean square projection. Take  $\mathcal{C}_{M,0} = \{\mathcal{C}_{j,k} : N \leq j \leq M + N, 0 \leq k < 2^{M+N-j}\}$  to be the collection of all upper layer cells down to the  $N$ -th layer, then

$$\Pi_1[\mathcal{C}_{M,0}(\mathbb{P}_Z, \rho)](g, r) = \Pi_0[\mathcal{C}_{M,0}(\mathbb{P}_Z, \rho)](gr), \quad g \in \mathcal{G}, r \in \mathcal{R}.$$

For the lower layers  $0 \leq j < N$ , suppose  $\tilde{\mathbb{P}}_Z$  is a mapping from  $\mathcal{B}(\mathbb{R}^{d+1})$  to  $[0, 1]$  such that

$$\begin{aligned} \tilde{\mathbb{P}}_Z(E) = \inf \left\{ \sum_{\ell=1}^{\mathfrak{L}} \sum_{0 \leq l < 2^M} \sum_{0 \leq m < 2^N} \mathbb{E}[\mathbb{1}(\mathbf{x}_i \in A_\ell) | \mathbf{x}_i \in \mathcal{X}_{0,l}] \cdot \mathbb{E}[\mathbb{1}(y_i \in B_\ell) | \mathbf{x}_i \in \mathcal{X}_{0,l}, y_i \in \mathcal{Y}_{l,0,m}] : \right. \\ \left. E \subseteq \sqcup_{\ell=1}^{\mathfrak{L}} A_\ell \times B_\ell \text{ with } A_\ell \times B_\ell, 1 \leq l \leq \mathfrak{L} \in \mathbb{N}, \text{ disjoint rectangles in } \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}) \right\}, \end{aligned}$$

with  $E \in \mathcal{B}(\mathbb{R}^{d+1})$ . It follows that  $\tilde{\mathbb{P}}_Z$  defines a probability measure on  $(\mathbb{R}^{d+1}, \mathcal{B}(\mathbb{R}^{d+1}))$ , and

$$\begin{aligned}
\gamma_{j,k}(g, r) &= \mathbb{E}[g(\mathbf{x}_i)|\mathbf{x}_i \in \mathcal{X}_{0,l}] \cdot \mathbb{E}[r(y_i)|\mathbf{x}_i \in \mathcal{X}_{0,l}, y_i \in \mathcal{Y}_{l,j,m}] \\
&= \sum_{m': \mathcal{Y}_{l,j,m'} \subseteq \mathcal{Y}_{l,j,m}} 2^{-j} \mathbb{E}[g(\mathbf{x}_i)|\mathbf{x}_i \in \mathcal{X}_{0,l}] \cdot \mathbb{E}[r(y_i)|\mathbf{x}_i \in \mathcal{X}_{0,l}, y_i \in \mathcal{Y}_{l,0,m'}] \\
&= \sum_{m': \mathcal{Y}_{l,j,m'} \subseteq \mathcal{Y}_{l,j,m}} 2^{-j} \mathbb{E}_{\tilde{\mathbb{P}}_Z}[g(\mathbf{x}_i)r(y_i)|\mathbf{x}_i \in \mathcal{X}_{0,l}, y_i \in \mathcal{Y}_{l,j,m'}] \\
&= \mathbb{E}_{\tilde{\mathbb{P}}_Z}[g(\mathbf{x}_i)r(y_i)|\mathbf{x}_i \in \mathcal{X}_{0,l}, y_i \in \mathcal{Y}_{l,j,m}] \\
&= \mathbb{E}_{\tilde{\mathbb{P}}_Z}[g(\mathbf{x}_i)r(y_i)|(\mathbf{x}_i, y_i) \in \mathcal{C}_{j,k}], \quad 0 \leq j < N, 0 \leq k < 2^{M+N-j},
\end{aligned}$$

where  $\mathbb{E}_{\tilde{\mathbb{P}}_Z}$  means the expectation is taken with  $(\mathbf{x}_i, y_i)$  following the law of  $\tilde{\mathbb{P}}_Z$  instead of  $\mathbb{P}_Z$ . This implies

$$\Pi_1[\mathcal{C}_{M,N}(\mathbb{P}_Z, \rho)](g, r) - \Pi_1[\mathcal{C}_{M,0}(\mathbb{P}_Z, \rho)](g, r) = \Pi_0[\mathcal{C}_{M,N}(\tilde{\mathbb{P}}_Z, \rho)](gr) - \Pi_0[\mathcal{C}_{M,0}(\tilde{\mathbb{P}}_Z, \rho)](gr), \quad g \in \mathcal{G}, r \in \mathcal{R}.$$

We can then express the  $\Pi_1$  projection of  $(g, r)$  as three  $L_2$  projections as follows:

$$\begin{aligned}
\Pi_1[\mathcal{C}_{M,N}(\mathbb{P}_Z, \rho)](g, r) &= \Pi_1[\mathcal{C}_{M,0}(\mathbb{P}_Z, \rho)](g, r) + \Pi_1[\mathcal{C}_{M,N}(\mathbb{P}_Z, \rho)](g, r) - \Pi_1[\mathcal{C}_{M,0}(\mathbb{P}_Z, \rho)](g, r) \\
&= \Pi_0[\mathcal{C}_{M,0}(\mathbb{P}_Z, \rho)](gr) + \Pi_0[\mathcal{C}_{M,N}(\tilde{\mathbb{P}}_Z, \rho)](gr) - \Pi_0[\mathcal{C}_{M,0}(\tilde{\mathbb{P}}_Z, \rho)](gr), \quad g \in \mathcal{G}, r \in \mathcal{R}.
\end{aligned}$$

Since  $\|\Pi_0[\mathcal{C}_{M,N}(\tilde{\mathbb{P}}_Z, \rho)]\|_{\mathcal{G} \times \mathcal{R}} \leq c_{v,\alpha} \mathbf{M}_{\mathcal{G},\mathcal{X}} N^\alpha$ , Claim 1 applies to all of the three terms:

$$\begin{aligned}
&J_{\mathcal{X} \times \mathcal{Y}}(\Pi_0[\mathcal{C}_{M,N}(\mathbb{P}_Z, \rho)](\mathcal{G} \times \mathcal{R}), c_{v,\alpha} \mathbf{M}_{\mathcal{G},\mathcal{X}} N^\alpha, \delta) + J_{\mathcal{X} \times \mathcal{Y}}(\Pi_0[\mathcal{C}_{M,N}(\tilde{\mathbb{P}}_Z, \rho)](\mathcal{G} \times \mathcal{R}), c_{v,\alpha} \mathbf{M}_{\mathcal{G},\mathcal{X}} N^\alpha, \delta) \\
&+ J_{\mathcal{X} \times \mathcal{Y}}(\Pi_0[\mathcal{C}_{M,0}(\tilde{\mathbb{P}}_Z, \rho)](\mathcal{G} \times \mathcal{R}), c_{v,\alpha} \mathbf{M}_{\mathcal{G},\mathcal{X}} N^\alpha, \delta) \lesssim J_{\mathcal{X} \times \mathcal{Y}}(\mathcal{G} \times \mathcal{R}, \mathbf{M}_{\mathcal{G},\mathcal{X}} M_{\mathcal{R},\mathcal{Y}}, \delta).
\end{aligned}$$

Then Claim 2 follows from Claim 1.

Putting together,

$$\begin{aligned}
&J_{\mathcal{X} \times \mathcal{Y}}((\mathcal{G} \times \mathcal{R}) \cup \Pi_0(\mathcal{G} \times \mathcal{R}) \cup \Pi_1(\mathcal{G} \times \mathcal{R}), \mathbf{M}_{\mathcal{G},\mathcal{X}} M_{\mathcal{R},\mathcal{Y}} + c_{v,\alpha} \mathbf{M}_{\mathcal{G},\mathcal{X}} N^\alpha, 1) \\
&\lesssim J_{\mathcal{X}}(\mathcal{G}, \mathbf{M}_{\mathcal{G},\mathcal{X}}, 1) + J_{\mathcal{Y}}(\mathcal{R}, M_{\mathcal{R},\mathcal{Y}}, 1) < \infty,
\end{aligned}$$

and the conclusion follows from separability of  $\mathcal{G}$  and  $\mathcal{R}$ , and [van der Vaart and Wellner \(2013, Corollary 2.2.9\)](#).  $\square$

The construction of the Gaussian process essentially follows from the arguments in Section [SA-II.1.3](#) with  $\mathbf{z}_i$ 's replacing  $\mathbf{x}_i$ 's. (Recall that  $\mathbf{z}_i = (\mathbf{x}_i, y_i)$  in this section.) We start with a Gaussian process indexed by  $(\mathcal{G} \cdot \mathcal{R}) \cup \Pi_0(\mathcal{G} \times \mathcal{R}) \cup \Pi_1(\mathcal{G} \times \mathcal{R})$  with almost sure continuous sample paths, and take conditional quantile transformations of Gaussian process indexed by  $\mathbb{1}_{\mathcal{C}_{j,k}}$  to construct counts of  $(\mathbf{x}_i, y_i)$ 's on the cells  $\mathcal{C}_{j,k}$ 's. By a Skorohod embedding argument, this Gaussian process can be taken on a possibly enriched probability space. More precisely, we have the following result.

**Lemma SA.13.** *Suppose Assumption [SA.2](#) holds and a cylindered dyadic expansion  $\mathcal{C}_{M,N}(\mathbb{P}_Z, 1)$  is given. Then on a possibly enlarged probability space, there exists a  $\mathbb{P}_Z$ -Brownian bridge  $B_n$  indexed by  $\mathcal{F} = (\mathcal{G} \cdot \mathcal{R}) \cup \Pi_0(\mathcal{G} \times \mathcal{R}) \cup \Pi_1(\mathcal{G} \times \mathcal{R})$  with almost sure continuous trajectories on  $(\mathcal{F}, \mathfrak{d}_{\mathbb{P}_Z})$  such that for any  $f \in \mathcal{F}$  and any*

$x > 0$ ,

$$\mathbb{P} \left( \left| \sum_{i=1}^n f(\mathbf{x}_i, y_i) - \sqrt{n} B_n(f) \right| \geq 24 \sqrt{\|f\|_{\mathfrak{E}_{M+N}}^2 x} + 4 \sqrt{\mathfrak{C}_{\{f\}, M+N} x} \right) \leq 2 \exp(-x),$$

where for both  $\|f\|_{\mathfrak{E}_{M+N}}^2$  and  $\mathfrak{C}_{\{f\}, M+N}$  are defined in Lemma SA.3.

**Proof of Lemma SA.13.** The result follows from Lemma SA.12 and Lemma SA.3 with  $(\mathbf{x}_i, y_i)$  replacing  $\mathbf{x}_i$ .  $\square$

**Lemma SA.14.** Suppose Assumption SA.2 holds and a cylindered quasi-dyadic expansion  $\mathfrak{C}_{M,N}(\mathbb{P}_Z, \rho)$  with  $\rho > 1$  is given. Then on a possibly enlarged probability space, there exists a  $\mathbb{P}_Z$ -Brownian bridge  $B_n$  indexed by  $\mathcal{F} = (\mathcal{G} \cdot \mathcal{R}) \cup \Pi_0(\mathcal{G} \times \mathcal{R}) \cup \Pi_1(\mathcal{G} \times \mathcal{R})$  with almost sure continuous trajectories on  $(\mathcal{F}, \mathfrak{d}_{\mathbb{P}_Z})$  such that for any  $f \in \mathcal{F}$  and  $x > 0$ ,

$$\mathbb{P} \left( \left| \sum_{i=1}^n f(\mathbf{x}_i, y_i) - \sqrt{n} B_n(f) \right| \geq C_\rho \sqrt{\|f\|_{\mathfrak{E}_{M+N}}^2 x} + C_\rho \sqrt{\mathfrak{C}_{\{f\}, M+N} x} \right) \leq 2 \exp(-x) + 2^{M+2} \exp(-C_\rho n 2^{-M}),$$

where  $C_\rho$  is a constant that only depends on  $\rho$ .

**Proof of Lemma SA.14.** Replacing  $\mathbf{x}_i$  by  $\mathbf{z}_i = (\mathbf{x}_i, y_i)$  in Section SA-II.1.3 (and with the help of the pregaussian lemma SA.12), suppose we constructed as therein on a possibly enlarged probability space the i.i.d standard Gaussian random variables  $(\tilde{\xi}_{j,k} : (j,k) \in \mathcal{I}_{M+N})$  and the Binomial counts  $(U_{j,k} : (j,k) \in \mathcal{J}_{M+N}) = (\sum_{i=1}^n e_{j,k}(\mathbf{z}_i) : (j,k) \in \mathcal{J}_{M+N})$ . Again, we take  $\tilde{U}_{j,k} = U_{j-1,2k} - U_{j-1,2k+1}$  for  $(j,k) \in \mathcal{I}_{M+N}$ . By Definition SA.7, the upper layer cells ( $N \leq j \leq M+N$ ) may not be dyadic with respect to  $\mathbb{P}_Z$ , but the lower layer cells ( $0 \leq j < N$ ) are. Tusnady's Lemma (Bretagnolle and Massart, 1989, Lemma 4) and Lemma SA.4 then imply whenever the event  $\mathcal{A}$  holds, with

$$\mathcal{A} = \{|\tilde{U}_{j,k}| \leq c_{1,\rho} U_{j,k}, \text{ for all } N \leq j \leq M+N, 0 \leq k < 2^{M+N-j}\},$$

we know the following relations hold almost surely in  $\mathbb{P}_Z$ ,

$$\begin{aligned} \left| \tilde{U}_{j,k} - \sqrt{U_{j,k} \frac{\mathbb{P}_Z(\mathcal{C}_{j-1,2k}) \mathbb{P}_Z(\mathcal{C}_{j-1,2k+1})}{\mathbb{P}_Z(\mathcal{C}_{j,k})^2}} \tilde{\xi}_{j,k} \right| &< c_{2,\rho} \tilde{\xi}_{j,k}^2 + c_{3,\rho}, \\ \left| \tilde{U}_{j,k} \right| &\leq 1/c_{0,\rho} + 2 \sqrt{\frac{\mathbb{P}_Z(\mathcal{C}_{j-1,2k}) \mathbb{P}_Z(\mathcal{C}_{j-1,2k+1})}{\mathbb{P}_Z(\mathcal{C}_{j,k})^2}} U_{j,k} |\tilde{\xi}_{j,k}|, \end{aligned}$$

for all  $(j,k) \in \mathcal{I}_{M+N}$ , and where  $c_{0,\rho}, c_{1,\rho}, c_{2,\rho}, c_{3,\rho}$  are constants that only depends on  $\rho$ . By similar argument as in the proof for Lemma SA.5,  $\mathbb{P}(\mathcal{A}^c) \leq 3 \cdot 2^M \exp(-\min\{c_{1,\rho}^2/3, 1/8\} \rho^{-1} n 2^{-M})$ . The rest of the proof follows from Lemma SA.5 by replacing  $\mathbf{x}_i$  with  $(\mathbf{x}_i, y_i)$ .  $\square$

The above two lemmas allow for constructions of Gaussian processes and projected Gaussian processes as counterparts of the empirical processes in Section SA-II.1.3. In particular, we take  $Z_n^G, \Pi_0 Z_n^G, \Pi_1 Z_n^G$  to be the empirical processes indexed by  $\mathcal{G} \times \mathcal{R}$  such that

$$Z_n^G(g, r) = B_n(g \cdot r), \quad (g, r) \in \mathcal{G} \times \mathcal{R}. \quad (\text{SA-13})$$



We also define the following ancillary processes for analysis:

$$\Pi_0 Z_n^G(g, r) = B_n(\Pi_0[g \cdot r]), \quad \Pi_1 Z_n^G(g, r) = B_n(\Pi_1[g, r]), \quad (g, r) \in \mathcal{G} \times \mathcal{R}. \quad (\text{SA-14})$$

In particular,  $(Z_n^G(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  has almost sure continuous trajectories in  $(\mathcal{G} \times \mathcal{R}, \mathfrak{d}_{\mathbb{P}_Z})$

The following ancillary lemma for uniform covering number and uniform entropy integrals is used in the proof of Lemma SA.12.

**Lemma SA.15** (Covering Number using Covariance Semi-metric). *Assume  $\mathcal{F}$  is a class of functions from a measurable space  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$  to  $\mathbb{R}$  with envelope function  $M_{\mathcal{F}, \mathcal{X}}$ . Let  $P$  be a probability measure on  $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ . Then, for any  $0 < \varepsilon < 1$ ,*

$$N(\mathcal{F}, \|\cdot\|_{P,2}, \varepsilon \|M_{\mathcal{F}, \mathcal{X}}\|_{P,2}) \leq \mathbb{N}_{\mathcal{F}, \mathcal{X}}(\varepsilon, M_{\mathcal{F}, \mathcal{X}}).$$

**Proof of Lemma SA.15.** The proof essentially follows from the arguments for (van der Vaart and Wellner, 2013, Theorem 2.5.2), but we present here for completeness. Define  $\mathcal{H} = \{(f - g)^2 : f, g \in \mathcal{F}\} \cup \{M_{\mathcal{F}, \mathcal{X}}\}$ . Then, for all  $0 < \varepsilon < 1$ ,

$$\sup_Q N(\mathcal{H}, \|\cdot\|_{Q,1}, \varepsilon \|M_{\mathcal{F}, \mathcal{X}}^2\|_{Q,1}) \leq \sup_Q N(\mathcal{H}, \|\cdot\|_{Q,1}, \varepsilon \|M_{\mathcal{F}, \mathcal{X}}^2\|_{Q,2}) \leq \sup_Q N(\mathcal{F}, \|\cdot\|_{Q,1}, \varepsilon \|M_{\mathcal{F}, \mathcal{X}}\|_{Q,1})^2,$$

where the supremums are all taken over finite discrete measures on  $\mathcal{X}$ . By Theorem 2.4.3 in van der Vaart and Wellner (2013),  $\mathcal{H}$  is Glivenko-Cantelli. Let  $X_1, X_2, \dots$  be a sequence of i.i.d. random variables with distribution  $P$ . Define  $Q_N = \frac{1}{N} \sum_{j=1}^N \delta_{X_j}$ . Let  $0 < \varepsilon < 1$  and  $\delta > 0$ . Then there exists  $N \in \mathbb{N}$  and a realization  $x_1, \dots, x_N$  of  $X_1, \dots, X_N$  such that if we denote  $P_N = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}$ , then for all  $f_1, f_2 \in \mathcal{F}$ ,

$$\begin{aligned} \left| \|f_1 - f_2\|_{P,2}^2 - \|f_1 - f_2\|_{P_N,2}^2 \right| &\leq \delta^2 \varepsilon^2 \|M_{\mathcal{F}, \mathcal{X}}\|_{P,2}^2, \\ \|M_{\mathcal{F}, \mathcal{X}}\|_{P,2} - \|M_{\mathcal{F}, \mathcal{X}}\|_{P_N,2} &\leq \delta \|M_{\mathcal{F}, \mathcal{X}}\|_{P,2}. \end{aligned}$$

Since  $P_N$  is a finite discrete measure on  $\mathcal{X}$ , there exists an  $\varepsilon \|M_{\mathcal{F}, \mathcal{X}}\|_{P_N}$ -net,  $\mathcal{G}$ , of  $\mathcal{F}$  with minimal cardinality such that for all  $f \in \mathcal{F}$ , there exists  $f_0 \in \mathcal{G}$  such that  $\|f - f_0\|_{P_N,2} \leq \varepsilon \|M_{\mathcal{F}, \mathcal{X}}\|_{P_N,2} \leq \varepsilon (\|M_{\mathcal{F}, \mathcal{X}}\|_{P,2} + \delta \|M_{\mathcal{F}, \mathcal{X}}\|_{P,2}) \leq (1 + \delta) \varepsilon \|M_{\mathcal{F}, \mathcal{X}}\|_{P,2}$ . It follows that for all  $f \in \mathcal{F}$ , there exists  $g \in \mathcal{G}$  such that

$$\|f - g\|_{P,2} \leq \|f - g\|_{P_N,2} + \|f - g\|_{P,2} - \|f - g\|_{P_N,2} \leq (1 + 2\delta) \varepsilon \|M_{\mathcal{F}, \mathcal{X}}\|_{P,2},$$

Hence,  $N(\mathcal{F}, \|\cdot\|_{P,2}, \varepsilon \|M_{\mathcal{F}, \mathcal{X}}\|_{P,2}) \leq \mathbb{N}_{\mathcal{F}, \mathcal{X}}(\varepsilon / (1 + 2\delta), M_{\mathcal{F}, \mathcal{X}})$ . Take  $\delta \rightarrow 0$  to obtain the desired result.  $\square$

#### SA-III.1.4 Meshing Error

To simplify notation, the parameters of  $\mathcal{G}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{X}$ , and the index  $\mathcal{X}$  is omitted where there is no ambiguity; the parameters of  $\mathcal{R}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{Y}$ , and the index  $\mathcal{Y}$  is omitted where there is no ambiguity; and the parameters of  $\mathcal{G} \times \mathcal{R}$  (Definitions 4 to 12, SA.3, SA.4) are taken with  $\mathcal{C} = \mathcal{X} \times \mathcal{Y}$ , and the index  $\mathcal{X} \times \mathcal{Y}$  is omitted where there is no ambiguity. We also define, for  $\delta \in (0, 1]$ ,

$$\mathbb{N}(\delta) = \mathbb{N}_{\mathcal{G}}(\delta/\sqrt{2}, M_{\mathcal{G}}) \mathbb{N}_{\mathcal{R}}(\delta/\sqrt{2}, M_{\mathcal{R}})$$

and

$$J(\delta) = \sqrt{2}J(\mathcal{G}, \mathbf{M}_{\mathcal{G}}, \delta/\sqrt{2}) + \sqrt{2}J(\mathcal{R}, M_{\mathcal{R}}, \delta/\sqrt{2}).$$

For  $0 < \delta \leq 1$ , consider a  $\delta\mathbf{M}_{\mathcal{G}}\|M_{\mathcal{R}}\|_{\mathbb{P}_{\mathcal{Y},2}}$ -net of  $(\mathcal{G} \times \mathcal{R}, \|\cdot\|_{\mathbb{P}_{\mathcal{Z},2}})$ , denoted by  $(\mathcal{G} \times \mathcal{R})_{\delta}$ , with cardinality at most  $N_{\mathcal{G} \times \mathcal{R}}(\delta, \mathbf{M}_{\mathcal{G}}\|M_{\mathcal{R}}\|_{\mathbb{P}_{\mathcal{Y},2}})$ . Define the projection onto the  $\delta$ -net as a mapping  $\pi_{(\mathcal{G} \times \mathcal{R})_{\delta}} : \mathcal{G} \times \mathcal{R} \rightarrow \mathcal{G} \times \mathcal{R}$  such that  $\|\pi_{(\mathcal{G} \times \mathcal{R})_{\delta}}(g, r) - gr\|_{\mathbb{P}_{\mathcal{Z},2}} \leq \delta\mathbf{M}_{\mathcal{G}}\|M_{\mathcal{R}}\|_{\mathbb{P}_{\mathcal{Y},2}}$  for all  $g \in \mathcal{G}$  and  $r \in \mathcal{R}$ .

**Lemma SA.16.** *Suppose Assumption SA.2 holds, a cylindered quasi-dyadic expansion  $\mathcal{C}_{M,N}(\mathbb{P}_{\mathcal{Z}}, \rho)$  is given,  $(Z_n^G(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  is the Gaussian process constructed as in Equation (SA-13) on a possibly enlarged probability space, and  $(\mathcal{G} \times \mathcal{R})_{\delta}$  is chosen in Section SA-III.1.4. For all  $t > 0$  and  $0 < \delta < 1$ ,*

$$\mathbb{P}[\|G_n - G_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}}\|_{\mathcal{G} \times \mathcal{R}} + \|Z_n^G \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}} - Z_n^G\|_{\mathcal{G} \times \mathcal{R}} > C_1 c_{v,\alpha} \mathbf{F}_n^G(t, \delta)] \leq 8 \exp(-t),$$

where  $c_{v,\alpha} = v(1 + (2\alpha)^{\frac{\alpha}{2}})$  and

$$\mathbf{F}_n^G(t, \delta) = J(\delta)\mathbf{M}_{\mathcal{G}} + \frac{(\log n)^{\alpha/2}\mathbf{M}_{\mathcal{G}}J^2(\delta)}{\delta^2\sqrt{n}} + \frac{\mathbf{M}_{\mathcal{G}}}{\sqrt{n}}t + (\log n)^{\alpha}\frac{\mathbf{M}_{\mathcal{G}}}{\sqrt{n}}t^{\alpha}.$$

**Proof of Lemma SA.16.** By standard empirical process arguments, we can show for any  $0 < \delta < 1$ ,  $N_{\mathcal{G} \times \mathcal{R}}(\delta, \mathbf{M}_{\mathcal{G}}M_{\mathcal{R}}) \leq N(\delta)$  and  $J(\delta, \mathcal{G} \times \mathcal{R}, \mathbf{M}_{\mathcal{G}}M_{\mathcal{R}}) \leq J(\delta)$ . By definition of  $\pi_{(\mathcal{G} \times \mathcal{R})_{\delta}}$ ,  $\|\pi_{(\mathcal{G} \times \mathcal{R})_{\delta}}h - h\|_{\mathbb{P}_{\mathcal{Z},2}} \leq \delta\|\mathbf{M}_{\mathcal{G}}M_{\mathcal{R}}\|_{\mathbb{P}_{\mathcal{Z},2}} = \delta\mathbf{M}_{\mathcal{G}}\|M_{\mathcal{R}}\|_{\mathbb{P}_{\mathcal{Y},2}}$ . Take  $\mathcal{L} = \{h - \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}}h : h \in \mathcal{G} \times \mathcal{R}\}$ . Then, by Theorem 5.2 in [Chernozhukov et al. \(2014\)](#),

$$\begin{aligned} \mathbb{E}[\|X_n\|_{\mathcal{L}}] &\lesssim J(\delta)\mathbf{M}_{\mathcal{G}}\|M_{\mathcal{R}}(y_i)\|_2 + \frac{\mathbf{M}_{\mathcal{G}}\|\max_{1 \leq i \leq n} M_{\mathcal{R}}(y_i)\|_2 J^2(\delta)}{\delta^2\sqrt{n}} \\ &\lesssim c_{v,\alpha}J(\delta)\mathbf{M}_{\mathcal{G}} + c_{v,\alpha}(\log n)^{\alpha/2}\frac{\mathbf{M}_{\mathcal{G}}J^2(\delta)}{\delta^2\sqrt{n}}. \end{aligned}$$

Moreover,  $\|\max_{1 \leq i \leq n} \sup_{g \in \mathcal{G}, r \in \mathcal{R}} |g(\mathbf{x}_i)r(y_i)|\|_{\psi_{\alpha-1}} \lesssim v\mathbf{M}_{\mathcal{G}}(\|\max_{1 \leq i \leq n} y_i\|_{\psi_1})^{\alpha} \lesssim v\mathbf{M}_{\mathcal{G}}(\log n)^{\alpha}$ . Hence, by Theorem 4 in [Adamczak \(2008\)](#), for any  $t > 0$ , with probability at least  $1 - 4 \exp(-t)$ ,

$$\|X_n\|_{\mathcal{L}} \lesssim c_{v,\alpha}J(\delta)\mathbf{M}_{\mathcal{G}} + c_{v,\alpha}\frac{\mathbf{M}_{\mathcal{G}}J^2(\delta)}{\delta^2\sqrt{n}} + c_{v,\alpha}\frac{\mathbf{M}_{\mathcal{G}}}{\sqrt{n}}t + c_{v,\alpha}(\log n)^{\alpha}\frac{\mathbf{M}_{\mathcal{G}}}{\sqrt{n}}t^{\alpha}.$$

In particular,  $\|X_n\|_{\mathcal{L}} = \|G_n - G_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}}\|_{\mathcal{G} \times \mathcal{R}}$ . The bound for  $\|Z_n^G - Z_n^G \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}}\|$  follows from a standard concentration inequality for Gaussian suprema.  $\square$

### SA-III.1.5 Strong Approximation Errors

To simplify notation, the parameters of  $\mathcal{G}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{X}$ , and the index  $\mathcal{X}$  is omitted where there is no ambiguity; the parameters of  $\mathcal{R}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{Y}$ , and the index  $\mathcal{Y}$  is omitted where there is no ambiguity; and the parameters of  $\mathcal{G} \times \mathcal{R}$  (Definitions 4 to 12, [SA.3](#), [SA.4](#)) are taken with  $\mathcal{C} = \mathcal{X} \times \mathcal{Y}$ , and the index  $\mathcal{X} \times \mathcal{Y}$  is omitted where there is no ambiguity. Recall we also define, for  $\delta \in (0, 1]$ ,

$$N(\delta) = N_{\mathcal{G}}(\delta/\sqrt{2}, \mathbf{M}_{\mathcal{G}})N_{\mathcal{R}}(\delta/\sqrt{2}, M_{\mathcal{R}})$$

and

$$J(\delta) = \sqrt{2}J(\mathcal{G}, \mathbf{M}_{\mathcal{G}}, \delta/\sqrt{2}) + \sqrt{2}J(\mathcal{R}, M_{\mathcal{R}}, \delta/\sqrt{2}).$$

**Lemma SA.17.** *Suppose Assumption SA.2 holds, a cylindered dyadic expansion  $\mathcal{C}_{M,N}(\mathbb{P}_Z, 1)$  is given,  $(\Pi_1 Z_n^G(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  is the Gaussian process constructed as in Equation (SA-14) on a possibly enlarged probability space, and  $(\mathcal{G} \times \mathcal{R})_{\delta}$  is chosen in Section SA-III.1.4. Then for all  $t > 0$ ,*

$$\mathbb{P} \left[ \|\Pi_1 G_n - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_{\delta}} > C_1 c_{v,\alpha} \sqrt{\frac{N^{2\alpha+1} 2^M \mathbf{E}_{\mathcal{G}} \mathbf{M}_{\mathcal{G}}}{n}} t + C_1 c_{v,\alpha} \sqrt{\frac{\mathbf{C}_{\Pi_1(\mathcal{G} \times \mathcal{R})_{\delta}, M+N}}{n}} t \right] \leq 2N(\delta) e^{-t},$$

where  $C_1 > 0$  is a universal constant.

**Proof of Lemma SA.17.** To simplify notation, we will use  $\mathbb{E}[\cdot | \mathcal{X}_{0,l}]$  in short for  $\mathbb{E}[\cdot | \mathbf{x}_i \in \mathcal{X}_{0,l}]$ , and  $\mathbb{E}[\cdot | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}]$  in short for  $\mathbb{E}[\cdot | (\mathbf{x}_i, y_i) \in \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}]$ .

**Layers  $N+1 \leq j \leq M+N$ :** For these layers,  $\mathcal{C}_{j,k} = \mathcal{X}_{j-N,k} \times \mathcal{Y}_{*,N,0}$ . By definition of  $\tilde{\gamma}_{j,k}$ ,

$$\begin{aligned} \sum_{N < j \leq M+N} \sum_{0 \leq k < 2^{M+N-j}} |\tilde{\gamma}_{j,k}(g, r)| &\leq \sum_{N < j < M+N} \sum_{0 \leq k < 2^{M+N-j}} \mathbb{E}[|g(\mathbf{x}_i) r(y_i)| | \mathbf{x}_i \in \mathcal{X}_{j-N,k}] \\ &\leq \sum_{N < j < M+N} \sum_{0 \leq k < 2^{M+N-j}} \mathbb{E}[|g(\mathbf{x}_i) \mathbb{E}[r(y_i) | \mathbf{x}_i]| | \mathbf{x}_i \in \mathcal{X}_{j-N,k}] \\ &\leq c_{v,\alpha} \sum_{N < j < M+N} \sum_{0 \leq k < 2^{M+N-j}} \mathbb{E}[|g(\mathbf{x}_i) \mathbb{1}(\mathbf{x}_i \in \mathcal{X}_{j-N,k})|] \mathbb{P}(\mathbf{x}_i \in \mathcal{X}_{j-N,k})^{-1} \\ &\leq c_{v,\alpha} \sum_{N < j < M+N} \mathbf{E}_{\mathcal{G}} 2^{M+N-j} \\ &\leq c_{v,\alpha} 2^M \mathbf{E}_{\mathcal{G}}, \end{aligned}$$

where in the third line we have used  $\mathbb{E}[|r(y_i)| | \mathbf{x}_i = \mathbf{x}] \leq c_{v,\alpha} = v(1 + (2\alpha)^{\alpha/2})$  for all  $\mathbf{x} \in \mathcal{X}$ . Moreover,  $|\tilde{\gamma}_{j,k}(g, r)| \leq 2c_{v,\alpha} \mathbf{M}_{\mathcal{G}}$  for all  $j \in (N, M+N]$ , hence

$$\sum_{N < j \leq M+N} \sum_{0 \leq k < 2^{M+N-j}} |\tilde{\gamma}_{j,k}(g, r)|^2 \leq 2c_{v,\alpha}^2 2^M \mathbf{E}_{\mathcal{G}} \mathbf{M}_{\mathcal{G}}.$$

**Layers  $1 \leq j \leq N$ :** By definition,  $\mathcal{C}_{j,k} = \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}$ , where  $k = 2^{N-j}l + m$ , for some unique  $l \in [0, 2^M)$  and  $m \in [0, 2^{N-j})$ . Denote  $k = (l, m)$ . Fix  $j$  and  $l$ , sum across  $m$ ,

$$\sum_{m=0}^{2^{N-j}-1} |\tilde{\gamma}_{j,(l,m)}(g, r)| = \sum_{m=0}^{2^{N-j}-1} |\mathbb{E}[g(\mathbf{x}_i) | \mathcal{X}_{0,l}] (\mathbb{E}[r(y_i) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j-1,2m}] - \mathbb{E}[r(y_i) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j-1,2m+1}])|.$$

Case 1:  $\alpha > 0$  in (iv) from Assumption SA.2. Then  $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(|y_i|) | \mathbf{x}_i = \mathbf{x}] \leq 2$ , and Markov's inequality implies  $\min\{|y| : y \in \mathcal{Y}_{l,0,0}\} \leq \log(\mathbb{E}[\exp(|y_i|) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,0,0}]) \leq \log(2 \cdot 2^N) \leq 2N$ , and similarly  $\min\{|y| : y \in \mathcal{Y}_{l,0,2^{N-1}}\} \leq 2N$ . Hence the middle cells satisfy  $\mathcal{Y}_{l,j,m} \subseteq [-2N, 2N]$  for all  $0 \leq j < N$ ,  $1 \leq m \leq 2^{N-j} - 2$ , and

$$\sum_{m=1}^{2^{N-j}-2} |\mathbb{E}[r(y_i) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j-1,2m}] - \mathbb{E}[r(y_i) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j-1,2m+1}]| \leq \mathbf{pTV}_{r|_{[-2N, 2N]}} \leq c_{v,\alpha} N^{\alpha},$$

and for the left-most cells,

$$\begin{aligned} & |\mathbb{E}[r(y_i)|\mathcal{X}_{0,l} \times \mathcal{Y}_{l,j-1,1}] - \mathbb{E}[r(y_i)|\mathcal{X}_{0,l} \times \mathcal{Y}_{l,j-1,0}]| \\ & \leq \max_{0 \leq m < 2^{N-j+1}} \mathbb{E}[r(y_i)|\mathcal{X}_{0,l} \times \mathcal{Y}_{l,j-1,m}] - \min_{0 \leq m < 2^{N-j+1}} \mathbb{E}[r(y_i)|\mathcal{X}_{0,l} \times \mathcal{Y}_{l,j-1,m}] \leq 2c_{v,\alpha} N^\alpha, \end{aligned}$$

and similarly for the right-most cells,

$$|\mathbb{E}[r(y_i)|\mathcal{X}_{0,l} \times \mathcal{Y}_{l,j-1,2^{N-j-1}}] - \mathbb{E}[r(y_i)|\mathcal{X}_{0,l} \times \mathcal{Y}_{l,j-1,2^{N-j-2}}]| \leq 2c_{v,\alpha} N^\alpha.$$

*Case 2:*  $\alpha = 0$  in (iv) from Assumption SA.2, since  $\text{pTV}_{\{r\}} \leq 2v$  and  $\mathbb{M}_{\{r\}} \leq 2v$  for all  $r \in \mathcal{R}$ , the above three inequality still hold. It follows that for all  $g \in \mathcal{G}$ ,  $r \in \mathcal{R}$ , fix  $j, l$  and sum across  $m$ ,

$$\sum_{m=0}^{2^{N-j}-1} |\tilde{\gamma}_{j,(l,m)}(g, r)| \leq 2c_{v,\alpha} N^\alpha |\mathbb{E}[g(\mathbf{x}_i)|\mathcal{X}_{0,l}]|.$$

Fix  $j$  and sum the above across  $l$ ,

$$\begin{aligned} \sum_{0 \leq k < 2^{M+N-j}} |\tilde{\gamma}_{j,(l,m)}(g, r)| &= \sum_{l=0}^{2^M-1} \sum_{m=0}^{2^{N-j}-1} |\tilde{\gamma}_{j,(l,m)}(g, r)| \\ &\leq 2c_{v,\alpha} N^\alpha \sum_{l=0}^{2^M-1} \mathbb{E}[|g(\mathbf{x}_i) \mathbb{1}(\mathbf{x}_i \in \mathcal{X}_{0,l})|] \mathbb{P}(\mathbf{x}_i \in \mathcal{X}_{0,l})^{-1} \\ &\leq 2c_{v,\alpha} N^\alpha 2^M \mathbf{E}_{\mathcal{G}}. \end{aligned}$$

We can now sum across  $j$  to get

$$\sum_{j=1}^N \sum_{0 \leq k < 2^{M+N-j}} |\tilde{\gamma}_{j,k}(g, r)| \leq 2c_{v,\alpha} N^{\alpha+1} 2^M \mathbf{E}_{\mathcal{G}}.$$

By Equation (SA-11),  $\sup_{g \in \mathcal{G}, r \in \mathcal{R}} \max_{(j,k) \in \mathcal{I}_{M+N}} |\tilde{\gamma}_{j,k}(g, r)| \leq 2c_{v,\alpha} N^\alpha \mathbf{M}_{\mathcal{G}}$ , and hence

$$\sum_{1 \leq j \leq N} \sum_{0 \leq k < 2^{M+N-j}} |\tilde{\gamma}_{j,k}(g, r)|^2 \leq 4c_{v,\alpha}^2 N^{2\alpha+1} 2^M \mathbf{E}_{\mathcal{G}} \mathbf{M}_{\mathcal{G}}, \quad g \in \mathcal{G}, r \in \mathcal{R}.$$

**Putting Together:** Putting together the previous two parts,

$$\sum_{j=1}^{M+N} \sum_{k=0}^{2^{M+N-j}} \tilde{\gamma}_{j,k}^2(g, r) \leq 6c_{v,\alpha}^2 N^{2\alpha+1} 2^M \mathbf{E}_{\mathcal{G}} \mathbf{M}_{\mathcal{G}}, \quad g \in \mathcal{G}, r \in \mathcal{R}.$$

By Lemma SA.13, we know for any  $(g, r) \in \mathcal{G} \times \mathcal{R}$ , for any  $x > 0$ , with probability at least  $1 - 2 \exp(-x)$ ,

$$|G_n \circ \Pi_1(g, r) - \Pi_1 Z_n^G(g, r)| \lesssim c_{v,\alpha} \sqrt{\frac{N^{2\alpha+1} 2^M \mathbf{E}_{\mathcal{G}} \mathbf{M}_{\mathcal{G}}}{n}} x + c_{v,\alpha} \sqrt{\frac{\mathbf{C}_{\Pi_1\{(g,r)\}, M+N}}{n}} x,$$

and the proof is complete.  $\square$

**Lemma SA.18.** *Suppose Assumption SA.2 holds, a cylindered dyadic expansion  $\mathcal{C}_{M,N}(\mathbb{P}_Z, \rho)$  is given with*

$\rho > 1$ ,  $(\Pi_1 Z_n^G(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  is the Gaussian process constructed as in Equation (SA-14) on a possibly enlarged probability space, and  $(\mathcal{G} \times \mathcal{R})_\delta$  is chosen in Section SA-III.1.4. Then for all  $t > 0$ ,

$$\begin{aligned} \mathbb{P} \left[ \|\Pi_1 G_n - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta} > C_1 C_\rho c_{\mathbf{v}, \alpha} \sqrt{\frac{N^{2\alpha+1} 2^M \mathbf{E}_{\mathcal{G}} \mathbf{M}_{\mathcal{G}}}{n} t} + C_1 C_\rho c_{\mathbf{v}, \alpha} \sqrt{\frac{\mathcal{C}_{\Pi_1(\mathcal{G} \times \mathcal{R})_\delta, M+N}}{n} t} \right] \\ \leq 2\mathbf{N}(\delta) e^{-t} + 2^M \exp(-C_\rho n 2^{-M}), \end{aligned}$$

where  $C_1 > 0$  is a universal constant,  $c_{\mathbf{v}, \alpha} = \mathbf{v}(1 + (2\alpha)^{\alpha/2})$  and  $C_\rho$  is a constant that only depends on  $\rho$ .

**Proof of Lemma SA.18.** Since  $\mathcal{C}_{M,N}$  is a cylindered quasi-dyadic expansion,  $\rho^{-1} 2^{-M-N+j} \leq \mathbb{P}_Z(\mathcal{C}_{j,k}) \leq \rho 2^{-M-N+j}$ , for all  $0 \leq j \leq M+N$ ,  $0 \leq k < 2^{M+N-j}$ . The same argument for Lemma SA.17 implies

$$\sum_{j=1}^{M+N} \sum_{k=0}^{2^{M+N-j}} \tilde{\gamma}_{j,k}^2(g, r) \leq c_\rho c_{\mathbf{v}, \alpha}^2 N^{2\alpha+1} 2^M \mathbf{E}_{\mathcal{G}} \mathbf{M}_{\mathcal{G}}, \quad g \in \mathcal{G}, r \in \mathcal{R},$$

where  $c_\rho$  is a constant that only depends on  $\rho$ . The result then follows from Lemma SA.14.  $\square$

### SA-III.1.6 Projection Error

To simplify notation, the parameters of  $\mathcal{G}$  (Definitions 4 to 12, SA.1, SA.2) are taken with  $\mathcal{C} = \mathcal{X}$ , and the index  $\mathcal{X}$  is omitted where there is no ambiguity; the parameters of  $\mathcal{R}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{Y}$ , and the index  $\mathcal{Y}$  is omitted where there is no ambiguity; and the parameters of  $\mathcal{G} \times \mathcal{R}$  (Definitions 4 to 12, SA.3, SA.4) are taken with  $\mathcal{C} = \mathcal{X} \times \mathcal{Y}$ , and the index  $\mathcal{X} \times \mathcal{Y}$  is omitted where there is no ambiguity. Recall we also define, for  $\delta \in (0, 1]$ ,

$$\mathbf{N}(\delta) = \mathbf{N}_{\mathcal{G}}(\delta/\sqrt{2}, \mathbf{M}_{\mathcal{G}}) \mathbf{N}_{\mathcal{R}}(\delta/\sqrt{2}, M_{\mathcal{R}})$$

and

$$J(\delta) = \sqrt{2} J(\mathcal{G}, \mathbf{M}_{\mathcal{G}}, \delta/\sqrt{2}) + \sqrt{2} J(\mathcal{R}, M_{\mathcal{R}}, \delta/\sqrt{2}).$$

To analyze the projection error, we employ the decomposition

$$\Pi_1 G_n(g, r) - G_n(g, r) = (\Pi_0 G_n(g, r) - G_n(g, r)) + (\Pi_1 G_n(g, r) - \Pi_0 G_n(g, r)),$$

where  $\Pi_0 G_n(g, r) - G_n(g, r)$  represents the  $L_2$  projection error, and  $\Pi_1 G_n(g, r) - \Pi_0 G_n(g, r)$  denotes the mis-specification error. Specifically, the  $L_2$  projection error captures the minimum loss incurred by projecting onto the class of piecewise constant functions over the cells  $\mathcal{E}_{M+N}$ . In contrast, the mis-specification error reflects the additional loss introduced when shifting from the  $L_2$  projection to the product-factorized projection.

First we bound the mis-specification error.

**Lemma SA.19.** Suppose Assumption SA.2 holds, a cylindered dyadic expansion  $\mathcal{C}_{M,N}(\mathbb{P}_Z, 1)$  is given,  $(\Pi_0 Z_n^G(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  and  $(\Pi_1 Z_n^G(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  are the Gaussian processes constructed as in Equation (SA-14) on a possibly enlarged probability space, and  $(\mathcal{G} \times \mathcal{R})_\delta$  is chosen in Section SA-III.1.4. Suppose  $\mathbb{P}_X$  admits a Lebesgue density  $f_X$  supported on  $\mathcal{X} \subseteq \mathbb{R}^d$ . Let  $\tau > 0$ . Define  $r_\tau = r \mathbb{1}([- \tau^{\frac{1}{\alpha}}, \tau^{\frac{1}{\alpha}}])$ .

Then, for any  $g \in \mathcal{G}, r \in \mathcal{R}$ ,

$$\mathbb{E} \left[ (\Pi_1 G_n(g, r_\tau) - \Pi_0 G_n(g, r_\tau))^2 \right] = \mathbb{E} \left[ (\Pi_1 Z_n^G(g, r_\tau) - \Pi_0 Z_n^G(g, r_\tau))^2 \right] \leq 4\mathbf{v}^2(1 + \tau)^2 N^2 \mathbf{V}_\mathcal{G},$$

where

$$\mathbf{V}_\mathcal{G} = \min\{2\mathbf{M}_\mathcal{G}, \mathbf{L}_\mathcal{G} \|\mathcal{V}_M\|_\infty\} \left( \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \right)^2 2^M \mathbf{m}(\mathcal{V}_M) \|\mathcal{V}_M\|_\infty \mathbf{TV}_\mathcal{G}^*,$$

and, as in Section SA-II.1.6,  $\mathcal{V}_M = \cup_{0 \leq l < 2^M} (\mathcal{X}_{0,l} - \mathcal{X}_{0,l})$  is the upper level quasi-dyadic variation set.

**Proof of Lemma SA.19.** To simplify notation, we will use  $\mathbb{E}[\cdot | \mathcal{X}_{0,l}]$  in short for  $\mathbb{E}[\cdot | \mathbf{x}_i \in \mathcal{X}_{0,l}]$ , and  $\mathbb{E}[\cdot | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}]$  in short for  $\mathbb{E}[\cdot | (\mathbf{x}_i, y_i) \in \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}]$  in this proof.

Expanding  $\Pi_1 G_n(g, r_\tau) - \Pi_0 G_n(g, r_\tau)$  by Haar basis representation,

$$\begin{aligned} \Pi_1 G_n(g, r_\tau) - \Pi_0 G_n(g, r_\tau) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \Delta_i(g, r_\tau), \\ \Delta_i(g, r_\tau) &= \sum_{1 \leq j \leq N} \sum_{0 \leq k < 2^{M+N-j}} \left( \tilde{\gamma}_{j,k}(g, r_\tau) - \tilde{\beta}_{j,k}(g, r_\tau) \right) \tilde{e}_{j,k}(\mathbf{x}_i, y_i), \end{aligned}$$

where we have used  $\tilde{\gamma}_{j,k}(g, r_\tau) = \tilde{\beta}_{j,k}(g, r_\tau)$  for  $j > N$ . Moreover,

$$\mathbb{E}[|\Delta_i(g, r_\tau)|] \leq 2 \sum_{0 \leq j < N} \sum_{0 \leq k < 2^{M+N-j}} |\gamma_{j,k}(g, r) - \beta_{j,k}(g, r)| \mathbb{P}((\mathbf{x}_i, y_i) \in \mathcal{C}_{j,k}).$$

Recall in Definition SA.7,  $\mathcal{C}_{j,k} = \mathcal{X}_{j-N,l} \times \mathcal{Y}_{l,j,m}$ , where  $k = 2^{N-j}l + m$ ,  $0 \leq l < 2^M$  and  $0 \leq m < 2^{N-j}$ . Definitions of  $\gamma_{j,k}$  and  $\beta_{j,k}$  from Section SA-III.1.2 imply

$$\begin{aligned} |\gamma_{j,k}(g, r_\tau) - \beta_{j,k}(g, r_\tau)| &= |\mathbb{E}[g(\mathbf{x}_i) | \mathcal{X}_{0,l}] \cdot \mathbb{E}[r_\tau(y_i) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}] - \mathbb{E}[g(\mathbf{x}_i) r_\tau(y_i) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}]| \\ &= |\mathbb{E}[(g(\mathbf{x}_i) - \mathbb{E}[g(\mathbf{x}_i) | \mathcal{X}_{0,l}]) r_\tau(y_i) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}]| \\ &\leq \mathbf{v}(1 + \tau) \mathbb{E}[|g(\mathbf{x}_i) - \mathbb{E}[g(\mathbf{x}_i) | \mathcal{X}_{0,l}]| | \mathcal{C}_{j,k}], \end{aligned}$$

where the first line is simply the definitions of  $\gamma_{j,k}$  and  $\beta_{j,k}$ ; the second line is because  $\sigma(\mathbb{1}(\mathbf{x}_i \in \mathcal{X}_{0,l})) \subseteq \sigma(\mathbb{1}((\mathbf{x}_i, y_i) \in \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}))$ ; and the third line is because Assumption SA.2 (iv) implies  $\sup_{y \in \mathbb{R}} |r_\tau(y)| \leq \mathbf{v}(1 + \tau)$  for all  $r \in \mathcal{R}$ . Summing across  $j$  and  $k$ , then by similar argument as in the proof of Lemma SA.9,

$$\begin{aligned} \mathbb{E}[|\Delta_i(g, r_\tau)|] &\leq 2\mathbf{v}(1 + \tau)N \mathbb{E}[|g(\mathbf{x}_i) - \Pi_0(\mathbf{p}_X[\mathcal{C}_{M,N}(\mathbb{P}, \rho)])g(\mathbf{x}_i)|] \\ &\leq 2\mathbf{v}(1 + \tau)N \left( \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \right)^2 2^M \mathbf{m}(\mathcal{V}_M) \|\mathcal{V}_M\|_\infty \mathbf{TV}_{\{g\}}^*. \end{aligned}$$

For each fixed  $j$ ,  $\tilde{e}_{j,k}(\mathbf{x}, y)$  can be non-zero for only one  $k$ . Hence, almost surely,

$$\begin{aligned}
|\Delta_i(g, r_\tau)| &= \left| \sum_{j=1}^N \sum_{0 \leq k < 2^{M+N-j}} (\tilde{\gamma}_{j,k}(g, r_\tau) - \tilde{\beta}_{j,k}(g, r_\tau)) \tilde{e}_{j,k}(\mathbf{x}_i, y_i) \right| \\
&\leq \sum_{j=1}^N \max_{0 \leq k < 2^{M+N-j}} \left| \tilde{\gamma}_{j,k}(g, r_\tau) - \tilde{\beta}_{j,k}(g, r_\tau) \right| \\
&\leq 2 \sum_{j=0}^{N-1} \max_{0 \leq k < 2^{M+N-j}} |\gamma_{j,k}(g, r_\tau) - \beta_{j,k}(g, r_\tau)| \\
&\leq 2\mathbf{v}(1 + \tau) \sum_{j=0}^{N-1} \max_{0 \leq k < 2^{M+N-j}} |\mathbb{E}[g(\mathbf{x}_i) - \mathbb{E}[g(\mathbf{x}_i)|\mathcal{X}_{0,l}] | \mathcal{C}_{j,k}]| \\
&\leq 2N\mathbf{v}(1 + \tau) \min\{2\mathbf{M}_\mathcal{G}, \mathbf{L}_\mathcal{G} \|\mathcal{V}_M\|_\infty\}.
\end{aligned}$$

This shows the results.  $\square$

Next we bound the  $L_2$  projection error.

**Lemma SA.20.** *Suppose Assumption SA.2 holds, a cylindered dyadic expansion  $\mathcal{C}_{M,N}(\mathbb{P}_Z, 1)$  is given,  $(Z_n^G(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  and  $(\Pi_0 Z_n^G(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  are the Gaussian processes constructed as in Equations (SA-13) and (SA-14) on a possibly enlarged probability space, and  $(\mathcal{G} \times \mathcal{R})_\delta$  is chosen in Section SA-III.1.4. Suppose  $\mathbb{P}_X$  admits a Lebesgue density  $f_X$  supported on  $\mathcal{X} \subseteq \mathbb{R}^d$ . Let  $\tau > 0$ . Define  $r_\tau = r \mathbb{1}([-\tau^{\frac{1}{\alpha}}, \tau^{\frac{1}{\alpha}}])$ . Then for any  $g \in \mathcal{G}, r \in \mathcal{R}$ ,*

$$\mathbb{E} \left[ (\Pi_0 Z_n^G(g, r_\tau) - Z_n^G(g, r_\tau))^2 \right] = \mathbb{E} \left[ (\Pi_0 G_n(g, r_\tau) - G_n(g, r_\tau))^2 \right] \leq 4\mathbf{v}^2(1 + \tau)^2 (2^{-N} \mathbf{M}_\mathcal{G}^2 + \mathbf{V}_\mathcal{G}),$$

where  $\mathbf{V}_\mathcal{G}$  is defined in Lemma SA.19.

**Proof of Lemma SA.20.** To simplify notation, we will use  $\mathbb{E}[\cdot | \mathcal{X}_{0,l}]$  in short for  $\mathbb{E}[\cdot | \mathbf{x}_i \in \mathcal{X}_{0,l}]$ , and  $\mathbb{E}[\cdot | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}]$  in short for  $\mathbb{E}[\cdot | (\mathbf{x}_i, y_i) \in \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}]$  in this proof.

Let  $\mathcal{B} = \sigma(\{\mathbb{1}((\mathbf{x}_i, y_i) \in \mathcal{C}_{0,k}) : 0 \leq k < 2^{M+N}\})$  be the  $\sigma$ -algebra generated by  $\{\mathbb{1}((\mathbf{x}_i, y_i) \in \mathcal{C}_{0,k}) : 0 \leq k < 2^{M+N}\}$ . Then the difference between the  $L_2$  projection and the original can be expressed as

$$\begin{aligned}
\Pi_0(g \cdot r_\tau)(\mathbf{x}_i, y_i) - g(\mathbf{x}_i)r_\tau(y_i) &= \mathbb{E}[g(\mathbf{x}_i)r_\tau(y_i) | \mathcal{B}] - g(\mathbf{x}_i)r_\tau(y_i) \\
&= \mathbb{E}[g(\mathbf{x}_i)r_\tau(y_i) | \mathcal{B}] - \mathbb{E}[g(\mathbf{x}_i) | \mathcal{B}]r_\tau(y_i) + \mathbb{E}[g(\mathbf{x}_i) | \mathcal{B}]r_\tau(y_i) - g(\mathbf{x}_i)r_\tau(y_i).
\end{aligned}$$

By Definition SA.7, each cell  $\mathcal{C}_{0,k}$  is of the form of a product, that is,

$$\mathcal{C}_{0,k} = \mathcal{X}_{0,l} \times \mathcal{Y}_{l,0,m} \text{ with } k = 2^N l + m,$$

where  $0 \leq k < 2^{M+N}$ ,  $0 \leq l < 2^M$  and  $0 \leq m < 2^N$ .

The first two terms  $\mathbb{E}[g(\mathbf{x}_i)r_\tau(y_i) | \mathcal{B}] - \mathbb{E}[g(\mathbf{x}_i) | \mathcal{B}]r_\tau(y_i)$  in the decomposition are driven by projection of  $r_\tau$  on grids  $\mathcal{Y}_{l,0,m}$ 's, and can be upper bounded through probability measure assigned to each grid ( $2^{-N}$ ) and total variation of  $r_\tau$ . We consider the positive and negative parts separately: Consider the function

$$q_{l,m}^+(y) = \mathbb{E}[g(\mathbf{x}_i) \mathbb{1}(g(\mathbf{x}_i) \geq 0) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,0,m}] r_\tau(y) - \mathbb{E}[g(\mathbf{x}_i) r_\tau(y_i) \mathbb{1}(g(\mathbf{x}_i) \geq 0) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,0,m}], \quad y \in \mathcal{Y}_{l,0,m}.$$

Either  $q_{l,m}^+$  is constantly zero on  $\mathcal{Y}_{l,0,m}$  or  $q_{l,m}^+$  takes both positive and negative values on  $\mathcal{Y}_{l,0,m}$ . Under either case, we have  $|q_{l,m}^+(y)| \leq \text{pTV}_{\{q_{l,m}^+\}, \mathcal{Y}_{l,0,m}}$  for all  $y \in \mathcal{Y}_{l,0,m}$ . Hence

$$\begin{aligned} \mathbb{E}[|q_{l,m}^+(y_i)| \mathbb{1}(y_i \in \mathcal{Y}_{l,0,m}) | \mathbf{x}_i = \mathbf{x}] &= \int_{\mathcal{Y}_{l,0,m}} |q_{l,m}^+(y)| d\mathbb{P}(y_i \leq y | \mathbf{x}_i = \mathbf{x}) \\ &\leq \mathbb{P}(y_i \in \mathcal{Y}_{l,0,m} | \mathbf{x}_i = \mathbf{x}) \text{pTV}_{\{q_{l,m}^+\}, \mathcal{Y}_{l,0,m}} \\ &\leq \mathbb{P}(y_i \in \mathcal{Y}_{l,0,m} | \mathbf{x}_i = \mathbf{x}) \mathbb{M}_{\{g\}} \text{pTV}_{\{r_\tau\}, \mathcal{Y}_{l,0,m}}, \quad \mathbf{x} \in \mathcal{X}_{0,l}. \end{aligned}$$

Similarly,

$$q_{l,m}^-(y) = \mathbb{E}[g(\mathbf{x}_i) \mathbb{1}(g(\mathbf{x}_i) < 0) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,0,m}] r_\tau(y) - \mathbb{E}[g(\mathbf{x}_i) r_\tau(y_i) \mathbb{1}(g(\mathbf{x}_i) < 0) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,0,m}], \quad y \in \mathcal{Y}_{l,0,m},$$

and we have

$$\mathbb{E}[|q_{l,m}^-(y_i)| \mathbb{1}(y_i \in \mathcal{Y}_{l,0,m}) | \mathbf{x}_i = \mathbf{x}] \leq \mathbb{P}(y_i \in \mathcal{Y}_{l,0,m} | \mathbf{x}_i = \mathbf{x}) \mathbb{M}_{\{g\}} \text{pTV}_{\{r_\tau\}, \mathcal{Y}_{l,0,m}}, \quad \mathbf{x} \in \mathcal{X}_{0,l}.$$

Combining the two parts, and integrate over the event  $\mathbf{x}_i \in \mathcal{X}_{0,l}$ ,

$$\begin{aligned} &\mathbb{E}\left[ \left| \mathbb{E}[g(\mathbf{x}_i) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,0,m}] r_\tau(y_i) - \mathbb{E}[g(\mathbf{x}_i) r_\tau(y_i) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,0,m}] \mathbb{1}(y_i \in \mathcal{Y}_{l,0,m}) \right| \mathbf{x}_i \in \mathcal{X}_{0,l} \right] \\ &\leq 2\mathbb{P}(y_i \in \mathcal{Y}_{l,0,m} | \mathbf{x}_i \in \mathcal{X}_{0,l}) \mathbb{M}_{\{g\}} \text{pTV}_{\{r_\tau\}, \mathcal{Y}_{l,0,m}} \leq 2 \cdot 2^{-N} \mathbb{M}_{\{g\}} \text{pTV}_{\{r_\tau\}, \mathcal{Y}_{l,0,m}}. \end{aligned}$$

Summing over  $m$ , we get for each  $0 \leq l < 2^M$ ,

$$\mathbb{E}\left[ \left| \mathbb{E}[g(\mathbf{x}_i) | \mathcal{B}] r_\tau(y_i) - \mathbb{E}[g(\mathbf{x}_i) r_\tau(y_i) | \mathcal{B}] \right| \mathbf{x}_i \in \mathcal{X}_{0,l} \right] \leq 2 \cdot 2^{-N} \mathbb{M}_{\{g\}} \text{pTV}_{\{r_\tau\}, \mathcal{Y}_{*,N,0}}.$$

Hence, using the polynomial growth of total variation,

$$\mathbb{E}\left[ \left| \mathbb{E}[g(\mathbf{x}_i) | \mathcal{B}] r_\tau(y_i) - \mathbb{E}[g(\mathbf{x}_i) r_\tau(y_i) | \mathcal{B}] \right| \right] \leq 2^{-N} \mathbb{M}_{\{g\}} \text{pTV}_{\{r_\tau\}, \mathcal{Y}_{*,N,0}} \leq 2 \cdot 2^{-N} \mathbb{M}_{\mathcal{G}} \mathbf{v}(1 + \tau).$$

Since  $\left| \mathbb{E}[g(\mathbf{x}_i) r_\tau(y_i) | \mathcal{B}] - \mathbb{E}[g(\mathbf{x}_i) | \mathcal{B}] r_\tau(y_i) \right| \leq 2\mathbb{M}_{\mathcal{G}} \mathbf{v}(1 + \tau)$  almost surely,

$$\mathbb{E}\left[ \left( \mathbb{E}[g(\mathbf{x}_i) | \mathcal{B}] r_\tau(y_i) - \mathbb{E}[g(\mathbf{x}_i) r_\tau(y_i) | \mathcal{B}] \right)^2 \right] \leq 4 \cdot 2^{-N} \mathbf{v}^2 (1 + \tau)^2 \mathbb{M}_{\mathcal{G}}^2.$$

Now we look at the last two terms  $\mathbb{E}[g(\mathbf{x}_i) | \mathcal{B}] r_\tau(y_i) - g(\mathbf{x}_i) r_\tau(y_i)$ , which are essentially driven by the  $L_2$ -projection error of  $g$ . Denote by  $\mathcal{A} = \sigma(\{\mathbb{1}(\mathbf{x}_i \in \mathcal{X}_{0,l}) : 0 \leq l < 2^M\})$  the  $\sigma$ -algebra generated by  $\{\mathbb{1}(\mathbf{x}_i \in \mathcal{X}_{0,l}) : 0 \leq l < 2^M\}$ . Then  $\mathcal{A} \subseteq \mathcal{B}$ . By Jensen's inequality and a similar argument as in the proof of Lemma SA.9,

$$\mathbb{E}\left[ \left( \mathbb{E}[g(\mathbf{x}_i) | \mathcal{B}] r_\tau(y_i) - g(\mathbf{x}_i) r_\tau(y_i) \right)^2 \right] \leq 4\mathbf{v}^2 (1 + \tau)^2 \mathbb{E}\left[ \left( g(\mathbf{x}_i) - \mathbb{E}[g(\mathbf{x}_i) | \mathcal{A}] \right)^2 \right] \leq 4\mathbf{v}^2 (1 + \tau)^2 \mathbb{V}_{\mathcal{G}}.$$

It then follows that  $\mathbb{E}\left[ \left( \Pi_0 G_n(g, r_\tau) - G_n(g, r_\tau) \right)^2 \right] \leq 4\mathbf{v}^2 (1 + \tau)^2 (2^{-N} \mathbb{M}_{\mathcal{G}}^2 + \mathbb{V}_{\mathcal{G}})$ .  $\square$

Using a truncation argument and the previous two lemmas, we get the bound on  $\Pi_1$ -projection error with tail control.

**Lemma SA.21.** *Suppose Assumption SA.2 holds, a cylindered dyadic expansion  $\mathcal{C}_{M,N}(\mathbb{P}_Z, 1)$  is given,*



$(Z_n^G(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  and  $(\Pi_1 Z_n^G(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  are the Gaussian processes constructed as in Equations (SA-13) and (SA-14) on a possibly enlarged probability space, and  $(\mathcal{G} \times \mathcal{R})_\delta$  is chosen in Section SA-III.1.4. Suppose  $\mathbb{P}_X$  admits a Lebesgue density  $f_X$  supported on  $\mathcal{X} \subseteq \mathbb{R}^d$ . Then for all  $t > N$ ,

$$\begin{aligned} \mathbb{P} \left[ \|G_n - \Pi_1 G_n\|_{(\mathcal{G} \times \mathcal{R})_\delta} > C_1 \sqrt{c_{v,2\alpha}} \sqrt{N^2 \mathbf{V}_\mathcal{G} + 2^{-N} \mathbf{M}_\mathcal{G}^2 t^{\alpha + \frac{1}{2}}} + C_1 c_{v,\alpha} \frac{\mathbf{M}_\mathcal{G}}{\sqrt{n}} t^{\alpha + 1} \right] &\leq 4N(\delta) n e^{-t}, \\ \mathbb{P} \left[ \|Z_n^G - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta} > C_1 \sqrt{c_{v,2\alpha}} \sqrt{N^2 \mathbf{V}_\mathcal{G} + 2^{-N} \mathbf{M}_\mathcal{G}^2 t^{\frac{1}{2}}} + C_1 c_{v,\alpha} \frac{\mathbf{M}_\mathcal{G}}{\sqrt{n}} t \right] &\leq 4N(\delta) n e^{-t}, \end{aligned}$$

where  $c_{v,\alpha} = v(1 + (2\alpha)^{\frac{\alpha}{2}})$ ,  $c_{v,2\alpha} = v^2(1 + (4\alpha)^\alpha)$ , and  $C_1$  is a universal constant.

**Proof of Lemma SA.21.** To simplify notation, we will use  $\mathbb{E}[\cdot | \mathcal{X}_{0,l}]$  in short for  $\mathbb{E}[\cdot | \mathbf{x}_i \in \mathcal{X}_{0,l}]$ , and  $\mathbb{E}[\cdot | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}]$  in short for  $\mathbb{E}[\cdot | (\mathbf{x}_i, y_i) \in \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}]$  in this proof. We will use a truncation argument and consider the cases of whether  $\alpha > 0$  in (iv) of Assumption SA.2 separately.

First, suppose  $\alpha > 0$  in (iv) of Assumption SA.2. Let  $\tau > 0$  such that  $\tau^{\frac{1}{\alpha}} > \log(2^{N+1})$ .

Projection error for truncated processes: By Lemmas SA.19 and SA.20, and using Bernstein inequality, for all  $t > 0$ , for each  $g \in \mathcal{G}, r \in \mathcal{R}$ ,

$$\mathbb{P} \left[ |G_n(g, r_\tau) - \Pi_1 G_n(g, r_\tau)| \geq 4v(1 + \tau) \sqrt{N^2 \mathbf{V}_\mathcal{G} + 2^{-N} \mathbf{M}_\mathcal{G}^2} \sqrt{t} + \frac{4}{3} v(1 + \tau) \frac{\mathbf{M}_\mathcal{G}}{\sqrt{n}} t \right] \leq 2e^{-t}.$$

Truncation Error: Recall Equation (SA-11) implies  $\max_{0 \leq k < 2^{M+N}} \mathbb{E}[|r(y_i)| | (\mathbf{x}_i, y_i) \in \mathcal{C}_{0,k}] \leq c_{v,\alpha} N^\alpha$ . The same argument implies  $\max_{0 \leq k < 2^{M+N}} \mathbb{E}[r(y_i)^2 | (\mathbf{x}_i, y_i) \in \mathcal{C}_{0,k}] \leq v^2(1 + (N \log(2) \sqrt{2\alpha})^{2\alpha}) \leq c_{v,2\alpha} N^{2\alpha}$ . Hence the following holds almost surely,

$$\begin{aligned} |\Pi_1 G_n(g, r) - \Pi_1 G_n(g, r_\tau)| &\leq \max_{0 \leq l < 2^M} \max_{0 \leq m < 2^N} \left| \mathbb{E}[g(\mathbf{x}_i) | \mathcal{X}_{0,l}] \mathbb{E}[|r(y_i)| \mathbb{1}(|y_i| \geq \tau^{1/\alpha}) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,0,m}] \right| \\ &\leq c_{v,\alpha} \mathbf{M}_\mathcal{G} N^\alpha. \end{aligned}$$

Since  $\tau^{\frac{1}{\alpha}} > \log(2^{N+1}) > 0.5N$ ,  $\gamma_{0,k} = \beta_{0,k}$  for all  $k$  corresponding to  $\mathcal{X}_{0,l} \times \mathcal{Y}_{l,0,m}$  for  $0 < m < 2^N - 1$ , that is, the mismatch only happens at edge cells of  $y_i$ , we have

$$\begin{aligned} \mathbb{E} \left[ |\Pi_1 G_n(g, r) - \Pi_1 G_n(g, r_\tau)|^2 \right] &\leq \mathbb{P}(\Pi_1 G_n(g, r) - \Pi_1 G_n(g, r_\tau) \neq 0) c_{v,2\alpha} \mathbf{M}_\mathcal{G}^2 N^{2\alpha} \\ &\leq c_{v,2\alpha} 2^{-N+1} \mathbf{M}_\mathcal{G}^2 N^{2\alpha}. \end{aligned}$$

Using Bernstein's inequality, for all  $t > 0$ , with probability at least  $1 - 2 \exp(-t)$ ,

$$\begin{aligned} |\Pi_1 G_n(g, r) - \Pi_1 G_n(g, r_\tau)| &\lesssim \sqrt{c_{v,2\alpha}} 2^{-N/2} \mathbf{M}_\mathcal{G} N^\alpha \sqrt{t} + c_{v,\alpha} \frac{\mathbf{M}_\mathcal{G} N^\alpha}{\sqrt{n}} t \\ &\lesssim \sqrt{c_{v,2\alpha}} 2^{-N/2} \mathbf{M}_\mathcal{G} N^\alpha \sqrt{t} + c_{v,\alpha} \frac{\mathbf{M}_\mathcal{G} \tau}{\sqrt{n}} t. \end{aligned}$$

Moreover, using  $\mathbb{P}(|y_i| \geq \tau) \leq 2 \cdot 2^{-N}$ , we have

$$\begin{aligned} \mathbb{E}[(G_n(g, r) - G_n(g, r_\tau))^2] &\leq \mathbf{M}_\mathcal{G}^2 \mathbb{E}[(r(y_i) - r_\tau(y_i))^2] \leq \mathbf{M}_\mathcal{G}^2 \mathbb{E}[r(y_i)^2 \mathbb{1}(|y_i| \geq \tau)] \\ &\leq 2 \cdot 2^{-N} \mathbf{M}_\mathcal{G}^2 \max_{0 \leq k < 2^{M+N}} \mathbb{E}[r(y_i)^2 | (\mathbf{x}_i, y_i) \in \mathcal{C}_{0,k}] \leq 2c_{v,2\alpha} \mathbf{M}_\mathcal{G}^2 N^{2\alpha} 2^{-N}. \end{aligned}$$

By Bernstein inequality and a truncation argument, for all  $t > 0$ ,

$$\begin{aligned} & \mathbb{P}(\sqrt{n}|G_n(g, r) - G_n(g, r_\tau)| \geq t) \\ & \leq \min_{y>0} \left\{ 2 \exp\left(-\frac{t^2}{2n\mathbb{V}[G_n(g, r) - G_n(g, r_\tau)] + \frac{2}{3}xy}\right) + 2\mathbb{P}\left(\max_{1 \leq i \leq n} |g(\mathbf{x}_i)(r(y_i) - r_\tau(y_i))| \geq y\right) \right\}. \end{aligned}$$

Taking  $y = M_{\mathcal{G}}t^\alpha$ , we get for all  $t > 0$ , with probability at least  $1 - 4\exp(-t)$ ,

$$|G_n(g, r) - G_n(g, r_\tau)| \lesssim \sqrt{c_{v,2\alpha}} 2^{-N/2} M_{\mathcal{G}} N^\alpha \sqrt{t} + C_{v,\alpha} \frac{M_{\mathcal{G}}}{\sqrt{n}} t^{\alpha+1}.$$

Putting Together: Taking  $\tau = t^\alpha > 0.5^\alpha N^\alpha$ , we get from the previous bounds on  $G_n(g, r_\tau) - \Pi_1 G_n(g, r_\tau)$ ,  $\Pi_1 G_n(g, r) - \Pi_1 G_n(g, r_\tau)$ , and  $G_n(g, r) - G_n(g, r_\tau)$  that for all  $g \in \mathcal{G}$ ,  $r \in \mathcal{R}$ , for all  $t > N$ , with probability at least  $1 - 4n\exp(-t)$ ,

$$|\Pi_1 G_n(g, r) - G_n(g, r)| \lesssim \sqrt{c_{v,2\alpha}} \sqrt{N^2 \mathbb{V}_{\mathcal{G}} + 2^{-N} M_{\mathcal{G}}^2} t^{\alpha+\frac{1}{2}} + c_{v,\alpha} \frac{M_{\mathcal{G}}}{\sqrt{n}} t^{\alpha+1}. \quad (\text{SA-15})$$

The bound for  $|\Pi_1 Z_n^G(g, r) - Z_n^G(g, r)|$  follows from the fact that it is a mean-zero Gaussian random variable with variance equal to  $\mathbb{V}[\Pi_1 G_n(g, r) - G_n(g, r)]$ . The result follows then follows from a union bound over  $(g, r) \in (\mathcal{G} \times \mathcal{R})_\delta$ .

Next, suppose  $\alpha = 0$  in (iv) of Assumption SA.2. This implies  $M_{\mathcal{R}} \leq 2v$ . Hence choosing  $\tau = 2v$ , then  $G_n(g, r) = G_n(g, r_\tau)$  almost surely for all  $g \in \mathcal{G}$ ,  $r \in \mathcal{R}$ , that is, there is no truncation error. Hence the bound on  $G_n(g, r_\tau) - \Pi_1 G_n(g, r_\tau)$  implies Equation (SA-15) holds with  $\alpha = 0$  and similarly for the  $Z_n^G$  counterpart.  $\square$

## SA-III.2 General Result

This section presents the main result for the  $G_n$ -process. To simplify notation, the parameters of  $\mathcal{G}$  and  $\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}$  (Definitions 4 to 12, SA.1, SA.2) are taken with  $\mathcal{C} = \mathcal{Q}_{\mathcal{G}}$ , and the index  $\mathcal{Q}_{\mathcal{G}}$  is omitted where there is no ambiguity; the parameters of  $\mathcal{R}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{Y}$ , and the index  $\mathcal{Y}$  is omitted where there is no ambiguity; and the parameters of  $\mathcal{G} \times \mathcal{R}$  (Definitions 4 to 12, SA.3, SA.4) are taken with  $\mathcal{C} = \mathcal{Q}_{\mathcal{G}} \times \mathcal{Y}$ , and the index  $\mathcal{Q}_{\mathcal{G}} \times \mathcal{Y}$  is omitted where there is no ambiguity.

**Theorem SA.1.** *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  $\mathbb{Q}_{\mathcal{G}} = \mathbf{m} \circ \phi_{\mathcal{G}}$ , where the normalizing transformation  $\phi_{\mathcal{G}} : \mathcal{Q}_{\mathcal{G}} \mapsto [0, 1]^d$  is a diffeomorphism.

(iii)  $M_{\mathcal{G}} < \infty$  and  $J(\mathcal{G}, M_{\mathcal{G}}, 1) < \infty$ .

(iv)  $\mathcal{R}$  is a real-valued pointwise measurable class of functions on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_Y)$ .

(v)  $J(\mathcal{R}, M_{\mathcal{R}}, 1) < \infty$ , where  $M_{\mathcal{R}}(y) + \mathbf{pTV}_{\mathcal{R}, (-|y|, |y|)} \leq \mathbf{v}(1 + |y|^\alpha)$  for all  $y \in \mathcal{Y}$ , for some  $\mathbf{v} > 0$ , and for some  $\alpha \geq 0$ . Furthermore, if  $\alpha > 0$ , then  $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(|y_i|) | \mathbf{x}_i = \mathbf{x}] \leq 2$ .

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

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

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

$$\mathbb{T}_n^G(t) = \min_{\delta \in (0, 1)} \{A_n^G(t, \delta) + F_n^G(t, \delta)\},$$

with

$$\begin{aligned} A_n^G(t, \delta) &= \sqrt{d} \min \left\{ \left( \frac{c_1^d \mathbb{E}_{\mathcal{G}} \text{TV}_{\mathcal{G}}^d M_{\mathcal{G}}^{d+1}}{n} \right)^{\frac{1}{2(d+1)}}, \left( \frac{c_1^d c_2^d \mathbb{E}_{\mathcal{G}}^2 M_{\mathcal{G}}^2 \text{TV}_{\mathcal{G}}^d L_{\mathcal{G}}^d}{n^2} \right)^{\frac{1}{2(d+2)}} \right\} (t + \log(nN(\delta)N^*))^{\alpha+1} \\ &\quad + \sqrt{\frac{\min\{M_{\mathcal{G}}^2(M^* + N^*), M_{\mathcal{G}}(c_3 K_{\mathcal{G}} \cdot \mathcal{V}_{\mathcal{R}} + M_{\mathcal{G}})\}}{n}} (\log n)^\alpha (t + \log(nN(\delta)N^*))^{\alpha+1}, \\ F_n^G(t, \delta) &= J(\delta)M_{\mathcal{G}} + \frac{(\log n)^{\alpha/2} M_{\mathcal{G}} J^2(\delta)}{\delta^2 \sqrt{n}} + \frac{M_{\mathcal{G}}}{\sqrt{n}} \sqrt{t} + (\log n)^\alpha \frac{M_{\mathcal{G}}}{\sqrt{n}} t^\alpha, \end{aligned}$$

where

$$c_1 = d \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{X}}} \prod_{j=1}^{d-1} \sigma_j(\nabla \phi_{\mathcal{H}}(\mathbf{x})), \quad c_2 = \sup_{\mathbf{x} \in \mathcal{Q}_{\mathcal{X}}} \frac{1}{\sigma_d(\nabla \phi_{\mathcal{H}}(\mathbf{x}))}, \quad c_3 = d^{-1/2} (2\sqrt{d})^{d-1} c_1 c_2^{d-1},$$

and

$$\begin{aligned} \mathcal{V}_{\mathcal{R}} &= \{\theta(\cdot, r) : r \in \mathcal{R}\}, \\ N(\delta) &= N_{\mathcal{G}}(\delta/\sqrt{2}, M_{\mathcal{G}}) N_{\mathcal{R}}(\delta/\sqrt{2}, M_{\mathcal{R}}), \quad \delta \in (0, 1], \\ J(\delta) &= \sqrt{2} J(\mathcal{G}, M_{\mathcal{G}}, \delta/\sqrt{2}) + \sqrt{2} J(\mathcal{R}, M_{\mathcal{R}}, \delta/\sqrt{2}), \quad \delta \in (0, 1], \\ M^* &= \left\lceil \log_2 \min \left\{ \left( \frac{c_1 n \text{TV}_{\mathcal{G}}}{\mathbb{E}_{\mathcal{G}}} \right)^{\frac{d}{d+1}}, \left( \frac{c_1 c_2 n L_{\mathcal{G}} \text{TV}_{\mathcal{G}}}{\mathbb{E}_{\mathcal{G}} M_{\mathcal{G}}} \right)^{\frac{d}{d+2}} \right\} \right\rceil, \\ N^* &= \left\lceil \log_2 \max \left\{ \left( \frac{n M_{\mathcal{G}}^{d+1}}{c_1^d \mathbb{E}_{\mathcal{G}} \text{TV}_{\mathcal{G}}^d} \right)^{\frac{1}{d+1}}, \left( \frac{n^2 M_{\mathcal{G}}^{2d+2}}{c_1^d c_2^d \text{TV}_{\mathcal{G}}^d L_{\mathcal{G}}^d \mathbb{E}_{\mathcal{G}}^2} \right)^{\frac{1}{d+2}} \right\} \right\rceil. \end{aligned}$$

**Proof of Theorem SA.1.** To simplify notation, we will use  $\mathbb{E}[\cdot | \mathcal{X}_{0,l}]$  in short for  $\mathbb{E}[\cdot | \mathbf{x}_i \in \mathcal{X}_{0,l}]$ , and  $\mathbb{E}[\cdot | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}]$  in short for  $\mathbb{E}[\cdot | (\mathbf{x}_i, y_i) \in \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}]$  in this proof.

First, we make a reduction via the surrogate measure and normalizing transformation. Since  $\text{Supp}(\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}) \subseteq \text{Supp}(\mathcal{G})$ , we know  $\mathcal{Q}_{\mathcal{G}}$  is also a surrogate measure for  $\mathbb{P}_X$  with respect to  $\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}$ , and  $\phi_{\mathcal{G}}$  remains a valid normalizing transformation. Let  $\mathcal{Z}_{\mathcal{G}} = \mathcal{X} \cap \text{Supp}(\mathcal{G})$ . Since  $\mathbb{Q}_{\mathcal{G}} = \mathbf{m} \circ \phi_{\mathcal{G}}$  by assumption (ii) in Theorem 1, and  $\mathbb{Q}_{\mathcal{G}}|_{\mathcal{Z}_{\mathcal{G}}} = \mathbb{P}_X|_{\mathcal{Z}_{\mathcal{G}}}$ ,

$$\mathbb{P}_X|_{\mathcal{Z}_{\mathcal{G}}} = \mathbf{m} \circ \phi_{\mathcal{G}}|_{\mathcal{Z}_{\mathcal{G}}}.$$

To define the  $\text{Uniform}([0, 1]^d)$  random variables on the probability space that  $(\mathbf{x}_i, y_i)$ 's live in, we define a joint probability measure  $\mathbb{O}$  on  $(\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^{2d+1}))$  such that for all  $A \in \mathcal{B}(\mathbb{R}^{2d+1})$ :

$$\begin{aligned}\mathbb{O}(A \cap (\mathcal{Z}_{\mathcal{H}} \times \mathbb{R} \times \mathcal{Z}_{\mathcal{H}})) &= \mathbb{P}_Z(\Pi_{1:d+1}(A \cap \{(\mathbf{x}, y, \mathbf{x}) : \mathbf{x} \in \mathcal{Z}_{\mathcal{H}}, y \in \mathbb{R}\})), \\ \mathbb{O}(A \cap (\mathcal{Z}_{\mathcal{H}} \times \mathbb{R} \times \mathcal{Z}_{\mathcal{H}}^c)) &= \mathbb{O}(A \cap (\mathcal{Z}_{\mathcal{H}}^c \times \mathbb{R} \times \mathcal{Z}_{\mathcal{H}})) = 0, \\ \mathbb{O}(A \cap (\mathcal{Z}_{\mathcal{H}}^c \times \mathbb{R} \times \mathcal{Z}_{\mathcal{H}}^c)) &= \int_{\mathcal{Z}_{\mathcal{H}}^c \cap \Pi_{d+2:2d+1}(A)} \frac{\mathbb{P}_Z(A^{\mathbf{u}} \cap (\mathcal{Z}_{\mathcal{H}}^c \times \mathbb{R}))}{\mathbb{P}_Z(\mathcal{Z}_{\mathcal{H}}^c \times \mathbb{R})} d(\mathbf{m} \circ \phi_{\mathcal{H}})(\mathbf{u}),\end{aligned}$$

where  $\Pi_{1:d+1}(A) = \{\mathbf{z} \in \mathbb{R}^{d+1} : (\mathbf{z}, \mathbf{u}) \in A \text{ for some } \mathbf{u} \in \mathbb{R}^d\}$ ,  $\Pi_{d+2:2d+1}(A) = \{\mathbf{u} \in \mathbb{R}^d : (\mathbf{z}, \mathbf{u}) \in A \text{ for some } \mathbf{z} \in \mathbb{R}^{d+1}\}$ , and  $A^{\mathbf{u}} = \{\mathbf{z} \in \mathbb{R}^{d+1} : (\mathbf{z}, \mathbf{u}) \in A\}$ .

Then we can check that (i) the marginals of  $\mathbb{O}$  are  $\mathbb{P}_Z$  and  $\mathbf{m} \circ \phi_{\mathcal{H}}$ , respectively; (ii)  $\mathbb{O}|_{\mathcal{Z}_{\mathcal{H}} \times \mathbb{R} \times \mathbb{R}^d \cup \mathbb{R}^d \times \mathbb{R} \times \mathcal{Z}_{\mathcal{H}}}$  is supported on  $\{(\mathbf{x}, y, \mathbf{x}) : \mathbf{x} \in \mathcal{Z}_{\mathcal{H}}, y \in \mathbb{R}\}$ . By Skorohod embedding (Dudley, 2014, Lemma 3.35), on a possibly enlarged probability space, there exists a  $\mathbf{u}_i, 1 \leq i \leq n$  i.i.d.  $\text{Uniform}([0, 1]^d)$  such that  $(\mathbf{z}_i = (\mathbf{x}_i, y_i), \phi_{\mathcal{H}}^{-1}(\mathbf{u}_i))$  has joint law  $\mathbb{O}$ . In particular, if  $\mathbf{x}_i \in \mathcal{Z}_{\mathcal{H}}$ , then  $\mathbf{x}_i = \phi_{\mathcal{H}}^{-1}(\mathbf{u}_i)$ ; if  $\mathbf{x}_i \in \mathcal{Z}_{\mathcal{H}}^c$ , then  $\phi_{\mathcal{H}}^{-1}(\mathbf{u}_i) \in \mathcal{Z}_{\mathcal{H}}^c$ , and since  $\mathcal{Q}_{\mathcal{H}} \subseteq \mathcal{X} \cup (\cap_{h \in \mathcal{H}} \text{Supp}(h)^c)$ ,  $\phi_{\mathcal{H}}^{-1}(\mathbf{u}_i) \in \cap_{h \in \mathcal{H}} \text{Supp}(h)^c$ . In particular,  $\sup_{\mathbf{u} \in [0, 1]^d} \mathbb{E}[\exp(|y_i|) | \mathbf{u}_i = \mathbf{u}] \leq 2$ .

By the same argument as in the proof for Theorem 1, assumption (ii) implies that on a possibly enriched probability space, there exists  $(\mathbf{u}_i : 1 \leq i \leq n)$  i.i.d distributed with law  $\mathbb{P}_U = \text{Uniform}([0, 1]^d)$ , and

$$g(\mathbf{x}_i) = g(\phi_{\mathcal{H}}^{-1}(\mathbf{u}_i)), \quad \forall g \in \mathcal{G}, 1 \leq i \leq n.$$

Define  $\tilde{G}_n$  to be the empirical process based on  $((\mathbf{u}_i, y_i) : 1 \leq i \leq n)$ , and

$$\tilde{G}_n(f, s) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [f(\mathbf{u}_i) s(y_i) - \mathbb{E}[f(\mathbf{u}_i) s(y_i)]],$$

and take  $\tilde{\mathcal{G}} = \{g \circ \phi_{\mathcal{H}}^{-1} : g \in \mathcal{G}\}$ , then

$$G_n(g, r) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [g(\mathbf{x}_i) r(y_i) - \mathbb{E}[g(\mathbf{x}_i) r(y_i)]] = \frac{1}{\sqrt{n}} \sum_{i=1}^n [\tilde{g}(\mathbf{u}_i) r(y_i) - \mathbb{E}[\tilde{g}(\mathbf{u}_i) r(y_i)]] = \tilde{G}_n(\tilde{g}, r).$$

The relation between constants for  $\tilde{\mathcal{G}}$  and constants for  $\mathcal{G}$  can be deduced from Lemma SA.10. Hence, without loss of generality, we assume  $(\mathbf{x}_i : 1 \leq i \leq n)$  are i.i.d under common law  $\mathbb{P}_X = \text{Uniform}([0, 1]^d)$  distributed and  $\mathcal{X} = [0, 1]^d$ .

Take  $\mathcal{A}_{M,N}(\mathbb{P}_Z, 1)$  to be an axis-aligned cylindered quasi-dyadic expansion of  $\mathbb{R}^{d+1}$ , of depth  $M$  for the main subspace  $\mathbb{R}^d$  and depth  $N$  for the multiplier subspace  $\mathbb{R}$  with respect to  $\mathbb{P}_Z$ . Take  $(Z_n^G(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  and  $(\Pi_1 Z_n^G(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  to be the mean-zero Gaussian processes constructed as in Equations (SA-13) and (SA-14). Let  $(\mathcal{G} \times \mathcal{R})_\delta$  be a  $\delta \|M_{\mathcal{G}} M_{\mathcal{R}}\|_{\mathbb{P}_Z}$ -net of  $\mathcal{G} \times \mathcal{R}$  with cardinality no greater than  $N_{\mathcal{G} \times \mathcal{R}}(\delta, M_{\mathcal{G}} M_{\mathcal{R}})$ . By standard empirical process argument,  $N_{\mathcal{G} \times \mathcal{R}}(\delta, M_{\mathcal{G}} M_{\mathcal{R}}) \leq N(\delta)$ . By Lemma SA.16, the meshing error can be bounded by: For all  $t > 0$ ,

$$\mathbb{P}[\|G_n - G_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta}\|_{\mathcal{G} \times \mathcal{R}} + \|Z_n^G - Z_n^G \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta}\|_{\mathcal{G} \times \mathcal{R}} > C_1 c_{v,\alpha} \mathbf{F}_n^G(t, \delta)] \leq 8 \exp(-t),$$

where  $C_1$  is a universal constant and  $c_{v,\alpha} = v(1 + (2\alpha)^{\frac{\alpha}{2}})$ . Lemma SA.17 implies that the strong approxi-

mation error for the projected process on  $\delta$ -net is bounded by: For all  $t > 0$ ,

$$\mathbb{P} \left[ \|\Pi_1 G_n - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta} > C_1 c_{v,\alpha} \sqrt{\frac{N^{2\alpha+1} 2^M \mathbf{E}_G \mathbf{M}_G}{n}} t + C_1 c_{v,\alpha} \sqrt{\frac{\mathbf{C}_{\Pi_1(\mathcal{G} \times \mathcal{R}), M+N}}{n}} t \right] \leq 2N(\delta) e^{-t}.$$

where

$$\mathbf{C}_{\Pi_1(\mathcal{G} \times \mathcal{R}), M+N} = \sup_{f \in \Pi_1(\mathcal{G} \times \mathcal{R})} \min \left\{ \sup_{(j,k) \in \mathcal{I}_{M+N}} \left[ \sum_{j' < j} (j-j')(j-j'+1) 2^{j'-j} \sum_{k': \mathcal{C}_{j',k'} \subseteq \mathcal{C}_{j,k}} \tilde{\beta}_{j',k'}^2(f) \right], \sup_{\mathbf{z} \in \mathcal{C}_{M+N,0}} f(\mathbf{z})^2 (M+N) \right\}.$$

Now we upper bound the left hand side of the minimum. Let  $f \in \Pi_1(\mathcal{G} \times \mathcal{R})$ . Then there exists  $g \in \mathcal{G}$  and  $r \in \mathcal{R}$  such that  $f = \Pi_1[g, r]$ . Since  $f$  is already piecewise-constant, by definition of  $\beta_{j,k}$ 's and  $\gamma_{j,k}$ 's, we know  $\tilde{\beta}_{l,m}(f) = \tilde{\gamma}_{l,m}(g, r)$ . Fix  $(j, k) \in \mathcal{I}_{M+N}$ . We consider two cases.

**Case 1:**  $j > N$ . By Definition SA.7,  $\mathcal{C}_{j,k} = \mathcal{X}_{j-N,k} \times \mathcal{Y}_{*,N,0}$ . By definition of  $\mathcal{A}_{M,N}(\mathbb{P}_Z, 1)$  and the assumption that  $\mathbf{x}_i$ 's are Uniform( $[0, 1]^d$ ) distributed,  $\|\mathcal{X}_{j-N,k}\|_\infty \leq 2^{-\frac{M+N-j}{d}+1}$ .

Consider  $j'$  such that  $N \leq j' \leq j$ . By definition of  $\mathcal{A}_{M,N}(\mathbb{P}_Z, 1)$  and the assumption that  $\mathbf{x}_i$ 's are Uniform( $[0, 1]^d$ ) distributed, the  $j'$ -th level difference set  $\mathcal{U}_{j'} = \cup_{0 \leq k < 2^{M+N-j'}} (\mathcal{C}_{j'-1, 2k+1} - \mathcal{C}_{j'-1, 2k})$  is contained in  $[-2^{-\frac{M+N-j'}{d}+2}, 2^{-\frac{M+N-j'}{d}+2}]^d$ . Let  $g \in \mathcal{G}$ ,  $r \in \mathcal{R}$ . By definition of  $\tilde{\gamma}_{j',m}$  and similar arguments to those in the proof of Lemma SA.7,

$$\begin{aligned} \sum_{m: \mathcal{C}_{j',m} \subseteq \mathcal{C}_{j,k}} |\tilde{\gamma}_{j',m}(g, r)| &\leq 2^{2(M+N-j')} \int_{\mathcal{U}_{j'}} \int_{\mathcal{X}_{j-N,k}} |g(\mathbf{x})\theta(\mathbf{x}, r) - g(\mathbf{x} + \mathbf{s})\theta(\mathbf{x} + \mathbf{s}, r)| d\mathbf{x} d\mathbf{s} \\ &\leq 2^{2(M+N-j')} \int_{\mathcal{U}_{j'}} \|\mathbf{s}\| \|\mathcal{X}_{j-N,k}\|_\infty^{d-1} d\mathbf{s} \mathbf{K}_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}}^* \\ &\leq 2^{2(M+N-j')} \mathbf{m}(\mathcal{U}_{j'}) \|\mathcal{U}_{j'}\|_\infty \|\mathcal{X}_{j-N,k}\|_\infty^{d-1} \mathbf{K}_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}}^* \\ &\leq 2^{\frac{d-1}{d}(j-j')} \mathbf{K}_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}}^*. \end{aligned}$$

Next, consider  $j'$  such that  $0 \leq j' < N$ , we know

$$\begin{aligned} &\sum_{k': \mathcal{C}_{j',k'} \subseteq \mathcal{C}_{j,k}} |\tilde{\gamma}_{j',k'}(g, r)| \\ &= \sum_{j': \mathcal{X}_{0,j'} \subseteq \mathcal{X}_{j-N,k}} \sum_{0 \leq m < 2^{j'}} |\mathbb{E}[g(\mathbf{x}_i) | \mathcal{X}_{0,j'}] \cdot |\mathbb{E}[r(y_i) | \mathcal{X}_{0,j'} \times \mathcal{Y}_{j',j-1,2m}] - \mathbb{E}[r(y_i) | \mathcal{X}_{0,j'} \times \mathcal{Y}_{j',j-1,2m+1}]| \\ &\leq c_{v,\alpha} \sum_{j': \mathcal{X}_{0,j'} \subseteq \mathcal{X}_{j-N,k}} |\mathbb{E}[g(\mathbf{x}_i) | \mathcal{X}_{0,j'}]| N^\alpha \\ &\leq c_{v,\alpha} 2^{j-N} \mathbf{M}_G N^\alpha. \end{aligned}$$

It follows that

$$\begin{aligned}
& \sum_{j' < j} (j - j')(j - j' + 1)2^{j' - j} \sum_{k': \mathcal{C}_{j', k'} \subseteq \mathcal{C}_{j, k}} |\tilde{\gamma}_{j', k'}(g, r)| \\
& \leq \sum_{N \leq j' < j} (j - j')(j - j' + 1)2^{-\frac{j-j'}{d}} \mathsf{K}_{\mathfrak{G}, \mathcal{V}_{\mathfrak{R}}}^* + c_{v, \alpha} \sum_{j' < N} (j - j')(j - j' + 1)2^{j' - N} \mathsf{M}_{\mathfrak{G}} N^\alpha \\
& \lesssim \mathsf{K}_{\mathfrak{G}, \mathcal{V}_{\mathfrak{R}}}^* + c_{v, \alpha} \mathsf{M}_{\mathfrak{G}} N^\alpha.
\end{aligned}$$

**Case 2:**  $j \leq N$ . Then  $\mathcal{C}_{j, k} = \mathcal{X}_{0, l} \times \mathcal{Y}_{l, j, m}$  with  $k = 2^{N-j}l + m$ , and  $\mathcal{C}_{j', k'} = \mathcal{X}_{0, l'} \times \mathcal{Y}_{l', j', m'}$  with  $k' = 2^{N-j'}l' + m'$ . In particular,  $\mathcal{C}_{j', k'} \subseteq \mathcal{C}_{j, k}$  implies  $l' = l$  and  $\mathcal{Y}_{l', j', m'} \subseteq \mathcal{Y}_{l, j, m}$ . By a similar argument to the proof in Lemma SA.17 (Layers  $1 \leq j \leq N$ ), for any  $0 \leq j' \leq j$ ,

$$\begin{aligned}
& \sum_{k': \mathcal{C}_{j', k'} \subseteq \mathcal{C}_{j, k}} |\tilde{\gamma}_{j', k'}(g, r)| \\
& = |\mathbb{E}[g(\mathbf{x}_i) | \mathcal{X}_{0, l}]| \sum_{m': \mathcal{Y}_{l, j', m'} \subseteq \mathcal{Y}_{l, j, m}} |\mathbb{E}[r(y_i) | \mathcal{X}_{0, l} \times \mathcal{Y}_{l, j-1, 2m}] - \mathbb{E}[r(y_i) | \mathcal{X}_{0, l} \times \mathcal{Y}_{l, j-1, 2m+1}]| \\
& \leq c_{v, \alpha} |\mathbb{E}[g(\mathbf{x}_i) | \mathcal{X}_{0, l}]| N^\alpha \\
& \leq c_{v, \alpha} \mathsf{M}_{\mathfrak{G}} N^\alpha.
\end{aligned}$$

Using the elementary inequality that  $x(x+1) \leq 30 \cdot 2^{x/4}$  for  $x > 0$ , we can get

$$\sum_{1 \leq j' < j} (j - j')(j - j' + 1)2^{j' - j} \sum_{k': \mathcal{C}_{j', k'} \subseteq \mathcal{C}_{j, k}} |\tilde{\gamma}_{j', k'}(g, r)| \leq 60c_{v, \alpha} \mathsf{M}_{\mathfrak{G}} N^\alpha.$$

Moreover, for all  $(j, k)$ , we have  $\tilde{\beta}_{j, k}(g, r) \leq c_{v, \alpha} \mathsf{M}_{\mathfrak{G}} N^\alpha$ . This implies that

$$\mathsf{C}_{\Pi_1(\mathfrak{G} \times \mathfrak{R}), M+N} \lesssim c_{v, \alpha}^2 \mathsf{M}_{\mathfrak{G}} N^\alpha \min\{\mathsf{K}_{\mathfrak{G}, \mathcal{V}_{\mathfrak{R}}}^* + \mathsf{M}_{\mathfrak{G}} N^\alpha, \mathsf{M}_{\mathfrak{G}} N^\alpha (M + N)\}.$$

Since  $\mathbf{x}_i \stackrel{i.i.d.}{\sim} \text{Uniform}([0, 1]^d)$  and the cells  $\mathcal{A}_{M, N}(\mathbb{P}_Z, 1)$  are obtained via *axis aligned dyadic expansion*, we have  $\|\mathcal{X}_{0, k}\|_\infty \leq 2^{-\lfloor M/d \rfloor}$  for all  $0 \leq k < 2^M$ . Then by Lemma SA.21, for all  $t > N$ ,

$$\begin{aligned}
\mathbb{P}\left[\|G_n - \Pi_1 G_n\|_{(\mathfrak{G} \times \mathfrak{R})_\delta} \gtrsim \sqrt{c_{v, 2\alpha}} \sqrt{N^2 \mathsf{V}_{\mathfrak{G}} + 2^{-N} \mathsf{M}_{\mathfrak{G}}^2 t^{\alpha + \frac{1}{2}}} + c_{v, \alpha} \frac{\mathsf{M}_{\mathfrak{G}}}{\sqrt{n}} t^{\alpha + 1}\right] &\leq 4\mathsf{N}(\delta) n e^{-t}, \\
\mathbb{P}\left[\|Z_n^G - \Pi_1 Z_n^G\|_{(\mathfrak{G} \times \mathfrak{R})_\delta} \gtrsim \sqrt{c_{v, 2\alpha}} \sqrt{N^2 \mathsf{V}_{\mathfrak{G}} + 2^{-N} \mathsf{M}_{\mathfrak{G}}^2 t^{\frac{1}{2}}} + c_{v, \alpha} \frac{\mathsf{M}_{\mathfrak{G}}}{\sqrt{n}} t\right] &\leq 4\mathsf{N}(\delta) n e^{-t},
\end{aligned}$$

where  $c_{v, \alpha} = v(1 + (2\alpha)^{\frac{\alpha}{2}})$  and  $c_{v, 2\alpha} = v^2(1 + (4\alpha)^\alpha)$ , and

$$\mathsf{V}_{\mathfrak{G}} = \sqrt{d} \min\{2\mathsf{M}_{\mathfrak{G}}, \mathsf{L}_{\mathfrak{G}} 2^{-\lfloor M/d \rfloor}\} 2^{-\lfloor M/d \rfloor} \mathsf{TV}_{\mathfrak{G}}.$$

We find the optimal parameters  $M^*$  and  $N^*$  by balancing the term  $\sqrt{\frac{2^M \mathsf{E}_{\mathfrak{G}} \mathsf{M}_{\mathfrak{G}}}{n}}$  from the bound on  $\|\Pi_1 G_n - \Pi_1 Z_n^G\|_{(\mathfrak{G} \times \mathfrak{R})_\delta}$  and the term  $\mathsf{V}_{\mathfrak{G}}$  from the bounds on  $\|G_n - \Pi_1 G_n\|_{(\mathfrak{G} \times \mathfrak{R})_\delta}$  and  $\|Z_n - \Pi_1 Z_n^G\|_{(\mathfrak{G} \times \mathfrak{R})_\delta}$ , choosing

$$2^{M^*} = \min \left\{ \left( \frac{n \mathsf{TV}_{\mathfrak{G}}}{\mathsf{E}_{\mathfrak{G}}} \right)^{\frac{d}{d+1}}, \left( \frac{n \mathsf{L}_{\mathfrak{G}} \mathsf{TV}_{\mathfrak{G}}}{\mathsf{E}_{\mathfrak{G}} \mathsf{M}_{\mathfrak{G}}} \right)^{\frac{d}{d+2}} \right\}, \quad 2^{N^*} = \max \left\{ \left( \frac{n \mathsf{M}_{\mathfrak{G}}^{d+1}}{\mathsf{E}_{\mathfrak{G}} \mathsf{TV}_{\mathfrak{G}}^d} \right)^{\frac{1}{d+1}}, \left( \frac{n^2 \mathsf{M}_{\mathfrak{G}}^{2d+2}}{\mathsf{TV}_{\mathfrak{G}}^d \mathsf{L}_{\mathfrak{G}}^d \mathsf{E}_{\mathfrak{G}}^2} \right)^{\frac{1}{d+2}} \right\}.$$

It follows that for all  $t > N_*$ , with probability at least  $1 - 4n\mathbf{N}(\delta) \exp(-t)$ ,

$$\begin{aligned} & \|G_n - Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta} \\ & \leq \sqrt{d}N^* \min \left\{ \left( \frac{\mathbf{E}_\mathcal{G} \text{TV}_\mathcal{G}^d M_\mathcal{G}^{d+1}}{n} \right)^{\frac{1}{2(d+1)}}, \left( \frac{\mathbf{E}_\mathcal{G}^2 M_\mathcal{G}^2 \text{TV}_\mathcal{G}^d L_\mathcal{G}^d}{n^2} \right)^{\frac{1}{2(d+2)}} \right\} t^{\alpha + \frac{1}{2}} + \sqrt{\frac{\mathbf{C}_{\Pi_1(\mathcal{G} \times \mathcal{R}), M+N}}{n}} t^{\alpha+1}. \end{aligned}$$

The result then follows from the decomposition that

$$\begin{aligned} \|G_n - Z_n^G\|_{\mathcal{G} \times \mathcal{R}} &= \|G_n - Z_n^G\|_{\mathcal{G} \times \mathcal{R}} \\ &\leq \|G_n - G_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta}\|_{\mathcal{G} \times \mathcal{R}} + \|Z_n^G - Z_n^G \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta}\|_{\mathcal{G} \times \mathcal{R}} \\ &\quad + \|G_n - \Pi_1 G_n\|_{(\mathcal{G} \times \mathcal{R})_\delta} + \|Z_n^G - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta} + \|\Pi_1 G_n - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta}, \end{aligned}$$

and Lemma SA.10 for the reduction to the case of  $\text{Uniform}([0, 1]^d)$  distributed  $\mathbf{x}_i$ 's.  $\square$

### SA-III.3 Additional Results

This section presents the additional result for the  $G_n$ -process under VC-type entropy conditions. To simplify notation, the parameters of  $\mathcal{G}$  and  $\mathcal{G} \cdot \mathcal{V}_\mathcal{R}$  (Definitions 4 to 12, SA.1, SA.2) are taken with  $\mathcal{C} = \mathcal{Q}_\mathcal{G}$ , and the index  $\mathcal{Q}_\mathcal{G}$  is omitted where there is no ambiguity; the parameters of  $\mathcal{R}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{Y}$ , and the index  $\mathcal{Y}$  is omitted where there is no ambiguity; and the parameters of  $\mathcal{G} \times \mathcal{R}$  (Definitions 4 to 12, SA.3, SA.4) are taken with  $\mathcal{C} = \mathcal{Q}_\mathcal{G} \times \mathcal{Y}$ , and the index  $\mathcal{Q}_\mathcal{G} \times \mathcal{Y}$  is omitted where there is no ambiguity.

**Corollary SA.4** (VC-Type Lipschitz Functions). *Suppose the conditions of Theorem SA.1 and the following additional conditions hold.*

- (i)  $\mathcal{G}$  is a VC-type class with respect to envelope  $M_\mathcal{G}$  with constant  $c_\mathcal{G} \geq e$  and exponent  $d_\mathcal{G} \geq 1$  over  $\mathcal{Q}_\mathcal{G}$ .
- (ii)  $\mathcal{R}$  is a VC-type class with respect to envelope  $M_\mathcal{R}$  with constant  $c_\mathcal{R} \geq e$  and exponent  $d_\mathcal{R} \geq 1$  over  $\mathcal{Y}$ .
- (iii) There exists a constant  $\mathbf{k}$  such that  $|\log_2 \mathbf{E}_\mathcal{G}| + |\log_2 \text{TV}| + |\log_2 M_\mathcal{G}| \leq \mathbf{k} \log_2 n$ , where we take  $\text{TV} = \max\{\text{TV}_\mathcal{G}, \text{TV}_{\mathcal{G} \cdot \mathcal{V}_\mathcal{R}}\}$ .

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

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

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

$$\begin{aligned} \mathbf{T}_n^G(t) &= \sqrt{d} \min \left\{ \left( \frac{c_1^d \mathbf{E}_\mathcal{G} \text{TV}_\mathcal{G}^d M_\mathcal{G}^{d+1}}{n} \right)^{\frac{1}{2(d+1)}}, \left( \frac{c_1^d c_2^d \mathbf{E}_\mathcal{G}^2 M_\mathcal{G}^2 \text{TV}_\mathcal{G}^d L_\mathcal{G}^d}{n^2} \right)^{\frac{1}{2(d+2)}} \right\} (t + \mathbf{k} \log_2(n) + d \log(cn))^{\alpha+1} \\ &\quad + \sqrt{\frac{\min\{\mathbf{k} \log_2(n) M_\mathcal{G}^2, M_\mathcal{G}(c_3 K_{\mathcal{G} \cdot \mathcal{V}_\mathcal{R}} + M_\mathcal{G})\}}{n}} (\log n)^\alpha (t + \mathbf{k} \log_2(n) + d \log(cn))^{\alpha+1}, \end{aligned}$$

with  $c = c_\mathcal{G} c_\mathcal{R}$ ,  $d = d_\mathcal{G} + d_\mathcal{R}$ .

**Proof of Corollary SA.4.** The proof follows by Theorem SA.1 with  $\delta = n^{-1/2}$ , and

$$\mathbb{N}(n^{-1/2}) = \mathbb{N}_{\mathcal{G}}(1/\sqrt{2n}, M_{\mathcal{G}})\mathbb{N}_{\mathcal{R}}(1/\sqrt{2n}, M_{\mathcal{R}}) \leq c_{\mathcal{G}}c_{\mathcal{R}}(2\sqrt{n})^{\mathbf{d}_{\mathcal{G}}+\mathbf{d}_{\mathcal{R}}} = c(2\sqrt{n})^{\mathbf{d}},$$

and

$$\begin{aligned} J(n^{-1/2}) &= \sqrt{2}J(\mathcal{G}, M_{\mathcal{G}}, 1/\sqrt{2n}) + \sqrt{2}J(\mathcal{R}, M_{\mathcal{R}}, 1/\sqrt{2n}) \\ &\leq 3n^{-1/2}\sqrt{\mathbf{d}_{\mathcal{G}}\log(c_{\mathcal{G}}\sqrt{n})} + 3\delta\sqrt{\mathbf{d}_{\mathcal{R}}\log(c_{\mathcal{R}}\sqrt{n})} \\ &\leq 3\delta\sqrt{(\mathbf{d}_{\mathcal{G}} + \mathbf{d}_{\mathcal{R}})\log(c_{\mathcal{G}}c_{\mathcal{R}}n)} \leq 3\delta\sqrt{\mathbf{d}\log(cn)}. \end{aligned}$$

The conclusion follows.  $\square$

## SA-IV Residual-Based Empirical Processes

Recall that  $\mathbf{z}_i = (\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$ ,  $i = 1, \dots, n$ , are i.i.d. random vectors supported on a background probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , and the *residual-based empirical process* is

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

In particular,  $(R_n(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  can be seen as a combination of the two empirical processes studied in the previous sections: for  $r \in \mathcal{R}$  and  $\mathbf{x} \in \mathcal{X}$ ,

$$R_n(g, r) = G_n(g, r) - X_n(g\theta(\cdot, r)), \quad \theta(\mathbf{x}, r) = \mathbb{E}[r(y_i)|\mathbf{x}_i = \mathbf{x}],$$

where

$$\begin{aligned} G_n(g, r) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n [g(\mathbf{x}_i)r(y_i) - \mathbb{E}[g(\mathbf{x}_i)r(y_i)]], \\ X_n(g\theta(\cdot, r)) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n [g(\mathbf{x}_i)\theta(\mathbf{x}_i, r) - \mathbb{E}[g(\mathbf{x}_i)\theta(\mathbf{x}_i, r)]]. \end{aligned}$$

Results for the  $X_n$  process (Section SA-II) and for the  $G_n$  process (Section SA-III) will be used to handle the terms above. The same error decomposition as in Sections SA-II and SA-III also applies here:

$$\begin{aligned} \|R_n - Z_n^R\|_{\mathcal{G} \times \mathcal{R}} &\leq \|R_n - Z_n^R\|_{(\mathcal{G} \times \mathcal{R})_{\delta}} + \|R_n - R_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}}\|_{\mathcal{G} \times \mathcal{R}} + \|Z_n^R \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}} - Z_n^R\|_{\mathcal{G} \times \mathcal{R}} \\ &\leq \|\Pi_2 Z_n^R - Z_n^R\|_{(\mathcal{G} \times \mathcal{R})_{\delta}} + \|R_n - \Pi_2 R_n\|_{(\mathcal{G} \times \mathcal{R})_{\delta}} + \|\Pi_2 R_n - \Pi_2 Z_n^R\|_{(\mathcal{G} \times \mathcal{R})_{\delta}} \\ &\quad + \|R_n - R_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}}\|_{\mathcal{G} \times \mathcal{R}} + \|Z_n^R \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}} - Z_n^R\|_{\mathcal{G} \times \mathcal{R}}, \end{aligned}$$

where  $(\mathcal{G} \times \mathcal{R})_{\delta}$  denotes a discretization (or meshing) of  $\mathcal{G} \times \mathcal{R}$  (i.e.,  $\delta$ -net of  $\mathcal{G} \times \mathcal{R}$ ), and the terms  $\|R_n - R_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}}\|_{\mathcal{G} \times \mathcal{R}}$  and  $\|Z_n^R \circ \pi_{(\mathcal{G} \times \mathcal{R})_{\delta}} - Z_n^R\|_{\mathcal{G} \times \mathcal{R}}$  capture the fluctuations (or oscillations) of  $R_n$  and  $Z_n^R$  relative to the meshing for each of the stochastic processes.  $\|\Pi_2 R_n - \Pi_2 Z_n^R\|_{(\mathcal{G} \times \mathcal{R})_{\delta}}$  and  $\|\Pi_2 Z_n^R - Z_n^R\|_{(\mathcal{G} \times \mathcal{R})_{\delta}}$  represent projections onto a Haar function space, where  $\Pi_2 R_n(h) = R_n \circ \Pi_2 h$ . The operator  $\Pi_2$  is a projection onto piecewise constant functions that respects the multiplicative structure of the  $R_n$  process. The final term



$\|\Pi_2 R_n - \Pi_2 Z_n^R\|_{(\mathcal{G} \times \mathcal{R})_\delta}$  captures the coupling between the empirical process and the Gaussian process (on a  $\delta$ -net of  $\mathcal{G} \times \mathcal{R}$ , after the projection  $\Pi_2$ ).

The general result under uniform entropy integral conditions is presented in Section SA-IV.2. Theorem 2 and Corollary 4 then follow from that general result. The proofs leverage the existence of a surrogate measure and a normalizing transformation of  $\mathcal{G}$  with respect to  $\mathbb{P}_X$ , the distribution of  $\mathbf{x}_1$ , as developed in Section SA-II.2. We will use the same class of cylindered quasi-dyadic cell expansions as in Section SA-III.1.1, which explicitly exploits the multiplicative structure of  $R_n$ . Bounds for each term in the error decomposition are provided in Section SA-IV.1, which boils down to handle the extra  $X_n(g\theta(\cdot, r))$  term compared to the results in Section SA-III.1 and is organized as follows:

- Section SA-IV.1.1 introduces the *conditional mean adjusted product-factorized projection* that combines the *product-factorized projection* for the  $G_n(g, r)$  part and the  $L_2$  projection for the  $X_n(g\theta(\cdot, r))$  part.
- Section SA-IV.1.2 constructs the Gaussian process  $(Z_n^R(g, r) : (g, r) \in \mathcal{G} \times \mathcal{R})$ . The construction is essentially the same as those in Section SA-II.1.3, relying on coupling binomial random variables with Gaussian random variables.
- Section SA-IV.1.3 handles the meshing errors  $\|R_n - R_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta}\|_{\mathcal{G} \times \mathcal{R}}$  and  $\|Z_n^R \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta} - Z_n^R\|_{\mathcal{G} \times \mathcal{R}}$  using standard empirical process results.
- Section SA-IV.1.4 addresses the strong approximation error  $\|\Pi_2 R_n - \Pi_2 Z_n^R\|_{(\mathcal{G} \times \mathcal{R})_\delta}$ . With the help of the relation between  $\Pi_1$  and  $\Pi_2$ , we can reuse results from Section SA-III.1.5.
- Section SA-IV.1.5 addresses the projection errors  $\|R_n - \Pi_2 R_n\|_{(\mathcal{G} \times \mathcal{R})_\delta}$  and  $\|Z_n^R - \Pi_1 Z_n^R\|_{(\mathcal{G} \times \mathcal{R})_\delta}$ . We use the results from Section SA-III.1.6 for  $\|G_n - \Pi_1 G_n\|_{(\mathcal{G} \times \mathcal{R})_\delta}$ , and deal with  $\|X_n(g\theta(\cdot, r)) - \Pi_0 X_n(g\theta(\cdot, r))\|_{(\mathcal{G} \times \mathcal{R})_\delta}$  using results from Section SA-II.1.6.

## SA-IV.1 Preliminary Technical Results

This section presents preliminary technical results that are used to prove Theorem SA.1. Whenever possible, these results are presented at a higher level of generality, and therefore may be of independent theoretical interest. Throughout this section, we assume the same set of conditions (Assumption SA.2) on data generate process as in Section SA-III.1.

Compared to the assumptions in Theorem 2, this assumption does not require the existence of a surrogate measure or a normalizing transformation. It will be applied in the analysis of terms in the error decomposition, where we work with the  $\mathbb{P}_Z$  distribution and extra condition on the existence of Lebesgue density of  $\mathbb{P}_X$  is assumed whenever necessary (Section SA-IV.1.5). The surrogate measure and the normalizing transformation will be used in the proof of Theorem SA.1 with the help of Section SA-II.2, providing greater flexibility in the data generating process.

### SA-IV.1.1 Projection onto Piecewise Constant Functions

For the residual empirical process, we tailor a projection to piecewise constant functions on the quasi-dyadic cells that differs from the mean square projection from Section SA-II.1.2 and the product-factorized projection from Section SA-III.1.2. Given a cylindered quasi-dyadic expansion of  $\mathbb{R}^{d+1}$ ,  $\mathcal{C}_{M,N}(\mathbb{P}, \rho)$  with  $\mathbb{P}$  the law of random vector  $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$ , and recall the definition of  $\mathcal{E}_{M+N}$  from Section SA-II.1.2, for any

real valued functions  $g$  on  $\mathbb{R}^d$  and  $r$  on  $\mathbb{R}$  such that  $\int_{\mathbb{R}^d} \int_{\mathbb{R}} g(\mathbf{x})^2 \mathbb{P}(dyd\mathbf{x}) < \infty$  and  $\int_{\mathbb{R}^d} \int_{\mathbb{R}} r(y)^2 \mathbb{P}(dyd\mathbf{x}) < \infty$ , the *conditional mean adjusted product-factorized projection* of  $g$  and  $r$  is defined as

$$\Pi_2(\mathcal{C}_{M,N}(\mathbb{P}, \rho))[g, r] = \Pi_1(\mathcal{C}_{M,N}(\mathbb{P}, \rho))[g, r] - \Pi_0(\mathbf{p}_X[\mathcal{C}_{M,N}(\mathbb{P}, \rho)])[g\theta(\cdot, r)], \quad (\text{SA-16})$$

where  $\theta(\mathbf{x}, r) = \mathbb{E}[r(Y)|\mathbf{X} = \mathbf{x}]$  for  $r \in \mathcal{R}$  and  $\mathbf{x} \in \mathcal{X}$ , and  $\mathbf{p}_X[\mathcal{C}_{M,N}(\mathbb{P}, \rho)] = \{\mathcal{X}_{l,k} : 0 \leq l \leq M, 0 \leq k < 2^{M-l}\}$  as defined in Definition SA.7. We denote the collection of conditional mean functions based on  $\mathcal{R}$  by  $\mathcal{V}_{\mathcal{R}} = \{\theta(\cdot, r) : r \in \mathcal{R}\}$ .

This projection can also be represented using the Haar basis as

$$\Pi_2(\mathcal{C}_{M,N}(\mathbb{P}, \rho))[g, r] = \eta_{M+N,0}(g, r)e_{M+N,0} + \sum_{1 \leq j \leq M+N} \sum_{0 \leq k < 2^{M+N-j}} \tilde{\eta}_{j,k}(g, r)\tilde{e}_{j,k},$$

with

$$\eta_{j,k}(g, r) = \begin{cases} 0, & \text{if } N \leq j \leq M+N, \\ \gamma_{j,k}(g, r), & \text{if } j < N. \end{cases} \quad (\text{SA-17})$$

We will use  $\Pi_2$  as shorthand for  $\Pi_2(\mathcal{C}_{M,N}(\mathbb{P}, \rho))$ .

Next, we define the empirical processes indexed by these projected functions. With a slight abuse of notation, let  $(X_n(f) : f \in \mathcal{F})$  be the empirical process based on a random sample  $((\mathbf{x}_i, y_i) : 1 \leq i \leq n)$ , where  $\mathcal{F}$  is a class of real-valued functions on  $\mathbb{R}^{d+1}$ . Specifically,  $X_n(f) = n^{-1/2} \sum_{i=1}^n (f(\mathbf{x}_i, y_i) - \mathbb{E}[f(\mathbf{x}_i, y_i)])$  for  $f \in \mathcal{F}$ . For any real valued functions  $g$  on  $\mathbb{R}^d$  and  $r$  on  $\mathbb{R}$  such that  $\int_{\mathbb{R}^d} \int_{\mathbb{R}} g(\mathbf{x})^2 \mathbb{P}(dyd\mathbf{x}) < \infty$  and  $\int_{\mathbb{R}^d} \int_{\mathbb{R}} r(y)^2 \mathbb{P}(dyd\mathbf{x}) < \infty$ , we define

$$\begin{aligned} \Pi_2 R_n(g, r) &= X_n \circ \Pi_2(g, r), \\ \Pi_0 R_n(g, r) &= X_n \circ \Pi_0[\mathcal{C}_{M,N}(\mathbb{P}, \rho)](gr) - X_n \circ \Pi_0(\mathbf{p}_X[\mathcal{C}_{M,N}(\mathbb{P}, \rho)])[g\theta(\cdot, r)]. \end{aligned} \quad (\text{SA-18})$$

### SA-IV.1.2 Strong Approximation Constructions

**Lemma SA.22.** *Suppose Assumption SA.2 holds, and a cylindered quasi-dyadic expansion  $\mathcal{C}_K(\mathbb{P}_Z, \rho)$  is given. Then,  $(\mathcal{G} \cdot \mathcal{R}) \cup (\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}) \cup \Pi_1(\mathcal{G} \times \mathcal{R}) \cup \Pi_2(\mathcal{G} \times \mathcal{R}) \cup \Pi_0[\mathbf{p}_X(\mathcal{C}_{M,N})](\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}})$  is  $\mathbb{P}_Z$ -pregaussian.*

*Proof.* Recall we have shown in the proof of Lemma SA.12 that for all  $0 < \delta < 1$ ,

$$\begin{aligned} J_{\mathcal{X} \times \mathcal{Y}}(\mathcal{G} \cdot \mathcal{R}, \mathbf{M}_{\mathcal{G}, \mathcal{X}} M_{\mathcal{R}, \mathcal{Y}}, \delta) &\lesssim \sqrt{2} J_{\mathcal{X}}(\mathcal{G}, \mathbf{M}_{\mathcal{G}, \mathcal{X}}, \delta/\sqrt{2}) + \sqrt{2} J_{\mathcal{Y}}(\mathcal{R}, M_{\mathcal{R}, \mathcal{Y}}, \delta/\sqrt{2}), \\ J_{\mathcal{X} \times \mathcal{Y}}(\Pi_1(\mathcal{G} \times \mathcal{R}), c_{v,\alpha} \mathbf{M}_{\mathcal{G}, \mathcal{X}} N^\alpha, \delta) &\lesssim \sqrt{2} J_{\mathcal{X}}(\mathcal{G}, \mathbf{M}_{\mathcal{G}, \mathcal{X}}, \delta/(3\sqrt{2})) + \sqrt{2} J_{\mathcal{Y}}(\mathcal{R}, M_{\mathcal{R}, \mathcal{Y}}, \delta/(3\sqrt{2})), \end{aligned}$$

where  $c_{v,\alpha} = v(1 + (2\alpha)^{\frac{\alpha}{2}})$ . Lemma SA.25 implies  $J_{\mathcal{X}}(\mathcal{V}_{\mathcal{R}}, \theta(\cdot, M_{\mathcal{R}, \mathcal{Y}}), \delta) \leq J_{\mathcal{Y}}(\mathcal{R}, M_{\mathcal{R}, \mathcal{Y}}, \delta)$ . Since Assumption SA.2 (iv) implies  $\sup_{\mathbf{x} \in \mathcal{X}} \theta(\cdot, M_{\mathcal{R}, \mathcal{Y}}) \leq c_{v,\alpha} \mathbf{M}_{\mathcal{G}}$ , we know for all  $0 < \delta < 1$ ,

$$\begin{aligned} J_{\mathcal{X}}(\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}, c_{v,\alpha} \mathbf{M}_{\mathcal{G}}, \delta) &\leq \sqrt{2} J_{\mathcal{X}}(\mathcal{G}, \mathbf{M}_{\mathcal{G}, \mathcal{X}}, \delta/\sqrt{2}) + \sqrt{2} J_{\mathcal{X}}(\mathcal{V}_{\mathcal{R}}, \theta(\cdot, M_{\mathcal{R}, \mathcal{Y}}), \delta/\sqrt{2}) \\ &\leq \sqrt{2} J_{\mathcal{X}}(\mathcal{G}, \mathbf{M}_{\mathcal{G}, \mathcal{X}}, \delta/\sqrt{2}) + \sqrt{2} J_{\mathcal{Y}}(\mathcal{R}, M_{\mathcal{R}, \mathcal{Y}}, \delta/\sqrt{2}). \end{aligned}$$

The same argument for Lemma SA.12 implies that for all  $0 < \delta < 1$ ,

$$J_{\mathcal{X}}(\Pi_0[\mathfrak{p}_X(\mathcal{C}_{M,N})](\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}), C_{v,\alpha} M_{\mathcal{G},\mathcal{X}} N^\alpha, \delta) \leq J_{\mathcal{X}}(\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}, C_{v,\alpha} M_{\mathcal{G},\mathcal{X}} N^\alpha, \delta).$$

Moreover Lemma SA.15 implies  $\Pi_2(\mathcal{G} \times \mathcal{R}) \subseteq \Pi_1(\mathcal{G} \times \mathcal{R}) + \Pi_0[\mathfrak{p}_X(\mathcal{C}_{M,N})](\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}})$ . It follows from pointwise separability of  $\mathcal{G}$  and  $\mathcal{R}$  and Corollary 2.2.9 in van der Vaart and Wellner (2013) that  $(\mathcal{G} \cdot \mathcal{R}) \cup (\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}) \cup \Pi_1(\mathcal{G} \times \mathcal{R}) \cup \Pi_2(\mathcal{G} \times \mathcal{R}) \cup \Pi_0[\mathfrak{p}_X(\mathcal{C}_{M,N})](\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}})$  is  $\mathbb{P}_Z$ -pregaussian.  $\square$

**Lemma SA.23.** *Suppose Assumption SA.2 holds and a cylindered dyadic expansion  $\mathcal{C}_{M,N}(\mathbb{P}_Z, 1)$  is given. Then on a possibly enlarged probability space, there exists a  $\mathbb{P}_Z$ -Brownian bridge  $B_n$  indexed by  $\mathcal{F} = (\mathcal{G} \cdot \mathcal{R}) \cup \Pi_0(\mathcal{G} \times \mathcal{R}) \cup \Pi_1(\mathcal{G} \times \mathcal{R})$  with almost sure continuous trajectories on  $(\mathcal{F}, \mathfrak{d}_{\mathbb{P}_Z})$  such that for any  $f \in \mathcal{F}$  and any  $x > 0$ ,*

$$\mathbb{P} \left( \left| \sum_{i=1}^n f(\mathbf{x}_i, y_i) - \sqrt{n} B_n(f) \right| \geq 24 \sqrt{\|f\|_{\mathcal{E}_{M+N}}^2 x} + 4 \sqrt{\mathcal{C}_{\{f\}, M+N} x} \right) \leq 2 \exp(-x),$$

where for both  $\|f\|_{\mathcal{E}_{M+N}}^2$  and  $\mathcal{C}_{\{f\}, M+N}$  are defined in Lemma SA.3.

**Proof of Lemma SA.23.** The result follows from Lemma SA.22 and the same argument as Lemma SA.13.  $\square$

**Lemma SA.24.** *Suppose Assumption SA.2 holds and a cylindered quasi-dyadic expansion  $\mathcal{C}_{M,N}(\mathbb{P}_Z, \rho)$  with  $\rho > 1$  is given. Then on a possibly enlarged probability space, there exists a Brownian bridge  $B_n$  indexed by  $\mathcal{F} = (\mathcal{G} \cdot \mathcal{R}) \cup (\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}) \cup \Pi_1(\mathcal{G} \times \mathcal{R}) \cup \Pi_2(\mathcal{G} \times \mathcal{R}) \cup \Pi_0[\mathfrak{p}_X(\mathcal{C}_{M,N})](\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}})$  with almost sure continuous trajectories on  $(\mathcal{F}, \mathfrak{d}_{\mathbb{P}_Z})$  such that for any  $f \in \mathcal{F}$  and any  $x > 0$ ,*

$$\begin{aligned} \mathbb{P} \left( \left| \sum_{i=1}^n f(\mathbf{x}_i, y_i) - \sqrt{n} B_n(f) \right| \geq C_\rho \sqrt{\|f\|_{\mathcal{E}_{M+N}}^2 x} + C_\rho \sqrt{\mathcal{C}_{\{f\}, M+N} x} \right) \\ \leq 2 \exp(-x) + 2^{M+2} \exp(-C_\rho n 2^{-M}), \end{aligned}$$

where  $C_\rho$  is a constant that only depends on  $\rho$ .

**Proof of Lemma SA.24.** The result follows from Lemma SA.22 and the same argument as Lemma SA.14.  $\square$

The above two lemmas enable the construction of Gaussian processes and their projected counterparts as analogs to the empirical processes defined in Section SA-II.1.3 and Section SA-III.1.3. In particular, we define  $Z_n^R$  and  $\Pi_2 Z_n^R$  as Gaussian processes indexed by  $\mathcal{G} \times \mathcal{R}$  such that, for any  $g \in \mathcal{G}$  and  $r \in \mathcal{R}$ ,

$$\begin{aligned} Z_n^R(g, r) &= B_n(g(r - \theta(\cdot, r))), \\ \Pi_2 Z_n^R(g, r) &= B_n(\Pi_2[g, r]). \end{aligned} \tag{SA-19}$$

We also define the following ancillary processes for analysis:

$$\begin{aligned} Z_n^G(g, r) &= B_n(gr), & \Pi_1 Z_n^G(g, r) &= B_n(\Pi_1[g, r]), \\ Z_n^X(g \theta(\cdot, r)) &= B_n(g \theta(\cdot, r)), & \Pi_0 Z_n^X(g \theta(\cdot, r)) &= B_n(\Pi_0[\mathfrak{p}_X(\mathcal{C}_{M,N})][g \theta(\cdot, r)]). \end{aligned} \tag{SA-20}$$

Since for any  $g_1, g_2 \in \mathcal{G}$ ,  $r_1, r_2 \in \mathcal{R}$ ,

$$\mathfrak{d}_{\mathbb{P}_Z}(g_1(r_1 - \theta(\cdot, r_1)), g_2(r_2 - \theta(\cdot, r_2))) \leq 2\mathfrak{d}_{\mathbb{P}_Z}(g_1 r_1, g_2 r_2),$$

and  $B_n$  has almost sure continuous sample trajectories on  $(\mathcal{G} \cdot \mathcal{R}, \mathfrak{d}_{\mathbb{P}_Z})$ , Equation (SA-19) also implies  $(Z_n^R(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  has almost sure continuous sample trajectories on  $(\mathcal{G} \times \mathcal{R}, \mathfrak{d}_{\mathbb{P}_Z})$ .

The following ancillary lemma on uniform covering number of the class of conditional means is used for the proof of Lemma SA.22.

**Lemma SA.25.** *Suppose  $\mathcal{S}$  is a class of functions from a measurable space  $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$  to  $\mathbb{R}$ , where  $\mathcal{Y} \subseteq \mathbb{R}$ , with envelope function  $M_{\mathcal{S}, \mathcal{Y}}$ . Let  $\mathcal{V}_{\mathcal{S}}$  be the class of conditional means  $\{\theta(\cdot, s) : s \in \mathcal{S}\}$  with  $\theta(\mathbf{x}, s) = \mathbb{E}[s(y_i) | \mathbf{x}_i = \mathbf{x}]$  for  $\mathbf{x} \in \mathcal{X}$ . Then*

$$N_{\mathcal{V}_{\mathcal{S}}, \mathcal{X}}(\delta, \theta(\cdot, M_{\mathcal{S}, \mathcal{Y}})) \leq N_{\mathcal{S}, \mathcal{Y}}(\delta, M_{\mathcal{S}, \mathcal{Y}}).$$

**Proof of Lemma SA.25.** Let  $\mathcal{Q}$  be a finite discrete measure on  $\mathbb{R}^d$ , and let  $r, s \in \mathcal{S}$ . Define a new probability measure  $\tilde{P}$  on  $\mathbb{R}$  by

$$\tilde{P}(A) = \int \mathbb{E}[\mathbb{1}((\mathbf{x}_i, y_i) \in \mathbb{R}^d \times A) | \mathbf{x}_i = \mathbf{x}] d\mathcal{Q}(\mathbf{x}), \quad \forall A \subseteq \mathbb{R}^d.$$

Then  $\int |\theta(\cdot, M_{\mathcal{S}, \mathcal{Y}})| d\tilde{P} \leq \int_{\mathbb{R}^d} \mathbb{E}[M_{\mathcal{S}, \mathcal{Y}}(y_i) | \mathbf{x}_i = \mathbf{x}] d\mathcal{Q}(z) < \infty$ , since  $\sup_{m \in \mathcal{V}_{\mathcal{S}}} \|m\|_{\infty} < \infty$ .

For  $r, s \in \mathcal{S}$ , we have

$$\int |\theta(\cdot, r) - \theta(\cdot, s)|^2 d\mathcal{Q} \leq \int_{\mathbb{R}^d} \mathbb{E}[|r(y_i) - s(y_i)|^2 | \mathbf{x}_i = \mathbf{x}] d\mathcal{Q}(x) = \int |r - s|^2 d\tilde{P}.$$

Here,  $\tilde{P}$  is not necessarily finite or discrete, but by a similar argument as in Lemma SA.15, there exists a subset  $\mathcal{S}_{\varepsilon} \subseteq \mathcal{S}$  with cardinality no greater than  $N_{\mathcal{S}, \mathcal{Y}}(\delta, M_{\mathcal{S}, \mathcal{Y}})$ , such that for any  $s \in \mathcal{S}$ , there exists  $r \in \mathcal{S}_{\varepsilon}$  with  $\|r - s\|_{\tilde{P}, 2} \leq \varepsilon \|\theta(\cdot, M_{\mathcal{S}, \mathcal{Y}})\|_{\tilde{P}, 2}$ . Hence,  $\|m_r - m_s\|_{\mathcal{Q}, 2} \leq \varepsilon \|\theta(\cdot, M_{\mathcal{S}, \mathcal{Y}})\|_{\tilde{P}, 2} = \varepsilon \|\theta(\cdot, M_{\mathcal{S}, \mathcal{Y}})\|_{\mathcal{Q}, 2}$ . The conclusion then follows.  $\square$

### SA-IV.1.3 Meshing Error

To simplify notation, the parameters of  $\mathcal{G}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{X}$ , and the index  $\mathcal{X}$  is omitted where there is no ambiguity; the parameters of  $\mathcal{R}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{Y}$ , and the index  $\mathcal{Y}$  is omitted where there is no ambiguity; and the parameters of  $\mathcal{G} \times \mathcal{R}$  (Definitions 4 to 12, SA.3, SA.4) are taken with  $\mathcal{C} = \mathcal{X} \times \mathcal{Y}$ , and the index  $\mathcal{X} \times \mathcal{Y}$  is omitted where there is no ambiguity. We also define

$$\begin{aligned} J(\delta) &= \sqrt{2}J(\mathcal{G}, \mathbf{M}_{\mathcal{G}}, \delta/\sqrt{2}) + \sqrt{2}J(\mathcal{R}, M_{\mathcal{R}}, \delta/\sqrt{2}), & \delta \in (0, 1], \\ \mathbf{N}(\delta) &= N_{\mathcal{G}}(\delta/\sqrt{2}, \mathbf{M}_{\mathcal{G}})N_{\mathcal{R}}(\delta/\sqrt{2}, M_{\mathcal{R}}), & \delta \in (0, 1]. \end{aligned}$$

For  $0 < \delta \leq 1$ , consider a  $\delta \mathbf{M}_{\mathcal{G}} \|M_{\mathcal{R}}\|_{\mathbb{P}_{\mathcal{Y}, 2}}$ -net of  $(\mathcal{G} \times \mathcal{R}, \|\cdot\|_{\mathbb{P}_{Z, 2}})$ , denoted by  $(\mathcal{G} \times \mathcal{R})_{\delta}$ , with cardinality at most  $N_{\mathcal{G} \times \mathcal{R}}(\delta, \mathbf{M}_{\mathcal{G}} \|M_{\mathcal{R}}\|_{\mathbb{P}_{\mathcal{Y}, 2}})$ . Define the projection onto the  $\delta$ -net as a mapping  $\pi_{(\mathcal{G} \times \mathcal{R})_{\delta}} : \mathcal{G} \times \mathcal{R} \rightarrow \mathcal{G} \times \mathcal{R}$  such that  $\|\pi_{(\mathcal{G} \times \mathcal{R})_{\delta}}(g, r) - g r\|_{\mathbb{P}_{Z, 2}} \leq \delta \mathbf{M}_{\mathcal{G}} \|M_{\mathcal{R}}\|_{\mathbb{P}_{\mathcal{Y}, 2}}$  for all  $g \in \mathcal{G}$  and  $r \in \mathcal{R}$ .

**Lemma SA.26.** *Suppose Assumption SA.2 holds, a cylindered quasi-dyadic expansion  $\mathcal{C}_{M,N}(\mathbb{P}_Z, \rho)$  is given,  $(Z_n^R(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  is the Gaussian process constructed as in (SA-19) on a possibly enlarged probability space, and  $(\mathcal{G} \times \mathcal{R})_\delta$  is chosen in Section SA-III.1.4. For all  $t > 0$  and  $0 < \delta < 1$ ,*

$$\mathbb{P}[\|R_n - R_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta}\|_{\mathcal{G} \times \mathcal{R}} + \|Z_n^R \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta} - Z_n^R\|_{\mathcal{G} \times \mathcal{R}} > C_1 c_{v,\alpha} F_n^R(t, \delta)] \leq \exp(-t),$$

where  $c_{v,\alpha} = v(1 + (2\alpha)^{\frac{\alpha}{2}})$  and

$$F_n^R(t, \delta) = J(\delta)M_{\mathcal{G}} + \frac{(\log n)^{\alpha/2} M_{\mathcal{G}} J^2(\delta)}{\delta^2 \sqrt{n}} + \frac{M_{\mathcal{G}}}{\sqrt{n}} t + (\log n)^\alpha \frac{M_{\mathcal{G}}}{\sqrt{n}} t^\alpha.$$

**Proof of Lemma SA.26.** Recall for any  $g \in \mathcal{G}$ ,  $r \in \mathcal{R}$ ,

$$R_n(g, r) = G_n(g, r) + X_n[\mathbb{P}_X(\mathcal{C}_{M,N}(\mathbb{P}_Z, \rho))](g\theta(\cdot, r)).$$

Lemma SA.16 implies that for any  $t > 0$  and  $0 < \delta < 1$ ,

$$\mathbb{P}[\|G_n - G_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta}\|_{\mathcal{G} \times \mathcal{R}} + \|Z_n^G \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta} - Z_n^G\|_{\mathcal{G} \times \mathcal{R}} > C_1 c_{v,\alpha} F_n^R(t, \delta)] \leq \exp(-t).$$

For  $g \in \mathcal{G}$ ,  $r \in \mathcal{R}$ , and take  $(g_0, r_0) = \pi_{(\mathcal{G} \times \mathcal{R})_\delta}$ , Jensen's inequality implies

$$\begin{aligned} & \|X_n(g\theta(\cdot, r)) - X_n(g_0\theta(\cdot, r_0))\|_2^2 \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(g(\mathbf{x}_i)\mathbb{E}[r(y_i)|\mathbf{x}_i] - g_0(\mathbf{x}_i)\mathbb{E}[r_0(y_i)|\mathbf{x}_i])^2] \\ &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(g(\mathbf{x}_i)r(y_i) - g_0(\mathbf{x}_i)r_0(y_i))^2] = \|G_n(g, r) - G_n(g_0, r_0)\|_2^2. \end{aligned}$$

Thus,

$$\| \|X_n(g\theta(\cdot, r)) - X_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta}(g\theta(\cdot, r))\|_2 \|_{\mathcal{G} \times \mathcal{R}} \leq \| \|G_n - G_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta}\|_2 \|_{\mathcal{G} \times \mathcal{R}}.$$

Lemma SA.25 implies that if we define  $\mathcal{G} \times \overline{\mathcal{R}} = \{g(r - \theta(\cdot, r)) : g \in \mathcal{G}, r \in \mathcal{R}\}$ , then

$$N_{\mathcal{G} \times \overline{\mathcal{R}}, \mathcal{X} \times \mathcal{Y}}(\delta, M_{\mathcal{G}} M_{\mathcal{R}}) \leq 2N(\delta).$$

The conclusion then follows by applying the same empirical process argument to  $\|X_n(g\theta(\cdot, r)) - X_n \circ \pi_{(\mathcal{G} \times \mathcal{R})_\delta}\|_{\mathcal{G} \times \mathcal{R}}$  as in Lemma SA.16.  $\square$

#### SA-IV.1.4 Strong Approximation Errors

To simplify notation, the parameters of  $\mathcal{G}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{X}$ , and the index  $\mathcal{X}$  is omitted where there is no ambiguity; the parameters of  $\mathcal{R}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{Y}$ , and the index  $\mathcal{Y}$  is omitted where there is no ambiguity; and the parameters of  $\mathcal{G} \times \mathcal{R}$  (Definitions 4 to 12, SA.3, SA.4) are taken with  $\mathcal{C} = \mathcal{X} \times \mathcal{Y}$ , and the index  $\mathcal{X} \times \mathcal{Y}$  is omitted where there is no ambiguity. Recall we

also define

$$\begin{aligned} J(\delta) &= \sqrt{2}J(\mathfrak{G}, \mathfrak{M}_{\mathfrak{G}}, \delta/\sqrt{2}) + \sqrt{2}J(\mathfrak{R}, M_{\mathfrak{R}}, \delta/\sqrt{2}), \quad \delta \in (0, 1], \\ \mathbf{N}(\delta) &= \mathbf{N}_{\mathfrak{G}}(\delta/\sqrt{2}, \mathfrak{M}_{\mathfrak{G}})\mathbf{N}_{\mathfrak{R}}(\delta/\sqrt{2}, M_{\mathfrak{R}}), \quad \delta \in (0, 1]. \end{aligned}$$

**Lemma SA.27.** *Suppose Assumption SA.2 holds, a cylindered dyadic expansion  $\mathfrak{C}_{M,N}(\mathbb{P}_Z, 1)$  is given,  $(Z_n^R(g, r) : g \in \mathfrak{G}, r \in \mathfrak{R})$  and  $(\Pi_2 Z_n^R(g, r) : g \in \mathfrak{G}, r \in \mathfrak{R})$  are the Gaussian processes constructed as in Equations (SA-13) and (SA-14) on a possibly enlarged probability space, and  $(\mathfrak{G} \times \mathfrak{R})_\delta$  is chosen in Section SA-III.1.4. Then for all  $t > 0$ ,*

$$\mathbb{P} \left[ \|\Pi_2 R_n - \Pi_2 Z_n^R\|_{(\mathfrak{G} \times \mathfrak{R})_\delta} > C_1 c_{v,\alpha} \sqrt{\frac{N^{2\alpha+1} 2^M \mathbf{E}_{\mathfrak{G}} \mathfrak{M}_{\mathfrak{G}}}{n} t} + C_1 c_{v,\alpha} \sqrt{\frac{\mathbf{C}_{\Pi_2(\mathfrak{G} \times \mathfrak{R})_\delta, M+N}}{n} t} \right] \leq 2\mathbf{N}(\delta) e^{-t},$$

where  $C_1 > 0$  is a universal constant and  $c_{v,\alpha} = v(1 + (2\alpha)^{\alpha/2})$ .

**Proof of Lemma SA.27.** We have shown in the proof of Lemma SA.17 that for any  $(g, r) \in \mathfrak{G} \times \mathfrak{R}$ ,

$$\sum_{j=1}^{M+N} \sum_{0 \leq k < 2^{M+N-j}} |\tilde{\gamma}_{j,k}(g, r)|^2 \leq c_{v,\alpha}^2 N^{2\alpha+1} 2^M \mathbf{E}_{\mathfrak{G}} \mathfrak{M}_{\mathfrak{G}}.$$

It then follows from the relation between  $\gamma_{j,k}$  and  $\eta_{j,k}$  in Equation (SA-17) that for any  $(g, r) \in \mathfrak{G} \times \mathfrak{R}$ ,

$$\sum_{j=1}^{M+N} \sum_{0 \leq k < 2^{M+N-j}} |\tilde{\eta}_{j,k}(g, r)|^2 \leq c_{v,\alpha}^2 N^{2\alpha+1} 2^M \mathbf{E}_{\mathfrak{G}} \mathfrak{M}_{\mathfrak{G}},$$

and hence by Lemma SA.23, for any  $x > 0$ , with probability at least  $1 - 2 \exp(-x)$ ,

$$|\Pi_2 R_n(g, r) - \Pi_2 Z_n(g, r)| \leq c_{v,\alpha} \sqrt{\frac{N^{2\alpha+1} 2^M \mathbf{E}_{\mathfrak{G}} \mathfrak{M}_{\mathfrak{G}}}{n} x} + c_{v,\alpha} \sqrt{\frac{\mathbf{C}_{\Pi_2\{(g,r)\}, M+N}}{n} x}.$$

The conclusion then follows from a union bound on  $(\mathfrak{G} \times \mathfrak{R})_\delta$ .  $\square$

**Lemma SA.28.** *Suppose Assumption SA.2 holds, a cylindered quasi-dyadic expansion  $\mathfrak{C}_{M,N}(\mathbb{P}_Z, \rho)$  is given with  $\rho > 1$ ,  $(Z_n^R(g, r) : g \in \mathfrak{G}, r \in \mathfrak{R})$  and  $(\Pi_2 Z_n^R(g, r) : g \in \mathfrak{G}, r \in \mathfrak{R})$  are the Gaussian processes constructed as in Equations (SA-13) and (SA-14) on a possibly enlarged probability space, and  $(\mathfrak{G} \times \mathfrak{R})_\delta$  is chosen in Section SA-III.1.4. Then for all  $t > 0$ ,*

$$\begin{aligned} \mathbb{P} \left[ \|\Pi_2 R_n - \Pi_2 Z_n^R\|_{(\mathfrak{G} \times \mathfrak{R})_\delta} > C_\rho c_{v,\alpha} \sqrt{\frac{N^{2\alpha+1} 2^M \mathbf{E}_{\mathfrak{G}} \mathfrak{M}_{\mathfrak{G}}}{n} t} + C_\rho c_{v,\alpha} \sqrt{\frac{\mathbf{C}_{\Pi_2(\mathfrak{G} \times \mathfrak{R})_\delta, M+N}}{n} t} \right] \\ \leq 2\mathbf{N}(\delta) e^{-t} + 2^M \exp(-C_\rho n 2^{-M}), \end{aligned}$$

where  $C_\rho > 0$  is a constant that only depends on  $\rho$  and  $c_{v,\alpha} = v(1 + (2\alpha)^{\alpha/2})$ .

*Proof.* Since  $\mathfrak{C}_{M,N}(\mathbb{P}_Z, \rho)$  is a cylindered quasi-dyadic expansion,  $\rho^{-1} 2^{-M-N+j} \leq \mathbb{P}_Z(\mathcal{C}_{j,k}) \leq \rho 2^{-M-N+j}$ , for all  $0 \leq j \leq M+N$ ,  $0 \leq k < 2^{M+N-j}$ . Hence following the argument in the proof for Lemma SA.17, for

any  $g \in \mathcal{G}, r \in \mathcal{R}$ ,

$$\sum_{j=1}^{M+N} \sum_{k=0}^{2^{M+N-j}} \tilde{\eta}_{j,k}^2(g, r) \leq \sum_{j=1}^{M+N} \sum_{k=0}^{2^{M+N-j}} \tilde{\gamma}_{j,k}^2(g, r) \leq c_{v,\alpha}^2 N^{2\alpha+1} 2^M \mathbf{E}_{\mathcal{G}} \mathbf{M}_{\mathcal{G}}.$$

The result then follows from Lemma SA.14.  $\square$

### SA-IV.1.5 Projection Error

To simplify notation, the parameters of  $\mathcal{G}$  and  $\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}$  (Definitions 4 to 12, SA.1, SA.2) are taken with  $\mathcal{C} = \mathcal{X}$ , and the index  $\mathcal{X}$  is omitted where there is no ambiguity; the parameters of  $\mathcal{R}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{Y}$ , and the index  $\mathcal{Y}$  is omitted where there is no ambiguity; and the parameters of  $\mathcal{G} \times \mathcal{R}$  (Definitions 4 to 12, SA.3, SA.4) are taken with  $\mathcal{C} = \mathcal{X} \times \mathcal{Y}$ , and the index  $\mathcal{X} \times \mathcal{Y}$  is omitted where there is no ambiguity. Recall we also define

$$\begin{aligned} J(\delta) &= \sqrt{2}J(\mathcal{G}, \mathbf{M}_{\mathcal{G}}, \delta/\sqrt{2}) + \sqrt{2}J(\mathcal{R}, M_{\mathcal{R}}, \delta/\sqrt{2}), \quad \delta \in (0, 1], \\ \mathbf{N}(\delta) &= \mathbf{N}_{\mathcal{G}}(\delta/\sqrt{2}, \mathbf{M}_{\mathcal{G}}) \mathbf{N}_{\mathcal{R}}(\delta/\sqrt{2}, M_{\mathcal{R}}), \quad \delta \in (0, 1]. \end{aligned}$$

The projection errors for the  $R_n$  and  $Z_n^R$  processes can be decomposed by the observation that, for any  $g \in \mathcal{G}$  and  $r \in \mathcal{R}$ ,

$$\begin{aligned} \Pi_2 R_n(g, r) - R_n(g, r) &= \left( \Pi_1 G_n(g, r) - G_n(g, r) \right) - \left( \Pi_0[\mathbf{p}_X(\mathcal{C}_{M,N})] X_n(g\theta(\cdot, r)) - X_n(g\theta(\cdot, r)) \right), \\ \Pi_2 Z_n^R(g, r) - Z_n^R(g, r) &= \left( \Pi_1 Z_n^G(g, r) - Z_n^G(g, r) \right) - \left( \Pi_0[\mathbf{p}_X(\mathcal{C}_{M,N})] Z_n^X(g\theta(\cdot, r)) - Z_n^X(g\theta(\cdot, r)) \right), \end{aligned} \quad (\text{SA-21})$$

where, in each line, the first term in parentheses is the projection error for the  $G_n$ -process, as discussed in Section SA-III.1.6, and the second term is the projection error for the  $X_n$ -process, detailed in Section SA-II.1.6, with

$$\begin{aligned} X_n(g\theta(\cdot, r)) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[ g(\mathbf{x}_i) \theta(\mathbf{x}_i, r) - \mathbb{E}[g(\mathbf{x}_i) \theta(\mathbf{x}_i, r)] \right], \\ \Pi_0[\mathbf{p}_X(\mathcal{C}_{M,N})] X_n(g\theta(\cdot, r)) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[ \Pi_0[\mathbf{p}_X(\mathcal{C}_{M,N})](g\theta(\cdot, r))(\mathbf{x}_i) - \mathbb{E}[\Pi_0[\mathbf{p}_X(\mathcal{C}_{M,N})](g\theta(\cdot, r))(\mathbf{x}_i)] \right]. \end{aligned}$$

This decomposition allows us to leverage previously established error bounds and convergence results for  $G_n$  and  $X_n$  processes, thus facilitating the analysis of the  $R_n$  and  $Z_n^R$  processes. By utilizing known results from Sections SA-III.1.6 and SA-II.1.6, this approach simplifies the treatment of the projection errors for these new processes.

**Lemma SA.29.** *Suppose Assumption SA.2 holds, a cylindered dyadic expansion  $\mathcal{C}_{M,N}(\mathbb{P}_Z, 1)$  is given,  $(Z_n^R(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  and  $(\Pi_2 Z_n^R(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  are the Gaussian processes constructed as in Equations (SA-19) on a possibly enlarged probability space, and  $(\mathcal{G} \times \mathcal{R})_{\delta}$  is chosen in Section SA-III.1.4. Suppose  $\mathbb{P}_X$  admits a Lebesgue density  $f_X$  supported on  $\mathcal{X} \subseteq \mathbb{R}^d$ . Then for all  $t > N$ , with probability at*

least  $1 - 4N(\delta)ne^{-t}$ ,

$$\begin{aligned} \|R_n - \Pi_2 R_n\|_{(\mathcal{G} \times \mathcal{R})_\delta} &\lesssim \sqrt{V_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}}} t^{\frac{1}{2}} + \sqrt{c_{v,2\alpha}} \sqrt{N^2 V_{\mathcal{G}} + 2^{-N} M_{\mathcal{G}}^2} t^{\alpha + \frac{1}{2}} + c_{v,\alpha} \frac{M_{\mathcal{G}}}{\sqrt{n}} t^{\alpha+1}, \\ \|Z_n^R - \Pi_2 Z_n^R\|_{(\mathcal{G} \times \mathcal{R})_\delta} &\lesssim \sqrt{V_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}}} t^{\frac{1}{2}} + \sqrt{c_{v,2\alpha}} \sqrt{N^2 V_{\mathcal{G}} + 2^{-N} M_{\mathcal{G}}^2} t^{\frac{1}{2}} + c_{v,\alpha} \frac{M_{\mathcal{G}}}{\sqrt{n}} t, \end{aligned}$$

where  $c_{v,\alpha} = v(1 + (2\alpha)^{\frac{\alpha}{2}})$ ,  $c_{v,2\alpha} = v^2(1 + (4\alpha)^\alpha)$ , and

$$\begin{aligned} V_{\mathcal{G}} &= \min\{2M_{\mathcal{G}}, L_{\mathcal{G}} \|\mathcal{V}_M\|_\infty\} \left( \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \right)^2 2^M \mathfrak{m}(\mathcal{V}_M) \|\mathcal{V}_M\|_\infty \text{TV}_{\mathcal{G}}^*, \\ V_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}} &= \min\{2M_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}}, L_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}} \|\mathcal{V}_M\|_\infty\} \left( \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x}) \right)^2 2^M \mathfrak{m}(\mathcal{V}_M) \|\mathcal{V}_M\|_\infty \text{TV}_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}}^*, \end{aligned}$$

with  $\mathcal{V}_M = \cup_{0 \leq l < 2^M} (\mathcal{X}_{0,l} - \mathcal{X}_{0,l})$  the upper level quasi-dyadic variation set as in Section SA-II.1.6.

*Proof.* By Equations (SA-16) and (SA-18), we can show the decomposition in Equation (SA-21) holds. The terms  $\Pi_1 G_n - G_n$  and  $\Pi_1 Z_n^G - Z_n^G$  can be bounded from results in Lemma SA.21. Recall  $\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}} = \{g\theta(\cdot, r) : g \in \mathcal{G}, r \in \mathcal{R}\}$ . We know from Lemma SA.9 for all  $t > 0$ ,

$$\begin{aligned} \mathbb{P} \left( |\Pi_0[\mathbb{P}_X(\mathcal{C}_{M,N})] X_n(g\theta(\cdot, r)) - X_n(g\theta(\cdot, r))| \geq 2\sqrt{V_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}}} t + \frac{4}{3} \cdot \frac{M_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}}}{\sqrt{n}} t \right) &\leq 2 \exp(-t), \\ \mathbb{P} \left( |\Pi_0[\mathbb{P}_X(\mathcal{C}_{M,N})] Z_n^X(g\theta(\cdot, r)) - Z_n^X(g\theta(\cdot, r))| \geq 2\sqrt{V_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}}} t \right) &\leq 2 \exp(-t). \end{aligned}$$

Moreover, suppose  $\alpha > 0$  in (iv) from Assumption SA.2,  $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(|y_i|) | \mathbf{x}_i = \mathbf{x}] \leq 2$ , hence by moment properties of sub-Gaussian random variables,

$$\sup_{r \in \mathcal{R}} \sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[|r(y_i)| | \mathbf{x}_i = \mathbf{x}] \leq v(1 + \sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[|y_i|^\alpha | \mathbf{x}_i = \mathbf{x}]) \leq c_{v,\alpha}.$$

Hence  $M_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}} \leq c_{v,\alpha} M_{\mathcal{G}}$ . Suppose  $\alpha = 0$  from (iv) from Assumption SA.2 holds,  $\sup_{r \in \mathcal{R}} \sup_{\mathbf{x} \in \mathcal{X}} |r(\mathbf{x})| \leq 2v$ , hence we also have  $M_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}} \leq c_{v,\alpha} M_{\mathcal{G}}$ . The result then follows from a union bound over  $(\mathcal{G} \times \mathcal{R})_\delta$ .  $\square$

## SA-IV.2 General Result

The following theorem presents a generalization of Theorem 2 in the paper. To simplify notation, the parameters of  $\mathcal{G}$  and  $\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}$  (Definitions 4 to 12, SA.1, SA.2) are taken with  $\mathcal{C} = \mathcal{Q}_{\mathcal{G}}$ , and the index  $\mathcal{Q}_{\mathcal{G}}$  is omitted where there is no ambiguity; the parameters of  $\mathcal{R}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{Y}$ , and the index  $\mathcal{Y}$  is omitted where there is no ambiguity; and the parameters of  $\mathcal{G} \times \mathcal{R}$  (Definitions 4 to 12, SA.3, SA.4) are taken with  $\mathcal{C} = \mathcal{Q}_{\mathcal{G}} \times \mathcal{Y}$ , and the index  $\mathcal{Q}_{\mathcal{G}} \times \mathcal{Y}$  is omitted where there is no ambiguity.

**Theorem SA.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)  $M_{\mathcal{G}} < \infty$  and  $J(\mathcal{G}, M_{\mathcal{G}}, 1) < \infty$ .

(iv)  $\mathcal{R}$  is a real-valued pointwise measurable class of functions on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_{\mathcal{Y}})$ .

(v)  $J(\mathcal{R}, M_{\mathcal{R}}, 1) < \infty$ , where  $M_{\mathcal{R}}(y) + \mathbf{pTV}_{\mathcal{R}, (-|y|, |y|)} \leq \mathbf{v}(1 + |y|^\alpha)$  for all  $y \in \mathcal{Y}$ , for some  $\mathbf{v} > 0$ , and for some  $\alpha \geq 0$ . Furthermore, if  $\alpha > 0$ , then  $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(|y_i|) | \mathbf{x}_i = \mathbf{x}] \leq 2$ .

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 surely continuous trajectories on  $(\mathcal{G} \times \mathcal{R}, \mathfrak{D}_{\mathbb{P}_{\mathcal{X}}, \mathbb{P}_{\mathcal{Y}}})$  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  $(g_1, r_1), (g_2, r_2) \in \mathcal{G} \times \mathcal{R}$ .
- $\mathbb{P}[\|R_n - Z_n^R\|_{\mathcal{G} \times \mathcal{R}} > C_1 C_{\mathbf{v}, \alpha} \Upsilon_n^R(t)] \leq C_2 e^{-t}$  for all  $t > 0$ ,

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

$$\Upsilon_n^R(t) = \min_{\delta \in (0, 1)} \{A_n^R(t, \delta) + F_n^R(t, \delta)\}$$

with

$$\begin{aligned} A_n^R(t, \delta) &= \sqrt{d} \min \left\{ \left( \frac{\mathbf{c}_1^d \mathbf{E}_{\mathcal{G}} \mathbf{TV}^d M_{\mathcal{G}}^{d+1}}{n} \right)^{\frac{1}{2(d+1)}}, \left( \frac{\mathbf{c}_1^d \mathbf{c}_2^d \mathbf{E}_{\mathcal{G}}^2 M_{\mathcal{G}}^2 \mathbf{TV}^d L^d}{n^2} \right)^{\frac{1}{2(d+2)}} \right\} (t + \log(nN_{\mathcal{G}}(\delta/2)N_{\mathcal{R}}(\delta/2)N_*))^\alpha + 1 \\ &\quad + \frac{M_{\mathcal{G}}}{\sqrt{n}} (\log n)^\alpha (t + \log(nN_{\mathcal{G}}(\delta/2)N_{\mathcal{R}}(\delta/2)N_*))^{\alpha+1}, \\ F_n^R(t, \delta) &= J(\delta)M_{\mathcal{G}} + \frac{\log(n)^{\alpha/2} M_{\mathcal{G}} J^2(\delta)}{\delta^2 \sqrt{n}} + \frac{M_{\mathcal{G}}}{\sqrt{n}} \sqrt{t} + (\log n)^\alpha \frac{M_{\mathcal{G}}}{\sqrt{n}} t^\alpha, \end{aligned}$$

and

$$\begin{aligned} \mathcal{V}_{\mathcal{R}} &= \{\theta(\cdot, r) : r \in \mathcal{R}\}, \\ \mathbf{TV} &= \max\{\mathbf{TV}_{\mathcal{G}}, \mathbf{TV}_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}}\}, \quad \mathbf{L} = \max\{\mathbf{L}_{\mathcal{G}}, \mathbf{L}_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}}\}, \\ M_* &= \left\lceil \log_2 \min \left\{ \left( \frac{n\mathbf{TV}}{\mathbf{E}_{\mathcal{G}}} \right)^{\frac{d}{d+1}}, \left( \frac{n\mathbf{L}\mathbf{TV}}{\mathbf{E}_{\mathcal{G}} M_{\mathcal{G}}} \right)^{\frac{d}{d+2}} \right\} \right\rceil, \\ N_* &= \left\lceil \log_2 \max \left\{ \left( \frac{nM_{\mathcal{G}}^{d+1}}{\mathbf{E}_{\mathcal{G}} \mathbf{TV}^d} \right)^{\frac{1}{d+1}}, \left( \frac{n^2 M_{\mathcal{G}}^{2d+2}}{\mathbf{TV}^d L^d \mathbf{E}_{\mathcal{G}}^2} \right)^{\frac{1}{d+2}} \right\} \right\rceil. \end{aligned}$$

**Proof of Theorem SA.2.** To simplify notation, we will use  $\mathbb{E}[\cdot | \mathcal{X}_{0,l}]$  in short for  $\mathbb{E}[\cdot | \mathbf{x}_i \in \mathcal{X}_{0,l}]$ , and  $\mathbb{E}[\cdot | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}]$  in short for  $\mathbb{E}[\cdot | (\mathbf{x}_i, y_i) \in \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}]$  in this proof.

First, we make a reduction via the surrogate measure and normalizing transformation. Since  $\text{Supp}(\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}) \subseteq \text{Supp}(\mathcal{G})$ , we know  $\mathcal{Q}_{\mathcal{G}}$  is also a surrogate measure for  $\mathbb{P}_{\mathcal{X}}$  with respect to  $\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}$ , and  $\phi_{\mathcal{G}}$  remains a valid normalizing transformation. By the same argument as in the proof for Theorem 1, assumption (ii) implies that on a possibly enriched probability space, there exists  $(\mathbf{u}_i : 1 \leq i \leq n)$  i.i.d distributed with law  $\mathbb{P}_U = \text{Uniform}([0, 1]^d)$ , and

$$g(\mathbf{x}_i) = g(\phi_{\mathcal{G}}^{-1}(\mathbf{u}_i)), \quad \forall g \in \mathcal{G}, 1 \leq i \leq n,$$

and if  $g(\mathbf{x}_i) \neq 0$  for any  $g \in \mathcal{G}$ , then  $\mathbf{x}_i = \phi_{\mathcal{G}}^{-1}(\mathbf{u}_i)$ ,  $1 \leq i \leq n$ .

Define  $\tilde{R}_n$  to be the empirical process based on  $((\mathbf{u}_i, y_i) : 1 \leq i \leq n)$ , and

$$\tilde{R}_n(f, s) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[ f(\mathbf{u}_i) s(y_i) - \mathbb{E}[f(\mathbf{u}_i) s(y_i) | \mathbf{u}_i] \right],$$

and take  $\tilde{\mathcal{G}} = \{g \circ \phi_{\mathcal{X}}^{-1} : g \in \mathcal{G}\}$ , then

$$R_n(g, r) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[ g(\mathbf{x}_i) r(y_i) - \mathbb{E}[g(\mathbf{x}_i) r(y_i) | \mathbf{x}_i] \right] = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[ \tilde{g}(\mathbf{u}_i) r(y_i) - \mathbb{E}[\tilde{g}(\mathbf{u}_i) r(y_i) | \mathbf{u}_i] \right] = \tilde{R}_n(\tilde{g}, r).$$

The relation between constants for  $\tilde{\mathcal{R}}$  and constants for  $\mathcal{R}$  can be deduced from Lemma SA.10. Hence, without loss of generality, we assume  $(\mathbf{x}_i : 1 \leq i \leq n)$  are i.i.d under common law  $\mathbb{P}_X = \text{Uniform}([0, 1]^d)$  distributed and  $\mathcal{X} = [0, 1]^d$ .

Take  $\mathcal{A}_{M,N}(\mathbb{P}_Z, 1)$  to be the axis-aligned cylindered quasi-dyadic expansion of  $\mathbb{R}^{d+1}$ . By Lemma SA.27 and Lemma SA.29, for all  $t > N$ ,

$$\begin{aligned} \mathbb{P} \left[ \|\Pi_2 R_n - \Pi_2 Z_n^R\|_{(\mathcal{G} \times \mathcal{R})_\delta} > C_{v,\alpha} \sqrt{\frac{N^{2\alpha+1} 2^M \mathbf{E}_\mathcal{G} \mathbf{M}_\mathcal{G}}{n}} t + C_{v,\alpha} \sqrt{\frac{\mathbf{C}_{\Pi_2(\mathcal{G} \times \mathcal{R}), M+N}}{n}} t \right] &\leq 2N(\delta) e^{-t}, \\ \mathbb{P} \left[ \|R_n - \Pi_2 R_n\|_{(\mathcal{G} \times \mathcal{R})_\delta} > C_{v,\alpha} \sqrt{2N^2 \mathbf{V} + 2^{-N} \mathbf{M}_\mathcal{G}^2 t^{\alpha+\frac{1}{2}}} + C_{v,\alpha} \frac{\mathbf{M}_\mathcal{G}}{\sqrt{n}} t^{\alpha+1} \right] &\leq 4N(\delta) n e^{-t}, \\ \mathbb{P} \left[ \|Z_n^R - \Pi_2 Z_n^R\|_{(\mathcal{G} \times \mathcal{R})_\delta} > C_{v,\alpha} \sqrt{2N^2 \mathbf{V} + 2^{-N} \mathbf{M}_\mathcal{G}^2 t^{\frac{1}{2}}} + C_{v,\alpha} \frac{\mathbf{M}_\mathcal{G}}{\sqrt{n}} t \right] &\leq 4N(\delta) n e^{-t}, \end{aligned}$$

where  $\mathbf{V} = \sqrt{d} \min \{2\mathbf{M}_\mathcal{G}, \mathbf{L} 2^{-\lfloor M/d \rfloor}\} 2^{-\lfloor M/d \rfloor} \mathbf{T}\mathbf{V}$ , and

$$\mathbf{C}_{\Pi_2(\mathcal{G} \times \mathcal{R})} = \sup_{f \in \Pi_2(\mathcal{G} \times \mathcal{R})} \min \left\{ \sup_{(j,k)} \left[ \sum_{j' < j} (j-j')(j-j'+1) 2^{j'-j} \sum_{k': \mathcal{C}_{j',k'} \subseteq \mathcal{C}_{j,k}} \tilde{\beta}_{j',k'}^2(f) \right], \|f\|_\infty^2 (M+N) \right\}.$$

Let  $f \in \Pi_2(\mathcal{G} \times \mathcal{R})$ . Then there exists  $g \in \mathcal{G}$  and  $r \in \mathcal{R}$  such that  $f = \Pi_2[g, r]$ . Since  $f$  is already piecewise-constant, by definition of  $\beta_{j,k}$ 's and  $\eta_{j,k}$ 's, we know  $\tilde{\beta}_{l,m}(f) = \tilde{\eta}_{l,m}(g, r)$ . Fix  $(j, k)$ . We consider two cases. **Case 1:**  $j > N$ . Then by the design of cell expansions (Section SA-III.1.1),  $\mathcal{C}_{j,k} = \mathcal{X}_{j-N,k} \times \mathcal{Y}_{*,N,0}$ . By definition of  $\eta_{l,m}$ , for any  $N \leq j' \leq j$ , we have  $(j-j')(j-j'+1) 2^{j'-j} \sum_{k': \mathcal{C}_{j',k'} \subseteq \mathcal{C}_{j,k}} \tilde{\eta}_{j',k'}^2(g, r) = 0$ . Now consider  $0 \leq j' < N$ . Then

$$\begin{aligned} &\sum_{k': \mathcal{C}_{j',k'} \subseteq \mathcal{C}_{j,k}} |\tilde{\eta}_{j',k'}(g, r)| \\ &= \sum_{l: \mathcal{X}_{0,l} \subseteq \mathcal{X}_{j-N,k}} \sum_{0 \leq m < 2^{j'}} |\mathbb{E}[g(\mathbf{x}_i) | \mathcal{X}_{0,l}] \cdot |\mathbb{E}[r(y_i) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j-1,2m}] - \mathbb{E}[r(y_i) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j-1,2m+1}]| \\ &\leq C_{v,\alpha} \sum_{l: \mathcal{X}_{0,l} \subseteq \mathcal{X}_{j-N,k}} |\mathbb{E}[g(\mathbf{x}_i) | \mathcal{X}_{0,l}]| N^\alpha \leq C_{v,\alpha} 2^{j-N} \mathbf{M}_\mathcal{G} N^\alpha. \end{aligned}$$

It follows that

$$\sum_{j' < j} (j-j')(j-j'+1) 2^{j'-j} \sum_{k': \mathcal{C}_{j',k'} \subseteq \mathcal{C}_{j,k}} |\tilde{\eta}_{j',k'}(g, r)| \leq \sum_{j' < j} (j-j')(j-j'+1) 2^{j'-N} C_{v,\alpha} \mathbf{M}_\mathcal{G} N^\alpha \lesssim C_{v,\alpha} \mathbf{M}_\mathcal{G} N^\alpha.$$

**Case 2:**  $j \leq N$ . Then  $\mathcal{C}_{j,k} = \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j,m}$ . Hence for any  $0 \leq j' \leq j$ , we have

$$\begin{aligned} \sum_{k': \mathcal{C}_{j',k'} \subseteq \mathcal{C}_{j,k}} |\tilde{\eta}_{j',k'}(g, r)| &= |\mathbb{E}[g(\mathbf{x}_i) | \mathcal{X}_{0,l}]| \sum_{m': \mathcal{Y}_{l,j',m'} \subseteq \mathcal{Y}_{l,j,m}} |\mathbb{E}[r(y_i) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j-1,2m}] - \mathbb{E}[r(y_i) | \mathcal{X}_{0,l} \times \mathcal{Y}_{l,j-1,2m+1}]| \\ &\lesssim C_{v,\alpha} |\mathbb{E}[g(\mathbf{x}_i) | \mathcal{X}_{0,l}]| N^\alpha \lesssim C_{v,\alpha} M_{\mathcal{G}} N^\alpha. \end{aligned}$$

It follows that

$$\sum_{j' < j} (j - j')(j - j' + 1) 2^{j'-j} \sum_{k': \mathcal{C}_{j',k'} \subseteq \mathcal{C}_{j,k}} |\tilde{\eta}_{j',k'}(g, r)| \lesssim C_{v,\alpha} M_{\mathcal{G}} N^\alpha.$$

Moreover, for all  $(j, k)$ , we have  $\tilde{\beta}_{j,k}(g, r) \lesssim C_{v,\alpha} M_{\mathcal{G}} N^\alpha$ . Hence  $\mathbf{c}_{\Pi_2(\mathcal{G} \times \mathcal{R})} \lesssim (C_{v,\alpha} M_{\mathcal{G}} N^\alpha)^2$ . The rest of the proofs follow from choosing optimal  $M, N$  and Lemma SA.16 in the same way as in the proof for Theorem SA.1.  $\square$

### SA-IV.3 Proof of Theorem 2

The proof follows by Theorem SA.2 with  $\delta = n^{-1/2}$ , and

$$\mathbb{N}(n^{-1/2}) = \mathbb{N}_{\mathcal{G}}(1/\sqrt{2n}, M_{\mathcal{G}}) \mathbb{N}_{\mathcal{R}}(1/\sqrt{2n}, M_{\mathcal{R}}) \leq c_{\mathcal{G}} c_{\mathcal{R}} (2\sqrt{n})^{\mathbf{d}_{\mathcal{G}} + \mathbf{d}_{\mathcal{R}}} = c(2\sqrt{n})^{\mathbf{d}},$$

and

$$\begin{aligned} J(n^{-1/2}) &= \sqrt{2} J(\mathcal{G}, M_{\mathcal{G}}, 1/\sqrt{2n}) + \sqrt{2} J(\mathcal{R}, M_{\mathcal{R}}, 1/\sqrt{2n}) \\ &\leq 3n^{-1/2} \sqrt{\mathbf{d}_{\mathcal{G}} \log(c_{\mathcal{G}} \sqrt{n})} + 3\delta \sqrt{\mathbf{d}_{\mathcal{R}} \log(c_{\mathcal{R}} \sqrt{n})} \\ &\leq 3\delta \sqrt{(\mathbf{d}_{\mathcal{G}} + \mathbf{d}_{\mathcal{R}}) \log(c_{\mathcal{G}} c_{\mathcal{R}} n)} \leq 3\delta \sqrt{\mathbf{d} \log(cn)}. \end{aligned}$$

This completes the proof.  $\square$

### SA-IV.4 Proof of Corollary 4

Take  $t = C \log n$  with  $C > 1$  in Theorem 2.  $\square$

### SA-IV.5 Example: Local Polynomial Estimators

The following lemma provides sufficient conditions for the rate of *non-linearity error* and *smoothing bias* claimed in Section 4.1.

**Lemma SA.30.** *Consider the setup of Section 4.1. Recall we assume that  $((\mathbf{x}_i, y_i) : 1 \leq i \leq n)$  are i.i.d random vectors taking values in  $(\mathbb{R}^{d+1}, \mathcal{B}(\mathbb{R}^{d+1}))$ , with  $\mathbf{x}_i \sim \mathbb{P}_X$  admitting a continuous Lebesgue density  $f_X$  on its support  $\mathcal{X} = [0, 1]^d$ . Assume in addition that  $\mathbf{w} \mapsto \theta(\mathbf{w}; r)$  is  $(\mathbf{p} + 1)$ -times continuously differentiable with  $(\mathbf{p} + 1)$ th partial derivatives bounded uniformly over  $\mathbf{w} \in \mathcal{W} \subseteq \mathcal{X}$  and  $r \in \mathcal{R}_l$ ,  $l = 1, 2$ , for some  $\mathbf{p} \geq 0$ .*

If  $(nb^d)^{-1} \log n \rightarrow 0$ , then

$$\begin{aligned} \sup_{\mathbf{w} \in \mathcal{W}, r \in \mathcal{R}_2} |\mathbf{e}_1^\top (\widehat{\mathbf{H}}_{\mathbf{w}}^{-1} - \mathbf{H}_{\mathbf{w}}^{-1}) \mathbf{S}_{\mathbf{w}, r}| &= O((nb^d)^{-1} \log n) \quad a.s., \quad \text{and} \\ \sup_{\mathbf{w} \in \mathcal{W}, r \in \mathcal{R}_l} |\mathbb{E}[\widehat{\theta}(\mathbf{w}, r) | \mathbf{x}_1, \dots, \mathbf{x}_n] - \theta(\mathbf{w}, r)| &= O(b^{1+p}) \quad a.s., \quad l = 1, 2. \end{aligned}$$

If, in addition,  $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(|y_i|) | \mathbf{x}_i = \mathbf{x}] \leq 2$ , then

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

**Proof of Lemma SA.30.** We concisely flash out the arguments that are standard from the empirical process literature.

**Convergence rate for each entry of  $\widehat{\mathbf{H}}_{\mathbf{w}} - \mathbf{H}_{\mathbf{w}}$ :** Consider  $\mathbf{u}_1^\top (\widehat{\mathbf{H}}_{\mathbf{w}} - \mathbf{H}_{\mathbf{w}}) \mathbf{u}_2$ , where  $\mathbf{u}_1, \mathbf{u}_2$  are multi-indices such that  $|\mathbf{u}_1|, |\mathbf{u}_2| \leq p$ . Take  $\mathbf{v} = \mathbf{u}_1 + \mathbf{u}_2$ . Define

$$g_n(\xi, \mathbf{w}) = \left( \frac{\xi - \mathbf{w}}{h} \right)^\mathbf{v} \frac{1}{h^d} K \left( \frac{\xi - \mathbf{w}}{h} \right), \quad \xi \in \mathcal{X}, \mathbf{w} \in \mathcal{W}.$$

Define  $\mathcal{F} = \{g_n(\cdot, \mathbf{w}) : \mathbf{w} \in \mathcal{W}\}$ . Then  $\sup_{\mathbf{w} \in \mathcal{W}} |\mathbf{u}_1^\top (\widehat{\mathbf{H}}_{\mathbf{w}} - \mathbf{H}_{\mathbf{w}}) \mathbf{u}_2| = \sup_{f \in \mathcal{F}} |\mathbb{E}_n[f(\mathbf{x}_i)] - \mathbb{E}[f(\mathbf{x}_i)]|$ . By standard arguments from kernel regression literature, we can show  $\mathcal{F}$  forms a VC-type class over  $\mathcal{X}$  with exponent  $d$  and constant  $\|\mathcal{X}\|_\infty/b$ , with  $\mathbf{M}_{\mathcal{F}, \mathcal{X}} = O(b^{-d})$ ,  $\sigma_n^2 = \sup_{f \in \mathcal{F}} \mathbb{V}[f(\mathbf{x}_i)] = O(b^{-d/2})$ . By Corollary 5.1 in Chernozhukov *et al.* (2014), we can show  $\mathbb{E}[\sup_{f \in \mathcal{F}} |\mathbb{E}_n[f(\mathbf{x}_i)] - \mathbb{E}[f(\mathbf{x}_i)]|] = O((nb^d)^{-1/2} \sqrt{\log n} + (nb^d)^{-1} \log n)$ . Since  $\mathcal{F}$  is separable, we can use Talagrand's inequality (Giné and Nickl, 2016, Theorem 3.3.9) to get for all  $t > 0$ ,

$$\mathbb{P} \left( \sup_{f \in \mathcal{F}} |\mathbb{E}_n[f(\mathbf{x}_i)] - \mathbb{E}[f(\mathbf{x}_i)]| \geq C_1 (nb^d)^{-1/2} \sqrt{t + \log n} + C_1 (nb^d)^{-1} (t + \log n) \right) \leq \exp(-t),$$

where  $C_1$  is a constant not depending on  $n$ . This shows for any multi-indices  $\mathbf{u}_1, \mathbf{u}_2$  with  $|\mathbf{u}_1|, |\mathbf{u}_2| \leq p$ ,

$$\sup_{\mathbf{w} \in \mathcal{W}} |\mathbf{u}_1^\top (\widehat{\mathbf{H}}_{\mathbf{w}} - \mathbf{H}_{\mathbf{w}}) \mathbf{u}_2| = O((nb^d)^{-1/2} \sqrt{\log n} + (nb^d)^{-1} \log n), \quad a.s.$$

**Convergence rate for  $\sup_{\mathbf{w} \in \mathcal{W}} \|\widehat{\mathbf{H}}_{\mathbf{w}}^{-1} - \mathbf{H}_{\mathbf{w}}^{-1}\|$ :** Since  $\mathbf{H}_{\mathbf{w}}$  and  $\widehat{\mathbf{H}}_{\mathbf{w}}$  are finite-dimensional,  $\sup_{\mathbf{w} \in \mathcal{W}} \|\widehat{\mathbf{H}}_{\mathbf{w}} - \mathbf{H}_{\mathbf{w}}\| = O((nb^d)^{-1/2} \sqrt{\log n} + (nb^d)^{-1} \log n)$  a.s.. By Weyl's Theorem,  $\sup_{\mathbf{w} \in \mathcal{W}} |\sigma_d(\widehat{\mathbf{H}}_{\mathbf{w}}) - \sigma_d(\mathbf{H}_{\mathbf{w}})| = O((nb^d)^{-1/2} \sqrt{\log n} + (nb^d)^{-1} \log n)$  a.s., which also implies  $\inf_{\mathbf{w} \in \mathcal{W}} \sigma_d(\widehat{\mathbf{H}}_{\mathbf{w}}) = \Omega(1)$  a.s.. Hence

$$\sup_{\mathbf{w} \in \mathcal{W}} \|\widehat{\mathbf{H}}_{\mathbf{w}}^{-1} - \mathbf{H}_{\mathbf{w}}^{-1}\| \leq \sup_{\mathbf{w} \in \mathcal{W}} \|\widehat{\mathbf{H}}_{\mathbf{w}}^{-1}\| \|\widehat{\mathbf{H}}_{\mathbf{w}} - \mathbf{H}_{\mathbf{w}}\| \|\mathbf{H}_{\mathbf{w}}^{-1}\| = O((nb^d)^{-1/2} \sqrt{\log n}), \quad a.s..$$

**Convergence rate for  $\sup_{\mathbf{w} \in \mathcal{W}} \sup_{r \in \mathcal{R}_\ell} \|\mathbf{S}_{\mathbf{w}, r}\|$ ,  $\ell = 1, 2$ :** Consider  $\mathbf{v}^\top \mathbf{S}_{\mathbf{w}, r}$  where  $|\mathbf{v}| \leq p$ . Define  $\mathcal{H}_\ell = \{(\mathbf{z}, y) \mapsto g_n(\mathbf{z}, \mathbf{w})(r(y) - \theta(\mathbf{z}, r)) : \mathbf{w} \in \mathcal{W}, r \in \mathcal{R}_\ell\}$ ,  $\ell = 1, 2$ . It is not hard to check both  $\mathcal{H}_1$  and  $\mathcal{H}_2$  are VC-type classes over  $\mathcal{X}$ . By similar arguments as in  $\widehat{\mathbf{H}}_{\mathbf{w}} - \mathbf{H}_{\mathbf{w}}$ , for all  $t > 0$ ,

$$\mathbb{P} \left( \sup_{h \in \mathcal{H}_2} |\mathbb{E}_n[h(\mathbf{x}_i, y_i)] - \mathbb{E}[h(\mathbf{x}_i, y_i)]| \geq C_2 (nb^d)^{-1/2} \sqrt{t + \log n} + C_2 (nb^d)^{-1} (t + \log n) \right) \leq \exp(-t),$$

where  $C_2$  is a constant that does not depend on  $n$ . And if we further assume  $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(|y_i|)|\mathbf{x}_i = \mathbf{x}] \leq 2$ , then for all  $t > 0$ ,

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}_1} |\mathbb{E}_n[h(\mathbf{x}_i, y_i)] - \mathbb{E}[h(\mathbf{x}_i, y_i)]| \geq C_2(nb^d)^{-1/2}\sqrt{t + \log n} + C_2(nb^d)^{-1}(\log n)(t + \log n)\right) \leq \exp(-t).$$

Together with finite dimensionality of the vector  $\mathbf{S}_{\mathbf{w}, r}$ ,

$$\begin{aligned} \sup_{\mathbf{w} \in \mathcal{W}} \sup_{r \in \mathcal{R}_1} \|\mathbf{S}_{\mathbf{w}, r}\| &= O((nb^d)^{-1/2}\sqrt{\log n} + (nb^d)^{-1}(\log n)^2), \quad a.s., \\ \sup_{\mathbf{w} \in \mathcal{W}} \sup_{r \in \mathcal{R}_2} \|\mathbf{S}_{\mathbf{w}, r}\| &= O((nb^d)^{-1/2}\sqrt{\log n}), \quad a.s. \end{aligned}$$

**Putting together for Non-Linearity Errors:**

$$\begin{aligned} \sup_{\mathbf{w} \in \mathcal{W}} \sup_{r \in \mathcal{R}_1} |\mathbf{e}_1^\top (\widehat{\mathbf{H}}_{\mathbf{w}}^{-1} - \mathbf{H}_{\mathbf{w}}^{-1}) \mathbf{S}_{\mathbf{w}, r}| &= O((nb^d)^{-1} \log n + (nb^d)^{-3/2}(\log n)^{5/2}), \quad a.s., \\ \sup_{\mathbf{w} \in \mathcal{W}} \sup_{r \in \mathcal{R}_2} |\mathbf{e}_1^\top (\widehat{\mathbf{H}}_{\mathbf{w}}^{-1} - \mathbf{H}_{\mathbf{w}}^{-1}) \mathbf{S}_{\mathbf{w}, r}| &= O((nb^d)^{-1} \log n), \quad a.s.. \end{aligned}$$

**Smoothing Error:** Take  $\mathbf{R}_{\mathbf{w}, r} = \mathbb{E}_n [\mathbf{r}_p(\frac{\mathbf{X}_i - \mathbf{w}}{h}) K_h(\mathbf{X}_i - \mathbf{w}) \mathbf{r}_{\mathbf{w}}(\mathbf{X}_i; r)]$  where

$$\mathbf{r}_{\mathbf{w}}(\xi; r) = \theta(\xi; r) - \sum_{0 \leq |\nu| \leq p} \frac{\partial_\nu \theta(\mathbf{w}; r)}{\nu!} (\xi - \mathbf{w})^\nu.$$

Since all  $\theta(\cdot; r), r \in \mathcal{R}_\ell$  are  $(p+1)$ -times continuously differentiable with derivatives bounded uniformly over  $\mathcal{X}$  and  $\mathcal{R}_\ell$ , we have almost surely  $\sup_{r \in \mathcal{R}_\ell} \sup_{\mathbf{w} \in \mathcal{W}} |\mathbf{R}_{\mathbf{w}, r}| = O(b^{p+1})$ ,  $\ell = 1, 2$ . We have proved that  $\inf_{\mathbf{w} \in \mathcal{W}} \sigma_d(\widehat{\mathbf{H}}_{\mathbf{w}}) = \Omega(1)$  a.s.. Hence

$$\sup_{r \in \mathcal{R}_\ell} \sup_{\mathbf{w} \in \mathcal{W}} |\mathbb{E}[\widehat{\theta}(\mathbf{w}, r)|\mathbf{x}_1, \dots, \mathbf{x}_n] - \theta(\mathbf{w}, r)| = \sup_{r \in \mathcal{R}_\ell} \sup_{\mathbf{w} \in \mathcal{W}} |\mathbf{e}_1^\top \widehat{\mathbf{H}}_{\mathbf{w}}^{-1} \mathbf{R}_{\mathbf{w}, r}| = O(b^{p+1}), \quad a.s., \text{ for } \ell = 1, 2.$$

This completes the proof.  $\square$

The following two examples provide the omitted details concerning uniform Gaussian strong approximation rates obtained via other methods, which are discussed in Section 4.1 of the paper.

**Example SA.1** (Strong Approximation via [Rio \(1994\)](#)). Consider the setup of Section 4.1, and assume the following regularity conditions hold:

- (a)  $(\mathbf{x}_i, y_i) = (\mathbf{x}_i, \varphi(\mathbf{x}_i, u_i))$ , where the law of  $\mathbf{b}_i = (\mathbf{x}_i, u_i)$ ,  $\mathbb{P}_B$ , has continuous and positive Lebesgue density  $f_B$  on its support  $\mathcal{B} = [0, 1]^{d+1}$ .
- (b)  $\mathbf{M}_{\{\varphi\}, \mathcal{B}} = O(1)$ ,  $\mathbf{K}_{\{\varphi\}, \mathcal{B}} = O(1)$ , and  $\sup_{g \in \mathcal{G}} \mathbf{TV}_{\{\varphi\}, \text{Supp}(g) \times [0, 1]} = O(\sup_{g \in \mathcal{G}} \mathbf{m}(\text{Supp}(g)))$ .
- (c)  $\sup_{g \in \mathcal{G}} \mathbf{TV}_{\mathcal{V}_{\mathcal{R}_l}, \text{Supp}(g)} = O(\sup_{g \in \mathcal{G}} \mathbf{m}(\text{Supp}(g)))$  and  $\mathbf{K}_{\mathcal{V}_{\mathcal{R}_l}, \mathcal{B}} = O(1)$ , for  $l = 1, 2$ .

Recall  $\mathcal{G} = \{b^{-d/2} \mathfrak{K}_{\mathbf{w}}(\frac{\cdot - \mathbf{w}}{b}) : \mathbf{w} \in \mathcal{W}\}$  with  $\mathfrak{K}_{\mathbf{w}}(\mathbf{u}) = \mathbf{e}_1^\top \mathbf{H}_{\mathbf{w}}^{-1} \mathbf{p}(\mathbf{u}) K(\mathbf{u})$ . For  $\mathcal{R}_1$ , take  $\mathcal{H}_1 = \{h \circ T_{\mathbb{P}_B}^{-1} : h \in \mathcal{H}_1^o\}$ , where  $\mathcal{H}_1^o = \{(\mathbf{x}, u) \in \mathcal{B} \mapsto g(\mathbf{x})\varphi(\mathbf{x}, u) - g(\mathbf{x})\theta(\mathbf{x}, \text{Id}) : g \in \mathcal{G}\}$ ,  $T_{\mathbb{P}_B}$  is the Rosenblatt transformation

based on  $\mathbb{P}_B$  given in Section 3.1. Recall we denote  $\mathcal{X} = [0, 1]^d$ . Then,

$$\begin{aligned}
M_{\mathcal{H}_1, \mathcal{B}} &\leq M_{\mathcal{G}, \mathcal{X}} M_{\{\varphi\}, \mathcal{B}} = O(b^{-d/2}), \\
\text{TV}_{\mathcal{H}_1, \mathcal{B}} &\leq \frac{\bar{f}_B^2}{\underline{f}_B} (\text{TV}_{\mathcal{G}, \mathcal{X}} + M_{\mathcal{G}, \mathcal{X}} \sup_{g \in \mathcal{G}} \mathbf{m}(\text{Supp}(g))) = O(b^{d/2-1}), \\
K_{\mathcal{H}_1, \mathcal{B}} &\leq (2\sqrt{d})^{d-1} \frac{\bar{f}_B^{d+1}}{\underline{f}_B^d} (M_{\{\varphi\}, \mathcal{B}} K_{\mathcal{G}, \mathcal{X}} + M_{\mathcal{G}, \mathcal{X}} K_{\{\varphi\}, \mathcal{B}} + M_{\mathcal{G}, \mathcal{X}} K_{\mathcal{V}_{\mathcal{R}_1}, \mathcal{X}}) = O(b^{-d/2}), \\
N_{\mathcal{H}_1, \mathcal{B}}(\delta, M_{\mathcal{H}_1, \mathcal{B}}) &= O(\delta^{-d-1}), \quad 0 < \delta < 1,
\end{aligned} \tag{SA-22}$$

where  $\bar{f}_B = \sup_{\mathbf{x} \in \mathcal{B}} f_B(\mathbf{x})$  and  $\underline{f}_B = \inf_{\mathbf{x} \in \mathcal{B}} f_B(\mathbf{x})$ . Rio (1994, Theorem 1.1) implies that  $(X_n(h) : h \in \mathcal{H}_1) = (\sqrt{nb^d} \mathbf{e}_1^\top \mathbf{H}_{\mathbf{w}}^{-1} \mathbf{S}_{\mathbf{w}, r} : \mathbf{w} \in \mathcal{W}, r \in \mathcal{R}_1)$  admits a uniform Gaussian strong approximation with rate

$$S_n(t) = C_{d, \varphi, \mathbb{P}_B} (nb^{d+1})^{-1/(2d+2)} \sqrt{t + d \log n} + C_{d, \varphi, \mathbb{P}_B} (nb^d)^{-1/2} (t + d \log n),$$

where  $C_{d, \varphi, \mathbb{P}_B}$  is a quantity that only depends on  $d$ ,  $\varphi$  and  $\mathbb{P}_B$ .

For  $\mathcal{R}_2$ , take  $\mathcal{H}_2 = \{h \circ T_{\mathbb{P}_B}^{-1} : h \in \mathcal{H}_2^o\}$ , where  $\mathcal{H}_2^o = \{(\mathbf{x}, u) \in \mathcal{B} \mapsto g(\mathbf{x})r \circ \varphi(\mathbf{x}, u) - g(\mathbf{x})\theta(\mathbf{x}, r) : g \in \mathcal{G}, r \in \mathcal{R}_2\}$ . Then

$$\begin{aligned}
M_{\mathcal{H}_2} &= M_{\mathcal{G}, \mathcal{X}} M_{\{\varphi\}, \mathcal{B}} = O(b^{-d/2}), \\
\text{TV}_{\mathcal{H}_2} &\leq \frac{\bar{f}_B^2}{\underline{f}_B} (\text{TV}_{\mathcal{G}, \mathcal{X}} + E_{\mathcal{G}, \mathcal{X}} + M_{\mathcal{G}, \mathcal{X}} \sup_{g \in \mathcal{G}} \mathbf{m}(\text{Supp}(g))) \max\{L_{\{\varphi\}, \mathcal{B}}, 1\}^{d-1} = O(b^{d/2-1}), \\
N_{\mathcal{H}_2, \mathcal{B}}(\delta, M_{\mathcal{H}_2, \mathcal{B}}) &= O(\delta^{-d-1}), \quad 0 < \delta < 1.
\end{aligned}$$

Rio (1994, Theorem 1.1) implies that  $(X_n(h) : h \in \mathcal{H}_2) = (\sqrt{nb^d} \mathbf{e}_1^\top \mathbf{H}_{\mathbf{w}}^{-1} \mathbf{S}_{\mathbf{w}, r} : \mathbf{w} \in \mathcal{W}, r \in \mathcal{R}_2)$  admits a Gaussian strong approximation with rate function

$$S_n(t) = C_{d, \varphi, \mathbb{P}_B} (nb^{d+1})^{-1/(2d+2)} \sqrt{t + d \log n} + C_{d, \varphi, \mathbb{P}_B} \sqrt{\frac{\log n}{nb^d}} (t + d \log n),$$

where  $C_{d, \varphi, \mathbb{P}_B}$  is a quantity that only depends on  $d$ ,  $\varphi$  and  $\mathbb{P}_B$ .

The strong approximation rates stated in Section 4.1 now follow directly from the strong approximation results above.  $\blacktriangle$

**Proof of Example SA.1.** Recall  $\mathcal{G} = \{b^{-d/2} \mathfrak{K}_{\mathbf{w}}(\cdot - \frac{\mathbf{w}}{b}) : \mathbf{w} \in \mathcal{W}\}$  with  $\mathfrak{K}_{\mathbf{w}}(\mathbf{u}) = \mathbf{e}_1^\top \mathbf{H}_{\mathbf{x}}^{-1} \mathbf{p}(\mathbf{u}) K(\mathbf{u})$ .

(1) **Properties of  $\mathcal{G}$**  Since  $\sup_{\mathbf{w} \in \mathcal{W}} \|\mathbf{H}_{\mathbf{w}}^{-1}\| \lesssim 1$  and  $K$  is continuous with compact support, we know

$$M_{\mathcal{G}, \mathcal{X}} = O(b^{-d/2}).$$

By a change of variables, we can show

$$E_{\mathcal{G}, \mathcal{X}} = \sup_{\mathbf{w} \in \mathcal{W}} \mathbb{E} \left[ \left| b^{-d/2} \mathfrak{K}_{\mathbf{w}} \left( \frac{\mathbf{x}_i - \mathbf{w}}{b} \right) \right| \right] = O(b^{d/2}).$$

And  $\sup_{\mathbf{w} \in \mathcal{W}} \sup_{\mathbf{u}, \mathbf{u}'} \frac{|\mathbf{r}_p(\frac{\mathbf{u}-\mathbf{w}}{b}) - \mathbf{r}_p(\frac{\mathbf{u}'-\mathbf{w}}{b})|}{\|\mathbf{u}-\mathbf{u}'\|_\infty} = O(b^{-1})$ , and  $\sup_{\mathbf{w} \in \mathcal{W}} \sup_{\mathbf{u}, \mathbf{u}'} \frac{|K(\frac{\mathbf{u}-\mathbf{w}}{b}) - K(\frac{\mathbf{u}'-\mathbf{w}}{b})|}{\|\mathbf{u}-\mathbf{u}'\|_\infty} = O(b^{-1})$ , hence

$$\mathbf{L}_{\mathcal{G}, \mathcal{X}} = O(b^{-\frac{d}{2}-1}).$$

Notice that the support of functions in  $\mathcal{G}$  has uniformly bounded volume, i.e.  $\sup_{g \in \mathcal{G}} \mathbf{m}(\text{Supp}(g)) = O(b^d)$ . Together with the rate for  $\mathbf{L}_{\mathcal{G}, \mathcal{X}}$ , we know

$$\mathbf{TV}_{\mathcal{G}, \mathcal{X}} \leq \mathbf{L}_{\mathcal{G}, \mathcal{X}} \sup_{g \in \mathcal{G}} \mathbf{m}(\text{Supp}(g)) = O(b^{\frac{d}{2}-1}).$$

Now we will show that  $\mathbf{M}_{\mathcal{G}, \mathcal{X}}^{-1} \mathcal{G}$  is a VC-type class. We know  $\sup_{\mathbf{w}, \mathbf{w}' \in \mathcal{W}} \|\mathbf{H}_{\mathbf{w}} - \mathbf{H}_{\mathbf{w}'}\| / \|\mathbf{w} - \mathbf{w}'\|_\infty = O(b^{-1})$ . Since  $\inf_{\mathbf{w} \in \mathcal{W}} \|\mathbf{H}_{\mathbf{w}}\| = \Omega(1)$ , we also have  $\sup_{\mathbf{w}, \mathbf{w}' \in \mathcal{W}} \|\mathbf{H}_{\mathbf{w}}^{-1} - \mathbf{H}_{\mathbf{w}'}^{-1}\| / \|\mathbf{w} - \mathbf{w}'\|_\infty = O(b^{-1})$ . It follows that

$$\mathbf{L}_{\mathcal{G}, \mathcal{X}} = \sup_{\mathbf{w} \in \mathcal{W}} \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} \left| b^{-d/2} \mathfrak{K}_{\mathbf{w}} \left( \frac{\mathbf{x} - \mathbf{w}}{b} \right) - b^{-d/2} \mathfrak{K}_{\mathbf{w}} \left( \frac{\mathbf{x}' - \mathbf{w}}{b} \right) \right| / \|\mathbf{x} - \mathbf{x}'\|_\infty = O(b^{-\frac{d}{2}-1}).$$

To upper bound  $\mathbf{K}_{\mathcal{G}, \mathcal{X}}$ , let  $\mathcal{D} \subseteq \mathcal{X}$  be a cube with edges of length  $\mathbf{a}$  parallel to the coordinate axes. Consider the following two cases: (i) if  $\mathbf{a} < b$ , then  $\mathbf{TV}_{\mathcal{G}, \mathcal{D}} \leq C_K b^{-d/2-1} \mathbf{a}^d \leq C_K b^{-d/2} \mathbf{a}^{d-1}$ ; (ii) if  $\mathbf{a} > b$ , then  $\mathbf{TV}_{\mathcal{G}, \mathcal{D}} \leq C_K \sup_{\mathbf{w} \in \mathcal{W}} \mathbf{m}(\text{Supp}(\mathfrak{K}_{\mathbf{w}})) \mathbf{L}_{\mathcal{G}, \mathcal{X}} \leq C_K b^d b^{-d/2-1} \leq C_K b^{-d/2} b^{d-1} \leq C_K b^{-d/2} \mathbf{a}^{d-1}$ . This shows

$$\mathbf{K}_{\mathcal{G}, \mathcal{X}} \leq C_K b^{-d/2}.$$

Consider  $h_{\mathbf{w}}(\cdot) = \sqrt{b^d} \mathbf{e}_1^T \mathbf{H}_{\mathbf{w}}^{-1} \mathbf{r}_p(\cdot) K(\cdot)$ ,  $\mathbf{w} \in \mathcal{W}$ . Then  $b^{-d/2} \mathfrak{K}_{\mathbf{w}}(\frac{\cdot - \mathbf{w}}{b}) = h_{\mathbf{w}}(\frac{\cdot - \mathbf{w}}{b})$ ,  $\mathbf{w} \in \mathcal{W}$ . Recall that  $\mathbf{x}_i$  has common law  $\mathbb{P}_X$  with Lebesgue density  $f_X$ . Then there exists a constant  $\mathbf{c}$  only depending on  $\sup_{\mathbf{x} \in \mathcal{X}} K(\mathbf{x})$ ,  $\mathbf{L}_{\{K\}, \mathcal{X}}$ ,  $\sigma_K = (\int K(\mathbf{x}) d\mathbf{x})^{1/2}$ ,  $\bar{f}_X = \sup_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x})$ ,  $\underline{f}_X = \inf_{\mathbf{x} \in \mathcal{X}} f_X(\mathbf{x})$  that

$$\begin{aligned} \sup_{\mathbf{w} \in \mathcal{W}} \|h_{\mathbf{w}}\|_\infty &\leq \mathbf{c}, \\ \sup_{\mathbf{w} \in \mathcal{W}} \sup_{\mathbf{u}, \mathbf{v} \in \mathcal{W}} \frac{|h_{\mathbf{w}}(\mathbf{u}) - h_{\mathbf{w}}(\mathbf{v})|}{\|\mathbf{u} - \mathbf{v}\|_\infty} &\leq \mathbf{c}, \\ \sup_{\mathbf{w}, \mathbf{w}' \in \mathcal{W}} \sup_{\mathbf{u} \in \mathcal{W}} \frac{|h_{\mathbf{w}}(\mathbf{u}) - h_{\mathbf{w}'}(\mathbf{u})|}{\|\mathbf{w} - \mathbf{w}'\|_\infty} &\leq \mathbf{c}. \end{aligned}$$

We can again apply Lemma 7 from [Cattaneo et al. \(2024\)](#) to show that, for all  $0 < \delta < 1$ ,

$$N_{\mathcal{X}}(\mathbf{M}_{\mathcal{G}, \mathcal{X}}^{-1} \mathcal{G}, \|\cdot\|_{\mathbb{P}_{X,2}}, \delta) \leq \mathbf{c} \delta^{-2d-2} + 1.$$

**(2) Properties of  $\mathcal{H}_1^a$**  Let  $g \in \mathcal{G}$ . Take  $\mathcal{H}_1^a = \{g \cdot \varphi : g \in \mathcal{G}\}$  and  $\mathcal{H}_1^b = \{g \cdot \theta(\cdot, \text{Id}) : g \in \mathcal{G}\}$ . Then

$$\mathbf{M}_{\mathcal{H}_1^a, \mathcal{B}} \leq \mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{M}_{\{\varphi\}, \mathcal{B}}.$$

We have shown that all functions in  $\mathcal{G}$  are Lipschitz and  $\mathbf{L}_{\mathcal{G}, \mathcal{X}} = O(b^{-d/2-1})$ , [Ambrosio et al. \(2000, Proposition 3.2 \(b\)\)](#) then implies

$$\mathbf{TV}_{\mathcal{H}_1^a, \mathcal{B}} \leq \mathbf{M}_{\{\varphi\}, \mathcal{B}} \mathbf{TV}_{\mathcal{G}, \mathcal{X}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \sup_{g \in \mathcal{G}} \mathbf{TV}_{\{\varphi\}, \text{Supp}(g) \times [0,1]}.$$

Let  $\mathcal{C}$  be any cube of side-length  $a$  in  $\mathbb{R}^{d+1}$ . By [Ambrosio et al. \(2000, Proposition 3.2 \(b\)\)](#),

$$\mathrm{TV}_{\mathcal{H}_1^c, \mathcal{C}} \leq \mathbf{M}_{\{\varphi\}, \mathcal{B}} \mathrm{TV}_{\mathcal{G}, \mathcal{C}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \sup_{g \in \mathcal{G}} \mathrm{TV}_{\{\varphi\}, \mathrm{Supp}(g) \times [0, 1] \cap \mathcal{C}} \leq \mathbf{M}_{\{\varphi\}, \mathcal{B}} \mathbf{K}_{\mathcal{G}, \mathcal{X}} a^d + \mathbf{K}_{\{\varphi\}, \mathcal{B}} \mathbf{M}_{\mathcal{G}, \mathcal{X}} a^d,$$

which implies

$$\mathbf{K}_{\mathcal{H}_1^c, \mathcal{B}} \leq \mathbf{M}_{\{\varphi\}, \mathcal{B}} \mathbf{K}_{\mathcal{G}, \mathcal{X}} + \mathbf{K}_{\{\varphi\}, \mathcal{B}} \mathbf{M}_{\mathcal{G}, \mathcal{X}}.$$

Similar argument shows

$$\begin{aligned} \mathbf{M}_{\mathcal{H}_1^b, \mathcal{X}} &\leq \mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{M}_{\{\varphi\}, \mathcal{B}}, & \mathrm{TV}_{\mathcal{H}_1^b, \mathcal{X}} &\leq \mathrm{TV}_{\mathcal{G}, \mathcal{X}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \sup_{g \in \mathcal{G}} \mathrm{TV}_{\{\theta(\cdot, \mathrm{Id})\}, \mathrm{Supp}(g)}, \\ \mathbf{K}_{\mathcal{H}_1^b, \mathcal{X}} &\leq \mathbf{M}_{\{\varphi\}, \mathcal{B}} \mathbf{K}_{\mathcal{G}, \mathcal{X}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{K}_{\{\theta(\cdot, \mathrm{Id})\}, \mathcal{X}}. \end{aligned}$$

Then by assumptions  $\sup_{g \in \mathcal{G}} \mathrm{TV}_{\{\varphi\}, \mathrm{Supp}(g) \times [0, 1]} = O(\sup_{g \in \mathcal{G}} \mathbf{m}(\mathrm{Supp}(g)))$  and  $\sup_{g \in \mathcal{G}} \mathrm{TV}_{\{\theta(\cdot, \mathrm{Id})\}, \mathrm{Supp}(g)} = O(\sup_{g \in \mathcal{G}} \mathbf{m}(\mathrm{Supp}(g)))$ , we have

$$\begin{aligned} \mathbf{M}_{\mathcal{H}_1^c} &\leq \mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{M}_{\{\varphi\}, \mathcal{B}}, & \mathrm{TV}_{\mathcal{H}_1^c} &= O(\mathrm{TV}_{\mathcal{G}, \mathcal{X}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \sup_{g \in \mathcal{G}} \mathbf{m}(\mathrm{Supp}(g))), \\ \mathbf{K}_{\mathcal{H}_1^c} &\leq \mathbf{M}_{\{\varphi\}, \mathcal{B}} \mathbf{K}_{\mathcal{G}, \mathcal{X}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{K}_{\{\varphi\}, \mathcal{B}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{K}_{\{\theta(\cdot, \mathrm{Id})\}, \mathcal{X}}. \end{aligned}$$

By standard empirical process argument,  $\mathcal{H}_1^c$  is a VC-type class with constant  $\mathbf{c}2^{d+1}$  and exponent  $2d + 2$  with respect to envelope function  $\mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{M}_{\{\varphi\}, \mathcal{B}}$  over  $\mathcal{B}$ .

**(3) Properties of  $\mathcal{H}_2^c$**  The main challenge is that  $\mathcal{R}_2$  contains non-differentiable indicator. First, we study properties of  $\mathcal{G} \cdot \mathcal{R}_2$ . By Definition 4,

$$\begin{aligned} \mathrm{TV}_{\mathcal{G} \cdot \mathcal{R}_2, \mathcal{B}} &= \sup_{g \in \mathcal{G}} \sup_{y \in \mathbb{R}} \sup_{\substack{\phi \in \mathcal{D}_{d+1}([0, 1]^{d+1}) \\ \|\phi\|_2 \|\infty \leq 1}} \int_{[0, 1]^d} \int_{[0, 1]} g(\mathbf{x}) \mathbb{1}(u \leq y) \mathrm{div}(\phi)(\mathbf{x}, u) du d\mathbf{x} \\ &\leq \sup_{g \in \mathcal{G}} \sup_{y \in \mathbb{R}} \sup_{\substack{\phi \in \mathcal{D}_d([0, 1]^d) \\ \|\phi\|_2 \|\infty \leq 1}} \sup_{\substack{\psi \in \mathcal{D}_1([0, 1] \\ \|\psi\|_\infty \leq 1}} \int_{[0, 1]^d} \int_{[0, 1]} g(\mathbf{x}) \mathbb{1}(u \leq y) (\mathrm{div} \phi(\mathbf{x}) + \psi'(u)) du d\mathbf{x} \\ &= \sup_{g \in \mathcal{G}} \sup_{y \in \mathbb{R}} \sup_{\substack{\phi \in \mathcal{D}_d([0, 1]^d) \\ \|\phi\|_2 \|\infty \leq 1}} \int_{[0, 1]^d} g(\mathbf{x}) \mathrm{div} \phi(\mathbf{x}) d\mathbf{x} + \sup_{g \in \mathcal{G}} \sup_{y \in \mathbb{R}} \sup_{\substack{\psi \in \mathcal{D}_1([0, 1] \\ \|\psi\|_\infty \leq 1}} \int_{[0, 1]^d} g(\mathbf{x}) d\mathbf{x} (\psi(1) - \psi(0)) \\ &\leq \mathrm{TV}_{\mathcal{G}, \mathcal{X}} + 2\mathbf{E}_{\mathcal{G}, \mathcal{X}}, \end{aligned}$$

where  $\mathcal{D}_{d+1}([0, 1]^{d+1})$  denotes the space of infinitely differentiable functions from  $[0, 1]^{d+1}$  to  $\mathbb{R}^{d+1}$ , and  $\mathcal{D}_d([0, 1]^d)$  is analogously defined. Similar argument as in the proof for properties of  $\mathcal{H}_1^c$  gives

$$\mathrm{TV}_{\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}, \mathcal{B}} \leq \mathrm{TV}_{\mathcal{G}, \mathcal{X}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \sup_{g \in \mathcal{G}} \mathrm{TV}_{\mathcal{V}_{\mathcal{R}}, \mathrm{Supp}(g)} = O(\mathrm{TV}_{\mathcal{G}, \mathcal{X}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \sup_{g \in \mathcal{G}} \mathbf{m}(\mathrm{Supp}(g))).$$

It follows that

$$\mathrm{TV}_{\mathcal{G} \cdot \mathcal{R}_2 + \mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}, \mathcal{B}} = O(\mathrm{TV}_{\mathcal{G}, \mathcal{X}} + \mathbf{E}_{\mathcal{G}, \mathcal{X}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \sup_{g \in \mathcal{G}} \mathbf{m}(\mathrm{Supp}(g))).$$



Consider the change of variables function  $T : [0, 1]^{d+1} \rightarrow \mathbb{R}^{d+1}$  given by  $T(\mathbf{x}, u) = (\mathbf{x}, \varphi(\mathbf{x}, u))$ . Since  $\mathbf{L}_{\{T\}, \mathcal{B}} \leq \max\{\mathbf{L}_{\{\varphi\}, \mathcal{B}}, 1\}$ , Theorem 3.16 from [Ambrosio et al. \(2000\)](#) implies

$$\mathbf{TV}_{\mathcal{H}_2^c, \mathcal{B}} \leq \mathbf{L}_{\{T\}, \mathcal{B}}^{d-1} \mathbf{TV}_{\mathcal{G}, \mathcal{R}_2 + \mathcal{G}, \mathcal{V}_{\mathcal{R}}, \mathcal{B}} = O(\max\{\mathbf{L}_{\{\varphi\}, \mathcal{B}}, 1\}^{d-1} \mathbf{TV}_{\mathcal{G}, \mathcal{R}_2 + \mathcal{G}, \mathcal{V}_{\mathcal{R}}, \mathcal{B}}).$$

By standard empirical process argument,  $\mathcal{H}_2^c$  is a VC-type class with constant  $C_1 2^{d+1}$  and exponent  $2d + 2$  with respect to envelope function  $\mathbf{M}_{\mathcal{G}, \mathcal{X}}$ , where  $C_1$  is a constant that does not depend on  $n$ .

**(4) Effects of Rosenblatt Transformation** By Lemma [SA.10](#) with  $\mathbb{Q}_{\mathcal{G}} = \mathbb{P}_{\mathcal{X}}$  and  $\phi_{\mathcal{G}} = \text{Id}$ , we have  $\mathbf{TV}_{\mathcal{H}_1} \leq \mathbf{TV}_{\mathcal{H}_1^c} \bar{f}_B^2 f_B^{-1}$ ,  $\mathbf{TV}_{\mathcal{H}_2} \leq \mathbf{TV}_{\mathcal{H}_2^c} \bar{f}_B^2 f_B^{-1}$ ,  $\mathbf{M}_{\mathcal{H}_1} = \mathbf{M}_{\mathcal{H}_1^c}$ ,  $\mathbf{M}_{\mathcal{H}_2} = \mathbf{M}_{\mathcal{H}_2^c}$ . Moreover,  $\mathcal{H}_1$  and  $\mathcal{H}_2$  are VC-type classes with constant  $C_2 2^{d+1}$  and exponent  $2d + 2$  with respect to envelope functions  $\mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{M}_{\{\varphi\}, \mathcal{B}}$  and  $\mathbf{M}_{\mathcal{G}, \mathcal{X}}$  respectively, with  $C_2$  a constant that does not depend on  $n$ .

**(5) Application of Theorem 1.1 in [Rio \(1994\)](#)** We can now apply Theorem 1.1 in [Rio \(1994\)](#) to get  $\{X_n(h) : h \in \mathcal{H}_1\}$  admits a Gaussian strong approximation with rate function

$$\begin{aligned} & C_{d, \varphi} \sqrt{\frac{d \bar{f}_B^2}{\underline{f}_B} \frac{\sqrt{\mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{M}_{\{\varphi\}, \mathcal{B}} (\mathbf{TV}_{\mathcal{G}, \mathcal{X}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \sup_{g \in \mathcal{G}} \mathbf{m}(\text{Supp}(g)))}}{n^{\frac{1}{2d+2}}} \sqrt{t + d \log n}} \\ & + C_{d, \varphi} \sqrt{\frac{\mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{M}_{\{\varphi\}, \mathcal{B}}}{n}} \min \left\{ \sqrt{\log(n) \mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{M}_{\{\varphi\}, \mathcal{B}}}, \sqrt{\frac{(2\sqrt{d})^{d-1} \bar{f}_B^{d+1}}{\underline{f}_B^d} (\mathbf{K}_{\mathcal{G}, \mathcal{X}} \mathbf{M}_{\{\varphi\}, \mathcal{B}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{K}_{\{\varphi\}, \mathcal{B}})} \right\} (t + d \log n), \end{aligned}$$

where  $C_{d, \varphi}$  is a quantity that only depends on  $d$  and  $\varphi$ . And  $\{X_n(h) : h \in \mathcal{H}_2\}$  admits a Gaussian strong approximation with rate function

$$C_{d, \varphi} \sqrt{\frac{d \bar{f}_B^2}{\underline{f}_B} \frac{\sqrt{\mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{TV}_{\mathcal{G}, \mathcal{X}, \{\varphi\}, \mathcal{B}}}}{n^{\frac{1}{2d+2}}} \sqrt{t + d \log n} + C_{d, \varphi} \frac{\mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{M}_{\{\varphi\}, \mathcal{B}}}{\sqrt{n}} (t + d \log n),$$

where  $\mathbf{TV}_{\mathcal{G}, \mathcal{X}, \{\varphi\}, \mathcal{B}} = \max\{\mathbf{L}_{\{\varphi\}, \mathcal{B}}, 1\}^{d-1} (\mathbf{TV}_{\mathcal{G}, \mathcal{X}} + \mathbf{E}_{\mathcal{G}, \mathcal{X}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \sup_{g \in \mathcal{G}} \mathbf{m}(\text{Supp}(g)))$ .  $\square$

**Example SA.2** (Strong Approximation via Theorem 1). *Consider the setup of Section 4.1, and assume the following regularity conditions hold:*

- (a)  $(\mathbf{x}_i, y_i) = (\mathbf{x}_i, \varphi(\mathbf{x}_i, u_i))$ , where the law of  $\mathbf{b}_i = (\mathbf{x}_i, u_i)$ ,  $\mathbb{P}_{\mathcal{B}}$ , has a continuous and positive Lebesgue density  $f_{\mathcal{B}}$  on its support  $\mathcal{B} = [0, 1]^{d+1}$ .
- (b)  $\mathbf{M}_{\{\varphi\}, \mathcal{B}} = O(1)$ ,  $\sup_{g \in \mathcal{G}} \mathbf{TV}_{\{\varphi\}, \text{Supp}(g)} = O(\sup_{g \in \mathcal{G}} \mathbf{m}(\text{Supp}(g)))$ ,  $\mathbf{K}_{\{\varphi\}, \mathcal{B}} = O(1)$ , and  $\mathbf{L}_{\{\varphi\}, \mathcal{B}} = O(1)$ .
- (c)  $\sup_{r \in \mathcal{R}_\ell} \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} |\theta(\mathbf{x}, r) - \theta(\mathbf{y}, r)| / \|\mathbf{x} - \mathbf{y}\|_{\infty} < \infty$  for  $\ell = 1, 2$ .

Recall  $\mathcal{G} = \{b^{-d/2} \mathfrak{K}_{\mathbf{w}}(\frac{\cdot - \mathbf{w}}{b}) : \mathbf{w} \in \mathcal{W}\}$  with  $\mathfrak{K}_{\mathbf{w}}(\mathbf{u}) = \mathbf{e}_1^{\top} \mathbf{H}_{\mathbf{w}}^{-1} \mathbf{p}(\mathbf{u}) K(\mathbf{u})$ . For  $\mathcal{R}_1$ , take  $\mathcal{H}_1 = \{h \circ T_{\mathbb{P}_{\mathcal{B}}}^{-1} : h \in \mathcal{H}_1^c\}$ , where  $\mathcal{H}_1^c = \{(\mathbf{x}, u) \in \mathcal{B} \mapsto g(\mathbf{x}) \varphi(\mathbf{x}, u) - g(\mathbf{x}) \theta(\mathbf{x}, \text{Id}) : g \in \mathcal{G}\}$ ,  $T_{\mathbb{P}_{\mathcal{B}}}$  is the Rosenblatt transformation based on  $\mathbb{P}_{\mathcal{B}}$  given in Section 3.1. Recall we denote  $\mathcal{X} = [0, 1]^d$ . Then, Equation [\(SA-22\)](#) holds, and

$$\mathbf{L}_{\mathcal{H}_1} \leq \mathbf{L}_{\mathcal{H}_1^c} \frac{\bar{f}_B}{\underline{f}_B} \leq (\mathbf{L}_{\mathcal{G}, \mathcal{X}} \mathbf{M}_{\{\varphi\}, \mathcal{B}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{L}_{\{\varphi\}, \mathcal{B}} + \mathbf{M}_{\mathcal{G}, \mathcal{X}} \mathbf{L}_{\mathcal{V}_{\mathcal{R}_1}, \mathcal{X}}) \frac{\bar{f}_B}{\underline{f}_B} = O(b^{-d/2-1}).$$

Theorem 1 implies  $(X_n(h) : h \in \mathcal{H}_1) = (\sqrt{nb^d} \mathbf{e}_1^\top \mathbf{H}_\mathbf{x}^{-1} \mathbf{S}_{\mathbf{w},r} : \mathbf{x} \in [0, 1]^d, r \in \mathcal{R}_1)$  admits a uniform Gaussian strong approximation with rate

$$S_n(t) = C_{d,\varphi,\mathbb{P}_B} (nb^{d+1})^{-1/(d+1)} \sqrt{t + d \log n} + C_{d,\varphi,\mathbb{P}_B} (nb^d)^{-1/2} (t + d \log n),$$

where  $C_{d,\varphi,\mathbb{P}_B}$  is a quantity that only depends on  $d$ ,  $\varphi$  and  $\mathbb{P}_B$ .

The strong approximation rate stated in Section 4.1 in the paper now follows directly from the strong approximation result above.  $\blacktriangle$

**Proof of Example SA.2.** Besides the properties given in the proof of Example SA.1, using product rule we can show  $L_{\mathcal{H}_1^c} \leq L_{\mathcal{G},\mathcal{X}} M_{\{\varphi\},\mathcal{B}} + M_{\mathcal{G},\mathcal{X}} L_{\{\varphi\},\mathcal{B}} + M_{\mathcal{G},\mathcal{X}} L_{\mathcal{V}_{\mathcal{R}_1},\mathcal{X}} = O(b^{-d/2-1})$ . By the discussion on Rosenblatt transformation in Section 3.1,  $L_{\mathcal{H}_1,\mathcal{X}} \leq L_{\mathcal{H}_1^c,\mathcal{X}} \bar{f}_B / \underline{f}_B$ . Take the surrogate measure to be  $\mathcal{Q}_{\mathcal{H}_1} = \text{Uniform}([0, 1]^{d+1})$  with  $\phi_{\mathcal{H}_1} = \text{Id}$ . The result then follows from application of Theorem SA.1 on

$$X_n(h) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [h(\mathbf{x}_i, u_i) - \mathbb{E}[h(\mathbf{x}_i, u_i)]], \quad h \in \mathcal{H}_1.$$

This completes the proof.  $\square$

**Example SA.3** (Strong Approximation via Theorem 2). Consider the setup of Section 4.1 and assume the following regularity conditions hold:

- (a)  $\mathbf{x}_i$  has  $\mathbb{P}_X$  with Lebesgue density  $f_X$  continuous on its support  $\mathcal{X}$ , which is a compact subset of  $\mathbb{R}^d$ , and  $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(|y_i|) | \mathbf{x}_i = \mathbf{x}] \leq 2$ .
- (b)  $\sup_{r \in \mathcal{R}_\ell} \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} |\theta(\mathbf{x}, r) - \theta(\mathbf{y}, r)| / \|\mathbf{x} - \mathbf{y}\|_\infty < \infty$  for  $\ell = 1, 2$ .

Recall that  $\mathcal{G} = \{b^{-d/2} \mathfrak{R}_\mathbf{x}(\cdot - \frac{\mathbf{x}}{b}) : \mathbf{x} \in \mathcal{X}\}$ . Take the surrogate measure  $\mathcal{Q}_\mathcal{G} = \mathbb{P}_X$  and the normalizing transformation  $\phi_\mathcal{G} = \text{Id}$ . Then, using the notation introduced in the paper,

$$\mathbf{c}_1 = d \frac{\bar{f}_X^2}{\underline{f}_X}, \quad \mathbf{c}_2 = \frac{\bar{f}_X}{\underline{f}_X},$$

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

$$\begin{aligned} M_\mathcal{G} &= O(b^{-d/2}), & E_\mathcal{G} &= O(b^{d/2}), & \text{TV}_\mathcal{G} &= O(b^{d/2-1}), & L_\mathcal{G} &= O(b^{-d/2-1}), \\ N_\mathcal{G}(\delta) &= O(\delta^{-d-1}), & 0 &< \delta < 1. \end{aligned}$$

Theorem 2 implies that  $(R_n(g, r) : g \in \mathcal{G}, r \in \mathcal{R}_1) = (\sqrt{nb^d} \mathbf{e}_1^\top \mathbf{H}_\mathbf{x}^{-1} \mathbf{S}_{\mathbf{w},r} : \mathbf{x} \in [0, 1]^d, r \in \mathcal{R}_1)$  admits a uniform Gaussian strong approximation with rate function

$$S_n(t) = \left( \frac{\bar{f}_X^3}{\underline{f}_X^2} \right)^{\frac{d}{2(d+2)}} \sqrt{d} (nb^d)^{-1/(d+2)} (t + d \log n)^{3/2} + (nb^d)^{-1/2} (t + d \log n).$$

If, in addition,  $\sup_{\mathbf{x} \in [0, 1]^d} \mathbb{E}[\exp(|y_i|) | \mathbf{x}_i = \mathbf{x}] \leq 2$ , then Theorem 2 implies  $(R_n(g, r) : g \in \mathcal{G}, r \in \mathcal{R}_1) =$

$(\sqrt{nb^d} \mathbf{e}_1^\top \mathbf{H}_X^{-1} \mathbf{S}_{\mathbf{w},r} : \mathbf{x} \in [0, 1]^d, r \in \mathcal{R}_1)$  admits a uniform Gaussian strong approximation with rate function

$$S_n(t) = \left( \frac{\bar{f}_X^3}{\bar{f}_X^2} \right)^{\frac{d}{2(d+2)}} \sqrt{d} (nb^d)^{-1/(d+2)} (t + d \log n)^{5/2} + (nb^d)^{-1/2} (t + d \log n).$$

The strong approximation rate stated in Section 4.1 in the paper now follow directly from the strong approximation result above.  $\blacktriangle$

**Proof of Example SA.3.** The conditions of  $\mathcal{G}$  can be verified from Part (1) Properties of  $\mathcal{G}$  in Section SA.1. It is easy to check that  $\mathcal{R}_1$  satisfies the conditions in Theorem 2 with  $c_{\mathcal{R}_1} = 1$ ,  $d_{\mathcal{R}_1} = 1$  and  $\alpha = 1$ . Moreover,  $\mathcal{R}_2$  satisfies the conditions in Theorem 2 with  $c_{\mathcal{R}_2}$  some universal constant and  $d_{\mathcal{R}_2} = 2$  by van der Vaart and Wellner (2013, Theorem 2.6.7). The results then follow from Theorem 2.  $\square$

## SA-V Quasi-Uniform Haar Basis

This section provides the proofs and additional results for Section 5. In Section SA-V.1, we present the proofs of Theorem 3 and Corollary 5, and verify the claims for Example 2. In Section SA-V.2, we present the proofs of Theorem 4, Corollary 6 and the Haar Partitioning-based Regression example in Section 5.3, with the additional results for  $M_n$  and  $R_n$  processes under generic entropy conditions.

### SA-V.1 General Empirical Process

#### SA-V.1.1 Proof of Theorem 3

First, we make a reduction through the surrogate measure  $\mathbb{Q}_{\mathcal{H}}$  (Definition 2). Denote  $\mathcal{E}_{\mathcal{H}} = \mathcal{X} \cap \text{Supp}(\mathcal{H})$ . The definition of surrogate measure implies  $\mathbb{P}_X|_{\mathcal{E}_{\mathcal{H}}} = \mathbb{Q}_{\mathcal{H}}|_{\mathcal{E}_{\mathcal{H}}}$ . We use a coupling argument similar to the proof of Theorem 1. Define a probability measure  $\mathbb{O}$  on  $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^{2d}))$  such that for all  $A \in \mathcal{B}(\mathbb{R}^{2d})$ ,  $\mathbb{O}|_{\mathcal{E}_{\mathcal{H}} \times \mathcal{E}_{\mathcal{H}}}(A) = \mathbb{P}_X(\Pi_{1:d}(A \cap \{(\mathbf{x}, \mathbf{x}) : \mathbf{x} \in \mathcal{E}_{\mathcal{H}}\}))$ ,  $\mathbb{O}|_{\mathcal{E}_{\mathcal{H}} \times \mathcal{E}_{\mathcal{H}}^c}(A) = \mathbb{O}|_{\mathcal{E}_{\mathcal{H}}^c \times \mathcal{E}_{\mathcal{H}}}(A) = 0$ ,  $\mathbb{O}|_{\mathcal{E}_{\mathcal{H}}^c \times \mathcal{E}_{\mathcal{H}}^c}(A) = \int_{\mathcal{E}_{\mathcal{H}}^c} \mathbb{P}_X(A^{\mathbf{y}} \cap \mathcal{E}_{\mathcal{H}}^c) d\mathbb{Q}(\mathbf{y})$  where  $A^{\mathbf{y}} = \{\mathbf{x} \in \mathbb{R}^d : (\mathbf{x}, \mathbf{y}) \in A\}$ , where we take  $\Pi_{1:d}(E) = \{\mathbf{x} \in \mathbb{R}^d : (\mathbf{x}, \mathbf{y}) \in E \text{ for some } \mathbf{y} \in \mathbb{R}^d\}$  for any  $E \in \mathcal{B}(\mathbb{R}^{2d})$ .

The definition of  $\mathbb{O}$  implies the marginals are  $\mathbb{P}_X$  and  $\mathbb{Q}_{\mathcal{H}}$ , respectively. By Skorohod embedding (Dudley, 2014, Lemma 3.35), on a possibly enlarged probability space, there exists  $(\mathbf{z}_i : 1 \leq i \leq n)$  i.i.d. with law  $\mathbb{Q}_{\mathcal{H}}$  such that  $(\mathbf{x}_i, \mathbf{z}_i)$  has joint law  $\mathbb{O}$  for each  $1 \leq i \leq n$ . In particular, when  $\mathbf{x}_i \in \mathcal{E}_{\mathcal{H}}$ ,  $\mathbf{z}_i = \mathbf{x}_i$ ; and  $\mathbb{O}(\{\mathbf{x}_i \in \mathcal{E}_{\mathcal{H}}\} \triangle \{\mathbf{z}_i \in \mathcal{E}_{\mathcal{H}}^c\}) = 0$ . Moreover, since  $\mathbb{Q}_{\mathcal{H}}(\text{Supp}(\mathcal{H}) \setminus \mathcal{X}) = 0$ , and the definition of  $\mathbb{O}$  on  $\mathcal{E}_{\mathcal{H}} \times \mathcal{E}_{\mathcal{H}}$  as a product measure between  $\mathbb{P}_X$  and  $\mathbb{Q}_{\mathcal{H}}$ , we know  $\mathbb{O}(\{\mathbf{x}_i \in \mathcal{E}_{\mathcal{H}}^c\} \triangle \{\mathbf{z}_i \in \text{Supp}(\mathcal{H})^c\}) = 0$ . This allows for the reduction to  $\mathbf{z}_i$ -based processes, since for  $1 \leq i \leq n$ , almost surely

$$\begin{aligned} h(\mathbf{x}_i) &= h(\mathbf{x}_i) \mathbb{1}(\mathbf{x}_i \in \mathcal{E}_{\mathcal{H}}) + 0 \cdot \mathbb{1}(\mathbf{x}_i \in \mathcal{E}_{\mathcal{H}}^c) \\ &= h(\mathbf{z}_i) \mathbb{1}(\mathbf{z}_i \in \mathcal{E}_{\mathcal{H}}) + 0 \cdot \mathbb{1}(\mathbf{z}_i \in \text{Supp}(\mathcal{H})^c) \\ &= h(\mathbf{z}_i) \mathbb{1}(\mathbf{z}_i \in \mathcal{E}_{\mathcal{H}}) + h(\mathbf{z}_i) \cdot \mathbb{1}(\mathbf{z}_i \in \text{Supp}(\mathcal{H})^c) \\ &= h(\mathbf{z}_i) \mathbb{1}(\mathbf{z}_i \in \mathcal{E}_{\mathcal{H}}) + h(\mathbf{z}_i) \cdot \mathbb{1}(\mathbf{z}_i \in \mathcal{E}_{\mathcal{H}}^c) \\ &= h(\mathbf{z}_i), \quad \forall h \in \mathcal{H}, \end{aligned}$$

where the first line is due to  $h = 0$  on  $\text{Supp}(\mathcal{H})^c$ , the second line is by  $\mathbb{O}(\{\mathbf{x}_i \in \mathcal{E}_{\mathcal{H}}^c\} \Delta \{\mathbf{z}_i \in \text{Supp}(\mathcal{H})^c\}) = 0$ , the third line is due to  $h = 0$  on  $\text{Supp}(\mathcal{H})^c$ , and the fourth line by  $\mathbb{Q}_{\mathcal{H}}(\text{Supp}(\mathcal{H}) \setminus \mathcal{X}) = 0$ . Almost surely,

$$X_n(h) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [h(\mathbf{x}_i) - \mathbb{E}[h(\mathbf{x}_i)]] = \frac{1}{\sqrt{n}} \sum_{i=1}^n [h(\mathbf{z}_i) - \mathbb{E}[h(\mathbf{z}_i)]], \quad \forall h \in \mathcal{H}.$$

Hence we reduce the problem to coupling for  $(\tilde{X}_n(h) : h \in \mathcal{H})$ , with the process defined by

$$\tilde{X}_n(h) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [h(\mathbf{z}_i) - \mathbb{E}[h(\mathbf{z}_i)]], \quad h \in \mathcal{H},$$

with  $(\mathbf{z}_i : 1 \leq i \leq n)$  i.i.d  $\sim \mathbb{Q}_{\mathcal{H}}$ . Suppose  $2^K \leq L < 2^{K+1}$ . For each  $l \in \{1, 2, \dots, d\}$ , we can divide at most  $2^K$  cells into two intervals of equal measure under  $\mathbb{Q}_{\mathcal{H}}$  such that we get a new partition of  $\mathbb{Q}_{\mathcal{H}} = \sqcup_{0 \leq j < 2^{K+1}} \Delta'_l$  and satisfies

$$\frac{\max_{0 \leq l < 2^{K+1}} \mathbb{Q}_{\mathcal{H}}(\Delta'_l)}{\min_{0 \leq l < 2^{K+1}} \mathbb{Q}_{\mathcal{H}}(\Delta'_l)} \leq 2\rho.$$

By construction, there exists an axis-aligned quasi-dyadic expansion  $\mathcal{A}_{K+1}(\mathbb{Q}_{\mathcal{H}}, 2\rho) = \{\mathcal{C}_{j,k} : 0 \leq j \leq K+1, 0 \leq k < 2^{K+1-j}\}$  such that

$$\{\mathcal{C}_{0,k} : 0 \leq k < 2^{K+1}\} = \{\Delta'_l : 0 \leq l < 2^{K+1}\},$$

and  $\mathcal{H} \subseteq \text{Span}\{\mathbb{1}_{\Delta_j} : 0 \leq j < L\} \subseteq \text{Span}\{\mathbb{1}_{\mathcal{C}_{0,k}} : 0 \leq k < 2^{K+1}\}$ . Now we consider the term  $\mathbf{C}_{\mathcal{H}}$  from Lemma SA.5. Let  $h \in \mathcal{H}$ . By definition of  $S$  and the step of splitting each cell into at most two, there exists  $l_1, \dots, l_{2S} \in \{0, \dots, 2^{K+1} - 1\}$  such that  $h = \sum_{q=1}^{2S} c_q \mathbb{1}(\Delta'_{l_q})$  where  $|c_q| \leq \mathbb{M}_{\{h\}}$ . Fix  $(j, k)$ . Let  $(l, m)$  be an index such that  $\mathcal{C}_{l,m} \subseteq \mathcal{C}_{j,k}$ . Since each  $\Delta'_{l_q}$  belongs to at most one  $\mathcal{C}_{l-1,k}$ ,  $\tilde{\beta}_{l,m}(\mathbb{1}(\Delta'_{l_q})) = 0$  if  $\Delta'_{l_q}$  is not contained in  $\mathcal{C}_{l,m}$  and  $\tilde{\beta}_{l,m}(\mathbb{1}(\Delta'_{l_q})) = 2^{-l+1}$  if  $\Delta'_{l_q} \subseteq \mathcal{C}_{l,m}$ . Hence

$$\sum_{m: \mathcal{C}_{l,m} \subseteq \mathcal{C}_{j,k}} |\tilde{\beta}_{l,m}(h)|^2 \leq 2S \sum_{q=1}^{2S} \sum_{m: \mathcal{C}_{l,m} \subseteq \mathcal{C}_{j,k}} (c_q \tilde{\beta}_{l,m}(\mathbb{1}(\Delta'_{l_q})))^2 \leq 2S \sum_{q=1}^{2S} c_q^2 2^{-2l} \leq 4S^2 \mathbb{M}_{\mathcal{H}}^2 2^{-2l}.$$

It follows that

$$\mathbf{C}_{\mathcal{H}} = \sup_{h \in \mathcal{H}} \min \left\{ \sup_{(j,k)} \left[ \sum_{l < j} (j-l)(j-l+1) 2^{l-j} \sum_{m: \mathcal{C}_{l,m} \subseteq \mathcal{C}_{j,k}} \tilde{\beta}_{l,m}^2(h) \right], \mathbb{M}_{\mathcal{H}}^2(K+1) \right\} \lesssim \mathbb{M}_{\mathcal{H}}^2 \min\{K, S^2\}.$$

Then apply Lemma SA.5, we get there exists a mean-zero Gaussian process  $\tilde{Z}_n^X$  with the same covariance structure as  $\tilde{X}_n$  such that with probability at least  $1 - 2 \exp(-t) - 2^{K+1} \exp(-C_\rho n 2^{-K-1})$ ,

$$\|\tilde{X}_n - \tilde{Z}_n^X\|_{\mathcal{H}} \leq \min_{\delta \in (0,1)} \left\{ C_\rho \sqrt{\frac{2^{K+2} \mathbb{M}_{\mathcal{H}} \mathbb{E}_{\mathcal{H}}}{n}} (t + \log N_{\mathcal{H}}(\delta, \mathbb{M}_{\mathcal{H}})) \right. \\ \left. + C_\rho \sqrt{\frac{\min\{K, S^2\}}{n}} \mathbb{M}_{\mathcal{H}} (t + \log N_{\mathcal{H}}(\delta, \mathbb{M}_{\mathcal{H}})) + F_n(t, \delta) \right\},$$

where  $K \leq \log_2(L)$ , and  $C_\rho$  is a constant that only depends on  $\rho$ . The conclusion then follows from taking  $(Z_n(h) : h \in \mathcal{H}) = (\tilde{Z}_n(h) : h \in \mathcal{H})$  and the fact that  $(X_n(h) : h \in \mathcal{H}) = (\tilde{X}_n(h) : h \in \mathcal{H})$  almost surely.  $\square$

### SA-V.1.2 Proof of Corollary 5

Take  $t = C \log n$  with  $C > 1$  and  $\delta = n^{-\frac{1}{2}}$  in Theorem 3. □

### SA-V.1.3 Example 2: Histogram Density Estimation

Recall for  $\mathbf{w} \in \mathcal{W}$ , we define

$$h_{\mathbf{w}}(\mathbf{u}) = \sqrt{L} \sum_{0 \leq l < P} \mathbb{1}(\mathbf{u} \in \Delta_l) \mathbb{1}(\mathbf{w} \in \Delta_l), \quad \mathbf{u} \in \mathcal{X}, \mathbf{w} \in \mathcal{W},$$

and  $\mathcal{H} = \{h_{\mathbf{w}}(\cdot) : \mathbf{w} \in \mathcal{W}\}$ . Then  $\mathcal{H} \subseteq \text{Span}(\mathbb{1}_{\Delta_l} : 0 \leq l < P)$ . In particular, for every  $\mathbf{u} \in \mathcal{X}$  and  $\mathbf{w} \in \mathcal{W}$ , at most one of  $\mathbb{1}(\mathbf{u} \in \Delta_l) \mathbb{1}(\mathbf{w} \in \Delta_l)$  will be non-zero. Hence  $M_{\mathcal{H}, \mathbb{R}^d} = L^{1/2}$ . Each function in  $\mathcal{H}$  can be written as  $c \mathbb{1}(\Delta_l)$  for some  $l \leq L$ , which implies we can take  $S_{\mathcal{H}} = 1$ .

If  $\mathcal{W} = \mathcal{X}$ , since we assume the partition is quasi-uniform of  $\mathcal{Q}_{\mathcal{H}} = \mathcal{X}$  with  $\mathbb{Q}_{\mathcal{H}} = \mathbb{P}_X$ , we know  $\max_{0 \leq l < P} \mathbb{P}_X(\Delta_l) \leq c_{\rho} L^{-1}$  for some constant  $c_{\rho} > 0$  that only depends on  $\rho$ , which implies

$$E_{\mathcal{H}} \leq \max_{0 \leq l < P} \mathbb{P}_X(\Delta_l) \cdot M_{\mathcal{H}} \leq c_{\rho} L^{-1} \sqrt{L} \leq c_{\rho} L^{-1/2},$$

where in this case  $P = L$ .

If  $\mathcal{W} \subsetneq \mathcal{X}$ , take  $\mathring{P}$  to be the unique number in  $\mathbb{Z}$  such that  $\mathring{P} \leq \frac{\mathbb{P}_X(\cup_{0 \leq l < P} \Delta_l)^c}{\min_{0 \leq l < P} \mathbb{P}_X(\Delta_l)} < \mathring{P} + 1$ . Then we consider the following two cases. Set  $L = P + \mathring{P}$ .

*Case 1:*  $\mathring{P} \geq 1$ . The construction in Example 2 implies for every  $P \leq l < L$ ,

$$1 \leq \frac{\mathbb{Q}_{\mathcal{H}}(\Delta_l)}{\min_{0 \leq k < P} \mathbb{P}_X(\Delta_k)} \leq 1 + \mathring{P}^{-1} \leq 2.$$

In particular,  $\min_{0 \leq k < L} \mathbb{P}_X(\Delta_l) = \min_{0 \leq k < P} \mathbb{P}_X(\Delta_l)$ . Combined with quasi-uniformity of  $\{\Delta_l : 0 \leq l < P\}$ , we have

$$\frac{\max_{0 \leq k < L} \mathbb{P}_X(\Delta_k)}{\min_{0 \leq k < L} \mathbb{P}_X(\Delta_k)} \leq \max\{\rho, 2\}.$$

Since  $\mathbb{P}_X$  agrees with  $\mathbb{Q}_{\mathcal{H}}$  on  $\sqcup_{0 \leq l < P} \Delta_l$ , and  $\sqcup_{P \leq l < L} \Delta_l \subseteq \mathcal{X} \cup \text{Supp}(\mathcal{H})^c$ ,  $\mathbb{Q}_{\mathcal{H}}$  is a surrogate measure of  $\mathbb{P}_X$  with respect to  $\mathcal{H}$ . And we verified that  $\{\Delta_l : 0 \leq l < L\}$  is a quasi-uniform partition of  $\mathcal{Q}_{\mathcal{H}}$  with respect to  $\mathbb{Q}_{\mathcal{H}}$ .

*Case 2:*  $\mathring{P} = 0$ . Then for any  $0 \leq l < P$ , there exists  $\mathring{P}_l \in \mathbb{N}$  such that

$$\mathring{P}_l \leq \frac{\mathbb{P}_X(\Delta_l)}{\mathbb{P}_X(\cup_{0 \leq l < P} \Delta_l)^c} < \mathring{P}_l + 1.$$

Taking arbitrary  $\Delta_P$  with  $\mathbb{P}_X(\Delta_P) = \mathbb{P}_X(\cup_{0 \leq l < P} \Delta_l)^c$ , and for  $0 \leq l < P$  break  $\Delta_l$  into  $\mathring{P}_l$  pieces of equal measure by  $\mathbb{P}_X$ , we can show by similar arguments as above that the refined cells with the additional  $\Delta_P$  together forms a quasi-uniform partition of  $\mathcal{X}$  with respect to  $\mathbb{P}_X$ . Suppose also in this case, the number of cells in the quasi-uniform partition is  $L$  after refinement.

In both cases, we know  $\max_{0 \leq l < L} \mathbb{P}_X(\Delta_l) \leq c_{\rho} L^{-1}$  for some constant  $c_{\rho}$  that only depends on  $\rho$ , which

implies

$$\mathbf{E}_{\mathcal{J}\mathcal{C}} \leq \max_{0 \leq l < P} \mathbb{P}_X(\Delta_l) \cdot M_{\mathcal{J}\mathcal{C}} \leq c_\rho L^{-1} \sqrt{L} \leq c_\rho L^{-1/2}.$$

We can then apply Theorem 3 to get the stated rates.  $\square$

## SA-V.2 Residual-Based (and Multiplicative Separable) Empirical Process

For  $\delta \in (0, 1]$ , define

$$N(\delta) = N_{\mathcal{G}}(\delta/\sqrt{2}, M_{\mathcal{G}}) N_{\mathcal{R}}(\delta/\sqrt{2}, M_{\mathcal{R}})$$

and

$$J(\delta) = \sqrt{2}J(\mathcal{G}, M_{\mathcal{G}}, \delta/\sqrt{2}) + \sqrt{2}J(\mathcal{R}, M_{\mathcal{R}}, \delta/\sqrt{2}).$$

To simplify notation, the parameters of  $\mathcal{G}$  and  $\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}$  (Definitions 4 to 12, SA.1, SA.2) are taken with  $\mathcal{C} = \mathcal{Q}_{\mathcal{G}}$ , and the index  $\mathcal{Q}_{\mathcal{G}}$  is omitted where there is no ambiguity; the parameters of  $\mathcal{R}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{Y}$ , and the index  $\mathcal{Y}$  is omitted where there is no ambiguity; and the parameters of  $\mathcal{G} \times \mathcal{R}$  (Definitions 4 to 12, SA.3, SA.4) are taken with  $\mathcal{C} = \mathcal{Q}_{\mathcal{G}} \times \mathcal{Y}$ , and the index  $\mathcal{Q}_{\mathcal{G}} \times \mathcal{Y}$  is omitted where there is no ambiguity.

**Theorem SA.3.** *Suppose  $(\mathbf{z}_i = (\mathbf{x}_i, y_i) : 1 \leq i \leq n)$  are i.i.d. random vectors taking values in  $(\mathbb{R}^{d+1}, \mathcal{B}(\mathbb{R}^{d+1}))$ , 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)  $\mathcal{G} \subseteq \text{Span}\{\mathbb{1}_{\Delta_l} : 0 \leq l < L\}$  is a class of Haar functions on  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P}_X)$ .
- (ii) There exists a surrogate measure  $\mathbb{Q}_{\mathcal{G}}$  for  $\mathbb{P}_X$  with respect to  $\mathcal{G}$  such that  $\{\Delta_l : 0 \leq l < L\}$  forms a quasi-uniform partition of  $\mathcal{Q}_{\mathcal{G}}$  with respect to  $\mathbb{Q}_{\mathcal{G}}$ :

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

- (iii)  $\mathcal{G}$  is a VC-type class with envelope function  $M_{\mathcal{G}}$  over  $\mathcal{Q}_{\mathcal{G}}$  with  $c_{\mathcal{G}} \geq e$  and  $d_{\mathcal{G}} \geq 1$ .

- (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  $c_{\mathcal{R}, \mathcal{Y}} \geq e$  and  $d_{\mathcal{R}, \mathcal{Y}} \geq 1$ , where  $M_{\mathcal{R}, \mathcal{Y}}(y) + \text{pTV}_{\mathcal{R}, (-|y|, |y|)} \leq v(1 + |y|^\alpha)$  for all  $y \in \mathcal{Y}$ , for some  $v > 0$ , and for some  $\alpha \geq 0$ . Furthermore, if  $\alpha > 0$ , then  $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(|y_i|)|\mathbf{x}_i = \mathbf{x}] \leq 2$ .

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

- $\mathbb{E}[G_n(g_1, r_1)G_n(g_2, r_2)] = \mathbb{E}[Z_n^G(g_1, r_1)Z_n^G(g_2, r_2)]$  for all  $(g_1, r_1), (g_2, r_2) \in \mathcal{G} \times \mathcal{R}$ , and
- $\mathbb{P}[\|G_n - Z_n^G\|_{\mathcal{G} \times \mathcal{R}} > C_1 C_{v, \alpha} C_\rho \min_{\delta \in (0, 1)} (H_n^G(t, \delta) + F_n^G(t, \delta))] \leq C_2 e^{-t} + L e^{-C_\rho n/L}$  for all  $t > 0$ ,

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

$$\begin{aligned} \mathbf{H}_n^G(t, \delta) &= \sqrt{\frac{LM_{\mathcal{G}}E_{\mathcal{G}}}{n}} (t + \log N_{\mathcal{G}}(\delta/2) + \log N_{\mathcal{R}}(\delta/2) + \log_2 N^*)^{\alpha + \frac{1}{2}} \\ &\quad + \sqrt{\frac{\min\{L + N^*, \mathbf{S}_{\mathcal{G}}^2\}}{n}} M_{\mathcal{G}} (\log n)^\alpha (t + \log N_{\mathcal{G}}(\delta/2) + \log N_{\mathcal{R}}(\delta/2) + \log_2 N^*)^{\alpha + 1}, \end{aligned}$$

and recall

$$\mathbf{F}_n^G(t, \delta) = J(\delta)M_{\mathcal{G}} + \frac{(\log n)^{\alpha/2} M_{\mathcal{G}} J^2(\delta)}{\delta^2 \sqrt{n}} + \frac{M_{\mathcal{G}}}{\sqrt{n}} \sqrt{t} + (\log n)^\alpha \frac{M_{\mathcal{G}}}{\sqrt{n}} t^\alpha,$$

with  $N^* = \left\lceil \log_2 \left( \frac{nM_{\mathcal{G}}}{2^L E_{\mathcal{G}}} \right) \right\rceil$ ,  $\mathbf{S}_{\mathcal{G}} = \sup_{g \in \mathcal{G}} \sum_{l=1}^L \mathbb{1}(\text{Supp}(g) \cap \Delta_l \neq \emptyset)$ .

**Proof of Theorem SA.3.** First, we make a reduction through the surrogate measure and normalizing transformation. Let  $\mathcal{Z}_{\mathcal{G}} = \mathcal{X} \cap \text{Supp}(\mathcal{G})$ . Definition 2 implies  $\mathbb{P}_X|_{\mathcal{Z}_{\mathcal{G}}} = \mathbb{Q}_{\mathcal{G}}|_{\mathcal{Z}_{\mathcal{G}}}$ . Define a joint probability measure  $\mathbb{O}$  on  $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^{2d}))$  such that for all  $A \in \mathcal{B}(\mathbb{R}^{2d})$

$$\begin{aligned} \mathbb{O}(A \cap (\mathcal{Z}_{\mathcal{G}} \times \mathcal{Z}_{\mathcal{G}})) &= \mathbb{P}_X(\Pi_{1:d}(A \cap \{(\mathbf{x}, \mathbf{x}) : \mathbf{x} \in \mathcal{Z}_{\mathcal{G}}\})), \\ \mathbb{O}(A \cap (\mathcal{Z}_{\mathcal{G}} \times \mathcal{Z}_{\mathcal{G}}^c)) &= \mathbb{O}(A \cap (\mathcal{Z}_{\mathcal{G}}^c \times \mathcal{Z}_{\mathcal{G}})) = 0, \\ \mathbb{O}(A \cap (\mathcal{Z}_{\mathcal{G}}^c \times \mathcal{Z}_{\mathcal{G}}^c)) &= \int_{\mathcal{Z}_{\mathcal{G}}^c} \mathbb{P}_X(A^{\mathbf{y}} \cap \mathcal{Z}_{\mathcal{G}}^c) d\mathbb{Q}_{\mathcal{G}}(\mathbf{y}), \end{aligned}$$

where for  $A \in \mathcal{B}(\mathbb{R}^{2d})$ ,  $\Pi_{1:d}(A) = \{\mathbf{x} \in \mathbb{R}^d : (\mathbf{x}, \mathbf{y}) \in A \text{ for some } \mathbf{y} \in \mathbb{R}^d\}$ ,  $A^{\mathbf{y}} = \{\mathbf{x} \in \mathbb{R}^d : (\mathbf{x}, \mathbf{y}) \in A\}$ .

Then we can check that (i) the marginals of  $\mathbb{O}$  are  $\mathbb{P}_X$  and  $\mathbb{Q}_{\mathcal{G}}$ , respectively; (ii)  $\mathbb{O}|_{\mathcal{Z}_{\mathcal{G}} \times \mathbb{R}^d \cup \mathbb{R}^d \times \mathcal{Z}_{\mathcal{G}}}$  is supported on  $\{(\mathbf{x}, \mathbf{x}) : \mathbf{x} \in \mathcal{Z}_{\mathcal{G}}\}$ . By Skorohod embedding (Dudley, 2014, Lemma 3.35), on a possibly enlarged probability space, there exists a  $\mathbf{u}_i, 1 \leq i \leq n$  i.i.d.  $\sim \mathbb{Q}_{\mathcal{G}}$  such that  $(\mathbf{x}_i, \mathbf{u}_i)$  has joint law  $\mathbb{O}$ . In particular, if  $\mathbf{x}_i \in \mathcal{Z}_{\mathcal{G}}$ , then  $\mathbf{x}_i = \mathbf{u}_i$ ; if  $\mathbf{x}_i \in \mathcal{Z}_{\mathcal{G}}^c$ , then  $\mathbf{u}_i \in \mathcal{Z}_{\mathcal{G}}^c$ , and since  $\mathcal{Q}_{\mathcal{G}} \subseteq \mathcal{X} \cup (\cap_{g \in \mathcal{G}} \text{Supp}(g)^c)$ ,  $\mathbf{u}_i \in \cap_{g \in \mathcal{G}} \text{Supp}(g)^c$ . Thus for any  $g \in \mathcal{G}, r \in \mathcal{R}$ ,

$$G_n(g, r) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [g(\mathbf{x}_i)r(y_i) - \mathbb{E}[g(\mathbf{x}_i)r(y_i)]] = \frac{1}{\sqrt{n}} \sum_{i=1}^n [g(\mathbf{u}_i)r(y_i) - \mathbb{E}[g(\mathbf{u}_i)r(y_i)]],$$

where the second equality follows because  $\mathbf{x}_i = \mathbf{u}_i$  on the event  $\{\mathbf{x}_i \in \mathcal{Z}_{\mathcal{G}}\}$ , and  $g(\mathbf{x}_i) = g(\mathbf{u}_i) = 0$  (a.s.) on the event  $\{\mathbf{x}_i \in \mathcal{Z}_{\mathcal{G}}^c\}$ . Hence, we work with an equivalent empirical process

$$\tilde{G}_n(g, r) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [g(\mathbf{u}_i)r(y_i) - \mathbb{E}[g(\mathbf{u}_i)r(y_i)]], \quad g \in \mathcal{G}, r \in \mathcal{R}.$$

In particular  $(\tilde{G}_n(g, r) : g \in \mathcal{G}, r \in \mathcal{R}) = (G_n(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$ . Hence w.l.o.g. assume  $\mathbb{Q}_{\mathcal{G}} = \mathbb{P}_X$  and we work with the  $(G_n(g, r) : g \in \mathcal{G}, r \in \mathcal{R})$  process.

Suppose  $2^M \leq L < 2^{M+1}$ . For each  $l \in \{1, 2, \dots, d\}$ , we can divide at most  $2^M$  cells into two intervals of equal measure under  $\mathbb{P}_X$  such that we get a new partition of  $\mathcal{X} = \sqcup_{0 \leq j < 2^{M+1}} \Delta'_l$  and satisfies

$$\frac{\max_{0 \leq l < 2^{M+1}} \mathbb{P}_X(\Delta'_l)}{\min_{0 \leq l < 2^{M+1}} \mathbb{P}_X(\Delta'_l)} \leq 2\rho.$$

By construction, for each  $N \in \mathbb{N}$ , there exists an axis-aligned quasi-dyadic expansion  $\mathcal{A}_{M+1,N}(\mathbb{P}_Z, 2\rho) = \{\mathcal{C}_{j,k} : 0 \leq j \leq M+1+N, 0 \leq k < 2^{M+1+N-j}\}$  such that

$$\{\mathcal{X}_{0,k} : 0 \leq k < 2^{M+1}\} = \{\Delta'_l : 0 \leq l < 2^{M+1}\},$$

and  $\mathcal{G} \subseteq \text{Span}\{\mathbb{1}_{\Delta_j} : 0 \leq j < J\} \subseteq \text{Span}\{\mathbb{1}_{\mathcal{X}_{0,k}} : 0 \leq k < 2^{M+1}\}$ . Hence

$$\Pi_0(g, r) = \Pi_1(g, r) = \sum_{0 \leq l < 2^{K+1}} \sum_{0 \leq m < 2^N} \mathbb{1}(\mathcal{X}_{0,l} \times \mathcal{Y}_{j,l,m}) g|_{\mathcal{X}_{0,l}} \mathbb{E}[r(y_i) | \mathbf{x}_i \in \mathcal{X}_{0,l}, y_i \in \mathcal{Y}_{j,l,m}]. \quad (\text{SA-23})$$

Again, consider  $(\mathcal{G} \times \mathcal{R})_\delta$  which is a  $\delta \|\mathbf{M}_{\mathcal{G}} M_{\mathcal{R}}\|_{\mathbb{P}_Z}$  of  $\mathcal{G} \times \mathcal{R}$  of cardinality no greater than  $N_{\mathcal{G} \times \mathcal{R}}(\delta, \mathbf{M}_{\mathcal{G}} M_{\mathcal{R}})$ ,  $0 < \delta \leq 1$ . The SA error for projected process on the  $\delta$ -net is given by Lemma SA.18: For all  $t > 0$ ,

$$\begin{aligned} & \mathbb{P}\left[\|\Pi_1 G_n - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta} > C_{v,\alpha} \sqrt{\frac{N^{2\alpha+1} 2^{M+1} \mathbf{E}_{\mathcal{G}} \mathbf{M}_{\mathcal{G}}}{n}} t + C_{v,\alpha} \sqrt{\frac{\mathbf{C}_{\Pi_1(\mathcal{G} \times \mathcal{R}), M+N}}{n}} t\right] \\ & \leq 2N_{\mathcal{G} \times \mathcal{R}}(\delta, \mathbf{M}_{\mathcal{G}} M_{\mathcal{R}}) e^{-t} + 2^M \exp(-C_\rho n 2^{-M}). \end{aligned}$$

Now we find an upper bound for  $\mathbf{C}_{\Pi_1(\mathcal{G} \times \mathcal{R}), M+N}$ . Consider the following two cases.

**Case 1:**  $j \geq N$  Let  $g \in \mathcal{G}, r \in \mathcal{R}$ . Fix  $(j, k)$ . Let  $(j', m')$  be an index such that  $\mathcal{C}_{j',m'} \subseteq \mathcal{C}_{j,k}$ . If  $N \leq j' \leq M+N$ , then by definition of  $S$  and the step of splitting each cell into at most two, there exists  $l_1, \dots, l_{2S} \in \{0, \dots, 2^{M+1} - 1\}$  with possible duplication such that  $g = \sum_{q=1}^{2S} c_q \mathbb{1}(\Delta'_{l_q})$  where  $|c_q| \leq \mathbf{M}_{\{g\}}$ . Since each  $\Delta'_{l_q}$  belongs to at most one  $\mathcal{X}_{j'-N,k}$ ,  $\tilde{\gamma}_{j',m'}(\mathbb{1}(\Delta'_{l_q}), r) = 0$  if  $\Delta'_{l_q}$  is not contained in  $\mathcal{X}_{j'-N,m'}$  and  $|\tilde{\gamma}_{j',m'}(\mathbb{1}(\Delta'_{l_q}), r)| \leq C_{v,\alpha} 2^{-l+1}$  if  $\Delta'_{l_q} \subseteq \mathcal{X}_{j'-N,m'}$  where  $C_{v,\alpha} = v(1 + (2\sqrt{\alpha})^\alpha)$ . For  $j'$  such that  $N \leq j' \leq j$ ,

$$\sum_{m': \mathcal{C}_{j',m'} \subseteq \mathcal{C}_{j,k}} |\tilde{\gamma}_{j',m'}(g, r)|^2 \leq 2S \sum_{q=1}^{2S} \sum_{m': \mathcal{C}_{j',m'} \subseteq \mathcal{C}_{j,k}} (c_q \tilde{\gamma}_{j',m'}(\mathbb{1}(\Delta_{l_q}), r))^2 \leq 2C_{v,\alpha}^2 S \sum_{q=1}^{2S} c_q^2 2^{-2l} \leq 4C_{v,\alpha}^2 S^2 \mathbf{M}_{\mathcal{G}}^2 2^{-2l}.$$

For  $0 \leq j' \leq j$ ,

$$\begin{aligned} & \sum_{k': \mathcal{C}_{j',k'} \subseteq \mathcal{C}_{j,k}} |\tilde{\gamma}_{j',k'}(g, r)| \\ & = \sum_{l: \mathcal{X}_{0,l} \subseteq \mathcal{X}_{j-N,k}} \sum_{0 \leq m < 2^{j'}} |\mathbb{E}[g(\mathbf{x}_i) | \mathbf{x}_i \in \mathcal{X}_{0,l}] \cdot |\mathbb{E}[r(y_i) | \mathbf{x}_i \in \mathcal{X}_{0,l}, y_i \in \mathcal{Y}_{l,j-1,2m}] \\ & \quad - \mathbb{E}[r(y_i) | \mathbf{x}_i \in \mathcal{X}_{0,l}, y_i \in \mathcal{Y}_{l,j-1,2m+1}]| \\ & \leq C_{v,\alpha} \sum_{l: \mathcal{X}_{0,l} \subseteq \mathcal{X}_{j-N,k}} |\mathbb{E}[g(\mathbf{x}_i) | \mathbf{x}_i \in \mathcal{X}_{0,l}]| N^\alpha \leq C_{v,\alpha} 2^{j-N} \mathbf{M}_{\mathcal{G}} N^\alpha. \end{aligned}$$

Since  $|\tilde{\gamma}_{l,m}(g, r)| \lesssim C_{v,\alpha} \mathbf{M}_{\mathcal{G}} N^\alpha$  for all  $(l, m)$ ,  $\sum_{k': \mathcal{C}_{j',k'} \subseteq \mathcal{C}_{j,k}} \tilde{\gamma}_{j',k'}^2(g, r) \leq C_{v,\alpha}^2 2^{j-N} \mathbf{M}_{\mathcal{G}}^2 N^{2\alpha}$ . Putting together

$$\sum_{j' < j} (j - j')(j - j' + 1) 2^{j'-j} \sum_{k': \mathcal{C}_{j',k'} \subseteq \mathcal{C}_{j,k}} \tilde{\gamma}_{j',k'}^2(g, r) \lesssim C_{v,\alpha}^2 S^2 \mathbf{M}_{\mathcal{G}}^2 + C_{v,\alpha}^2 \mathbf{M}_{\mathcal{G}}^2 N^{2\alpha}.$$



**Case 2:**  $l < N$  Hence for any  $0 \leq j' \leq j$ , we have

$$\begin{aligned} \sum_{k': \mathcal{C}_{j',k'} \subseteq \mathcal{C}_{j,k}} |\tilde{\gamma}_{j',k'}(g, r)| &= |\mathbb{E}[g(\mathbf{x}_i) | \mathbf{x}_i \in \mathcal{X}_{0,l}]| \sum_{m': \mathcal{Y}_{l,j',m'} \subseteq \mathcal{Y}_{l,j,m}} |\mathbb{E}[r(y_i) | \mathbf{x}_i \in \mathcal{X}_{0,l}, y_i \in \mathcal{Y}_{l,j-1,2m}]| \\ &\quad - |\mathbb{E}[r(y_i) | \mathbf{x}_i \in \mathcal{X}_{0,l}, y_i \in \mathcal{Y}_{l,j-1,2m+1}]| \\ &\leq C_{v,\alpha} |\mathbb{E}[g(\mathbf{x}_i) | \mathbf{x}_i \in \mathcal{X}_{0,l}]| N^\alpha \leq C_{v,\alpha} M_{\mathcal{G}} N^\alpha. \end{aligned}$$

It follows that

$$\sum_{j' < j} (j - j')(j - j' + 1) 2^{j'-j} \sum_{k': \mathcal{C}_{j',k'} \subseteq \mathcal{C}_{j,k}} |\tilde{\gamma}_{j',k'}(g, r)| \leq C_{v,\alpha} M_{\mathcal{G}} N^\alpha.$$

It follows that

$$\begin{aligned} \mathbf{C}_{\Pi_1(\mathcal{G} \times \mathcal{R}), M+N} &= \sup_{h \in \mathcal{H}} \min \left\{ \sup_{(j,k)} \left[ \sum_{l < j} (j - l)(j - l + 1) 2^{l-j} \sum_{m: \mathcal{C}_{l,m} \subseteq \mathcal{C}_{j,k}} \tilde{\gamma}_{l,m}^2(h) \right], \mathbf{M}_{\Pi_1(\mathcal{G} \times \mathcal{R})}^2(M + N) \right\} \\ &\leq C_{v,\alpha}^2 M_{\mathcal{G}}^2 N^{2\alpha} \min\{M + N, S^2 + 1\}. \end{aligned}$$

By the characterization of projections in Equation SA-23, we know the mis-specification error is zero, that is,  $\Pi_1 G_n(g, r) = \Pi_0 G_n(g, r)$  and  $\Pi_1 Z_n^G(g, r) = \Pi_0 Z_n^G(g, r)$ . Since  $g$  is already piecewise-constant on  $\mathcal{X}_{0,l}$ 's, the  $L_2$ -projection error is solely contributed from  $r$ . Consider  $\mathcal{B} = \sigma(\{\mathbb{1}_{\mathcal{C}_{0,k}} : 0 \leq k < 2^{M+N+1}\})$ . Denote  $r_\tau = r|_{[-\tau^{1/\alpha}, \tau^{1/\alpha}]}$ . Then

$$|\mathbb{E}[g(\mathbf{x}_i) r_\tau(y_i) | \mathcal{B}] - g(\mathbf{x}_i) r_\tau(y_i)| \leq M_{\mathcal{G}} |r_\tau(y_i) - \mathbb{E}[r_\tau(y_i) | \mathcal{B}]|.$$

Then by the same argument as in the proof for Lemma SA.20 and the argument for truncation error in the proof for Lemma SA.21, for all  $t > N$ ,

$$\mathbb{P} \left( \|G_n - \Pi_1 G_n\|_{(\mathcal{G} \times \mathcal{R})_\delta} + \|Z_n^G - \Pi_1 Z_n^G\|_{(\mathcal{G} \times \mathcal{R})_\delta} \geq N \sqrt{2^{-N} M_{\mathcal{G}}^2} t^{\alpha + \frac{1}{2}} + \frac{M_{\mathcal{G}}}{\sqrt{n}} t^{\alpha+1} \right) \leq 4N_{\mathcal{G} \times \mathcal{R}}(\delta, M_{\mathcal{G}} M_{\mathcal{R}}) n e^{-t}. \quad (\text{SA-24})$$

Then apply Lemma SA.18, we get there exists a mean-zero Gaussian process  $Z_n^G$  with the same covariance structure as  $G_n$  such that with probability at least  $1 - 2 \exp(-t) - 2^{M+1} \exp(-C_\rho n 2^{-M-1})$ ,

$$\begin{aligned} \|\Pi_1 G_n - \Pi_1 Z_n^G\|_{\mathcal{G} \times \mathcal{R}} &\leq C_\rho \min_{\delta \in (0,1)} \left\{ \sqrt{\frac{2^{M+2} M_{\mathcal{G}} E_{\mathcal{G}}}{n}} (t + \log N_{\mathcal{G} \times \mathcal{R}}(\delta, M_{\mathcal{G}} M_{\mathcal{R}}))^{\alpha + \frac{1}{2}} \right. \\ &\quad \left. + \sqrt{\frac{\mathbf{C}_{\Pi_1(\mathcal{G} \times \mathcal{R}), M+N}}{n}} (t + \log N_{\mathcal{G} \times \mathcal{R}}(\delta, M_{\mathcal{G}} M_{\mathcal{R}}))^{\alpha+1} + F_n(t, \delta) \right\}, \end{aligned}$$

where  $C_\rho > 0$  is a constant that only depends on  $\rho$ . □

The following theorem presents a generalization of Theorem 4 in the paper. To simplify notation, the parameters of  $\mathcal{G}$  and  $\mathcal{G} \cdot \mathcal{V}_{\mathcal{R}}$  (Definitions 4 to 12, SA.1, SA.2) are taken with  $\mathcal{C} = \mathcal{Q}_{\mathcal{G}}$ , and the index  $\mathcal{Q}_{\mathcal{G}}$  is omitted where there is no ambiguity; the parameters of  $\mathcal{R}$  (Definitions 4 to 12) are taken with  $\mathcal{C} = \mathcal{Y}$ , and the index  $\mathcal{Y}$  is omitted where there is no ambiguity; and the parameters of  $\mathcal{G} \times \mathcal{R}$  (Definitions 4 to 12, SA.3,

SA.4) are taken with  $\mathcal{C} = \mathcal{Q}_{\mathcal{G}} \times \mathcal{Y}$ , and the index  $\mathcal{Q}_{\mathcal{G}} \times \mathcal{Y}$  is omitted where there is no ambiguity.

**Theorem SA.4.** *Suppose  $(\mathbf{z}_i = (\mathbf{x}_i, y_i) : 1 \leq i \leq n)$  are i.i.d. random vectors taking values in  $(\mathbb{R}^{d+1}, \mathcal{B}(\mathbb{R}^{d+1}))$ , 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)  $\mathcal{G} \subseteq \text{Span}\{\mathbb{1}_{\Delta_l} : 0 \leq l < L\}$  is a class of Haar functions on  $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P}_X)$ .
- (ii) There exists a surrogate measure  $\mathbb{Q}_{\mathcal{G}}$  for  $\mathbb{P}_X$  with respect to  $\mathcal{G}$  such that  $\{\Delta_l : 0 \leq l < L\}$  forms a quasi-uniform partition of  $\mathcal{Q}_{\mathcal{G}}$  with respect to  $\mathbb{Q}_{\mathcal{G}}$ :

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

- (iii)  $\mathcal{G}$  is a VC-type class with envelope function  $M_{\mathcal{G}}$  over  $\mathcal{Q}_{\mathcal{G}}$  with  $c_{\mathcal{G}} \geq e$  and  $d_{\mathcal{G}} \geq 1$ .
- (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  $c_{\mathcal{R}, \mathcal{Y}} \geq e$  and  $d_{\mathcal{R}, \mathcal{Y}} \geq 1$ , where  $M_{\mathcal{R}, \mathcal{Y}}(y) + \text{pTV}_{\mathcal{R}, (-|y|, |y|)} \leq v(1 + |y|^\alpha)$  for all  $y \in \mathcal{Y}$ , for some  $v > 0$ , and for some  $\alpha \geq 0$ . Furthermore, if  $\alpha > 0$ , then  $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(|y_i|) | \mathbf{x}_i = \mathbf{x}] \leq 2$ .

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

- $\mathbb{E}[R_n(g_1, r_1)R_n(g_2, r_2)] = \mathbb{E}[Z_n^R(g_1, r_1)Z_n^R(g_2, r_2)]$  for all  $(g_1, r_1), (g_2, r_2) \in \mathcal{G} \times \mathcal{R}$ , and
- $\mathbb{P}[\|R_n - Z_n^R\|_{\mathcal{G} \times \mathcal{R}} > C_1 C_{v, \alpha} C_\rho \min_{\delta \in (0, 1)} (\mathbf{H}_n^R(t, \delta) + \mathbf{F}_n^R(t, \delta)) + \mathbf{W}_n(t)] \leq C_2 e^{-t} + L e^{-C_\rho n/L}$  for all  $t > 0$ ,

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

$$\begin{aligned} \mathbf{H}_n^R(t, \delta) &= \sqrt{\frac{LM_{\mathcal{G}}E_{\mathcal{G}}}{n}} (t + \log N_{\mathcal{G}}(\delta/2) + \log N_{\mathcal{R}}(\delta/2) + \log_2 N^*)^{\alpha + \frac{1}{2}} \\ &\quad + \frac{M_{\mathcal{G}}}{\sqrt{n}} (\log n)^\alpha (t + \log N_{\mathcal{G}}(\delta/2) + \log N_{\mathcal{R}}(\delta/2) + \log_2 N^*)^{\alpha + 1}, \\ \mathbf{W}_n(t) &= \mathbb{1}(|\mathcal{R}| > 1) \sqrt{M_{\mathcal{G}}E_{\mathcal{G}}} \left( \max_{0 \leq l < L} \|\Delta_l\|_\infty \right) L_{\mathcal{V}_{\mathcal{R}}} \sqrt{t + \log N_{\mathcal{G}}(\delta/2) + \log N_{\mathcal{R}}(\delta/2) + \log_2 N^*}. \end{aligned}$$

with  $\mathcal{V}_{\mathcal{R}} = \{\theta(\cdot, r) : \mathbf{x} \mapsto \mathbb{E}[r(y_i) | \mathbf{x}_i = \mathbf{x}], \mathbf{x} \in \mathcal{X}, r \in \mathcal{R}\}$  and  $N^* = \lceil \log_2(\frac{nM_{\mathcal{G}}}{2LE_{\mathcal{G}}}) \rceil$ .

**Proof of Theorem SA.4.** By the same reduction through surrogate measure, we can w.l.o.g. assume  $\mathbb{Q}_{\mathcal{G}} = \mathbb{P}_X$ . Suppose  $2^M \leq J < 2^{M+1}$ . By the same cell divisions in the proof for Theorem SA.3, there exists a quasi-dyadic expansion  $\mathcal{C}_{M+1, N}$  such that

$$\text{Span}(\{\mathbb{1}(\Delta_j) : 0 \leq j < J\}) \subseteq \text{Span}(\{\mathbb{1}(\mathcal{X}_{0,l}) : 0 \leq l < 2^{M+1}\}).$$

By definition, the projection error can be decomposed as

$$R_n(g, r) - \Pi_2 R_n(g, r) = G_n(g, r) - \Pi_1 G_n(g, r) + X_n(g\theta(\cdot, r)) - \Pi_0 X_n(g\theta(\cdot, r)),$$

where  $\Pi_0$  denotes the  $L_2$ -projection from  $L_2(\mathbb{R}^d)$  to  $\text{Span}(\{\mathbb{1}(\mathcal{X}_{0,l}) : 0 \leq l < 2^{M+1}\})$ . For any  $g \in \mathcal{G}$ , since  $g \in \text{Span}(\{\mathbb{1}(\mathcal{X}_{0,l}) : 0 \leq l < 2^{M+1}\})$ ,

$$\begin{aligned} \mathbb{E} [(X_n(g\theta(\cdot, r)) - \Pi_0 X_n(g\theta(\cdot, r)))^2] &= \sum_{0 \leq j < J} \mathbb{P}_X(\Delta_j) g^2|_{\Delta_j} \mathbb{E} [(\theta(\mathbf{x}_i, \mathbf{x}) - \Pi_0 \theta(\mathbf{x}_i, \mathbf{x}))^2 | \mathbf{x}_i \in \Delta_j] \\ &\leq \mathbb{E}[g(\mathbf{x}_i)^2] \max_{0 \leq j < J} \|\Delta_j\|_{\infty}^2 \mathbb{L}_{\mathcal{V}_{\mathcal{R}}}^2 \\ &\leq \mathbb{M}_{\mathcal{G}} \mathbb{E}_{\mathcal{G}} \max_{0 \leq j < J} \|\Delta_j\|_{\infty}^2 \mathbb{L}_{\mathcal{V}_{\mathcal{R}}}^2. \end{aligned}$$

Then  $X_n(g\theta(\cdot, r)) - \Pi_0 X_n(g\theta(\cdot, r))$  is bounded through Bernstein inequality and union bound, for all  $t > 0$ ,

$$\mathbb{P} \left( \|X_n(g\theta(\cdot, r)) - \Pi_0 X_n(g\theta(\cdot, r))\|_{(\mathcal{G} \times \mathcal{R})_{\delta}} \geq \frac{4}{3} \sqrt{\mathbb{M}_{\mathcal{G}} \mathbb{E}_{\mathcal{G}}} \max_{0 \leq j < J} \|\Delta_j\|_{\infty} \mathbb{L}_{\mathcal{V}_{\mathcal{R}}} \sqrt{t} + 2C_{v,\alpha} \frac{\mathbb{M}_{\mathcal{G}}}{\sqrt{n}} t \right) \leq 2 \exp(-t).$$

Combining Lemma SA.18 and Equation (SA-24), and the same calculation as in the proof for Theorem SA.2 to get  $\mathbb{C}_{\Pi_2(\mathcal{G}, \mathcal{R})} \lesssim (C_{v,\alpha} \mathbb{M}_{\mathcal{G}} N^{\alpha})^2$ , for all  $t > N_*$ , with probability at least  $1 - 2\mathbb{N}_{\mathcal{G} \times \mathcal{R}}(\delta, \mathbb{M}_{\mathcal{G}} M_{\mathcal{R}}) e^{-t} - 2^M \exp(-C_{\rho} n 2^{-M})$ ,

$$\|R_n - Z_n^R\|_{(\mathcal{G} \times \mathcal{R})_{\delta}} \leq \frac{4}{3} \sqrt{\mathbb{M}_{\mathcal{G}} \mathbb{E}_{\mathcal{G}}} \max_{0 \leq j < J} \|\Delta_j\|_{\infty} \mathbb{L}_{\mathcal{V}_{\mathcal{R}}} \sqrt{t} + C_{v,\alpha} N_*^{\alpha + \frac{1}{2}} \sqrt{\frac{J \mathbb{E}_{\mathcal{G}} \mathbb{M}_{\mathcal{G}}}{n}} \sqrt{t} + C_{v,\alpha} \frac{\mathbb{M}_{\mathcal{G}}}{\sqrt{n}} t^{\alpha+1},$$

The rest follows from the error for fluctuation off the  $\delta$ -net given in Lemma SA.16. The ‘‘bias’’ term  $\sqrt{\mathbb{M}_{\mathcal{G}} \mathbb{E}_{\mathcal{G}}} \max_{0 \leq j < J} \|\Delta_j\|_{\infty} \mathbb{L}_{\mathcal{V}_{\mathcal{R}}} \sqrt{t}$  comes from  $X_n(g\theta(\cdot, r)) - \Pi_0 X_n(g\theta(\cdot, r))$  in the decomposition.

In the special case that we have a singleton  $\mathcal{R} = \{r\}$ , we can get rid of the ‘‘bias’’ term by redefining  $\varepsilon_i = \text{sign}(r(y_i) - \mathbb{E}[r(y_i) | \mathbf{x}_i]) |r(y_i) - \mathbb{E}[r(y_i) | \mathbf{x}_i]|^{1/\alpha}$ . Take  $\tilde{r}(u) = \text{sign}(u) |u|^{\alpha}$ ,  $u \in \mathbb{R}$ . In particular,  $\mathbb{E}[\tilde{r}(\varepsilon_i) | \mathbf{x}_i] = 0$  almost surely. Either  $r$  is bounded and we can take  $\alpha = 0$ , which makes  $\tilde{r}$  also bounded; or  $\alpha > 0$  and  $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(|y_i|) | \mathbf{x}_i = \mathbf{x}] \leq 2$  and  $|r(u)| \lesssim 1 + |u|^{\alpha}$ , which implies  $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(|\varepsilon_i|) | \mathbf{x}_i = \mathbf{x}] \lesssim 2$  and  $\tilde{r}$  has polynomial growth. Then for any  $g \in \mathcal{G}$ ,

$$R_n(g, r) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g(\mathbf{x}_i) \tilde{r}(\varepsilon_i) - \mathbb{E}[g(\mathbf{x}_i) \tilde{r}(\varepsilon_i)] = G'_n(g, \tilde{r}),$$

where  $G'_n$  denotes the empirical process based on random sample  $((\mathbf{x}_i, \varepsilon_i) : 1 \leq i \leq n)$ . The result then follows from Theorem SA.3. By similar arguments as in the proof of Theorem SA.4,

$$\mathbb{C}_{\Pi_1(\mathcal{G}, \{\tilde{r}\})} = \sup_{f \in \Pi_1(\mathcal{G}, \{\tilde{r}\})} \min \left\{ \sup_{(j,k)} \left[ \sum_{j' < j} (j - j')(j - j' + 1) 2^{j'-j} \sum_{k' : \mathcal{C}_{j', k'} \subseteq \mathcal{C}_{j,k}} \tilde{\beta}_{j', k'}^2(f) \right], \|f\|_{\infty}^2 (M + N) \right\},$$

but  $\tilde{\beta}_{j,k}(f)$  vanishes for all  $j > N$  and we obtain similarly  $\mathbb{C}_{\Pi_1(\mathcal{G}, \{\tilde{r}\})} \lesssim (C_{v,\alpha} \mathbb{M}_{\mathcal{G}} N^{\alpha})^2$ .  $\square$

#### SA-V.2.1 Proof of Theorem 4

By standard empirical process arguments,  $\mathbb{N}_{\mathcal{G}}(\delta) \leq c_{\mathcal{G}} \delta^{-d_{\mathcal{G}}}$  and  $\mathbb{N}_{\mathcal{R}}(\delta) \leq c_{\mathcal{R}} \delta^{-d_{\mathcal{R}}}$  for  $\delta \in (0, 1]$ , and the result follows by Lemma SA.4.  $\square$

### SA-V.2.2 Proof of Corollary 6

Take  $t = C \log n$  with  $C > 1$  in Theorem 4. □

### SA-V.3 Example: Haar Partitioning-based Regression

The following lemma gives precise regularity conditions for the example in Section 5.3 of the paper.

**Lemma SA.31** (Haar Basis Regression Estimators). *Consider the setup in Example 3, and assume in addition that  $\sup_{r \in \mathcal{R}_\ell} \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} |\theta(\mathbf{x}, r) - \theta(\mathbf{y}, r)| / \|\mathbf{x} - \mathbf{y}\|_\infty < \infty$  for  $\ell = 1, 2$ .*

*If  $\log(nL)L/n \rightarrow 0$ , then*

$$\begin{aligned} \sup_{r \in \mathcal{R}_2} \sup_{\mathbf{w} \in \mathcal{W}} |\mathbf{p}(\mathbf{w})^\top (\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1}) \mathbf{T}_r| &= O(\log(nL)L/n) \quad a.s., \quad \text{and} \\ \sup_{r \in \mathcal{R}_\ell} \sup_{\mathbf{w} \in \mathcal{W}} |\mathbb{E}[\check{\theta}(\mathbf{w}, r) | \mathbf{x}_1, \dots, \mathbf{x}_n] - \theta(\mathbf{w}, r)| &= O\left(\max_{0 \leq l < L} \|\Delta_l\|_\infty\right) \quad a.s., \quad l = 1, 2. \end{aligned}$$

*If, in addition,  $\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\exp(|y_i|) | \mathbf{x}_i = \mathbf{x}] \leq 2$ , then*

$$\sup_{r \in \mathcal{R}_2} \sup_{\mathbf{w} \in \mathcal{W}} |\mathbf{p}(\mathbf{w})^\top (\widehat{\mathbf{Q}}^{-1} - \mathbf{Q}^{-1}) \mathbf{T}_r| = O(\log(nL)L/n + (\log n)(\log(nL)L/n)^{3/2}) \quad a.s.$$

**Proof of Lemma SA.31.** We use the notation  $\mathbb{P}_X(\Delta_l) = \mathbb{P}(\mathbf{x}_i \in \Delta_l)$ , and  $\widehat{\mathbb{P}}_X(\Delta_l) = n^{-1} \sum_{i=1}^n \mathbb{1}(\mathbf{x}_i \in \Delta_l)$ ,  $0 \leq l < L$ .

**Non-linearity Errors:** For  $\ell = 1, 2$ ,  $\mathbf{w} \in \mathcal{W}$ ,  $r \in \mathcal{R}_\ell$ , we have

$$\mathbf{p}(\mathbf{w})^\top (\widehat{\mathbf{J}}^{-1} - \mathbf{J}^{-1}) \mathbf{T}_r = \sum_{0 \leq l < L} \mathbb{1}(\mathbf{w} \in \Delta_l) (L^{-1} \widehat{\mathbb{P}}_X(\Delta_l)^{-1} - L^{-1} \mathbb{P}_X(\Delta_l)^{-1}) \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}(\mathbf{x}_i \in \Delta_l)}{L^{-1}} \epsilon_i(r),$$

where  $\epsilon_i(r) = r(y_i) - \mathbb{E}[r(y_i) | \mathbf{x}_i]$ . By maximal inequality for sub-Gaussian random variables (van der Vaart and Wellner, 2013, Lemma 2.2.2),  $\max_{0 \leq l < L} |L \widehat{\mathbb{P}}_X(\Delta_l) - L \mathbb{P}_X(\Delta_l)| = O(\sqrt{\frac{\log(nL)}{n/L}})$  a.s.. Since  $\{\Delta_l : 0 \leq l < L\}$  is a quasi-uniform partition of  $\mathcal{X}$  with respect to  $\mathbb{P}_X$ ,  $\min_{0 \leq l < L} L \mathbb{P}_X(\Delta_l) = \Omega(1)$ . Hence

$$\max_{0 \leq l < L} |L^{-1} \widehat{\mathbb{P}}_X(\Delta_l)^{-1} - L^{-1} \mathbb{P}_X(\Delta_l)^{-1}| = O(\sqrt{(n/L)^{-1} \log(nL)}), \quad a.s.. \quad (\text{SA-25})$$

Take  $\mathcal{H}_\ell = \{(\mathbf{w}, y) \mapsto L \mathbb{1}(\mathbf{w} \in \Delta_l)(r(y) - \theta(\mathbf{w}, r)) : 0 \leq l < L, r \in \mathcal{R}_\ell\}$ , for  $\ell = 1, 2$ . In particular, if we take  $\mathcal{G} = \{L \mathbb{1}(\cdot \in \Delta_l) : 0 \leq l < L\}$ , then  $\mathcal{G}$  is a VC-type class w.r.p. constant envelope  $L$  with constant  $\mathbf{c}_\mathcal{G} = L$  and exponent  $\mathbf{d}_\mathcal{G} = 1$ . In the main text, we explained that both  $\mathcal{R}_1$  and  $\mathcal{R}_2$  are VC-type class with  $\mathbf{c}_{\mathcal{R}_1} = 1$ ,  $\mathbf{d}_{\mathcal{R}_1} = 1$  and  $\mathbf{c}_{\mathcal{R}_2}$  some universal constant,  $\mathbf{d}_{\mathcal{R}_2} = 2$ . By standard empirical process arguments, both  $\mathcal{H}_\ell$ 's are VC-type class with  $\mathbf{c}_{\mathcal{H}_1} = L$ ,  $\mathbf{d}_{\mathcal{H}_1} = 1$ ,  $\mathbf{c}_{\mathcal{H}_2} = O(L)$ ,  $\mathbf{d}_{\mathcal{H}_2} = 2$ . Since  $\sup_{r \in \mathcal{R}_\ell} \max_{0 \leq l < L} \frac{1}{n} \sum_{i=1}^n L \mathbb{1}(\mathbf{x}_i \in \Delta_l) \epsilon_i(r) = \sup_{h \in \mathcal{H}_\ell} |\mathbb{E}_n[h(\mathbf{x}_i, y_i)] - \mathbb{E}[h(\mathbf{x}_i, y_i)]|$  is the suprema of empirical process, by Corollary 5.1 in Chernozhukov *et al.* (2014),

$$\begin{aligned} \sup_{r \in \mathcal{R}_1} \max_{0 \leq l < L} \left| \frac{1}{n} \sum_{i=1}^n L \mathbb{1}(\mathbf{x}_i \in \Delta_l) \epsilon_i(r) \right| &= O\left(\sqrt{\frac{\log(nL)}{n/L}} + \log(n) \frac{\log(nL)}{n/L}\right) \quad a.s., \\ \sup_{r \in \mathcal{R}_2} \max_{0 \leq l < L} \left| \frac{1}{n} \sum_{i=1}^n L \mathbb{1}(\mathbf{x}_i \in \Delta_l) \epsilon_i(r) \right| &= O\left(\sqrt{\frac{\log(nL)}{n/L}}\right) \quad a.s.. \end{aligned} \quad (\text{SA-26})$$

Putting together Equations (SA-25), (SA-26), we have

$$\sup_{\mathbf{w} \in \mathcal{W}} \sup_{r \in \mathcal{R}_\ell} \left| \mathbf{p}(\mathbf{w})^\top (\widehat{\mathbf{J}}^{-1} - \mathbf{J}^{-1}) \mathbf{T}_r \right| = O\left(\frac{\log(nL)}{n/L}\right) + \mathbb{1}(\ell = 1) O\left(\log(n) \left(\frac{\log(nL)}{n/L}\right)^{3/2}\right).$$

**Smoothing Bias:** Since we have assumed that  $\sup_{r \in \mathcal{R}_\ell} \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} |\mu(\mathbf{x}, r) - \mu(\mathbf{y}, r)| / \|\mathbf{x} - \mathbf{y}\|_\infty < \infty$ ,  $\ell = 1, 2$ ,

$$\begin{aligned} \sup_{\mathbf{x} \in \mathcal{X}} \sup_{r \in \mathcal{R}_\ell} |\mathbb{E}[\widehat{\mu}(\mathbf{x}, r) | \mathbf{x}_1, \dots, \mathbf{x}_n] - \mu(\mathbf{x}, r)| &= \left| \sum_{0 \leq l < L} \mathbb{1}(\mathbf{x} \in \Delta_l) \frac{\sum_{i=1}^n \mathbb{1}(\mathbf{x}_i \in \Delta_l) \mu(\mathbf{x}_i, r)}{\sum_{i=1}^n \mathbb{1}(\mathbf{x}_i \in \Delta_l)} - \mu(\mathbf{x}, r) \right| \\ &= O\left(\max_{0 \leq l < L} \|\Delta_l\|_\infty\right). \end{aligned}$$

□

## References

- Adamczak, R. (2008). “A tail inequality for suprema of unbounded empirical processes with applications to Markov chains,” *Electronic Journal of Probability*, 13(34), 1000–1034.
- Ambrosio, L., Fusco, N., and Pallara, D. (2000). *Functions of bounded variation and free discontinuity problems*: Oxford university press.
- Bretagnolle, J. and Massart, P. (1989). “Hungarian Constructions from the Nonasymptotic Viewpoint,” *Annals of Probability*, 17(1), 239–256.
- Brown, L. D., Cai, T. T., and Zhou, H. H. (2010). “Nonparametric regression in exponential families,” *Annals of Statistics*, 38(4), 2005–2046.
- Cattaneo, M. D., Chandak, R., Jansson, M., and Ma, X. (2024). “Local Polynomial Conditional Density Estimators,” *Bernoulli*, 30(4), 3193–3223.
- Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). “Gaussian approximation of suprema of empirical processes,” *Annals of Statistics*, 42(4), 1564–1597.
- Dudley, R. M. (2014). *Uniform central limit theorems, 142*: Cambridge university press.
- Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-dimensional Statistical Models*: Cambridge University Press.
- Rio, E. (1994). “Local Invariance Principles and Their Application to Density Estimation,” *Probability Theory and Related Fields*, 98(1), 21–45.
- Sakhanenko, A. (1996). “Estimates for the accuracy of coupling in the central limit theorem,” *Siberian Mathematical Journal*, 37(4), 811–823.
- van der Vaart, A. and Wellner, J. (2013). *Weak convergence and empirical processes: with applications to statistics*: Springer Science & Business Media.