Kernel Ridge Regression Inference
with Applications to Preference Data

Rahul Singh
Harvard University

Suhas Vijaykumar*
Amazon Science

Original draft: February 2023. This draft: October 2023

Abstract

We provide uniform inference and confidence bands for kernel ridge regression (KRR), a widely-used non-parametric regression estimator for general data types including rankings, images, and graphs. Despite the prevalence of these data—e.g., ranked preference lists in school assignment—the inferential theory of KRR is not fully known, limiting its role in economics and other scientific domains. We construct sharp, uniform confidence sets for KRR, which shrink at nearly the minimax rate, for general regressors. To conduct inference, we develop an efficient bootstrap procedure that uses symmetrization to cancel bias and limit computational overhead. To justify the procedure, we derive finite-sample, uniform Gaussian and bootstrap couplings for partial sums in a reproducing kernel Hilbert space (RKHS). These imply strong approximation for empirical processes indexed by the RKHS unit ball with logarithmic dependence on the covering number. Simulations verify coverage. We use our procedure to construct a novel test for match effects in school assignment, an important question in education economics with consequences for school choice reforms.

Keywords: Gaussian approximation, nonparametric regression, reproducing kernel Hilbert space, preference data, school choice

*We thank Alberto Abadie, Joshua Angrist, Victor Chernozhukov, Anna Mikusheva, Whitney Newey, Parag Pathak, and Vasilis Syrgkanis for helpful discussions. We are particularly grateful to Anna Mikusheva for guidance. Both authors received support from the Hausman Dissertation Fellowship, and part of this work was done while Rahul Singh visited the Simons Insitute for the Theory of Computing.

1 Introduction

We study a regularized, non-parametric regression estimator for general data—e.g. permutations, images, and graphs—called kernel ridge regression. Kernel ridge regression (KRR) is ubiquitous in data science, and several recent works advocate for its use in econometric problems.[1] However, its inferential theory is not fully known, limiting its role in empirical economic research. Our research question is how to compute sharp, uniform confidence bands for KRR, with arbitrary regressors.

Our results allow economists to flexibly conduct estimation and inference with ranked preference data. Individual preferences, truthfully reported as rankings over a set of choices, are increasingly available for empirical research due to the widespread adoption of strategy-proof matching mechanisms in education, medicine, and technology. However, they remain challenging for traditional econometric methods due to their massive ambient dimension.[2] Our method exploits latent, low-dimensional structure in ranked preferences, without needing to estimate an underlying choice model.

We apply these ideas to an important question in education economics: do students choose schools based upon "vertical" school quality or "horizontal" student-school compatibility? This question speaks to the welfare consequences of school choice reforms (Bau, 2022; Angrist et al., 2023). We provide a test that uses KRR to directly analyze ranked preference data from centralized school assignment, with valid inference.

1.1 Contributions

We provide uniform confidence bands for KRR with finite sample guarantees, allowing for non-parametric estimation and inference with general regressors. Our key assumptions, which we formalize in Section 3, are that (i) the effective dimension of the approximating basis is low, even if its ambient dimension is infinite; and (ii) the true regression function

[1] See e.g. Kasy (2018) for an application to optimal taxation, as well as Nie and Wager (2021) and Singh et al. (2023) for estimation of conditional average treatment effects.

[2] For example, an indicator for each possible preference over p schools would require $p!$ indicators.

f_0 is smooth. We highlight three aspects of our contribution.

First, we propose an efficient, *symmetrized* bootstrap process $\mathfrak{B}$ to sample from the distribution of $\sqrt{n}(\hat{f} - f_0)$. Here, $\hat{f}$ is the well-known KRR estimator and f_0 is the true regression function. We show that in finite samples, the process $\mathfrak{B}$ approximates the distribution of the entire KRR function in a uniform sense. Our proposal is practical: $\mathfrak{B}$ has a closed form solution and may be evaluated without re-computing the estimator. These properties follow from a symmetric construction that cancels bias, which appears new and simplifies both analysis and computation.[3] We provide inference results for any sup-norm continuous functional of $\sqrt{n}(\hat{f} - f_0)$, allowing for non-parametric inference and variable-width uniform confidence bands.

Second, to justify $\mathfrak{B}$, we derive non-asymptotic, uniform Gaussian couplings for partial sums in a reproducing kernel Hilbert space (RKHS). A major advantage of kernel methods is their flexibility with respect to new forms of data. By deriving Gaussian couplings directly in the RKHS, we avoid typical arguments that restrict the structure of the kernel and the form of the regressors. Formally, given a centered i.i.d. sequence $U = (U_1, U_2, \ldots)$ in an RKHS H, we construct a Gaussian random variable Z such that $\left\| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n U_i \right) - Z \right\|_H$ is small with high probability. This implies approximation in sup-norm when H has a bounded kernel.[4] We similarly construct bootstrap couplings, to sample from Z conditional upon data.

Finally, we illustrate how these results can provide insight into economic phenomena through the analysis of ranked choice data. A persistent question in the economics of education, particularly with regard to school choice reforms, concerns *match effects*: do student preferences reflect heterogeneous match quality? This question has significant implications for both the short-term welfare gains from school choice, and the long-term

[3]Our arguments rely only upon convergence of the sample covariance and thus generalize to other bootstrap procedures or efficient approximations thereof (i.e. sketches).

[4]This form of Gaussian approximation is stronger than approximation of the sup-norm, i.e. control of $\|n^{-1/2} \sum_{i=1}^n U_i\|_\infty - \|Z\|_\infty$(cf. Chernozhukov et al. 2014b, 2016; Chernozhuokov et al. 2022), yet in our setting we pay a smaller price in terms of model complexity parameters.

effect of choice reforms on education supply (Narita, 2018; Bau, 2022). In a synthetic exercise calibrated to ranked preference data from the Boston Public Schools system, we study whether the effects of pilot school attendance depend systematically on student preferences, which we interpret as a test for match effects. We find that KRR adapts to latent choice structure inherent in the data, and that our inferential procedure delivers a powerful, computationally efficient, non-parametric test for match effects.

The structure of this paper is as follows. Section 2 situates our contributions within the context of related work. Section 3 formalizes our inferential goal and our key assumptions. Section 4 proposes our inferential procedure and demonstrates its performance in simulations. Section 5 derives general results for partial sums, which we use to theoretically justify our procedure in Section 6. Section 7 presents the semi-synthetic exercise and discusses limitations. Section 8 concludes.

2 Related work

2.1 Gaussian approximation and confidence bands

A fundamental goal in non-parametric statistics is to construct uniform confidence bands for estimators. Many results have focused on density estimation where, under suitable conditions, one can verify the Donsker property and establish a central limit theorem for the deviations of the estimator (Giné and Nickl, 2008, 2009). One can then construct confidence bands using classical ideas of Bickel and Rosenblatt (1973).

In non-parametric regression, which we study, and in other inverse problems, the Donsker property often fails. In this case one may use *non-asymptotic* Gaussian couplings that hold in the absence of a stable Gaussian limit; see Remark 3.8. Here the so-called "Hungarian construction" of Komlós et al. (1975) has played a major role (see also Massart, 1989; Koltchinskii, 1994). However, applying these results in a sharp manner requires strong restrictions on the data (e.g. uniform distribution on the unit interval or cube, or a smooth density), which are not appropriate in our setting. Indeed,

flexibility with respect to the data type is a major advantage of RKHS methods.

Chernozhukov et al. (2014b, 2016, 2017) established a major breakthrough on both fronts. The authors developed nonasymptotic Gaussian couplings for maxima of sums with excellent dependence on the dimension, under weak conditions, by building on Stein's method (Stein, 1972). The authors applied these Gaussian couplings to nonparametric confidence bands via bootstrap couplings. Although we base our Gaussian approximation on a different method, we build on the arguments of Chernozhukov et al. (2014a, 2016) to establish bootstrap inference for KRR.

For Gaussian approximation, we build on results of Zaitsev (1987a,b) and Buzun et al. (2022) that give couplings in Euclidean norm with sharp dependence on the ambient dimension, applicable to a wide range of data types. Previously, Berthet and Mason (2006) (see also Rio, 1994; Einmahl and Mason, 1997) used these results in Banach spaces. Using an idea of Götze and Zaitsev (2011), we show that in Hilbert spaces, one can deploy these results with very good dependence on the covering number, comparable to Chernozhukov et al. (2016), while obtaining stronger forms of approximation.

2.2 Kernel methods and Tikhonov regularization

KRR and its variants generalize linear regression to an abstract setting while retaining computational and analytical tractability. Early work focused on spline models (Wahba, 1978) and support vector machines (Boser et al., 1992). In the abstract setting, minimax convergence rates for KRR in $L^2(\mathbb{P})$ are well established; see, e.g., Caponnetto and De Vito (2007); Smale and Zhou (2007); Mendelson and Neeman (2010), among many others, and van der Vaart and van Zanten (2008a,b) for related work on Gaussian process regression. More recently, Fischer and Steinwart (2020) establish minimax rates in interpolation spaces H' where $H \subseteq H' \subseteq L^2(\mathbb{P})$, which imply sup-norm rates.

Various authors have investigated Gaussian approximation for KRR and its variants. Hable (2012) establishes asymptotic Gaussian approximation in the "parametric" setting where the regularization parameter λ is bounded away from 0, which is at odds with

the consistency results cited above. For splines (or Sobolev RKHS), Shang and Cheng (2013) establish finite-sample Gaussian approximation where the data are uniform on $[0, 1]$. Our results complement those of Yang et al. (2017), who study posterior coverage and sup-norm credible sets in Gaussian process regression, under additional assumptions motivated by the spline setting. To our knowledge, our results are the first to provide valid, non-conservative inference in the fully non-parametric regime where the regularization parameter λ vanishes, applicable to the full range of settings in which KRR is applied in practice, e.g. ranking data.

Kernel methods have found many other uses in machine learning: computationally-efficient probability metrics for generative modeling, two-sample testing, and independence testing (Bach and Jordan, 2002; Gretton et al., 2005, 2012); structured prediction (Ciliberto et al., 2020); nonparametric causal inference (Nie and Wager, 2021; Singh et al., 2023); and several others. Motivated by density estimation, Sriperumbudur (2016) establishes asymptotic Gaussian approximation for sums in a restricted set of RKHSs using the Donsker property. Future work may use our nonasymptotic couplings to considerably strengthen and generalize results in these domains.

The ridge penalty in KRR is a type of Tikhonov regularization, which is widely used in the econometrics literature on ill-posed inverse problems. Formally, KRR replaces the traditional Tikhonov regularization over $L^2(\mathbb{P})$ with regularization over the RKHS—a stronger penalty, leading to more regular solutions. This stabilizes ill-posed inversion of the covariance operator. See e.g. Newey and Powell (2003); Hall and Horowitz (2005); Horowitz and Lee (2005); Carrasco et al. (2007); Darolles et al. (2011); Chen and Pouzo (2012); Singh et al. (2019), and references therein, for non-parametric instrumental variable regression estimators that employ Tikhonov regularization to address ill-posedness of inverting a conditional expectation operator.

3 Model and assumptions

3.1 General data types

We study non-linear regression where the covariates X_i may take values in an arbitrary space, S. For example, S could consist of rankings, images, or graphs. Kernel ridge regression extends linear regression to this abstract setting, while remaining practical.

Throughout the paper, our regression model H is a space of functions $f : S \to \mathbb{R}$, with additional structure. In particular, H is derived from a kernel function, $k : S \times S \to \mathbb{R}$. The functions $k_s(-) = k(s, -)$ span H, and $k(s, t) = \langle k_s, k_t \rangle$ defines an inner product in H. These conditions imply the reproducing property: $f(s) = \langle f, k_s \rangle$, for any $f \in H$.[5] We maintain that the kernel is bounded, i.e. $k(s, t) \leq \kappa^2$.

By analogy to classical regression, $k_{X_i}(-)$ may be viewed as an expansion of the data point X_i with respect to a non-linear basis. The kernel function also encodes similarity between data points. The implied norm in H, given by $\|f\| = \langle f, f \rangle^{\frac{1}{2}}$, encodes a corresponding notion of smoothness: similar points are assigned similar values.

In practice, we perform computations by evaluating the kernel, k, at pairs of data points. As such, we may perform regression in general covariate spaces S, provided we may find a suitable kernel. Due to the popularity of kernel methods, there is extensive theoretical and practical guidance on how to choose k. We consider some examples.[6]

Example 3.1 (Linear or polynomial kernel)**.** If $S = \mathbb{R}^p$, so that covariates X_i are finite dimensional vectors, then choosing $k(x, x') = x^\top x'$ recovers linear models: H is the set of linear functions $f_\beta(x) = \beta^\top x$ for $\beta \in \mathbb{R}^p$. Here, k_x corresponds to x, a trivial basis expansion. On the other hand, if $k(x, x') = (x^\top x' + 1)^d$, then the entries of k_x are

[5] Formally, we require that k is a positive-definite function, so that the inner-product is well defined, and we define the RKHS H to be the closure of $\{k_s \mid s \in S\}$ with respect to $\langle -, - \rangle$. We require S to be a separable, complete metric space, so that H is separable. We refer the reader to Berlinet and Thomas-Agnan (2004) for further background on the RKHS setting.

[6] For further examples, including the classical Sobolev spaces, we refer the reader to Rasmussen and Williams (2006). For network data, see Vishwanathan et al. (2010).

polynomials in x of degree at most d, and H contains all such polynomials.

Example 3.2 (Ranking data)**.** Suppose that S consists of rankings of 25 high schools in Boston. The set H of all functions $S \to \mathbb{R}$ now has dimension 25!, which is larger than 10^{25}. Nonetheless, if we take k to be the Mallows kernel, $k_{\mathrm{m}}(\pi, \pi') \coloneqq e^{-N(\pi,\pi')}$, where $N(\pi, \pi')$ counts the number of pairwise comparisons in which the rankings π and π' disagree, then we recover this H as the set of functions defined by k_{m} (Mania et al., 2018). Note that k_{m} may be easily computed from pairs of rankings. We provide further details when discussing our application in Section 7.

3.2 Regression setting

Consider an i.i.d. sequence of n data points (X_i, Y_i) in $S \times \mathbb{R}$, which are supported on a background probability space $(\mathbb{P}, \Omega, \mathcal{F})$[7]. Our goal is to learn the regression function $f_0 \in \operatorname{argmin}_{f \in H} \mathbb{E}[\{Y_i - f(X_i)\}^2]$. We consider the *kernel ridge regression* (KRR) estimator $\hat{f}$, given by

$$\hat{f} \in \operatorname*{argmin}_{f \in H} \left[\frac{1}{n} \sum_{i=1}^{n} \{Y_i - f(X_i)\}^2 + \lambda \|f\|^2 \right],$$

where $\lambda > 0$ is the regularization parameter and $\|f\| = \langle f, f \rangle^{\frac{1}{2}}$ denotes the norm in H. The solution to the optimization problem has a closed form (Kimeldorf and Wahba, 1971), which makes KRR practical:

$$\hat{f}(x) = K_x (K + n\lambda I)^{-1} Y.$$

This involves the kernel matrix $K = \{k(X_i, X_j)\}_{ij}$ in $\mathbb{R}^{n \times n}$, the kernel vector $K_x = \{k(x, X_1), \ldots k(x, X_n)\}$ in $\mathbb{R}^{1 \times n}$ and the outcome vector $Y = (Y_1, \ldots, Y_n)^\top$ in $\mathbb{R}^{n \times 1}$.

Example 3.1 (Linear kernel, continued)**.** Consider the linear kernel $k(s, t) = s^\top t$. Let X be the design matrix. Then the estimator is $\hat{f}(x) = x^\top X^\top (XX^\top + n\lambda I)^{-1} Y$, which is standard ridge regression. Here, K_x is $x^\top X^\top$ and K is the Gram matrix, $XX^\top$.

[7]We assume the probability space is sufficiently rich so that we may construct couplings, e.g. that it supports a countable sequence of i.i.d. Gaussians independent of the data.

Example 3.2 (Ranking data, continued)**.** For ranking data regression with the Mallows kernel, computing K requires computing $K_{ij} = k_{\mathrm{m}}(X_i, X_j)$ for each pair of rankings. Once K has been computed, estimation is similar to the linear case.

In general we replace X_i with k_{X_i}, which may be viewed as a dictionary of non-linear transformations applied to the data. Whereas OLS inverts the covariance matrix $X^\top X \in \mathbb{R}^{p\times p}$, we now invert $K \in \mathbb{R}^{n\times n}$, which is the Gram matrix of the transformed data, generalizing $XX^\top \in \mathbb{R}^{n\times n}$. By focusing on the latter, KRR allows the dictionary to be infinite-dimensional while retaining computational tractability.[8]

Ridge regularization ensures stability of the estimator even when its dimension exceeds the number of samples, but the regularization introduces bias. To analysis bias and variance, it helps to define the *pseudo-true parameter*

$$f_\lambda \in \operatorname*{argmin}_{f\in H} \left(\mathbb{E}[\{Y_i - f(X_i)\}^2] + \lambda \|f\|^2 \right).$$

We denote the regression error by $\varepsilon_i \coloneqq Y_i - f_0(X_i)$, and for simplicity we assume it is bounded: $|\varepsilon_i| \leq \bar{\sigma}$ almost surely. We write $D = \{(X_i, Y_i)\}_{i=1}^n$ for the observed sample.

3.3 Why is uniform inference challenging?

We would like a procedure that provides uniform inference for KRR, a widely used and flexible non-parametric regression estimator, while allowing the regularization to converge to zero. We would also like a procedure that handles general data such as rankings and that is easy to compute, departing from previous work. Finally, we would like the procedure to be valid and sharp, i.e. to provide coverage that is at least, but not much more than, the nominal level.[9]

Definition 3.3 (Validity)**.** The confidence sets $\hat{S}_n$ are τ-valid at level α if $\mathbb{P}(f_0 \in \hat{S}_n) \geq 1 - \alpha - \tau$.

[8] This is the celebrated "kernel trick;" see Aizerman (1964).

[9] In the non-parametrics literature, "validity" is also known as "honesty." See Section 6 for further discussion of sharpness, which modifies the usual definition of "exactness."

Definition 3.4 (Sharpness)**.** The confidence sets $\hat{S}_n$ are (δ, τ)-sharp at level α if for some positive $\delta, \tau > 0$, $\mathbb{P}\{f_0 \in \delta \hat{f} + (1-\delta)\hat{S}_n\} \leq 1 - \alpha + \tau$.

Why is uniform inference hard for KRR? Several challenges arise. First, our motivating interest in KRR is its adaptability to general regressors such as rankings; how can we provide results for many data types simultaneously? Our answer is to derive Gaussian and bootstrap couplings in the RKHS that adapt to the effective dimension, departing from previous work that places direct assumptions on the approximating basis. Second, a computationally intensive inference procedure may undermine the practicality of KRR; how can we avoid computing and inverting kernel matrices $K \in \mathbb{R}^{n \times n}$ at each bootstrap iteration? Our answer is to propose a symmetrized bootstrap with a closed form solution that re-uses the kernel evaluations and matrix inversions of KRR. Third, to cover f_0 we require $\lambda \downarrow 0$, but in this regime asymptotic central limit theorems based upon e.g. the Donsker property do not apply; how can we establish validity in the absence of a well-behaved Gaussian limit?[10] Our answer is to develop a nonasymptotic framework to transfer sharp, finite sample Gaussian and bootstrap coupling results to our proposed inferential procedure.

It is not obvious that we can choose a sequence $\lambda \downarrow 0$ that vanishes slowly enough for Gaussian approximation and yet quickly enough to cover f_0. Table 7 summarizes the corollaries of our main result, demonstrating that it is indeed possible in many settings. In fact, both can often be achieved alongside a near-optimal rate of estimation.[11]

3.4 Formal notation

We denote the Euclidean norm in $\mathbb{R}^n$ by $\|-\|_{\mathbb{R}^n}$. For $u, v \in H$, we denote by $u \otimes v^* : H \to H$ the tensor product, i.e. the rank one operator with $(u \otimes v^*)t = \langle v, t \rangle u$. For any $A : H \to H$ we use $\|A\|_{op}$ to denote the operator norm, $\|A\|_{\text{HS}}$ to denote the Hilbert-Schmidt (or Frobenius) norm, and $\operatorname{tr} A$ to denote the trace. If A is compact and

[10] Indeed, as $\lambda \downarrow 0$, the resulting stochastic process is not totally bounded in $L^2(\mathbb{P})$.

[11] Specifically, we consider minimax-optimal rates in H-norm.

self-adjoint, then H admits an orthonormal basis of A-eigenvectors $\{e_1(A), e_2(A), \ldots\}$ and corresponding eigenvalues $\{\nu_1(A), \nu_2(A), \ldots\}$. We suppress the operator A when it is clear from context.

We use C, C', etc. to denote sufficiently large, positive universal constants whose value may change across displays; $C(t)$ denotes a large enough number that depends only on the parameter t. Similarly, c, c', $c(t)$ denote sufficiently small positive quantities. We also use the notation $\lesssim$ (or $\lesssim_t$) to denote an inequality that holds up to a universal constant (or function of t). In summary, for $a, b > 0$, $a \lesssim b \iff a \leq Cb \iff ca \leq b$. We do not optimize such constants appearing in our analysis, but we do provide numerical constants wherever possible in the proof text.

We state our formal results in terms of a high probability parameter $\eta \in (0, 1)$; they hold on an event of probability at least $1 - \eta$. We also write $o_p(1)$ for quantities that converge to 0 for any fixed $\eta \in (0, 1)$.

3.5 Key assumption: Local width

To derive finite-sample Gaussian and bootstrap couplings for our estimator, we must limit model complexity. In the RKHS setting, this complexity is elegantly captured by the covariance operator.[12] We assume that the eigenvalues of this operator decay rapidly. This means that H has a low *effective dimension*: although it is an infinite dimensional space of functions, it has limited capacity to overfit the data.

Formally, let $U = (U_1, U_2, \ldots, U_n)$ be an i.i.d. sequence of n random variables taking values in H, such that $\mathbb{E}(U_i) = 0$ and $\mathbb{E}\|U_i\|^2 < \infty$, and let $\Sigma : H \to H$ defined by $\Sigma \coloneqq \mathbb{E}(U_i \otimes U_i^*)$ denote the associated covariance operator, which is self-adjoint and has finite trace. The complexity of Σ plays a central role in our results for partial sums. In particular, our bounds are in terms of the following quantity.

Definition 3.5 (Local width)**.** Given $m \geq 0$, the *local width* of Σ, written $\sigma(\Sigma, m)$, is

[12] By working with the covariance operator, we avoid placing explicit regularity conditions upon kernel or its eigenfunctions.

given by $\sigma^2(\Sigma, m) = \sum_{s>m} \nu_s(\Sigma)$, where the eigenvalues $\{\nu_1(\Sigma), \nu_2(\Sigma), \ldots\}$ are listed in decreasing order.

Remark 3.6 (Intuition)**.** The local width is the tail sum of eigenvalues. It quantifies how much of the covariance is not explained by the top eigenfunctions.[13] To provide further intuition it is helpful to consider the case $U_i = k_{X_i} - \mathbb{E}(k_X)$. In this case the eigenfunctions form an orthonormal basis of $L^2(\mathbb{P})$. If the local width is small, a small set of these basis functions may explain most of the variation in $f(X_i)$, for any $f \in H$. In other words, the model H is almost contained within a low-dimensional subspace of $L^2(\mathbb{P})$, hence it has low effective dimension. Equivalently, the data X_i have an approximate non-linear factor structure, with smooth factors (Kutateladze, 2022).

Remark 3.7 (Generality)**.** The local width $\sigma^2(\Sigma, m)$ converges to 0 for large m when $\mathbb{E}\|U_i\|^2 < \infty$, a very weak regularity condition that is satisfied in all of the settings we consider. Under polynomial decay of eigenvalues (see Section 6), our results allow $\sigma^2(\Sigma, m) \asymp m^{-c}$ for any positive $c > 0$, corresponding to large function classes H. Thus, the local width assumption primarily allows us to quantify rates of convergence, and does little to restrict the applicability of our results.

Remark 3.8 (Donsker property, empirical processes)**.** If $U_i = k_{X_i} - \mathbb{E}(k_{X_i})$, then partial sums correspond to the empirical processes $S_n(f) \coloneqq \frac{1}{n}\sum_i f(X_i) - \mathbb{E}f(X_i)$ indexed by $\|f\|_H \leq 1$.[14] In this context, the Donsker property often holds. For KRR, we show in Section 6 that $U_i = (T+\lambda)^{-1}\{(k_{X_i} \otimes k^*_{X_i} - T)(f_0 - f_\lambda) + \varepsilon_i k_{X_i}\}$, where $T = \mathbb{E}(k_{X_i} \otimes k^*_{X_i})$, which corresponds to a non-standard, *conditional* empirical processes (Stute, 1986). As $\lambda \downarrow 0$, the Donsker property does not hold because $(T+\lambda)^{-1}$ does not remain bounded. Therefore we provide non-asymptotic arguments via local width.

Remark 3.9 (Bounds on local width)**.** In Appendix I, we further characterize Definition 3.5 in leading cases: polynomial and exponential decay of eigenvalues. For these cases, we provide concrete upper bounds on local width in Appendix I.1.

[13] These eigenfunctions depend implicitly upon the kernel k and the distribution of the data.

[14] In the kernel methods literature, this setting corresponds to *kernel mean embeddings*.

The *entropy number* of the ellipsoid $\mathfrak{E} = \Sigma^{\frac{1}{2}} B$, where B is the unit ball in H, is $\mathrm{ent}_m(\mathfrak{E}) := \inf_{|A|\leq 2^m} \sup_{t\in\mathfrak{E}} \inf_{a\in A} \|t-a\|$. The set $\mathfrak{E}$ is isometric to the ellipsoid $\{\langle f, U_i\rangle \,|\, \|f\| \leq 1\} \subset L^2(\mathbb{P})$. We verify in Appendix I.2 that $\sigma(\Sigma, m)$ is roughly the local Gaussian complexity of $\mathfrak{E}$ at scale $\mathrm{ent}_m(\mathfrak{E})$, using the results of Wei et al. (2020). This facilitates comparison with other non-asymptotic Gaussian couplings.

3.6 Key assumption: Source condition

Our next key assumption, which is necessary to prove that our uniform confidence bands cover the true regression function f_0 despite bias, is that f_0 is a sufficiently smooth element of H. Formally, let $T : H \to H$ defined by $T := \mathbb{E}(k_{X_i} \otimes k^*_{X_i})$ denote the covariance operator induced by the data, with eigenfunctions $\{e_1(T), e_2(T), \ldots\}$ and corresponding eigenvalues $\{\nu_1(T), \nu_2(T), \ldots\}$. We quantify the complexity of f_0 with respect to the spectrum of T using the well-known source condition, which plays a central role in the minimax analysis of KRR both in L^2 and in H (Caponnetto and De Vito, 2007; Fischer and Steinwart, 2020).

Definition 3.10 (Source condition)**.** The true regression f_0 satisfies $f_0 \in H^r$ for some $r \in (1, 3]$, where we define $H^r \subseteq H \subseteq L^2$ as $H^r := \{f = \sum_{s=1}^{\infty} f_s e_s(T) : \sum_{s=1}^{\infty} f_s^2 \nu_s^{-r}(T) < \infty\}$.

Remark 3.11 (Intuition)**.** Taking $r = 0$ recovers square summability: $\sum_{s=1}^{\infty} f_s^2 < \infty$, which defines L^2. Taking $r = 1$ produces the condition $\sum_{s=1}^{\infty} f_s^2/\nu_s(T) < \infty$, which is equivalent to correct specification. For $r > 1$, the smoothness of the true regression f_0 exceeds the worst-case smoothness of H; f_0 is approximated well by the leading terms in the series $\{e_1(T), e_2(T), \ldots\}$. This notion of smoothness depends on the kernel and data. For example, we consider two rankings to be similar if they make similar pairwise comparisons, and define smoothness with respect to induced covariance T.

Example 3.12 (Sobolev space)**.** For example, denote by $\mathbb{H}^s_2$ the Sobolev space with $s > p/2$ square integrable derivatives over $[0, 1]^p$. If $H = \mathbb{H}^s_2$ and $f_0 \in \mathbb{H}^{s_0}_2$ then $r = s_0/s$ and $\mathbb{H}^{s_0}_2 = (\mathbb{H}^s_2)^r$.

Remark 3.13 (Bias control)**.** The source condition implies a bound on regularization bias: $\|f_\lambda - f_0\| \leq \kappa^{1-r}\lambda^{(r-1)/2}\|f_0\|$ (Smale and Zhou, 2005, Theorem 4). This bound is standard in the literature. It appears prominently in the proof of consistency and minimax rates for KRR, and we rely upon it to prove valid inference in Section 6.

4 Fixed and variable width confidence bands

4.1 Overview of the procedure

We state our procedure at a high level before filling in details.

- For each bootstrap iteration, draw Gaussians and compute $\mathfrak{B}(x)$ and $\hat{\sigma}^2(x)$.
- Across bootstrap iterations, compute the α-quantile, $\hat{t}_\alpha$, of $\sup_{x\in S}\left|n^{1/2}\hat{\sigma}(x)^{-1}\mathfrak{B}(x)\right|$.
- Calculate the band $\hat{C}_\alpha$ where $\hat{C}_\alpha(s) = \hat{f}(s) \pm \hat{t}_\alpha \cdot n^{-1/2}\hat{\sigma}(s)$ for $s \in S$.

This overall structure is familiar. Our contributions are to propose $\mathfrak{B}(x)$ and $\hat{\sigma}^2(x)$ in closed form, with low computational overhead, using symmetrization to correct regularization bias; our proposals depart from e.g. Yang et al. (2017), who provide frequentist analysis of the Bayesian posterior. We now fill in the details.

Estimator 4.1 (Uniform confidence band)**.** Given a sample $D = \{(X_i, Y_i)\}_{i=1}^n$, a kernel k, and regularization parameter $\lambda > 0$:

1. Compute the kernel matrix $K \in \mathbb{R}^{n\times n}$ with entries $K_{ij} \coloneqq k(X_i, X_j)$ and the kernel vector $K_x \in \mathbb{R}^{1\times n}$ with entries $k(x, X_i)$. Set $v_x^\top = K_x(K + n\lambda I)^{-1} \in \mathbb{R}^{1\times n}$.
2. Estimate KRR as $\hat{f}(x) = v_x^\top Y$ and compute the residual vector $\hat{\varepsilon} \in \mathbb{R}^n$ with entries $\hat{\varepsilon}_i = Y_i - \hat{f}(X_i)$. Set $\hat{\sigma}^2(x) = n\left\|v_x^\top \operatorname{diag}(\hat{\varepsilon})\right\|_{\mathbb{R}^n}^2$.
3. For each bootstrap iteration,
   (a) draw a matrix $h \in \mathbb{R}^{n\times n}$ of i.i.d. standard Gaussians;

(b) set $\mathfrak{B}(x) = v_x^\top \operatorname{diag}(\hat{\varepsilon})(h^\top - h)\mathbf{1}/\sqrt{2}$ where $\mathbf{1} \in \mathbb{R}^n$ is a vector of ones;

(c) compute $M = \sup_{x \in S} \left| n^{1/2} \hat{\sigma}(x)^{-1} \mathfrak{B}(x) \right|$.

4. Across bootstrap iterations, compute the α-quantile, $\hat{t}_\alpha$, of M.

5. Calculate the band $\hat{C}_\alpha$ where $\hat{C}_\alpha(s) = \hat{f}(s) \pm \hat{t}_\alpha \cdot n^{-1/2} \hat{\sigma}(s)$ for $s \in S$.

For fixed-width bands rather than variable-width bands, replace $\hat{\sigma}(x) n^{-1/2}$ with 1.

Remark 4.2 (Symmetry cancels bias)**.** Within the formula for $\hat{f}$, the quantity v_x incorporates the regularized inverse of K. For intuition, using the linear kernel $k(s,t) = s^\top t$, $v_x = (XX^\top + n\lambda I)^{-1} Xx$. Regularization ensures stability of the estimator and introduces bias. Our proposal $\mathfrak{B}(x)$ involves the anti-symmetric multipliers $(h^\top - h)/\sqrt{2}$ to cancel bias, which appears to be an innovation.

Remark 4.3 (Same overhead as KRR)**.** The primary computational cost of KRR comes from computing the vector $v_x \in \mathbb{R}^n$, which encodes *all* of the kernel evaluations as well as the inversion of the regularized kernel matrix. The same vector v_x of KRR appears in our proposed objects $\mathfrak{B}(x)$ and $\hat{\sigma}^2(x)$. Our procedure involves no additional kernel evaluations or matrix inversions beyond KRR. In this sense, the procedure is relatively efficient from a computational perspective. See Appendix F for the derivation and for an alternative $\hat{\sigma}^2(x)$ for small samples.

Remark 4.4 (Varying objects across iterations)**.** Across each bootstrap iteration, the only new object is a fresh draw of Gaussians h. Every other factor in the matrix multiplication that delivers $\mathfrak{B}(x)$ remains the same across iterations.

4.2 Robust performance with nonlinearity

We demonstrate that our procedure performs well in nonlinear simulations both with standard data and with non-standard data such as rankings.

To begin, we illustrate key concepts with a nonlinear regression simulation with standard data. By construction, the true regression function f_0 is the third eigenfunction of

the Gaussian kernel $k(x, x') = \exp\left\{-\frac{1}{2}\frac{(x-x')^2}{\iota^2}\right\}$, which is the third Hermite polynomial (Rasmussen and Williams, 2006, Section 4.3).[15] See Appendix K for details of the data generating process (DGP), results for the mis-specified setting $f_0 \notin H$, and results for classical Sobolev spaces. The mis-specified setting showcases how our inferential results deliver meaningful guarantees without any assumptions on the bias $f_\lambda - f_0$.

Figure 1 plots the true f_0 in solid red and the pseudo-true f_λ in dashed red. As expected, f_λ is smoother than f_0. We implement KRR $\hat{f}$ in solid blue using $n = 500$ observations, with our variable-width sup-norm confidence band in light blue. We also present a fixed-width H-norm band in dashed, medium blue; in this example the two bands nearly coincide. While Estimator 4.1 gives a sup-norm confidence band, the theory in Sections 5 and 6 also justifies the H-norm band. The bands contain both f_0 and f_λ across values of $x \in S$.

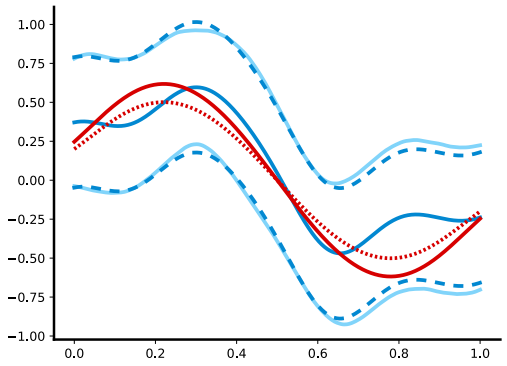


Figure 1: Warm up: Standard data with nonlinearity

Figure 1 suggests that in one draw from the data-generating process, the bands seem to cover f_0 and f_λ. Using the simulation described above, we now implement our procedure 500 times, collecting 500 estimates $\hat{f}$. We visualize the empirical distribution of $\|\sqrt{n}(\hat{f} - f_0)\|_\infty$ across repeated simulations. We also take data from only the first simulation, and generate 500 draws of $\|\mathfrak{B}\|_\infty$ conditional upon this data. Qualitatively, in Figure 2, the distribution of $\|\sqrt{n}(\hat{f} - f_0)\|_\infty$ across multiple simulation draws closely

[15]The lengthscale ι is a kernel hyperparameter with well-known heuristics; see Appendix K.

resembles the conditional distribution of $\|\mathfrak{B}\|_\infty$ given data from only the first simulation draw. This illustrates fidelity of our bootstrap approximation.

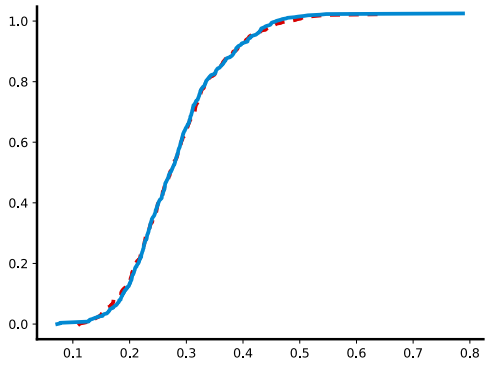


Figure 2: Our approach succeeds with nonlinearity. We compare the distribution of $\|\mathfrak{B}\|_\infty$ using our procedure and a single draw of the data (blue) with the distribution of $\sqrt{n}\|\hat{f} - f_0\|_\infty$ across many draws (red, dashed). Coverage of f_0 is 94.8%.

We quantify performance in coverage tables. In Table 1, different rows correspond to different sample sizes. Initially, we consider the theoretical tuning of the regularization hyperparameter $\lambda = n^{-1/3}$, following Section 6. For each sample size, we record four notions of coverage. In the first and second column, we evaluate coverage in terms of sup norm for the true regression f_0 and the pseudo-true regression f_λ, across 500 draws from the DGP. In the third and fourth column, we evaluate coverage in terms of the H norm for both quantities across 500 draws from the DGP. Our confidence bands are the correct width, since about 95% of them include their target functions.

Table 2 revisits the issue of tuning the hyperparameter λ. We fix the sample size to $n = 500$. Different rows correspond to different tunings of λ. Across regularization values, the confidence bands are again the correct width, attaining nominal coverage. When regularization strongly deviates from theoretical guidelines, coverage breaks down.

Finally, we fix $n = 500$ and $\lambda = n^{-1/3}$ and examine additional metrics of confidence band performance beyond coverage. Table 3 records bias and width of the confidence bands across draws from the DGP. We find that the H norm bands have slightly more

sample	sup norm		H norm	
	true	pseudo	true	pseudo
50	0.964	0.964	0.924	0.926
100	0.972	0.976	0.940	0.942
500	0.948	0.958	0.932	0.940
1000	0.974	0.972	0.946	0.958

Table 1: Coverage is nominal across sample sizes with nonlinearity. Across rows, we vary n and set $\lambda = n^{-1/3}$, following Section 6.

reg.	sup norm		H norm	
	true	pseudo	true	pseudo
0.500	0.836	0.948	0.802	0.932
0.100	0.942	0.952	0.930	0.940
0.050	0.974	0.964	0.952	0.956
0.010	0.970	0.968	0.968	0.968
0.005	0.942	0.944	0.934	0.934
0.001	0.920	0.922	0.954	0.954

Table 2: Coverage is nominal across regularization values with nonlinearity. Across rows, we fix $n = 500$ and vary λ across a reasonable range.

bias and width than the sup norm bands. Both types of bands achieve nominal coverage for both f_0 and f_λ. Since we use the same band for both f_0 and f_λ, the width is the same for both quantities, but the bias for f_0 reflects the difference $f_\lambda - f_0$.

4.3 Robust performance with ranking data

We repeat this exercise with ranking data. See Section 7 for details on the appropriate kernel. As before, the true regression f_0 is the third eigenfunction, which ensures correct specification. See Section K for details on the DGP. The extremely high-dimensional

metric	sup norm		H norm	
	true	pseudo	true	pseudo
coverage	0.976	0.978	0.968	0.968
bias	0.019	0.000	0.022	0.000
width	0.899	0.899	0.943	0.943

Table 3: A detailed look: Coverage, bias, and width with nonlinearity. Across rows, we fix $n = 500$ and set $\lambda = n^{-1/3}$. We examine additional metrics in addition to coverage.

regression is not easy to visualize, so we simply visualize the empirical distribution of $\|\sqrt{n}(\hat{f} - f_0)\|_\infty$ in Figure 3. As before, the distribution resembles that of $\|\mathfrak{B}\|_\infty$, suggesting that our bootstrap approximation is reasonable.

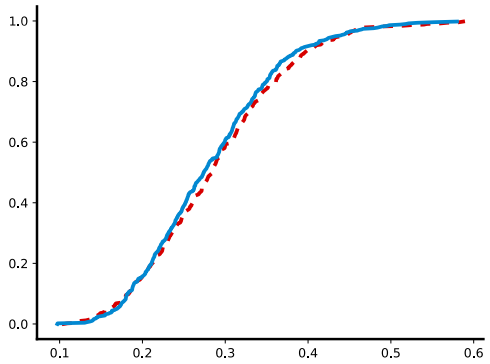


Figure 3: Our approach succeeds with ranking data. We compare the distribution of $\|\mathfrak{B}\|_\infty$ using our procedure and a single draw of the data (blue) with the distribution of $\sqrt{n}\|\hat{f} - f_0\|_\infty$ across many draws (red, dashed). Coverage of f_0 is 95.2%.

We generate similar coverage tables: Table 4 varies n across rows to evaluate coverage across sample sizes; Table 5 varies λ across rows to evaluate coverage across regularization values; and Table 6 fixes n and λ to study additional metrics beyond coverage. We obtain similar results to the standard data setting, despite the additional challenge of using ranking data. For regularization values that are relatively close

to the theoretical value of $\lambda = n^{-1/3}$, the confidence bands are the correct width, attaining nominal coverage. Comparing Tables 2 and 5, we see that tuning λ close to the theoretical value is more important for ranking data than for standard data. Our bands for KRR achieve high quality performance for general data types. As such, our bands provide reliable uncertainty quantification for challenging economic applications, e.g. individual preference data.

sample	sup norm		H norm	
	true	pseudo	true	pseudo
50	0.986	0.982	0.972	0.970
100	0.982	0.984	0.970	0.978
500	0.948	0.958	0.934	0.952
1000	0.972	0.978	0.946	0.958

Table 4: Coverage is nominal across sample sizes with ranking data. Across rows, we vary n and set $\lambda = n^{-1/3}$, following Section 6.

reg.	sup norm		H norm	
	true	pseudo	true	pseudo
0.500	0.836	0.966	0.836	0.974
0.100	0.974	0.980	0.988	0.984
0.050	0.976	0.974	0.960	0.958
0.010	0.956	0.956	0.820	0.818
0.005	0.958	0.958	0.628	0.628
0.001	0.994	0.994	0.414	0.414

Table 5: Coverage is nominal for regularization values near $n^{-1/3}$ with ranking data. Across rows, we fix $n = 500$ and vary λ across a reasonable range.

metric	sup norm		H norm	
	true	pseudo	true	pseudo
coverage	0.952	0.95	0.942	0.940
bias	0.018	0.000	0.027	0.000
width	0.942	0.942	2.342	2.342

Table 6: A detailed look: Coverage, bias, and width with ranking data. Across rows, we fix $n = 500$ and set $\lambda = n^{-1/3}$. We examine additional metrics in addition to coverage.

5 Finite sample analysis of partial sums

5.1 Overview of results

Recall the three reasons why uniform inference for KRR is challenging. First, we would like to provide results for many data types simultaneously. We introduce $\mathfrak{B}$ as an algorithmic solution and demonstrate its performance in simulations with standard and non-standard data, yet we still need to provide theoretical justification. Second, a computationally intensive bootstrap may undermine KRR's practicality. We propose a closed form solution for $\mathfrak{B}$ that re-uses the kernel operations of KRR and uses symmetry to cancel bias, yet we still need to prove that this closed form is correct. Third, as $\lambda \downarrow 0$, there is no uniform convergence to a Gaussian limit. Nonetheless, we must argue that our procedure is valid.

We propose a nonasymptotic framework to overcome these theoretical challenges. In this section, we analyze Gaussian and bootstrap couplings for general partial sums $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i$, where $U = (U_1, U_2, \ldots, U_n)$ is an i.i.d. sequence of n random variables taking values in H, such that $\mathbb{E}(U_i) = 0$, $\mathbb{E}\left\|U_i\right\|^2 < \infty$, and $\Sigma \coloneqq \mathbb{E}(U_i \otimes U_i^*)$. Gaussian and bootstrap couplings in Hilbert norms are "easy": the rates have excellent dependence on the $L^2(\mathbb{P})$ covering number. Moreover, approximation in H norm implies approximation in sup-norm: by the Cauchy-Schwarz inequality, $k \leq \kappa^2$ implies $\sup_{x \in S} |f(x)| \leq \kappa \left\|f\right\|$. We therefore study Gaussian and bootstrap couplings in H to justify our uniform

confidence bands. Future work may apply these results to kernel methods beyond KRR.

In Section 6, we apply the general results of this section to the inferential procedure proposed in Section 4. In particular, we demonstrate that KRR is well approximated by the partial sum with $U_i = (T+\lambda)^{-1}\{(k_{X_i} \otimes k^*_{X_i} - T)(f_0 - f_\lambda) + \varepsilon_i k_{X_i}\}$, where $T = \mathbb{E}(k_{X_i} \otimes k^*_{X_i})$. This characterization may be viewed as a functional Bahadur representation. Nonasymptotic analysis in this section allows $\lambda \downarrow 0$ in Section 6, for the fully nonparametric regime and for general data types.

In addition to our key assumption on the local width, defined in Section 3, we place regularity conditions on the distribution of U_i. We consider two possible regularity conditions, which lead to different rates.

Definition 5.1 (Regularity conditions)**.** U_i is *a-bounded* if $\|U_i\| \le a$ almost surely. U_i is *b-sub-Gaussian* if for all $t \in H$, $\log \mathbb{E}(\exp \langle U_i, t\rangle) \le \frac{b^2}{2}\mathbb{E}(\langle U_i, t\rangle^2)$.

Remark 5.2 (Intuition)**.** Boundedness is straightforward. For kernel mean embeddings, $U_i = k_{X_i} - \mathbb{E}(k_{X_i})$ and $k \le \kappa$ imply $a = 2\kappa$. For KRR, $U_i = (T+\lambda)^{-1}\{(k_{X_i} \otimes k^*_{X_i} - T)(f_0 - f_\lambda) + \varepsilon_i k_{X_i}\}$, so the bound diverges as $\lambda \downarrow 0$, which we capture via finite sample analysis. Sub-Gaussianity means that the tails are no worse than a Gaussian distribution; the moments of U_i are well behaved. For both kernel mean embeddings and KRR, we directly assume k_{X_i} are sub-Gaussian, which sharpens the analysis.

5.2 RKHS-valued Gaussian coupling

We begin with a finite-sample Gaussian approximation result, which is motivated by our interest in KRR.

Theorem 5.3 (Gaussian coupling)**.** Suppose the U_i are a-bounded. Then for all $m \ge 1$ there exists a Gaussian random variable Z taking values in H, with $\mathbb{E}(Z \otimes Z^*) = \Sigma$, such that with probability at least $1 - \eta$,

$$\left\| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \right) - Z \right\| \lesssim \sqrt{\log(6/\eta)}\sigma(\Sigma, m) + \frac{am^2 \log(m^2/\eta)}{\sqrt{n}}.$$

Similarly, if the U_i are b-sub-Gaussian then for all $m \leq n$ there exists Z with the same distribution as before, such that with probability at least $1 - \eta$,

$$\left\| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \right) - Z \right\| \lesssim b\sqrt{\log(6/\eta)}\sigma(\Sigma, m) + b^3 \left\|\Sigma\right\|_{op}^{\frac{1}{2}} \left\{ \frac{m^{\frac{3}{2}} \log(n)^2}{\sqrt{n}} \right\} (3/\eta)^{1/\log(mn)}.$$

Example 5.4 (KMT rate for smooth, radial kernels)**.** Suppose $k(s,t) = \psi(\|s - t\|_{\mathbb{R}^p})$ for some ψ with $(d^r/dt^r)\sqrt{\psi} \leq r!\, C^r$, e.g. the Gaussian and inverse multi-quadratic kernels (Wendland, 2004, Chapter 11). If $U_i = k_{X_i} - \mathbb{E}(k_{X_i})$, then it follows from Belkin (2018, Theorem 5) that $\nu_s(\Sigma) \leq C \exp(-cs^\gamma)$, where $\gamma = \gamma(p)$, so the bounded version of Theorem 5.3 implies

$$\sup_{\|f\| \leq 1} \left| \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(X_i) \right\} - \langle Z,\, f \rangle \right| = \left\| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} k_{X_i} \right) - Z \right\| = \tilde{O}(1/\sqrt{n}),$$

with high probability, where $\tilde{O}(-)$ hides poly-logarithmic factors in n. Thus we obtain "KMT type" approximation for the empirical process indexed by the unit ball in H, at the optimal rate (Koltchinskii, 1994).

Remark 5.5 (Interpretation)**.** Theorem 5.3 holds for many choices of m. The condition $1 \leq m \leq n$ simplifies the presentation and does not bind; one must choose $1 \ll m \ll n^{\frac{1}{4}}$ for the former bound to be $o_p(1)$, and $1 \ll m \ll n^{\frac{1}{3}}$ for the latter. Note that $\sigma(\Sigma, m) \downarrow 0$ holds under only the assumption that $\mathbb{E}\left\|U_i\right\|_H^2 < \infty$, a weak and well-known condition (Hoffmann-Jørgensen and Pisier, 1976). So for many choices of m, under almost no further assumptions, the bounds vanish.

However, without further assumptions the rate may be arbitrarily slow, which does not suffice for our application to KRR. Therefore we choose m as function of n to optimize the bounds: in Appendix C.6, we specialize and optimize Theorem 5.3 for leading cases, corresponding to different regimes for $m \mapsto \sigma(\Sigma, m)$. As in Example 5.4, such bounds follow from properties of k, e.g. for Gaussian and Sobolev (a.k.a. Matérn) kernels. We derive general results for polynomial and exponential eigenvalue decay.

Remark 5.6 (Comparison)**.** Theorem 5.3, though highly specific to Hilbert norms, is a strong form of approximation that can be applied to approximate the distribution

of any continuous functional $F : H \to \mathbb{R}$. In particular, unlike Chernozhukov et al. (2014b) and related work, it is not limited to suprema of linear functionals. For our specific problem, we seem to improve the mode and rate of convergence.

For a rough comparison of rates, the method used by Berthet and Mason (2006) and also Chernozhukov et al. (2014b, 2016) is to cover the index set with $p = 2^m$ points at resolution $\delta(p)$ and incur the approximation error $\sigma(\Sigma, \log p)$, which is the local Gaussian complexity at the scale $\delta(p)$. In this manner, Berthet and Mason (2006) obtain the same form of approximation as we do, but with the exponentially worse rate $p^2/\sqrt{n} + \sigma(\Sigma, \log p)$. Applying the recent result of Chernozhuokov et al. (2022, Theorem 2.1) in the above argument gives the comparable rate $\{\log^5(np)/n\}^{1/4} + \sigma(\Sigma, \log p)$, with a weaker form of approximation.

By contrast, we obtain the rate $\log^2(p)/\sqrt{n} + \sigma(\Sigma, \log p)$ for strong approximation in H-norm, hence also in L^∞. Such improvements are specific to the RKHS setting, where we can exploit compatibility between the projection technique of Götze and Zaitsev (2011) and the finite dimensional ℓ^2 couplings of Zaitsev (1987a) and Buzun et al. (2022); see the proof sketch below. It seems this combination had not been applied to the RKHS before, where it is especially powerful, implying approximation in sup-norm with very good dependence on complexity.

An open question for future work is how to sharply apply the improved max-norm couplings in Chernozhuokov et al. (2022), and similar results, to the RKHS setting. For example, it appears that the finite dimensional results on max-statistics of Lopes et al. (2020); Lopes (2022) are not directly applicable.[16]

Remark 5.7 (Strong approximation)**.** Our result for bounded U_i actually implies strong approximation in the traditional sense, i.e. control of $\max_{j \le n} \| n^{-1/2} \sum_{i=1}^{j} U_i - Z_i \|$, at the same rate. This follows from combining our more detailed Proposition C.7 with

[16]To apply finite-dimensional results on max-statistics, reduce $\sup_{x \in S} |S_n(x)| = \sup_{x \in S} |\langle S_n, k_x \rangle|$ to a finite dimensional max by covering the index set $\{k_x | x \in S\}$. Lopes et al. (2020, Assumption 2.2(ii)) requires coordinates to have correlation at most $1 - \epsilon_0$, for all n; it is well-suited for $\sup_{k \in \mathbb{N}} |\langle S_n, e_k \rangle|$ rather than $\sup_{x \in S} |S_n(x)|$.

de la Peña and Giné (1999, Theorem 1.1.5).

5.3 Symmetrized bootstrap coupling

In order to perform inference, we need some way to sample from the approximating Gaussian distribution. To do so, we propose the following symmetrized multiplier bootstrap $Z_{\mathfrak{B}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}\left(\frac{V_i - V_j}{\sqrt{2}}\right)$, where the h_{ij} are i.i.d. Gaussian multipliers and $V_i = U_i + \mu$, for some unknown, deterministic bias μ. The symmetrized bootstrap is motivated by our application to KRR, where we aim to provide valid uncertainty quantification even when the bias is significant. It is based on the observation that U_i has the same covariance as $(U_i - U_j)/\sqrt{2} = (V_i - V_j)/\sqrt{2}$.

Theorem 5.8 (Symmetrized bootstrap coupling)**.** Suppose the U_i are a-bounded and $n \geq 2$. Then for all m, there exists a random variable Z' such that the conditional law of Z' given U is almost surely Gaussian with covariance Σ, and with probability $1 - \eta$,

$$\mathbb{P}\left(\|Z' - Z_{\mathfrak{B}}\| \geq C \log^{3/2}(C/\eta) \left[\left\{ \frac{a^2\sigma^2(\Sigma, 0)m}{n} + \frac{a^4 m}{n^2} \right\}^{\frac{1}{4}} + \sigma(\Sigma, m) \right] \middle| U \right) \leq \eta.$$

If the U_i are instead b-sub-Gaussian, for any n, the identical result holds with

$$\mathbb{P}\left(\|Z' - Z_{\mathfrak{B}}\| \geq C \log^{3/2}(C/\eta) \left[\left\{ \frac{b^4\sigma^4(\Sigma, 0)m}{n} \right\}^{\frac{1}{4}} + \sigma(\Sigma, m) \right] \middle| U \right) \leq \eta.$$

Remark 5.9 (Interpretation)**.** On a high-probability event, we can approximately sample from the distribution of Z'—and hence, from the distribution of the empirical process $n^{-1/2}\sum_{i=1}^{n} U_i$—by sampling from the bootstrap process $Z_{\mathfrak{B}}$ conditional on the realized data, U. These results presented a significant technical challenge, and our methods differ substantially from those used by, e.g. Chernozhukov et al. (2016). As before, we heavily exploit the Euclidean structure in H to derive a stronger form of approximation. In our setting, strong approximation does not incur slower rates.

Remark 5.10 (Traditional multiplier bootstrap)**.** Our technical results only use bounds on $\mathbb{E}(Z_{\mathfrak{B}} \otimes Z_{\mathfrak{B}}^* | U) - \Sigma$, so they may apply to other settings. For example, one may

consider the traditional Gaussian multiplier bootstrap $n^{-1/2}\sum_{i=1}^n h_i U_i$, subsampled data, or more modern covariance matrix approximations.

5.4 Proof sketch

Gaussian coupling. We sketch the proof here, and provide the formal argument in Appendix C. The key ingredients in our proof are the following Euclidean norm couplings due to Zaitsev (1987a) and Buzun et al. (2022), which cover the bounded and sub-Gaussian settings, respectively.

Lemma 5.11 (Zaitsev, 1987a, Theorem 1.1)**.** Let $\xi_1, \xi_2, \ldots$ be an independent sequence of random variables in $\mathbb{R}^m$ whose Euclidean norms satisfy $\|\xi_i\|_{\mathbb{R}^m} \leq a$. For each n and for each i, there exists a Gaussian $\theta_i \in \mathbb{R}^m$ with the same covariance as ξ_i, such that

$$\mathbb{P}\left\{\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n(\xi_i-\theta_i)\right\|_{\mathbb{R}^m} > \delta\right\} \lesssim m^2 \exp\left(\frac{-c\delta\sqrt{n}}{m^2 a}\right).$$

Lemma 5.12 (Buzun et al., 2022, Theorem 3)**.** Set $\xi = \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i$, and suppose that each ξ_i is b-sub-Gaussian. Then, for any $m \leq n$ there exists a Gaussian random variable $\theta \in \mathbb{R}^m$ with covariance $\Sigma = \mathbb{E}(\xi_i \xi_i^\top)$ such that with probability $1-\eta$

$$\|\theta - \xi\|_{\mathbb{R}^m} \lesssim \|\Sigma\|_{op}^{\frac{1}{2}} b^3 \left\{\frac{m^{\frac{3}{2}}\log(n)^2}{\sqrt{n}}\right\}(1/\eta)^{1/\log(mn)}.$$

Equipped with the above results, we proceed by finite-dimensional approximation. However, instead of the standard covering arguments which are necessary in Banach spaces, we choose a different, and rather simple, method due to Götze and Zaitsev (2011) that exploits the ℓ^2 structure in H.

Let Π_m denote the projection of H onto the subspace E_m spanned by the top m eigenvectors of Σ. Note that $\Pi_m = A^*A$ where $A : H \to \mathbb{R}^m$ projects E_m to $\mathbb{R}^m$ and $A^* : \mathbb{R}^m \to H$ isometrically embeds $\mathbb{R}^m$ as E_m.

We may then: (i) use the preceding results to couple $n^{-1/2}\sum_{i=1}^n AU_i$ and θ in $\mathbb{R}^m$; (ii) apply A^* to deduce a coupling of $n^{-1/2}\sum_{i=1}^n U_i$ and $A^*\theta$ in H; (iii) use concentration to show that the error from projecting onto E_m is at most of order $\sigma(m)$.

The key observation is that the standard covering argument requires roughly $\exp(m)$ points to achieve an approximation error of $\sigma(\Sigma, m)$. Thus, in the RKHS setting, we get a logarithmic dependence on the covering number comparable to Chernozhukov et al. (2014b, 2016) "for free," and with a stronger form of approximation.

Bootstrap coupling. We sketch the proof here, and provide the formal argument in Appendix D. Our proof of Theorem 5.8 builds on the observation that if g_1 and g_2 are Gaussian random vectors in $\mathbb{R}^m$ with covariances Σ_1 and Σ_2, then one can construct a trivial coupling between the two. Take h to be a standard normal vector, so that $g_1 \sim \Sigma_1^{1/2} h$, $g_2 \sim \Sigma_2^{1/2} h$. Then

$$\mathbb{E}\|\Sigma_1^{1/2} h - \Sigma_2^{1/2} h\|_{\mathbb{R}^m}^2 = \mathbb{E}\{h^\top (\Sigma_1^{1/2} - \Sigma_2^{1/2})^\top (\Sigma_1^{1/2} - \Sigma_2^{1/2}) h\} = \|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|_F^2.$$

Conditional upon the data, $Z_{\mathfrak{B}}$ is Gaussian with a covariance $\hat{\Sigma} = \mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}_n[U_i] \otimes (\mathbb{E}_n[U_i])^*$ that should converge to $\Sigma = \mathbb{E}(U_i \otimes U_i^*)$, in the sense that $\|\hat{\Sigma} - \Sigma\|_{\mathrm{HS}} = o_p(1)$. The primary challenge is to show that if $\|\hat{\Sigma} - \Sigma\|_{\mathrm{HS}}$ is small then $\|\hat{\Sigma}^{1/2} - \Sigma^{1/2}\|_{\mathrm{HS}}$ is also small, thereby reducing the problem to covariance estimation in a suitable sense. We overcome this challenge with the help of the following lemma.

Lemma 5.13 (Wihler (2009, Theorem 1.1))**.** If $A, B \in \mathbb{R}^{m \times m}$, then $\|A^{\frac{1}{2}} - B^{\frac{1}{2}}\|_F \leq m^{\frac{1}{4}} \|A - B\|_F^{\frac{1}{2}}$.

By using a truncation argument along the lines of the previous subsection's proof, we deduce the following. Here Π_m denotes the projection operator for the top m eigenvectors of Σ, and $\Pi_m^\perp = (I - \Pi_m)$.

Proposition 5.14 (Abstract bootstrap coupling)**.** For any $m \geq 1$ there exists a random variable Z' that is conditonally Gaussian with covariance Σ, such that with probability at least $1 - 3\eta$,

$$\|Z' - Z_{\mathfrak{B}}\| \leq \left\{1 + \sqrt{2\log(1/\eta)}\right\} \left\{m^{\frac{1}{4}} \Delta_1^{1/2} + \Delta_2^{1/2} + 2\sigma(\Sigma, m)\right\},$$

where $\Delta_1 \coloneqq \|\hat{\Sigma} - \Sigma\|_{\mathrm{HS}}$ and $\Delta_2 \coloneqq \operatorname{tr} \Pi_m^\perp (\hat{\Sigma} - \Sigma) \Pi_m^\perp$.

Finally, we control the quantities Δ_1 and Δ_2 using concentration. From the results above, whenever we can control the quantities Δ_1 and Δ_2, we can establish a version of Theorem 5.8. Thus, our technique does indeed accommodate e.g. sub-sampled data or the traditional bootstrap; in both cases $\hat{\Sigma}$ concentrates around Σ.

Remark 5.15 (Comparison)**.** In our setting, with $m = \log p$, Chernozhukov et al. (2016, Theorem 3.3) implies $|\|Z'\|_\infty - \|Z_{\mathfrak{B}}\|_\infty| \lesssim \sqrt{m}\|\hat{\Sigma} - \Sigma\|_{op}^{1/2} + \sigma(\Sigma, m)$ using an argument based on Stein interpolation. This can give a better rate when $\|\hat{\Sigma} - \Sigma\|_{op} \ll \|\hat{\Sigma} - \Sigma\|_{\mathrm{HS}}$, though with a weaker form of approximation.

6 Finite sample analysis of uniform confidence bands

6.1 Overview of results

We apply Section 5's general results to conduct uniform inference for KRR. We approximate the distribution of $\sqrt{n}(\hat{f} - f_0)$ with the distribution of a feasible symmetrized bootstrap process $\mathfrak{B}$, whose conditional distribution can be efficiently simulated by reusing artifacts from KRR estimation, as described in Section 4. Under standard conditions for KRR consistency, we prove that Estimator 4.1 yields valid and sharp H-norm confidence sets $\hat{S}_n$ for the true parameter f_0, which shrink at nearly the minimax rate. Since k is bounded, these immediately imply uniform confidence bands.

We provide two versions of our KRR inference result: (i) an infinitesimal factor approach in Section 6.2, and (ii) an anti-concentration approach in Section 6.3. The former justifies the near-minimax H-norm confidence sets described above, and allows for more generality. The latter justifies sharper, uniform and variable-width confidence bands with asymptotically exact coverage, under an additional anti-concentration assumption. We contribute a practical diagnostic of whether anti-concentration holds. In summary, the novel contributions of this section are: providing sharp H-norm confidence sets without anti-concentration, studying general functionals F such as variable-width uniform bands with anti-concentration, and proposing a practical diagnostic to assess

anti-concentration; see Section 2 for further discussion.

For both types of results, we give concrete corollaries according to whether the spectrum $\nu_s(T)$ decays polynomially or exponentially (e.g. Sobolev versus Gaussian kernels), and whether the data are bounded or sub-Gaussian (e.g. $k_{X_i} \leq \kappa$ versus Assumption 6.12 below). For clarity of exposition, we state our results abstractly, then fill in the corollaries with a table in Section 6.4.

Definition 6.1 (Key quantities)**.** To streamline notation, we define the functions (Q, R, L, B) so that the following statements hold. We verify that these conditions hold in Section 6.4 under appropriate assumptions.

- Gaussian approximation. There exists a Gaussian random element Z in H such that with probability at least $1 - \eta$, $\left\| \sqrt{n}(\hat{f} - f_\lambda) - Z \right\| \leq Q(n, \lambda, \eta)$. This condition will be verified with Theorem 5.3 using local width.

- Bootstrap approximation. There exists a random element Z' in H whose conditional distribution given D is almost surely Gaussian with covariance Σ, and with probability at least $1 - \eta$, $\mathbb{P}\{ \|\mathfrak{B} - Z'\| \leq R(n, \lambda, \eta) | D \} \geq 1 - \eta$. This condition will be verified with Theorem 5.8 using local width.

- Variance lower bound. It holds with probability $1 - \eta$ that $\|Z\| \geq L(\lambda, \eta)$ for some function L which is strictly increasing in η. This condition will be verified with a lemma, again using local width.

- Bias upper bound. It holds that $\sqrt{n}\|f_\lambda - f_0\| \leq B(\lambda, n)$. This condition will be verified with a lemma using the source condition.

Finally let $\Delta(n, \lambda, \eta) \coloneqq Q(n, \lambda, \eta) + R(n, \lambda, \eta)$. We abbreviate these quantities by suppressing their arguments.

6.2 Uniform inference via infinitesimal factor

The former version of our KRR inference result builds on the infinitesimal factor approach proposed by Andrews and Shi (2013). In high-dimensional or non-parameteric

inference, the distribution of the test statistic may be extremely concentrated, in which case small perturbations will cause $\mathbb{P}(f_0 \in \hat{S}_n)$ to fluctuate dramatically around the desired coverage level $1-\alpha$.

We therefore expand the size of the confidence set $\hat{S}_n$ by an infinitesimal factor $\delta = o(1)$ to guarantee valid and sharp confidence bands in the sense of Definitions 3.3 and 3.4. This expansion may lead to conservative confidence bands. In practice, however, expanding the confidence set by an infinitesimal factor is unlikely to impact conclusions drawn from inference in a meaningful way, and allows us to derive very general results.

Theorem 6.2 (Uniform inference via infinitesimal factor)**.** For $\alpha \in (0,1)$ define $\hat{t}_\alpha(D)$ by $\mathbb{P}(\|\mathfrak{B}\| > \hat{t}_\alpha | D) = \alpha$.[17] Suppose that the infinitesimal factor δ satisfies

$$\frac{1}{2} \geq \delta \geq \frac{\Delta(\eta, n, \lambda) + B(n, \lambda)}{L(1-\alpha-2\eta, \lambda) - \Delta(\eta, n, \lambda)}.$$

Then $\hat{S}^{\text{inc}}_{n,\delta} = \left\{ \hat{f} + h \,\middle|\, \|h\| \leq (1+\delta)\hat{t}_\alpha n^{-1/2} \right\}$ is (3η)-valid and $(2\delta, 2\eta)$-sharp.

We summarize the main implications. See Section 6.4 for exact rates.

Corollary 6.3 (Polynomial decay)**.** With polynomial decay and sub-Gaussian data, choose $\delta \asymp \log^{-c}(n)$ and $\lambda \asymp n^{-(\beta+\epsilon)/(r\beta+1)}$ for some small $\epsilon > 0$, where r is from Assumption 3.10 and β is the polynomial rate of decay formalized in Table 7. Then $\hat{S}^{\text{inc}}_{n,\delta}$ is $O(1/n)$ valid, $\{2\delta, O(1/n)\}$ sharp, and shrinks at nearly the minimax rate. With bounded data, the same statement holds provided $r > 1 + 2/\beta$.

Corollary 6.4 (Exponential decay)**.** With exponential decay and either bounded or sub-Gaussian data, choose $\delta \asymp \log^{-c}(n)$ and $\lambda \asymp n^{-(1+\epsilon)/r}$ for some small $\epsilon > 0$, where r is from Assumption 3.10. Then $\hat{S}^{\text{inc}}_{n,\delta}$ is $O(1/n)$ valid, $\{2\delta, O(1/n)\}$ sharp, and shrinks at nearly the minimax rate.

[17] Conditional upon D, $\|\mathfrak{B}\|$ is a non-degenerate quadratic form of a finite dimensional Gaussian, so it has a density.

6.3 Uniform inference via anti-concentration

The latter version of our KRR inference result builds on the anti-concentration approach of Chernozhukov et al. (2014a). This approach explicitly rules out concentration of the test statistic to ensure valid and exact sup-norm confidence bands $\hat{S}_n$. For many supremeum-type statistics, the required anti-concentration property may be derived explicitly (Chernozhukov et al., 2014a, 2015). We propose what appears to be a novel, data-driven verification of the anti-concentration property in settings where such results are not available; in our simulations, the required property does hold.

We state our results for an arbitrary sup-norm continuous functional F. In particular, let $F : H \to \mathbb{R}$ be a uniformly continuous functional, i.e. $|F(u) - F(v)| \leq \psi(\|u - v\|)$ for some modulus of continuity $\psi : \mathbb{R} \to \mathbb{R}$. If $F(f) = \sup_{x \in S} |f(x)|$, then $\psi(x) \leq \kappa x$.

Assumption 6.5 (Anti-concentration)**.** For a Gaussian W in H with covariance Σ, there exists $\zeta > 0$ such that $\zeta := \sup_{\delta > 0} \sup_{t \in \mathbb{R}} \left[\frac{1}{\delta} \mathbb{P} \left\{ |F(W) - t| \leq \delta \right\} \right] < \infty$.

Remark 6.6 (Interpretation)**.** Assumption 6.5 is a technical condition which ensures that the random variable $F(W)$ is not too concentrated. For some F, one may bound ζ using ideas of Chernozhukov et al. (2014a, 2015). We show that ζ can be replaced by a data driven quantity in general.

Theorem 6.7 (Uniform inference via anti-concentration)**.** Suppose Assumption 6.5 holds. Then with probability $1 - \eta$

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P} \left[F\left\{ \sqrt{n}(\hat{f} - f_0) \right\} \leq t \right] - \mathbb{P} \left\{ F(\mathfrak{B}) \leq t \middle| D \right\} \right| \leq 2(\zeta \psi(\Delta + B) + \eta). \tag{1}$$

Moreover, in the absence Assumption 6.5, we have the following the data-driven bound

$$\begin{aligned} &\sup_{t \in \mathbb{R}} \left| \mathbb{P} \left[F\left\{ \sqrt{n}(\hat{f} - f_0) \right\} \leq t \right] - \mathbb{P} \left\{ F(\mathfrak{B}) \leq t \middle| D \right\} \right| \\ &\quad \leq 2 \left[\sup_{t \in \mathbb{R}} \mathbb{P} \left\{ |F(\mathfrak{B}) - t| \leq 2\psi(\Delta + B) \middle| D \right\} + \eta \right]. \end{aligned}$$

Remark 6.8 (Nonparametrics versus semiparametrics)**.** We provide Gaussian and bootstrap couplings for the entire KRR function in supremum norm as a stochastic

process on the input space S; see Propositions 6.14 and 6.15 in the proof sketch. S can be any complete, separable metric space. For uniform confidence bands, we bootstrap the single-valued statistic $\sup_{x\in S}|\sqrt{n}\{\hat{f}(x) - f_0(x)\}|$. The key point distinguishing our result from semiparametric inference is that we allow F to be any sup-norm continuous functional, which is more general and allows for uniform confidence bands.

Example 6.9 (Uniform confidence band)**.** Let $w : S \to (0, \infty)$ be a bounded weight function. Then $F_w(f) = \sup_{x\in S}|f(x)w(x)|$ is uniformly continuous with modulus $x \mapsto x\kappa\|w\|_\infty$. Moreover, $F_w\{\sqrt{n}(\hat{f} - f_0)\} \le t$ is precisely the event that for all $x \in S$,

$$\hat{f}(x) - \frac{t/w(x)}{\sqrt{n}} \le f_0(x) \le \hat{f}(x) + \frac{t/w(x)}{\sqrt{n}}.$$

Thus whenever Theorem 6.7 implies an $o_p(1)$ bound in (1)—e.g. under Assumption 6.5, when k is a "smooth, radial kernel" as in Corollary 5.4, $\lambda = n^{-1/3-\epsilon}$, and $\zeta \ll B$—we can simulate to find $\hat{t}_\alpha$ such that $\mathbb{P}\{F_w(\mathfrak{B}) \le \hat{t}_\alpha | D\} = \alpha$. Taking $t = \hat{t}_\alpha$ above we obtain an asymptotically exact $(1 - \alpha)$ confidence band for f_0; see Section 4.

6.4 Proof sketch

Notation. Our KRR inference results in Section 6 build on more abstract results for partial sums in Section 5, which are in terms of local width. When appealing to these general results, we control the local width of $\Sigma = \mathbb{E}(U_i \otimes U_i^*)$ in terms of the local width of the operators $(T + \lambda)^{-2}T$ and T, where $T = \mathbb{E}(k_{X_i} \otimes k_{X_i}^*)$ is the covariance operator induced by the data. Slightly abusing notation, we write

$$\mathfrak{n}(\lambda) \coloneqq \sigma^2\left\{(T + \lambda)^{-2}T, 0\right\} \quad \text{and} \quad \sigma^2(m) \coloneqq \sigma^2(T, m).$$

The quantity $\mathfrak{n}(\lambda) = \operatorname{tr}\{(T + \lambda)^{-2}T\}$ is closely related to the *effective dimension* in the kernel methods literature. See Appendix I for details.

In addition to our key assumption on the local width, we place a key assumption called the source condition on the smoothness of f_0, defined in Section 3. Finally, we list additional regularity assumptions to simplify the exposition and facilitate comparisons.

Assumption 6.10 (Rate condition)**.** The sample size n, regularization parameter λ, and kernel k, are such that $n \geq 16\kappa^2 \ln(4/\eta)^2\{\mathfrak{n}(\lambda) \vee \lambda^{-1} \vee \lambda^{-2}\mathfrak{n}(\lambda)^{-1}\}$.

Assumption 6.11 (Noise lower bound)**.** The regression errors satisfy $\mathbb{E}(\varepsilon_i^2|X_i) \geq \underline{\sigma}^2$.

Assumption 6.12 (Sub-Gaussianity)**.** Each k_{X_i} is b sub-Gaussian.

Remark 6.13 (Interpretation)**.** Assumption 6.10 holds whenever our bounds are $o_p(1)$. The fundamental condition within Assumption 6.10 is $\mathfrak{n}(\lambda)/n \downarrow 0$, which is necessary for KRR to be consistent in H-norm.[18] Intuitively, $\mathfrak{n}(\lambda)/n \downarrow 0$ means that the effective dimension is smaller than the sample size. It implies Assumption 6.10 under the weak regularity condition that $\mathfrak{n}(\lambda) \geq \lambda^{-1}$.

Assumption 6.11 imposes a lower bound on the variance of the regression error, which is needed to guarantee coverage via under-smoothing.

Assumption 6.12 requires $\{f(X_i) \,|\, f \in H\}$ to be a sub-Gaussian process: $\|f\|_{L^2} \leq \|f\|_{\psi_2} \lesssim b\,\|f\|_{L^2}$ (cf. Lecué and Mendelson, 2013). This strong assumption leads to stronger guarantees, primarily for comparison. It is not necessary for our main results.

Verifying high level conditions. The proofs of Theorems 6.2 and 6.7 share several steps in common. These common steps provide concrete values for the functions (Q, R, L, B) in Definition 6.1. We state results for bounded and sub-Gaussian data. We specialize the results for polynomial and exponential decay in the rate tables below.

We being with Q. First, we show that $\sqrt{n}(\hat{f} - f_\lambda)$ is close to a partial sum, similar to the Bahadur representation of Shang and Cheng (2013) for splines (Appendix B).[19] In particular, we bound the difference

$$\left\| \sqrt{n}(\hat{f} - f_\lambda) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \right\|, \quad U_i = (T + \lambda)^{-1}\{(k_{X_i} \otimes k_{X_i}^* - T)(f_0 - f_\lambda) + \varepsilon_i k_{X_i}\}.$$

[18] Consistency in H-norm can be avoided by assuming the existence of an interpolation space $H \subseteq H' \subseteq L^2(X)$ which embeds continuously in L^∞ (Fischer and Steinwart, 2020). Future work may derive a Gaussian approximation in H' by extending our arguments.

[19] Yang et al. (2017) provide a stronger result assuming uniform boundedness of eigenfunctions.

We construct a Gaussian coupling Z to the partial sum $\frac{1}{\sqrt{n}}\sum_{i=1}^n U_i$ using the general results stated in Section 5.2 (proved in Appendix C). These initial steps establish a Gaussian approximation Z of $\sqrt{n}(\hat{f} - f_\lambda)$, and they are summarized by the following intermediate result. Set $M \coloneqq \kappa(\kappa\,\|f_0\| + \bar{\sigma})$ and $\bar{M} \coloneqq (\underline{\sigma}^{-1} \vee \kappa)(\kappa\,\|f_0\| + \bar{\sigma})$.

Proposition 6.14 (Gaussian approximation)**.** Suppose Assumption 6.10 holds. Then there exists a Gaussian Z in H, with covariance Σ, such that with probability $1-\eta$,

$$\left\|\sqrt{n}(\hat{f} - f_\lambda) - Z\right\| \lesssim Q(n,\lambda,\eta) = M\log^2(C/\eta)\left[\frac{1}{\lambda}\inf_{m\geq 1}\left\{\sigma(m) + \frac{m^2\log(m^2)}{\sqrt{n}}\right\} + \frac{\mathfrak{n}(\lambda)}{\sqrt{n}}\right].$$

Under Assumption 6.12, the identical statement holds with

$$Q(n,\lambda,\eta) = \bar{M}^4 b^3 (1/\eta)^{1/\log(n)}\left[\inf_{m\geq 1}\left\{\frac{\sigma(m)}{\lambda} + \frac{m^{\frac{3}{2}}\log(n)^2}{\sqrt{n\lambda}}\right\} + \frac{\mathfrak{n}(\lambda)}{\sqrt{n}}\right].$$

Next, we turn to R. For inference, we wish to sample from the law of the approximating Gaussian Z. We construct a bootstrap approximation to Z given by

$$Z_{\mathfrak{B}} = \frac{1}{n}\sum_{i,j=1}^n \left(\frac{U_i - U_j}{\sqrt{2}}\right) h_{ij} = \frac{1}{n}\sum_{i,j=1}^n \left(\frac{V_i - V_j}{\sqrt{2}}\right) h_{ij}$$

using the general results stated in Section 5.3 (proved in Appendix D). For KRR,

$$V_i = (T+\lambda)^{-1}\{(k_{X_i}\otimes k_{X_i}^*)(f_0 - f_\lambda) + \varepsilon_i k_{X_i}\} = (T+\lambda)^{-1}\{Y_i - f_\lambda(X_i)\}k_{X_i}.$$

Note that $V_i = U_i + \mu$ where $\mu = (T+\lambda)^{-1}T(f_0 - f_\lambda)$ is a bias term due to regularization. Its observed counterpart is

$$\hat{V}_i = (\hat{T}+\lambda)^{-1}\{(k_{X_i}\otimes k_{X_i}^*)(f_0 - \hat{f}) + \varepsilon_i k_{X_i}\} = (\hat{T}+\lambda)^{-1}\{Y_i - \hat{f}(X_i)\}k_{X_i}.$$

Therefore another step (Appendix E) is necessary to argue that it suffices to sample from the observed process. We give a high-probability bound for $\|Z_{\mathfrak{B}} - \mathfrak{B}\|$, where

$$\mathfrak{B} \coloneqq \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n h_{ij}\frac{\hat{V}_i - \hat{V}_j}{\sqrt{2}},$$

with explicit computations in Estimator 4.1. These steps prove that the process $\mathfrak{B}$ allows us to sample from Z conditional upon the data, and they are summarized by the following intermediate result.

Proposition 6.15 (Bootstrap approximation)**.** Suppose Assumption 6.10 holds. Then, there exists a random variable Z' whose distribution conditional upon the data is Gaussian with covariance Σ, and such that with probability at least $1-\eta$, conditional upon the data D, $\mathbb{P}\{\|\mathfrak{B}-Z'\| \lesssim R(n,\lambda,\eta)|D\} \geq 1-\eta$ where

$$R(n,\lambda,\eta) = M\log^2(C/\eta)\left(\inf_{m\geq 1}\left[\left\{\frac{m\,\mathfrak{n}(\lambda)}{\lambda^2 n}+\frac{m}{\lambda^4 n^2}\right\}^{\frac{1}{4}}+\frac{\sigma(m)}{\lambda}\right]+\frac{\kappa\,\mathfrak{n}(\lambda)}{\sqrt{n}}\right).$$

If in addition Assumption 6.12 holds, the identical statement holds with

$$R(n,\lambda,\eta) = \bar{M}^2 b\log^2(C/\eta)\left(\inf_{m\geq 1}\left[\left\{\frac{m\,\mathfrak{n}^2(\lambda)}{n}\right\}^{\frac{1}{4}}+\frac{\sigma(m)}{\lambda}\right]+\frac{\mathfrak{n}(\lambda)}{\sqrt{n}}\right).$$

Finally, we characterize (L,B).

Lemma 6.16 (Variance lower bound)**.** Let Z be a Gaussian random element of H with covariance Σ, and suppose Assumption 6.11 holds. Let $M' \coloneqq \kappa^2\|f_0\|^2+\bar{\sigma}^2$. Then with probability at least $1-\eta$, $\|Z\| \geq L(\lambda,\eta) = \sqrt{\underline{\sigma}^2\mathfrak{n}(\lambda)}-\left\{2+\sqrt{2\log(1/\eta)}\right\}\sqrt{M'/\lambda}$.

Lemma 6.17 (Bias upper bound; cf. Theorem 4 of Smale and Zhou, 2005)**.** If Assumption 3.10 holds then $\sqrt{n}\|f_\lambda - f_0\| \leq B(n,\lambda) = \sqrt{n}\kappa^{1-r}\lambda^{(r-1)/2}\|f_0\|$.

Using (Q,R,L,B), we directly establish Theorem 6.2 via elementary arguments in Appendix G. We then establish Theorem 6.7 by combining (Q,R) to establish bounds on $\mathbb{P}[|F\{\sqrt{n}(f_\lambda-\hat{f})\}-F(Z)|>\delta]$ and $\mathbb{P}\{|F(Z')-F(\mathfrak{B})|>\delta|D\}$ (also in Appendix G). We finally use anti-concentration to derive bounds on the difference between distribution functions, following Chernozhukov et al. (2014a, 2016).

Corollaries. We summarize our results with concrete corollaries according to whether the spectrum $\nu_s(T)$ decays polynomially or exponentially (e.g. Sobolev versus Gaussian kernels), and whether the data are bounded or sub-Gaussian (e.g. $k_{X_i} \leq \kappa$ versus Assumption 6.12 above). For each regime, we characterize (Q,R,L,B) suppressing log factors. We then characterize the restrictions on the regularization parameter λ implied by the conditions that $B \ll L$ (undersmoothing) and $(Q+R) \ll L$ (valid approximation). Finally, we evaluate whether the these conditions allow λ to approach the minimax optimal choice for learning. They generally do.

Spectrum	Poly. : $\nu_s(T) \asymp \omega s^{-\beta}$		Exp. : $\nu_s(T) \asymp \omega \exp(-\alpha s^\gamma)$	
Data	Bounded	sub-Gaussian	Bounded	sub-Gaussian
B	$n^{\frac{1}{2}}\lambda^{\frac{r-1}{2}}$			
L	$\lambda^{-\frac{1}{2}-\frac{1}{2\beta}}$		$\lambda^{-\frac{1}{2}}$	
$B \ll L$	$\lambda \ll n^{-\beta/(r\beta+1)}$		$\lambda \ll n^{-1/r}$	
Q	$\lambda^{-1}n^{\frac{1-\beta}{6+2\beta}}$	$\lambda^{\frac{-(5+\beta)}{4+2\beta}}n^{\frac{1-\beta}{4+2\beta}}$	$\lambda^{-1}n^{-\frac{1}{2}}$	$(\lambda n)^{-\frac{1}{2}}$
R	$\left(\lambda^{3+\frac{1}{\beta}+\frac{2}{\beta-1}}n\right)^{\frac{1-\beta}{4\beta-2}}$	$\left(\lambda^{2+\frac{2}{\beta}+\frac{2}{\beta-1}}n\right)^{\frac{1-\beta}{4\beta-2}}$	$\lambda^{-3/4}n^{-1/4}$	$\lambda^{-1/2}n^{-1/4}$
$Q+R \ll L$	$\lambda \gg n^{-\beta/(\beta+3)}$	$\lambda \gg n^{-\beta/2}$	$\lambda \gg n^{-1}$	$\lambda \gg 0$
Minimax?	$r > 1 + 2/\beta$	✓	✓	✓

Table 7: Summary of results under different assumptions, suppressing log factors. The initial two rows present rates for the bias upper bound B and variance lower bound L. The third row presents the restriction on λ implied by $B \ll L$. The fourth and fifth rows present rates for the Gaussian coupling Q and bootstrap coupling R. The sixth row presents the restriction on λ implied by $Q + R \ll L$. The final row evaluates whether the allowed path for λ approaches the minimax optimal choice for learning. As $\beta \to \infty$, the rates under polynomial decay recover the rates under exponential decay.

7 Case study: Heterogeneity by preferences

7.1 Economic research question

In the past forty years, centralized school assignment mechanisms have proliferated, alongside research on alternatives to traditional neighborhood-based education in the United States. A thorough investigation of these mechanisms led many districts to adopt variants of the classical Deferred Acceptance (DA) or Random Serial Dictatorship (RSD) mechanisms (Abdulkadiroğlu and Sönmez, 2003; Abdulkadiroğlu et al., 2005; Pathak and Sönmez, 2008). Students submit rank-ordered lists of schools to a central authority, who uses the lists to assign schools. These mechanisms are strategy-proof:

under rational behavior, students truthfully report their preferences.

Centralized assignment, and school choice reforms more broadly (e.g. vouchers), are nonetheless controversial. Learning and information frictions may limit the benefits of choice (Narita, 2018), and choice reforms may have nuanced supply-side effects (Bau, 2022). An important question in this literature concerns match effects: is there systematic heterogeneity in learning quality across students at a given school, and, if so, is this heterogeneity reflected in student preferences? The answer has implications for the potential welfare gains from choice, and the extent to which they are realized.

Prior approaches are somewhat indirect: some examples are (i) testing the heterogeneity explained by the gap between a student's baseline achievement level and the school's median level (Angrist et al., 2023), or (ii) estimating structural models of demand to explain student preferences in terms of observable quantities (Bau, 2022; Narita, 2018). We instead pursue a direct approach, using KRR inference to model and test for heterogeneity in school effects depending upon reported preferences. Building on Abdulkadiroğlu et al. (2017), we use unified enrollment lotteries to identify school effects conditional upon reported preferences. Thus, it is possible to study directly whether students' ranked preferences explain heterogeneity in school effects, without estimating an underlying choice model. Because KRR exploits latent, low-dimensional structure in ranking data despite their dimensionality, and because we provide new methods for inference, this direct approach becomes tractable.

7.2 Semi-synthetic preference data

We generate synthetic student ranked preference lists for Boston Public School students by combining publicly available data sets. Though the underlying student-level micro data are not publicly available, Pathak and Shi (2021) report coefficients from models estimated using real preference data. Pathak and Shi (2021) show that their model reproduces the distribution of student preferences in subsequent years with reasonable accuracy, motivating this procedure.

We generate each student ranked preference list as follows. First, we draw a location and covariate vector for that student from the distribution represented in the American Community Survey (ACS). We also obtain school locations and other covariates from the Massachusetts Department of Elementary and Secondary Education (DESE). Next, we compute walking distances from that student's location to various schools, using the Google Maps API for distances. We then use the random utility model of Pathak and Shi (2021) to generate the preference list. See Appendix L for details. We use these preference lists and the standard RSD mechanism to assign students to schools.

7.3 A kernel for rankings

The data consist of 4000 synthetic students' rankings of 25 Boston high-schools, so the ambient dimension is larger than 10^{25}. For this reason, the direct approach is intractable with standard methods. We now describe how our KRR inference procedure makes the direct approach tractable, i.e. relatively efficient to implement and adaptive to latent, low-dimensional structure in student preferences. In particular, we demonstrate that our inferential procedure detects heterogeneity with non-trivial power.

Previous work shows the effectiveness of KRR for ranking data (Kondor and Barbosa, 2010; Jiao and Vert, 2015; Mania et al., 2018), though without the uniform inference guarantees necessary to test for match effects. Following Mania et al. (2018), in our implementation we use the Mallows kernel for ranking data.

Definition 7.1 (Mallows kernel; Mania et al., 2018)**.** Given a set of q alternatives $\mathcal{A}$ and two complete ranked preference lists $\pi, \pi' : \mathcal{A} \to \{1, \dots, q\}$, the number of discordant pairs is $N(\pi, \pi') := \#\left\{(s,t) \in \mathcal{A} \times \mathcal{A} \,|\, \pi'(s) > \pi(s), \pi'(t) < \pi(t)\right\}$. It is the number of unordered pairs $\{s,t\}$ such that π prefers s to t but π' does not, or vice versa. The Mallows kernel is $k(\pi, \pi') := \exp\{-\ell N(\pi, \pi')\}$, where ℓ is called the length-scale.

Intuitively, two preference lists are similar if they make similar pairwise comparisons. The following result justifies this kernel choice. It implies that the RKHS for the Mallows kernel is automatically well-specified.

Lemma 7.2 (Universality; Mania et al., 2018, Theorem 5)**.** The Mallows kernel is universal. Formally, if μ is a probability distribution over the set of rank-ordered lists, then the RKHS for the Mallows kernel contains all functions $\text{supp}(\mu) \to \mathbb{R}$.

Remark 7.3 (Universality)**.** Conventionally, a kernel is universal if and only if H is dense in $L^2(\mu)$. In the case of rankings, this stronger statement is possible because the set of rank-ordered lists is finite.

7.4 Do students who highly rank schools benefit more?

Finally, we use our inferential procedure to detect heterogeneity in school sector effects. Under the RSD lottery, sector assignment is random conditional upon a student's ranked preference list (Abdulkadiroğlu et al., 2017). Moreover, the treatment propensity is determined completely by the RSD lottery, and may be computed to arbitrary precision by running the RSD lottery many times. Thus, conditional school-sector effects are non-parametrically identified as a function of student preferences.

We use KRR inference to distinguish between two counterfactual scenarios: no match effects versus match effects. In the no match effect scenario, we generate student preferences as described above. We draw school effects as described in Appendix L. Here, student taste parameters, which correspond to random coefficients from the model of Pathak and Shi (2021), do not incorporate information about student-specific match effects. In the match effect scenario, we modify student preferences to incorporate a random, latent taste parameter χ_i which represents a preference for pilot sector schools. This taste parameter also determines sector effects, so that students who prefer pilot sector schools also benefit more from them. See Appendix L for details.

To begin, Figure 4 verifies our key assumption that the local width is small. Using the student rank lists generated by the mixed logit model of Pathak and Shi (2021), and using the Mallows kernel, we calculate the kernel matrix K as described in Section 3. Following Mania et al. (2018), we set the hyperparameter ℓ equal to the median number of discordant pairs. We plot the top 50 eigenvalues of K. After the top 25 eigenvalues,

corresponding to the number of schools, the remaining eigenvalues are negligible. Our arguments based upon local width exploit this fact.

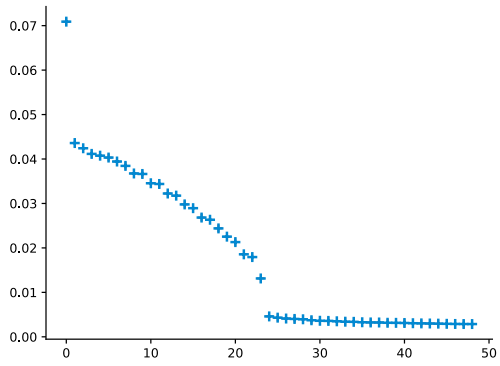


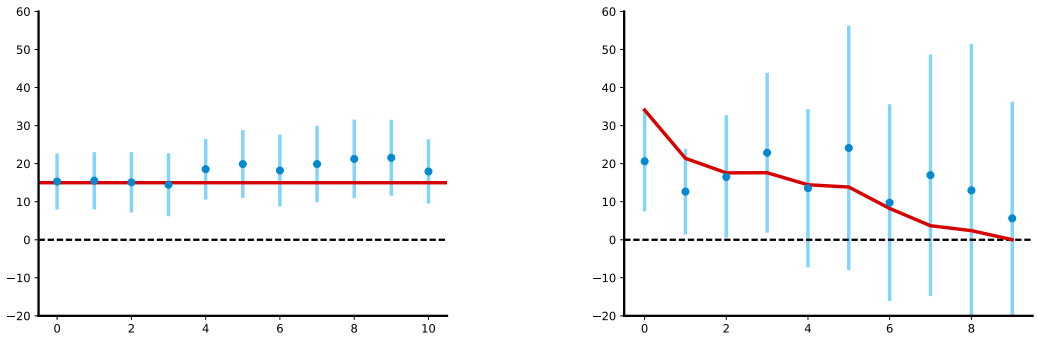
Figure 4: KRR exploits latent low-dimensional structure in student rankings.

To estimate the conditional average treatment effect function, given that the treatment propensity score is exactly known, we regress the inverse propensity weighted outcome $\tilde{Y}_i = \frac{Y_i D_i}{p(X_i)} - \frac{Y_i(1-D_i)}{1-p(X_i)}$ on student reported rankings X_i using KRR. This provides an unbiased signal of the conditional average treatment effect, which is γ_1 in the former scenario or $\gamma_1 + \gamma_2 \mathbb{E}[\chi_i | X_i]$ in the latter scenario, where $\gamma_1, \gamma_2 \in \mathbb{R}$. The simplicity of this function loosely corresponds to our assumption that f_0 is smooth.

To perform interpretable inference, we divide student preference lists into groups S_ρ, where ρ is the highest ranking assigned to a pilot school in a student's preference list. We then perform simultaneous inference on the full vector of group averages, using our theoretical results to study the statistic $F(\hat{f} - f_0) = \max_{1 \leq \rho \leq 25} \mathbb{E}_n\{(\hat{f} - f_0)(X_i) | X_i \in S_\rho\}$, which is continuous with respect to the sup norm.

Figure 5 demonstrates that our KRR inference procedure successfully distinguishes between the two scenarios. We display group average treatment effects, where groups are defined by the highest rank given to a pilot school. Light blue bars are simultaneous confidence bands obtained using our procedure, and red lines denote the true group averages. In both scenarios, we cover the truth. In the match effect scenario, we reject the null hypothesis of no effect for students who highly rank pilot schools.

Interestingly, the average treatment effect is not identified by simple group averages, since the treatment propensity is non-constant within the group for $\rho > 1$. Therefore comparing treatment and control means directly within rank strata is not sufficient to identify heterogeneity in this experiment.[20] See Appendix L for further discussion.



(a) No match effects.

(b) Match effects.

Figure 5: KRR inference detects match effects.

8 Conclusion

The increasing availability of truthfully reported preference data, as well as other complex types of economic data, motivates us to study KRR inference. Theoretically, our goal is to prove uniform inference guarantees; practically, our goal is to create new tools for economic data. We propose new uniform confidence bands that have simple closed form solutions, strong statistical guarantees, and robust empirical performance across nonlinear settings. While Section 6 focuses on KRR, Section 5 builds a framework to use Gaussian and bootstrap couplings for more general kernel methods. Future research may extend our results to non-parametric estimands beyond regression.

[20] Group average treatment effects, as reported above, are identified by averaging inverse propensity weighted outcomes within rank strata. We find in Appendix L that this method is not precise enough to detect heterogeneity; KRR dramatically improves precision by conditioning upon the full rank list.

References

Abdulkadiroğlu, A., Angrist, J. D., Narita, Y., and Pathak, P. A. (2017). Research design meets market design: Using centralized assignment for impact evaluation. *Econometrica*, 85(5):1373–1432.

Abdulkadiroğlu, A., Pathak, P. A., Roth, A. E., and Sönmez, T. (2005). The Boston public school match. *American Economic Review*, 95(2):368–371.

Abdulkadiroğlu, A. and Sönmez, T. (2003). School choice: A mechanism design approach. *American Economic Review*, 93(3):729–747.

Aizerman, A. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837.

Andrews, D. W. and Shi, X. (2013). Inference based on conditional moment inequalities. *Econometrica*, 81(2):609–666.

Angrist, J. D., Pathak, P. A., and Zarate, R. A. (2023). Choice and consequence: Assessing mismatch at Chicago exam schools. *Journal of Public Economics*, 223:104892.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48.

Bau, N. (2022). Estimating an equilibrium model of horizontal competition in education. *Journal of Political Economy*, 130(7):1717–1764.

Belkin, M. (2018). Approximation beats concentration? An approximation view on inference with smooth radial kernels. In *Conference On Learning Theory*, pages 1348–1361. PMLR.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media.

Berthet, P. and Mason, D. M. (2006). Revisiting two strong approximation results of Dudley and Philipp. In *High Dimensional Probability*, volume 51, pages 155–173. Institute of Mathematical Statistics.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics*, 1(6):1071–1095.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Computational Learning Theory*, pages 144–152.

Buzun, N., Shvetsov, N., and Dylov, D. V. (2022). Strong Gaussian approximation for the sum of random vectors. In *Conference on Learning Theory*, pages 1693–1715. PMLR.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368.

Carrasco, M., Florens, J.-P., and Renault, E. (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of Econometrics*, 6:5633–5751.

Chen, X. and Pouzo, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321.

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

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

Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, 162:47–70.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2016). Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related Gaussian couplings. *Stochastic Processes and their Applications*, 126(12):3632–3651.

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

Chernozhuokov, V., Chetverikov, D., Kato, K., and Koike, Y. (2022). Improved central limit theorem and bootstrap approximations in high dimensions. *The Annals of Statistics*, 50(5):2562–2586.

Ciliberto, C., Rosasco, L., and Rudi, A. (2020). A general framework for consistent structured prediction with implicit loss embeddings. *Journal of Machine Learning Research*, 21(98):1–67.

Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565.

de la Peña, V. H. and Giné, E. (1999). Sums of independent random variables. *Decoupling: From Dependence to Independence*, pages 1–50.

Dirksen, S. (2015). Tail bounds via generic chaining. *Electronic Journal of Probability*, 20(53):1–29.

Einmahl, U. and Mason, D. M. (1997). Gaussian approximation of local empirical processes indexed by functions. *Probability Theory and Related Fields*, 107:283–311.

Fischer, S. and Steinwart, I. (2020). Sobolev norm learning rates for regularized least-squares algorithms. *The Journal of Machine Learning Research*, 21(1):8464–8501.

Gale, D. and Shapley, L. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15.

Giné, E. and Nickl, R. (2008). Uniform central limit theorems for kernel density estimators. *Probability Theory and Related Fields*, 141(3-4):333–387.

Giné, E. and Nickl, R. (2009). Uniform limit theorems for wavelet density estimators. *The Annals of Probability*, 37(4):1605–1646.

Giné, E. and Nickl, R. (2021). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press.

Götze, F. and Zaitsev, A. Y. (2011). Estimates for the rate of strong approximation in Hilbert space. *Siberian Mathematical Journal*, 52:628–638.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory*, pages 63–77. Springer.

Hable, R. (2012). Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis*, 106:92–117.

Hall, P. and Horowitz, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904–2929.

Hoffmann-Jørgensen, J. and Pisier, G. (1976). The law of large numbers and the central limit theorem in Banach spaces. *The Annals of Probability*, 4(4):587–599.

Horowitz, J. L. and Lee, S. (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, 100(472):1238–1249.

Jiao, Y. and Vert, J.-P. (2015). The Kendall and Mallows kernels for permutations. In *International Conference on Machine Learning*, pages 1935–1944. PMLR.

Kasy, M. (2018). Optimal taxation and insurance using machine learning—sufficient statistics and beyond. *Journal of Public Economics*, 167:205–219.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95.

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.

Komlós, J., Major, P., and Tusnády, G. (1975). An approximation of partial sums of independent R.V.'s, and the sample D.F. I. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32:111–131.

Kondor, R. and Barbosa, M. S. (2010). Ranking with kernels in Fourier space. In *Conference on Learning Theory*, pages 451–463.

Kutateladze, V. (2022). The kernel trick for nonlinear factor modeling. *International Journal of Forecasting*, 38(1):165–177.

Lecué, G. and Mendelson, S. (2013). Learning sub-Gaussian classes: Upper and minimax bounds. *arXiv:1305.4825*.

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

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

Mania, H., Ramdas, A., Wainwright, M. J., Jordan, M. I., and Recht, B. (2018). On kernel methods for covariates that are rankings. *Electronic Journal of Statistics*, 12(2):2537–2577.

Massart, P. (1989). Strong approximation for multivariate empirical and related processes, via KMT constructions. *The Annals of Probability*, 17(1):266–291.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884.

Mendelson, S. and Neeman, J. (2010). Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565.

Monrad, D. and Philipp, W. (1991). Nearby variables with nearby conditional laws and a strong approximation theorem for Hilbert space valued martingales. *Probability Theory and Related Fields*, 88(3):381–404.

Narita, Y. (2018). Match or mismatch? Learning and inertia in school choice. *SSRN:3198417*.

Natalini, P. and Palumbo, B. (2000). Inequalities for the incomplete Gamma function. *Mathematical Inequalities and Applications*, 3(1):69–77.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.

Pathak, P. A. and Shi, P. (2021). How well do structural demand models work? Counterfactual predictions in school choice. *Journal of Econometrics*, 222(1):161–195.

Pathak, P. A. and Sönmez, T. (2008). Leveling the playing field: Sincere and sophisticated players in the Boston mechanism. *American Economic Review*, 98(4):1636–1652.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian Processes for Machine Learning*, volume 2. MIT Press.

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

Shang, Z. and Cheng, G. (2013). Local and global asymptotic inference in smoothing spline models. *The Annals of Statistics*, 41(5):2608–2638.

Singh, R., Sahani, M., and Gretton, A. (2019). Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, pages 4593–4605.

Singh, R., Xu, L., and Gretton, A. (2023). Kernel methods for causal functions: Dose, heterogeneous, and incremental response curves. *Biometrika (forthcoming)*.

Smale, S. and Zhou, D.-X. (2005). Shannon sampling II: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302.

Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172.

Sriperumbudur, B. (2016). On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, pages 583–602.

Stute, W. (1986). Conditional empirical processes. *The Annals of Statistics*, pages 638–647.

Talagrand, M. (2014). *Upper and Lower Bounds for Stochastic Processes*. Springer.

van der Vaart, A. and van Zanten, J. (2008a). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463.

van der Vaart, A. W. and van Zanten, J. H. (2008b). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, pages 200–222. Institute of Mathematical Statistics.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press.

Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. (2010). Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(3):364–372.

Wei, Y., Fang, B., and Wainwright, M. J. (2020). From Gauss to Kolmogorov: Localized measures of complexity for ellipses. *Electronic Journal of Statistics*, 14:2988–3031.

Wendland, H. (2004). *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

Wihler, T. P. (2009). On the Hölder continuity of matrix functions for normal matrices. *Journal of Inequalities in Pure and Applied Mathematics*, 10:1–5.

Yang, Y., Bhattacharya, A., and Pati, D. (2017). Frequentist coverage and sup-norm convergence rate in Gaussian process regression. *arXiv:1708.04753*.

Yurinsky, V. (1995). *Sums and Gaussian Vectors*. Lecture Notes in Mathematics. Springer.

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

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

Part
Appendix

Table of Contents

A Glossary

Symbol	Meaning	Definition
H	RKHS	Sec. 3
k	kernel function	Sec. 3
k_x	$y \mapsto k(x, y)$	Sec. 3
κ	$\|k\|_\infty^{1/2}$	Sec. 3
$\langle -, - \rangle$	H inner product	Sec. 3
$\|-\|$	H norm	Sec. 3
$\|-\|_{\mathbb{R}^n}$	Euclidean norm	Sec. 3
$\|-\|_{\mathrm{HS}}$	Hilbert-Schmidt norm	Sec. 3
$\|-\|_{op}$	operator norm	Sec. 3
$u \otimes v^*$	$w \mapsto u \langle v, w \rangle$	Sec. 3
$\nu_s(A)$	s^{th} eigenvalue of A	Sec. 3
$e_s(A)$	eigenvector corresponding to $\nu_s(A)$	Sec. 3
f_0	square-loss minimizer	Sec. 3
λ	regularization parameter	Sec. 3
$\hat{f}$	KRR estimator	Sec. 3
f_λ	pseudo-true parameter	Sec. 3
$D = \{(X_i, Y_i)\}_{i=1}^n$	dataset	Sec. 3
ε_i	$Y_i - f_0(X_i)$	Sec. 3
$\bar{\sigma}^2$	ess sup ε_i^2	Sec. 3
ent_m	m^{th} entropy number	Sec. 3
$\sigma(\Sigma, m)$	local width	Sec. 3
K	sample kernel matrix	Sec. 4
K_x	$\{k(x, X_i)\}_{i=1}^n$	Sec. 4
$\mathbf{1}$	ones vector	Sec. 4
$\mathfrak{B}$	feasible bootstrap	Sec. 4

Table 8: General and estimator notation

Symbol	Meaning	Definition
U_i	centered i.i.d. summand	Sec. 5
Σ	$\mathbb{E}(U_i \otimes U_i^*)$	Sec. 5
Z	approximating Gaussian r.v.	Sec. 5
$Z_{\mathfrak{B}}$	symmetrized bootstrap	Sec. 5
V_i	un-centered i.i.d. summand	Sec. 5
$\hat{\Sigma}$	$\mathbb{E}(Z_{\mathfrak{B}} \otimes Z_{\mathfrak{B}}^* \vert U)$	Sec. 5
T	$\mathbb{E}(k_{X_i} \otimes k_{X_i}^*)$	Sec. 6
$\hat{T}$	$\frac{1}{n}\sum_i (k_{X_i} \otimes k_{X_i}^*)$	Sec. 6
$\sigma(m)$	$\sigma(T, m)$	Sec. 6
$\mathfrak{n}(\lambda)$	$\sigma^2\{(T+\lambda)^{-2}T, 0\}$	Sec. 6
F	functional $H \to \mathbb{R}$	Sec. 6
ψ	modulus of continuity of F	Sec. 6
Q_{bd}, Q_{sg}	Gaussian coupling rate	Sec. 6
R_{bd}, R_{sg}	bootstrap coupling rate	Sec. 6
$\underline{\sigma}^2$	ess inf $\mathbb{E}(\varepsilon_i^2 \vert X_i)$	Sec. 6
$\hat{\varepsilon}_i$	$Y_i - \hat{f}(X_i)$	Sec. 6
$\hat{V}$	$\{Y_i - \hat{f}(X_i)\}(\hat{T}+\lambda)^{-1}k_{X_i}$	Sec. 6

Table 9: Theoretical notation

Symbol	Meaning	Definition
$\mathbb{E}_n$	sample expectation $\frac{1}{n}\sum_i(-)$	Apx. B
T_λ	$(T+\lambda)$	Apx. B
$\hat{T}_\lambda$	$(\hat{T}+\lambda)$	Apx. B
Π_m	projection onto $\mathrm{span}\{e_1(\Sigma),\ldots,e_s(\Sigma)\}$	Apx. C
$I, 1$	identity operator	Apx. C
$\omega, \alpha, \beta, \gamma$	spectral decay parameters	Apx. C
$\Pi_m^\perp$	$I - \Pi_m$	Apx. D
$\overset{U}{\sim}$	equal in $\sigma(U)$-conditional law	Apx. D
$g = (g_i)_{i=1}^\infty$,	i.i.d. standard Gaussian variables	Apx. D
$(h_{ij})_{i,j=1}^\infty$	i.i.d. standard Gaussian variables	Apx. D
h	$n \times n$ Gaussian random matrix $(h_{ij})_{i,j=1}^n$	Apx. D
$(q_i)_{i=1}^\infty$,	non-i.i.d. Gaussian variables	Apx. D
$\hat{\varepsilon}_i^\lambda$	$Y_i - f_\lambda(X_i)$	Apx. E
z_0	$\|\hat{f} - f_0\|$	Apx. E
Φ	sampling operator $f \mapsto (\langle k_{X_i}, f\rangle)_{i=1}^n$	Apx. F
$\mathcal{G}$	Gaussian complexity	Apx. I
$\psi(a, m)$	$\sum_{s>m} \nu_s(T)/\{\nu_s(T)+\lambda\}^a$	Apx. I
$d(-,-)$	metric	Apx. J
$\gamma_2(T, d)$	Talagrand's γ_2 functional	Apx. J
$\|-\|_p$	$L^p(\mathbb{P})$ norm	Apx. J

Table 10: Appendix notation

B Bahadur representation

The sketch in Section 6 has the step

$$\sqrt{n}(\hat{f} - f_\lambda) \approx \sqrt{n}T_\lambda^{-1}\{(\hat{T} - T)(f_0 - f_\lambda) + \mathbb{E}_n[k_{X_i}\varepsilon_i]\} = \sqrt{n}\mathbb{E}_n[U_i].$$

We prove this approximation in what follows.

B.1 First order lemmas

Lemma B.1 (Higher-order resolvent)**.** Let V be a vector space and $A, B : V \to V$ be invertible linear operators. Then, for all $l \geq 1$, it holds

$$A^{-1} - B^{-1} = A^{-1}\{(B-A)B^{-1}\}^l + \sum_{r=1}^{l-1} B^{-1}\{(B-A)B^{-1}\}^r.$$

Proof. Note that when $l = 1$ this reduces to the familiar "resolvent identity"

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1} \iff A^{-1} = A^{-1}(B - A)B^{-1} + B^{-1}.$$

We then prove by induction. Supposing the inequality holds for $l - 1$, we have

$$A^{-1} - B^{-1} = A^{-1}\{(B - A)B^{-1}\}^{l-1} + \sum_{r=1}^{l-2} B^{-1}\{(B - A)B^{-1}\}^r.$$

Plugging in the resolvent identity again for the left-most appearance of A^{-1} gives

$$\begin{aligned}
&= \{A^{-1}(B - A)B^{-1} + B^{-1}\}\{(B - A)B^{-1}\}^{l-1} + \sum_{r=1}^{l-2} B^{-1}\{(B - A)B^{-1}\}^r \\
&= A^{-1}\{(B - A)B^{-1}\}^l + B^{-1}\{(B - A)B^{-1}\}^{l-1} + \sum_{r=1}^{l-2} B^{-1}\{(B - A)B^{-1}\}^r \\
&= A^{-1}\{(B - A)B^{-1}\}^l + \sum_{r=1}^{l-1} B^{-1}\{(B - A)B^{-1}\}^r.
\end{aligned}$$

This proves the inductive hypothesis. □

Lemma B.2. Suppose $\|T_\lambda^{-1}(\hat{T} - T)\|_{\text{HS}} \leq \delta < 1$. Then for all $k \geq 1$

$$(\hat{T}_\lambda^{-1} - T_\lambda^{-1})u = A_1 u + A_2 T_\lambda^{-1} u + A_3 T_\lambda^{-1} u$$

where

$$\|A_1\|_{\text{HS}} \leq \frac{\delta^k}{\lambda}, \quad \|A_2\|_{\text{HS}} \leq \delta, \quad \|A_3\|_{\text{HS}} \leq \frac{\delta^2}{1 - \delta}.$$

Proof. By applying Lemma B.1 with $A = \hat{T}_\lambda = \hat{T} + \lambda$ and $B = T_\lambda = T + \lambda$, we obtain

$$\begin{aligned}
\hat{T}_\lambda^{-1} - T_\lambda^{-1} &= \hat{T}_\lambda^{-1}\{(T - \hat{T})T_\lambda^{-1}\}^k + \sum_{r=1}^{k-1} T_\lambda^{-1}\{(T - \hat{T})T_\lambda^{-1}\}^r \\
&= \hat{T}_\lambda^{-1}\{(T - \hat{T})T_\lambda^{-1}\}^k + T_\lambda^{-1}(T - \hat{T})T_\lambda^{-1} + T_\lambda^{-1}\sum_{r=2}^{k-1}\{(T - \hat{T})T_\lambda^{-1}\}^r \\
&= \hat{T}_\lambda^{-1}\{(T - \hat{T})T_\lambda^{-1}\}^k + T_\lambda^{-1}(T - \hat{T})T_\lambda^{-1} + \left[\sum_{r=2}^{k-1}\{T_\lambda^{-1}(T - \hat{T})\}^r\right] T_\lambda^{-1} \\
&=: A_1 + A_2 T_\lambda^{-1} + A_3 T_\lambda^{-1},
\end{aligned}$$

where in the third equality we've used that $B(AB)^r = (BA)^r B$. Since $\|(T-\hat{T})T_\lambda^{-1}\|_{\text{HS}} \leq \delta < 1$,

$$\|A_1\|_{\text{HS}} = \|\hat{T}_\lambda^{-1}\{(T-\hat{T})T_\lambda^{-1}\}^k\|_{\text{HS}} \leq \frac{\delta^k}{\lambda}, \quad \|A_2\|_{\text{HS}} = \|T_\lambda^{-1}(T-\hat{T})\|_{\text{HS}} \leq \delta,$$

where in the first inequality we used that $\|AB\|_{\text{HS}} \leq \|A\|_{op}\|B\|_{\text{HS}} \leq \|A\|_{\text{HS}}\|B\|_{\text{HS}}$. Using this again along with the triangle inequality for $\|-\|_{\text{HS}}$, we find

$$\|A_3\|_{\text{HS}} = \left\|\sum_{r=2}^{k-1}\{T_\lambda^{-1}(T-\hat{T})\}^r\right\|_{\text{HS}} \leq \sum_{r=2}^{k-1}\left\|\{T_\lambda^{-1}(T-\hat{T})\}\right\|_{\text{HS}}^r \leq \sum_{r=2}^{k-1}\delta^r \leq \sum_{r=2}^{\infty}\delta^r = \frac{\delta^2}{1-\delta}.$$

□

B.2 Main result

Proposition B.3. Suppose $\|T_\lambda^{-1}(\hat{T}-T)\|_{\text{HS}} \leq \delta \leq \frac{1}{2}$, and $\left\|T_\lambda^{-1}\mathbb{E}_n[k_{X_i}\varepsilon_i]\right\| \leq \gamma$. Then

$$\hat{f} - f_\lambda = T_\lambda^{-1}\mathbb{E}_n[k_{X_i}\varepsilon_i] + T_\lambda^{-1}(\hat{T}-T)(f_0 - f_\lambda) + u \tag{2}$$

for some u with

$$\|u\| \leq \frac{\delta}{1-\delta}\left(\gamma + \delta\|f_0 - f_\lambda\|\right) \leq 2\delta\left(\gamma + \delta\|f_0 - f_\lambda\|\right).$$

Proof. We proceed in steps.

1. Decomposition. Write

$$\hat{f} - f_\lambda = \hat{T}_\lambda^{-1}\mathbb{E}_n[k_{X_i}Y_i] - T_\lambda^{-1}Tf_0 = \hat{T}_\lambda^{-1}\hat{T}f_0 + \hat{T}_\lambda^{-1}\mathbb{E}_n[k_{X_i}\varepsilon_i] - T_\lambda^{-1}Tf_0.$$

Adding and subracting the terms $T_\lambda^{-1}\mathbb{E}_n[k_{X_i}\varepsilon_i]$ and $T_\lambda^{-1}\hat{T}f_0 - T_\lambda^{-1}Tf_0$, this is

$$\begin{aligned}
&= (\hat{T}_\lambda^{-1} - T_\lambda^{-1})\mathbb{E}_n[k_{X_i}\varepsilon_i] + T_\lambda^{-1}\mathbb{E}_n[k_{X_i}\varepsilon_i] \\
&\quad + (\hat{T}_\lambda^{-1} - T_\lambda^{-1})\hat{T}f_0 + T_\lambda^{-1}\hat{T}f_0 - T_\lambda^{-1}Tf_0 \\
&= (\hat{T}_\lambda^{-1} - T_\lambda^{-1})\mathbb{E}_n[k_{X_i}\varepsilon_i] + T_\lambda^{-1}\mathbb{E}_n[k_{X_i}\varepsilon_i] \\
&\quad + (\hat{T}_\lambda^{-1} - T_\lambda^{-1})(\hat{T}-T)f_0 + (\hat{T}_\lambda^{-1} - T_\lambda^{-1})Tf_0 + T_\lambda^{-1}\hat{T}f_0 - T_\lambda^{-1}Tf_0 \\
&= (\hat{T}_\lambda^{-1} - T_\lambda^{-1})\{\mathbb{E}_n[k_{X_i}\varepsilon_i] \\
&\quad + (\hat{T}-T)f_0\} + T_\lambda^{-1}\mathbb{E}_n[k_{X_i}\varepsilon_i] \\
&\quad - \hat{T}_\lambda^{-1}(\hat{T}-T)f_\lambda + T_\lambda^{-1}(\hat{T}-T)f_0
\end{aligned}$$

where we used the transformation

$$(\hat{T}_\lambda^{-1} - T_\lambda^{-1})Tf_0 = \hat{T}_\lambda^{-1}(T - \hat{T})T_\lambda^{-1}Tf_0 = \hat{T}_\lambda^{-1}(T - \hat{T})f_\lambda = -\hat{T}_\lambda^{-1}(\hat{T} - T)f_\lambda.$$

Focusing on the final two terms

$$\begin{aligned}
-\hat{T}_\lambda^{-1}(\hat{T} - T)f_\lambda + T_\lambda^{-1}(\hat{T} - T)f_0 &= -\hat{T}_\lambda^{-1}(\hat{T} - T)f_\lambda + T_\lambda^{-1}(\hat{T} - T)f_0 \pm T_\lambda^{-1}(\hat{T} - T)f_\lambda \\
&= (T_\lambda^{-1} - \hat{T}_\lambda^{-1})(\hat{T} - T)f_\lambda + T_\lambda^{-1}(\hat{T} - T)(f_0 - f_\lambda) \\
&= (\hat{T}_\lambda^{-1} - T_\lambda^{-1})(\hat{T} - T)(-f_\lambda) + T_\lambda^{-1}(\hat{T} - T)(f_0 - f_\lambda).
\end{aligned}$$

In summary,

$$\begin{aligned}
\hat{f} - f_\lambda = (\hat{T}_\lambda^{-1} - T_\lambda^{-1})\{\mathbb{E}_n[k_{X_i}\varepsilon_i] + (\hat{T} - T)(f_0 - f_\lambda)\} \\
+ T_\lambda^{-1}\mathbb{E}_n[k_{X_i}\varepsilon_i] + T_\lambda^{-1}(\hat{T} - T)(f_0 - f_\lambda)
\end{aligned}$$

i.e.

$$\hat{f} - f_\lambda = (\hat{T}_\lambda^{-1} - T_\lambda^{-1})S_n + T_\lambda^{-1}S_n, \quad S_n \coloneqq \mathbb{E}_n[k_{X_i}\varepsilon_i] + (\hat{T} - T)(f_0 - f_\lambda).$$

2. What remains is to control the first term:

$$(\hat{T}_\lambda^{-1} - T_\lambda^{-1})S_n, \quad S_n \coloneqq \mathbb{E}_n[k_{X_i}\varepsilon_i] + (\hat{T} - T)(f_0 - f_\lambda).$$

Recall Lemma B.2:

$$(\hat{T}_\lambda^{-1} - T_\lambda^{-1})u = A_1u + A_2T_\lambda^{-1}u + A_3T_\lambda^{-1}u$$

where

$$\|A_1\|_{\text{HS}} \leq \frac{\delta^k}{\lambda}, \quad \|A_2\|_{\text{HS}} \leq \delta, \quad \|A_3\|_{\text{HS}} \leq \frac{\delta^2}{1-\delta}.$$

Therefore

$$\begin{aligned}
\|(\hat{T}_\lambda^{-1} - T_\lambda^{-1})S_n\| &\leq \|A_1\|_{\text{HS}} \cdot \|S_n\| + \|A_2\|_{\text{HS}} \cdot \|T_\lambda^{-1}S_n\| + \|A_3\|_{\text{HS}} \cdot \|T_\lambda^{-1}S_n\| \\
&\leq \frac{\delta^k}{\lambda}\|S_n\| + \left(\delta + \frac{\delta^2}{1-\delta}\right)\|T_\lambda^{-1}S_n\| \\
&= \frac{\delta^k}{\lambda}\|S_n\| + \frac{\delta}{1-\delta}\|T_\lambda^{-1}S_n\| \\
&\leq \frac{\delta^k}{\lambda}\|S_n\| + \frac{\delta}{1-\delta}\left(\gamma + \delta\|f_0 - f_\lambda\|\right).
\end{aligned}$$

Taking $k \uparrow \infty$ removes the first term, which finishes the proof.

□

Theorem B.4. Suppose that $n \geq 16\kappa^2 \ln(4/\eta)^2[\mathfrak{n}(\lambda) \vee \lambda^{-1}]$. It then holds with probability $1 - \eta$ that

$$\hat{f} - f_\lambda = \mathbb{E}_n[T_\lambda^{-1} k_{X_i} \varepsilon_i] + T_\lambda^{-1}(\hat{T} - T)(f_0 - f_\lambda) + u$$

for some u with

$$\|u\| \leq 8(\kappa^2 \|f_0 - f_\lambda\| + \bar{\sigma}\kappa) \ln(4/\eta)^2 \left\{ \sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \vee \frac{2\kappa}{n\lambda} \right\}^2.$$

Proof. By combining Lemmas H.4 and H.5 with a union bound, we have with probability at least $1 - \eta$ that

$$\left\| \frac{1}{n} \sum_{i=1}^n T_\lambda^{-1} \varepsilon_i k_{X_i} \right\| \leq 2\bar{\sigma} \ln(4/\eta) \left\{ \sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \vee \frac{2\kappa}{n\lambda} \right\} = \gamma$$
$$\left\| \frac{1}{n} \sum_{i=1}^n T_\lambda^{-1}(T_i - T) \right\|_{\mathrm{HS}} \leq 2\kappa \ln(4/\eta) \left\{ \sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \vee \frac{2\kappa}{n\lambda} \right\} = \delta.$$

On the event that these inequalities hold, and provided that $\delta < \frac{1}{2}$, we have from Proposition B.3 that (2) holds with

$$\|u\| \leq 8(\kappa^2 \|f_0\| + \bar{\sigma}\kappa) \ln(4/\eta)^2 \left\{ \sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \vee \frac{2\kappa}{n\lambda} \right\}^2.$$

Moreover, the condition $\delta \leq \frac{1}{2}$ can be seen to hold whenever

$$n \geq 16\kappa^2 \ln(4/\eta)^2 \mathfrak{n}(\lambda) \vee 8\kappa^2 \lambda^{-1} \ln(4/\eta).$$

In particular, consider the two cases

$$2\kappa \ln(4/\eta) \sqrt{\frac{\mathfrak{n}(\lambda)}{n}} < \frac{1}{2} \iff 4\kappa \ln(4/\eta) \sqrt{\mathfrak{n}(\lambda)} < \sqrt{n} \iff 16\kappa^2 \ln(4/\eta)^2 \mathfrak{n}(\lambda) < n$$

and

$$2\kappa \ln(4/\eta) \frac{2\kappa}{n\lambda} < \frac{1}{2} \iff 8\kappa^2 \ln(4/\eta) \frac{1}{\lambda} < n.$$

Since $\eta \in [0, 1]$, $4/\eta \geq e$ and so $\ln(4/\eta)^2 \geq \ln(4/\eta)$. Thus, we see that the condition $n \geq 16\kappa^2 \ln(4/\eta)^2[\mathfrak{n}(\lambda) \vee \lambda^{-1}]$ is sufficient. □

C Gaussian couplings

The sketch in Section 6 claims a high-probability bound on $\left\| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n U_i \right) - Z \right\|$, where Z is a Gaussian random variable in H with covariance $\Sigma = \mathbb{E}[U_i \otimes U_i^*]$ and

$$U_i = T_\lambda^{-1}\{(k_{X_i} \otimes k_{X_i}^* - T)(f_0 - f_\lambda) + \varepsilon_i k_{X_i}\}.$$

We prove this approximation in what follows, via results for general summands U_i.

C.1 Notation and preliminaries

Let $U = (U_1, U_2, \ldots)$ be an i.i.d. sequence of centered random elements in H. Let $\Sigma : H \to H$ denote the covariance operator associated to the summands U_i, so that for all $u, v \in H$ and $i \geq 1$, $\mathbb{E}[\langle u, U_i \rangle \langle v, U_i \rangle] = \langle u, \Sigma v \rangle$. Let us also suppose that $\mathbb{E}\|U_i\|^2 < \infty$. Then Σ is trace-class and self-adjoint, so that we may choose an H-orthonormal basis of Σ-eigenvectors $(e_1, e_2, \ldots)$, with corresponding eigenvalues $(\nu_1, \nu_2, \ldots)$. We also define the numbers $\sigma^2(m) := \sigma^2(\Sigma, m) = \sum_{k>m} \nu_k$, which play a role similar to the metric entropy, giving a quantitative measure of the compactness of Σ and hence the tightness of μ. Given some m, let $\Pi_m = \sum_{i=1}^m e_i \otimes e_i^*$ be the self-adjoint and idempotent projection onto the span of $(e_1, e_2, \ldots, e_m)$.

In this section, our aim will be to construct a Gaussian process $x \mapsto Z(x)$ indexed by $x \in S$ such that the quantity

$$\sup_{x \in S} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n U_i(x) - Z(x) \right| \tag{3}$$

is bounded with high probability in terms of n and Σ; in other words, we wish to construct nonasymptotic Gaussian couplings for RKHS-valued partial sums in L^∞.

C.1.1 Coupling technicalities

In order to construct couplings, we require that the background probability space $(\mathbb{P}, \mathcal{F}, \Omega)$ is sufficiently rich. For example, it is enough to assume that there exists a countable sequence of standard normal random variable independent of U. This is a

minor technicality; ultimately, the couplings in this paper are a logical device used to quantify similarity of distributions and establish validity of inference procedures.

C.2 Discussion

The nonasymptotic nature of our bounds is crucial in inverse problems such as kernel ridge regression, where the complexity of Σ increases with n, and thus the limiting process is non-Donsker. Moreover, the bounds we derive are *universal* in that they only depend on n, the spectrum of Σ, and boundedness of the kernel: $\sup_{x\in S}\|k_x\| = \sup_{x\in S}\sqrt{k(x,x)} \leq \kappa$. Thus, they are applicable across a broad range of settings in which RKHS methods find use. This is to be contrasted with couplings based upon the Hungarian construction for the uniform empirical process.

The first, rather basic observation we make is that when the kernel is bounded, the Gaussian approximation problem (3) can be reduced to Gaussian approximation in H.

Lemma C.1. Let Z be a random element of the RKHS H. Then

$$\sup_{x\in S}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i(x) - Z(x)\right| \leq \kappa\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i - Z\right\|.$$

Proof. Using the reproducing property, we may write

$$\begin{aligned}\sup_{x\in S}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i(x) - Z(x)\right| &= \sup_{x\in S}\left|\left\langle\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i - Z,\, k_x\right\rangle\right| \\ &\leq \left(\sup_{x\in S}\|k_x\|\right)\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i - Z\right\|,\end{aligned}$$

where the last step is Cauchy-Schwartz. □

C.3 Coupling

The crucial ingredient is the following finite-dimensional coupling lemma due to Zaitsev.

Lemma C.2 (Theorem 1.1 of Zaitsev, 1987a)**.** Let $\xi_1, \xi_2, \ldots$ be an independent sequence of random variables in $\mathbb{R}^m$ satisfying $\|\xi_i\|_{\mathbb{R}^m} \leq a$. For each n there is a construction

such that for each i, ζ_i is a Gaussian with the same covariance as ξ_i and

$$\mathbb{P}\left\{\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\xi_i-\zeta_i)\right\|_{\mathbb{R}^m}>\delta\right\}\leq Cm^2\exp\left(\frac{-c\delta\sqrt{n}}{m^2a}\right).$$

In fact, if the vectors are sub-Gaussian in the sense that for all $t\in\mathbb{R}^m$,

$$\log\{\mathbb{E}(\exp\langle\xi_i,\,t\rangle_{\mathbb{R}^m})\}\leq\frac{b^2}{2}\mathbb{E}(\langle\xi_i,\,t\rangle_{\mathbb{R}^m}^2),$$

then we can apply the following, slightly stronger result due to Buzun et al. (2022). The following is an immediate consequence of applying (Buzun et al., 2022, Theorem 3) with $L=2$ and $\nu_0=b$.

Lemma C.3 (Buzun et al., 2022, Theorem 3)**.** Set $\xi=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i$, and suppose $m\leq n$. There then exists a Gaussian random variable $\zeta\in\mathbb{R}^m$ with covariance $\Sigma=\mathbb{E}[\xi_i\xi_i^\top]$ such that with probability $1-\eta$

$$\|\zeta-\xi\|_{\mathbb{R}^m}\lesssim\|\Sigma\|_{op}^{\frac{1}{2}}\,b^3\left(\frac{m^{\frac{3}{2}}\log(n)^2}{\sqrt{n}}\right)(1/\eta)^{1/\log(mn)}.$$

Proof. To lighten notation, we suppress the subscript $\mathbb{R}^m$ for the Euclidean norm and inner product. The precise claim made was that if $X_i=\frac{1}{b}\Sigma^{-\frac{1}{2}}\xi_i$ satisfies

$$\log\{\mathbb{E}(\exp\langle X_i,\,t\rangle)\}\leq\frac{\|t\|^2}{2}$$

for all $t\in\mathbb{R}^m$ such that $\|t\|\leq\mathsf{g}$, where $0.3\mathsf{g}\geq\sqrt{m}$, then with probability $1-e^{-t}$,

$$\|\xi-\zeta\|\leq\|\Sigma\|_{op}^{\frac{1}{2}}\left(\frac{Cb^2m\log(mn)^{\frac{3}{2}}}{\sqrt{n}}+\frac{5b^3m^{\frac{3}{2}}\log(mn)\log(2n)}{\sqrt{n}}\right)e^{t/\log(mn)}.$$

Firstly, note that under our sub-Gaussianity assumption

$$\begin{aligned}\log\{\mathbb{E}(\exp\langle X_i,\,t\rangle)\}&=\log\left\{\mathbb{E}\left(\exp\left\langle\frac{1}{b}\Sigma^{-\frac{1}{2}}\xi_i,\,t\right\rangle\right)\right\}\\&\leq\frac{b^2}{2}\mathbb{E}\left(\left\langle\xi_i,\,\frac{1}{b}\Sigma^{-\frac{1}{2}}t\right\rangle^2\right)=\frac{\|t\|^2}{2},\end{aligned}$$

where we have used the fact that $\Sigma^{-\frac{1}{2}}\xi_i$ is isotropic (by construction). Thus, the stated result indeed applies to our setting.

Since we have assumed $m \le n$ and since $\eta = e^{-t} \iff t = \log(1/\eta)$, we recover

$$\begin{aligned}\|\xi - \zeta\| &\le \|\Sigma\|_{op}^{\frac{1}{2}} \left(\frac{Cb^2 m \log(mn)^{\frac{3}{2}}}{\sqrt{n}} + \frac{5b^3 m^{\frac{3}{2}} \log(mn)\log(2n)}{\sqrt{n}} \right) e^{\log(1/\eta)/\log(mn)} \\ &\lesssim \|\Sigma\|_{op}^{\frac{1}{2}} b^3 \left(\frac{m^{\frac{3}{2}} \log(n)^2}{\sqrt{n}} \right) (1/\eta)^{1/\log(mn)},\end{aligned}$$

where we use $\log(mn) \lesssim \log(n)$ and $e^{\log(1/\eta)/\log(mn)} = \{e^{\log(1/\eta)}\}^{1/\log(mn)} = (1/\eta)^{1/\log(mn)}$.

□

In our setting, we wish to construct a Gaussian element Z such that the distance $\Delta_n := \left\| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n U_i \right) - Z \right\|$ is small with high probability and such that U_i and Z have the same covariance structure. We will do this by projecting the U_i onto an appropriate m-dimensional subspace, then appealing to the m-dimensional coupling results.

C.4 Tail bounds

Lemma C.4 (Tail bound: Bounded)**.** Suppose that $\|U_i\| \le a$ almost surely. Then with probability at least $1 - \eta$,

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n (I - \Pi_m) U_i \right\| \le 2\sigma(m)\sqrt{\log(2/\eta)} \vee \frac{4a \log(2/\eta)}{\sqrt{n}}.$$

Proof. We use Lemma H.1 with $\xi_i = (I - \Pi_m)U_i$. Then $\mathbb{E}[\xi_i] = \mathbb{E}\left[(I - \Pi_m)U_i\right] = (I - \Pi_m)\mathbb{E}\left[U_i\right] = 0$. Moreover, $\|(I - \Pi_m)U_i\| \le \|U_i\| \le a$ and

$$\mathbb{E}\,\|(I - \Pi_m)U_i\|^2 = \mathbb{E}\langle U_i, (I - \Pi_m)U_i\rangle = \mathbb{E}\,\mathrm{tr}\{(I - \Pi_m)U_i \otimes U_i^*\} = \mathrm{tr}\{(I - \Pi_m)\Sigma\} = \sigma^2(m).$$

Combining these, we can bound

$$\mathbb{E}\,\|(I - \Pi_m)U_i\|^\ell = \sigma^2(m)a^{\ell-2} \le \sigma^2(m)a^{\ell-2} \le \frac{\ell!}{2}\sigma^2(m)a^{\ell-2}$$

for $\ell \ge 2$. Thus, by Bernstein's inequality (Lemma H.2) with $A = 2a$, $B = \sqrt{\sigma^2(m)}$, we obtain

$$\left\| \frac{1}{n} \sum_{i=1}^n (I - \Pi_m)U_i \right\| \le 2\left(\sqrt{\frac{\sigma^2(m)\log(2/\eta)}{n}} \vee \frac{2a\log(2/\eta)}{n} \right).$$

Multiplying by $\sqrt{n}$ completes the proof. □

Lemma C.5 (Tail bound: sub-Gaussian)**.** If U_i is b sub-Gaussian then

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1 - \Pi_m) U_i \right\| \lesssim b\sigma(m)\sqrt{\log(2/\eta)} + \frac{b\sigma(m)\log(2/\eta)}{\sqrt{n}}.$$

Proof. By Lemma J.8, we have $(\mathbb{E}\,\|(I - \Pi_m)U_i\|^p)^{\frac{1}{p}} \leq Cb\sqrt{p}\sigma(m)$. Thus we can bound

$$\mathbb{E}\,\|(I - \Pi_m)U_i\|^\ell = \ell^{\ell/2} C^\ell b^\ell \sigma^\ell(m) \leq \ell^{\ell/2} C^\ell b^\ell \sigma^\ell(m) \leq \frac{\ell!}{2} [2Cb\sigma(m)]^\ell$$

for $\ell \geq 2$, using $(1/\sqrt{2})^\ell \ell^{\ell/2} = (\ell/2)^{\ell/2} \leq \ell!$ which follows by taking logs, and $2 \leq (\sqrt{2})^\ell$ for $\ell \geq 2$, in the last step. Thus, applying Bernstein's inequality (Lemma H.2) with $A/2 = B = 2Cb\sigma(m)$, we obtain that with probability $1 - \eta$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (1 - \Pi_m) U_i \right\| \leq 2 \left(\sqrt{\frac{4C^2 b^2 \sigma^2(m) \log(2/\eta)}{n}} \vee \frac{4Cb\sigma(m)\log(2/\eta)}{n} \right).$$

Multiplying by $\sqrt{n}$ and removing universal constants completes the proof. □

Lemma C.6 (Tail bound: Gaussian)**.** For each i, let Z_i be an independent Gaussian element with covariance Σ. Then with probability at least $1 - \eta$,

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1 - \Pi_m) Z_i \right\| \leq \left(1 + \sqrt{2\log(1/\eta)} \right) \sigma(m).$$

Proof. Since we are concerned with the norm of a Gaussian random element, it suffices to compute

$$\begin{aligned} \mathbb{E} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1 - \Pi_m) Z_i \right\|^2 &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}\,\langle (1 - \Pi_m) Z_i,\, (1 - \Pi_m) Z_j \rangle \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\,\|(1 - \Pi_m) Z_i\|^2 = \sigma^2(m), \end{aligned}$$

where the last step is identical to the proof in Lemma C.4 since Z_i has covariance Σ. The result then follows from Gaussian concentration (Lemma D.5). □

C.5 Technical result

Proposition C.7 (Bounded coupling)**.** Suppose $\|U_i\| \leq a$ almost surely. Suppose the probability space supports a countable sequence of standard Gaussian random variables

$(h_{i,s})_{i,s\geq 1}$, independent of the U_i. Then, for all m, there exists a sequence of independent Gaussians (Z_i), each with covariance Σ, such that with probability at least $1-\eta$,

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(U_i - Z_i)\right\| \lesssim \sqrt{\log(6/\eta)}\sigma(m) + \frac{am^2\log(m^2/\eta)}{\sqrt{n}}.$$

Proof. Note that we may identify the range of Π_m, which is spanned by the top m eigenvectors of Σ, with $\mathbb{R}^m$. In particular, let A be the orthogonal projection of H onto $\mathbb{R}^m$ defined by $A : f \mapsto (\langle f, e_i\rangle)_{1\leq i\leq m}$. Then the adjoint A^* is an isometric embedding of $\mathbb{R}^m$ into the range of Π_m, and $\Pi_m = A^*A$.

Using (C.2), then, we can construct a coupling of $\frac{1}{\sqrt{n}}\sum_{i=1}^n AU_i$ and $\frac{1}{\sqrt{n}}\sum_{i=1}^n \zeta_i$, where the ζ_i are independent Gaussian vectors in $\mathbb{R}^m$, such that

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A^*AU_i - A^*\zeta_i\right\| > \delta\right) \leq Cm^2\exp\left(\frac{-c\delta\sqrt{n}}{am^2}\right).$$

Inverting the bound and noting that $A^*A = \Pi_m$, we find that with probability at least $1-\eta$ we have

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Pi_m U_i - A^*\zeta_i\right\| \leq \frac{am^2\log(Cm^2/\eta)}{c\sqrt{n}}. \tag{4}$$

Finally, we set $Z_i = A^*\zeta_i + \sum_{s=m+1}^{\infty} h_{i,s}\sqrt{\nu_s}e_s$. This ensures that the Z_i are independent and that Z_i and U_i both have the same covariance, namely Σ. To complete the argument, we apply the triangle inequality to see that

$$\begin{aligned}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(U_i - Z_i)\right\| &\leq \left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Pi_m(U_i - Z_i)\right\| + \left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\Pi_m)(U_i - Z_i)\right\| \\ &\leq \left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Pi_m(U_i - Z_i)\right\| + \left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\Pi_m)U_i\right\| + \left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\Pi_m)Z_i\right\|.\end{aligned}$$

The first term is controlled by our construction above. For the second term, by Lemma C.4, it holds with probability at least $1-\eta$ that

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\Pi_m)U_i\right\| \leq 2\sigma(m)\sqrt{\log(2/\eta)} \vee \frac{4a\log(2/\eta)}{\sqrt{n}}. \tag{5}$$

Similarly, by Lemma C.6, with probability at least $1-\eta$

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\Pi_m)Z_i\right\| \leq \left(1+\sqrt{2\log(1/\eta)}\right)\sigma(m). \tag{6}$$

Applying a union bound to the events in (4), (5) and (6) then consolidating terms, we find that with probability at least $1 - 3\eta$,

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (U_i - Z_i) \right\| \leq \left(1 + 4\sqrt{\log(2/\eta)}\right) \sigma(m) + \frac{5am^2 \log(Cm^2/\eta)}{\sqrt{n}}.$$

The result follows by replacing η with $\eta/3$ and suppressing universal constants. □

Proposition C.8 (Sub-Gaussian coupling)**.** Suppose the U_i are b sub-Gaussian. Then there exists a Gaussian random variable Z such that for any $n \geq \log(2/\eta)$, it holds with probability $1 - \eta$ that

$$\left\| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \right) - Z \right\| \lesssim b\sigma(m)\sqrt{\log(6/\eta)} + \|\Sigma\|_{op}^{\frac{1}{2}} b^3 \left(\frac{m^{\frac{3}{2}} \log(n)^2}{\sqrt{n}} \right) (3/\eta)^{1/\log(mn)}.$$

Proof. The proof is identical to the above, with the following changes. Firstly, we instead use Lemma C.3 to couple $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} AU_i$ and ζ in $\mathbb{R}^m$. This leads to the bound

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Pi_m U_i - A^* \zeta_i \right\| \lesssim \|\Sigma\|_{op}^{\frac{1}{2}} b^3 \left(\frac{m^{\frac{3}{2}} \log(n)^2}{\sqrt{n}} \right) (1/\eta)^{1/\log(mn)}. \tag{7}$$

We then set $Z = A^*\zeta + \sum_{s=m+1}^{\infty} h_{1,s} \sqrt{\nu_s}\, e_s$, and decompose

$$\left\| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \right) - Z \right\| \leq \left\| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Pi_m U_i \right) - \Pi_m Z \right\| + \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1 - \Pi_m) U_i \right\| + \left\| (1 - \Pi_m) Z \right\|.$$

The first term we have controlled above in (7). For the second term, we apply Lemma C.5 to find that

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1 - \Pi_m) U_i \right\| \lesssim b\sigma(m)\sqrt{\log(2/\eta)} + \frac{b\sigma(m) \log(2/\eta)}{\sqrt{n}}.$$

The latter is dominated by the former for n sufficiently large: $\sqrt{\log(2/\eta)}/\sqrt{n} \leq 1 \iff n \geq \log(2/\eta)$.

For the final term, we apply Lemma C.6 (with $n = 1$) to find that

$$\|(1 - \Pi_m) Z\| \leq \left(1 + \sqrt{2 \log(1/\eta)}\right) \sigma(m).$$

Putting it all together, we have

$$\left\| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \right) - Z \right\| \lesssim b\sigma(m)\sqrt{\log(2/\eta)} + \|\Sigma\|_{op}^{\frac{1}{2}} b^3 \left(\frac{m^{\frac{3}{2}} \log(n)^2}{\sqrt{n}} \right) (1/\eta)^{1/\log(mn)}.$$

The result follows by replacing η with $\eta/3$ and suppressing universal constants. □

C.6 Leading cases

In Section I as well as the corollaries below, we specialize this bound for leading cases, including Sobolev and Gaussian RKHS.

C.6.1 Sobolev-type RKHS

In a "Sobolev-type RKHS," the eigenvalues satisfy $\nu_m \leq \omega m^{-\beta}$ for some $\beta > 1$. In this case, $\sigma^2(m) \lesssim_\beta \omega m^{1-\beta}$. The optimal trade-off will depend on whether the summands are bounded or sub-Gaussian, as demonstrated by the following results.

Corollary C.9 (Sobolev-type RKHS: Bounded data)**.** Suppose that $\nu_m \leq \omega m^{-\beta}$ for some $\beta > 1$ and $\|U_i\| \leq a$. Then there exists a Gaussian random variable Z with covariance Σ such that

$$\left\| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \right) - Z \right\| \lesssim_\beta \omega^{\frac{2}{\beta+3}} \left(\frac{a}{\sqrt{n}} \right)^{\frac{\beta-1}{\beta+3}} \log \left\{ \frac{6}{\eta} (\sqrt{\omega n}/a)^{4/(\beta+3)} \right\}.$$

Proof. In this case $\sigma^2(m) \lesssim_\beta \omega m^{1-\beta}$ by Proposition I.2 in Section I. It suffices to choose m correctly in Proposition C.7. In particular, Proposition C.7 implies that for any m, we may construct a Gaussian random variable Z such that

$$\left\| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \right) - Z \right\| \lesssim_\beta \sqrt{\log(6/\eta)} \omega^{1/2} m^{(1-\beta)/2} + \frac{a m^2 \log(m^2/\eta)}{\sqrt{n}}.$$

Solving for m that balances the two dominating terms,

$$\omega^{1/2} m^{(1-\beta)/2} = \frac{a m^2}{\sqrt{n}} \iff m = (\sqrt{\omega n}/a)^{2/(\beta+3)}.$$

Then

$$\sqrt{\log(6/\eta)} \omega^{1/2} m^{(1-\beta)/2} = \sqrt{\log(6/\eta)} (\sqrt{n}/a)^{(1-\beta)/(\beta+3)} \omega^{2/(\beta+3)}$$

and

$$\begin{aligned} \frac{a m^2 \log(m^2/\eta)}{\sqrt{n}} &= \frac{a}{\sqrt{n}} (\sqrt{\omega n}/a)^{4/(\beta+3)} \log\{(\sqrt{\omega n}/a)^{4/(\beta+3)}/\eta\} \\ &= \omega^{2/(\beta+3)} (\sqrt{n}/a)^{(1-\beta)/(\beta+3)} \cdot \log\{(\sqrt{\omega n}/a)^{4/(\beta+3)}/\eta\}. \end{aligned}$$

Combining constants and logarithmic factors gives us the desired rate. □

Corollary C.10 (Sobolev-type RKHS: Sub-Gaussian data)**.** Suppose that $\nu_m \leq \omega m^{-\beta}$ for some $\beta > 1$ and U_i are b sub-Gaussian. Then there exists a Gaussian random variable Z with covariance Σ such that

$$\left\| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \right) - Z \right\| \lesssim_\beta b^{\frac{3\beta}{\beta+2}} \cdot n^{\frac{1}{2}\frac{1-\beta}{2+\beta}} \cdot \|\Sigma\|_{op}^{\frac{1}{2}\frac{\beta-1}{2+\beta}} \omega^{\frac{3}{2(\beta+2)}} \cdot \log(n^2)(3/\eta)^{\frac{1}{\log(n)}}.$$

Proof. In this case $\sigma^2(m) \lesssim_\beta \omega m^{1-\beta}$ by Proposition I.2 in Section I. It suffices to choose m correctly in Proposition C.8. In particular, Proposition C.8 implies that for any m, we may construct a Gaussian random variable Z with covariance Σ such that

$$\left\| \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \right) - Z \right\| \lesssim_\beta b\sqrt{\log(6/\eta)}\omega^{1/2} m^{(1-\beta)/2} + \|\Sigma\|_{op}^{\frac{1}{2}} b^3 \left(\frac{m^{\frac{3}{2}} \log(n)^2}{\sqrt{n}} \right) (3/\eta)^{1/\log(mn)}.$$

Solving for m that balances the main terms,

$$b\,\omega^{1/2} m^{(1-\beta)/2} = \frac{\|\Sigma\|_{op}^{\frac{1}{2}} b^3 m^{3/2}}{\sqrt{n}} \iff m = (\sqrt{\omega n} b^{-2} \|\Sigma\|_{op}^{-\frac{1}{2}})^{2/(\beta+2)}.$$

Then

$$\begin{aligned} b\omega^{1/2} m^{(1-\beta)/2} &= b\omega^{1/2} (\sqrt{\omega n} b^{-2} \|\Sigma\|_{op}^{-\frac{1}{2}})^{(1-\beta)/(\beta+2)} \\ &= b^{\frac{3\beta}{\beta+2}} \cdot n^{\frac{1}{2}\frac{1-\beta}{2+\beta}} \cdot \|\Sigma\|_{op}^{-\frac{1}{2}\frac{1-\beta}{2+\beta}} \omega^{\frac{3}{2(\beta+2)}}, \end{aligned}$$

and

$$\begin{aligned} \frac{\|\Sigma\|_{op}^{\frac{1}{2}} b^3 m^{3/2}}{\sqrt{n}} &= \frac{\|\Sigma\|_{op}^{\frac{1}{2}} b^3 (\sqrt{\omega n} b^{-2} \|\Sigma\|_{op}^{-\frac{1}{2}})^{3/(\beta+2)}}{\sqrt{n}} \\ &= b^{\frac{3\beta}{\beta+2}} \cdot n^{\frac{1}{2}\frac{1-\beta}{2+\beta}} \cdot \|\Sigma\|_{op}^{-\frac{1}{2}\frac{1-\beta}{2+\beta}} \omega^{\frac{3}{2(\beta+2)}}. \end{aligned}$$

Combining constants and logarithmic factors gives us the desired rate. In particular note that $\sqrt{\log(6/\eta)}$ and $(3/\eta)^{1/\log(mn)}$ are dominated by $(3/\eta)^{1/\log(n)}$. □

C.6.2 Gaussian-type RKHS

In a "Gaussian-type RKHS," $\sigma(m)$ is exponentially decaying. We obtain a $\sqrt{n}$ rate of Gaussian approximation up to logarithmic factors. Notably, this includes the case where $U_i = k_{X_i}$ for k a smooth, radial kernel on $\mathbb{R}^d$, see (Belkin, 2018; Wendland, 2004).

Corollary C.11 (Gaussian-type RKHS: Bounded data)**.** Suppose that $\nu_m \leq \omega \exp(-\alpha m^\gamma)$ for some $\alpha, \gamma > 0$ and $\|U_i\| \leq a$ almost surely. Then there exists a sequence of independent Gaussians (Z_i) such that

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (U_i - Z_i) \right\|_H \lesssim_{\alpha,\gamma} \frac{a}{\sqrt{n}} \left[\frac{1}{\alpha} \log\left(\frac{\omega n}{a^2} \right) \right]^{\frac{2}{\gamma}} \cdot \log\left\{ \frac{6}{\alpha} \log\left(\frac{\omega n}{a^2} \right) / \eta^{\gamma/2} \right\}.$$

Proof. In this case $\sigma^2(m) \lesssim_{\alpha,\gamma} \omega m^{1-\gamma} \exp(-\alpha m^\gamma)$ by Proposition I.2 in Section I. Proposition C.7 then implies that there exists a sequence of independent Gaussians $(Z_i)_{1 \leq i \leq m}$ such that

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (U_i - Z_i) \right\|_H \lesssim_{\gamma,\alpha} \sqrt{\log(6/\eta)} \left(\omega^{1/2} m^{(1-\gamma)/2} \exp(-\alpha m^\gamma / 2) \right) + \frac{a m^2 \log(m^2/\eta)}{\sqrt{n}}.$$

Solving for m that balances the two dominating terms,

$$\omega^{1/2} \exp(-\alpha m^\gamma / 2) = \frac{a}{\sqrt{n}} \iff m = \left[\frac{1}{\alpha} \log\left(\frac{\omega n}{a^2} \right) \right]^{\frac{1}{\gamma}}.$$

Then

$$\begin{aligned}
&\sqrt{\log(6/\eta)} \omega^{1/2} m^{(1-\gamma)/2} \exp(-\alpha m^\gamma / 2) \\
&= \sqrt{\log(6/\eta)} \omega^{1/2} \left[\frac{1}{\alpha} \log\left(\frac{\omega n}{a^2} \right) \right]^{\frac{1-\gamma}{2\gamma}} \cdot \exp\left\{ -\alpha \left[\frac{1}{\alpha} \log\left(\frac{\omega n}{a^2} \right) \right] / 2 \right\} \\
&= \sqrt{\log(6/\eta)} \omega^{1/2} \left[\frac{1}{\alpha} \log\left(\frac{\omega n}{a^2} \right) \right]^{\frac{1-\gamma}{2\gamma}} \cdot \exp\left\{ \left[\log\left(\frac{a}{\sqrt{\omega n}} \right) \right] \right\} \\
&= \sqrt{\log(6/\eta)} \omega^{1/2} \left[\frac{1}{\alpha} \log\left(\frac{\omega n}{a^2} \right) \right]^{\frac{1-\gamma}{2\gamma}} \cdot \frac{a}{\sqrt{\omega n}} \\
&= \sqrt{\log(6/\eta)} \left[\frac{1}{\alpha} \log\left(\frac{\omega n}{a^2} \right) \right]^{\frac{1-\gamma}{2\gamma}} \cdot \frac{a}{\sqrt{n}}
\end{aligned}$$

and

$$\begin{aligned}
\frac{a m^2 \log(m^2/\eta)}{\sqrt{n}} &= \frac{a}{\sqrt{n}} \left[\frac{1}{\alpha} \log\left(\frac{\omega n}{a^2} \right) \right]^{\frac{2}{\gamma}} \log\left\{ \left[\frac{1}{\alpha} \log\left(\frac{\omega n}{a^2} \right) \right]^{\frac{2}{\gamma}} / \eta \right\} \\
&= \frac{a}{\sqrt{n}} \left[\frac{1}{\alpha} \log\left(\frac{\omega n}{a^2} \right) \right]^{\frac{2}{\gamma}} \cdot \frac{2}{\gamma} \log\left\{ \frac{1}{\alpha} \log\left(\frac{\omega n}{a^2} \right) / \eta^{\gamma/2} \right\}.
\end{aligned}$$

Combining constants and logarithmic factors gives us the desired rate. In particular, note that $\gamma > 0$ implies $\frac{2}{\gamma} > \frac{1-\gamma}{2\gamma}$. □

Corollary C.12 (Gaussian-type RKHS: Sub-Gaussian data)**.** Suppose that $\nu_m \leq \omega \exp(-\alpha m^\gamma)$ for some $\alpha, \gamma > 0$ and U_i are b sub-Gaussian. Then there exists a Gaussian Z such that

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i - Z \right\| \lesssim \|\Sigma\|_{op}^{\frac{1}{2}} \frac{b^3}{\sqrt{n}} \left[\frac{1}{\alpha} \log \left(\frac{\omega n}{\|\Sigma\|_{op} b^4} \right) \right]^{\frac{3}{2\gamma}} \cdot \log(n)^2 \left(\frac{3}{\eta} \right)^{1/\log(n)}.$$

Proof. In this case $\sigma^2(m) \lesssim_{\gamma,\alpha} \omega m^{1-\gamma} \exp(-\alpha m^\gamma)$ by Proposition I.2 in Section I. Proposition C.8 then implies that there exists a sequence of independent Gaussians $(Z_i)_{1 \leq i \leq m}$ such that

$$\begin{aligned} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i - Z \right\| \lesssim_{\gamma,\alpha} & \sqrt{\log(6/\eta)} b \omega^{1/2} \left\{ m^{(1-\gamma)/2} \exp(-\alpha m^\gamma/2) \right\} \\ & + \|\Sigma\|_{op}^{\frac{1}{2}} b^3 \left(\frac{m^{\frac{3}{2}} \log(n)^2}{\sqrt{n}} \right) (3/\eta)^{1/\log(mn)}. \end{aligned}$$

Solving for m that balances the two dominating terms,

$$b\omega^{1/2} \exp(-\alpha m^\gamma/2) = \frac{\|\Sigma\|_{op}^{\frac{1}{2}} b^3}{\sqrt{n}} \iff m = \left[\frac{1}{\alpha} \log \left(\frac{\omega n}{\|\Sigma\|_{op} b^4} \right) \right]^{\frac{1}{\gamma}}.$$

Then

$$\begin{aligned} & b\omega^{1/2} m^{(1-\gamma)/2} \exp(-\alpha m^\gamma/2) \\ &= b\omega^{1/2} \left[\frac{1}{\alpha} \log \left(\frac{\omega n}{\|\Sigma\|_{op} b^4} \right) \right]^{\frac{1-\gamma}{2\gamma}} \cdot \exp \left\{ -\alpha \left[\frac{1}{\alpha} \log \left(\frac{\omega n}{\|\Sigma\|_{op} b^4} \right) \right] /2 \right\} \\ &= b\omega^{1/2} \left[\frac{1}{\alpha} \log \left(\frac{\omega n}{\|\Sigma\|_{op} b^4} \right) \right]^{\frac{1-\gamma}{2\gamma}} \cdot \exp \left\{ \left[\log \left(\frac{\|\Sigma\|_{op}^{\frac{1}{2}} b^2}{\sqrt{\omega n}} \right) \right] \right\} \\ &= \left[\frac{1}{\alpha} \log \left(\frac{\omega n}{\|\Sigma\|_{op} b^4} \right) \right]^{\frac{1-\gamma}{2\gamma}} \cdot \frac{\|\Sigma\|_{op}^{1/2} b^3}{\sqrt{n}} \end{aligned}$$

and

$$\|\Sigma\|_{op}^{1/2} b^3 \left(\frac{m^{\frac{3}{2}} \log(n)^2}{\sqrt{n}} \right) = \|\Sigma\|_{op}^{1/2} \frac{b^3}{\sqrt{n}} \left[\frac{1}{\alpha} \log \left(\frac{\omega n}{\|\Sigma\|_{op} b^4} \right) \right]^{\frac{3}{2\gamma}} \log(n)^2.$$

Combining constants and logarithmic factors gives us the desired rate. In particular, note that $\gamma > 0$ implies $\frac{3}{2\gamma} > \frac{1-\gamma}{2\gamma}$. The final comparison is between $\sqrt{\log(6/\eta)}$ and $(3/\eta)^{1/\log(mn)}$, both of which are dominated by $(3/\eta)^{1/\log(n)}$. □

C.7 Supremum norm

Theorem C.13 (Sup-norm coupling)**.** With probability $1-\eta$, there exists a Gaussian process Z such that either:

1. If $\|U_i\| \leq a$, then
$$\sup_{x\in S}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i(x) - Z(x)\right| \lesssim \kappa\left(\sqrt{\log(6/\eta)}\sigma(m) + \frac{am^2\log(m^2/\eta)}{\sqrt{n}}\right);$$
2. If instead U_i are b sub-Gaussian and $n \geq \log(2/\eta)$, then
$$\sup_{x\in S}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i(x) - Z(x)\right| \lesssim \kappa\Bigg\{b\sigma(m)\sqrt{\log(6/\eta)} + \|\Sigma\|_{op}^{\frac{1}{2}} b^3\left(\frac{m^{\frac{3}{2}}\log(n)^2}{\sqrt{n}}\right)(3/\eta)^{1/\log(mn)}\Bigg\}.$$
These bounds can be further specialized for Sobolev and Gaussian RKHSs using Corollaries C.9, C.10, C.11, and C.12.

Proof. The proof is immediate from Lemma C.1 and Propositions C.7 and C.8. □

D Bootstrap couplings

According to the sketch in Section 6, in this section we consider $Z_{\mathfrak{B}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}\frac{V_i - V_j}{\sqrt{2}}$ and show that its conditional distribution given the data U is approximately Gaussian with covariance $\Sigma = \mathbb{E}(U_i \otimes U_i^*)$. This is done by the following sequence of steps:

Re-writing the symmetrized bootstrap process We show that the conditional distribution of $Z_{\mathfrak{B}}$ is Gaussian with covariance $\hat{\Sigma}$, where $\hat{\Sigma} = \mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}_n[U_i] \otimes (\mathbb{E}_n[U_i])^*$. Note that we must use the empirically centered operator $\hat{\Sigma}$ because of the symmetrized bootstrap, as detailed below.

Covariance estimation Using concentration and approximation arguments, we prove a bound on $\|\hat{\Sigma}^{\frac{1}{2}} - \Sigma^{\frac{1}{2}}\|$ and show that this implies a bound on the distance between Gaussian distributions.

Main result We use an extension of the Dudley-Strassen theorem due to Monrad and Philipp (1991) (our Lemma D.8) and Gaussian concentration to construct Z' that is conditionally Gaussian with covariance Σ, such that $\|Z' - Z_{\mathfrak{B}}\|$ is small w.h.p. conditional upon U.

D.1 Notation

As before, let $(U_1, U_2, \ldots)$ denote an i.i.d. sequence of centered random elements in H, and let $\Sigma = \mathbb{E}[U_i \otimes U_i^*]$. Let $V_i = U_i + \mu$ for some arbitrary (deterministic) $\mu \in H$.

Given an independent sequence of standard normal variables $(g_1, g_2, \ldots)$ we may then consider the random series $g = \sum_{i=1}^{\infty} g_i e_i$ and $\Sigma^{\frac{1}{2}} g = \sum_{i=1}^{\infty} \sqrt{\nu_i} g_i e_i$, which belong almost surely to $L^2(U_i)$ and H, respectively.

Finally, we will use the notation $X \overset{U}{\sim} Y$ to denote that the $\sigma(U)$-conditional distributions of the random variables X and Y are equal, i.e. for any Borel set A, $\mathbb{P}(X \in A|U) = \mathbb{P}(Y \in A|U)$ holds U-almost surely.

D.2 Rewriting the symmetrized bootstrap process

To begin, we argue $\hat{\Sigma}^{1/2} g \overset{U}{\sim} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n h_{ij} \frac{V_i - V_j}{\sqrt{2}}$.

Lemma D.1 (Covariance of symmetrized bootstrap process)**.** We have that

$$\hat{\Sigma}^{1/2} g \overset{U}{\sim} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n h_{ij} \left(\frac{V_i - V_j}{\sqrt{2}} \right).$$

Proof. Since $\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n h_{ij} \left(\frac{V_i - V_j}{\sqrt{2}} \right)$ is jointly Gaussian conditional upon $V = (V_1, V_2, \ldots, V_n)$,

it suffices to compute

$$\begin{aligned}
&\mathbb{E}_h\left[\left\{\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n h_{ij}\left(\frac{V_i-V_j}{\sqrt{2}}\right)\right\}\otimes\left\{\frac{1}{n}\sum_{k=1}^n\sum_{\ell=1}^n h_{k\ell}\left(\frac{V_k-V_\ell}{\sqrt{2}}\right)\right\}^*\right]\\
&=\mathbb{E}_h\left[\left\{\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n h_{ij}\left(\frac{U_i-U_j}{\sqrt{2}}\right)\right\}\otimes\left\{\frac{1}{n}\sum_{k=1}^n\sum_{\ell=1}^n h_{k\ell}\left(\frac{U_k-U_\ell}{\sqrt{2}}\right)\right\}^*\right]\\
&=\frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n\{(U_i-U_j)\otimes(U_i-U_j)^*\}\\
&=\frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n\{(U_i\otimes U_i^*)-(U_i\otimes U_j^*)+(U_j\otimes U_j^*)-(U_j\otimes U_i^*)\}.
\end{aligned}$$

Here we have used the fact that each summand (i,j) is associated with an independent standard Gaussian multiplier h_{ij}. By symmetry under transposition of i and j, this reduces to

$$\begin{aligned}
&=\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\{(U_i\otimes U_i^*)-(U_i\otimes U_j^*)\}\\
&=\frac{1}{n}\sum_{i=1}^n\left\{(U_i\otimes U_i^*)-\left(U_i\otimes\frac{1}{n}\sum_{j=1}^n U_j^*\right)\right\}\\
&=\frac{1}{n}\sum_{i=1}^n(U_i\otimes U_i^*)-\left(\frac{1}{n}\sum_{i=1}^n U_i\otimes\frac{1}{n}\sum_{j=1}^n U_j^*\right)\\
&=\hat{\Sigma}.
\end{aligned}$$

Since the conditional distribution of $\hat{\Sigma}^{\frac{1}{2}}g$ is also jointly Gaussian with the same covariance, the two are equal in conditional distribution. □

An alternative expression will be helpful later on.

Lemma D.2 (Single sum symmetrized bootstrap process)**.**

$$\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n h_{ij}\left(\frac{V_i-V_j}{\sqrt{2}}\right)=\sqrt{n}\mathbb{E}_n[q_iU_i],\quad q=\frac{1}{\sqrt{2n}}(h-h^\top)\mathbf{1},\quad var(q_i)<1.$$

Here, (q_i) are Gaussians whose definitions are expressed in terms of (h_{ij}). Note that (q_i) are not independent while (h_{ij}) are.

Proof. Write

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}\left(\frac{V_i - V_j}{\sqrt{2}}\right) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}\left(\frac{U_i - U_j}{\sqrt{2}}\right) \\
&= \frac{1}{\sqrt{2}n}\sum_{i=1}^{n}\sum_{j=1}^{n}(U_i h_{ij} - U_j h_{ij}) \\
&= \frac{1}{\sqrt{2}n}\sum_{i=1}^{n} U_i \sum_{j=1}^{n} h_{ij} - \frac{1}{\sqrt{2}n}\sum_{j=1}^{n} U_j \sum_{i=1}^{n} h_{ij} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i \left(\frac{1}{\sqrt{2n}}\sum_{j=1}^{n} h_{ij}\right) - \frac{1}{\sqrt{n}}\sum_{j=1}^{n} U_j \left(\frac{1}{\sqrt{2n}}\sum_{i=1}^{n} h_{ij}\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i q_i \\
&= \sqrt{n}\mathbb{E}_n[U_i q_i]
\end{aligned}$$

where

$$q_i = \frac{1}{\sqrt{2n}}\sum_{j=1}^{n}\left(h_{ij} - h_{ji}\right), \quad q = \frac{1}{\sqrt{2n}}(h - h^\top)\mathbf{1}, \quad var(q_i) < 1.$$

Note that the variance of q_i is strictly less than one since $h_{ij} - h_{ji} = 0$ when $i = j$. □

Corollary D.3 (Combining results)**.** In summary,

$$\hat{\Sigma}^{1/2} g \overset{U}{\sim} \sqrt{n}\mathbb{E}_n[q_i U_i], \quad q = \frac{1}{\sqrt{2n}}(h - h^\top)\mathbf{1}, \quad var(q_i) < 1.$$

Proof. The result is immediate from Lemmas D.1 and D.2. □

D.3 Covariance estimation

Next, we argue $\Sigma^{1/2} g \approx \hat{\Sigma}^{1/2} g$. As before, we will make our argument by way of finite dimensional approximation. Our approximation uses the top m eigenvectors, thus avoiding calculations based upon the metric entropy.

We write this subsection with some additional generality to accommodate alternative bootstraps. Throughout this subsection, $\hat{\Sigma}$ denotes some feasible covariance estimator, with the property that $\hat{\Sigma}^{1/2} g \overset{U}{\sim} \sqrt{n}\mathbb{E}_n[q_i U_i]$ for some jointly Gaussian random variables q_i that have variance at most one and that may be correlated.

The structure of the argument is as follows: (i) technical lemmas, (ii) abstract bound (agnostic to covariance operator), and (iii) bounding key terms (for $\hat{\Sigma}$).

D.3.1 Technical lemmas

We quote the following inequality due to Borell and independently to Sudakov, Ibragamov and Tsirelson.

Lemma D.4 (Giné and Nickl 2021, Theorem 2.5.8)**.** Let G_t be a centered Gaussian process, a.s. bounded on T. Then for $u > 0$,

$$\mathbb{P}\left(\sup_{t\in T} G_t - \mathbb{E}\sup_{t\in T} G_t > u\right) \vee \mathbb{P}\left(\sup_{t\in T} G_t - \mathbb{E}\sup_{t\in T} G_t < -u\right) \leq \exp\left(\frac{-u^2}{2\sigma_T^2}\right)$$

where $\sigma_T^2 := \sup_{t\in T} \mathbb{E}G_t^2$.

The following lemma follows from Gaussian concentration.

Lemma D.5. Let Z be a Gaussian random element in a Hilbert space H such that $\mathbb{E}\left\|Z\right\|^2 < \infty$. Then, with probability at least $1-\eta$,

$$\|Z\| \leq \left(1 + \sqrt{2\log(1/\eta)}\right)\sqrt{\mathbb{E}\left\|Z\right\|^2}.$$

In particular, if $A : H \to H$ is a trace-class operator, then with probability at least $1-\eta$ w.r.t. g,

$$\|Ag\| \leq \left(1 + \sqrt{2\log(1/\eta)}\right)\|A\|_{\mathrm{HS}}.$$

Proof. We proceed in steps.

1. For the first claim, we express $\|Z\|$ as the supremum of a separable Gaussian process, in particular $\|Z\| = \sup_{t\in B_H}\langle t,\, Z\rangle = \sup_{t\in T} G_t$. The result follows from Gaussian concentration (Lemma D.4) provided we can estimate the quantities

$$\mathbb{E}\sup_{t\in T} G_t = \mathbb{E}\sup_{t\in B_H}\langle t,\, Z\rangle = \mathbb{E}\|Z\|, \quad \sigma_T^2 = \sup_{t\in B_H}\mathbb{E}\left\langle t,\, Z\right\rangle^2,$$

   and show that the process is a.s. bounded.

By Jensen's inequality we have $(\mathbb{E}\|Z\|)^2 \leq \mathbb{E}\|Z\|^2 < \infty$, so we may deduce from Markov's inequality that $\langle t,\, Z\rangle$ is a.s. bounded. Also by Jensen's inequality

$$\sup_{t\in B_H} \mathbb{E}\,\langle t,\, Z\rangle^2 \leq \mathbb{E}\sup_{t\in B_H}\langle t,\, Z\rangle^2 = \mathbb{E}\,\|Z\|^2\,.$$

Plugging these estimates into Borell's inequality then gives

$$\mathbb{P}\left(\|Z\| \geq \sqrt{\mathbb{E}\,\|Z\|^2} + u\right) \leq \exp\left(\frac{-u^2}{2\mathbb{E}\,\|Z\|^2}\right).$$

Choosing $u = \sqrt{2\log(1/\eta)\mathbb{E}\,\|Z\|^2}$ and squaring gives the desired result.

2. For the second claim, take $Z = Ag$. We need to check that $\mathbb{E}\,\|Ag\|^2 = \|A\|_{\mathrm{HS}}^2$. Indeed, since $g = \sum_s g_s e_s$, we have that $Ag = \sum_s g_s Ae_s$ and

$$\|Ag\|^2 = \langle \sum_s g_s Ae_s, \sum_t g_t Ae_t\rangle = \sum_{s,t} g_s g_t \langle Ae_s, Ae_t\rangle.$$

Taking the expectation and denoting by ι_s the s-th eigenvalue of A

$$\mathbb{E}\|Ag\|^2 = \sum_s \langle Ae_s, Ae_s\rangle = \sum_s \langle \iota_s e_s, \iota_s e_s\rangle = \sum_s \iota_s^2 = \|A\|_{\mathrm{HS}}^2.$$

□

Thus, the bootstrap provides a good approximation to the distribution of $\Sigma^{\frac{1}{2}} g$ whenever we can control the quantity $\|\hat{\Sigma}^{\frac{1}{2}} - \Sigma^{\frac{1}{2}}\|_{\mathrm{HS}}$. To accomplish this, we make use of the following finite-dimensional result.

Lemma D.6 (cf. Wihler (2009, Theorem 1.1))**.** Let A and B be $m \times m$ real matrices. Then $\|A^{\frac{1}{2}} - B^{\frac{1}{2}}\|_F \leq m^{\frac{1}{4}}\|A - B\|_F^{\frac{1}{2}}$.

Proof. According to Wihler (2009, Theorem 1.1), we have $\|f(A) - f(B)\|_F \leq [f]_{0,\frac{1}{2}} m^{\frac{1}{4}} \|A - B\|_F^{\frac{1}{2}}$ where $[f]_{0,\frac{1}{2}}$ is the Hölder constant $[f]_{0,\frac{1}{2}} = \sup_{x\neq y} \frac{|f(x)-f(y)|}{|x-y|^{\frac{1}{2}}}$. To show that for $f(x) = \sqrt{x}$ this constant is at most 1, assume w.l.o.g. that $x > y$. In this case we have

$$(\sqrt{x} - \sqrt{y})^2 = x(1 - \sqrt{y/x})^2 \leq x(1 - y/x) = x - y,$$

since $|1 - a| \leq \sqrt{|1 - a^2|}$ for $a = \sqrt{y/x}$, as the graph of the semicircle is concave over $[-1, 1]$. The proof for the case $y > x$ is symmetric. □

To prove the abstract bound, we prove some additional helpful lemmas.

Lemma D.7. We have $\Sigma^{\frac{1}{2}} = (\Pi_m \Sigma \Pi_m)^{\frac{1}{2}} + (\Pi_m^{\perp} \Sigma \Pi_m^{\perp})^{\frac{1}{2}}$.

Proof. Since

$$(e_i \otimes e_i^*)(e_j \otimes e_j^*) = \langle e_i,\, e_j \rangle \, (e_i \otimes e_j^*) = \delta_{ij}(e_i \otimes e_j^*),$$

it follows by definition of $\Sigma^{1/2}$ and Π_m that

$$\Sigma^{\frac{1}{2}} \Pi_m = \left(\sum_{i=1}^{\infty} \sqrt{\nu_i} e_i \otimes e_i^* \right) \left(\sum_{j=1}^{m} e_j \otimes e_j^* \right) = \sum_{j=1}^{m} \sqrt{\nu_j} (e_j \otimes e_j^*)$$

which is self-adjoint, so

$$\Sigma^{\frac{1}{2}} \Pi_m = (\Sigma^{\frac{1}{2}} \Pi_m)^* = \Pi_m^* (\Sigma^{\frac{1}{2}})^* = \Pi_m \Sigma^{\frac{1}{2}}$$

and $\Pi_m C^{\frac{1}{2}} \Pi_m = \Pi_m^2 C^{\frac{1}{2}} = \Pi_m \Sigma^{\frac{1}{2}}$. Therefore

$$\Sigma^{\frac{1}{2}} \Pi_m = \Pi_m C^{\frac{1}{2}} = \Pi_m \Sigma^{\frac{1}{2}} \Pi_m = (\Pi_m \Sigma \Pi_m)^{\frac{1}{2}},$$

and likewise for $\Pi_m^{\perp}$, replacing the indexing from $j \in [m]$ to $j > m$. The result follows after noting that $\Sigma^{\frac{1}{2}} = \Sigma^{\frac{1}{2}}(\Pi_m + \Pi_m^{\perp})$. □

Lemma D.8 (Conditional Strassen's Lemma; Monrad and Philipp 1991, Theorem 4)**.** Let X be a random variable on a probability space $(\Omega, \mathcal{S}, \mathbb{P})$, and suppose that X takes values in a complete metric space (S, d). Let $\mathcal{F} \subset \mathcal{S}$ be countably generated as a σ-algebra, and assume that there exists a random variable R on $(\Omega, \mathcal{S}, \mathbb{P})$ that is independent of $\mathcal{F} \vee \sigma(X)$. Let $G(-|\mathcal{F})$ be a regular conditional distribution on the Borel sets of (S, d) and suppose that for some non-negative numbers α and β

$$\mathbb{E} \sup_{A \in \mathcal{S}} \Big(\mathbb{P}(X \in A | \mathcal{F}) - G(\mathrm{cl}(A^{\alpha}) | \mathcal{F}) \Big) \leq \beta$$

where A^{α} is the α-extension of A and the randomness in the expectation is over $\mathcal{F}$.

Then there exists a random variable Y with values in S, defined on $(\Omega, \mathcal{S}, \mathbb{P})$ with conditional distribution G satisfying $\mathbb{P}\left(d(X, Y) > \alpha\right) \leq \beta$.

Corollary D.9. Under the same conditions as Lemma D.8, suppose there exist random variables X' and Y' such that (i) X and X' have the same distribution conditional upon $\mathcal{F}$, and (ii) $\mathbb{P}(d(X', Y') > \alpha) \leq \beta$. Then there exists some Y with the same conditional distribution as Y' such that $\mathbb{P}(d(X, Y) > \alpha) \leq \beta$.

Proof. According to Lemma D.8, it suffices to bound

$$\mathbb{E} \sup_{A \in \mathcal{S}} \left(\mathbb{P}(X \in A | \mathcal{F}) - \mathbb{P}(Y' \in \mathrm{cl}(A^\alpha) | \mathcal{F}) \right) = \mathbb{E} \sup_{A \in \mathcal{S}} \left(\mathbb{P}(X' \in A | \mathcal{F}) - \mathbb{P}(Y' \in \mathrm{cl}(A^\alpha) | \mathcal{F}) \right),$$

where we have first used the fact that X' and X are equal in conditional distribution.

Now, consider the event $E = \{d(X', Y') \leq \alpha\}$. By our assumption $\mathbb{P}(E) \geq 1 - \beta$. Also, by construction, for any Borel set A, we have $\{X' \in A\} \cap E \subseteq \{Y' \in \mathrm{cl}(A^\alpha)\}$. It follows that on the event E, $\sup_{A \in \mathcal{S}} \mathbb{1}\{X' \in A\} - \mathbb{1}\{Y' \in \mathrm{cl}(A^\alpha)\} \leq 0$. Moreover, since the expression inside the supremum is a difference of two probabilities, it is at most 1 everywhere. In particular, this crude bound holds on the complement of E. Thus, by the conditional version of Jensen's inequality,

$$\begin{aligned}
\mathbb{E} \sup_{A \in \mathcal{S}} \left(\mathbb{P}(X' \in A | \mathcal{F}) - \mathbb{P}(Y' \in \mathrm{cl}(A^\alpha) | \mathcal{F}) \right) &= \mathbb{E} \left[\sup_{A \in \mathcal{S}} \mathbb{E} \left[\mathbb{1}\left\{ X' \in A \right\} - \mathbb{1}\left\{ Y' \in \mathrm{cl}(A^\alpha) \right\} \middle| \mathcal{F} \right] \right] \\
&\leq \mathbb{E} \left[\sup_{A \in \mathcal{S}} \mathbb{1}\left\{ X' \in A \right\} - \mathbb{1}\left\{ Y' \in \mathrm{cl}(A^\alpha) \right\} \right] \\
&\leq 0 \cdot \mathbb{P}(E) + 1 \cdot \{1 - \mathbb{P}(E)\} \leq \beta.
\end{aligned}$$

Thus, we have verified that

$$\mathbb{E} \sup_{A \in \mathcal{S}} \left(\mathbb{P}(X \in A | \mathcal{F}) - \mathbb{P}(Y' \in \mathrm{cl}(A^\alpha) | \mathcal{F}) \right) = \mathbb{E} \sup_{A \in \mathcal{S}} \left(\mathbb{P}(X' \in A | \mathcal{F}) - \mathbb{P}(Y' \in \mathrm{cl}(A^\alpha) | \mathcal{F}) \right) \leq \beta,$$

and may conclude by applying Lemma D.8. □

Lemma D.10 (Applying operators)**.** Let (q_i) be a sequence of jointly Gaussian random variables and $A : H \to H$ be a self-adjoint operator. If $\frac{1}{\sqrt{n}} \sum_{i=1}^n q_i U_i \overset{U}{\sim} \hat{\Sigma}^{\frac{1}{2}} g$ then $\frac{1}{\sqrt{n}} \sum_{i=1}^n q_i A U_i \overset{U}{\sim} (A \hat{\Sigma} A)^{\frac{1}{2}} g$.

Proof. Since $\frac{1}{\sqrt{n}} \sum_{i=1}^n q_i U_i \overset{U}{\sim} \hat{\Sigma}^{\frac{1}{2}} g$, both vectors must have covariance operator $\hat{\Sigma}$. Now, conditional upon U, both random vectors in the conclusion of the Lemma are jointly

Gaussian in $\mathrm{span}(U_1, U_2, \ldots, U_n) \subset H$. Therefore it suffices to compute the covariance

$$\mathbb{E}\left[\left\langle \frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i A U_i,\, u\right\rangle \left\langle \frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i A U_i,\, v\right\rangle\right] = \left\langle u,\, A\hat{\Sigma}Av\right\rangle$$

which follows by repeatedly using self-adjointness of A and the definition of the covariance. □

Lemma D.11. Suppose $\mathbb{P}\{\|(\Pi_m^{\perp}\hat{\Sigma}\Pi_m^{\perp})^{\frac{1}{2}}g\| > \alpha\} \leq \beta$. Then, there exists a random variable $G \overset{U}{\sim} \hat{\Sigma}^{\frac{1}{2}}g$ such that with probability at least $1-\beta$ $\left\|(\Pi_m\hat{\Sigma}\Pi_m)^{\frac{1}{2}}g - G\right\| \leq \alpha$.

Proof. We verify by computing the covariance operator, which determines the law of jointly Gaussian variables, that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i U_i \overset{U}{\sim} \hat{\Sigma}^{\frac{1}{2}}g, \quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i \Pi_m U_i \overset{U}{\sim} (\Pi_m\hat{\Sigma}\Pi_m)^{\frac{1}{2}}g, \quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i \Pi_m^{\perp} U_i \overset{U}{\sim} (\Pi_m^{\perp}\hat{\Sigma}\Pi_m^{\perp})^{\frac{1}{2}}g. \tag{8}$$

The full computation is carried out in Lemma D.10. Moreover,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i U_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i \Pi_m U_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i \Pi_m^{\perp} U_i. \tag{9}$$

Thus we may apply Corollary D.9 with $\mathcal{F} = \sigma(U)$, choosing the random variables to be

$$X = (\Pi_m\hat{\Sigma}\Pi_m)^{\frac{1}{2}}g, \quad X' = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i \Pi_m U_i, \quad Y' = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i U_i.$$

In particular, note that

$$\mathbb{P}\left(\|X' - Y'\| > \alpha\right) = \mathbb{P}\left(\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_i \Pi_m^{\perp} U_i\right\| > \alpha\right) = \mathbb{P}\left(\|(\Pi_m^{\perp}\hat{\Sigma}\Pi_m^{\perp})^{\frac{1}{2}}g\| > \alpha\right) \leq \beta,$$

where the first equality is by equality of the random variables in (9), and the second is by equality in distribution (from equality in conditional distribution in (8)).

Thus Corollary D.9 guarantees the existence of some random variable Y (G in the statement of this Lemma) such that Y has the same conditional distribution as Y', which is the same as that of $\hat{\Sigma}^{\frac{1}{2}}g$, and $\mathbb{P}(\|X - Y\| > \alpha) \leq \beta$, which is precisely what was claimed. □

D.3.2 Abstract bound

In the abstract bound, we wish to argue that $\|\Sigma^{1/2}g \approx \hat{\Sigma}^{1/2}g\|$ is small w.h.p., where the gap depends on two key quantities:

$$\Delta_1 := \|\hat{\Sigma} - \Sigma\|_{\mathrm{HS}}, \quad \Delta_2 := \operatorname{tr} \Pi_m^{\perp}(\hat{\Sigma} - \Sigma)\Pi_m^{\perp}.$$

The bounds on these key quantities will depend on further assumptions, which we reserve for later.

Proposition D.12 (Abstract bound)**.** There exists a random variable $G \overset{U}{\sim} \hat{\Sigma}^{\frac{1}{2}}g$ such that with probability at least $1 - 3\eta$, it holds that

$$\left\|\Sigma^{\frac{1}{2}}g - G\right\| \leq \left\{1 + \sqrt{2\log(1/\eta)}\right\}\left\{m^{\frac{1}{4}}\Delta_1^{1/2} + \Delta_2^{1/2} + 2\sigma(m)\right\}.$$

Proof. We proceed in steps.

1. First we show that if $\mathbb{P}\{\|(\Pi_m^{\perp}\hat{\Sigma}\Pi_m^{\perp})^{\frac{1}{2}}g\| > \alpha\} \leq \beta$, then with probability at least $1 - \beta$ it holds that

$$\left\|\Sigma^{\frac{1}{2}}g - G\right\| \leq \left\|(\Pi_m C \Pi_m)^{\frac{1}{2}}g - (\Pi_m\hat{\Sigma}\Pi_m)^{\frac{1}{2}}g\right\| + \left\|(\Pi_m^{\perp}\Sigma\Pi_m^{\perp})^{\frac{1}{2}}g\right\| + \alpha. \tag{10}$$

   Let G be constructed as in Lemma D.11. By Lemma D.7 we have

$$\Sigma^{\frac{1}{2}}g = (\Pi_m\Sigma\Pi_m)^{\frac{1}{2}}g + (\Pi_m^{\perp}\Sigma\Pi_m^{\perp})^{\frac{1}{2}}g.$$

   Thus, by adding and subtracting, we have

$$\begin{aligned}\Sigma^{\frac{1}{2}}g - G = (\Pi_m\Sigma\Pi_m)^{\frac{1}{2}}g - (\Pi_m\hat{\Sigma}\Pi_m)^{\frac{1}{2}}g \\ + (\Pi_m^{\perp}\Sigma\Pi_m^{\perp})^{\frac{1}{2}}g \\ + (\Pi_m\hat{\Sigma}\Pi_m)^{\frac{1}{2}}g - G.\end{aligned}$$

   The equation (10) then follows by applying the triangle inequality and noting that, by Lemma D.11, the final term has norm at most α with probability at least $1 - \beta$.

2. Second, we show that if $\mathbb{P}\{\|(\Pi_m^\perp \hat{\Sigma} \Pi_m^\perp)^{\frac{1}{2}} g\| > \alpha\} \leq \beta$, then with probability at least $1 - \beta - 2\eta$ it holds that

$$\left\| \Sigma^{\frac{1}{2}} g - G \right\| \leq \left\{ m^{\frac{1}{4}} \|\Sigma - \hat{\Sigma}\|_{\mathrm{HS}}^{\frac{1}{2}} + \sigma(m) \right\} \left\{ 1 + \sqrt{2 \log(1/\eta)} \right\} + \alpha. \tag{11}$$

To deduce (11) from (10), apply Lemma D.5 to each term, conditional upon the data. For the first term, we have by Lemma D.5 that with conditional probability $1 - \eta$,

$$\|(\Pi_m \Sigma \Pi_m)^{\frac{1}{2}} g - (\Pi_m \hat{\Sigma} \Pi_m)^{\frac{1}{2}} g\| \leq \|(\Pi_m \Sigma \Pi_m)^{\frac{1}{2}} - (\Pi_m \hat{\Sigma} \Pi_m)^{\frac{1}{2}}\|_{\mathrm{HS}} \left(1 + \sqrt{2 \log(1/\eta)} \right).$$

By Lemma D.6, this is

$$\begin{aligned}
&\leq m^{\frac{1}{4}} \|(\Pi_m \Sigma \Pi_m) - (\Pi_m \hat{\Sigma} \Pi_m)\|_{\mathrm{HS}}^{\frac{1}{2}} \left(1 + \sqrt{2 \log(1/\eta)} \right) \\
&\leq m^{\frac{1}{4}} \|\Pi_m\|_{op}^{\frac{1}{2}} \|\Sigma - \hat{\Sigma}\|_{\mathrm{HS}}^{\frac{1}{2}} \|\Pi_m\|_{op}^{\frac{1}{2}} \left(1 + \sqrt{2 \log(1/\eta)} \right) \\
&\leq m^{\frac{1}{4}} \|\Sigma - \hat{\Sigma}\|_{\mathrm{HS}}^{\frac{1}{2}} \left(1 + \sqrt{2 \log(1/\eta)} \right)
\end{aligned}$$

since $\|\Pi_m\|_{op} = 1$. For the other term we apply Lemma D.5 similarly, taking note that since $\Pi_m^\perp \Sigma^{\frac{1}{2}} \Pi_m^\perp$ is self-adjoint we have

$$\|(\Pi_m^\perp \Sigma \Pi_m^\perp)^{\frac{1}{2}}\|_{\mathrm{HS}} = \sqrt{\operatorname{tr} \Pi_m^\perp \Sigma \Pi_m^\perp} = \sqrt{\sigma^2(m)}.$$

After a union bound over events w.p. η, η, and β, the proof is complete.

3. Finally, we show the proposition statement. By Lemma D.5, w.p. $1 - \eta$

$$\|(\Pi_m^\perp \hat{\Sigma} \Pi_m^\perp)^{\frac{1}{2}} g\| \leq \|(\Pi_m^\perp \hat{\Sigma} \Pi_m^\perp)^{\frac{1}{2}}\|_{\mathrm{HS}} \left(1 + \sqrt{2 \log(1/\eta)} \right).$$

Moreover

$$\begin{aligned}
\|(\Pi_m^\perp \hat{\Sigma} \Pi_m^\perp)^{\frac{1}{2}}\|_{\mathrm{HS}} &= \{\operatorname{tr}(\Pi_m^\perp \hat{\Sigma} \Pi_m^\perp)\}^{1/2} \\
&= [\operatorname{tr}\{\Pi_m^\perp (\hat{\Sigma} - \Sigma) \Pi_m^\perp\} + \operatorname{tr}(\Pi_m^\perp \Sigma \Pi_m^\perp)]^{1/2} \\
&\leq [\operatorname{tr}\{\Pi_m^\perp (\hat{\Sigma} - \Sigma) \Pi_m^\perp\}]^{1/2} + \{\operatorname{tr}(\Pi_m^\perp \Sigma \Pi_m^\perp)\}^{1/2} \\
&= [\operatorname{tr}\{\Pi_m^\perp (\hat{\Sigma} - \Sigma) \Pi_m^\perp\}]^{1/2} + \sigma(m).
\end{aligned}$$

In summary, w.p. $1-\eta$

$$\|(\Pi_m^\perp \hat{\Sigma} \Pi_m^\perp)^{\frac{1}{2}} g\| \leq \left\{1+\sqrt{2\log(1/\eta)}\right\}\left([\mathrm{tr}\{\Pi_m^\perp(\hat{\Sigma}-\Sigma)\Pi_m^\perp\}]^{1/2}+\sigma(m)\right) = \alpha.$$

Therefore by (11), w.p. $1-3\eta$

$$\begin{aligned}\left\|\Sigma^{\frac{1}{2}} g - G\right\| &\leq \left\{1+\sqrt{2\log(1/\eta)}\right\}\left\{m^{\frac{1}{4}}\|\Sigma-\hat{\Sigma}\|_{\mathrm{HS}}^{\frac{1}{2}}+\sigma(m)\right\}+\alpha \\ &= \left\{1+\sqrt{2\log(1/\eta)}\right\}\left\{m^{\frac{1}{4}}\|\Sigma-\hat{\Sigma}\|_{\mathrm{HS}}^{\frac{1}{2}}+[\mathrm{tr}\{\Pi_m^\perp(\hat{\Sigma}-\Sigma)\Pi_m^\perp\}]^{1/2}+2\sigma(m)\right\}.\end{aligned}$$

□

D.3.3 Bounding key terms

The abstract bound shows with high probability, $\Sigma^{1/2} g \approx \hat{\Sigma}^{1/2} g$ where the gap depends on $\Delta_1 := \|\hat{\Sigma}-\Sigma\|_{\mathrm{HS}}$ and $\Delta_2 := \mathrm{tr}\,\Pi_m^\perp(\hat{\Sigma}-\Sigma)\Pi_m^\perp$. We bound these key quantities under alternative assumptions. The bounds are with high probability, with respect to the randomness in U. Since these arguments are technical and self-contained, we present them in Appendix J. We summarize them in the following lemma.

Lemma D.13 (Bounding key terms for covariance estimation)**.** Under a-boundedness, w.p. $1-3\eta$, both of the following bounds hold:

$$\Delta_1 \leq 2\log(2/\eta)^2\left\{\sqrt{\frac{a^2\sigma^2(0)}{n}} \vee \frac{4a^2}{n} \vee \frac{8a^2}{n^2}\right\}, \quad \Delta_2 \leq 2\log(2/\eta)\left(\sqrt{\frac{a^2\sigma^2(m)}{n}} \vee \frac{2a^2}{n}\right).$$

Under b-sub-Gaussianity, w.p. $1-3\eta$, both of the following bounds hold:

$$\Delta_1 \leq C\log(2/\eta)^2\frac{b^2\sigma^2(0)}{\sqrt{n}}, \quad \Delta_2 \leq C\log(2/\eta)\frac{b^2\sigma^2(m)}{\sqrt{n}}$$

where C is a universal constant.

Proof. See Appendix J. □

D.4 Main result

Theorem D.14. Suppose a-boundedness holds and $n \geq 2$. Then there exists a random element G with the same conditional distribution as the multiplier bootstrap process $\hat{\Sigma}^{\frac{1}{2}} g$ such that with total probability at least $1 - \eta$

$$\left\| \Sigma^{\frac{1}{2}} g - G \right\| \leq C \log(2/\eta)^{3/2} \inf_{m \geq 1} \left\{ m^{\frac{1}{4}} \left(\frac{a^2 \sigma^2(0)}{n} + \frac{a^4}{n^2} \right)^{\frac{1}{4}} + \sigma(m) \right\}. \tag{12}$$

Suppose b-sub-Gaussianity holds. Then the identical statement holds with

$$\left\| \Sigma^{\frac{1}{2}} g - G \right\| \leq C \log(2/\eta)^{3/2} \inf_{m \geq 1} \left[m^{\frac{1}{4}} \left\{ \frac{b^4 \sigma^4(0)}{n} \right\}^{\frac{1}{4}} + \sigma(m) \right]. \tag{13}$$

Proof. We proceed in steps, appealing to Proposition D.12: w.p. $1 - 3\eta$,

$$\left\| \Sigma^{\frac{1}{2}} g - G \right\| \leq \left\{ 1 + \sqrt{2 \log(1/\eta)} \right\} \left\{ m^{\frac{1}{4}} \Delta_1^{1/2} + \Delta_2^{1/2} + 2\sigma(m) \right\}.$$

In particular, we substitute in different bounds from Lemma D.13.

1. Proving (12). Since Δ_1 dominates Δ_2, w.p. $1 - 6\eta$,

   $$\left\| \Sigma^{\frac{1}{2}} g - G \right\| \leq C \log(2/\eta)^{3/2} \left[m^{\frac{1}{4}} \left\{ \sqrt{\frac{a^2 \sigma^2(0)}{n}} \vee \frac{4a^2}{n} \vee \frac{8a^2}{n^2} \right\}^{\frac{1}{2}} + \sigma(m) \right].$$

   If $n \geq 2$ then $8a^2/n^2 \leq 4a^2/n$. In summary,

   $$\left\| \Sigma^{\frac{1}{2}} g - G \right\| \leq C \log(2/\eta)^{3/2} \left[m^{\frac{1}{4}} \left\{ \frac{a^2 \sigma^2(0)}{n} + \frac{16a^4}{n^2} \right\}^{\frac{1}{4}} + \sigma(m) \right].$$

   Optimizing over m and absorbing constants then yields the result.

2. Proving (13). Since Δ_1 dominates Δ_2, w.p. $1 - 6\eta$

   $$\begin{aligned} \left\| \Sigma^{\frac{1}{2}} g - G \right\| &\leq C \log(2/\eta)^{3/2} \left[m^{\frac{1}{4}} \left\{ \frac{b^2 \sigma^2(0)}{\sqrt{n}} \right\}^{\frac{1}{2}} + \sigma(m) \right] \\ &= C \log(2/\eta)^{3/2} \left[m^{\frac{1}{4}} \left\{ \frac{b^4 \sigma^4(0)}{n} \right\}^{\frac{1}{4}} + \sigma(m) \right]. \end{aligned}$$

□

To simplify the exposition, we will introduce shorthand for the finite sample rate in our bootstrap couplings:

$$R_{\mathrm{bd}}(n) \coloneqq \inf_{m \geq 1} \left\{ m^{\frac{1}{4}} \left(\frac{a^2 \sigma^2(0)}{n} + \frac{a^4}{n^2} \right)^{\frac{1}{4}} + \sigma(m) \right\},$$

$$R_{\mathrm{sg}}(n) \coloneqq \inf_{m \geq 1} \left\{ \left(\frac{m b^4 \sigma^4(0)}{n} \right)^{\frac{1}{4}} + \sigma(m) \right\}.$$

We state the following corollary, which is useful in case we cannot sample from the multiplier bootstrap process $\hat{\Sigma}^{\frac{1}{2}} g$ directly, but can sample from a proxy for it.

Corollary D.15. Suppose W and W' are random variables taking values in H, such that (i) $\mathbb{P}(\|W' - W\| > \delta) \leq \eta$ and (ii) $W \overset{U}{\sim} \hat{\Sigma}^{\frac{1}{2}} g$. Then, under the same conditions as Theorem D.14, there exists a random element Z with the same conditional distribution as $\Sigma^{\frac{1}{2}} g$ such that with probability at least $1 - \eta$, either:

$$\mathbb{P}\left[\|Z - W'\| \geq C' \log(6/\eta)^{3/2} R_{\mathrm{bd}}(n) + \delta \middle| U \right] \leq \eta$$

under a-boundedness with $n \geq 4a^2 \vee 2$ and $\sigma(0) \geq 1$; or

$$\mathbb{P}\left[\|Z - W'\| \geq C' \log(6/\eta)^{3/2} R_{\mathrm{sg}}(n) + \delta \middle| U \right] \leq \eta$$

under b-sub-Gaussianity.

Proof. We cover the bounded case; the sub-Gaussian case is completely analogous. We proceed in steps.

1. Let $G \overset{U}{\sim} \hat{\Sigma}^{\frac{1}{2}} g$ be constructed as in Theorem D.14, so that

$$\mathbb{P}\left\{ \left\| \Sigma^{\frac{1}{2}} g - G \right\| \geq C \log(12/\eta)^{3/2} R_{\mathrm{bd}}(n) \right\} \leq \eta.$$

2. Find $Z \overset{U}{\sim} \Sigma^{\frac{1}{2}} g$ such that

$$\mathbb{P}\{\|Z - W\| \geq C \log(12/\eta)^{3/2} R_{\mathrm{bd}}(n)\} \leq \eta.$$

   The existence of such a Z follows immediately from Corollary D.9 given the above. In particular, let $X = W$, $X' = G$, $Y' = \Sigma^{\frac{1}{2}} g$, and $Y = Z$.

3. By the triangle inequality, since we have assumed $\mathbb{P}(\|W' - W\| \geq \delta) \leq \eta$, it holds with probability at least $1 - 2\eta$ that

$$\|Z - W'\| \leq \|Z - W\| + \|W - W'\| \leq C \log(12/\eta)^{3/2} R_{\mathrm{bd}}(n) + \delta.$$

Replacing η by $\eta/2$, the above may be rewritten

$$\mathbb{P}\left(\|Z - W'\| > C \log(24/\eta)^{3/2} R_{\mathrm{bd}}(n) + \delta\right) \leq \eta.$$

Thus we have

$$\begin{aligned}\eta &\geq \mathbb{P}\left\{ \|Z - W'\| \geq C \log(24/\eta)^{3/2} R_{\mathrm{bd}}(n) \right\} \\ &= \mathbb{E}\left[\mathbb{P}\left\{ \|Z - W'\| \geq C \log(24/\eta)^{3/2} R_{\mathrm{bd}}(n) \middle| U \right\} \right].\end{aligned}$$

4. We apply Markov's inequality to the random variable

$$A(\eta) = \mathbb{P}\left\{ \|Z - W'\| \geq C \log(24/\eta)^{3/2} R_{\mathrm{bd}}(n) \middle| U \right\}.$$

Then, for any $\eta \in (0, 1)$ we have

$$\mathbb{P}_U\{A(\eta) > t\} \leq \mathbb{E}_U[A(\eta)]/t \leq \eta/t.$$

Choose $t = \sqrt{\eta}$. Then, the above says that w.p. $1 - \sqrt{\eta}$

$$\mathbb{P}\left\{ \|Z - W'\| \geq C \log(24/\eta)^{3/2} R_{\mathrm{bd}}(n) \middle| U \right\} \leq \sqrt{\eta}$$

so that with w.p. $1 - \eta$

$$\mathbb{P}\left\{ \|Z - W'\| \geq C \log(24/\eta^2)^{3/2} R_{\mathrm{bd}}(n) \middle| U \right\} \leq \eta.$$

5. Finally note that

$$C\{\log(24/\eta^2)\}^{\frac{3}{2}} \leq C[\log\{(6/\eta)^2\}]^{\frac{3}{2}} \leq 2^{\frac{3}{2}} C\{\log(6/\eta)\}^{3/2} = C'\{\log(6/\eta)\}^{3/2}.$$

□

E Feasible bootstrap

The sketch in Section 6 claims a high-probability bound on $\|\mathfrak{B} - Z_{\mathfrak{B}}\|$, i.e.

$$\left\| \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}\frac{V_i - V_j}{\sqrt{2}} - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}\frac{\hat{V}_i - \hat{V}_j}{\sqrt{2}} \right\|.$$

We prove this approximation in what follows.

E.1 Decomposition

Define $\varepsilon_i^\lambda = Y_i - f_\lambda(X_i)$ and recall that

$$V_i = T_\lambda^{-1}\{(k_{X_i} \otimes k_{X_i}^*)(f_0 - f_\lambda) + \varepsilon_i k_{X_i}\} = T_\lambda^{-1}\{Y_i - f_\lambda(X_i)\}k_{X_i} = T_\lambda^{-1}\varepsilon_i^\lambda k_{X_i}$$

and that its feasible counterpart is

$$\hat{V}_i = \hat{T}_\lambda^{-1}\{(k_{X_i} \otimes k_{X_i}^*)(f_0 - \hat{f}) + \varepsilon_i k_{X_i}\} = \hat{T}_\lambda^{-1}\{Y_i - \hat{f}(X_i)\}k_{X_i} = \hat{T}_\lambda^{-1}\hat{\varepsilon}_i k_{X_i}.$$

To lighten notation, let

$$w_{ij} = \frac{1}{\sqrt{2}}(\varepsilon_i^\lambda k_{X_i} - \varepsilon_j^\lambda k_{X_j}), \quad \hat{w}_{ij} = \frac{1}{\sqrt{2}}(\hat{\varepsilon}_i k_{X_i} - \hat{\varepsilon}_j k_{X_j})$$

so that the comparison of interest is

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}\hat{T}_\lambda^{-1}\hat{w}_{ij} \approx \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}T_\lambda^{-1}w_{ij}.$$

Lemma E.1. We have that

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}\frac{\hat{V}_i - \hat{V}_j}{\sqrt{2}} - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}\frac{V_i - V_j}{\sqrt{2}} = \Delta_1 + \Delta_2$$

where

$$\Delta_1 = (\hat{T}_\lambda^{-1} - T_\lambda^{-1})\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{w}_{ij}h_{ij}\right), \quad \Delta_2 = T_\lambda^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(\hat{w}_{ij} - w_{ij})h_{ij}\right\}.$$

Proof. Write

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}h_{ij}\frac{\hat{V}_i-\hat{V}_j}{\sqrt{2}}-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}h_{ij}\frac{V_i-V_j}{\sqrt{2}}=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}h_{ij}\hat{T}_\lambda^{-1}\hat{w}_{ij}-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}h_{ij}T_\lambda^{-1}w_{ij}$$
$$=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}h_{ij}\{\hat{T}_\lambda^{-1}\hat{w}_{ij}-T_\lambda^{-1}w_{ij}\}$$
$$=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}h_{ij}\{\hat{T}_\lambda^{-1}\hat{w}_{ij}\pm T_\lambda^{-1}\hat{w}_{ij}-T_\lambda^{-1}w_{ij}\}.$$

□

E.2 Finite sample rate

The following is a special case of Fischer and Steinwart (2020, Theorem 16). We give a self-contained proof, as the version needed here is an easy consequence of our Proposition B.3.

Lemma E.2 (cf. Fischer and Steinwart (2020, Theorem 16))**.** Suppose $\|T_\lambda(\hat{T}-T)\|_{\mathrm{HS}}\leq \delta\leq\frac{1}{2}$ and $\left\|T_\lambda^{-1}\mathbb{E}_n[k_{X_i}\varepsilon_i]\right\|\leq\gamma$. Then we have $\|\hat{f}-f_\lambda\|\leq 2(\gamma+\delta\|f_0\|)$.

Proof. Recall Proposition B.3, which says that if $\|T_\lambda(\hat{T}-T)\|_{\mathrm{HS}}\leq\delta\leq\frac{1}{2}$ and $\left\|T_\lambda^{-1}\mathbb{E}_n[k_{X_i}\varepsilon_i]\right\|\leq\gamma$, then

$$\hat{f}-f_\lambda=T_\lambda^{-1}\mathbb{E}_n[k_{X_i}\varepsilon_i]+T_\lambda^{-1}(\hat{T}-T)(f_0-f_\lambda)+u$$

for some u with

$$\|u\|\leq 2\delta\left(\gamma+\delta\|f_0-f_\lambda\|\right).$$

Under these events, by the triangle inequality,

$$\|\hat{f}-f_\lambda\|\leq\gamma+\delta\|f_0-f_\lambda\|+2\delta(\gamma+\delta\|f_0-f_\lambda\|).$$

Since $f_0-f_\lambda=(I-T_\lambda^{-1}T)f_0$ and $0\preceq(I-T_\lambda^{-1}T)\preceq I$, it further follows that

$$\|\hat{f}-f_\lambda\|\leq\gamma+\delta\|f_0\|+2\delta(\gamma+\delta\|f_0\|)\leq 2(\gamma+\delta\|f_0\|),$$

where we again used our assumption that $\delta\leq\frac{1}{2}$. □

E.3 First term

We initially focus on $\Delta_1 = (\hat{T}_\lambda^{-1} - T_\lambda^{-1})v$ where $v = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \hat{w}_{ij} h_{ij}$.

Lemma E.3. Suppose $\|T_\lambda^{-1}(\hat{T} - T)\|_{\text{HS}} \leq \delta \leq \frac{1}{2}$. Then $\|\Delta_1\| \leq 2\delta \|T_\lambda^{-1} v\|$.

Proof. Recall Lemma B.2:

$$(\hat{T}_\lambda^{-1} - T_\lambda^{-1})v = A_1 v + A_2 T_\lambda^{-1} v + A_3 T_\lambda^{-1} v$$

where

$$\|A_1\|_{\text{HS}} \leq \frac{\delta^k}{\lambda}, \quad \|A_2\|_{\text{HS}} \leq \delta, \quad \|A_3\|_{\text{HS}} \leq \frac{\delta^2}{1-\delta}.$$

Therefore

$$\begin{aligned}
\|\Delta_1\| &= \|(\hat{T}_\lambda^{-1} - T_\lambda^{-1})v\| \\
&\leq \|A_1\|_{\text{HS}} \cdot \|v\| + \|A_2\|_{\text{HS}} \cdot \|T_\lambda^{-1} v\| + \|A_3\|_{\text{HS}} \cdot \|T_\lambda^{-1} v\| \\
&\leq \frac{\delta^k}{\lambda}\|v\| + \left(\delta + \frac{\delta^2}{1-\delta}\right) \|T_\lambda^{-1} v\| \\
&= \frac{\delta^k}{\lambda}\|v\| + \frac{\delta}{1-\delta}\|T_\lambda^{-1} v\|.
\end{aligned}$$

Taking the limit as $k \uparrow \infty$ removes the first term. Then $\delta \leq 1/2$ implies $\frac{1}{1-\delta} \leq 2$ so that

$$\frac{\delta}{1-\delta}\|T_\lambda^{-1} v\| \leq 2\delta \|T_\lambda^{-1} v\|.$$

□

Lemma E.4. Suppose $\|f - f_0\| \leq z_0$ and $\varepsilon_i \leq \bar{\sigma}$. Conditional on the data, w.p. $1 - \eta$,

$$\left\|T_\lambda^{-1} v\right\| \leq 8 \log(4/\eta)(\bar{\sigma}\kappa + \kappa^2 z_0) \left[\sqrt{\mathfrak{n}(\lambda)} + \left\{\frac{1}{n^{1/2}\lambda} \vee \frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}}\right\}\right].$$

Proof. We proceed in steps.

1. Borell's inequality (Lemma D.5). Since $Z = T_\lambda^{-1} v$ is Gaussian we know that with probability at least $1 - \eta$, $\|Z\| \leq \left(1 + \sqrt{2\log(1/\eta)}\right) \sqrt{\mathbb{E}\|Z\|^2}$. Therefore it suffices to control $\mathbb{E}\left\|T_\lambda^{-1} v\right\|^2$. Note that

$$\mathbb{E}_h \left\|T_\lambda^{-1} v\right\|^2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \|T_\lambda^{-1} \hat{w}_{ij}\|^2$$

where $\hat{w}_{ij} = \frac{1}{\sqrt{2}}(\hat{\varepsilon}_i k_{X_i} - \hat{\varepsilon}_j k_{X_j})$.

2. Single sum representation. Note that

$$\begin{aligned}\|T_\lambda^{-1}\hat{w}_{ij}\|^2 &= \left\|T_\lambda^{-1}\frac{1}{\sqrt{2}}(\hat{\varepsilon}_i k_{X_i} - \hat{\varepsilon}_j k_{X_j})\right\|^2 \\ &= \frac{1}{2}\left\|T_\lambda^{-1}(\hat{\varepsilon}_i k_{X_i} - \hat{\varepsilon}_j k_{X_j})\right\|^2 \\ &\leq \|T_\lambda^{-1}\hat{\varepsilon}_i k_{X_i}\|^2 + \|T_\lambda^{-1}\hat{\varepsilon}_j k_{X_j}\|^2\end{aligned}$$

by Cauchy-Schwartz. Hence

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\|T_\lambda^{-1}\hat{w}_{ij}\|^2 &\leq \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\|T_\lambda^{-1}\hat{\varepsilon}_i k_{X_i}\|^2 + \|T_\lambda^{-1}\hat{\varepsilon}_j k_{X_j}\|^2 \\ &= \frac{2}{n}\sum_{i=1}^n\|T_\lambda^{-1}\hat{\varepsilon}_i k_{X_i}\|^2 = 2\mathbb{E}_n[\xi_i]\end{aligned}$$

where $\xi_i = \|T_\lambda^{-1}\hat{\varepsilon}_i k_{X_i}\|^2$.

3. Concentration. We control ξ_i, $\mathbb{E}[\xi_i]$ and hence $\mathbb{E}[\xi_i^2] \leq \xi_i\mathbb{E}(\xi_i)$. Then we apply Bernstein's inequality (Lemma H.2). To begin, write

$$\hat{\varepsilon}_i k_{X_i} = \{Y_i - \hat{f}(X_i)\}k_{X_i} = \{\varepsilon_i + f_0(X_i) - \hat{f}(X_i)\}k_{X_i} = \varepsilon_i k_{X_i} + T_i(f_0 - \hat{f}).$$

Hence

$$\begin{aligned}\xi_i &= \|T_\lambda^{-1}\hat{\varepsilon}_i k_{X_i}\|^2 \\ &= \|T_\lambda^{-1}\{\varepsilon_i k_{X_i} + T_i(f_0 - \hat{f})\}\|^2 \\ &\leq \frac{1}{\lambda^2}(\bar{\sigma}\kappa + \kappa^2\|\hat{f} - f_0\|)^2 \\ &\leq \frac{1}{\lambda^2}(\bar{\sigma}\kappa + \kappa^2 z_0)^2.\end{aligned}$$

Moreover

$$\mathbb{E}\xi_i = \mathbb{E}\|T_\lambda^{-1}\hat{\varepsilon}_i k_{X_i}\|^2 \leq 2\mathbb{E}\|T_\lambda^{-1}\varepsilon_i k_{X_i}\|^2 + 2\mathbb{E}\|T_\lambda^{-1}T_i(f_0 - \hat{f})\|^2.$$

In the first term

$$\begin{aligned}\mathbb{E}\|T_\lambda^{-1}\varepsilon_i k_{X_i}\|^2 &= \int \left[\varepsilon_i^2 \langle k_{X_i}, T_\lambda^{-2} k_{X_i}\rangle\right] \mathrm{d}\mathbb{P} \\ &\leq \bar{\sigma}^2 \int \left[\langle k_{X_i}, T_\lambda^{-2} k_{X_i}\rangle\right] \mathrm{d}\mathbb{P} \\ &= \bar{\sigma}^2 \int \left[\mathrm{tr}(T_\lambda^{-2} T_i)\right] \mathrm{d}\mathbb{P} \\ &= \bar{\sigma}^2 \,\mathrm{tr}(T_\lambda^{-2} T) \\ &= \bar{\sigma}^2 \mathfrak{n}(\lambda).\end{aligned}$$

In the second term

$$\mathbb{E}\|T_\lambda^{-1} T_i (f_0 - \hat{f})\}\|^2 = \mathbb{E}\|T_\lambda^{-1} k_{X_i}\{f_0(X_i) - \hat{f}(X_i)\}\|^2 \leq \|f_0 - \hat{f}\|_\infty^2 \mathbb{E}\|T_\lambda^{-1} k_{X_i}\|^2.$$

Note that, as argued above,

$$\|f_0 - \hat{f}\|_\infty^2 \leq \kappa^2 \|f_0 - \hat{f}\|^2, \quad \mathbb{E}\|T_\lambda^{-1} k_{X_i}\|^2 = \int [\langle k_{X_i}, T_\lambda^{-2} k_{X_i}\rangle] \mathrm{d}\mathbb{P} = \mathfrak{n}(\lambda).$$

In summary, the second term is bounded as $\mathbb{E}\|T_\lambda^{-1} T_i (f_0 - \hat{f})\}\|^2 \leq \kappa^2 \mathfrak{n}(\lambda) \|f_0 - \hat{f}\|^2$ and hence

$$\begin{aligned}\mathbb{E}[\xi_i] &\leq 2\left(\bar{\sigma}^2 \mathfrak{n}(\lambda) + \kappa^2 \mathfrak{n}(\lambda) \|f_0 - \hat{f}\|^2\right) \\ &= 2\mathfrak{n}(\lambda)\left(\bar{\sigma}^2 + \kappa^2 \|f_0 - \hat{f}\|^2\right) \\ &\leq 2\mathfrak{n}(\lambda)\left(\bar{\sigma}^2 + \kappa^2 z_0^2\right).\end{aligned}$$

It follows that

$$\mathbb{E}[\xi_i^2] \leq 2\mathfrak{n}(\lambda)\left(\bar{\sigma}^2 + \kappa^2 z_0^2\right) \cdot \frac{1}{\lambda^2}(\bar{\sigma}\kappa + \kappa^2 z_0)^2.$$

Therefore by Bernstein's inequality (Lemma H.2), w.p. $1 - \eta$

$$|\mathbb{E}_n[\xi_i] - \mathbb{E}[\xi_i]| \leq 2\ln(2/\eta) \left\{ \frac{2}{n\lambda^2}(\bar{\sigma}\kappa + \kappa^2 z_0)^2 \vee \sqrt{2\mathfrak{n}(\lambda)\left(\bar{\sigma}^2 + \kappa^2 z_0^2\right) \cdot \frac{1}{n\lambda^2}(\bar{\sigma}\kappa + \kappa^2 z_0)^2} \right\}.$$

In particular,

$$
\begin{aligned}
&\mathbb{E}_n[\xi_i] \\
&\le \mathbb{E}[\xi_i] + 2\ln(2/\eta)\left\{\frac{2}{n\lambda^2}(\bar\sigma\kappa+\kappa^2 z_0)^2 \vee \sqrt{2\mathfrak{n}(\lambda)\left(\bar\sigma^2+\kappa^2 z_0^2\right)\cdot\frac{1}{n\lambda^2}(\bar\sigma\kappa+\kappa^2 z_0)^2}\right\} \\
&\le 2\mathfrak{n}(\lambda)\left(\bar\sigma^2+\kappa^2 z_0^2\right) \\
&\qquad + 2\ln(2/\eta)\left\{\frac{2}{n\lambda^2}(\bar\sigma\kappa+\kappa^2 z_0)^2 \vee \sqrt{2\mathfrak{n}(\lambda)\left(\bar\sigma^2+\kappa^2 z_0^2\right)\cdot\frac{1}{n\lambda^2}(\bar\sigma\kappa+\kappa^2 z_0)^2}\right\} \\
&\le 4\ln(2/\eta)(\bar\sigma\kappa+\kappa^2 z_0)^2\left[\mathfrak{n}(\lambda)+\left\{\frac{1}{n\lambda^2}\vee\sqrt{\frac{\mathfrak{n}(\lambda)}{n\lambda^2}}\right\}\right].
\end{aligned}
$$

4. Collecting results. We have shown by concentration that, w.p. $1-\eta$,

$$
\begin{aligned}
\mathbb{E}_h\left\|T_\lambda^{-1}v\right\|^2 &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\|T_\lambda^{-1}\hat w_{ij}\|^2 \\
&\le 2\mathbb{E}_n[\xi_i] \\
&\le R := 8\ln(2/\eta)(\bar\sigma\kappa+\kappa^2 z_0)^2\left[\mathfrak{n}(\lambda)+\left\{\frac{1}{n\lambda^2}\vee\sqrt{\frac{\mathfrak{n}(\lambda)}{n\lambda^2}}\right\}\right].
\end{aligned}
$$

Therefore by Borell's inequality, w.p. $1-2\eta$

$$
\begin{aligned}
\|T_\lambda v\| &\le \left(1+\sqrt{2\log(1/\eta)}\right)\sqrt{R} \\
&\le \left(1+\sqrt{2\log(1/\eta)}\right)\sqrt{8}\ln(2/\eta)^{1/2}(\bar\sigma\kappa+\kappa^2 z_0)\left[\sqrt{\mathfrak{n}(\lambda)}+\left\{\frac{1}{n^{1/2}\lambda}\vee\frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}}\right\}\right] \\
&\le 8\log(2/\eta)(\bar\sigma\kappa+\kappa^2 z_0)\left[\sqrt{\mathfrak{n}(\lambda)}+\left\{\frac{1}{n^{1/2}\lambda}\vee\frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}}\right\}\right].
\end{aligned}
$$

□

Proposition E.5. Suppose that $\|T_\lambda(\hat T - T)\|_{\mathrm{HS}} \le \delta \le \frac{1}{2}$, $\|f-f_0\| \le z_0$, and $\varepsilon \le \bar\sigma$. It then holds w.p. $1-\eta$ that

$$
\|\Delta_1\| \le 16\delta\log(4/\eta)(\bar\sigma\kappa+\kappa^2 z_0)\left[\sqrt{\mathfrak{n}(\lambda)}+\left\{\frac{1}{n^{1/2}\lambda}\vee\frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}}\right\}\right].
$$

Proof. By Lemma E.4, w.p. $1-\eta$

$$
\left\|T_\lambda^{-1}v\right\| \le 8\log(4/\eta)(\bar\sigma\kappa+\kappa^2 z_0)\left[\sqrt{\mathfrak{n}(\lambda)}+\left\{\frac{1}{n^{1/2}\lambda}\vee\frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}}\right\}\right].
$$

Therefore by Lemma E.3,

$$\begin{aligned}\|\Delta_1\| &\le 2\delta\|T_\lambda^{-1}v\| \\ &\le 16\delta\log(4/\eta)(\bar{\sigma}\kappa+\kappa^2 z_0)\left[\sqrt{\mathfrak{n}(\lambda)}+\left\{\frac{1}{n^{1/2}\lambda}\vee\frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}}\right\}\right].\end{aligned}$$

□

E.4 Second term

Next, we turn to $\Delta_2 = T_\lambda^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n(\hat{w}_{ij}-w_{ij})h_{ij}\right\}$. Let $T_i = k_{X_i}\otimes k_{X_i}^*$.

Lemma E.6. $\Delta_2 = T_\lambda^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\frac{1}{\sqrt{2}}(T_i-T_j)(f_\lambda-\hat{f})h_{ij}\right\}$.

Proof. Write

$$\begin{aligned}&\hat{w}_{ij}-w_{ij} \\ &= \frac{1}{\sqrt{2}}\left[\{Y_i-\hat{f}(X_i)\}k_{X_i}-\{Y_j-\hat{f}(X_j)\}k_{X_j}\right]-\frac{1}{\sqrt{2}}\left[\{Y_i-f_\lambda(X_i)\}k_{X_i}-\{Y_j-f_\lambda(X_j)\}k_{X_j}\right] \\ &= \frac{1}{\sqrt{2}}\left[\{f_\lambda(X_i)-\hat{f}(X_i)\}k_{X_i}-\{f_\lambda(X_j)-\hat{f}(X_j)\}k_{X_j}\right] \\ &= \frac{1}{\sqrt{2}}\left[T_i\{f_\lambda-\hat{f}\}-T_j\{f_\lambda-\hat{f}\}\right].\end{aligned}$$

□

Our strategy is then to write

$$\|\Delta_2\|\le\left\|T_\lambda^{-1}\left\{\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\frac{1}{\sqrt{2}}(T_i-T_j)h_{ij}\right\}\right\|_{op}\cdot\|\hat{f}-f_\lambda\|.$$

To control the second factor, we use Lemma E.2. For the former factor, we have the following.

Lemma E.7. With probability at least $1-\eta$, we have

$$\left\|\frac{1}{n}\sum_{i,j}^n T_\lambda^{-1}\left(\frac{T_i-T_j}{\sqrt{2}}\right)h_{ij}\right\|_{\mathrm{HS}}\le 8\kappa^2\ln(4/\eta)\left[\sqrt{\mathfrak{n}(\lambda)}+\frac{1}{\sqrt{n}\lambda}\vee\frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}}\right].$$

Proof. We proceed in steps.

1. Second moment. Consider

$$\mathbb{E}_h \left\| \frac{1}{n} \sum_{i,j}^{n} T_\lambda^{-1} \left(\frac{T_i - T_j}{\sqrt{2}} \right) h_{ij} \right\|_{\mathrm{HS}}^2 \leq \frac{1}{n^2} \sum_{i,j}^{n} \left\| T_\lambda^{-1} \left(\frac{T_i - T_j}{\sqrt{2}} \right) \right\|_{\mathrm{HS}}^2 = \frac{1}{2n^2} \sum_{i,j}^{n} \|B_i - B_j\|_{\mathrm{HS}}^2$$

where $B_i = T_\lambda^{-1} T_i$. By triangle inequality and AM-GM inequality, $\|B_i - B_j\|_{\mathrm{HS}}^2 \leq 2(\|B_i\|_{\mathrm{HS}}^2 + \|B_j\|_{\mathrm{HS}}^2)$. Therefore

$$\frac{1}{2n^2} \sum_{i,j}^{n} \|B_i - B_j\|_{\mathrm{HS}}^2 \leq \frac{1}{n^2} \sum_{i,j}^{n} \|B_i\|_{\mathrm{HS}}^2 + \|B_j\|_{\mathrm{HS}}^2 = \frac{2}{n} \sum_{i=1}^{n} \|B_i\|_{\mathrm{HS}}^2 .$$

In summary

$$\mathbb{E}_h \left\| \frac{1}{n} \sum_{i,j}^{n} T_\lambda^{-1} \left(\frac{T_i - T_j}{\sqrt{2}} \right) h_{ij} \right\|_{\mathrm{HS}}^2 \leq \frac{2}{n} \sum_{i=1}^{n} \|B_i\|_{\mathrm{HS}}^2 .$$

2. Concentration. Next we bound the sample mean $\mathbb{E}_n[\xi_i]$ where $\xi_i = \|B_i\|_{\mathrm{HS}}^2$. Note that $\xi_i = \|B_i\|_{\mathrm{HS}}^2 = \left\|T_\lambda^{-1} T_i\right\|_{\mathrm{HS}}^2 \leq \frac{\kappa^4}{\lambda^2}$. Moreover $\mathbb{E}(\xi_i^2) \leq \xi_i \mathbb{E}(\xi_i) \leq \frac{\kappa^4}{\lambda^2} \mathbb{E}(\xi_i)$. Within the final expression,

$$\begin{aligned}
\mathbb{E}(\xi_i) &= \mathbb{E}(\left\|T_\lambda^{-1} T_i\right\|_{\mathrm{HS}}^2) \\
&= \int \mathrm{tr}(T_i T_\lambda^{-2} T_i) \mathrm{d}\mathbb{P}(x) \\
&\leq \kappa^2 \int \mathrm{tr}(T_\lambda^{-2} T_i) \mathrm{d}\mathbb{P}(x) \\
&= \kappa^2 \,\mathrm{tr}(T_\lambda^{-2} T) \\
&= \kappa^2 \mathfrak{n}(\lambda).
\end{aligned}$$

Hence $\mathbb{E}(\xi_i^2) \leq \frac{\kappa^6}{\lambda^2} \mathfrak{n}(\lambda)$. Therefore by Bernstein's inequality (Lemma H.2), w.p. $1 - \eta$

$$|\mathbb{E}_n[\xi_i] - \mathbb{E}[\xi_i]| \leq 2 \ln(2/\eta) \left\{ \frac{2\kappa^4}{n\lambda^2} \vee \sqrt{\frac{\kappa^6 \mathfrak{n}(\lambda)}{n\lambda^2}} \right\}.$$

In particular,

$$\mathbb{E}_n[\xi_i] \leq \mathbb{E}[\xi_i] + 2 \ln(2/\eta) \left\{ \frac{2\kappa^4}{n\lambda^2} \vee \sqrt{\frac{\kappa^6 \mathfrak{n}(\lambda)}{n\lambda^2}} \right\} \leq \kappa^2 \mathfrak{n}(\lambda) + 2 \ln(2/\eta) \left\{ \frac{2\kappa^4}{n\lambda^2} \vee \sqrt{\frac{\kappa^6 \mathfrak{n}(\lambda)}{n\lambda^2}} \right\}.$$

Thus w.p. $1-\eta$

$$\mathbb{E}_h \left\| \frac{1}{n} \sum_{i,j}^{n} T_\lambda^{-1} \left(\frac{T_i - T_j}{\sqrt{2}} \right) h_{ij} \right\|_{\mathrm{HS}}^2 \leq \frac{2}{n} \sum_{i=1}^{n} \|B_i\|_{\mathrm{HS}}^2$$
$$\leq R \coloneqq 2\kappa^2 \mathfrak{n}(\lambda) + 4\ln(2/\eta) \left\{ \frac{2\kappa^4}{n\lambda^2} \vee \sqrt{\frac{\kappa^6 \mathfrak{n}(\lambda)}{n\lambda^2}} \right\}.$$

3. Borell's inequality (Lemma D.5). Since $Z = \frac{1}{n} \sum_{i,j}^{n} T_\lambda^{-1} \left(\frac{T_i - T_j}{\sqrt{2}} \right) h_{ij}$ is Gaussian, we know that with probability at least $1-\eta$, $\|Z\| \leq \left(1 + \sqrt{2\log(1/\eta)}\right) \sqrt{\mathbb{E}\|Z\|^2}$. Therefore it suffices to use the high probability bound $\mathbb{E}\|Z\|_{\mathrm{HS}}^2 \leq R$ derived above. In particular, w.p. $1-2\eta$

$$\left\| \frac{1}{n} \sum_{i,j}^{n} T_\lambda^{-1} \left(\frac{T_i - T_j}{\sqrt{2}} \right) h_{ij} \right\|_{\mathrm{HS}}$$
$$\leq \{1 + \sqrt{2\ln(1/\eta)}\} R^{1/2}$$
$$= \{1 + \sqrt{2\ln(1/\eta)}\} \left[2\kappa^2 \mathfrak{n}(\lambda) + 4\ln(2/\eta) \left\{ \frac{2\kappa^4}{n\lambda^2} \vee \sqrt{\frac{\kappa^6 \mathfrak{n}(\lambda)}{n\lambda^2}} \right\} \right]^{1/2}$$
$$\leq \{1 + \sqrt{2\ln(1/\eta)}\} \left[\sqrt{2}\kappa \sqrt{\mathfrak{n}(\lambda)} + 2\ln(2/\eta)^{1/2} \left\{ \frac{\sqrt{2}\kappa^2}{\sqrt{n}\lambda} \vee \frac{\kappa^{3/2} \mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}} \right\} \right]$$
$$\leq 8\ln(2/\eta) \left[\kappa \sqrt{\mathfrak{n}(\lambda)} + \left\{ \frac{\kappa^2}{\sqrt{n}\lambda} \vee \frac{\kappa^{3/2} \mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}} \right\} \right].$$

□

Proposition E.8. It holds w.p. $1-\eta$ that

$$\|\Delta_2\| \leq 8\kappa^2 \ln(4/\eta) \left[\sqrt{\mathfrak{n}(\lambda)} + \left\{ \frac{1}{\sqrt{n}\lambda} \vee \frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}} \right\} \right] \cdot \|\hat{f} - f_\lambda\|.$$

Proof. By Lemmas E.6 and E.7, it holds w.p. $1-\eta$ that

$$\|\Delta_2\| \leq \left\| T_\lambda^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{\sqrt{2}} (T_i - T_j) h_{ij} \right\} \right\|_{\mathrm{HS}} \cdot \|\hat{f} - f_\lambda\|$$
$$\leq 8\kappa^2 \ln(4/\eta) \left[\sqrt{\mathfrak{n}(\lambda)} + \left\{ \frac{1}{\sqrt{n}\lambda} \vee \frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}} \right\} \right] \cdot \|\hat{f} - f_\lambda\|.$$

□

E.5 Main result

Theorem E.9. Suppose that $n \geq 16\kappa^2 \ln(4/\eta)^2[\mathfrak{n}(\lambda) \vee \lambda^{-1}]$ and $\varepsilon_i \leq \bar{\sigma}$. It then holds w.p. $1-\eta$ that

$$\left\| \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n h_{ij} \frac{\hat{V}_i - \hat{V}_j}{\sqrt{2}} - \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n h_{ij} \frac{V_i - V_j}{\sqrt{2}} \right\|$$
$$\leq 96\kappa^2[\bar{\sigma} + \kappa\|f_0\|] \ln(16/\eta)^2 \left\{ \sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \vee \frac{2\kappa}{n\lambda} \right\} \cdot \left[\sqrt{\mathfrak{n}(\lambda)} + \left\{ \frac{1}{\sqrt{n}\lambda} \vee \frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}} \right\} \right].$$

Proof. By Lemma E.1 as well as Propositions E.5 and E.8, w.p. $1-2\eta$

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n h_{ij} \frac{\hat{V}_i - \hat{V}_j}{\sqrt{2}} - \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n h_{ij} \frac{V_i - V_j}{\sqrt{2}} \right\| \\
&\leq \|\Delta_1\| + \|\Delta_2\| \\
&\leq 16\delta(\bar{\sigma}\kappa + \kappa^2 z_0) \log(4/\eta) \left[\sqrt{\mathfrak{n}(\lambda)} + \left\{ \frac{1}{n^{1/2}\lambda} \vee \frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}} \right\} \right] \\
&\quad + 8\kappa^2 \|\hat{f} - f_\lambda\| \log(4/\eta) \left[\sqrt{\mathfrak{n}(\lambda)} + \left\{ \frac{1}{\sqrt{n}\lambda} \vee \frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}} \right\} \right].
\end{aligned}$$

Therefore it suffices to control $16\delta(\bar{\sigma}\kappa + \kappa^2 z_0)$ and $8\kappa^2\|\hat{f} - f_\lambda\|$. Recall that, as argued in Lemma E.2, when $\delta \leq 1/2$, $\|\hat{f} - f_\lambda\| \leq 2(\gamma + \delta\|f_0\|)$, so that

$$8\kappa^2\|\hat{f} - f_\lambda\| \leq 16\kappa^2(\gamma + \delta\|f_0\|) = 16\kappa^2\gamma + 16\kappa^2\delta\|f_0\|.$$

Moreover, as argued in Lemma E.2

$$z_0 = \|\hat{f} - f_\lambda\| + \|f_\lambda - f_0\| \leq \|\hat{f} - f_\lambda\| + \|f_0\| \leq 2(\gamma + \delta\|f_0\|) + \|f_0\| \leq 2(\gamma + \|f_0\|)$$

so that

$$\begin{aligned}
16\delta(\bar{\sigma}\kappa + \kappa^2 z_0) &\leq 16\delta\{\bar{\sigma}\kappa + 2\kappa^2(\gamma + \|f_0\|)\} \\
&= 16\delta\bar{\sigma}\kappa + 32\delta\kappa^2\gamma + 32\delta\kappa^2\|f_0\| \\
&\leq 16\delta\bar{\sigma}\kappa + 16\kappa^2\gamma + 32\delta\kappa^2\|f_0\|.
\end{aligned}$$

Therefore

$$\begin{aligned}
8\kappa^2\|\hat{f} - f_\lambda\| + 16\delta(\bar{\sigma}\kappa + \kappa^2 z_0) &\leq 16\kappa^2\gamma + 16\kappa^2\delta\|f_0\| + 16\delta\bar{\sigma}\kappa + 16\kappa^2\gamma + 32\delta\kappa^2\|f_0\| \\
&= [32\kappa^2]\gamma + [48\kappa^2\|f_0\| + 16\bar{\sigma}\kappa]\delta.
\end{aligned}$$

By combining Lemmas H.4 and H.5, we have with probability at least $1 - 2\eta$ that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} T_\lambda^{-1} \varepsilon_i k_{X_i} \right\| \leq 2\bar{\sigma} \ln(2/\eta) \left\{ \sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \vee \frac{2\kappa}{n\lambda} \right\} = \gamma$$
$$\left\| \frac{1}{n} \sum_{i=1}^{n} T_\lambda^{-1} (T_i - T) \right\|_{\mathrm{HS}} \leq 2\kappa \ln(2/\eta) \left\{ \sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \vee \frac{2\kappa}{n\lambda} \right\} = \delta.$$

Therefore

$$\begin{aligned}
&[32\kappa^2]\gamma + [48\kappa^2 \|f_0\| + 16\bar{\sigma}\kappa]\delta \\
&\leq [32\kappa^2\bar{\sigma} + 48\kappa^3\|f_0\| + 16\kappa^2\bar{\sigma}] 2 \ln(2/\eta) \left\{ \sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \vee \frac{2\kappa}{n\lambda} \right\} \\
&= [48\kappa^2\bar{\sigma} + 48\kappa^3\|f_0\|] 2 \ln(2/\eta) \left\{ \sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \vee \frac{2\kappa}{n\lambda} \right\} \\
&= 96\kappa^2[\bar{\sigma} + \kappa\|f_0\|] \ln(2/\eta) \left\{ \sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \vee \frac{2\kappa}{n\lambda} \right\}.
\end{aligned}$$

We conclude that w.p. $1 - 4\eta$

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij} \frac{\hat{V}_i - \hat{V}_j}{\sqrt{2}} - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij} \frac{V_i - V_j}{\sqrt{2}} \right\| \\
&\leq 96\kappa^2[\bar{\sigma} + \kappa\|f_0\|] \ln(4/\eta)^2 \left\{ \sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \vee \frac{2\kappa}{n\lambda} \right\} \cdot \left[\sqrt{\mathfrak{n}(\lambda)} + \left\{ \frac{1}{\sqrt{n\lambda}} \vee \frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}} \right\} \right].
\end{aligned}$$

Finally note that the condition on n ensures $\delta \leq 1/2$, as argued in the proof of Theorem B.4. □

F Practitioner's guide

The sketch in Section 6 has the step

$$\mathfrak{B}(x) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij} \frac{\hat{V}_i - \hat{V}_j}{\sqrt{2}} = K_x (K + n\lambda)^{-1} \operatorname{diag}(\hat{\varepsilon}) \frac{1}{\sqrt{2}} (h^\top - h) \mathbf{1},$$

where $(K_x)_i = k(x, X_i)$, $K_{ij} = k(X_i, X_j)$, $\hat{\varepsilon}_i = Y_i - \hat{f}(X_i)$, $\mathbf{1}_i = 1$, and h is a matrix whose entries are i.i.d. standard Gaussians. We prove this equality in what follows.

F.1 Simulation

Our goal is to sample from the symmetrized empirical process. Let us construct the estimated residuals $\hat{\varepsilon}_i = Y_i - \hat{f}(X_i)$, which are collected into the vector $\hat{\varepsilon} \in \mathbb{R}^n$. Let $h = (h_{ij})$ be an $n \times n$ matrix consisting of independent, standard Gaussian multipliers.

Theorem F.1. We have that $\mathfrak{B}(x) = K_x(K + n\lambda)^{-1} \operatorname{diag}(\hat{\varepsilon}) \frac{1}{\sqrt{2}}(h^\top - h)\mathbf{1}$.

Proof. We proceed in steps.

1. Notation. To begin, write

$$\mathfrak{B} = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n h_{ij} \frac{\hat{V}_i - \hat{V}_j}{\sqrt{2}} = \hat{T}_\lambda^{-1} \left[\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \left(\frac{k_{X_i}\hat{\varepsilon}_i - k_{X_j}\hat{\varepsilon}_j}{\sqrt{2}} \right) h_{ij} \right].$$

   Focusing on the inner expression

$$\begin{aligned}
\sum_{i=1}^n \sum_{j=1}^n \left(\frac{k_{X_i}\hat{\varepsilon}_i - k_{X_j}\hat{\varepsilon}_j}{\sqrt{2}} \right) h_{ij} &= \frac{1}{\sqrt{2}} \left(\sum_{i=1}^n k_{X_i}\hat{\varepsilon}_i \sum_{j=1}^n h_{ij} - \sum_{j=1}^n k_{X_j}\hat{\varepsilon}_j \sum_{i=1}^n h_{ij} \right) \\
&= \frac{1}{\sqrt{2}} \sum_{i=1}^n k_{X_i}\hat{\varepsilon}_i \left(\sum_{j=1}^n h_{ij} - h_{ji} \right) \\
&= \sum_{i=1}^n \beta_i k_{X_i}
\end{aligned}$$

   for $\beta = \operatorname{diag}(\hat{\varepsilon}) \frac{1}{\sqrt{2}}(h^\top - h)\mathbf{1}$. The argument above shows

$$\sum_{i=1}^n \sum_{j=1}^n \left(\frac{k_{X_i}\hat{\varepsilon}_i - k_{X_j}\hat{\varepsilon}_j}{\sqrt{2}} \right) h_{ij} = \Phi^* \beta, \quad \beta = \operatorname{diag}(\hat{\varepsilon}) \frac{1}{\sqrt{2}}(h^\top - h)\mathbf{1}$$

   where $\Phi : H \to \mathbb{R}^n$ is the sampling operator $\Phi : f \mapsto (\langle k_{X_i}, f \rangle)_{i=1}^n$, so $\hat{T} = \Phi^*\Phi$.

2. Sample representation. Therefore

$$\mathfrak{B} = \left(\frac{1}{n}\Phi^*\Phi + \lambda \right)^{-1} \left(\frac{1}{n}\Phi^*\beta \right) = \left(\Phi^*\Phi + n\lambda \right)^{-1} \Phi^*\beta = \Phi^*(\Phi\Phi^* + n\lambda)^{-1}\beta$$

   and hence $\mathfrak{B}(x) = K_x(K + n\lambda)^{-1}\beta$.

□

F.2 Inference

F.2.1 Fixed width

By simulation, we may sample from the distribution of

$$\mathrm{M} = \sup_{x \in S} |\mathfrak{B}(x)| = \sup_{x \in S} |K_x (K + n\lambda)^{-1} \operatorname{diag}(\hat{\varepsilon}) \frac{1}{\sqrt{2}} (h^\top - h)\mathbf{1}|.$$

Let $\hat{\tau}_\alpha$ denote the α quantile of this distribution, which is computed by simulation of h. Then, we may construct a fixed-width α-sensitivity band as the set

$$\hat{C}_\alpha = \left\{ (s,t) \in S \times \mathbb{R} \,\middle|\, |t - \hat{f}(s)| \leq \hat{\tau}_\alpha \right\}.$$

F.2.2 Variable width

We may also wish to incorporate information about the pointwise variance of our process. To do this, we first compute the estimate

$$\hat{\sigma}^2(x) = n \left\| K_x (K + n\lambda)^{-1} \operatorname{diag}(\hat{\varepsilon}) \right\|_{\mathbb{R}^n}^2.$$

We then sample from the distribution of

$$\mathrm{M}' = \sup_{x \in S} \left| \frac{\mathfrak{B}(x)}{\hat{\sigma}(x) n^{-1/2}} \right| = \sup_{x \in S} \left| \frac{K_x (K + n\lambda)^{-1} \operatorname{diag}(\hat{\varepsilon}) \frac{1}{\sqrt{2}} (h^\top - h)\mathbf{1}}{\| K_x (K + n\lambda)^{-1} \operatorname{diag}(\hat{\varepsilon}) \|_{\mathbb{R}^n}} \right|$$

to compute the α quantile, $\hat{\tau}'_\alpha$, again by simulation of h. Having done so, we may construct the variable width confidence band

$$\begin{aligned} \hat{C}'_\alpha &= \left\{ (s,t) \in S \times \mathbb{R} \,\middle|\, |t - \hat{f}(s)| \leq \hat{\sigma}(x) n^{-1/2} \tau'_\alpha \right\} \\ &= \left\{ (s,t) \in S \times \mathbb{R} \,\middle|\, |t - \hat{f}(s)| \leq \left\| K_x (K + n\lambda)^{-1} \operatorname{diag}(\hat{\varepsilon}) \right\|_{\mathbb{R}^n} \tau'_\alpha \right\}. \end{aligned}$$

The derivation of $\hat{\sigma}^2(x)$ is as follows. Recall that

$$\begin{aligned} \mathfrak{B}(x) &= K_x (K + n\lambda)^{-1} \beta \\ \beta &= \operatorname{diag}(\hat{\varepsilon}) \frac{1}{\sqrt{2}} (h^\top - h)\mathbf{1} = \sqrt{n} \operatorname{diag}(\hat{\varepsilon}) q \\ q &= \frac{1}{\sqrt{2n}} (h - h^\top)\mathbf{1}, \quad \operatorname{var}(q_i) < 1. \end{aligned}$$

Proposition F.2 (Pointwise variance estimation in small samples)**.** Taking the expectation over randomness in h,

$$\mathbb{E}_h\{\mathfrak{B}(x)\}^2 = nK_x(K+n\lambda)^{-1}\operatorname{diag}(\hat{\varepsilon})(I-\mathbf{1}\mathbf{1}^\top/n)\operatorname{diag}(\hat{\varepsilon})(K+n\lambda)^{-1}K_x^\top.$$

For large n, $\mathbf{1}\mathbf{1}^\top/n$ is close to $\mathbf{0}\mathbf{0}^\top$, and hence $\mathbb{E}_h\{\mathfrak{B}(x)\}^2$ is close to $\hat{\sigma}^2(x)$. For small samples, the more complicated expression may be preferred.

Proof. Recall that

$$q_i = \frac{1}{\sqrt{2n}}\sum_{j=1}^{n}(h_{ij}-h_{ji}), \quad q = \frac{1}{\sqrt{2n}}(h-h^\top)\mathbf{1}, \quad var(q_i) < 1.$$

We proceed in steps, suppressing the subscript h which mean integrating over h.

1. For diagonal terms, fix i and write

$$\begin{aligned}
\mathbb{E}(q_i^2) &= \frac{1}{2n}\sum_{j=1}^{n}\sum_{k=1}^{n}\mathbb{E}\left\{(h_{ij}-h_{ji})(h_{ik}-h_{ki})\right\} \\
&= \frac{1}{2n}\sum_{j=1}^{n}\sum_{k=1}^{n}\mathbb{E}\left(h_{ij}h_{ik}-h_{ij}h_{ki}-h_{ji}h_{ik}+h_{ji}h_{ki}\right) \\
&= \frac{1}{2n}\sum_{j=1}^{n}\sum_{k=1}^{n}\left(1_{j=k}-1_{i=j=k}-1_{i=j=k}+1_{j=k}\right) \\
&= 1-1/n.
\end{aligned}$$

2. For off diagonal terms, fix $i \neq \ell$ and write

$$\begin{aligned}
\mathbb{E}(q_iq_\ell) &= \frac{1}{2n}\sum_{j=1}^{n}\sum_{k=1}^{n}\mathbb{E}\left\{(h_{ij}-h_{ji})(h_{\ell k}-h_{k\ell})\right\} \\
&= \frac{1}{2n}\sum_{j=1}^{n}\sum_{k=1}^{n}\mathbb{E}\left(h_{ij}h_{\ell k}-h_{ij}h_{k\ell}-h_{ji}h_{\ell k}+h_{ji}h_{k\ell}\right) \\
&= \frac{1}{2n}\sum_{j=1}^{n}\sum_{k=1}^{n}\left(1_{i=\ell,j=k}-1_{i=k,j=\ell}-1_{j=\ell,i=k}+1_{j=k,i=\ell}\right) \\
&= -1/n.
\end{aligned}$$

3. We conclude that $\mathbb{E}(qq^\top) = I-\mathbf{1}\mathbf{1}^\top/n$.

4. Collecting results, we have

$$\mathbb{E}\{\mathfrak{B}(x)\}^2 = K_x(K+n\lambda)^{-1}\mathbb{E}(\beta\beta^\top)(K+n\lambda)^{-1}K_x^\top$$

where

$$\mathbb{E}(\beta\beta^\top) = n\,\mathrm{diag}(\hat{\varepsilon})\mathbb{E}(qq^\top)\,\mathrm{diag}(\hat{\varepsilon}) = n\,\mathrm{diag}(\hat{\varepsilon})(I - \mathbf{1}\mathbf{1}^\top/n)\,\mathrm{diag}(\hat{\varepsilon}).$$

□

G Uniform confidence bands

G.1 Incremental factor approach

Lemma G.1 (One-sided error bound)**.** Let V, W be random variables such that $\mathbb{P}(|V-W|> r_1|\mathcal{A}) \leq r_2$ for some $r_1, r_2 > 0$, where $\mathcal{A}$ is σ-subalgebra of $\mathbb{P}$. Then, for any $t \in \mathbb{R}$

$$\mathbb{P}(V > t|\mathcal{A}) \leq \mathbb{P}(W > t - r_1|\mathcal{A}) + r_2.$$

Proof of Lemma G.1. Note that if $V > t$ then either $W > t - r_1$ or $|W - V| \geq r_1$. Thus, for any $A \in \mathcal{A}$, $\mathbb{1}\{V > t\}\mathbb{1}_A \leq (\mathbb{1}\{W > t - r_1\} + \mathbb{1}\{|W - V| \geq r_1\})\mathbb{1}_A$. The result follows by taking expectations, noting that $A \in \mathcal{A}$ was arbitrary, and using the definition of conditional expectation. □

Definition G.2 (Lighter notation)**.** To streamline notation, we define the functions Q, R, Δ, L, and B so that the following statements hold:

1. There exists a Gaussian random element Z in H such that with probability at least $1-\eta$

$$\left\|\sqrt{n}(\hat{f} - f_\lambda) - Z\right\| \leq Q(n, \lambda, \eta).$$

2. There exists a random element Z' in H whose conditional distribution given U is almost surely Gaussian with covariance Σ, and with probability at least $1-\eta$,

$$\mathbb{P}\left[\, \|\mathfrak{B} - Z'\| \leq R(n, \lambda, \eta)\Big|U\right] \geq 1 - \eta.$$

3. We define Δ by $\Delta(n, \lambda, \eta) \coloneqq Q(n, \lambda, \eta) + R(n, \lambda, \eta)$

4. It holds with probability $1 - \eta$ that $\|Z\| \geq L(\lambda, \eta)$ for some strictly increasing function L.

5. $\sqrt{n}\|f_\lambda - f_0\| \leq B(\lambda, n)$.

We abbreviate these quantities by Q, R, Δ, L, and B, respectively.

Proposition G.3 (High probability events)**.** With probability $1 - \eta$ it simultaneously holds that

$$\mathbb{P}(\|Z'\| > t + \Delta | U) \leq \mathbb{P}(\|\mathfrak{B}\| > t | U) + \eta \tag{14}$$

$$\mathbb{P}(\sqrt{n}\|\hat{f} - f_0\| > t + \Delta + B) \leq \mathbb{P}(\|\mathfrak{B}\| > t | U) + 2\eta \tag{15}$$

$$\mathbb{P}(\|\mathfrak{B}\| > t + \Delta + B | U) \leq \mathbb{P}(\sqrt{n}\|\hat{f} - f_0\| > t) + 2\eta. \tag{16}$$

Proof. The proof is entirely analogous to Theorem G.6, only we replace the use of Lemma G.4 by Lemma G.1:

$$\mathbb{P}(|V - W| > r_1 | \mathcal{A}) \leq r_2 \implies \mathbb{P}(V > t | \mathcal{A}) \leq \mathbb{P}(W > t - r_1 | \mathcal{A}) + r_2.$$

We proceed in steps.

1. First, recall that

$$\left\|\sqrt{n}(\hat{f} - f_0) - Z\right\| \leq \left\|\sqrt{n}(f_\lambda - f_0)\right\| + \left\|\sqrt{n}(\hat{f} - f_\lambda) - Z\right\| \leq B + Q$$

with probability at least $1 - \eta$, for some Gaussian random variable Z with covariance Σ. By Lemma G.1 with $\mathcal{A}$ chosen to be the trivial σ-algebra, $V = \sqrt{n}\|\hat{f} - f_0\|$, $W = \|Z\|$, $r_1 = Q + B$, and $r_2 = \eta$ we then have

$$\mathbb{P}\left[\sqrt{n}\|\hat{f} - f_0\| > t\right] \leq \mathbb{P}\left(\|Z\| > t - (Q + B)\right) + \eta.$$

Taking $t = s + Q + R + B$,

$$\mathbb{P}\left[\sqrt{n}\|\hat{f} - f_0\| > s + Q + R + B\right] \leq \mathbb{P}\left(\|Z\| > s + R\right) + \eta.$$

2. Next, by Proposition G.15 on an event $\mathcal{E}_1$ with probability $1-\eta$,

$$\mathbb{P}\left[\, \|\mathfrak{B} - Z'\| > R \middle| U \right] \leq \eta$$

for some Z' whose conditional distribution is Gaussian with covariance Σ. By Lemma G.1, with $\mathcal{A} = \sigma(U)\}$, with $V = \|Z'\|$, $W = \|\mathfrak{B}\|$, $r_1 = R$, and $r_2 = \eta$, it follows that on $\mathcal{E}_1$,

$$\mathbb{P}\left[\|Z'\| > t | U\right] \leq \mathbb{P}\left(\|\mathfrak{B}\| > t - R | U\right) + \eta.$$

Taking $t = s + R$, on $\mathcal{E}_1$,

$$\mathbb{P}\left[\|Z'\| > s + R | U\right] \leq \mathbb{P}\left(\|\mathfrak{B}\| > s | U\right) + \eta.$$

In light of the fact that $R \leq \Delta$, this immediately implies (14).

3. Since the conditional distribution of Z' given U is Gaussian with covariance Σ, as is the marginal distribution of Z, we can combine the previous two steps to deduce that on $\mathcal{E}_1$

$$\begin{aligned}
\mathbb{P}\left[\sqrt{n}\|\hat{f} - f_0\| > s + Q + R + B\right] &\leq \mathbb{P}\left(\|Z\| > s + R\right) + \eta \\
&= \mathbb{P}\left(\|Z'\| > s + R | U\right) + \eta \\
&\leq \mathbb{P}\left(\|\mathfrak{B}\| > s | U\right) + 2\eta.
\end{aligned}$$

Noting that $\Delta = Q + R$ gives (15).

4. Reversing the argument with the roles of V and W interchanged in Lemma G.1 yields (16). In particular, on $\mathcal{E}_1$,

$$\mathbb{P}\left[\|\mathfrak{B}\| > t | U\right] \leq \mathbb{P}\left(\|Z'\| > t - R | U\right) + \eta.$$

Taking $t = s + Q + R + B$, then on $\mathcal{E}_1$,

$$\mathbb{P}\left[\|\mathfrak{B}\| > s + Q + R + B | U\right] \leq \mathbb{P}\left(\|Z'\| > s + Q + B | U\right) + \eta.$$

Moreover,

$$\mathbb{P}\left[\|Z\| > t\right] \leq \mathbb{P}\left(\sqrt{n}\|\hat{f} - f_0\| > t - (Q + B)\right) + \eta.$$

Taking $t = s + Q + B$,

$$\mathbb{P}\left[\|Z\| > s + Q + B\right] \leq \mathbb{P}\left(\sqrt{n}\|\hat{f} - f_0\| > s\right) + \eta.$$

Therefore on $\mathcal{E}_1$

$$\begin{aligned}\mathbb{P}\left[\|\mathfrak{B}\| > s + Q + R + B|U\right] &\leq \mathbb{P}\left(\|Z'\| > s + Q + B|U\right) + \eta \\ &= \mathbb{P}\left(\|Z\| > s + Q + B\right) + \eta \\ &\leq \mathbb{P}\left(\sqrt{n}\|\hat{f} - f_0\| > s\right) + 2\eta.\end{aligned}$$

□

Proof of Theorem 6.2. Let $\mathcal{E}_1$ denote the event in Proposition G.3. We proceed in steps.

1. Let us first work conditioned upon this $\mathcal{E}_1$. By (14) in Proposition G.3, and since the law of Z is the same as the conditional law of Z' given U, it holds that

   $$\mathbb{P}(\|Z\| > \hat{t}_\alpha + \Delta) = \mathbb{P}(\|Z'\| > \hat{t}_\alpha + \Delta|U) \leq \alpha + \eta.$$

   Meanwhile, by Lemma G.13 and the fact that L is strictly increasing in its latter argument, we have that $L(\lambda, 1 - \alpha - 2\eta) < L(\lambda, 1 - \alpha - \eta)$ and hence

   $$\mathbb{P}\{\|Z\| > L(\lambda, 1 - \alpha - 2\eta)\} \geq \mathbb{P}\{\|Z\| \geq L(\lambda, 1 - \alpha - \eta)\} \geq \alpha + \eta.$$

   Thus, putting $\tilde{L} \coloneqq L(\lambda, 1 - \alpha - 2\eta)$, we can conclude that

   $$\mathbb{P}\{\|Z\| > \tilde{L}\} \geq \mathbb{P}(\|Z\| > \hat{t}_\alpha + \Delta)$$

   and therefore $\hat{t}_\alpha + \Delta \geq \tilde{L} \iff \hat{t}_\alpha \geq \tilde{L} - \Delta$.

2. Consider any δ that satisfies $\frac{1}{2} \geq \delta \geq \frac{\Delta + B(n,\lambda)}{\tilde{L} - \Delta}$. Then

$$\delta \hat{t}_\alpha \geq \frac{\Delta + B(n,\lambda)}{\tilde{L} - \Delta} \tilde{L} - \Delta = \Delta + B(n,\lambda).$$

It follows that $(1+\delta)\hat{t}_\alpha \geq \hat{t}_\alpha + B + \Delta$, so by (15) in Proposition G.3, we have

$$\mathbb{P}\{\sqrt{n}\|\hat{f} - f_0\| > (1+\delta)\hat{t}_\alpha\} \leq \alpha + 2\eta.$$

Another implication is that $(1-\delta)\hat{t}_\alpha \leq \hat{t}_\alpha - B - \Delta$, so by (16) in Proposition G.3 we have that

$$\mathbb{P}\{\sqrt{n}\|\hat{f} - f_0\| > (1-\delta)\hat{t}_\alpha\} \geq \alpha - 2\eta.$$

3. Finally, note that on the complementary event $\mathcal{E}_1^c$, $\mathbb{P}\{\sqrt{n}\|\hat{f} - f_0\| > -\}$ must take values between 0 and 1.

One implication is that, unconditionally, we must have

$$\mathbb{P}\{\sqrt{n}\|\hat{f} - f_0\| > (1+\delta)\hat{t}_\alpha\} \leq \alpha + 3\eta \iff \mathbb{P}\{\sqrt{n}\|\hat{f} - f_0\| \leq (1+\delta)\hat{t}_\alpha\} > 1 - \alpha - 3\eta.$$

Recall that $\hat{S}_n$ is τ-*honest* at level α if

$$\mathbb{P}(f_0 \in \hat{S}_n) \geq 1 - \alpha - \tau.$$

Therefore, unconditionally, the confidence set given by $\hat{S}_\alpha$ is 3η honest.

Another implication is that, unconditionally,

$$\mathbb{P}\{\sqrt{n}\|\hat{f} - f_0\| > (1-\delta)\hat{t}_\alpha\} \geq \alpha - 2\eta \iff \mathbb{P}\{\sqrt{n}\|\hat{f} - f_0\| \leq (1-\delta)\hat{t}_\alpha\} < 1 - \alpha + 2\eta.$$

Recall that $\hat{S}_n$ is (δ', τ)-*sharp* at level α if

$$\mathbb{P}\{f_0 \in \delta' \hat{f} + (1-\delta')\hat{S}_n\} \leq 1 - \alpha + \tau.$$

We finally show that, unconditionally, the confidence set given by $\hat{S}_\alpha$ is $(2\delta, 2\eta)$-sharp. It suffices to show that

$$\sqrt{n}\|\hat{f} - f_0\| \leq (1-\delta)\hat{t}_\alpha \implies f_0 \in 2\delta\hat{f} + (1-2\delta)\hat{S}_n.$$

The latter expression may be rewritten as

$$\|\hat{f} - f_0\| \leq (1-2\delta)\|h\| \leq (1-2\delta)(1+\delta)\hat{t}_\alpha n^{-1/2} \leq (1-\delta)\hat{t}_\alpha n^{-1/2}$$

by the definition of $\hat{S}_n$ and the fact that $\delta \leq 1/2$ implies $(1-2\delta)(1+\delta) \leq (1-\delta)$.

□

G.2 Anti-concentration approach

Having coupled the distributions of the ridge error process $\sqrt{n}(\hat{f} - f_\lambda)$ and that of the feasible bootstrap process $\mathfrak{B}$ with respect to the norm in H, we obtain valid inference for any functional $F : H \mapsto \mathbb{R}$ provided that uniform continuity and anti-concentration conditions are satisfied.

Lemma G.4 (cf. Chernozhukov et al. (2016, Lemma 2.1))**.** Let V, W be real-valued random variables such that $\mathbb{P}(|V - W| > r_1 | \mathcal{A}) \leq r_2$ for some constants $r_1, r_2 > 0$ where $\mathcal{A}$ is a σ-algebra comprised of Borel sets. Then we have

$$\sup_{t \in \mathbb{R}} \Big| \mathbb{P}(V \leq t | \mathcal{A}) - \mathbb{P}(W \leq t | \mathcal{A}) \Big| \leq \sup_{t \in \mathbb{R}} \mathbb{P}\left(|W - t| \leq r_1 \Big| \mathcal{A} \right) + r_2.$$

Proof. We proceed in steps.

1. To begin, we show that for fixed (V, W, t, z), where $z > 0$,

$$\mathbb{1}\{V \leq t\} - \mathbb{1}\{W \leq t\} \leq \mathbb{1}\{|W - t| \leq z\} \vee \mathbb{1}\{|V - W| > z\}.$$

   If the left-hand side is 1, then $V \leq t < W$, so $|W - t| < |W - V|$. Thus, it is not possible $|W - t| > z \geq |W - V|$, so the right-hand side is also 1. Otherwise the left hand side is at most 0 and the right hand side is at least 0.

2. Taking $z = r_1$, we write

$$\begin{aligned} \mathbb{1}\{V \leq t\} - \mathbb{1}\{W \leq t\} &\leq \mathbb{1}\{|W - t| \leq r_1\} \vee \mathbb{1}\{|V - W| > r_1\} \\ &\leq \mathbb{1}\{|W - t| \leq r_1\} + \mathbb{1}\{|V - W| > r_1\}. \end{aligned}$$

Thus, for any $A \in \mathcal{A}$, multiplication by $\mathbb{1}_A$ yields

$$\mathbb{E}\left[\mathbb{1}_A \mathbb{1}\{V \le t\}\right] - \mathbb{E}\left[\mathbb{1}_A \mathbb{1}\{W \le t\}\right] \le \mathbb{E}\left[\mathbb{1}_A \mathbb{1}\{|W - t| \le r_1\}\right] + \mathbb{E}\left[\mathbb{1}_A \mathbb{1}\{|V - W| > r_1\}\right]$$

so that by the definition of conditional expectation

$$\begin{aligned}\mathbb{P}(V \le t|\mathcal{A}) - \mathbb{P}(W \le t|\mathcal{A}) &\le \mathbb{P}(|W - t| \le r_1|\mathcal{A}) + \mathbb{P}(|V - W| > r_1|\mathcal{A}) \\ &\le \mathbb{P}(|W - t| \le r_1|\mathcal{A}) + r_2.\end{aligned}$$

3. Now, note that the same bound holds with the roles of V and W interchanged, so that we may replace the left-hand side by its absolute value. Finally, we take the supremum over $t \in \mathbb{R}$.

□

Definition G.5 (Heavier notation)**.** In what follows, define

$$\Delta_{\text{bd}} := \psi\left[C' M \log(16/\eta)^2 \left\{Q_{\text{bd}}(T, n, \lambda) + R_{\text{bd}}(T, n, \lambda) + \frac{\kappa \mathfrak{n}(\lambda)}{\sqrt{n}}\right\}\right]$$
$$\Delta_{\text{sg}} := \psi\left[C' \bar{M}^4 b^3 (16/\eta)^{1/\log(n)} \left\{Q_{\text{sg}}(T, n, \lambda) + R_{\text{sg}}(T, n, \lambda) + \frac{\mathfrak{n}(\lambda)}{\sqrt{n}}\right\}\right]$$

where C' is a sufficiently large universal constant and

$$Q_{\text{bd}}(n, \lambda) = \frac{1}{\lambda} \inf_{m \ge 1}\left\{\sigma(m) + \frac{m^2 \log(m^2)}{\sqrt{n}}\right\}, \qquad Q_{\text{sg}}(n, \lambda) = \inf_{m \ge 1}\left\{\frac{\sigma(m)}{\lambda} + \frac{m^{\frac{3}{2}} \log(n)^2}{\sqrt{n\lambda}}\right\},$$
$$R_{\text{bd}}(n, \lambda) = \inf_{m \ge 1}\left[\left\{\frac{m\, \mathfrak{n}(\lambda)}{\lambda^2 n} + \frac{m}{\lambda^4 n^2}\right\}^{\frac{1}{4}} + \frac{\sigma(m)}{\lambda}\right], \quad R_{\text{sg}}(n, \lambda) = \inf_{m \ge 1}\left[\left\{\frac{m\, \mathfrak{n}^2(\lambda)}{n}\right\}^{\frac{1}{4}} + \frac{\sigma(m)}{\lambda}\right].$$

Theorem G.6 (Inference)**.** Suppose F is uniformly continuous.

1. If in addition the conditions of Propositions G.14 and G.15 hold, set $\Delta = \Delta_{\text{bd}}$.

2. If in addition the conditions of Proposition G.16 and G.17 hold, set $\Delta = \Delta_{\text{sg}}$.

Then with probability $1 - \eta$:

$$\sup_{t \in \mathbb{R}}\left|\mathbb{P}\left[F\left\{\sqrt{n}(\hat{f} - f_\lambda)\right\} \le t\right] - \mathbb{P}\left\{F(\mathfrak{B}) \le t \middle| U\right\}\right| \le 2\left[\sup_{t \in \mathbb{R}} \mathbb{P}\left\{|F(\mathfrak{B}) - t| \le 2\Delta \middle| U\right\} + \eta\right].$$

If in addition Assumption 6.5 holds, then on this same event we have

$$\sup_{t \in \mathbb{R}}\left|\mathbb{P}\left[F\left\{\sqrt{n}(\hat{f} - f_\lambda)\right\} \le t\right] - \mathbb{P}\left\{F(\mathfrak{B}) \le t \middle| U\right\}\right| \le 2(\zeta\Delta + \eta).$$

Proof. We proceed in steps, focusing on the case where $\Delta = \Delta_{\mathrm{bd}}$.

1. First, recall that by Proposition G.14,

$$\left\|\sqrt{n}(\hat{f} - f_\lambda) - Z\right\| \lesssim Q_{\mathrm{bd}}(T, n, \lambda) M \log(12/\eta) + \frac{\mathfrak{n}(\lambda)}{\sqrt{n}} M \log(8/\eta)^2$$

with probability at least $1 - \eta$, for some Gaussian random variable Z with covariance Σ. Using uniform continuity of the functional F, it follows that on this same event,

$$|F\{\sqrt{n}(\hat{f} - f_\lambda)\} - F(Z)| \leq \psi\left[C' M \log(12/\eta)^2 \left\{Q_{\mathrm{bd}}(T, n, \lambda) + \frac{\mathfrak{n}(\lambda)}{\sqrt{n}}\right\}\right] \leq \Delta.$$

By Lemma G.4 (with $\mathcal{A}$ chosen to be the trivial σ-algebra) with $V = F\{\sqrt{n}(\hat{f} - f_\lambda)\}$, $W = F(Z)$, $r_1 = \Delta$, and $r_2 = \eta$ we then have

$$\sup_{t \in \mathbb{R}} \left|\mathbb{P}\left[F\left\{\sqrt{n}(\hat{f} - f_\lambda)\right\} \leq t\right] - \mathbb{P}\left\{F(Z) \leq t\right\}\right| \leq \sup_{t \in \mathbb{R}} \mathbb{P}\left\{|F(Z) - t| \leq \Delta \middle| \mathcal{A}\right\} + \eta. \tag{17}$$

The same argument holds for $\Delta = \Delta_{\mathrm{sg}}$, instead appealing to Proposition G.16.

2. Next, by Proposition G.15 with probability $1 - \eta$,

$$\mathbb{P}\left[\|\mathfrak{B} - Z'\| \lesssim M \log(16/\eta)^2 \left\{R_{\mathrm{bd}}(T, n, \lambda) + \frac{\kappa \mathfrak{n}(\lambda)}{\sqrt{n}}\right\} \middle| U\right] \geq 1 - \eta$$

for some Z' whose conditional distribution is Gaussian with covariance Σ. Again, by uniform continuity of F, we have on this event that

$$\mathbb{P}\left\{|F(\mathfrak{B}) - F(Z')| > \Delta \middle| U\right\} \leq \eta \tag{18}$$

where C' in the definition of Δ may have increased to absorb the universal constant in the preceding display.

By Lemma G.4, {with $\mathcal{A} = \sigma(U)$} with $V = F(\mathfrak{B})$, $W = F(Z')$, $r_1 = \Delta$, and $r_2 = \eta$, it follows that with probability at least $1 - \eta$,

$$\sup_{t \in \mathbb{R}} \left|\mathbb{P}\left\{F(\mathfrak{B}) \leq t | U\right\} - \mathbb{P}\left\{F(Z') \leq t | U\right\}\right| \leq \sup_{t \in \mathbb{R}} \mathbb{P}\left\{|F(Z') - t| \leq \Delta \middle| U\right\} + \eta. \tag{19}$$

The same argument holds for $\Delta = \Delta_{\mathrm{sg}}$, instead appealing to Proposition G.17.

3. To recover the first statement, we provide bounds on (17) and (19). We begin by showing that, for any $\delta > 0$,

$$\mathcal{E}_1 = \{|F(Z')-t|\leq \delta \mid U\} \subset \{|F(\mathfrak{B})-t|\leq 2\delta \mid U\} \cup \{|F(\mathfrak{B})-F(Z')|> \delta \mid U\} = \mathcal{E}_2.$$

We prove this result by contradiction; we argue that $\mathcal{E}_2^c \subset \mathcal{E}_1^c$. To begin, write

$$\begin{aligned}\mathcal{E}_2^c &= \{|F(\mathfrak{B}) - t|\leq 2\delta \mid U\}^c \cap \{|F(\mathfrak{B}) - F(Z')|> \delta \mid U\}^c \\ &= \{|F(\mathfrak{B}) - t|> 2\delta \mid U\} \cap \{|F(\mathfrak{B}) - F(Z')|\leq \delta \mid U\}.\end{aligned}$$

Under $\mathcal{E}_2^c$, conditional on U,

$$2\delta < |F(\mathfrak{B}) - t|\leq |F(\mathfrak{B}) - F(Z')|+|F(Z') - t|\leq \delta + |F(Z') - t|$$

which implies $\delta < |F(Z') - t|$, i.e. $\mathcal{E}_1^c$. Therefore

$$\begin{aligned}\mathbb{P}\left\{|F(Z) - t|\leq \delta\right\} &= \mathbb{P}\left\{|F(Z') - t|\leq \delta \middle| U\right\} \\ &= \mathbb{P}(\mathcal{E}_1) \\ &\leq \mathbb{P}(\mathcal{E}_2) \\ &\leq \mathbb{P}\left\{|F(\mathfrak{B}) - t|\leq 2\delta \middle| U\right\} + \mathbb{P}\left\{|F(Z') - F(\mathfrak{B})|> \delta \middle| U\right\} \\ &\leq \mathbb{P}\left\{|F(\mathfrak{B}) - t|\leq 2\delta \middle| U\right\} + \eta\end{aligned}$$

where the last inequality hold with probability $1 - \eta$ conditional on our chosen event and (18). Taking the supremum over $t \in \mathbb{R}$ and setting $\delta = \Delta$ yields

$$\sup_{t\in\mathbb{R}} \mathbb{P}\left\{|F(Z) - t|\leq \Delta\right\} = \sup_{t\in\mathbb{R}} \mathbb{P}\left\{|F(Z') - t|\leq \Delta \middle| U\right\} \leq \sup_{t\in\mathbb{R}} \mathbb{P}\left\{|F(\mathfrak{B}) - t|\leq 2\Delta \middle| U\right\}+\eta.$$

Thus, by the triangle inequality for the sup-norm, with probability $1 - \eta$,

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\left[F\left\{\sqrt{n}(\hat{f} - f_\lambda)\right\} \leq t\right] - \mathbb{P}\left\{F(\mathfrak{B}) \leq t \middle| U\right\}\right| \leq 2\left[\sup_{t\in\mathbb{R}} \mathbb{P}\left\{|F(\mathfrak{B}) - t|\leq 2\Delta \middle| U\right\} + \eta\right],$$

proving the first statement.

4. To recover the second statement, we use Assumption 6.5 to provide alternative bounds for the right hand side of (17) and (19).

In particular for (17), recall that there exists $\zeta > 0$, potentially depending on n and λ, such that

$$\zeta \coloneqq \sup_{\delta>0} \sup_{t\in\mathbb{R}} \left[\frac{1}{\delta}\mathbb{P}\left\{|F(Z) - t| \leq \delta \middle| \mathcal{A}\right\}\right] < \infty.$$

Hence for $\delta = \Delta$

$$\sup_{t\in\mathbb{R}} \left[\frac{1}{\Delta}\mathbb{P}\left\{|F(Z) - t| \leq \Delta \middle| \mathcal{A}\right\}\right] < \zeta.$$

In summary

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\left[F\left\{\sqrt{n}(\hat{f} - f_\lambda)\right\} \leq t\right] - \mathbb{P}\left\{F(Z) \leq t\right\}\right| \leq \zeta\Delta + \eta.$$

Since Z' is conditionally Gaussian with covariance Σ, we appeal to Assumption 6.5 to further bound the right hand side of (19) as well. In particular, recall that there exists $\zeta > 0$, potentially depending on n and λ, such that

$$\zeta \coloneqq \sup_{\delta>0} \sup_{t\in\mathbb{R}} \left[\frac{1}{\delta}\mathbb{P}\left\{|F(Z') - t| \leq \delta \middle| U\right\}\right] < \infty.$$

Hence for $\delta = \Delta$

$$\sup_{t\in\mathbb{R}} \left[\frac{1}{\Delta}\mathbb{P}\left\{|F(Z') - t| \leq \Delta \middle| \mathcal{A}\right\}\right] < \zeta.$$

We conclude that, with probability $1 - \eta$,

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\left\{F(\mathfrak{B}) \leq t|U\right\} - \mathbb{P}\left\{F(Z') \leq t|U\right\}\right| \leq \zeta\Delta + \eta.$$

Thus, by the triangle inequality for the sup-norm, since $Z \sim Z'|U$, it holds with probability $1 - \eta$ that

$$\sup_{t\in\mathbb{R}} \left|\mathbb{P}\left[F\left\{\sqrt{n}(\hat{f} - f_\lambda)\right\} \leq t\right] - \mathbb{P}\left\{F(\mathfrak{B}) \leq t \middle| U\right\}\right| \leq 2(\zeta\Delta + \eta),$$

proving the second statement.

□

G.3 Bounding key terms

G.3.1 Covariance operator

To this end, set

$$U_i = T_\lambda^{-1}[(k_{X_i} \otimes k_{X_i}^* - T)(f_0 - f_\lambda) + \varepsilon_i k_{X_i}], \quad \Sigma = \mathbb{E}[U_i \otimes U_i^*].$$

In order to apply our coupling results, we will need to make the following estimates on the covariance of U_i.

Lemma G.7 (Upper bounding the covariance)**.** We have $0 \preceq \Sigma \preceq (\kappa^2 \|f_0\|^2 + \bar{\sigma}^2) T_\lambda^{-2} T$.

Proof. We begin by using bi-linearity of the tensor product to expand

$$\begin{aligned}
\mathbb{E}[U_i \otimes U_i^*] = \mathbb{E}\left[\left\{T_\lambda^{-1}(k_{X_i} \otimes k_{X_i}^* - T)(f_0 - f_\lambda)\right\} \otimes \left\{T_\lambda^{-1}(k_{X_i} \otimes k_{X_i}^* - T)(f_0 - f_\lambda)\right\}^*\right] \\
+ \mathbb{E}\left[\left\{T_\lambda^{-1}(k_{X_i} \otimes k_{X_i}^* - T)(f_0 - f_\lambda)\right\} \otimes \left\{\varepsilon_i T_\lambda^{-1} k_{X_i}\right\}^*\right] \\
+ \mathbb{E}\left[\left\{\varepsilon_i T_\lambda^{-1} k_{X_i}\right\} \otimes \left\{T_\lambda^{-1}(k_{X_i} \otimes k_{X_i}^* - T)(f_0 - f_\lambda)\right\}^*\right] \\
+ \mathbb{E}\left[\left\{\varepsilon_i T_\lambda^{-1} k_{X_i}\right\} \otimes \left\{\varepsilon_i T_\lambda^{-1} k_{X_i}\right\}^*\right].
\end{aligned}$$

Since $\mathbb{E}[\epsilon_i | X_i] = 0$, second and third terms are zero. For the first term, since

$$0 \preceq \mathbb{E}[(f - \mathbb{E}f) \otimes (f - \mathbb{E}f)^*] = \mathbb{E}[f \otimes f^*] - (\mathbb{E}[f] \otimes \mathbb{E}[f])^* \preceq \mathbb{E}[f \otimes f^*],$$

we have

$$\begin{aligned}
0 &\preceq \mathbb{E}\left[\left\{T_\lambda^{-1}(k_{X_i} \otimes k_{X_i}^* - T)(f_0 - f_\lambda)\right\} \otimes \left\{T_\lambda^{-1}(k_{X_i} \otimes k_{X_i}^* - T)(f_0 - f_\lambda)\right\}^*\right] \\
&\preceq \mathbb{E}\left[\left\{T_\lambda^{-1}(k_{X_i} \otimes k_{X_i}^*)(f_0 - f_\lambda)\right\} \otimes \left\{T_\lambda^{-1}(k_{X_i} \otimes k_{X_i}^*)(f_0 - f_\lambda)\right\}^*\right] \\
&= \mathbb{E}\left[(T_\lambda^{-1} k_{X_i}) \otimes \left(T_\lambda^{-1} k_{X_i}\right)^* \{f_0(X_i) - f_\lambda(X_i)\}^2\right].
\end{aligned}$$

Since the fourth term is also clearly positive definite, we have

$$0 \preceq \Sigma \preceq \mathbb{E}\left(\left\{(T_\lambda^{-1} k_{X_i}) \otimes (T_\lambda^{-1} k_{X_i})^*\right\}\left[\{f_0(X_i) - f_\lambda(X_i)\}^2 + \varepsilon_i^2\right]\right).$$

Finally, it suffices to use the almost sure bounds given by $|\varepsilon_i| \leq \bar{\sigma}$ and

$$|f_0(X_i) - f_\lambda(X_i)| = |\langle f_0 - f_\lambda, k_{X_i}\rangle| = \left|\left\langle (I - T_\lambda^{-1} T) f_0, k_{X_i}\right\rangle\right| \leq \kappa \|f_0\|,$$

the latter of which follows by Cauchy-Schwartz and the fact that $\left\|I - T_\lambda^{-1}T\right\|_{op} \leq 1$ since it is a difference of two positive definite operators of norm at most 1. Lastly, we have

$$\mathbb{E}\left\{(T_\lambda^{-1}k_{X_i}) \otimes \left(T_\lambda^{-1}k_{X_i}\right)^*\right\} = \mathbb{E}\{T_\lambda^{-1}(k_{X_i} \otimes k_{X_i}^*)T_\lambda^{-1}\} = T_\lambda^{-1}TT_\lambda^{-1},$$

since $\mathbb{E}[k_{X_i} \otimes k_{X_i}^*] = T$. To arrive at the result, note that T commutes with T_λ^{-1}. □

Lemma G.8 (Lower bounding the covariance)**.** If $\mathbb{E}[\varepsilon_i^2|X_i] \geq \underline{\sigma}^2$ then $\Sigma \succeq \underline{\sigma}^2 T_\lambda^{-2}T$.

Proof. As argued in the proof of Lemma G.7

$$\begin{aligned}
\Sigma &= \Sigma_1 + \Sigma_2 \\
\Sigma_1 &= \mathbb{E}\left[\left\{T_\lambda^{-1}(k_{X_i} \otimes k_{X_i}^* - T)(f_0 - f_\lambda)\right\} \otimes \left\{T_\lambda^{-1}(k_{X_i} \otimes k_{X_i}^* - T)(f_0 - f_\lambda)\right\}^*\right] \succeq 0 \\
\Sigma_2 &= \mathbb{E}\left[\left\{\varepsilon_i T_\lambda^{-1}k_{X_i}\right\} \otimes \left\{\varepsilon_i T_\lambda^{-1}k_{X_i}\right\}^*\right] \succeq 0.
\end{aligned}$$

Hence

$$\begin{aligned}
\Sigma &\succeq \Sigma_2 \\
&= \mathbb{E}\left[\left\{\varepsilon_i T_\lambda^{-1}k_{X_i}\right\} \otimes \left\{\varepsilon_i T_\lambda^{-1}k_{X_i}\right\}^*\right] \\
&\succeq \underline{\sigma}^2\mathbb{E}\left[\left\{T_\lambda^{-1}k_{X_i}\right\} \otimes \left\{T_\lambda^{-1}k_{X_i}\right\}^*\right] \\
&= \underline{\sigma}^2 T_\lambda^{-1}TT_\lambda^{-1} \\
&= \underline{\sigma}^2 T_\lambda^{-2}T.
\end{aligned}$$

□

Lemma G.9 (Spectral ordering)**.** If $A \preceq B$ for some trace-class, self-adjoint operators A and B, then $\sigma^2(A, m) \leq \sigma^2(B, m)$. In particular, taking $m = 0$, we recover $\mathrm{tr}(A) \leq \mathrm{tr}(B)$.

Proof. Recall the definition

$$\sigma^2(A, m) = \sum_{s=m+1}^{\infty} \nu_s(A) = \sum_{s=m+1}^{\infty} \langle e_s(A),\, Ae_s(A)\rangle\,,$$

for a self-adjoint operator A with finite trace, where $\{\nu_1(A), \nu_2(A), \ldots\}$ are the eigenvalues of A listed in decreasing order and $\{e_1(A), e_2(A), \ldots\}$ are the corresponding eigenvectors. Now, let $(f_1, f_2, \ldots)$ denote any other orthonormal basis of H. It follows by the variational representation of the top m eigenvectors that $\sum_{s=1}^{m} \nu_s(A) \geq \sum_{s=1}^{m} \langle f_s,\, Af_s \rangle$. Thus, if $A \preceq B$ for some trace-class, self-adjoint operator B, and taking $f_s = e_s(B)$,

$$\begin{aligned}
\sigma^2(A, m) &= \sum_{s=m+1}^{\infty} \nu_s(A) \\
&= \operatorname{tr}(A) - \sum_{s=1}^{m} \nu_s(A) \\
&\leq \operatorname{tr}(A) - \sum_{s=1}^{m} \langle e_s(B),\, Ae_s(B) \rangle \\
&= \sum_{s=m+1}^{\infty} \langle e_s(B),\, Ae_s(B) \rangle \,,
\end{aligned}$$

since the trace is independent of the choice of orthonormal basis. Since $A \preceq B$, this is

$$\leq \sum_{s=m+1}^{\infty} \langle e_s(B),\, Be_s(B) \rangle = \sigma^2(B, m).$$

□

Lemma G.10 (Local width bounds)**.** In our setting we have the bounds

$$\sigma(\Sigma, m) \leq \left(\frac{\kappa \, \|f_0\| + \bar{\sigma}}{\lambda} \right) \sigma(T, m), \quad \sigma(\Sigma, 0) \leq (\kappa \, \|f_0\| + \bar{\sigma}) \sqrt{\mathfrak{n}(\lambda)}.$$

Proof. We appeal to Lemmas G.7 and G.9. Recall the bound on $\Sigma = \mathbb{E}[U_i \otimes U_i^*]$ is

$$0 \preceq \Sigma \preceq (\kappa \, \|f_0\| + \bar{\sigma})^2 T_\lambda^{-2} T \preceq \left(\frac{\kappa \, \|f_0\| + \bar{\sigma}}{\lambda} \right)^2 T.$$

Then, since $\mathfrak{n}(\lambda) \coloneqq \operatorname{tr}(T_\lambda^{-2} T)$, we have

$$\sigma^2(\Sigma, 0) = \operatorname{tr}(\Sigma) \leq (\kappa \, \|f_0\| + \bar{\sigma})^2 \operatorname{tr}(T_\lambda^{-2} T) = (\kappa \, \|f_0\| + \bar{\sigma})^2 \mathfrak{n}(\lambda)$$

and

$$\sigma^2(\Sigma, m) \leq \left(\frac{\kappa \, \|f_0\| + \bar{\sigma}}{\lambda} \right)^2 \sigma^2(T, m).$$

Taking square roots recovers the second and third stated inequalities. □

G.3.2 Summands

Lemma G.11 (Bounded summands)**.** We may appeal to the bounded summand results using $\|U_i\| \leq \left(\frac{\kappa^2\|f_0\|+\kappa\bar{\sigma}}{\lambda}\right)$.

Proof. Write

$$\begin{aligned}\|U_i\| &= \left\|T_\lambda^{-1}[(k_{X_i}\otimes k_{X_i}^* - T)(f_0 - f_\lambda) + \varepsilon_i k_{X_i}]\right\| \\ &\leq \left\|T_\lambda^{-1}\right\|_{op}\left\{\left\|k_{X_i}\otimes k_{X_i}^* - T\right\|_{op}\|f_0 - f_\lambda\| + \|\varepsilon_i k_{X_i}\|\right\} \\ &\leq \frac{1}{\lambda}\left\{\left\|k_{X_i}\otimes k_{X_i}^* - T\right\|_{op}\|f_0 - f_\lambda\| + \bar{\sigma}\kappa\right\}.\end{aligned}$$

Note that $k_{X_i}\otimes k_{X_i}^* - T$ is a difference of two positive definite operators, so $\left\|k_{X_i}\otimes k_{X_i}^* - T\right\|_{op} \leq \left\|k_{X_i}\otimes k_{X_i}^*\right\|_{op} \vee \|T\|_{op}$. Further, $\|T\|_{op} = \left\|\mathbb{E}[k_{X_i}\otimes k_{X_i}^*]\right\|_{op} \leq \mathbb{E}\left\|k_{X_i}\otimes k_{X_i}^*\right\|_{op}$ by Jensen's inequality and $\left\|k_{X_i}\otimes k_{X_i}^*\right\|_{op} \leq \kappa^2$ almost surely, so the above is

$$\leq \frac{1}{\lambda}\left\{\kappa^2\|f_0 - f_\lambda\| + \bar{\sigma}\kappa\right\}.$$

Finally, $f_0 - f_\lambda = (I - T_\lambda^{-1}T)f_0$ and $(I - T_\lambda^{-1}T)$ is a difference of two positive definite operators of norm at most 1, so the above is

$$\begin{aligned}&\leq \frac{1}{\lambda}\left\{\kappa^2\left\|I - T_\lambda^{-1}T\right\|_{op}\|f_0\| + \bar{\sigma}\kappa\right\} \\ &\leq \frac{1}{\lambda}\left\{\kappa^2\|f_0 - f_\lambda\| + \bar{\sigma}\kappa\right\}.\end{aligned}$$

□

Lemma G.12 (Sub-Gaussian summands)**.** Suppose k_{X_i} is sub-Gaussian with parameter b and $\mathbb{E}[\varepsilon_i^2|X_i] \geq \underline{\sigma}^2$. Then U_i is sub-Gaussian with parameter $\frac{(\kappa\|f_0\|+\bar{\sigma})}{\underline{\sigma}}b$.

Proof. We proceed in steps.

1. Equivalent expressions. By hypothesis, for all $t \in H$, k_{X_i} satisfies

$$\mathbb{P}\left(\langle t,\, k_{X_i}\rangle > ub\left(\mathbb{E}\langle t,\, k_{X_i}\rangle^2\right)^{\frac{1}{2}}\right) \leq 2e^{-u^2}.$$

As argued in Appendix J, this definition implies $\langle t,\, k_{X_i}\rangle$ is sub-Gaussian with $\|\langle k_{X_i},\, t\rangle\|_{\psi_2} \le Cb\,\langle t,\, Tt\rangle^{\frac{1}{2}}$. To see why, notice that in our sub-Gaussian assumption, $b\left(\mathbb{E}\,\langle t,\, k_{X_i}\rangle^2\right)^{\frac{1}{2}} = b\,\langle t,\, Tt\rangle^{\frac{1}{2}}$. Therefore we are effectively assuming

$$\left\|\frac{\langle t,\, k_{X_i}\rangle}{b\,\langle t,\, Tt\rangle^{\frac{1}{2}}}\right\|_{\psi_2} \le C \iff \|\langle t,\, k_{X_i}\rangle\|_{\psi_2} \le Cb\,\langle t,\, Tt\rangle^{\frac{1}{2}}.$$

2. Comparison. Recall that $U_i = \{f_\lambda(X_i) - f_0(X_i) + \varepsilon_i\}T_\lambda^{-1}k_{X_i}$. Thus, we have

$$\begin{aligned}
|\langle U_i,\, t\rangle| &\le |f_\lambda(X_i) - f_0(X_i) + \varepsilon_i| \cdot |\left\langle T_\lambda^{-1}k_{X_i},\, t\right\rangle| \\
&\le (\kappa\,\|f_0\| + \bar{\sigma})|\left\langle T_\lambda^{-1}k_{X_i},\, t\right\rangle| \\
&= (\kappa\,\|f_0\| + \bar{\sigma})|\left\langle k_{X_i},\, T_\lambda^{-1}t\right\rangle|.
\end{aligned}$$

Hence for any $u \ge (\kappa\,\|f_0\| + \bar{\sigma})\left\|\left\langle k_{X_i},\, T_\lambda^{-1}t\right\rangle\right\|_{\psi_2}$ we have

$$\mathbb{E}\{\exp(\langle U_i,\, t\rangle^2/u^2)\} \le \mathbb{E}\left\{\exp\left(\left\langle k_{X_i},\, T_\lambda^{-1}t\right\rangle^2 / \left\|\left\langle k_{X_i},\, T_\lambda^{-1}t\right\rangle\right\|_{\psi_2}^2\right)\right\} \le 2,$$

so we conclude that $\|\langle U_i,\, t\rangle\|_{\psi_2} \le (\kappa\,\|f_0\| + \bar{\sigma})\left\|\left\langle k_{X_i},\, T_\lambda^{-1}t\right\rangle\right\|_{\psi_2}$ by definition of ψ_2 norm as the smallest u such that the above holds.

3. Lower bound. In summary, we have shown

$$\|\langle U_i,\, t\rangle\|_{\psi_2} \le (\kappa\,\|f_0\| + \bar{\sigma})\left\|\left\langle k_{X_i},\, T_\lambda^{-1}t\right\rangle\right\|_{\psi_2} \le C(\kappa\,\|f_0\| + \bar{\sigma})b\left\langle t,\, T_\lambda^{-1}TT_\lambda^{-1}t\right\rangle^{\frac{1}{2}}.$$

We wish to show there exists some $\tilde{b}$ such that $\|\langle U_i,\, t\rangle\|_{\psi_2} \le C\tilde{b}\,\langle t,\, \Sigma t\rangle^{\frac{1}{2}}$. It would suffice to argue $T_\lambda^{-1}TT_\lambda^{-1} \preceq \Sigma$. By Lemma G.8, $T_\lambda^{-1}TT_\lambda^{-1} \preceq \frac{1}{\underline{\sigma}^2}\Sigma$ and hence

$$\|\langle U_i,\, t\rangle\|_{\psi_2} \le C(\kappa\,\|f_0\| + \bar{\sigma})\frac{1}{\underline{\sigma}}b\,\langle t,\, \Sigma t\rangle^{\frac{1}{2}}$$

i.e. $\tilde{b} = \frac{(\kappa\|f_0\| + \bar{\sigma})}{\underline{\sigma}}b$.

□

G.4 Bounding (Q,R,L,B)

G.4.1 Bounding L

First, we lower bound the size of the Gaussian process Z (with covariance Σ) that approximates the deviations of the estimator.

Lemma G.13 (Variance lower bound)**.** Let Z be a Gaussian random element of H with covariance Σ, and suppose $\mathbb{E}[\varepsilon_i^2|X_i] \geq \underline{\sigma}^2$ almost surely. Put $M \coloneqq \kappa^2 \|f_0\|^2 + \bar{\sigma}^2$. Then with probability at least $1-\eta$,

$$\|Z\| \geq L(\lambda, \eta) := \sqrt{\underline{\sigma}^2 \mathfrak{n}(\lambda)} - \left\{2 + \sqrt{2\log(1/\eta)}\right\}\sqrt{M/\lambda}.$$

Proof of Lemma G.13. We will lower bound $\mathbb{E}\|Z\|$ using the identity,

$$\{\mathbb{E}(\|Z\|)\}^2 = \mathbb{E}(\|Z\|^2) - \mathbb{E}\{\|Z\| - \mathbb{E}(\|Z\|)\}^2, \tag{20}$$

by bounding both terms on the right-hand side. Finally we will apply Gaussian concentration for a high-probability bound.

1. Upper bounding $\mathbb{E}(\|Z\| - \mathbb{E}\|Z\|)^2$. Firstly, recall that $\|Z\|$ is the supremum of the Gaussian process $\langle Z, t\rangle$ indexed by $T = \{t \in H \mid \|t\| \leq 1\}$. Moreover, we have $\sigma_T^2 \coloneqq \sup_{t\in T} \mathbb{E}\langle Z, t\rangle^2 = \|\Sigma\|_{op}$.

   By Lemma G.7, $\Sigma \preceq M T_\lambda^{-2} T$, it follows that

   $$\sigma_T^2 = \|\Sigma\|_{op} \leq M \left\|T_\lambda^{-2} T\right\|_{op} \leq M/\lambda,$$

   where the bound $\|T_\lambda^{-2}T\|_{op} \leq 1/\lambda$ follows by maximizing $s \mapsto (s+\lambda)^{-2}s$. Similarly,

   $$\mathbb{E}\|Z\|^2 = \operatorname{tr}\Sigma \leq M \operatorname{tr}(T_\lambda^{-2}T) = M\mathfrak{n}(\lambda),$$

   so by Markov's inequality $\langle Z, t\rangle$ is a.s. bounded on T. Thus, by combining the two inequalities of Lemma D.4 with a union bound we have

   $$\mathbb{P}\left\{(\|Z\| - \mathbb{E}\|Z\|)^2 \geq u\right\} = \mathbb{P}\left\{|\|Z\| - \mathbb{E}\|Z\|| \geq \sqrt{u}\right\} \leq 2\exp\left(-\frac{u}{2\|\Sigma\|_{op}}\right).$$

By integrating the tail,

$$\begin{aligned}\mathbb{E}(\|Z\| - \mathbb{E}\,\|Z\|)^2 &= \int_0^\infty \mathbb{P}\left\{(\|Z\| - \mathbb{E}\,\|Z\|)^2 \geq u\right\} du \\ &\leq \int_0^\infty 2\exp\left(-\frac{u}{2\|\Sigma\|_{op}}\right) du \\ &= 4\|\Sigma\|_{op} \\ &\leq 4M/\lambda.\end{aligned}$$

2. Lower bounding $\mathbb{E}(\|Z\|^2)$. Similarly using the upper bound $\Sigma \succeq \underline{\sigma}^2 T_\lambda^{-2} T$ from Lemma G.8, we find that

$$\mathbb{E}\|Z\|^2 = \operatorname{tr}\Sigma \geq \underline{\sigma}^2 \operatorname{tr}(T_\lambda^{-2} T) = \underline{\sigma}^2 \mathfrak{n}(\lambda).$$

3. Combining. Combining these estimates in (20), we have

$$\mathbb{E}\|Z\| \geq \sqrt{\underline{\sigma}^2 \mathfrak{n}(\lambda) - 4M/\lambda}.$$

   By a final application of the lower tail bound in Lemma D.4 (in inverted form) we have that w.p. $1-\eta$,

$$\|Z\| \geq \mathbb{E}\|Z\| - \sqrt{2\|\Sigma\|_{op}\log(1/\eta)} \geq \sqrt{\underline{\sigma}^2 \mathfrak{n}(\lambda) - 4M/\lambda} - \sqrt{2M\log(1/\eta)/\lambda}.$$

   Since $\sqrt{a-b} \geq \sqrt{a} - \sqrt{b}$ (rearrange and square both sides), this simplifies to

$$\|Z\| \geq \sqrt{\underline{\sigma}^2 \mathfrak{n}(\lambda)} - \left\{2 + \sqrt{2\log(1/\eta)}\right\}\sqrt{M/\lambda}.$$

□

G.4.2 Q and R for bounded summands

In what follows, define $M := \kappa^2\|f_0\| + \bar{\sigma}\kappa$.

Proposition G.14 (Gaussian approximation: Bounded)**.** Suppose Assumption 6.10 holds. Then there exists a sequence $(Z_i)_{1\leq i\leq n}$ of Gaussian random elements in H, with covariance Σ, such that with probability $1-\eta$,

$$\left\|\sqrt{n}(\hat{f} - f_\lambda) - \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i\right\| \lesssim Q_{\mathrm{bd}}(T,n,\lambda) M \log(12/\eta) + \frac{\mathfrak{n}(\lambda)}{\sqrt{n}} M \log(8/\eta)^2,$$

where

$$Q_{\mathrm{bd}}(T,n,\lambda) = \frac{1}{\lambda}\inf_{m\geq 1}\left\{\sigma(T,m) + \frac{m^2\log(m^2)}{\sqrt{n}}\right\}.$$

Proof. By Theorem B.4, with probability at least $1-\eta$,

$$\left\|\sqrt{n}(\hat{f}-f_\lambda) - \frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\right\| \lesssim \frac{\mathfrak{n}(\lambda)}{\sqrt{n}} M\log(4/\eta)^2$$

since

$$\frac{2\kappa}{n\lambda} \leq \sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \iff n \geq \frac{4\kappa^2}{\mathfrak{n}(\lambda)\lambda^2}.$$

Then, using Proposition C.7 together with the bounds deduced in Lemmas G.10 and G.11, we deduce that

$$\begin{aligned}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n (U_i - Z_i)\right\| &\lesssim \inf_{m\geq 1}\left\{\sqrt{\log(6/\eta)}\sigma(\Sigma,m) + \frac{am^2\log(m^2/\eta)}{\sqrt{n}}\right\} \\ &\lesssim \inf_{m\geq 1}\left\{\frac{M\sigma(T,m)}{\lambda}\sqrt{\log(6/\eta)} + m^2\frac{M\log(m^2/\eta)}{\lambda\sqrt{n}}\right\}\end{aligned}$$

holds with probability at least $1-\eta$. Thus, by a union bound, it holds with probability $1-\eta$ that

$$\left\|\sqrt{n}(\hat{f}-f_\lambda) - \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i\right\| \lesssim Q_{\mathrm{bd}}(T,n,\lambda)M\log(12/\eta) + \frac{\mathfrak{n}(\lambda)}{\sqrt{n}}M\log(8/\eta)^2,$$

where we have simplified and consolidated log factors. □

Recall the bootstrap process

$$\mathfrak{B} = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n h_{ij}\left(\frac{\hat{V}_i - \hat{V}_j}{\sqrt{2}}\right), \quad \hat{V}_i = \{Y_i - \hat{f}(X_i)\}\hat{T}_\lambda^{-1}k_{X_i}.$$

Proposition G.15 (Bootstrap approximation: Bounded)**.** Suppose Assumption 6.10 holds. Then, there exists a random variable Z whose conditional distribution given U is Gaussian with covariance Σ, and such that with probability at least $1-\eta$

$$\mathbb{P}\left[\|\mathfrak{B}-Z\| \lesssim M\log(16/\eta)^2\left\{R_{\mathrm{bd}}(T,n,\lambda) + \frac{\kappa\mathfrak{n}(\lambda)}{\sqrt{n}}\right\}\middle| U\right] \geq 1-\eta$$

where

$$R_{\mathrm{bd}}(T,n,\lambda) \coloneqq \inf_{m\geq 1}\left\{\left(\frac{m\mathfrak{n}(\lambda)}{\lambda^2 n} + \frac{m}{\lambda^4 n^2}\right)^{\frac{1}{4}} + \frac{1}{\lambda}\sigma(T,m)\right\}.$$

Proof. Under Assumption 6.10, Theorem E.9 implies that

$$\left\| \mathfrak{B} - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij} \frac{U_i - U_j}{\sqrt{2}} \right\| \lesssim \frac{\mathfrak{n}(\lambda)}{\sqrt{n}} \kappa M \ln(16/\eta)^2$$

holds with probability at least $1 - \eta$ since $n \geq \frac{4\kappa^2}{\mathfrak{n}(\lambda)\lambda^2}$ implies

$$\frac{2\kappa}{n\lambda} \leq \sqrt{\frac{\mathfrak{n}(\lambda)}{n}}, \quad \sqrt{\mathfrak{n}(\lambda)} \geq \frac{1}{\sqrt{n}\lambda}, \quad \sqrt{\mathfrak{n}(\lambda)} \geq \frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}}.$$

We use Corollary D.15 along with the bounds in Lemmas G.10 and G.11. In particular, set $W = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij} \frac{U_i - U_j}{\sqrt{2}}$, $W' = \mathfrak{B}$, and $\delta = \frac{\mathfrak{n}(\lambda)}{\sqrt{n}} \kappa M \ln(16/\eta)^2$. Then there must exist Z with the desired conditional distribution, such that with probability at least $1 - \eta$, the $\sigma(U)$-conditional probability of the event

$$\begin{aligned}
\|Z - \mathfrak{B}\| &\lesssim \log(6/\eta)^{3/2} \inf_{m \geq 1} \left\{ m^{\frac{1}{4}} \left(\frac{M^2 \cdot M^2 \mathfrak{n}(\lambda)}{\lambda^2 n} + \frac{M^4}{\lambda^4 n^2} \right)^{\frac{1}{4}} \right. \\
&\qquad \left. + \frac{M}{\lambda} \sigma(T, m) \right\} + \frac{\mathfrak{n}(\lambda)}{\sqrt{n}} \kappa M \log(16/\eta)^2 \\
&\lesssim M \log(16/\eta)^2 \left(\inf_{m \geq 1} \left\{ m^{\frac{1}{4}} \left(\frac{\mathfrak{n}(\lambda)}{\lambda^2 n} + \frac{1}{\lambda^4 n^2} \right)^{\frac{1}{4}} + \frac{\sigma(T, m)}{\lambda} \right\} + \frac{\kappa \mathfrak{n}(\lambda)}{\sqrt{n}} \right) \\
&= M \log(16/\eta)^2 \left(R_{\mathrm{bd}}(T, n, \lambda) + \frac{\kappa \mathfrak{n}(\lambda)}{\sqrt{n}} \right)
\end{aligned}$$

is at least $1 - \eta$ when $n \geq 2$. □

G.4.3 Q and R for sub-Gaussian summands

In what follows, define $\tilde{M} := \frac{1}{\underline{\sigma}}(\kappa \|f_0\| + \bar{\sigma})$ and $\bar{M} = M \vee \tilde{M}$.

Proposition G.16 (Gaussian approximation: Sub-Gaussian)**.** Suppose Assumptions 6.10 and 6.12 hold. Then there exists a Gaussian random element in H, with covariance Σ, such that with probability $1 - \eta$,

$$\left\| \sqrt{n}(\hat{f} - f_\lambda) - Z \right\| \lesssim Q_{\mathrm{sg}}(T, n, \lambda) \bar{M}^4 b^3 (6/\eta)^{1/\log(mn)} + \frac{\mathfrak{n}(\lambda)}{\sqrt{n}} M \log(8/\eta)^2,$$

where

$$Q_{\mathrm{sg}}(T, n, \lambda) = \inf_{m \geq 1} \left\{ \frac{\sigma(T, m)}{\lambda} + \frac{m^{3/2} \log(n^2)}{\sqrt{n\lambda}} \right\}.$$

Proof. By Theorem B.4, with probability at least $1-\eta$,

$$\left\|\sqrt{n}(\hat{f}-f_\lambda)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i\right\| \lesssim \frac{\mathfrak{n}(\lambda)}{\sqrt{n}}M\log(4/\eta)^2$$

since

$$\frac{2\kappa}{n\lambda}\leq\sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \iff n\geq\frac{4\kappa^2}{\mathfrak{n}(\lambda)\lambda^2}.$$

Then, using Proposition C.8 together with the bounds deduced in Lemmas G.10 and G.12, we deduce that

$$\begin{aligned}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i-Z\right\| &\lesssim \inf_{m\geq 1}\left\{\tilde{b}\sigma(\Sigma,m)\sqrt{\log(6/\eta)}+\|\Sigma\|_{op}^{\frac{1}{2}}\tilde{b}^3\left(\frac{m^{\frac{3}{2}}\log(n)^2}{\sqrt{n}}\right)(3/\eta)^{1/\log(mn)}\right\}\\ &\lesssim \inf_{m\geq 1}\left\{\tilde{M}b\frac{M\sigma(T,m)}{\lambda}\sqrt{\log(6/\eta)}+\frac{M}{\sqrt{\lambda}}\tilde{M}^3b^3\left(\frac{m^{\frac{3}{2}}\log(n)^2}{\sqrt{n}}\right)(3/\eta)^{1/\log(mn)}\right\}\end{aligned}$$

holds with probability at least $1-\eta$. Note that we use the bound

$$\|\Sigma\|_{op}\leq M^2\|T_\lambda^{-2}T\|_{op}\leq M^2\frac{1}{\lambda}$$

which is justified in the proof of Proposition I.4 with $s=1$ and $c=2$.

Thus, by a union bound, it holds with probability $1-\eta$ that

$$\left\|\sqrt{n}(\hat{f}-f_\lambda)-Z\right\| \lesssim Q_{\mathrm{bd}}(T,n,\lambda)\bar{M}^4b^3(6/\eta)^{1/\log(mn)}+\frac{\mathfrak{n}(\lambda)}{\sqrt{n}}M\log(8/\eta)^2,$$

where we have simplified and consolidated log factors. □

Proposition G.17 (Bootstrap approximation: Sub-Gaussian)**.** Suppose Assumptions 6.10 and 6.12 hold. Then, there exists a random variable Z whose conditional distribution given U is Gaussian with covariance Σ, and such that with probability at least $1-\eta$

$$\mathbb{P}\left[\|\mathfrak{B}-Z\|\lesssim\bar{M}^2b\cdot\log(16/\eta)^2\left\{R_{\mathrm{sg}}(T,n,\lambda)+\frac{\mathfrak{n}(\lambda)}{\sqrt{n}}\right\}\middle|U\right]\geq 1-\eta$$

where

$$R_{\mathrm{sg}}(T,n,\lambda)\coloneqq\inf_{m\geq 1}\left\{\left(\frac{m\mathfrak{n}(\lambda)^2}{n}\right)^{\frac{1}{4}}+\frac{1}{\lambda}\sigma(T,m)\right\}.$$

Proof. Under Assumption 6.10, Theorem E.9 implies that

$$\left\| \mathfrak{B} - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij} \frac{U_i - U_j}{\sqrt{2}} \right\| \lesssim \frac{\mathfrak{n}(\lambda)}{\sqrt{n}} \kappa M \ln(16/\eta)^2$$

holds with probability at least $1 - \eta$ since $n \geq \frac{4\kappa^2}{\mathfrak{n}(\lambda)\lambda^2}$ implies

$$\frac{2\kappa}{n\lambda} \leq \sqrt{\frac{\mathfrak{n}(\lambda)}{n}}, \quad \sqrt{\mathfrak{n}(\lambda)} \geq \frac{1}{\sqrt{n}\lambda}, \quad \sqrt{\mathfrak{n}(\lambda)} \geq \frac{\mathfrak{n}(\lambda)^{1/4}}{n^{1/4}\lambda^{1/2}}.$$

We use Corollary D.15 along with the bounds in Lemmas G.10 and G.12. In particular, set $W = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij} \frac{U_i - U_j}{\sqrt{2}}$, $W' = \mathfrak{B}$, and $\delta = \frac{\mathfrak{n}(\lambda)}{\sqrt{n}} \kappa M \ln(16/\eta)^2$. Then there must exist Z with the desired conditional distribution, such that with probability at least $1 - \eta$, the $\sigma(U)$-conditional probability of the event

$$\begin{aligned}
\|Z - \mathfrak{B}\| &\lesssim \log(6/\eta)^{3/2} \inf_{m \geq 1} \left\{ \left(\frac{m}{n} \tilde{M}^4 b^4 M^4 \mathfrak{n}(\lambda)^2 \right)^{\frac{1}{4}} + \frac{M}{\lambda} \sigma(T, m) \right\} + \frac{\mathfrak{n}(\lambda)}{\sqrt{n}} \kappa M \log(16/\eta)^2 \\
&\lesssim \bar{M}^2 b \cdot \log(16/\eta)^2 \left(\inf_{m \geq 1} \left\{ \left(\frac{m \mathfrak{n}(\lambda)^2}{n} \right)^{\frac{1}{4}} + \frac{\sigma(T, m)}{\lambda} \right\} + \frac{\mathfrak{n}(\lambda)}{\sqrt{n}} \right) \\
&= \bar{M}^2 b \cdot \log(16/\eta)^2 \left(R_{\mathrm{sg}}(T, n, \lambda) + \frac{\mathfrak{n}(\lambda)}{\sqrt{n}} \right)
\end{aligned}$$

is at least $1 - \eta$. □

G.5 Corollaries

Recall the definitions of quantities appearing in Theorem G.6:

$$\begin{aligned}
Q_{\mathrm{bd}}(n, \lambda) &= \frac{1}{\lambda} \inf_{m \geq 1} \left\{ \sigma(m) + \frac{m^2 \log(m^2)}{\sqrt{n}} \right\}, & Q_{\mathrm{sg}}(n, \lambda) &= \inf_{m \geq 1} \left\{ \frac{\sigma(m)}{\lambda} + \frac{m^{\frac{3}{2}} \log(n)^2}{\sqrt{n\lambda}} \right\}, \\
R_{\mathrm{bd}}(n, \lambda) &= \inf_{m \geq 1} \left[\left\{ \frac{m\, \mathfrak{n}(\lambda)}{\lambda^2 n} + \frac{m}{\lambda^4 n^2} \right\}^{\frac{1}{4}} + \frac{\sigma(m)}{\lambda} \right], & R_{\mathrm{sg}}(n, \lambda) &= \inf_{m \geq 1} \left[\left\{ \frac{m\, \mathfrak{n}^2(\lambda)}{n} \right\}^{\frac{1}{4}} + \frac{\sigma(m)}{\lambda} \right].
\end{aligned}$$

We will now give a simplified asymptotic characterization for each of these expressions under two different assumptions on the spectrum of T, namely (i) polynomial decay,

i.e. $\nu_s(T) \leq \omega s^{-\beta}$; (ii) exponential decay, i.e. $\nu_s(T) \leq \omega \exp(-\alpha s^\gamma)$. We suppress dependence on T in the notation. Note that, within R_{bd},

$$\frac{m\mathfrak{n}(\lambda)}{\lambda^2 n} \geq \frac{m}{\lambda^4 n^2} \iff n \geq \frac{1}{\lambda^2 \mathfrak{n}(\lambda)}$$

which is implied by Assumption 6.10.

Sobolev RKHS: Bounded data

We proceed in steps.

1. If $\nu_s \leq \omega s^{-\beta}$ then by Proposition I.2 we have $\sigma(m) \lesssim_{\beta,\omega} m^{1/2-\beta/2}$, and by Proposition I.4 we have $\mathfrak{n}(\lambda) \lesssim_{\beta,\omega} \lambda^{-1-1/\beta}$.

2. First we study $Q_{\text{bd}}(n, \lambda)$. Plugging in our bound on $\sigma(m)$ and then equating the main terms (ignoring constants and log factors) gives

$$m^{1/2-\beta/2} = \frac{m^2}{\sqrt{n}} \iff m = n^{1/(3+\beta)}.$$

   Plugging this in and simplifying exponents in the logarithm gives

$$\begin{aligned} Q_{\text{bd}}(n, \lambda) &\lesssim_{\beta,\omega} \lambda^{-1} \cdot (n^{\frac{1}{3+\beta}})^{\frac{1-\beta}{2}} \cdot \log(n) \\ &= \lambda^{-1} n^{(1-\beta)/(6+2\beta)} \log(n) \\ &= \log(n) \left(\frac{1}{n \lambda^{\frac{6+2\beta}{\beta-1}}} \right)^{(\beta-1)/(6+2\beta)}. \end{aligned}$$

3. Next we study $R_{\text{bd}}(n, \lambda)$. Plugging in our bounds and equating main terms gives

$$\begin{aligned} m^{\frac{1}{4}} \left\{ \frac{\lambda^{-1-1/\beta}}{\lambda^2 n} \right\}^{\frac{1}{4}} &= \frac{m^{1/2-\beta/2}}{\lambda} \\ \iff \left\{ \frac{\lambda^{1-1/\beta}}{n} \right\}^{\frac{1}{1-2\beta}} &= m. \end{aligned}$$

Using this value of m gives the bound

$$\begin{aligned} R_{\mathrm{bd}}(n,\lambda) &\lesssim_{\beta,\omega} \left\{\frac{m^{1/2-\beta/2}}{\lambda}\right\} \\ &= \frac{1}{\lambda}\left\{\frac{\lambda^{1-1/\beta}}{n}\right\}^{\frac{1-\beta}{2-4\beta}} \\ &= \left\{\frac{1}{\lambda^{3+1/\beta+2/(\beta-1)}n}\right\}^{\frac{1-\beta}{2-4\beta}}. \end{aligned}$$

$$\left(\frac{\lambda^{\frac{6+2\beta}{\beta-1}}}{\log(n)} \wedge \lambda^{3+1/\beta+2/(\beta-1)}\right) n \gg 1.$$

Sobolev RKHS: Sub-Gaussian data

We proceed in steps.

1. Once again, if $\nu_s \leq \omega s^{-\beta}$ then by Proposition I.2 we have $\sigma(m) \lesssim_{\beta,\omega} m^{1/2-\beta/2}$, and by Proposition I.4 we have $\mathfrak{n}(\lambda) \lesssim_{\beta,\omega} \lambda^{-1-1/\beta}$.

2. For $Q_{\mathrm{sg}}(n,\lambda)$, equating main terms gives

$$\frac{m^{1/2-\beta/2}}{\lambda} = \frac{m^{\frac{3}{2}}}{\sqrt{\lambda n}} \iff m^{-1-\beta/2} = \sqrt{\frac{\lambda}{n}} \iff m = (n/\lambda)^{1/(2+\beta)}.$$

   Plugging this in gives

$$\begin{aligned} Q_{\mathrm{sg}}(n,\lambda) &\lesssim_{\beta,\omega} \log^2(n) n^{3/(4+2\beta)-1/2} \lambda^{-3/(4+2\beta)-1/2} \\ &= \log^2(n) n^{\frac{1-\beta}{4+2\beta}} \lambda^{\frac{-(5+\beta)}{4+2\beta}} \\ &= \log^2(n)\left(\frac{1}{n\lambda^{\frac{5+\beta}{\beta-1}}}\right)^{\frac{\beta-1}{4+2\beta}}. \end{aligned}$$

3. For $R_{\mathrm{sg}}(n,\lambda)$, equating main terms gives

$$m^{\frac{1}{4}}\left\{\frac{\lambda^{-2-2/\beta}}{n}\right\}^{\frac{1}{4}} = \frac{m^{1/2-\beta/2}}{\lambda} \iff \left\{\frac{\lambda^{2-2/\beta}}{n}\right\}^{\frac{1}{1-2\beta}} = m.$$

Using this value of m gives the bound

$$\begin{aligned} R_{\mathrm{sg}}(n,\lambda) &\lesssim_{\beta,\omega} \left\{\frac{m^{1/2-\beta/2}}{\lambda}\right\} \\ &= \frac{1}{\lambda}\left\{\frac{\lambda^{2-2/\beta}}{n}\right\}^{\frac{1-\beta}{2-4\beta}} \\ &= \left\{\frac{1}{\lambda^{2+2/\beta+2/(\beta-1)}n}\right\}^{\frac{1-\beta}{2-4\beta}}. \end{aligned}$$

Gaussian RKHS: Bounded data

We proceed in steps.

1. If $\nu_s \leq \omega \exp(-\alpha m^\gamma)$ then by Proposition I.2 we have $\sigma(m) \lesssim_{\omega,\alpha,\gamma} m^{1/2-\gamma/2}\exp(-\alpha m^\gamma/2)$, and by Proposition I.4 we have $\mathfrak{n}(\lambda) \lesssim_{\omega,\alpha,\gamma} \lambda^{-1}\log(1/\lambda)^{1/\gamma}$.

2. First we study $Q_{\mathrm{bd}}(n,\lambda)$. Plugging in our bound on $\sigma(m)$ and then equating the main terms (ignoring constants and log factors) gives

$$\exp(-\alpha m^\gamma/2) = \frac{1}{\sqrt{n}} \iff m = \left(\frac{\log(n)}{\alpha}\right)^{\frac{1}{\gamma}}.$$

$$\begin{aligned} \sigma(m) &\lesssim_{\omega,\alpha,\gamma} m^{\frac{1-\gamma}{2}}\exp(-\alpha m^\gamma/2) \\ &= \left(\frac{\log(n)}{\alpha}\right)^{\frac{1-\gamma}{2\gamma}} \exp\left(-\alpha\frac{\log(n)}{\alpha}/2\right) \\ &= \left(\frac{\log(n)}{\alpha}\right)^{\frac{1-\gamma}{2\gamma}} \exp\left(\log(n^{-1/2})\right) \\ &= \left(\frac{\log(n)}{\alpha}\right)^{\frac{1-\gamma}{2\gamma}} n^{-1/2} \\ &\lesssim_{\omega,\alpha,\gamma} \log(n)^{\frac{1-\gamma}{2\gamma}} n^{-1/2}. \end{aligned}$$

Moreover,

$$\frac{m^2\log(m^2)}{\sqrt{n}} = n^{-1/2}\left(\frac{\log(n)}{\alpha}\right)^{\frac{2}{\gamma}} \log\left\{\left(\frac{\log(n)}{\alpha}\right)^{\frac{2}{\gamma}}\right\} \lesssim_{\omega,\gamma,\alpha} n^{-1/2}\log(n)^{\frac{2}{\gamma}}\log(\log(n)).$$

Therefore, since $\gamma > 0$ implies $\frac{2}{\gamma} > \frac{1-\gamma}{2\gamma}$

$$Q_{\mathrm{bd}}(n,\lambda) \lesssim_{\omega,\alpha,\gamma} \frac{1}{\lambda\sqrt{n}}\log(n)^{\frac{2}{\gamma}}\log(\log(n)).$$

3. Next we study $R_{\mathrm{bd}}(n,\lambda)$. Plugging in our bounds and equating main terms gives

$$\begin{aligned}
\left\{\frac{\lambda^{-1}\log(1/\lambda)^{1/\gamma}}{\lambda^2 n}\right\}^{\frac{1}{4}} &= \frac{\exp(-\alpha m^\gamma/2)}{\lambda} \\
\iff \left\{\frac{\lambda\log(1/\lambda)^{1/\gamma}}{n}\right\}^{\frac{1}{4}} &= \exp(-\alpha m^\gamma/2) \\
\iff \left\{\frac{1}{2\alpha}\log\left[\frac{n}{\lambda\log(1/\lambda)^{1/\gamma}}\right]\right\}^{\frac{1}{\gamma}} &= m.
\end{aligned}$$

Using this value of m,

$$\begin{aligned}
\frac{\sigma(m)}{\lambda} &\lesssim_{\omega,\alpha,\gamma} \frac{1}{\lambda} m^{\frac{1-\gamma}{2}}\exp(-\alpha m^\gamma/2) \\
&= \frac{1}{\lambda}\left\{\frac{1}{2\alpha}\log\left[\frac{n}{\lambda\log(1/\lambda)^{1/\gamma}}\right]\right\}^{\frac{1-\gamma}{2\gamma}} \exp\left(-\alpha\frac{1}{2\alpha}\log\left[\frac{n}{\lambda\log(1/\lambda)^{1/\gamma}}\right]/2\right) \\
&= \frac{1}{\lambda}\left\{\frac{1}{2\alpha}\log\left[\frac{n}{\lambda\log(1/\lambda)^{1/\gamma}}\right]\right\}^{\frac{1-\gamma}{2\gamma}} \exp\left(-\frac{1}{4}\log\left[\frac{n}{\lambda\log(1/\lambda)^{1/\gamma}}\right]\right) \\
&= \frac{1}{\lambda}\left\{\frac{1}{2\alpha}\log\left[\frac{n}{\lambda\log(1/\lambda)^{1/\gamma}}\right]\right\}^{\frac{1-\gamma}{2\gamma}} \left\{\frac{n}{\lambda\log(1/\lambda)^{1/\gamma}}\right\}^{-\frac{1}{4}} \\
&\lesssim_{\omega,\gamma,\alpha} \frac{1}{\lambda}\left\{\log\left[\frac{n}{\lambda\log(1/\lambda)^{1/\gamma}}\right]\right\}^{\frac{1-\gamma}{2\gamma}} \left\{\frac{n}{\lambda\log(1/\lambda)^{1/\gamma}}\right\}^{-\frac{1}{4}}.
\end{aligned}$$

Meanwhile

$$\begin{aligned}
\left\{\frac{m\mathfrak{n}(\lambda)}{n\lambda^2}\right\}^{\frac{1}{4}} &\lesssim_{\omega,\gamma,\alpha} \left[\frac{1}{n\lambda^2}\left\{\frac{1}{2\alpha}\log\left[\frac{n}{\lambda\log(1/\lambda)^{1/\gamma}}\right]\right\}^{\frac{1}{\gamma}} \lambda^{-1}\log(1/\lambda)^{1/\gamma}\right]^{\frac{1}{4}} \\
&=\lesssim_{\omega,\gamma,\alpha} \left\{\log\left[\frac{n}{\lambda\log(1/\lambda)^{1/\gamma}}\right]\right\}^{\frac{1}{4\gamma}} n^{-1/4}\lambda^{-1}\{\lambda\log(1/\lambda)^{1/\gamma}\}^{1/4}.
\end{aligned}$$

Therefore since $\gamma > 0$ implies $\frac{1}{2\gamma} > \frac{1}{4\gamma}$ and $\frac{1}{2\gamma} > \frac{1-\gamma}{2\gamma}$, we conclude that

$$R_{\mathrm{bd}}(n,\lambda) \lesssim_{\omega,\gamma,\alpha} \left\{\log\left[\frac{n}{\lambda\log(1/\lambda)^{1/\gamma}}\right]\right\}^{\frac{1}{2\gamma}} \left\{\frac{n\lambda^3}{\log(1/\lambda)^{1/\gamma}}\right\}^{-\frac{1}{4}}.$$

Gaussian RKHS: Sub-Gaussian data

We proceed in steps.

1. If $\nu_s \le \omega\exp(-\alpha m^\gamma)$ then by Proposition I.2 we have $\sigma(m) \lesssim_{\omega,\alpha,\gamma} m^{\frac{1-\gamma}{2}}\exp(-\alpha m^\gamma/2)$, and by Proposition I.4 we have $\mathfrak{n}(\lambda) \lesssim_{\omega,\alpha,\gamma} \lambda^{-1}\log(1/\lambda)^{1/\gamma}$.

2. For $Q_{\mathrm{sg}}(n,\lambda)$, equating main terms gives

$$\frac{\exp(-\alpha m^\gamma/2)}{\lambda} = \frac{1}{\sqrt{\lambda n}} \iff m = \left\{\frac{1}{\alpha}\log\left(\frac{n}{\lambda}\right)\right\}^{\frac{1}{\gamma}}.$$

Using this value of m,

$$\begin{aligned}
\frac{\sigma(m)}{\lambda} &\lesssim_{\omega,\alpha,\gamma} \lambda^{-1} m^{\frac{1-\gamma}{2}} \exp(-\alpha m^\gamma/2) \\
&= \lambda^{-1}\left\{\frac{1}{\alpha}\log\left(\frac{n}{\lambda}\right)\right\}^{\frac{1-\gamma}{2\gamma}} \exp(-\alpha\frac{1}{\alpha}\log\left(\frac{n}{\lambda}\right)/2) \\
&= \lambda^{-1}\left\{\frac{1}{\alpha}\log\left(\frac{n}{\lambda}\right)\right\}^{\frac{1-\gamma}{2\gamma}} \exp\left\{\log\left(\frac{\lambda^{1/2}}{n^{1/2}}\right)\right\} \\
&= \left\{\frac{1}{\alpha}\log\left(\frac{n}{\lambda}\right)\right\}^{\frac{1-\gamma}{2\gamma}} \frac{1}{\lambda^{1/2}n^{1/2}} \\
&\lesssim_{\omega,\alpha,\gamma} \log\left(\frac{n}{\lambda}\right)^{\frac{1-\gamma}{2\gamma}} \frac{1}{\lambda^{1/2}n^{1/2}}.
\end{aligned}$$

Meanwhile

$$\frac{m^{3/2}\log(n)^2}{\sqrt{n\lambda}} = \frac{\log(n)^2}{\sqrt{n\lambda}}\left\{\frac{1}{\alpha}\log\left(\frac{n}{\lambda}\right)\right\}^{\frac{3}{2\gamma}} \lesssim_{\omega,\alpha,\gamma} \frac{\log(n)^2}{\sqrt{n\lambda}}\log\left(\frac{n}{\lambda}\right)^{\frac{3}{2\gamma}}.$$

Since $\gamma > 0$ implies $\frac{3}{2\gamma} > \frac{1-\gamma}{2\gamma}$, w conclude that

$$Q_{\mathrm{sg}}(n,\lambda) \lesssim_{\omega,\alpha,\gamma} \frac{\log(n)^2}{\sqrt{n\lambda}}\log\left(\frac{n}{\lambda}\right)^{\frac{3}{2\gamma}}.$$

3. Next we study $R_{\mathrm{sg}}(n,\lambda)$. Plugging in our bounds and equating main terms gives

$$\begin{aligned}
\left\{\frac{\lambda^{-2}\log(1/\lambda)^{2/\gamma}}{n}\right\}^{\frac{1}{4}} &= \frac{\exp(-\alpha m^\gamma/2)}{\lambda} \\
\iff \left\{\frac{\lambda^{2}\log(1/\lambda)^{2/\gamma}}{n}\right\}^{\frac{1}{4}} &= \exp(-\alpha m^\gamma/2) \\
\iff \left[\frac{1}{\alpha}\log\left\{\frac{n^{1/2}}{\lambda\log(1/\lambda)^{1/\gamma}}\right\}\right]^{\frac{1}{\gamma}} &= m.
\end{aligned}$$

Using this value of m,

$$\begin{aligned}
\frac{\sigma(m)}{\lambda} &\lesssim_{\omega,\gamma,\alpha} \lambda^{-1} m^{\frac{1-\gamma}{2}} \exp(-\alpha m^{\gamma}/2) \\
&= \lambda^{-1} \left[\frac{1}{\alpha} \log\left\{\frac{n^{1/2}}{\lambda \log(1/\lambda)^{1/\gamma}}\right\}\right]^{\frac{1-\gamma}{2\gamma}} \exp\left[-\alpha \frac{1}{\alpha} \log\left\{\frac{n^{1/2}}{\lambda \log(1/\lambda)^{1/\gamma}}\right\}/2\right] \\
&= \lambda^{-1} \left[\frac{1}{\alpha} \log\left\{\frac{n^{1/2}}{\lambda \log(1/\lambda)^{1/\gamma}}\right\}\right]^{\frac{1-\gamma}{2\gamma}} \exp\left[\log\left\{\frac{\lambda^{1/2} \log(1/\lambda)^{1/2\gamma}}{n^{1/4}}\right\}\right] \\
&= \lambda^{-1} \left[\frac{1}{\alpha} \log\left\{\frac{n^{1/2}}{\lambda \log(1/\lambda)^{1/\gamma}}\right\}\right]^{\frac{1-\gamma}{2\gamma}} \frac{\lambda^{1/2} \log(1/\lambda)^{1/2\gamma}}{n^{1/4}} \\
&\lesssim_{\omega,\alpha,\gamma} \log\left\{\frac{n^{1/2}}{\lambda \log(1/\lambda)^{1/\gamma}}\right\}^{\frac{1-\gamma}{2\gamma}} \frac{\log(1/\lambda)^{1/2\gamma}}{\lambda^{1/2} n^{1/4}}.
\end{aligned}$$

Meanwhile

$$\begin{aligned}
\left\{\frac{m\mathfrak{n}(\lambda)^2}{n}\right\}^{1/4} &\lesssim_{\omega,\gamma,\alpha} \left[\frac{1}{\alpha} \log\left\{\frac{n^{1/2}}{\lambda \log(1/\lambda)^{1/\gamma}}\right\}\right]^{\frac{1}{4\gamma}} n^{-1/4} \left\{\lambda^{-1} \log(1/\lambda)^{1/\gamma}\right\}^{1/2} \\
&\lesssim_{\omega,\gamma,\alpha} \log\left\{\frac{n^{1/2}}{\lambda \log(1/\lambda)^{1/\gamma}}\right\}^{\frac{1}{4\gamma}} n^{-1/4} \left\{\lambda^{-1} \log(1/\lambda)^{1/\gamma}\right\}^{1/2}.
\end{aligned}$$

Since $\gamma > 0$ implies $\frac{1}{2\gamma} > \frac{1}{4\gamma}$ and $\frac{1}{2\gamma} > \frac{1-\gamma}{2\gamma}$, we conclude that

$$R_{\mathrm{bd}}(n,\lambda) \lesssim_{\omega,\alpha,\gamma} \log\left\{\frac{n^{1/2}}{\lambda \log(1/\lambda)^{1/\gamma}}\right\}^{\frac{1}{2\gamma}} \frac{\log(1/\lambda)^{1/2\gamma}}{\lambda^{1/2} n^{1/4}}.$$

Summary

Spectrum	Poly. : $\nu_s \le \omega s^{-\beta}$		Exp. : $\nu_s \le \omega \exp(-\alpha s^{\gamma})$	
Data	Bounded	sub-Gaussian	Bounded	sub-Gaussian
$Q_\bullet$	$\lambda^{-1} n^{\frac{1-\beta}{6+2\beta}}$	$\lambda^{\frac{-(5+\beta)}{4+2\beta}} n^{\frac{1-\beta}{4+2\beta}}$	$\lambda^{-1} n^{-\frac{1}{2}}$	$(\lambda n)^{-\frac{1}{2}}$
$R_\bullet$	$\left(\lambda^{3+\frac{1}{\beta}+\frac{2}{\beta-1}} n\right)^{\frac{1-\beta}{4\beta-2}}$	$\left(\lambda^{2+\frac{2}{\beta}+\frac{2}{\beta-1}} n\right)^{\frac{1-\beta}{4\beta-2}}$	$\lambda^{-3/4} n^{-1/4}$	$\lambda^{-1/2} n^{-1/4}$

Table 11: Summary of results under different assumptions. The initial two rows present rates (suppressing log factors) for the quantities Q_{bd}, Q_{sg}, R_{bd}, and R_{sg}.

These values correspond to (Q, R) in Table 7. We now derive the remaining entries in that table. We proceed in steps.

1. L. Lemma G.13 implies $L \asymp \mathfrak{n}(\lambda)^{1/2} - \lambda^{-1/2}$. Note that $\mathfrak{n}(\lambda) = \sigma^2\{(T+\lambda)^{-2}T, 0\} = \psi(0,2)$, so we appeal to Proposition I.5. For polynomial decay, we show $\psi(0,2) \gtrsim \lambda^{-1-1/\beta}$. Hence, for $\beta > 1$,

$$L \asymp \mathfrak{n}(\lambda)^{1/2} - \lambda^{-1/2} \gtrsim \lambda^{-\frac{1}{2}-\frac{1}{2\beta}} - \lambda^{-1/2} = \lambda^{-1/2}(\lambda^{-\frac{1}{2\beta}} - 1) \gtrsim \lambda^{-\frac{1}{2}-\frac{1}{2\beta}}.$$

For exponential decay, we show $\psi(0,2) \gtrsim \lambda^{-1}$, ignoring log factors. Hence

$$L \asymp \mathfrak{n}(\lambda)^{1/2} - \lambda^{-1/2} \gtrsim \lambda^{-\frac{1}{2}} - \lambda^{-\frac{1}{2}} \asymp \lambda^{-\frac{1}{2}}.$$

2. $B \ll L$. For polynomial decay $n\lambda^{r-1} \ll \lambda^{-1-1/\beta} \iff \lambda^{r+1/\beta} \ll n^{-1}$. For exponential decay $n\lambda^{r-1} \ll \lambda^{-1} \iff \lambda^r \ll n^{-1}$.

3. $Q + R \ll L$.

   (a) Polynomial decay, bounded data. We have

$$\begin{aligned} Q \ll L &\iff \lambda^{-2} n^{\frac{1-\beta}{3+\beta}} \ll \lambda^{-1-1/\beta} \iff n^{\frac{1-\beta}{3+\beta}} \ll \lambda^{1-1/\beta} \\ &\iff n^{\frac{1-\beta}{3+\beta}} \ll \lambda^{(\beta-1)/\beta} \iff n^{\frac{1}{3+\beta}} \gg \lambda^{-1/\beta} \iff n^{\frac{-\beta}{3+\beta}} \ll \lambda. \end{aligned}$$

   Then

$$\begin{aligned} R \ll L &\iff \{\lambda^{3+\frac{1}{\beta}+\frac{2}{\beta-1}} n\}^{\frac{1-\beta}{2\beta-1}} \ll \lambda^{-1-1/\beta} \\ &\iff \{\lambda^{3+\frac{1}{\beta}+\frac{2}{\beta-1}} n\}^{\frac{1-\beta}{2\beta-1}} \ll \lambda^{-(\beta+1)/\beta} \\ &\iff \lambda^{3+\frac{1}{\beta}+\frac{2}{\beta-1}} n \gg \lambda^{-\frac{(\beta+1)(2\beta-1)}{\beta(1-\beta)}} \\ &\iff \lambda^{3+\frac{1}{\beta}+\frac{2}{\beta-1}} n \gg \lambda^{\frac{(\beta+1)(2\beta-1)}{\beta(\beta-1)}} \\ &\iff \lambda^{\frac{3\beta(\beta-1)+(\beta-1)+2\beta}{\beta(\beta-1)}} n \gg \lambda^{\frac{(\beta+1)(2\beta-1)}{\beta(\beta-1)}} \\ &\iff n \gg \lambda^{\frac{(\beta+1)(2\beta-1)-[3\beta(\beta-1)+(\beta-1)+2\beta]}{\beta(\beta-1)}} \end{aligned}$$

   To simplify the numerator, note that

$$\begin{aligned} &(\beta+1)(2\beta-1) - [3\beta(\beta-1) + (\beta-1) + 2\beta] \\ &= 2\beta^2 + \beta - 1 - [3\beta^2 - 3\beta + \beta - 1 + 2\beta] \\ &= -\beta^2 + \beta = \beta(1-\beta). \end{aligned}$$

Therefore the above simplifies to $n \gg \lambda^{-1} \iff \lambda \gg 1/n$.

Since $1 > \beta/(3+\beta)$, the Q condition binds. Combining $Q + R \ll L$ with $B \ll L$,

$$n^{\frac{-\beta}{3+\beta}} \ll \lambda \ll n^{-\beta/(r\beta+1)} \iff 3 + \beta \leq r\beta + 1 \iff r \geq 1 + 2/\beta.$$

(b) Polynomial decay, sub-Gaussian data. We have

$$\begin{aligned} Q \ll L &\iff \lambda^{-\frac{5+\beta}{2+\beta}} n^{\frac{1-\beta}{2+\beta}} \ll \lambda^{-1-1/\beta} \\ &\iff \lambda^{-\frac{3}{2+\beta}} n^{\frac{1-\beta}{2+\beta}} \ll \lambda^{-1/\beta} \\ &\iff \lambda^{-3} n^{1-\beta} \ll \lambda^{-(2+\beta)/\beta} \\ &\iff n^{-1(\beta-1)} \ll \lambda^{(2\beta-2)/\beta} \\ &\iff \lambda \gg n^{-\beta/2}. \end{aligned}$$

Then

$$\begin{aligned} R \ll L &\iff \{\lambda^{2+\frac{2}{\beta}+\frac{2}{\beta-1}} n\}^{\frac{1-\beta}{2\beta-1}} \ll \lambda^{-1-1/\beta} \\ &\iff \{\lambda^{2+\frac{2}{\beta}+\frac{2}{\beta-1}} n\}^{\frac{1-\beta}{2\beta-1}} \ll \lambda^{-(\beta+1)/\beta} \\ &\iff \lambda^{2+\frac{2}{\beta}+\frac{2}{\beta-1}} n \gg \lambda^{-\frac{(\beta+1)(2\beta-1)}{\beta(1-\beta)}} \\ &\iff \lambda^{2+\frac{2}{\beta}+\frac{2}{\beta-1}} n \gg \lambda^{\frac{(\beta+1)(2\beta-1)}{\beta(\beta-1)}} \\ &\iff \lambda^{\frac{2\beta(\beta-1)+2(\beta-1)+2\beta}{\beta(\beta-1)}} n \gg \lambda^{\frac{(\beta+1)(2\beta-1)}{\beta(\beta-1)}} \\ &\iff n \gg \lambda^{\frac{(\beta+1)(2\beta-1)-[2\beta(\beta-1)+2(\beta-1)+2\beta]}{\beta(\beta-1)}} \end{aligned}$$

To simplify the numerator, note that

$$\begin{aligned} (\beta+1)(2\beta-1) - [2\beta(\beta-1) + 2(\beta-1) + 2\beta] &= 2\beta^2 + \beta - 1 - [2\beta^2 - 2\beta + 2\beta - 2 + 2\beta] \\ &= 1 - \beta. \end{aligned}$$

Therefore the above simplifies to $n \gg \lambda^{-1/\beta} \iff \lambda \gg n^{-\beta}$.

Since $\beta > \beta/2$, the Q condition binds. Combining $Q + R \ll L$ with $B \ll L$,

$$n^{\frac{-\beta}{2}} \ll \lambda \ll n^{-\beta/(r\beta+1)} \iff \frac{1}{2} \geq \frac{1}{r\beta+1} \iff r \geq 1/\beta.$$

(c) Exponential decay, bounded data. The two conditions are equivalent:

$$Q \ll L \iff \lambda^{-2}n^{-1} \ll \lambda^{-1} \iff \lambda \gg n^{-1};$$

$$R \ll L \iff \lambda^{-\frac{3}{2}}n^{-\frac{1}{2}} \ll \lambda^{-1} \iff \lambda^{1/2} \gg n^{-1/2}.$$

Combining $Q + R \ll L$ with $B \ll L$,

$$n^{-1} \ll \lambda \ll n^{-1/r} \iff 1 \geq \frac{1}{r} \iff r \geq 1.$$

(d) Exponential decay, sub-Gaussian data. The two conditions are equivalent:

$$Q \ll L \iff \lambda^{-1}n^{-1} \ll \lambda^{-1} \iff 1 \gg n^{-1};$$

$$R \ll L \iff \lambda^{-1}n^{-\frac{1}{2}} \ll \lambda^{-1} \iff 1 \gg n^{-1/2}.$$

Combining $Q + R \ll L$ with $B \ll L$, $0 \ll \lambda \ll n^{-1/r}$.

H Concentration inequalities for Bahadur representation (Sections B and E)

H.1 Concentration

We quote a version of Bernstein's inequality for random variables in a Hilbert space.

Lemma H.1 (Theorem 3.3.4 of Yurinsky (1995))**.** Let $\xi_1, \xi_2, \ldots$ be an independent sequence of random variables in a Hilbert space that satisfy $\mathbb{E}[\xi_i] = 0$. If, for some $B, A > 0$ and all $\ell \geq 2$ it holds that $\sum_{i=1}^{n} \mathbb{E}\|\xi_i\|^{\ell} \leq \frac{\ell!}{2} B^2 A^{\ell-2}$, then for $\delta > 0$

$$\mathbb{P}\left(\max_{m\in[n]} \left\|\sum_{i=1}^{m} \xi_i\right\| \geq \delta B\right) \leq 2\exp\left\{\frac{-\delta^2/2}{1+\delta A/B}\right\}.$$

A useful corollary, frequently used in the kernel ridge regression literature (Caponnetto and De Vito, 2007), is as follows.

Lemma H.2. Suppose that ξ_i are i.i.d. random elements of a Hilbert space, which satisfy, for all $\ell \geq 2$ $\mathbb{E}\left\|\xi_i - \mathbb{E}\xi_i\right\|^\ell \leq \frac{1}{2}\ell!\, B^2(A/2)^{\ell-2}$. Then for any $0 < \eta < 1$ it holds with probability at least $1-\eta$ that

$$\left\|\frac{1}{n}\sum_{i=1}^n \xi_i - \mathbb{E}[\xi_i]\right\| \leq 2\left(\sqrt{\frac{B^2\log(2/\eta)}{n}} \vee \frac{A\log(2/\eta)}{n}\right) \leq 2\log(2/\eta)\left\{\frac{A}{n} \vee \sqrt{\frac{B^2}{n}}\right\}.$$

In particular, this holds if $\mathbb{E}(\|\xi_i\|^2) \leq B^2$ and $\|\xi_i\| \leq A/2$ almost surely.

Remark H.3. Note that the result of Lemma H.2 for bounded random vectors may be recovered from Talagrand's concentration inequality for empirical processes (see Massart, 2000), by considering the special case where the sample paths are linear and the parameter space is an ellipsoid.

H.2 Bounds for sums

Lemma H.4. With probability $1-\eta$,

$$\left\|\frac{1}{n}\sum_{i=1}^n T_\lambda^{-1}\varepsilon_i k_{X_i}\right\| \leq 2\bar{\sigma}\ln(2/\eta)\left\{\sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \vee \frac{2\kappa}{n\lambda}\right\}.$$

Proof. Note that $\mathbb{E}[T_\lambda^{-1}\varepsilon_i k_{X_i}] = \mathbb{E}[T_\lambda^{-1}k_{X_i}\mathbb{E}(\varepsilon_i|X_i)] = 0$. It therefore suffices to show $\left\|T_\lambda^{-1}\varepsilon_i k_{X_i}\right\| \leq \left\|T_\lambda^{-1}\right\|_{op}\|\varepsilon_i k_{X_i}\| \leq \frac{\kappa\bar{\sigma}}{\lambda}$ and

$$\mathbb{E}\left\|T_\lambda^{-1}\varepsilon_i k_{X_i}\right\|^2 \leq \bar{\sigma}^2\mathbb{E}\left\|T_\lambda^{-1}k_{X_i}\right\|^2 = \bar{\sigma}^2\sum_{s=1}^\infty \frac{\mathbb{E}\left\langle k_{X_i}, e_s\right\rangle^2}{(\nu_s+\lambda)^2} = \bar{\sigma}^2\sum_{s=1}^\infty \frac{\nu_s}{(\nu_s+\lambda)^2} = \bar{\sigma}^2\mathfrak{n}(\lambda).$$

Plugging these estimates into Lemma H.2 gives the result. □

Lemma H.5. With probability $1-\eta$,

$$\left\|\frac{1}{n}\sum_{i=1}^n T_\lambda^{-1}(T_i - T)\right\|_{\mathrm{HS}} \leq 2\kappa\ln(2/\eta)\left\{\sqrt{\frac{\mathfrak{n}(\lambda)}{n}} \vee \frac{2\kappa}{n\lambda}\right\}.$$

Proof. Note that $\mathbb{E}[T_\lambda^{-1}(T_i - T)] = T_\lambda^{-1}\mathbb{E}[T_i - T] = 0$. Also,

$$\left\|T_\lambda^{-1}(T_i - T)\right\|_{\mathrm{HS}} \leq \left\|T_\lambda^{-1}\right\|_{op}\|T_i - T\|_{\mathrm{HS}} \leq \frac{2\kappa^2}{\lambda}.$$

Finally, using the fact that T, T_i and T_λ^{-1} are self-adjoint,

$$\begin{aligned}\mathbb{E}\left\|T_\lambda^{-1}(T_i - T)\right\|_{\mathrm{HS}}^2 &= \mathbb{E}\,\mathrm{tr}[T_\lambda^{-1}(T_i - T)^2 T_\lambda^{-1}] \\ &= \mathrm{tr}\,\mathbb{E}[T_\lambda^{-1}(T_i - T)^2 T_\lambda^{-1}] \\ &= \mathrm{tr}\{\mathbb{E}[T_\lambda^{-1}T_i^2 T_\lambda^{-1}] - T_\lambda^{-1}TT_\lambda^{-1}\}.\end{aligned}$$

Since both operators are positive definite, this is

$$\leq \mathrm{tr}\{\mathbb{E}[T_\lambda^{-1}T_i^2 T_\lambda^{-1}].$$

Since $T_i = k_{X_i} \otimes k_{X_i}^*$, $\|T_i\|_{op} \leq \|T_i\|_{\mathrm{HS}} = \|k_{X_i}\|^2 \leq \kappa^2$, and hence

$$\leq \kappa^2 \mathbb{E}[|\mathrm{tr}\{T_\lambda^{-1}T_i T_\lambda^{-1}\}|].$$

Since the trace is almost surely positive, this is

$$\begin{aligned}&= \kappa^2 \mathbb{E}[\mathrm{tr}\{T_\lambda^{-1}T_i T_\lambda^{-1}\}] \\ &= \kappa^2\,\mathrm{tr}[T_\lambda^{-1}TT_\lambda^{-1}].\end{aligned}$$

Finally, since T commutes with T_λ^{-1} by construction, this is

$$= \kappa^2\,\mathrm{tr}[T_\lambda^{-2}T] = \kappa^2\mathfrak{n}(\lambda).$$

Plugging these estimates into Lemma H.2 gives the result. □

I Spectral bounds for Gaussian couplings (Section C)

We characterize the behavior of $\sigma^2(m) = \sum_{s>m}\nu_s$ and $\mathfrak{n}(\lambda) = \mathrm{tr}(T_\lambda^{-2}T)$ under various regimes for the spectrum of Σ. Such regimes can be deduced directly from regularity of the kernel function k (Belkin, 2018).

I.1 Spectral decay

We need the following technical lemma on the size of the incomplete Gamma function.

Lemma I.1 (Natalini and Palumbo 2000, Sec 3.1 and eq. 3.5)**.** Let the (upper) incomplete gamma function Γ be given by $\Gamma(z,x) \coloneqq \int_x^\infty u^{z-1}e^{-u}\,du$. For $z > 1$ and $x \geq 0$ we have

$$x^{z-1}e^{-x} < \Gamma(z,x) \leq \frac{x^z e^{-x}}{x-z}.$$

Proposition I.2. Suppose $\nu_s \leq \bar{\nu}(s) : \mathbb{R} \to \mathbb{R}$ for some non-increasing positive function $\bar{\nu}(s)$. Then, it holds that $\sigma^2(m) = \sum_{s=m+1}^\infty \nu_s \leq \int_m^\infty \bar{\nu}(s)\,ds$. In particular, if $\nu_s \leq \omega s^{-\beta}$ for $\beta > 1$, then we have $\sigma^2(m) \leq \frac{\omega m^{1-\beta}}{\beta-1}$. Similarly, if $\nu_s \leq \omega \exp(-\alpha s^\gamma)$ for $\alpha, \omega, \gamma > 0$, then we have $\sigma^2(m) \leq C(\gamma,\alpha)\omega m^{1-\gamma}\exp(-\alpha m^\gamma)$, for a constant $C(\gamma,\alpha)$ which depends only on α and γ (and not on m).

Proof. We upper bound $\sigma^2(m) = \sum_{s=m+1}^\infty \nu_s \leq \int_m^\infty \bar{\nu}(s)\,ds$, where $\bar{\nu} : \mathbb{R} \to \mathbb{R}$ is any non-increasing function with $\bar{\nu}(s) \geq \nu_s$. The first bound follows from taking $\bar{\nu}(s) = \omega s^{-\beta}$ and computing the resulting integral exactly. For the second bound, we take $\bar{\nu}(s) = \omega\exp(-\alpha s^\gamma)$ and obtain $\frac{\sigma^2(m)}{\omega} \leq \int_m^\infty \exp(-\alpha s^\gamma)\,ds$. To bound the right hand side, we proceed in steps.

1. Making the substitution $u = \alpha s^\gamma$ gives us

$$\int_m^\infty \exp(-\alpha s^\gamma)\,ds = \gamma^{-1}\alpha^{\frac{-1}{\gamma}} \int_{\alpha m^\gamma}^\infty \exp(-u)u^{(1-\gamma)/\gamma}\,du.$$

2. We have $\int_{\alpha m^\gamma}^\infty \exp(-u)u^{(1-\gamma)/\gamma}\,du = \Gamma(1/\gamma, \alpha m^\gamma)$ where $\Gamma(z,x)$ is the incomplete Gamma function (cf. Lemma I.1). By Lemma I.1, $\Gamma(z,x) \leq \frac{x^z e^{-x}}{x-z}$, so

$$\frac{\sigma^2(m)}{\omega} \leq \gamma^{-1}\alpha^{\frac{-1}{\gamma}}\Gamma\Big(1/\gamma, \alpha m^\gamma\Big) \leq \gamma^{-1}\alpha^{\frac{-1}{\gamma}}\frac{(\alpha m^\gamma)^{1/\gamma}e^{-\alpha m^\gamma}}{\alpha m^\gamma - 1/\gamma} = \frac{m\exp(-\alpha m^\gamma)}{\alpha\gamma m^\gamma - 1}.$$

3. Finally we absorb constants. When $m \geq \{2/(\alpha\gamma)\}^{1/\gamma}$, we have that $\alpha\gamma m^\gamma \geq 2$ and hence $\alpha\gamma m^\gamma - 1 \geq \alpha\gamma m^\gamma/2$. Thus, the final expression is at most $(2/\alpha\gamma)m^{1-\gamma}\exp(-\alpha m^\gamma)$. For $m < \{2/(\alpha\gamma)\}^{1/\gamma}$, we decompose the sum as

$$\sum_{s=m+1}^{\lfloor\{2/(\alpha\gamma)\}^{1/\gamma}\rfloor} \nu_s + \sum_{s=\lceil\{2/(\alpha\gamma)\}^{1/\gamma}\rceil}^{\infty} \nu_s.$$

The latter sum is bounded by the previous case. Focusing on the former sum,

$$\sum_{s=m+1}^{\{2/(\alpha\gamma)\}^{1/\gamma}} \nu_s \leq \sum_{s=1}^{\{2/(\alpha\gamma)\}^{1/\gamma}} \nu_s \leq \{2/(\alpha\gamma)\}^{1/\gamma}\nu_1 \leq \{2/(\alpha\gamma)\}^{1/\gamma} \cdot \omega \exp(-\alpha)$$

since $\nu_1 \leq \bar{\nu}(1) \leq \omega \exp(-\alpha 1^\gamma) = \omega \exp(-\alpha)$. These terms may be absorbed into a constant depending only on ω, α, γ, and which is linear in ω.

□

I.2 Complexity measures

Recall the definition of the ellipse $\mathfrak{E} = \Sigma^{\frac{1}{2}} B$, where B is the unit ball in $L^2(\mathbb{P})$. We use the results of Wei et al. (2020) to briefly sketch the claim that, under suitable regularity conditions, the local width $\sigma(\Sigma, m)$ is roughly comparable to the local Gaussian complexity of $\mathfrak{E}$ at scale $\delta = \text{ent}_m(\mathfrak{E})$, which we denote by $\mathcal{G}(\mathfrak{E} \cap \delta B)$.

First, some technicalities. For $S_m \subset \mathbb{R}^m$, the Gaussian width and Gaussian complexity are given by $\mathbb{E} \sup_{t \in S_m} \langle g, s \rangle$ and $\mathbb{E} \sup_{t \in S_m} |\langle g, s \rangle|$, respectively, where g is an isotropic Gaussian vector. These quantities immediately generalize to compact subsets of a separable Hilbert space by finite-dimensional approximation. Moreover, when S_m contains the origin (as in our setting) they are equivalent up to a constant (Vershynin, 2018, Chapter 7). Thus, as is standard, we can use the results of Wei et al. (2020), who study the Gaussian width of finite-dimensional sets, to characterize the Gaussian complexity in our setting.

Lemma I.3. Suppose either

1. $\nu_s(\Sigma) \asymp s^{-\beta}$ for $\beta > 1$, or
2. $\nu_s(\Sigma) \asymp \exp(-\alpha s^\gamma)$ for $\alpha > 0$ and $\gamma \in (0, 1)$.

Then, for sufficiently small $\delta = \text{ent}_m(\mathfrak{E})$, we have $\sigma(\Sigma, m) \lesssim \mathcal{G}(\mathfrak{E} \cap \delta B)$.

Proof. Both cases were covered by Wei et al. (2020, Example 4), and the subsequent discussion in Section 4 of that paper.

In particular, it was shown that given a scale parameter δ, one can compute the "critical dimension" m^* as the smallest m such that $\sqrt{\nu_m(\Sigma)} \leq 9\delta/10$. Then, according to regularity conditions which were checked in the aforementioned Example 4, one has by Wei et al. (2020, Theorem 2) that for small enough δ, $\delta\sqrt{m^*} \lesssim \mathcal{G}(\mathfrak{E} \cap \delta B)$. It then follows from our bounds in Proposition I.2 above that

$$\delta\sqrt{m^*} \gtrsim \sqrt{m^*\nu_{m^*}(\Sigma)} \gtrsim \sigma(\Sigma, m^*)$$

in both of the above examples.

Next, Wei et al. (2020, Corollary 2) implies (after inverting the entropy function) that in the same setting

$$\delta \lesssim \mathrm{ent}_{m^*}(\mathfrak{E} \cap \delta B) \leq \mathrm{ent}_{m^*}(\mathfrak{E}) = \delta'.$$

Our sketch is then complete by monotonicity of $\mathcal{G}$ since

$$\sigma(\Sigma, m^*) \lesssim \mathcal{G}(\mathfrak{E} \cap \delta B) \leq \mathcal{G}(\mathfrak{E} \cap \delta' B).$$

□

I.3 Effective dimension

In this subsection, we will derive bounds on the quantities

$$\psi(m, c) = \sum_{s=m+1}^{\infty} \frac{\nu_s}{(\nu_s + \lambda)^c},$$

where $c \geq 1$, which are crucial for our analysis. Such quantities arise frequently in studies of ridge regression; see, for example, the "effective dimension" appearing in Caponnetto and De Vito (2007). In particular note that according to our definition, $\mathfrak{n}(\lambda) = \psi(0, 2)$. We provide matching upper and lower bounds.

Proposition I.4 (Upper bound on effective dimension)**.** Suppose $\nu_s \leq \bar{\nu}(s) : \mathbb{R} \to \mathbb{R}$ for some non-increasing positive function $\bar{\nu}(s)$, and let $c \geq 1$ be given. Then

$$\psi(m, c) \leq \lambda^{1-c} \inf_{s \geq m} \left\{ (s - m + 1) + \frac{1}{\lambda} \int_s^{\infty} \bar{\nu}(t)\, dt \right\}.$$

In particular, if $\nu_s \le \omega s^{-\beta}$ for some $\omega > 0$ and $\beta > 1$, and if $\lambda \le \omega$,

$$\psi(m, c) \le \frac{\omega(2\beta - 1)}{\beta - 1} \left(\frac{1}{\lambda^{c+1/\beta-1}} \wedge \frac{m^{1-\beta}}{\lambda^c} \right).$$

If $\nu_s \le \omega e^{-\alpha s^\gamma}$ for some $\alpha, \omega, \gamma > 0$, and if $\lambda \le \omega / e^\alpha$, then

$$\psi(m, c) \le C(\alpha, \omega, \gamma) \left\{ \frac{1}{\lambda^{c-1}} \left(\frac{\log(\omega/\lambda)}{\alpha} \right)^{1/\gamma} \wedge \frac{m^{1-\gamma} \exp(-\alpha m^\gamma)}{\lambda^c} \right\}$$

for some constant $C(\alpha, \omega, \gamma)$ which depends only on α, ω, and γ.

Proof of Proposition I.4. We proceed in steps.

1. First we note

$$\frac{\nu_s}{(\nu_s + \lambda)^c} \le \frac{\nu_s}{\lambda^c}, \quad \frac{\nu_s}{(\nu_s + \lambda)^c} \le \lambda^{1-c}.$$

   The former holds since $\nu_s \ge 0$. The latter follows from maximizing the function $f(u) = \frac{u}{(u+\lambda)^c}$ at $u = \lambda/(c-1)$.

2. Combining the two, we have for any $k \ge m$ that

$$\begin{aligned} \psi(m, c) &= \sum_{s=m+1}^{\infty} \frac{\nu_s}{(\nu_s + \lambda)^c} \\ &\le \sum_{s=m+1}^{k} \lambda^{1-c} + \sum_{s=k+1}^{\infty} \frac{\nu_s}{\lambda^c} \\ &\le \lambda^{1-c} \left\{ (k - m) + \frac{1}{\lambda} \int_k^\infty \bar{\nu}(t)\, dt \right\}. \end{aligned}$$

   Since k was arbitrary, we can further deduce that

$$\begin{aligned} \psi(m, c) &\le \lambda^{1-c} \inf_{k \ge m} \left\{ (k - m) + \frac{1}{\lambda} \int_k^\infty \bar{\nu}(t)\, dt \right\} \\ &= \lambda^{1-c} \inf_{s \ge m} \left\{ (\lceil s \rceil - m) + \frac{1}{\lambda} \int_{\lceil s \rceil}^\infty \bar{\nu}(t)\, dt \right\} \\ &\le \lambda^{1-c} \inf_{s \ge m} \left\{ (\lceil s \rceil - m) + \frac{1}{\lambda} \int_s^\infty \bar{\nu}(t)\, dt \right\}, \end{aligned}$$

   where the latter infimum is over real numbers s.

3. Optimizing the bounds. In the infimum, consider $s = m$. Then

$$\psi(m, c) \leq \lambda^{-c} \int_m^\infty \bar{\nu}(t)\, dt.$$

Alternatively, consider the choice $s^* = \bar{\nu}^{-1}(\lambda) \vee m$. Then

$$\psi(m, c) \leq \lambda^{1-c} \left\{ 1 + s^* - m + \frac{1}{\lambda} \int_{s^*}^\infty \bar{\nu}(t)\, dt \right\}.$$

In particular, choosing $m = 0$ and $s^* = \bar{\nu}^{-1}(\lambda)$, we see

$$\psi(m, c) \leq \psi(0, c) \leq \lambda^{1-c} \left\{ 1 + \bar{\nu}^{-1}(\lambda) + \frac{1}{\lambda} \int_{\bar{\nu}^{-1}(\lambda)}^\infty \bar{\nu}(t)\, dt \right\}.$$

Combining both bounds,

$$\psi(m, c) \leq \lambda^{1-c} \left\{ 1 + \bar{\nu}^{-1}(\lambda) + \frac{1}{\lambda} \int_{\bar{\nu}^{-1}(\lambda)}^\infty \bar{\nu}(t)\, dt \right\} \wedge \frac{1}{\lambda^c} \int_m^\infty \bar{\nu}(t)\, dt.$$

4. Polynomial decay. Now, if $\bar{\nu}(s) = \omega s^{-\beta}$ then we have $\bar{\nu}^{-1}(\lambda) = (\omega/\lambda)^{1/\beta}$, and as argued in Proposition I.2, $\int_s^\infty \bar{\nu}(t)\, dt = \frac{\omega s^{1-\beta}}{\beta - 1}$. Plugging this into the first term in the optimized bound,

$$\frac{1}{\lambda} \int_{\bar{\nu}^{-1}(\lambda)}^\infty \bar{\nu}(t)\, dt = \frac{1}{\lambda} \frac{\omega \{(\omega/\lambda)^{1/\beta}\}^{1-\beta}}{\beta - 1} = \lambda^{-1-(1-\beta)/\beta} \omega^{1+(1-\beta)/\beta} \frac{1}{\beta - 1} = \lambda^{-1/\beta} \cdot \omega^{1/\beta} \cdot (\beta - 1)^{-1}$$

so that, when $\omega \geq \lambda$,

$$\begin{aligned}
\lambda^{1-c} \left\{ 1 + \bar{\nu}^{-1}(\lambda) + \frac{1}{\lambda} \int_{\bar{\nu}^{-1}(\lambda)}^\infty \bar{\nu}(t)\, dt \right\} &= \lambda^{1-c} \left\{ 1 + (\omega/\lambda)^{1/\beta} + \lambda^{-1/\beta} \cdot \omega^{1/\beta} \cdot (\beta - 1)^{-1} \right\} \\
&\leq \lambda^{1-c} \left\{ 2(\omega/\lambda)^{1/\beta} + \lambda^{-1/\beta} \cdot \omega^{1/\beta} \cdot (\beta - 1)^{-1} \right\} \\
&= \left(2\omega^{1/\beta} + \frac{\omega^{1/\beta}}{\beta - 1} \right) \lambda^{1-c-1/\beta} \\
&= \frac{2\beta - 1}{\beta - 1} \omega^{1/\beta} \lambda^{1-c-1/\beta}.
\end{aligned}$$

Therefore the overall bound is

$$\frac{2\beta - 1}{\beta - 1} \omega^{1/\beta} \lambda^{1-c-1/\beta} \wedge \frac{\omega m^{1-\beta}}{\lambda^c (\beta - 1)} \leq \frac{\omega(2\beta - 1)}{\beta - 1} \left(\frac{1}{\lambda^{c+1/\beta-1}} \wedge \frac{m^{1-\beta}}{\lambda^c} \right).$$

5. Exponential decay. If $\bar{\nu}(t) \leq \omega \exp(-\alpha t^\gamma)$ then we have $\bar{\nu}^{-1}(\lambda) = \{\log(\omega/\lambda)/\alpha\}^{1/\gamma}$ and, as argued in Proposition I.2, $\int_s^\infty \bar{\nu}(t)\,dt \leq C\omega s^{1-\gamma}\exp(-\alpha s^\gamma)$. Plugging this into the first term in the optimized bound,

$$\begin{aligned}
\frac{1}{\lambda}\int_{\bar{\nu}^{-1}(\lambda)}^\infty \bar{\nu}(t)\,dt &= \frac{1}{\lambda}C\omega[\{\log(\omega/\lambda)/\alpha\}^{1/\gamma}]^{1-\gamma}\cdot\exp(-\alpha[\{\log(\omega/\lambda)/\alpha\}^{1/\gamma}]^\gamma) \\
&= \frac{1}{\lambda}C\omega\{\log(\omega/\lambda)/\alpha\}^{(1-\gamma)/\gamma}\cdot\exp(-\alpha\{\log(\omega/\lambda)/\alpha\}) \\
&= \frac{1}{\lambda}C\omega\{\log(\omega/\lambda)/\alpha\}^{(1-\gamma)/\gamma}\cdot\frac{\lambda}{\omega} \\
&= C\{\log(\omega/\lambda)/\alpha\}^{(1-\gamma)/\gamma}
\end{aligned}$$

so that, when $\omega/e^\alpha \geq \lambda$,

$$\begin{aligned}
&\lambda^{1-c}\left\{1+\bar{\nu}^{-1}(\lambda)+\frac{1}{\lambda}\int_{\bar{\nu}^{-1}(\lambda)}^\infty \bar{\nu}(t)\,dt\right\} \\
&= \lambda^{1-c}\left[1+\{\log(\omega/\lambda)/\alpha\}^{1/\gamma}+C\{\log(\omega/\lambda)/\alpha\}^{(1-\gamma)/\gamma}\right] \\
&\leq \lambda^{1-c}\left[2\{\log(\omega/\lambda)/\alpha\}^{1/\gamma}+C\{\log(\omega/\lambda)/\alpha\}^{(1-\gamma)/\gamma}\right] \\
&\leq C'\lambda^{1-c}\{\log(\omega/\lambda)/\alpha\}^{1/\gamma}
\end{aligned}$$

since $\gamma > 0$ implies $1/\gamma > (1-\gamma)/\gamma$. Therefore the overall bound is

$$\begin{aligned}
&C'\lambda^{1-c}\{\log(\omega/\lambda)/\alpha\}^{1/\gamma} \wedge \frac{1}{\lambda^c}C\omega m^{1-\gamma}\exp(-\alpha m^\gamma) \\
&\leq C'\omega\left(\frac{\{\log(\omega/\lambda)/\alpha\}^{1/\gamma}}{\lambda^{c-1}} \wedge \frac{m^{1-\gamma}\exp(-\alpha m^\gamma)}{\lambda^c}\right).
\end{aligned}$$

□

Proposition I.5 (Lower bound on effective dimension)**.** Suppose $\nu_s \geq \underline{\nu}(s) : \mathbb{R} \to \mathbb{R}$ for some non-increasing positive function $\underline{\nu}(s)$. Then, if s^* is the smallest positive integer with $\underline{\nu}(s^*) \leq \lambda$,

$$\psi(m,2) \geq \int_{s^*\vee(m+1)}^\infty \frac{\underline{\nu}(s)}{(\underline{\nu}(s)+\lambda)^2}\,ds \geq \frac{1}{4\lambda^2}\int_{s^*\vee(m+1)}^\infty \underline{\nu}(s)\,ds.$$

Moreover, if we have $\underline{\nu}(s) = \omega s^{-\beta}$ for $\beta > 1$ (polynomial decay), then, whenever $\lambda \leq \omega$,

$$\psi(0,2) \geq \frac{1}{c_\beta}\left(\frac{\omega^{1/\beta}}{\lambda^{1+1/\beta}}\right).$$

If $\underline{\nu}(s) = \omega \exp(-\alpha s^\gamma)$ for $\gamma < 1$ (exponential decay), then, whenever $\lambda \leq \omega / e^\alpha$,

$$\psi(0,2) \geq \frac{1}{c_{\alpha,\gamma}} \left(\frac{\log(\omega/\lambda)^{(1-\gamma)/\gamma}}{\lambda} \right).$$

Proof. We proceed in steps.

1. Note that the function $t \mapsto (t+\lambda)^{-2} t$ is strictly increasing for all $t < \lambda$. Thus, for each integer s such that $\underline{\nu}(s) \leq \lambda$ we have

$$\frac{\nu_s}{(\nu_s + \lambda)^2} \geq \int_s^{s+1} \frac{\underline{\nu}(t)}{(\underline{\nu}(t) + \lambda)^2} \, dt.$$

   So we may bound the sum by an integral to show

$$\psi(m,2) = \sum_{s=m+1}^{\infty} \frac{\nu_s}{(\nu_s + \lambda)^2} \geq \int_{s^* \vee (m+1)}^{\infty} \frac{\underline{\nu}(t)}{(\underline{\nu}(t) + \lambda)^2} \, dt \geq \frac{1}{4\lambda^2} \int_{s^* \vee (m+1)}^{\infty} \underline{\nu}(t) \, dt,$$

   where s^* is the smallest positive integer such that $\underline{\nu}(s) \leq \lambda$.

2. Polynomial decay. If $\underline{\nu}(s) = \omega s^{-\beta}$ then $s^* \leq (\lambda/\omega)^{-1/\beta} + 1$. Using the exact integral from Proposition I.2,

$$\psi(0,2) \geq \frac{1}{4\lambda^2} \int_{s^*}^{\infty} \omega s^{-\beta} \, ds \geq \frac{\omega(\beta-1)^{-1}}{4\lambda^2} [s^*]^{1-\beta} \geq \frac{\omega(\beta-1)^{-1}}{4\lambda^2} [(\lambda/\omega)^{-1/\beta} + 1]^{1-\beta}.$$

   Under the condition that $\lambda \leq \omega$, the bracketed term is at most $2(\lambda/\omega)^{-1/\beta}$. This yields

$$\psi(0,2) \geq \frac{(\beta-1)^{-1} \omega^{1/\beta}}{2^{1+\beta} \lambda^{1+1/\beta}} \geq \frac{1}{c_{\beta,\omega}} \left(\frac{1}{\lambda^{1+1/\beta}} \right).$$

3. Exponential decay. If $\underline{\nu}(s) = \omega \exp(-\alpha s^\gamma)$ then we have

$$\{\log(\omega/\lambda)/\alpha\}^{1/\gamma} \leq s^* \leq \{\log(\omega/\lambda)/\alpha\}^{1/\gamma} + 1 \leq 2\{\log(\omega/\lambda)/\alpha\}^{1/\gamma}$$

   following from the condition that $\lambda \leq \omega / e^\alpha$. We can compute

$$4\lambda^2 \cdot \psi(0,2) \geq \int_{s^*}^{\infty} \omega \exp(-\alpha s^\gamma) \, ds.$$

Making the substitution $u = \alpha s^\gamma$ as in Proposition I.2 gives us

$$\begin{aligned}
&= \omega\gamma^{-1}\alpha^{\frac{-1}{\gamma}} \int_{\alpha(s^*)^\gamma}^{\infty} \exp(-u) u^{(1-\gamma)/\gamma}\, du \\
&= \omega\gamma^{-1}\alpha^{\frac{-1}{\gamma}} \Gamma(1/\gamma, \alpha[s^*]^\gamma).
\end{aligned}$$

Using Lemma I.1, which says that $\Gamma(z,x) \geq x^{z-1}e^{-x}$ for $z > 1$, this is

$$\begin{aligned}
&\geq \omega\gamma^{-1}\alpha^{\frac{-1}{\gamma}} (\alpha[s^*]^\gamma)^{1/\gamma - 1} e^{-\alpha[s^*]^\gamma} \\
&= \omega\gamma^{-1}\alpha^{-1}(s^*)^{1-\gamma} \exp(-\alpha[s^*]^\gamma).
\end{aligned}$$

Finally we use the bounds on s^* in to find that this is

$$\begin{aligned}
&\geq \omega\gamma^{-1}\alpha^{-1} \left[\{\log(\omega/\lambda)/\alpha\}^{1/\gamma}\right]^{1-\gamma} \exp(-\alpha[2\{\log(\omega/\lambda)/\alpha\}^{1/\gamma}]^\gamma) \\
&= \omega\gamma^{-1}\alpha^{-1}\{\log(\omega/\lambda)/\alpha\}^{(1/\gamma)-1} \frac{\lambda}{\omega} e^{2^\gamma} \\
&= \lambda\gamma^{-1}\alpha^{-1} e^{2^\gamma} \{\log(\omega/\lambda)/\alpha\}^{(1/\gamma)-1}.
\end{aligned}$$

After rearranging, we obtain

$$\psi(0,2) \geq \frac{1}{c_{\alpha,\gamma}} \left(\frac{\log(\omega/\lambda)^{(1-\gamma)/\gamma}}{\lambda} \right).$$

□

J Bounding key terms for bootstrap couplings (Section D)

The abstract bound is in terms of $\Delta_1 := \|\hat{\Sigma} - \Sigma\|_{\text{HS}}$ and $\Delta_2 := \operatorname{tr} \Pi_m^\perp (\hat{\Sigma} - \Sigma) \Pi_m^\perp$, where

$$\Sigma = \mathbb{E}[U_i \otimes U_i^*], \quad \hat{\Sigma} = \mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}_n[U_i] \otimes (\mathbb{E}_n[U_i])^*.$$

We bound these key quantities under different assumptions. In particular, we prove *unconditional* bounds, i.e., high probability bounds with respect to U.

Assumption J.1 (boundedness)**.** The random variables U_i satisfy $\|U_i\| \leq a$ almost surely.

Assumption J.2 (sub-Gaussianity)**.** For all $f \in H$, the random variables U_i satisfy

$$\mathbb{P}\left\{\langle f,\, U_i\rangle > ub\left(\mathbb{E}\,\langle f,\, U_i\rangle^2\right)^{\frac{1}{2}}\right\} \leq 2e^{-u^2}.$$

J.1 Boundedness

Lemma J.3. Under Assumption J.1, the following event holds with probability at least $1-2\eta$:

$$\|\hat{\Sigma}-\Sigma\|_{\mathrm{HS}} \leq 2\log(2/\eta)^2\left\{\sqrt{\frac{a^2\sigma^2(0)}{n}} \vee \frac{4a^2}{n} \vee \frac{8a^2}{n^2}\right\}.$$

Proof. We proceed in steps.

1. Decomposition. Write

$$\begin{aligned}\hat{\Sigma}-\Sigma &= \mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}_n[U_i]\otimes(\mathbb{E}_n[U_i])^* - \Sigma \\ &= \left\{\mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}[U_i \otimes U_i^*]\right\} - \left\{\mathbb{E}_n[U_i]\otimes(\mathbb{E}_n[U_i])^*\right\}.\end{aligned}$$

2. Focusing on the former term, write $\Sigma_i = U_i \otimes U_i^*$. Since $\|\Sigma_i\|_{\mathrm{HS}} = \|U_i\|^2 \leq a^2$ and

$$\mathbb{E}\|\Sigma_i\|_{\mathrm{HS}}^2 = \int \mathrm{tr}(\Sigma_i^*\Sigma_i)\mathrm{d}\mathbb{P} \leq \int \|\Sigma_i\|_{op}\mathrm{tr}(\Sigma_i)\mathrm{d}\mathbb{P} \leq a^2\,\mathrm{tr}(\Sigma) = a^2\sigma^2(0),$$

by Lemma H.2, with probability $1-\eta$,

$$\|\mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}[U_i \otimes U_i^*]\|_{\mathrm{HS}} \leq 2\log(2/\eta)\left\{\sqrt{\frac{a^2\sigma^2(0)}{n}} \vee \frac{2a^2}{n}\right\}.$$

3. Focusing on the latter term, write

$$\|\mathbb{E}_n[U_i]\otimes(\mathbb{E}_n[U_i])^*\|_{\mathrm{HS}} = \|\mathbb{E}_n[U_i]\|^2 = \|\mathbb{E}_n[U_i]-0\|^2.$$

Notice that $\|U_i\| \leq a$ and $\mathbb{E}\|U_i\|^2 \leq a^2$ so by Lemma H.2, with probability $1-\eta$,

$$\|\mathbb{E}_n[U_i]-0\| \leq 2\log(2/\eta)\left\{\sqrt{\frac{a^2}{n}} \vee \frac{2a}{n}\right\}$$

and therefore

$$\|\mathbb{E}_n[U_i] \otimes (\mathbb{E}_n[U_i])^*\|_{\mathrm{HS}} \leq 4\log(2/\eta)^2 \left\{ \frac{a^2}{n} \vee \frac{4a^2}{n^2} \right\} = 2\log(2/\eta)^2 \left\{ \frac{2a^2}{n} \vee \frac{8a^2}{n^2} \right\}.$$

4. In summary

$$\begin{aligned}\|\hat{\Sigma} - \Sigma\|_{\mathrm{HS}} &\leq \|\mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}[U_i \otimes U_i^*]\|_{\mathrm{HS}} + \|\mathbb{E}_n[U_i] \otimes (\mathbb{E}_n[U_i])^*\|_{\mathrm{HS}} \\ &\leq 2\log(2/\eta) \left\{ \sqrt{\frac{a^2\sigma^2(0)}{n}} \vee \frac{2a^2}{n} \right\} + 2\log(2/\eta)^2 \left\{ \frac{2a^2}{n} \vee \frac{8a^2}{n^2} \right\}.\end{aligned}$$

□

Lemma J.4. Under Assumption J.1, the following event holds with probability at least $1 - \eta$:

$$\operatorname{tr} \Pi_m^\perp (\hat{\Sigma} - \Sigma) \Pi_m^\perp \leq 2\log(2/\eta) \left(\sqrt{\frac{a^2\sigma^2(m)}{n}} \vee \frac{2a^2}{n} \right).$$

Proof. We proceed in steps.

1. Decomposition. As before

$$\Pi_m^\perp (\hat{\Sigma} - \Sigma) \Pi_m^\perp = \Pi_m^\perp \left\{ \mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}[U_i \otimes U_i^*] \right\} \Pi_m^\perp - \Pi_m^\perp \left\{ \mathbb{E}_n[U_i] \otimes (\mathbb{E}_n[U_i])^* \right\} \Pi_m^\perp.$$

2. Focusing on the former term, notice that

$$\operatorname{tr}[\Pi_m^\perp \left\{ \mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}[U_i \otimes U_i^*] \right\} \Pi_m^\perp] = \mathbb{E}_n \xi_i - \mathbb{E}\xi_i, \quad \xi_i = \operatorname{tr} \Pi_m^\perp \Sigma_i \Pi_m^\perp.$$

Then we have

$$\begin{aligned}\xi_i &= \operatorname{tr} \Pi_m^\perp \Sigma_i \Pi_m^\perp = \operatorname{tr}\{\Pi_m^\perp U_i \otimes U_i^* \Pi_m^\perp\} = \|\Pi_m^\perp U_i\|^2 \leq \|U_i\|^2 \leq a^2, \\ \mathbb{E}\xi_i^2 &\leq a^2 \mathbb{E}\xi_i = a^2 \int \operatorname{tr} \Pi_m^\perp \Sigma_i \Pi_m^\perp \mathrm{d}\mathbb{P} = a^2 \int \operatorname{tr} \Pi_m^\perp \Sigma_i \mathrm{d}\mathbb{P} = a^2 \operatorname{tr} \Pi_m^\perp \Sigma = a^2 \sigma^2(m),\end{aligned}$$

so by Lemma H.2, with probability $1 - \eta$,

$$\operatorname{tr}[\Pi_m^\perp \left\{ \mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}[U_i \otimes U_i^*] \right\} \Pi_m^\perp] \leq 2\log(2/\eta) \left\{ \sqrt{\frac{a^2\sigma^2(m)}{n}} \vee \frac{2a^2}{n} \right\}.$$

3. Focusing on the latter term, notice that

$$\operatorname{tr}[\Pi_m^{\perp}\{\mathbb{E}_n[U_i]\otimes(\mathbb{E}_n[U_i])^*\}\Pi_m^{\perp}] = \|\Pi_m^{\perp}\mathbb{E}_n[U_i]\|^2 \geq 0.$$

4. In summary

$$\begin{aligned}\operatorname{tr}\Pi_m^{\perp}(\hat{\Sigma}-\Sigma)\Pi_m^{\perp} &= \operatorname{tr}[\Pi_m^{\perp}\{\mathbb{E}_n[U_i\otimes U_i^*]-\mathbb{E}[U_i\otimes U_i^*]\}\Pi_m^{\perp}] \\ &\quad - \operatorname{tr}[\Pi_m^{\perp}\{\mathbb{E}_n[U_i]\otimes(\mathbb{E}_n[U_i])^*\}\Pi_m^{\perp}] \\ &\leq \operatorname{tr}[\Pi_m^{\perp}\{\mathbb{E}_n[U_i\otimes U_i^*]-\mathbb{E}[U_i\otimes U_i^*]\}\Pi_m^{\perp}] \\ &\leq 2\log(2/\eta)\left\{\sqrt{\frac{a^2\sigma^2(m)}{n}}\vee\frac{2a^2}{n}\right\}.\end{aligned}$$

□

J.2 Sub-Gaussianity

The sub-Gaussian case is more involved. We quote some helpful results before bounding the key terms.

J.2.1 Preliminaries

Lemma J.5 (Dirksen (2015, Theorem 3.2))**.** Let $(X_t)_{t\in T}$ denote a real-valued stochastic process on the separable metric space (T,d), whose increments are sub-Gaussian with respect to d, i.e. $\mathbb{P}\left(|X_s-X_t|> ud(s,t)\right)\leq 2e^{-u^2}$. Define the γ functional

$$\gamma(T,d)=\inf_S\sup_{t\in T}\sum_{n=0}^{\infty}2^{n/2}d(S_n,t),$$

where the infimum is over sequences $S=(S_0,S_1,\ldots)$ of subsets of T satisfying the growth condition $\#S_0=1$, $\#S_n\leq 2^{2^n}$, and $d(S_n,t)\coloneqq\inf_{s\in S_n}d(s,t)$. Then, there exist universal constants C,D such that the following inequalities hold, for any $t_0\in T$:

$$\left(\mathbb{E}\sup_{t\in T}|X_t-X_{t_0}|^p\right)^{\frac{1}{p}}\leq C\gamma(T,d)+2\sup_{t\in T}\left(\mathbb{E}|X_t-X_{t_0}|^p\right)^{\frac{1}{p}}.$$

Lemma J.6 (Talagrand (2014, Theorem 2.4.1))**.** Let $(Z_t)_{t\in T}$ be a Gaussian process on the separable metric space (T,d), where $d(s,t)=(\mathbb{E}|Z_s-Z_t|^2)^{\frac{1}{2}}$. Then there exists a universal constant $C>0$ such that $C^{-1}\gamma(T,d)\leq\mathbb{E}\sup_{t\in T}|Z_t|\leq C\gamma(T,d)$.

J.2.2 Bounding moments

Lemma J.7. Under Assumption J.2, there exists a universal constant C such that $(\mathbb{E}\,\|U_i\|^p)^{\frac{1}{p}} \leq Cb\sqrt{p}\sigma(0)$.

Proof. We proceed in steps.

1. Matching symbols. We consider the process $X_t = \langle U_i,\, t\rangle$ indexed by $t \in B_H$. Note in particular that if $d(s,t) = \langle (s-t),\, b^2C(s-t)\rangle^{\frac{1}{2}}$ then $(X_t)_{t\in B_H}$ satisfies the sub-Gaussian increments condition with respect to d:

$$\mathbb{P}\left(\langle s-t,\, U_i\rangle > ub\left(\mathbb{E}\,\langle s-t,\, U_i\rangle^2\right)^{\frac{1}{2}}\right) \leq 2e^{-u^2}$$

implies $\mathbb{P}\left(|X_s - X_t| > ud(s,t)\right) \leq 2e^{-u^2}$ when

$$d(s,t) = b\left(\mathbb{E}\,\langle s-t,\, U_i\rangle^2\right)^{\frac{1}{2}} = \left\langle (s-t),\, b^2C(s-t)\right\rangle^{\frac{1}{2}}.$$

Note also that $\sup_{t\in B_H}|X_t| = \sup_{t\in B}\langle U_i,\, t\rangle = \|U_i\|$. Applying Lemma J.5 to $(X_t)_{t\in B_H}$ with $t_0 = 0$ therefore yields

$$(\mathbb{E}\,\|U_i\|^p)^{\frac{1}{p}} \leq C\gamma(B_H, d) + 2\sup_{t\in B_H}\left(\mathbb{E}|\langle t,\, U_i\rangle|^p\right)^{\frac{1}{p}}.$$

2. First term. To bound $\gamma(B_H, d)$, note that if $Z_t = b\langle \Sigma^{\frac{1}{2}}g,\, t\rangle$ so that $\mathbb{E}|Z_s - Z_t|^2 = d(s,t)^2$ by construction, it holds by Lemma J.6 that

$$\gamma(B_H, d) \leq C\mathbb{E}\sup_{t\in B_H}|b\langle \Sigma^{\frac{1}{2}}g,\, t\rangle| = bC\mathbb{E}\|\Sigma^{\frac{1}{2}}g\| \leq bC\left(\mathbb{E}\|\Sigma^{\frac{1}{2}}g\|^2\right)^{\frac{1}{2}} = bC\sigma(0).$$

In particular

$$\mathbb{E}\|\Sigma^{\frac{1}{2}}g\|^2 = \mathbb{E}[\operatorname{tr} g^*\Sigma g] = \mathbb{E}[\operatorname{tr}\Sigma g\otimes g^*] = \operatorname{tr}\Sigma = \sigma^2(0).$$

3. Second term. What remains is control of the latter term $\sup_{t\in B_H}\left(\mathbb{E}|\langle t,\, U_i\rangle|^p\right)^{\frac{1}{p}}$. Initially fix t. To begin, we argue that $\langle t,\, U_i\rangle$ is sub-Gaussian with $\|\langle U_i,\, t\rangle\|_{\psi_2} \leq Cb\,\langle t,\, \Sigma t\rangle^{\frac{1}{2}}$. To see why, notice that in our sub-Gaussian assumption, $b\left(\mathbb{E}\,\langle t,\, U_i\rangle^2\right)^{\frac{1}{2}} = b\,\langle t,\, \Sigma t\rangle^{\frac{1}{2}}$. Therefore we are effectively assuming

$$\left\|\frac{\langle t,\, U_i\rangle}{b\,\langle t,\, \Sigma t\rangle^{\frac{1}{2}}}\right\|_{\psi_2} \leq C \iff \|\langle t,\, U_i\rangle\|_{\psi_2} \leq Cb\,\langle t,\, \Sigma t\rangle^{\frac{1}{2}}.$$

Since $\langle t,\, U_i\rangle$ is sub-Gaussian, we can appeal to Vershynin (2018, Proposition 2.5.2) to control its moments. In particular,

$$(\mathbb{E}|\langle t,\, U_i\rangle\,|^p)^{\frac{1}{p}} \leq C\sqrt{p}\|\langle t,\, U_i\rangle\,\|_{\psi_2} \leq Cb\sqrt{p}\,\langle t,\, \Sigma t\rangle^{\frac{1}{2}}\,.$$

Taking the supremum over $t \in B_H$,

$$\sup_{t\in B_H}(\mathbb{E}|\langle t,\, U_i\rangle\,|^p)^{\frac{1}{p}} \leq Cb\sqrt{p}\sup_{t\in B_H}\langle t,\, \Sigma t\rangle^{\frac{1}{2}} = Cb\sqrt{p}\nu_1^{1/2} \leq Cb\sqrt{p}\sigma(0).$$

4. Collecting results. In summary,

$$(\mathbb{E}\,\|U_i\|^p)^{\frac{1}{p}} \leq C\gamma(B_H, d) + 2\sup_{t\in B_H}(\mathbb{E}|\langle t,\, U_i\rangle\,|^p)^{\frac{1}{p}} \leq bC\sigma(0) + Cb\sqrt{p}\sigma(0) = Cb\sqrt{p}\sigma(0).$$

□

Lemma J.8. Under Assumption J.2, there exists a universal constant C such that $\left(\mathbb{E}\left\|\Pi_m^{\perp}U_i\right\|^p\right)^{\frac{1}{p}} \leq Cb\sqrt{p}\sigma(m)$.

Proof. We use as identical proof to Lemma J.7, replacing U_i with $\Pi_m^{\perp}U_i$. In particular, we consider the process $X_t = \left\langle \Pi_m^{\perp}U_i,\, t\right\rangle = \left\langle U_i,\, \Pi_m^{\perp}t\right\rangle$ which corresponds restricting the index set to $t \in \Pi_m^{\perp}B_H$. Since this process is simply a restriction of the original process, the increment condition is again satisfied, and it suffices to plug in the improved bounds

$$\gamma(\Pi_m^{\perp}B_H, d) \leq bC\mathbb{E}\left\|\Pi_m^{\perp}\Sigma^{\frac{1}{2}}g\right\| \leq bC\sigma^2(m),$$
$$\sup_{t\in\Pi_m^{\perp}B_H}\|\langle U_i,\, t\rangle\|_{\psi_2} \leq Cb\sup_{t\in B_H}\left\langle \Pi_m^{\perp}t,\, \Sigma\Pi_m^{\perp}t\right\rangle^{\frac{1}{2}} = Cb\sqrt{\nu_m} \leq b\sigma(m).$$

□

J.2.3 Main result

Lemma J.9. Under Assumption J.2, the following event holds with probability at least $1-2\eta$, $\|\hat{\Sigma}-\Sigma\|_{\mathrm{HS}} \leq C\log(2/\eta)^2\frac{b^2\sigma^2(0)}{\sqrt{n}}$.

Proof. We proceed in steps.

1. Decomposition. As before, write

$$\begin{aligned}\hat{\Sigma} - \Sigma &= \mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}_n[U_i] \otimes (\mathbb{E}_n[U_i])^* - \Sigma \\ &= \{\mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}[U_i \otimes U_i^*]\} - \{\mathbb{E}_n[U_i] \otimes (\mathbb{E}_n[U_i])^*\}.\end{aligned}$$

2. Focusing on the first term, write $\Sigma_i = U_i \otimes U_i^*$. Notice that

$$\|\|\Sigma_i - \Sigma\|_{\text{HS}}\|_p \leq \|\|\Sigma_i\|_{\text{HS}} + \|\Sigma\|_{\text{HS}}\|_p \leq \|\|\Sigma_i\|_{\text{HS}}\|_p + \|\|\Sigma\|_{\text{HS}}\|_p.$$

   Note that the former term is $\|\|\Sigma_i\|_{\text{HS}}\|_p^p = \mathbb{E}(\|\Sigma_i\|_{\text{HS}}^p)$. Within the latter term $\|\Sigma\|_{\text{HS}} = \|\mathbb{E}(\Sigma_i)\|_{HS} \leq \mathbb{E}(\|\Sigma_i\|_{\text{HS}})$ hence

$$\|\|\Sigma\|_{\text{HS}}\|_p^p \leq \mathbb{E}[\{\mathbb{E}(\|\Sigma_i\|_{\text{HS}})\}^p] \leq \mathbb{E}\{\mathbb{E}(\|\Sigma_i\|_{\text{HS}}^p)\} = \mathbb{E}(\|\Sigma_i\|_{\text{HS}}^p).$$

   In summary $\|\|\Sigma_i - \Sigma\|_{\text{HS}}\|_p \leq 2\|\|\Sigma_i\|_{\text{HS}}\|_p$. Finally, we have $\|\Sigma_i\|_{\text{HS}}^p = \|U_i\|^{2p}$. Thus, we can estimate using Lemma J.7 that

$$\mathbb{E}\,\|\Sigma_i - \Sigma\|_{\text{HS}}^p \leq 2^p \mathbb{E}(\|U_i\|^{2p}) \leq (2b^2C^2)^p (\sqrt{2p})^{2p} \sigma^{2p}(0)$$

   since $\mathbb{E}\,\|U_i\|^q \leq \{Cb\sqrt{q}\sigma(0)\}^q$ where $q = 2p$.

3. By Stirling's approximation, we have $p! \geq (p/e)^p$, so $(2e)^p p! \geq (2p)^p = (\sqrt{2p})^{2p}$. Hence

$$\mathbb{E}\,\|\Sigma_i - \Sigma\|_{\text{HS}}^p \leq (2b^2C^2)^p (2e)^p p!\, \sigma^{2p}(0) = (2^{1/p} 4eb^2C^2)^p \frac{1}{2} p!\, \sigma^{2p}(0) = (C'')^p \frac{1}{2} p!\, \sigma^{2p}(0)$$

   and therefore $\mathbb{E}\left\|\frac{\Sigma_i - \Sigma}{C''}\right\|_{\text{HS}}^p \leq \frac{1}{2} p!\,\{\sigma^2(0)\}^p$ where $C'' = 2^{1/p} 4eb^2C^2$. Therefore by applying Lemma H.2 with $A = B = \sigma^2(0)$, we see that with probability at least $1 - \eta$,

$$\begin{aligned}\|\mathbb{E}_n[\Sigma_i] - \Sigma\|_{\text{HS}} &= C'' \left\| \frac{\mathbb{E}_n[\Sigma_i] - \Sigma}{C''} \right\|_{\text{HS}} \\ &\leq C'' 2\ln(2/\eta) \left\{ \frac{\sigma^2(0)}{n} \vee \sqrt{\frac{\sigma^4(0)}{n}} \right\} \\ &\leq 2C'' \ln(2/\eta) \frac{\sigma^2(0)}{\sqrt{n}} \\ &\leq 16eb^2C^2 \ln(2/\eta) \frac{\sigma^2(0)}{\sqrt{n}}.\end{aligned}$$

4. Turning to the second term, write

$$\|\mathbb{E}_n[U_i] \otimes (\mathbb{E}_n[U_i])^*\|_{\text{HS}} = \|\mathbb{E}_n[U_i]\|^2 = \|\mathbb{E}_n[U_i] - 0\|^2.$$

By Lemma J.7,

$$\mathbb{E}\|U_i\|^p \leq \{Cb\sqrt{p}\sigma(0)\}^p = (bC)^p \sqrt{p}^p \{\sigma(0)\}^p.$$

By Stirling's approximation, $p! \geq (p/e)^p$ so $\sqrt{p}^p \leq (e^p p!\,)^{1/2} \leq e^{p/2} p!$. Therefore

$$\mathbb{E}\|U_i\|^p \leq (bCe^{1/2})^p p! \{\sigma(0)\}^p = (C')^p \frac{1}{2} p! \{\sigma(0)\}^p, \quad C' = 2^{1/p} bCe^{1/2}.$$

Applying Lemma H.2 with $A = B = \sigma(0)$, with probability at least $1 - \eta$,

$$\begin{aligned}
\|\mathbb{E}_n[U_i]\| &= C' \left\| \frac{\mathbb{E}_n[U_i]}{C'} \right\| \\
&\leq C' 2 \ln(2/\eta) \left\{ \frac{\sigma(0)}{n} \vee \sqrt{\frac{\sigma^2(0)}{n}} \right\} \\
&\leq 2C' \ln(2/\eta) \frac{\sigma(0)}{\sqrt{n}} \\
&\leq 4bCe^{1/2} \ln(2/\eta) \frac{\sigma(0)}{\sqrt{n}}.
\end{aligned}$$

Therefore

$$\|\mathbb{E}_n[U_i] \otimes (\mathbb{E}_n[U_i])^*\|_{\text{HS}} \leq 16 b^2 C^2 e \ln(2/\eta)^2 \frac{\sigma^2(0)}{n}.$$

5. In summary

$$\begin{aligned}
\|\hat{\Sigma} - \Sigma\|_{\text{HS}} &\leq \|\mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}[U_i \otimes U_i^*]\|_{\text{HS}} + \|\mathbb{E}_n[U_i] \otimes (\mathbb{E}_n[U_i])^*\|_{\text{HS}} \\
&\leq 16 e b^2 C^2 \ln(2/\eta) \frac{\sigma^2(0)}{\sqrt{n}} + 16 b^2 C^2 e \ln(2/\eta)^2 \frac{\sigma^2(0)}{n}.
\end{aligned}$$

□

Lemma J.10. Under Assumption J.2, the following event holds with probability at least $1 - \eta$: $\operatorname{tr} \Pi_m^\perp (\hat{\Sigma} - \Sigma) \Pi_m^\perp \leq C \log(2/\eta) \frac{b^2 \sigma^2(m)}{\sqrt{n}}$.

Proof. We proceed in steps.

1. Decomposition. As before

$$\Pi_m^\perp(\hat{\Sigma}-\Sigma)\Pi_m^\perp = \Pi_m^\perp \left\{\mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}[U_i \otimes U_i^*]\right\} \Pi_m^\perp - \Pi_m^\perp \left\{\mathbb{E}_n[U_i] \otimes (\mathbb{E}_n[U_i])^*\right\} \Pi_m^\perp.$$

2. Focusing on the first term, notice that

$$\operatorname{tr}[\Pi_m^\perp \left\{\mathbb{E}_n[U_i \otimes U_i^*] - \mathbb{E}[U_i \otimes U_i^*]\right\} \Pi_m^\perp] = \mathbb{E}_n\xi_i - \mathbb{E}\xi_i, \quad \xi_i = \operatorname{tr} \Pi_m^\perp \Sigma_i \Pi_m^\perp.$$

Moreover $\|\xi_i - \mathbb{E}\xi_i\|_p \leq \|\xi_i\|_p + \|\mathbb{E}\xi_i\|_p$. The former term is $\|\xi_i\|_p^p = \mathbb{E}(|\xi_i|^p)$. Within the latter term

$$\|\mathbb{E}\xi_i\|_p^p = \mathbb{E}[|\mathbb{E}\xi_i|^p] \leq \mathbb{E}[\mathbb{E}(|\xi_i|^p)] = \mathbb{E}(|\xi_i|^p).$$

In summary $\|\xi_i - \mathbb{E}\xi_i\|_p \leq 2\|\xi_i\|_p$. Then

$$\xi_i = \operatorname{tr} \Pi_m^\perp \Sigma_i \Pi_m^\perp = \operatorname{tr}\{\Pi_m^\perp U_i \otimes U_i^* \Pi_m^\perp\} = \|\Pi_m^\perp U_i\|^2.$$

By Lemma J.8,

$$\mathbb{E}(|\xi_i - \mathbb{E}\xi_i|^p) \leq 2^p \mathbb{E}(|\xi_i|^p) = 2^p \mathbb{E}(\|\Pi_m^\perp U_i\|^{2p}) \leq 2^p \{Cb\sqrt{2p}\sigma(m)\}^{2p}.$$

3. By Stirling's approximation, $p! \geq (p/e)^p$, so $(2e)^p p! \geq (2p)^p = (\sqrt{2p})^{2p}$. Hence

$$\begin{aligned}
\mathbb{E}(|\xi_i|^p) &\leq 2^p \{Cb\}^{2p} (2e)^p p! \left\{\sigma^2(m)\right\}^p \\
&= (2^{1/p} 4eb^2C^2)^p \frac{1}{2} p! \left\{\sigma^2(m)\right\}^p \\
&= (C''')^p \frac{1}{2} p! \left\{\sigma^2(m)\right\}^p
\end{aligned}$$

and therefore

$$\mathbb{E}\left(\left|\frac{\xi_i}{C'''}\right|^p\right) \leq \frac{1}{2} p! \left\{\sigma^2(m)\right\}^p, \quad C''' = 2^{1/p} 4eb^2C^2.$$

Applying Lemma H.2 with $A = B = \sigma^2(m)$, we see that with probability $1 - \eta$

$$\begin{aligned}
|\mathbb{E}_n\xi_i - \mathbb{E}\xi_i| &= C''' \left|\frac{\mathbb{E}_n\xi_i - \mathbb{E}\xi_i}{C'''}\right| \\
&\leq C'''2\ln(2/\eta) \left\{\frac{\sigma^2(m)}{n} \vee \sqrt{\frac{\sigma^4(m)}{n}}\right\} \\
&\leq 2C''' \ln(2/\eta) \frac{\sigma^2(m)}{\sqrt{n}}. \\
&\leq 16eb^2C^2 \ln(2/\eta) \frac{\sigma^2(m)}{\sqrt{n}}.
\end{aligned}$$

4. Focusing on the second term, notice that

$$\operatorname{tr}[\Pi_m^{\perp}\{\mathbb{E}_n[U_i]\otimes(\mathbb{E}_n[U_i])^*\}\Pi_m^{\perp}] = \|\Pi_m^{\perp}\mathbb{E}_n[U_i]\|^2 \geq 0.$$

5. In summary

$$\begin{aligned}\operatorname{tr}\Pi_m^{\perp}(\hat{\Sigma}-\Sigma)\Pi_m^{\perp} &= \operatorname{tr}[\Pi_m^{\perp}\{\mathbb{E}_n[U_i\otimes U_i^*]-\mathbb{E}[U_i\otimes U_i^*]\}\Pi_m^{\perp}] \\ &\quad - \operatorname{tr}[\Pi_m^{\perp}\{\mathbb{E}_n[U_i]\otimes(\mathbb{E}_n[U_i])^*\}\Pi_m^{\perp}] \\ &\leq \operatorname{tr}[\Pi_m^{\perp}\{\mathbb{E}_n[U_i\otimes U_i^*]-\mathbb{E}[U_i\otimes U_i^*]\}\Pi_m^{\perp}] \\ &\leq 16eb^2C^2\ln(2/\eta)\frac{\sigma^2(m)}{\sqrt{n}}.\end{aligned}$$

□

K Simulation details

K.1 Robust performance in Sobolev spaces

In this appendix, we present additional results in Sobolev spaces. The Sobolev space corresponds to the Matern kernel. We verify nominal coverage across smoothness degrees, sample sizes, and regularization values. We document additional metrics such as bias and width of the confidence bands.

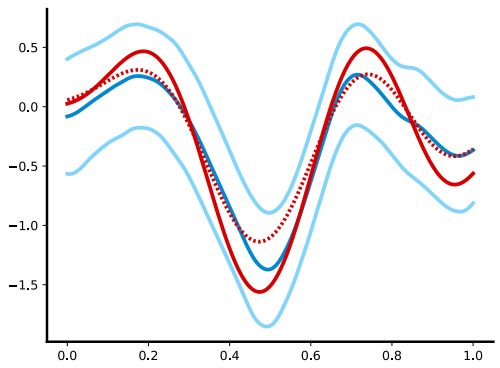


Figure 6: Regression in $\mathbb{H}_2^1$ with standard data.

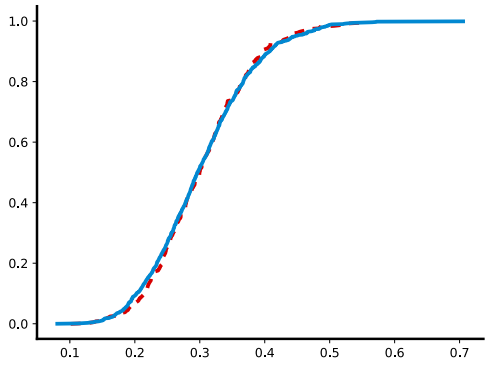


Figure 7: Our approach succeeds in $\mathbb{H}_2^1$.

sample	sup norm		H norm	
	true	pseudo	true	pseudo
50	0.942	0.946	0.938	0.944
100	0.968	0.972	0.934	0.940
500	0.928	0.924	0.916	0.926
1000	0.970	0.970	0.955	0.960

Table 12: Coverage is nominal across sample sizes in $\mathbb{H}_2^1$. Across rows, we vary n and set $\lambda = n^{-1/3}$, following Section 6.

reg.	sup norm		H norm	
	true	pseudo	true	pseudo
0.500	0.784	0.920	0.750	0.918
0.100	0.938	0.942	0.934	0.942
0.050	0.932	0.932	0.920	0.924
0.010	0.942	0.942	0.954	0.954
0.005	0.934	0.934	0.914	0.914
0.001	0.956	0.956	0.922	0.922

Table 13: Coverage is nominal across regularization values in $\mathbb{H}_2^1$. Across rows, we fix $n = 500$ and vary λ across a reasonable range.

metric	sup norm		H norm	
	true	pseudo	true	pseudo
coverage	0.952	0.954	0.942	0.946
bias	0.019	0.000	0.022	0.000
width	0.878	0.878	0.889	0.889

Table 14: A detailed look: Coverage, bias, and width in $\mathbb{H}_2^1$. Across rows, we fix $n = 500$ and set $\lambda = n^{-1/3}$. We examine additional metrics in addition to coverage.

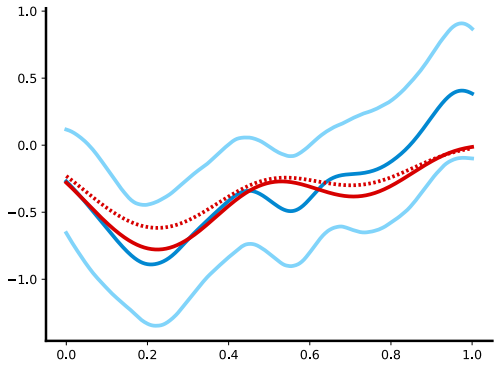


Figure 8: Regression in $\mathbb{H}_2^2$ with standard data.

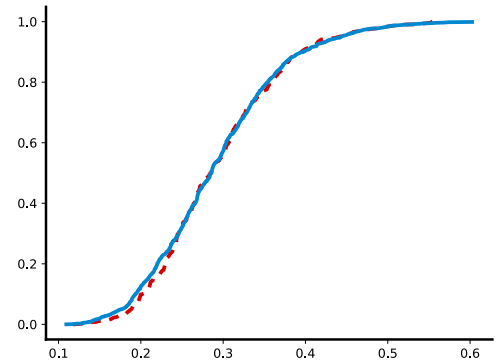


Figure 9: Our approach succeeds in $\mathbb{H}_2^2$.

sample	sup norm		H norm	
	true	pseudo	true	pseudo
50	0.884	0.878	0.746	0.752
100	0.942	0.940	0.940	0.946
500	0.976	0.972	0.980	0.978
1000	0.980	0.980	0.970	0.975

Table 15: Coverage is nominal across sample sizes in $\mathbb{H}_2^2$. Across rows, we vary n and set $\lambda = n^{-1/3}$, following Section 6.

reg.	sup norm		H norm	
	true	pseudo	true	pseudo
0.500	0.884	0.970	0.810	0.970
0.100	0.930	0.940	0.930	0.942
0.050	0.948	0.948	0.940	0.940
0.010	0.964	0.966	0.940	0.940
0.005	0.950	0.948	0.958	0.958
0.001	0.928	0.928	0.970	0.970

Table 16: Coverage is nominal across regularization values in $\mathbb{H}_2^2$. Across rows, we fix $n = 500$ and vary λ across a reasonable range.

metric	sup norm		H norm	
	true	pseudo	true	pseudo
coverage	0.938	0.944	0.944	0.944
bias	0.017	0.000	0.024	0.000
width	0.855	0.855	0.872	0.872

Table 17: A detailed look: Coverage, bias, and width in $\mathbb{H}_2^2$. Across rows, we fix $n =$ and set $\lambda = n^{-1/3}$. We examine additional metrics in addition to coverage.

K.2 Pseudo true coverage under mis-specification

In this appendix, we present additional results for the mis-specified Gaussian kernel. While coverage for the true parameter $f_0 \notin H$ breaks down, coverage for the pseudo true parameter $f_\lambda \in H$ remains nominal across sample sizes and regularization values. These simulations showcase how our inferential theory delivers meaningful guarantees without any assumptions on the bias $f_\lambda - f_0$. Again, we document additional metrics such as bias and width of the confidence bands.

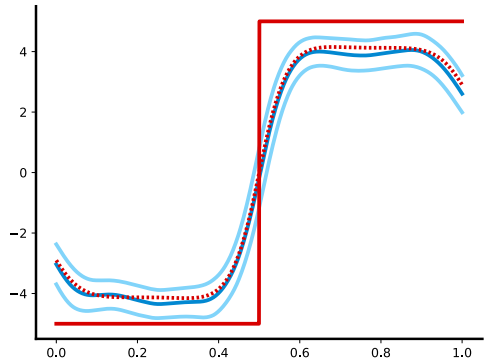


Figure 10: Standard data with mis-specification

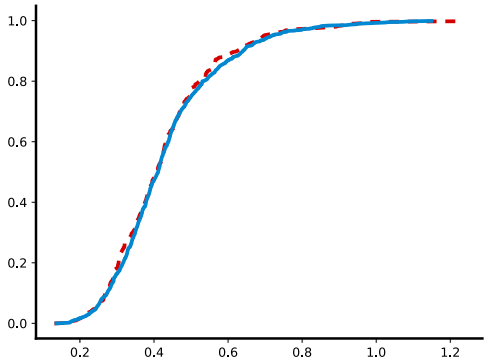


Figure 11: Our approach succeeds with mis-specification

sample	sup norm		H norm	
	true	pseudo	true	pseudo
50	0.000	0.974	0.000	0.966
100	0.000	0.946	0.000	0.952
500	0.000	0.948	0.000	0.956
1000	0.000	0.955	0.000	0.950

Table 18: Coverage is nominal across sample sizes with mis-specification. Across rows, we vary n and set $\lambda = n^{-1/3}$, following Section 6.

reg.	sup norm		H norm	
	true	pseudo	true	pseudo
0.500	0.000	0.952	0.000	0.946
0.100	0.000	0.956	0.000	0.964
0.050	0.000	0.960	0.000	0.932
0.010	0.000	0.948	0.000	0.944
0.005	0.000	0.960	0.000	0.960
0.001	0.000	0.942	0.000	0.954

Table 19: Coverage is nominal across regularization values with mis-specification. Across rows, we fix $n = 500$ and vary λ across a reasonable range.

metric	sup norm		H norm	
	true	pseudo	true	pseudo
coverage	0.000	0.960	0.000	0.976
bias	4.973	0.000	NA	0.000
width	1.366	1.366	1.976	1.976

Table 20: A detailed look: Coverage, bias, and width with mis-specification. Across rows, we fix $n = 500$ and set $\lambda = n^{-1/3}$. We examine additional metrics in addition to coverage.

K.3 Implementation details

In Section 4.2, each observation is generated as follows. Draw $X_i \sim \text{Unif}([0, 1])$ and $\varepsilon_i \sim \text{Unif}([-2, 2])$ independently for each $1 \leq i \leq n$. Then set $Y_i = f_0(X_i) + \varepsilon_i$ where

$$f_0 = \frac{1}{\sqrt{5}} \sum_{i=1}^{5} e_i g_i,$$

where e_i denote the eigenfunctions of the integral operator $\mathbb{E}[k_{X_i} k_{X_i}^*]$, and the g_i are independent standard normal multipliers. We implement Estimator 4.1 with the

Gaussian kernel $k(x, y) = \exp\{-\|x - y\|/(2\iota^2)\}$, with the scale ι set to 0.1.

In Section 4.3, each observation is generated as in section 4.2, with the following changes. The covariates X_i are a uniform random permutation of $\{1, 2, \cdots, 7\}$, and the kernel k is chosen to be the Mallows kernel $k(\pi, \pi') = \exp\{-\tau(\pi, \pi')/(2\iota^2)\}$; here τ is Kendall's-τ function which counts the number of discordant pairs. We employ the standard heuristic and choose ι to be the median of the $\tau(X_i, X_j)$ for independent draws of the data.

In Section K.1, each observation is generated as in Section 4.2, replacing f_0 with the third eigenfunction of the Matern kernel. We implement Estimator 4.1 with the length-scale parameter set to 0.1.

In Section K.2, each observation is generated as in Section 4.2, replacing f_0 with the step function $f_0(x) = \mathbb{1}\{x \geq 1/2\}$. We implement Estimator 4.1, with the length-scale parameter set to 0.1.

L Application details

L.1 Random utility model

In order to evaluate our proposed direct approach, we generate synthetic data which resemble real student preferences using the discrete choice model estimated and reported in Pathak and Shi (2021). In that paper, the authors fit several common discrete choice models to students' reported preference data from the Boston Public Schools centralized match, and then assess each model's ability to forecast demand. As Boston Public Schools used a student-proposing variant of the deferred acceptance (DA) algorithm of Gale and Shapley (1962), strategy-proofness holds and students can reasonably be assumed to report preferences truthfully (Pathak and Sönmez, 2008).

In the random utility model of Pathak and Shi (2021), the utility of student i at school s is

$$u_{is} = \hat{\alpha}_s + \hat{\beta}_1^\top X_1 + \beta_2^\top X_2 + \epsilon_i.$$

In this model, X_1 and X_2 are student-school covariates such as student race, school demographics, walking distance to the school, student English language learner status ,and the availability of English instructional programs in the student's language. The vector $\hat{\beta}_1$ contains estimated coefficients, while $\beta_2 \sim N(\hat{\mu}_2, \hat{\Sigma}_2)$ are random coefficients fit using the mixed multinomial logit procedure outlined in the paper. Finally, $\hat{\alpha}_s$ is a school fixed-effect, and ϵ_i is an independent standard Gumbel random variable.

For the scenario of no match effects, we generate student preferences as described above. We draw school effects as

$$Y_{is} = \mu + \sigma\delta_i + \gamma_1 P_s,$$

where δ_i is a standard normal random variable. We truncate Y_{is} to lie in $[0, 100]$.

For the scenario of match effects, we generate student preferences with an additional taste parameter:

$$u_{is} = \hat{\alpha}_s + \hat{\beta}_1^\top X_1 + \beta_2^\top X_2 + \chi_i P_s \epsilon_i,$$

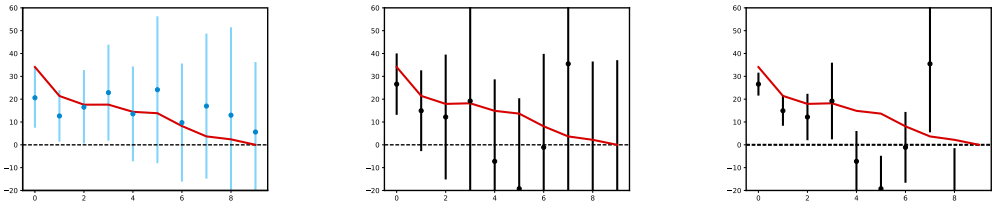
where P_s indicated whether s is a pilot school and χ_i represents preference for pilot sector schools. Here, χ_i determines sector effects: we generate counterfactual outcomes by calculating and then truncating

$$Y_{is} = \mu + \sigma\delta_i + \gamma\chi_i P_s,$$

so students who prefer pilot sector schools also benefit more from them.

L.2 Comparison to propensity score conditioning

As remarked in the main text, group average treatment effects for rank strata (and, indeed, individual preference types) are also identified by simple averages. However, when the strata are relatively small, the resulting measurements are far too noisy for meaningful statistical inference. KRR improves precision by using the full ranked list to predict outcomes. To illustrate this point, we compare our methodology using KRR to the standard inverse propensity-weighted averages below.



(a) KRR (uniform bands) (b) IPW (uniform bands) (c) IPW (pointwise bands)

Figure 12: KRR inference improves power. In our experiment, simply averaging the inverse propensity-weighted score frequently produces the incorrect sign (with pointwise signifiance), and uniform confidence bands are uninformative.