Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Data-driven confidence bands for distributed nonparametric regression

Valeriy Avanesov

submitted: June 9, 2020

Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: valeriy.avanesov@wias-berlin.de

No. 2729
Berlin 2020

2010 *Mathematics Subject Classification.* 62G15, 62F40, 60G15.

Key words and phrases. Gaussian process regression, kernel ridge regression, nonparametric regression, distributed regression, confidence bands, bootstrap.

The research of “Project Approximative Bayesian inference and model selection for stochastic differential equations (SDEs)” has been partially funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1294 “Data Assimilation”, “Project Approximative Bayesian inference and model selection for stochastic differential equations (SDEs)”. Further, we would like to thank Vladimir Spokoiny, Evgeniya Sokolova, the three anonymous reviewers and the area chair for the discussions, suggestions, criticism and/or proofreading which have greatly improved the manuscript.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Data-driven confidence bands for distributed nonparametric regression

Valeriy Avanesov

Abstract

Gaussian Process Regression and Kernel Ridge Regression are popular nonparametric regression approaches. Unfortunately, they suffer from high computational complexity rendering them inapplicable to the modern massive datasets. To that end a number of approximations have been suggested, some of them allowing for a distributed implementation. One of them is the divide and conquer approach, splitting the data into a number of partitions, obtaining the local estimates and finally averaging them. In this paper we suggest a novel computationally efficient fully data-driven algorithm, quantifying uncertainty of this method, yielding frequentist L_2-confidence bands. We rigorously demonstrate validity of the algorithm. Another contribution of the paper is a minimax-optimal high-probability bound for the averaged estimator, complementing and generalizing the known risk bounds.

1 Introduction

The problem of nonparametric regression arises in numerous applications including but not limited to finance [12, 48, 43, 2], neuroimaging [22, 34], climate [20, 31, 30], geology [14, 26, 23] and optimization [40, 11]. The frequentist setting of such a problem considers n response-covariate pairs (y_i, X_i) from $\mathbb{R} \times \mathcal{X}$ are being observed such that

$$y_i = f^*(X_i) + \varepsilon_i$$

for a compact $\mathcal{X} \subseteq \mathbb{R}^d$, centered independent sub-Gaussian noise ε_i of variance σ^2. Throughout the paper we presume X_i are drawn independently w.r.t. some unknown continuous measure π. In the paper we investigate the behaviour of one of the most popular non-parametric approaches to estimation of f^* – the Gaussian Process Regression (GPR) [32, 5, 19, 7]. GPR, being a Bayesian procedure, has been predominantly examined from a Bayesian point of view, i.e. no existence of f^* has been presumed, contrary to the frequentist setting we consider here. Namely, the contraction rate of the posterior distribution has been typically in focus [45, 46, 3].

Commonly, in order to analyse the frequentist behaviour of GPR, researchers turn to the Kernel Ridge Regression (KRR), whose point estimate $\hat{f}$ coincides with the mean of GPR posterior [6, 28, 44, 25, 52]. Most recently researchers have also developed interest for frequentist confidence sets. Namely, the authors of [49] suggest an approach to construct sup-norm confidence bands.

Unfortunately, KRR and GPR suffer from $O(n^3)$ time complexity, which renders them inapplicable to the datasets containing more than several thousands elements. To that end numerous approximations emerged. A wide variety of them rely on a low-rank approximation of the kernel driving the GP prior, e.g. PCA [38] or Nyström approximation [47]. The latter has been demonstrated to achieve nearly minimax-optimal performance [1, 16, 37]. Another strategy is to split the dataset into P partitions,

DOI 10.20347/WIAS.PREPRINT.2729 Berlin 2020

obtain the local estimates $\hat{f}_p$ separately and then average them, yielding $\bar{f}$. This does not only reduce the complexity to $O(n^3/P^2)$, but also makes a distributed implementation trivial. In [35] the method is given theoretical treatment for parametric families of dimenionality $p(n)/n \to const \in (0,1)$. [13] employed the idea for GPR, in [53] the idea is applied to KRR and further studied in [27]. A broad range of distributed non-parametric methods is analysed in [42].

One of the properties making GPR the instrument of choice is the ability to quantify the uncertainty of prediction. Only recently [49] have demonstrated that GPR posterior can be used to construct sup-norm frequentist confidence bands. At the same time, data driven-techniques based on Kernel Density Estimator were suggested [18, 9]. To the best of the authors' knowledge there has not yet been a distributed approach to the problem of nonparametric regression yielding frequentist uncertainty estimates.

The main contribution of this paper is enrichment of the divide and conquer approach suggested by [53] with a highly cost-effective novel bootstrap algorithm constructing confidence bands for f^*. We rigorously demonstrate the validity of the approach. The main result is established for undersmoothed prior, which is an assumption commonly employed to establish the validity of confidence sets [24, 41, 33, 49]. Moreover, we also obtain a minimax-optimal high-probability bound for $\left\|\bar{f} - f^*\right\|_2^2$, extending the earlier results [53, 27, 29], where the authors control expectation of the norm or of its positive power.

1.1 Notation

In the paper we heavily rely on a spectral decomposition of the kernel operator $k(\cdot,\cdot)$ w.r.t. π. Mercer's theorem [32] provides existence of normalized eigenfunctions $\phi_j \in L_2(\mathcal{X},\pi)$ along with the corresponding eigenvalues μ_j (in decreasing order). For a function $\|\cdot\|_2$ denotes an $L_2(\mathcal{X},\pi)$-norm, namely $\|f\|_2^2 = \int f^2 d\pi$, the dot-product is also defined w.r.t. π: $\langle f,g\rangle = \int fg d\pi$. The kernel $k(\cdot,\cdot)$ induces a RKHS $\mathcal{H}_k$ endowed with a norm

$$\|f\|_{\mathcal{H}_k}^2 = \sum_{j=1}^{\infty} \frac{\langle f,\phi_j\rangle^2}{\mu_j}. \tag{1.1}$$

For a vector $\|\cdot\|$ denotes an ℓ_2-norm, while for a matrix it denotes its maximum absolute eigenvalue. Frobenius norm is denoted as $\|\cdot\|_F$. We denote j-th largest eigenvalue of an operator A as $\lambda_j(A)$ and $\|A\|_1 \coloneqq \sum_{j=1}^{\infty} |\lambda_j(A)|$. I stands for an identity operator. We also use c and C as generalized positive constants, whose values may differ from line to line and depend only on $k(\cdot,\cdot)$. We use $\asymp$ to denote equality up to a multiplicative constant – namely, $a_i \asymp b_i$ implies $\mathrm{c}b_i \le a_i \le \mathrm{C}b_i$ for all i. We will also write $\mathcal{H}^s$ to denote a RKHS induced by a kernel exhibiting polynomial eigendecay of form $\mu_j \asymp j^{-2s}$ (Assumption 3.1).

2 The algorithm

GPR attains bias-variance trade-off via imposing a Gaussian Process (GP) prior over the function in question. A GP prior is driven by its mean (typically, assumed to be constant and zero) and covariance function

$$f \sim \mathcal{GP}(0, \sigma^2(n\rho)^{-1}k(\cdot,\cdot)),$$

where $\rho > 0$ is a regularization parameter. In the current study we focus on Matérn kernels, yet the results are also applicable to any covariance function demonstrating polynomial eigendecay and boundness of its eigenfunctions. Posterior distribution over f is also a GP with mean

$$\hat{f}(x) = k^*(x) \left(K + n\rho I\right)^{-1} \boldsymbol{y},$$

where $\boldsymbol{y} \coloneqq [y_i]_{i=1..n}$, $k^*(x) = [k(x, x_i)]_{i=1..n}$ and $K = [k(x_i, x_j)]_{i,j=1..n}$.

Alternatively, one can arrive to the same point estimate via Kernel Ridge Regression (KRR) [32]

$$\hat{f} \coloneqq \arg\max_f \left\{ -\frac{1}{2n} \sum_{i=1}^{n} \left(y_i - f(X_i)\right)^2 - \frac{\rho}{2} \left\|f\right\|^2_{\mathcal{H}_k} \right\}, \tag{2.1}$$

where $\left\|\cdot\right\|_{\mathcal{H}_k}$ refers to the RKHS norm, induced by the kernel $k(\cdot, \cdot)$ (see (1.1) for the definition).

The problem (2.1) is notorious for its high computational complexity being $O(n^3)$, rendering it impossible to scale. As [53] suggests, split the set of indices $\{1, 2, .., n\}$ into P disjoint sets $\{S_p\}_{p=1}^P$ of size $S \coloneqq |S_p| = n/P$ (we presume n/P is natural for simplicity). Now define P local estimators

$$\hat{f}_p \coloneqq \arg\max_f \left\{ -\frac{1}{2S} \sum_{i \in S_p} \left(y_i - f(X_i)\right)^2 - \frac{\rho}{2} \left\|f\right\|^2_{\mathcal{H}_k} \right\} \tag{2.2}$$

and the averaged one

$$\bar{f} \coloneqq \frac{1}{P} \sum_{p=1}^{P} \hat{f}_p.$$

Of course, P cannot grow linearly with n, yet for highly smooth classes of functions it can grow as a power of n close to 1, making the overall complexity $O(n^3/P^2)$ nearly linear (see (3.6) for details).

Distribution of $\bar{f}$ has a complicated nature, while its limiting distribution involves the spectral decomposition of $k(\cdot, \cdot)$, which is time-consuming to obtain. This significantly complicates the problem of constructing confidence bands for a confidence level β of sort

$$\mathbb{P}\left\{ \left\| \bar{f} - f^* \right\|_2 \leq r_\beta \right\} = \beta. \tag{2.3}$$

To that end we suggest a non-trivial bootstrap procedure. Classic bootstrap schemes [15] suggest to re-sample the input data. In our case it means solving the problems (2.2) from scratch for each bootstrap iteration, which is time-consuming. In order to avoid that we suggest to re-sample $\hat{f}_p$ directly, achieving $O(P)$ time complexity. Formally, we draw $\hat{f}_p^\flat$ independently and uniformly from $\left\{ \hat{f}_p \right\}_{p=1}^P$ for $p = 1..P$ and define a bootstrap counterpart of the averaged estimator $\bar{f}$

$$\bar{f}^\flat \coloneqq \frac{1}{P} \sum_{p=1}^{P} \hat{f}_p^\flat.$$

Denoting the bootstrap measure as $\mathbb{P}^\flat$ we can now obtain $r_\beta^\flat$ as

$$\mathbb{P}^\flat\left\{ \left\| \bar{f}^\flat - \bar{f} \right\|_2 \leq r_\beta^\flat \right\} = \beta.$$

We establish closeness of $\mathbb{P}$ and $\mathbb{P}^\flat$ in some sense (see Theorem 3.3), justifying the use of the bootstrap quantile $r_\beta^\flat$ instead of the real-world r_β in (2.3).

Remark 2.1. *From a distributed implementation standpoint it may be more convenient and efficient to employ multipliers instead of sampling with return. Namely, the bootstrap counterpart of $\bar{f}$ can be constructed as $\bar{f}^w := \frac{1}{P}\sum_{p=1}^{P} u_p \hat{f}_p$ for i.i.d. weights u_p with unit expectation and variance, e. g. $u_p \sim \mathcal{N}(1,1)$. All the results demonstrated for $\bar{f}^b$ are also valid for $\bar{f}^w$, as we only rely on the first two moments of the bootstrap estimate.*

3 Theoretical analysis

3.1 Assumptions

First of all, we impose a polynomial rate of decay on the eigenvalues of $k(\cdot,\cdot)$.

Assumption 3.1 (Polynomial eigendecay)**.** *Let there exist a constant $s > 1/2$ s.t. for the j-th largest eigenvalue μ_j of $k(\cdot,\cdot)$*

$$\mu_j \asymp j^{-2s}.$$

As demonstrated in [49], Assumption 3.1 holds for Matérn kernel with smoothness α in a d-dimensional space with $s = (2\alpha + d)/2$. Another popular example is a Squared Exponential kernel, which is known to exhibit an exponential rate of eigendecay. With some abuse of formality, our results can be applied in this case with $s = \infty$. Alternatively, the argument can be carefully repeated with minimal augmentation in this case as well.

We also assume the eigenfunctions of the kernel to be bounded. The analysis in [50, 4] proves the assumption holds for Matérn kernel under uniform and normal distributions of covariates on a compact.

Assumption 3.2 (Boundness of eigenfunctions)**.** *Denote a normalized eigenfunction corresponding to the j-th largest eigenvalue as $\phi_j(\cdot)$. Let there exist a positive constant C_ϕ s.t. $\sup_j \max_{X\in\mathcal{X}} |\phi_j(X)| \le C_\phi$.*

In conclusion we impose sub-Gaussianity assumption over noise, that being a common relaxation of Gaussianity.

Assumption 3.3 (Sub-Gaussianity)**.** *Let there exist a constant $\mathfrak{g}^2$ s.t. for all $a \in \mathbb{R}$*

$$\mathbb{E}\left[\exp(a\varepsilon_1)\right] \le \exp\left(\frac{\mathfrak{g}^2 a^2}{2}\right).$$

3.2 Theoretical results

We open the section with a consistency result for $\hat{f}$. A similar bound is obtained in [6]. We extend it, explicitly covering undersmoothed priors and providing a slightly tighter bound for that case. Note, it gives a bound in terms of $L_2(\mathcal{X},\pi)$-norm, which is natural, as an increase of density of X_i in some subset of $\mathcal{X}$ leads to better predictions on the subset.

Theorem 3.1. *Impose Assumption 3.1, Assumption 3.2, Assumption 3.3 and let $f^* \in \mathcal{H}^{s_\circ}$ for $s_\circ \ge s$. Choose*

$$\rho = n^{-\frac{2s}{2s+1}}.$$

Then for any $\mathrm{x} > 1$ *and any* $\mathrm{t} > 2.6$ *on a set of probability at least* $1 - e^{-\mathrm{x}} - e^{-\mathrm{t}/2}$

$$\left\| \hat{f} - f^* \right\|_2 \leq \mathrm{C}\sqrt{\mathrm{tx}}\mathfrak{g} n^{-\frac{s}{2s+1}} + \mathrm{C} \left\| f^* \right\|_{\mathcal{H}^{s_\circ}} n^{-\frac{\min\{s_\circ, 2s\}}{2s+1}}$$

for some $\mathrm{C} > 0$ *depending only on* s*.*

The proof (deferred to Appendix A) relies on the bound, established on a set of high probability. Namely, we define a class of designs we are satisfied with (Assumption A.1), next we demonstrate the measure of the class is high (Lemma D.2) and establish a consistency result under Assumption A.1 (Lemma A.3). Here we choose the regularization parameter ρ in a classical manner, acquiring balance between bias and variance in case $s_\circ = s$.

Under mild assumptions the same minimax-optimal bound may be established for $\bar{f}$.

Theorem 3.2. *Impose Assumption 3.1, Assumption 3.2, Assumption 3.3 and let* $f^* \in \mathcal{H}^{s_\circ}$, $s_\circ \geq s$ *and*

$$P \leq \mathrm{c} \frac{n^{\frac{2s-1}{2s+1}}}{\log n}. \tag{3.1}$$

Choose

$$\rho = n^{-\frac{2s}{2s+1}}.$$

Then for all $\mathrm{x} > 1$ *and* $\mathrm{t} > 2.6$ *with probability at least* $1 - e^{-\mathrm{x}} - e^{-\mathrm{t}/2}$

$$\left\| \bar{f} - f^* \right\|_2 \leq \mathrm{C}\sqrt{\mathrm{tx}}\mathfrak{g} n^{-\frac{s}{2s+1}} + \mathrm{C} \left\| f^* \right\|_{\mathcal{H}^{s_\circ}} n^{-\frac{\min\{s_\circ, 2s\}}{2s+1}}. \tag{3.2}$$

This theorem is a direct corollary of Lemma B.1. The strategy of the proof is to consider Fisher expansion (see Lemma E.1 proven by [39]) for each $\hat{f}_p$, expressing the discrepancy between the sample-level parameter and its penalized population-level counterpart in terms of Hessian and gradient of the likelihood. Next, we bound the Hessian by Lemma D.2, sum up the expansions and employ additivity of the gradient. Finally, we obtain the concentration via Hanson-Wright inequality.

The expression (3.1) dictates the maximum number of partitions P allowed for the minimax-optimal bound (B.2) to hold. It does indeed match the condition obtained in [53].

Having obtained the high-probability bound with exponential tail, we can apply integrated tail probability expectation formula to produce the following corollary, repeating the result by [29].

Corollary 3.1. *Impose assumptions of Theorem 3.2. Then for any positive* η

$$\mathbb{E}\left[\left\| \bar{f} - f^* \right\|_2^\eta \right] = O\left(n^{-\frac{s\eta}{2s+1}} \right).$$

Finally, we turn to analysis of the suggested bootstrap scheme. The idea is usual for bootstrap validity results [10, 8]. First, we establish Gaussian Approximation for the estimator $\bar{f}$. In order to do so we first notice that by CLT

$$\sup_{r>0} \left| \mathbb{P}\left\{ \left\| \bar{f} - f_\rho^* \right\|_2^2 < r \right\} - \mathbb{P}\left\{ \|\boldsymbol{\gamma}\|^2 < r \right\} \right| \to 0$$

for $n, P \to +\infty$, where $f_\rho^* = \mathbb{E}\left[\bar{f}\right]$ and $\boldsymbol{\gamma}$ is a centered Gaussian element of a Hilbert space with covariance operator $\mathrm{Var}\left[\bar{f}\right]$. As we are interested in a concentration around f^* and not f_ρ^*, we also have to account for the mis-tie between the two, making use of Gaussian Comparison, arriving to

$$\sup_{r>0} \left| \mathbb{P}\left\{ \left\| \bar{f} - f^* \right\|_2^2 < r \right\} - \mathbb{P}\left\{ \|\boldsymbol{\gamma}\|^2 < r \right\} \right| \to 0. \tag{3.3}$$

Here we will have to impose undersmoothness of the prior ($s_\circ > s$) in order to make the remainder term negligible.

Turning to the bootstrap estimator $\bar{f}^\flat$, we will use CLT again, which yields

$$\sup_{r>0} \left| \mathbb{P}\left\{ \left\| \bar{f}^\flat - \bar{f} \right\|_2^2 < r \right\} - \mathbb{P}\left\{ \|\hat{\boldsymbol{\gamma}}\|^2 < r \right\} \right| \to 0 \tag{3.4}$$

for a centered Gaussian element of a Hilbert space $\hat{\boldsymbol{\gamma}}$ with covariance operator $\operatorname{Var}\left[\bar{f}^\flat\right]$. The final step is to establish closeness of covariance operators of $\boldsymbol{\gamma}$ and $\hat{\boldsymbol{\gamma}}$ and apply Gaussian Comparison obtaining

$$\sup_{r>0} \left| \mathbb{P}\left\{ \|\boldsymbol{\gamma}\|^2 < r \right\} - \mathbb{P}\left\{ \|\hat{\boldsymbol{\gamma}}\|^2 < r \right\} \right| \to 0. \tag{3.5}$$

Combining (3.3), (3.4) and (3.5) will constitute the following claim.

Theorem 3.3. *Impose Assumption 3.1, Assumption 3.2, Assumption 3.3 and let $f^* \in \mathcal{H}^{s_\circ}$ for $s_\circ > s$. Choose*

$$\rho = n^{-\frac{2s}{2s+1}}.$$

Then

$$\begin{aligned} \mathrm{R}^\flat &\coloneqq \sup_{r>0} \left| \mathbb{P}\left\{ \left\| \bar{f} - f^* \right\|_2^2 < r \right\} - \mathbb{P}^\flat\left\{ \left\| \bar{f}^\flat - \bar{f} \right\|_2^2 < r \right\} \right| \\ &\leq \mathrm{C}\left(\frac{\sigma^2 n^{\frac{2}{2s+1}} \log^2 n}{P} \right)^{\frac{4s-1}{8s}} + \mathrm{C} n^{-\frac{2\min\{s, s_\circ - s\}}{2s+1}} \left\| f^* \right\|_{\mathcal{H}^{s_\circ}}^2. \end{aligned}$$

The sketched proof is implemented in Appendix C.

Naturally, the remainder gets smaller for larger P, as it implies a richer set to sample from. On the other hand, Theorem 3.2 imposes an upper bound on P. Up to logarithmic terms, the choice of $P = P(n)$ implying both $\mathrm{R}^\flat = o(1)$ and (3.1) must satisfy

$$n^{\frac{2}{2s_\circ+1}} \ll P(n) \ll n^{\frac{2s_\circ-1}{2s_\circ+1}} \tag{3.6}$$

in order for us to have both high-probability bound and credible bands. Clearly, it is possible only for $s_\circ > 3/2$. In case of Matérn kernels this translates to $\alpha + d/2 > 3/2$, prohibiting only the case $d = 1$ and $\alpha \in (1/2, 1]$.

As discussed in Section 1, choice of an undersmoothed prior is a common way to trade optimality of an estimator for a possibility to construct a confidence interval. But how much do we have to pay? Consider the two summands in the claim of Theorem 3.3. The choice $s = \frac{2}{3} s_\circ$ implies the former term dominates the latter, hence this choice is the largest reasonable sacrifice. The concentration rate for $\bar{f}$ then would be $n^{-\frac{s_\circ}{2s_\circ+3/2}}$ (instead of the minimax $n^{-\frac{s_\circ}{2s_\circ+1}}$) which is only marginally suboptimal.

4 Simulation study

In this section we study the suggested algorithm experimentally. We choose $\mathcal{X} = [0, 1]$ and $f^*(x) = \sin(\tau x)$, where $\tau \approx 6.28\ldots$ denotes the number of radians in a turn. The design is uniformly and identically sampled from $\mathcal{X}$. The noise ε_i is Gaussian with variance $\sigma^2 = 1$, meaning the signal-to-noise ratio is $1/\sqrt{2}$. Nominal confidence level is set to $\beta = 0.95$. The number of bootstrap iterations is

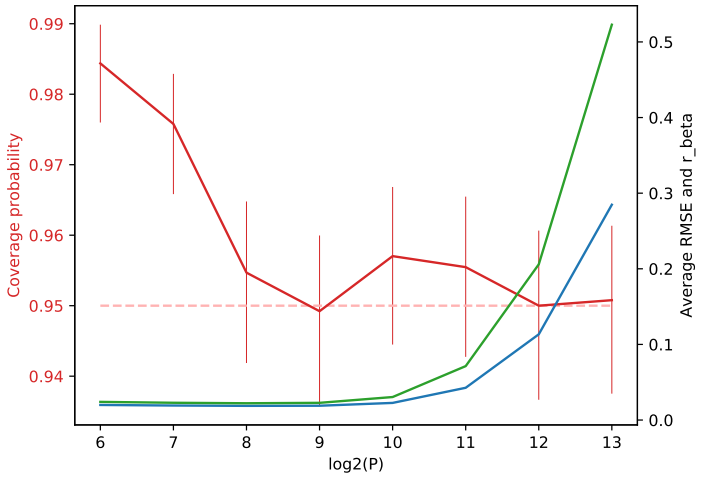

Figure 1: The horizontal axis uses log-2 scale. The nominal significance level is $\beta = 0.95$, shown with a dashed line. The estimated coverage probability is shown with a red line, the error bars correspond to Wilson 95% point-wise confidence intervals. The green and the blue lines depict the quantile r_β (see (2.3)) and the average root-mean-square error respectively. Each point of the plot is averaged over 1280 trials.

chosen as 1000. $k(\cdot, \cdot)$ is chosen to be Matérn kernel with a smoothness index $\alpha = 5/2$. We choose sample size $n = 2^{17}$ and let P vary from 2^6 to 2^{13}. The results are shown in Figure 1. As Theorem 3.3 suggests, the number of partitions needs to be large enough and we observe, the method matches the nominal confidence level for $P \geq 2^8$ and does not diverge from it even when excessively large P (e.g. above 2^{10}) renders the averaged estimator $\bar{f}$ sub-optimal. The latter effect is described by Theorem 3.2. Thus, there is a wide range to choose P from, enjoying both minimax optimal estimator $\bar{f}$ and valid confidence bands.

5 Conclusion and future work

The problem of distributed nonparametric regression being of great importance in the light of the ever-growing datasets has earlier received a consistent estimator. Namely, the Fast-KRR approach, being an application of divide and conquer paradigm to KRR. Its consistency has been demonstrated in terms of risk. In this paper we complement these results with a high-probability bound. Our main contribution is a novel enhancement of the method, providing a confidence band in addition to the point estimate. The time complexity of the procedure, being sub-linear in sample size, is dwarfed by the complexity of the Fast-KRR itself, so calculation of the confidence bands is virtually free of charge. The theoretical analysis of the procedure is powered by the recent results on Gaussian Comparison [17] and a familiar Central Limit Theorem in a Hilbert space [51].

Future research should also explore the posteriors of KRRs

$$f \,|\, \{(X_i, y_i)\}_{i \in S_p} \propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(X_i))^2 \right\} \exp\left\{ -\frac{\rho}{2} \|f\|_{\mathcal{H}_k}^2 \right\},$$

justifying aggregation of both posterior mean and covariance. A major obstacle here would be the

need to obtain a spectral decomposition of the kernel (as it is involved in the posterior covariance), which is computationally difficult.

There is also a promising alternative. So far all the consistency results for Fast-KRR deal with L_2-measure. Obtaining a concentration with respect to a stronger L_∞-norm may turn out highly beneficial in the light of the recent research [49]. There the authors have recognized the posterior covariance of Gaussian Process Regression as an M-estimator and employed the observation to justify its use in construction of sup-norm frequentist confidence sets.

References

[1] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209, 2013.

[2] K Basu and Maria C Mariani. Interpolating techniques and nonparametric regression methods applied to geophysical and financial data analysis. *Handbook of High-Frequency Trading and Modeling in Finance*, pages 251–294, 2016.

[3] Anirban Bhattacharya and Debdeep Pati. Posterior contraction in gaussian process regression using wasserstein approximations. *Information and Inference: A Journal of the IMA*, 6(4):416–440, 2017.

[4] Anirban Bhattacharya and Debdeep Pati. Posterior contraction in Gaussian process regression using Wasserstein approximations. *Information and Inference*, 6(4):416–440, 2017.

[5] Ilias Bilionis and Nicholas Zabaras. Multi-output local gaussian process regression: Applications to uncertainty quantification. *Journal of Computational Physics*, 231(17):5718–5746, 2012.

[6] A Caponnetto and E De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. 368:331–368, 2007.

[7] Junghui Chen, Lester Lik Teck Chan, and Yi-Cheng Cheng. Gaussian process regression based optimal design of combustion systems using flame images. *Applied energy*, 111:153–160, 2013.

[8] Xiaohui Chen. Gaussian and bootstrap approximations for high-dimensional u-statistics and their applications. *Ann. Statist.*, 46(2):642–678, 04 2018.

[9] Gang Cheng and Yen Chi Chen. Nonparametric inference via bootstrapping the debiased estimator. *Electronic Journal of Statistics*, 13(1):2194–2256, 2019.

[10] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, 45(4):2309–2352, 2017.

[11] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. *34th International Conference on Machine Learning, ICML 2017*, 2:1397–1422, 2017.

[12] Marta Degl'Innocenti, Roman Matousek, Zeljko Sevic, and Nickolaos G Tzeremes. Bank efficiency and financial centres: Does geographical location matter? *Journal of International Financial Markets, Institutions and Money*, 46:188–198, 2017.

[13] Marc Peter Deisenroth and Jun Wei Ng. Distributed Gaussian processes. *32nd International Conference on Machine Learning, ICML 2015*, 2:1481–1490, 2015.

[14] Marco Di Marzio, Agnese Panzera, and Charles C Taylor. Nonparametric regression for spherical data. *Journal of the American Statistical Association*, 109(506):748–763, 2014.

[15] B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 01 1979.

[16] Shai Fine and Katya Scheinberg. Efficient svm training using low-rank kernel representations. *Journal of Machine Learning Research*, 2(Dec):243–264, 2001.

[17] Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, and Vladimir Ulyanov. Large ball probabilities, Gaussian comparison and anti-concentration. *Bernoulli*, 25(4 A):2538–2563, 2019.

[18] Peter Hall and Joel Horowitz. A simple bootstrap method for constructing nonparametric confidence bands for functions. *Annals of Statistics*, 41(4):1892–1921, 2013.

[19] He He and Wan-Chi Siu. Single image super-resolution using gaussian process regression. In *CVPR 2011*, pages 449–456. IEEE, 2011.

[20] Yasushi Honda, Masahide Kondo, Glenn McGregor, Ho Kim, Yue-Leon Guo, Yasuaki Hijioka, Minoru Yoshikawa, Kazutaka Oka, Saneyuki Takano, Simon Hales, et al. Heat-related mortality risk model for climate change impact projection. *Environmental health and preventive medicine*, 19(1):56, 2014.

[21] Daniel Hsu, Sham Kakade, and Tong Zhang. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electron. Commun. Probab.*, 17:13 pp., 2012.

[22] Jung Won Hyun, Yimei Li, John H Gilmore, Zhaohua Lu, Martin Styner, and Hongtu Zhu. Sgpp: spatial gaussian predictive process models for neuroimaging data. *NeuroImage*, 89:70–80, 2014.

[23] Charlie Kirkwood, David Beamish, Bob Lister, Mark Cave, Antonio Ferreira, and Paul Everett. Geological mapping using high resolution regression modelled soil geochemistry. 2015.

[24] Bartek T Knapik, Aad W Van Der Vaart, J Harry van Zanten, et al. Bayesian inverse problems with gaussian priors. *The Annals of Statistics*, 39(5):2626–2657, 2011.

[25] Vladimir Koltchinskii et al. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[26] David J Lary, Amir H Alavi, Amir H Gandomi, and Annette L Walker. Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1):3–10, 2016.

[27] Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *The Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

[28] Shahar Mendelson. Geometric parameters of kernel machines. In *International Conference on Computational Learning Theory*, pages 29–43. Springer, 2002.

[29] Nicole Mücke and Gilles Blanchard. Parallelizing spectrally regularized kernel algorithms. *Journal of Machine Learning Research*, 19:1–29, 2018.

[30] Manfred Mudelsee. Trend analysis of climate time series: A review of methods. *Earth-science reviews*, 190:310–322, 2019.

[31] Isidro A Pérez, M Luisa Sánchez, M Ángeles García, and Nuria Pardo. Features of the annual evolution of co2 and ch4 in the atmosphere of a mediterranean climate site studied using a nonparametric and a harmonic function. *Atmospheric Pollution Research*, 7(6):1013–1021, 2016.

[32] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, January 2006.

[33] Kolyan Ray et al. Adaptive bernstein–von mises theorems in gaussian white noise. *The Annals of Statistics*, 45(6):2511–2536, 2017.

[34] Philip T Reiss, Lei Huang, Yin-Hsiu Chen, Lan Huo, Thaddeus Tarpey, and Maarten Mennes. Massively parallel nonparametric regression, with an application to developmental brain mapping. *Journal of Computational and Graphical Statistics*, 23(1):232–248, 2014.

[35] Jonathan D. Rosenblatt and Boaz Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference*, 5(4):379–404, 2016.

[36] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18:9 pp., 2013.

[37] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 2015-January:1657–1665, 2015.

[38] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[39] Vladimir Spokoiny and Maxim Panov. Accuracy of Gaussian approximation in nonparametric Bernstein – von Mises Theorem. 2019.

[40] Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *In International Conference on Machine Learning*, 2010.

[41] Botond Szabó, Aad W Van Der Vaart, JH van Zanten, et al. Frequentist coverage of adaptive nonparametric bayesian credible sets. *The Annals of Statistics*, 43(4):1391–1428, 2015.

[42] Botond Szabó and Harry Van Zanten. An asymptotic analysis of distributed nonparametric methods. *Journal of Machine Learning Research*, 20, 6 2019.

[43] Nickolaos Tzeremes. Financial development and countries' production efficiency: A nonparametric analysis. *Journal of Risk and Financial Management*, 11(3):46, 2018.

[44] van de Geer. *Empirical Processes in M-estimation*. Cambridge UP, 2006.

[45] Aad van der Vaart and Harry van Zanten. Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12(Jun):2095–2119, 2011.

[46] Aad W van der Vaart, J Harry van Zanten, et al. Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, 37(5B):2655–2675, 2009.

[47] Christopher KI Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2001.

[48] BaiLing Xu. Nonparametric model research on price fluctuation of financial assets. In *2016 2nd International Conference on Education, Social Science, Management and Sports (ICESSMS 2016)*. Atlantis Press, 2017.

[49] Yun Yang and Debdeep Pati. Bayesian model selection consistency and oracle inequality with intractable marginal likelihood. pages 1–38, 2017.

[50] Yun Yang and Debdeep Pati. Bayesian model selection consistency and oracle inequality with intractable marginal likelihood. pages 1–38, 2017.

[51] B. A. Zalesskii, V. V. Sazonov, and V. V. Ul'yanov. A precise estimate of the rate of convergence in the central limit theorem in hilbert space. *Mathematics of the USSR - Sbornik*, 68(2):453–482, 1991.

[52] Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.

[53] Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16:3299–3340, 2015.

A Proof of $\hat{f}$ consistency

All the claims presented in Appendices A – D implicitly impose Assumption 3.1, Assumption 3.2 and Assumption 3.3. We also let $\rho = \rho(n) = o(1)$ and $f^* \in \mathcal{H}^{s_\circ}$. In this section we let $s_\circ \geq s$.

We open the section with some notation. Consider the eigenvalues $\{\mu_j\}_{j=1}^{\infty}$ and normalized eigenfunctions $\{\phi_j(\cdot)\}_{j=1}^{\infty}$ of $k(\cdot,\cdot)$ w.r.t. continuous measure π. Now for $f \in \mathcal{H}_k$ we have the vector $\mathbf{f}$ of expansion coefficients $\mathbf{f}_j := \langle f, \phi_j \rangle$. Further, we introduce the design matrix $\Phi \in \mathbb{R}^{n\times\infty}$ s.t. $\Phi_{ij} = \phi_j(X_i)$. At this point the estimator (2.1) can be rewritten as

$$\hat{\mathbf{f}} := \arg\max_{\mathbf{f}} \left(-\frac{1}{2n} \|\Phi\mathbf{f} - y\|^2 - \frac{\rho}{2} \sum_j \frac{\mathbf{f}_j^2}{\mu_j} \right).$$

Next, consider a diagonal matrix $M \in \mathbb{R}^{\infty\times\infty}$ s. t. $M_{jj} = \sqrt{\mu_j}$ and define $\boldsymbol{\theta} := M^{-1}\mathbf{f}$, $\Psi := \Phi M$, rewriting (2.1) again

$$\hat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}} \left(\underbrace{-\frac{1}{2n} \|\Psi\boldsymbol{\theta} - y\|^2}_{L(\boldsymbol{\theta})} - \frac{\rho}{2} \|\boldsymbol{\theta}\|^2 \right).$$

We also define

$$\boldsymbol{\theta}^* := \arg\max_{\boldsymbol{\theta}} \mathbb{E}\left[L(\boldsymbol{\theta})\right]$$

and its penalized counterpart

$$\boldsymbol{\theta}^*_{\boldsymbol{\rho}} := \arg\max_{\boldsymbol{\theta}} \mathbb{E}\left[L(\boldsymbol{\theta})\right] - \frac{\rho}{2} \|\boldsymbol{\theta}\|^2. \tag{A.1}$$

Similarly we define $\mathbf{f}^*_\rho$ and f^*_ρ. We also introduce a vector $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ s.t. $\boldsymbol{\varepsilon}_i = \varepsilon_i$. By the means of trivial calculus we have

$$\nabla\zeta \coloneqq \nabla\left(L(\boldsymbol{\theta}) - \mathbb{E}_\varepsilon[L(\boldsymbol{\theta})]\right) = \frac{1}{n}\Psi^T\boldsymbol{\varepsilon}$$

and

$$D^2_\rho \coloneqq -\nabla^2\left(L(\boldsymbol{\theta}) - 2\rho\left\|\boldsymbol{\theta}\right\|\right) = \frac{1}{n}\Psi^T\Psi + \rho I.$$

Now we are ready to formulate an assumption we impose on the design $\{X_i\}_{i=1}^n$.

Assumption A.1 (Design regularity)**.** *Let there exist some positive δ s. t.*

$$\left\|\left(M^2+\rho I\right)^{-1/2} D^2_\rho \left(M^2+\rho I\right)^{-1/2} - I\right\| \leq \delta < 1.$$

Assumption A.1 can seem to be obscure, but Lemma D.2 guarantees that it holds for a random design with X_i being i.i.d. and distributed w.r.t. a continuous measure.

First, we bound the bias term.

Lemma A.1. *Let $f^* \in \mathcal{H}^{s_\circ}$ and $s_\circ \geq s$. Then*

$$\left\|\mathbf{f}^* - \mathbf{f}^*_\rho\right\|^2 \leq \rho^{\min\left\{\frac{s_\circ}{s},2\right\}} \left\|f^*\right\|^2_{\mathcal{H}^{s_\circ}}.$$

Proof. From the stationarity condition for (A.1) one gets

$$\left(\boldsymbol{\theta}^*_{\boldsymbol{\rho}} - \boldsymbol{\theta}^*\right) = \rho\left(M^2+\rho I\right)^{-1}\boldsymbol{\theta}^*$$

and hence

$$\left(\mathbf{f}^*_\rho - \mathbf{f}^*\right) = \rho\left(M^2+\rho I\right)^{-1}\mathbf{f}^*. \tag{A.2}$$

Now using the fact that $f^* \in \mathcal{H}^{s_\circ}$

$$\left\|\rho\left(M^2+\rho I\right)^{-1}\mathbf{f}^*\right\|^2 \leq \mathrm{C}\rho^2 \sum_j \frac{j^{-2s_\circ}}{\left(j^{-2s}+\rho\right)^2}\frac{\left(\mathbf{f}^*{}_j\right)^2}{j^{-2s_\circ}}.$$

Maximization of $\frac{j^{-2s_\circ}}{\left(j^{-2s}+\rho\right)^2}$ over $j > 0$ yields

$$\left\|\rho\left(M^2+\rho I\right)^{-1}\mathbf{f}^*\right\|^2 \leq \rho^{\min\left\{\frac{s_\circ}{s},2\right\}} \left\|f^*\right\|^2_{\mathcal{H}^{s_\circ}}. \tag{A.3}$$

Combining (A.2) and (A.3) finalizes the proof. □

The next lemma bounds the right-hand side of Fisher expansion (Lemma E.1).

Lemma A.2. *For all* $\mathrm{x} > 0$

$$\mathbb{P}\left\{\left\|D_\rho^{-1}\nabla\zeta\right\| \leq \mathrm{C}_{\mathfrak{g}}\sqrt{\frac{1+\sqrt{\mathrm{x}}+2\mathrm{x}}{\rho^{1/(2s)}n(1-\delta)}}\right\} \geq 1 - e^{-\mathrm{x}}.$$

Proof. By the definition of $\nabla\zeta$

$$\left\|D_\rho^{-1}\nabla\zeta\right\|^2 = \frac{1}{n^2}\boldsymbol{\varepsilon}^T\Psi^T\left(\frac{1}{n}\Psi^T\Psi+\rho I\right)^{-1}\Psi\boldsymbol{\varepsilon}.$$

Clearly,

$$\left(\frac{1}{n}\Psi^T\Psi+\rho I\right) \geq (1-\delta)\left(M^2+\rho I\right)$$

and hence

$$\left\|D_\rho^{-1}\nabla\zeta\right\|^2 \leq \boldsymbol{\varepsilon}^T\frac{1}{n^2}\Psi^T\left((1-\delta)\left(M^2+\rho I\right)\right)^{-1}\Psi\boldsymbol{\varepsilon}.$$

Applying Lemma D.1 completes the argument. □

Now we are ready to establish a consistency result.

Lemma A.3. *Impose Assumption A.1. Then on a set of probability at least* $1-e^{-\mathrm{x}}$ *for all* $\mathrm{x}>1$

$$\left\|\hat{f}-f^*\right\|_2 \leq \mathrm{C}\sqrt{\underbrace{\frac{\mathfrak{g}^2\mathrm{x}}{(1-\delta)\rho^{1/(2s)}n}}_{Variance}+\underbrace{\rho^{\min\left\{\frac{s_\circ}{s},2\right\}}\left\|f^*\right\|^2_{\mathcal{H}^{s_\circ}}}_{Bias}}.$$

Proof. We apply Lemma A.2 along with Lemma E.1, which yield for some positive C on a set of probability at least $1-e^{-\mathrm{x}}$ for any positive x

$$\left\|D_\rho\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*_{\boldsymbol{\rho}}\right)\right\| \leq \mathrm{C}\mathfrak{g}\sqrt{\frac{1+\sqrt{\mathrm{x}}+2\mathrm{x}}{\rho^{1/(2s)}n(1-\delta)}}.$$

But clearly,

$$(1-\delta)\left\|\hat{\mathbf{f}}-\mathbf{f}^*_\rho\right\|_2^2 = \left\|\sqrt{1-\delta}M(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\right\|^2 \leq (1-\delta)\left\|D_\rho\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right\|^2.$$

In order to bound the bias term we apply Lemma A.1, constituting the claim. □

The proof of Theorem 3.1 is almost trivial now.

Proof of Theorem 3.1. The proof consists in applying Lemma D.2 followed by applying Lemma A.3. □

B Proof of $\bar{f}$ consistency

Denote a map from $\mathcal{X}$ to $\mathbb{R}^\infty$

$$\psi(X) \coloneqq \left(\sqrt{\mu_1}\phi_1(X)\ \sqrt{\mu_2}\phi_2(X)\ \sqrt{\mu_3}\phi_3(X)\ ...\right)^T \in \mathbb{R}^\infty$$

and $\psi_i \coloneqq \psi(X_i)$.

We have earlier introduced the objects related to the global estimator $\hat{f}$, such as Ψ, D^2_ρ, $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{f}}$, $\boldsymbol{\varepsilon}$. In this section we make use of their local counterparts related to $\hat{f}_p$ such as Ψ_p, $D^2_\rho(p)$, $\hat{\boldsymbol{\theta}}_p$, $\hat{\mathbf{f}}(p)$, $\boldsymbol{\varepsilon}(p)$. Also define $\bar{\boldsymbol{\theta}} \coloneqq \frac{1}{P}\sum_p\hat{\boldsymbol{\theta}}_p$ and $\bar{\mathbf{f}} \coloneqq \frac{1}{P}\sum_p\hat{\mathbf{f}}(p)$. Throughout the section we let $s_\circ \geq s$.

Lemma B.1. *Let*

$$P \leq \mathrm{c}\frac{n^{\frac{2s-1}{2s+1}}}{\log n}. \tag{B.1}$$

Choose

$$\rho = n^{-\frac{2s}{2s+1}}.$$

Then for all $\mathrm{x} > 0$ *and* $\mathrm{t} > 0$ *with probability at least* $1 - e^{-\mathrm{x}} - \left(e^{\mathrm{t}} - \mathrm{t} - 1\right)^{-1}$

$$\left\|\bar{f} - f^*\right\|_2 \leq \mathrm{C}\sqrt{\mathrm{t}\left(1 + 2\mathrm{x} + \sqrt{\mathrm{x}}\right)}\mathfrak{g} n^{-\frac{s}{2s+1}} + \mathrm{C}\left\|f^*\right\|_{\mathcal{H}^{s_\circ}} n^{-\frac{\min\{s_\circ, 2s\}}{2s+1}}. \tag{B.2}$$

Proof. Using Fisher expansion (Lemma E.1)

$$\begin{aligned} D_\rho(p)\left(\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}^*_{\boldsymbol{\rho}}\right) &= \frac{1}{S} D_\rho^{-1}(p)\Psi_p \boldsymbol{\varepsilon}(p) \\ &= \frac{1}{S} D_\rho^{-1}(p)\left(\sum_{i \in S_p} \psi_i \varepsilon_i\right). \end{aligned}$$

Now by Lemma D.2 with probability at least $1 - \left(e^{\mathrm{t}} - \mathrm{t} - 1\right)^{-1}$ we have

$$\begin{aligned} \left\|(1 - \delta(S,\mathrm{t}))^{1/2}\left(M^2 + \rho I\right)^{1/2}\left(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*_{\boldsymbol{\rho}}\right)\right\| &\leq \left\|\frac{1}{P}\sum_{p=1}^{P} D_\rho(p)\left(\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}^*_{\boldsymbol{\rho}}\right)\right\| \\ &\leq \left\|\frac{1}{n}(1 - \delta(S,\mathrm{t}))^{-1/2}\left(M^2 + \rho I\right)^{-1/2}\Psi\boldsymbol{\varepsilon}\right\| \end{aligned}$$

and hence

$$\left\|\bar{\mathbf{f}} - \mathbf{f}^*_\rho\right\| \leq \frac{1}{1 - \delta(S,\mathrm{t})}\left\|\frac{1}{n}\left(M^2 + \rho I\right)^{-1/2}\Psi\boldsymbol{\varepsilon}\right\|.$$

Next apply Lemma D.1. On a set of probability at least $1 - e^{-\mathrm{x}}$ for all $\mathrm{x} > 0$ we have

$$\left\|\bar{\mathbf{f}} - \mathbf{f}^*_\rho\right\| \leq \frac{\mathrm{C}\mathfrak{g}}{1 - \delta(S,\mathrm{t})}\sqrt{\frac{1 + \sqrt{\mathrm{x}} + 2\mathrm{x}}{\rho^{1/(2s)} n}}.$$

Assumption (B.1) implies $1 - \delta(S,\mathrm{t}) > 1/2$. The bias term is controlled by Lemma A.1. □

C Bootstrap validity proof

Define P i.i.d. vectors $\mathbf{g}(p) \coloneqq \hat{\mathbf{f}}(p) - \mathbf{f}^*_\rho$, denote $\Sigma = \mathbb{E}\left[\mathbf{g}(p)\mathbf{g}(p)^T\right]$ and $\hat{\Sigma} \coloneqq \frac{1}{P}\sum_p \mathbf{g}(p)\mathbf{g}(p)^T$. Throughout the section we let $s_\circ > s$.

Lemma C.1. *Consider a centered Gaussian element of a Hilbert space* $\boldsymbol{\gamma}$ *with a covariance operator* Σ*. Then for all positive* $\mathrm{t} > 2.6$ *and* $\delta = \delta(S,\mathrm{t})$ *coming from Lemma D.2*

$$(1+\delta)^{-1}\Sigma^* \leq n\mathrm{Var}\left[\bar{\mathbf{f}} - \mathbf{f}^*_\rho\right] \leq (1-\delta)^{-1}\Sigma^*,$$

$$(1+\delta)^{-1}\Sigma^* \leq S\mathrm{Var}\left[\hat{\mathbf{f}}(p) - \mathbf{f}^*_\rho\right] \leq (1-\delta)^{-1}\Sigma^*$$

and

$$\operatorname{tr}(\Sigma) \leq \mathrm{C}\sigma^2 S^{-1}\rho^{-1/(2s)},$$

where $\Sigma^* := \sigma^2 \left(M^2+\rho I\right)^{-2} M^4$. *Morover, uniformly for positive* r

$$\begin{aligned}&\left|\mathbb{P}\left\{\left\|\bar{f}-f^*\right\|_2 \leq r\right\} - \mathbb{P}\left\{P^{-1/2}\|\boldsymbol{\gamma}\| \leq r\right\}\right| \\ &\leq \mathrm{C}P^{-1/2} + \mathrm{C}n^{-\frac{2\min\{s,s_\circ-s\}}{2s+1}}\|f^*\|^2_{\mathcal{H}^{s_\circ}}.\end{aligned}$$

Proof. Using Fisher expansion (Lemma E.1) we have

$$\left(\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta}^*_{\boldsymbol{\rho}}\right) = \frac{1}{S}D_\rho^{-2}(p)\left(\sum_{i\in S_p}\psi_i\varepsilon_i\right)$$

and hence by Lemma D.2 we have

$$M\left(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*_{\boldsymbol{\rho}}\right) \geq \frac{1}{n(1+\delta)}M\left(M^2+\rho I\right)^{-1}\left(\sum_i \psi_i\varepsilon_i\right).$$

In the same way we obtain the less-or-equal inequality constituting the first part of the claim. The bound for $\operatorname{tr}(\Sigma)$ is obtained by Lemma D.3

$$\operatorname{tr}(S\Sigma) \leq \mathrm{C}\sigma^2(1-\delta)^{-1}\sum_j \frac{\mu_j^2}{(\mu_j+\rho)^2} \leq \mathrm{C}\sigma^2\rho^{-1/(2s)}.$$

This also bounds the six largest eigenvalues of $S\Sigma$ away from zero.

Next we use Lemma A.2, Lemma D.2 and Lemma E.1 in the same way we did in the proof of Lemma A.3 and have for an arbitrary p and all positive x

$$\mathbb{P}\left\{\|\mathbf{g}(p)\| \geq \mathrm{C}\mathfrak{g}\sqrt{\frac{\mathrm{x}}{\rho^{1/(2s)}S}}\right\} \leq e^{-\mathrm{x}}.$$

Hence, using integrated tail probability expectation formula we have

$$\mathbb{E}\left[\|\mathbf{g}(p)\|^3\right] \leq \left(\frac{\mathrm{C}^2\mathfrak{g}^2}{\rho^{1/(2s)}S}\right)^{3/2}.$$

By Lemma D.3

$$\mathbb{E}\left[\|\mathbf{g}(p)\|^2\right] \asymp \frac{\mathrm{C}^2\mathfrak{g}^2}{\rho^{1/(2s)}S}.$$

Now we are ready to apply Lemma E.3 which yields

$$\left|\mathbb{P}\left\{\left\|\bar{f}-f^*_\rho\right\|_2 \leq r\right\} - \mathbb{P}\left\{P^{-1/2}\|\boldsymbol{\gamma}\| \leq r\right\}\right| \leq \mathrm{C}\frac{\mathfrak{g}^3}{\sigma^3}P^{-1/2}.$$

The last step is to apply Lemma E.2 accounting for the mis-tie between f^* and f^*_ρ, which is controlled by Lemma A.1. □

Lemma C.2. *For all* $\mathrm{t} > 2.6$ *on a set of probability at least* $1 - e^{-\mathrm{t}/2} - P^{-3}$

$$S\left\|\hat{\Sigma}-\Sigma\right\| \leq \Delta(P,\mathrm{t}) := \mathrm{C}\sigma^2\sqrt{\mathrm{t}}\rho^{-1/(2s)}P^{-1/2}\log P.$$

Proof. Consider P i.i.d. matrices

$$\Omega_p = \mathbf{g}(p)\mathbf{g}(p)^T - \Sigma,$$

Lemma C.1 yields

$$(1+\delta)^{-1}M^4\big(M^2+\rho I\big)^{-2} \leq \Sigma \leq (1-\delta)^{-1}M^4\big(M^2+\rho I\big)^{-2}.$$

On a set of probability at least $1-P^{-3}$ for all p we have

$$\operatorname{tr}\left(\mathbf{g}(p)\mathbf{g}(p)^T\right) \leq \mathrm{C}\frac{\mathfrak{g}^2\log P}{\rho^{1/(2s)}S},$$

at the same time by Lemma C.1

$$\operatorname{tr}(\Sigma) \leq \mathrm{C}\sigma^2\rho^{-1/(2s)}S^{-1}$$

and hence

$$\operatorname{tr}(\Omega_p) \leq \mathrm{C}\sigma^2\rho^{-1/(2s)}S^{-1}\log P.$$

Clearly,

$$\left\|\Omega_p^2\right\| \leq \operatorname{tr}\left(\Omega_p^2\right) \leq C\sigma^2\rho^{-1/s}S^{-2}\log^2 P.$$

The rest is due to Lemma E.4. □

Lemma C.3. *On the same set which the claim of Lemma C.2 holds on (of probability at least* $1-e^{-\mathrm{t}/2}-P^{-3}$*) for an arbitrary* $\mathrm{t}>2.6$

$$S\left\|\Sigma-\hat{\Sigma}\right\|_1 \leq \mathrm{C}\frac{\Delta(P,\mathrm{t})^{1-1/(4s)}}{\rho^{1/(2s)}}.$$

Proof. Denote $\Delta := \Delta(P,\mathrm{t})$.

$$\begin{aligned}
S\left\|\Sigma-\hat{\Sigma}\right\|_1 &= S\left(\sum_{j^{2s}\leq 1/(\rho\sqrt{\Delta})} + \sum_{j^{2s}>1/(\rho\sqrt{\Delta})}\right)\left|\lambda_j(\Sigma-\hat{\Sigma})\right| \\
&\leq \mathrm{C}\frac{\Delta^{1-1/(4s)}}{\rho^{1/(2s)}} + \frac{\mathrm{C}}{\rho^2}\int_{u^{2s}>1/(\rho\sqrt{\Delta})}\frac{du}{u^4 s} \\
&\leq \mathrm{C}\frac{\Delta^{1-1/(4s)}}{\rho^{1/(2s)}}.
\end{aligned}$$

□

Lemma C.4. *Consider* $\boldsymbol{\gamma}^\flat \sim \mathcal{N}(0,\hat{\Sigma})$. *Then uniformly for positive* r

$$\left|\mathbb{P}^\flat\left\{\left\|\bar{\mathbf{f}}^\flat-\bar{\mathbf{f}}\right\|<r\right\} - \mathbb{P}^\flat\left\{P^{-1/2}\left\|\boldsymbol{\gamma}^\flat\right\|<r\right\}\right| \leq \mathrm{C}\frac{\mathfrak{g}^3}{\sigma^3}P^{-1/2}.$$

Proof. The proof consists in applying CLT. In order to estimate the moments involved in its residual term we apply Theorem 3.1 to the local estimates $\hat{f}_p$ and obtain on a set of probability at least $1-Pe^{-\mathrm{x}}-Pe^{-\mathrm{t}/2}$ (choosing $2\mathrm{x}=\mathrm{t}=4\log P$) for S large enough

$$\|\mathbf{g}(p)\| \leq 1,$$

which enables us to apply Hoeffding's inequality, yielding concentration for u chosen as $\sqrt{\log P}$

$$\mathbb{P}\left\{\left|\mathbb{E}^\flat\left[\|\mathbf{g}(p)\|^2\right]-\mathbb{E}\left[\|\mathbf{g}(p)\|^2\right]\right|>\mathrm{u}P^{-1/2}\right\} \leq 2e^{-2\mathrm{u}^2}$$

and also

$$\mathbb{P}\left\{\left|\mathbb{E}^\flat\left[\|\mathbf{g}(p)\|^3\right]-\mathbb{E}\left[\|\mathbf{g}(p)\|^3\right]\right|>\mathrm{u}P^{-1/2}\right\}\leq 2e^{-2\mathrm{u}^2}.$$

Now we can bound the moments involved in Lemma E.3

$$\mathbb{E}^\flat\left[\left\|\hat{\mathbf{f}}^\flat(p)-\bar{\mathbf{f}}\right\|^2\right]\asymp\mathbb{E}\left[\left\|\hat{\mathbf{f}}(p)-\mathbf{f}^*\right\|^2\right]\asymp\left(\sigma\rho^{-1/(2s)}\right)^2,$$

$$\mathbb{E}^\flat\left[\left\|\hat{\mathbf{f}}^\flat(p)-\bar{\mathbf{f}}\right\|^3\right]\asymp\mathbb{E}\left[\left\|\hat{\mathbf{f}}(p)-\mathbf{f}^*\right\|^3\right]\asymp\left(\mathfrak{g}\rho^{-1/(2s)}\right)^3.$$

Lemma C.2 (choose $\mathrm{t}=2\log P$) demonstrates boundness of the six largest eigenvalues of $\hat{\Sigma}$ away from zero. Finally, we use Lemma D.4 to account for the conditioning (probability of the set is at least $1-1/P$). □

Proof of Theorem 3.3. Finally, we are in position to demonstrate closeness of $\mathbb{P}$ and $\mathbb{P}^\flat$. We apply Lemma C.4 and Lemma C.1, obtaining two Gaussian approximations and compare them by the means of Lemma E.2. We use Lemma C.3 (choose $\mathrm{t}=2\log P$) to establish closeness of the covariance operators of the limiting distributions. Finally, account for conditioning (Lemma D.4). □

D Technical Results

The following lemma aids to bound the right-hand side of Fisher expansion (Lemma E.1).

Lemma D.1. *For all positive* x

$$\mathbb{P}\left\{\left\|\frac{1}{n}\left(M^2+\rho I\right)^{-1/2}\Psi\boldsymbol{\varepsilon}\right\|\leq\mathrm{C}\mathfrak{g}\sqrt{\frac{1+\sqrt{\mathrm{x}}+2\mathrm{x}}{\rho^{1/(2s)}n}}\right\}\geq 1-e^{-\mathrm{x}}.$$

Proof. Trivially,

$$\left\|\frac{1}{n}\left(M^2+\rho I\right)^{-1/2}\Psi\boldsymbol{\varepsilon}\right\|^2=\left\|\boldsymbol{\varepsilon}^T\underbrace{\frac{1}{n^2}\Psi^T\left(M^2+\rho I\right)^{-1}\Psi}_{A}\boldsymbol{\varepsilon}\right\|.$$

Now we bound every diagonal element of the matrix A.

$$\max_i A_{ii}\leq\frac{C_\phi^2}{n^2}\sum_{j=1}^{\infty}\frac{\mu_j}{\mu_j+\rho}.$$

Now employ Lemma D.3, which holds due to Assumption 3.1, and obtain

$$\max_i A_{ii}\leq\mathrm{C}\rho^{-1/(2s)}n^{-2}.$$

Therefore,

$$\mathrm{tr}\left(A\right)\leq\mathrm{C}\rho^{-1/(2s)}n^{-1}$$

and trivially

$$\mathrm{tr}\left(A^2\right)\leq\left(\mathrm{C}\rho^{-1/(2s)}n^{-1}\right)^2,$$

$$\|A\|_F=\sqrt{\mathrm{tr}\left(A^TA\right)}\leq\mathrm{C}\rho^{-1/(2s)}n^{-1}.$$

Finally, we are ready to employ Hanson-Wright inequality [36], constituting the claim. □

Below we demonstrate that Assumption A.1 holds with high probability under general assumptions.

Lemma D.2. *Let $\{\phi_j\}$ and $\{\mu_j\}$ be eigenfunctions and eigenvalues of $k(\cdot,\cdot)$ w.r.t. π. Then Assumption A.1 holds for some $\mathrm{C} > 0$ and arbitrary $\mathrm{t} > 2.6$ with*

$$\delta = \mathrm{C}\rho^{-1/(2s)}\left(\sqrt{\frac{\mathrm{t}}{n}} + \frac{\mathrm{t}}{n}\right)$$

on a set of probability at least $1 - e^{-\mathrm{t}/2}$.

Proof. Consider matrices $\Psi^i \in \mathbb{R}^{\infty\times\infty}$ s. t. $\Psi^i_{jk} = \sqrt{\mu_j\mu_k}\phi_j(X_i)\phi_k(X_i)$. Denote

$$\Omega_i = \left(M^2 + \rho I\right)^{-1/2}\left(\Psi^i + \rho I\right)\left(M^2 + \rho I\right)^{-1/2} - I.$$

Observe

$$\frac{1}{n}\sum_i^n \Omega_i = \left(M^2 + \rho I\right)^{-1/2}\left(\frac{1}{n}\Psi^T\Psi + \rho I\right)\left(M^2 + \rho I\right)^{-1/2} - I.$$

Due to the fact that the eigenfunctions are normalized, $\mathbb{E}\left[\Omega_i\right] = 0$.

Below j and k, being summation indexes, always run from 1 to ∞ unless specified otherwise. Clearly the maximum eigenvalue of a p.s.d. matrix does not exceed its trace. Hence, using Lemma D.3

$$\begin{aligned}\left\|\left(M^2 + \rho I\right)^{-1/2}\Psi^i\left(M^2 + \rho I\right)^{-1/2}\right\| &\le \mathrm{C}\sum_j \frac{\mu_j}{\mu_j + \rho}\\ &\le \mathrm{C}\rho^{-1/(2s)}\end{aligned}$$

and

$$\begin{aligned}\left\|\left(M^2 + \rho I\right)^{-1/2}\rho I\left(M^2 + \rho I\right)^{-1/2}\right\| &\le \rho\max_j \frac{1}{\mu_j + \rho}\\ &\le 1.\end{aligned}$$

Further, using the choice of ρ we have

$$\begin{aligned}\|\Omega_i\| &\le \mathrm{C}\rho^{-1/(2s)} + 2\\ &\le \mathrm{C}\rho^{-1/(2s)}.\end{aligned}$$

Now the goal is to bound $\mathrm{tr}\left(\mathbb{E}\left[\Omega_i^2\right]\right)$. First, we observe

$$\mathbb{E}\left[\left(M^2 + \rho I\right)^{-1/2}\left(\Psi^i + \rho I\right)\left(M^2 + \rho I\right)^{-1/2}\right] = I$$

due to the fact that $\mathbb{E}\left[\Psi^i\right] = M^2$. Therefore,

$$\mathbb{E}\left[\Omega_i^2\right] = \mathbb{E}\left[\underbrace{\left(\left(M^2 + \rho I\right)^{-1/2}\left(\Psi^i + \rho I\right)\left(M^2 + \rho I\right)^{-1/2}\right)^2}_{A}\right] - I.$$

For its trace, using $\mathbb{I}[\cdot]$ to denote an indicator, we have

$$\begin{aligned}\operatorname{tr}\left(\mathbb{E}\left[\Omega_i^2\right]\right) &\le \sum_j \left(\sum_k \frac{\left(C_\phi^2 \sqrt{\mu_j \mu_k} + \rho \mathbb{I}[j=k]\right)^2}{(\mu_j+\rho)(\mu_k+\rho)} - 1 \right) \\ &\le \sum_j \sum_k \frac{\left(C_\phi^2 \sqrt{\mu_j \mu_k}\right)^2}{(\mu_j+\rho)(\mu_k+\rho)} + \left| \sum_j \left(\left(\frac{C_\phi^2 \mu_j + \rho}{\mu_j + \rho} \right)^2 - 1 \right) \right| \\ &=: T_1 + T_2.\end{aligned}$$

Using Lemma D.3 twice, relying on the choice of ρ we have

$$\begin{aligned}T_1 &\le \mathrm{C} \sum_j \left(\frac{\mu_j}{\mu_j+\rho} \sum_k \frac{\mu_k}{\mu_k+\rho} \right) \\ &\le \mathrm{C} \sum_j \frac{\mu_j}{\mu_j+\rho} \times \rho^{-1/(2s)} \\ &\le \mathrm{C} \rho^{-1/s}.\end{aligned}$$

The treatment of the second term uses decay of μ_j

$$\begin{aligned}T_2 &\le \mathrm{C} \left| \sum_j \frac{\mu_j^2 + \rho \mu_j}{(\mu_j + \rho \mu_j)^2} \right| \\ &\le \mathrm{C} \sum_j \frac{\rho \mu_j}{\mu_j^2 + 2\mu_j \rho + \rho^2} \\ &\le \mathrm{C} \sum_j \frac{\mu_j}{\mu_j + \rho} \\ &\le \mathrm{C} \rho^{-1/(2s)},\end{aligned}$$

where Lemma D.3 was used again. Therefore, we have

$$\left\| \mathbb{E}\left[\Omega_i^2\right] \right\| \le \operatorname{tr}\left(\mathbb{E}\left[\Omega_i^2\right]\right) \le \mathrm{C} \rho^{-1/s}.$$

Finally, apply Lemma E.4, demonstrating that Assumption A.1 holds for

$$\delta = \mathrm{C} \rho^{-1/(2s)} \left(\sqrt{\frac{\mathrm{t}}{n}} + \frac{\mathrm{t}}{n} \right)$$

with probability at least $1 - e^{-\mathrm{t}/2}$. □

The next lemma, being almost folklore (see [53, 49, 39]), bounds the effective dimensionality.

Lemma D.3.

$$\sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \rho} \asymp \rho^{-1/(2s)}.$$

Proof.

$$\sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \rho} \asymp \underbrace{\sum_{j^{2s} \le 1/\rho} \frac{1}{1 + j^{2s}\rho}}_{T_1} + \underbrace{\sum_{j^{2s} > 1/\rho} \frac{1}{1 + j^{2s}\rho}}_{T_2}.$$

$$
\begin{aligned}
T_1 &\asymp \rho^{-1/(2s)}.\\
T_2 = \frac{1}{\rho}\sum_{j^{2s}>1/\rho}\frac{1}{1/\rho + j^{2s}} &\\
&\asymp \frac{1}{\rho}\sum_{j^{2s}>1/\rho}\frac{1}{j^{2s}}\\
&\asymp \frac{1}{\rho}\int_{u=\rho^{-1/(2s)}}^{+\infty}\frac{du}{u^{2s}}\\
&\asymp \rho^{-1/2s}.
\end{aligned}
$$

□

The next lemma quantifies how much a measure of a set changes after conditioning.

Lemma D.4. *Consider a measure $\mathbb{P}$ and two measurable sets A and B. Then*

$$
\left|\mathbb{P}\left\{A\right\} - \mathbb{P}\left\{A|B\right\}\right| \leq 2\mathbb{P}\left\{\bar{B}\right\}.
$$

Proof.

$$
\begin{aligned}
\left|\mathbb{P}\left\{A\right\} - \mathbb{P}\left\{A|B\right\}\right| &= \left|\mathbb{P}\left\{A|B\right\}\mathbb{P}\left\{B\right\} + \mathbb{P}\left\{A|\bar{B}\right\}\mathbb{P}\left\{\bar{B}\right\} - \mathbb{P}\left\{A|B\right\}\right|\\
&= \left|\mathbb{P}\left\{A|B\right\}\left(\mathbb{P}\left\{B\right\} - 1\right) + \mathbb{P}\left\{A|\bar{B}\right\}\mathbb{P}\left\{\bar{B}\right\}\right|\\
&\leq 2\mathbb{P}\left\{\bar{B}\right\}.
\end{aligned}
$$

□

E Tools

This section briefly cites the results we relied upon.

E.1 Consistency of penalized maximum likelihood estimation

Consider a quadratic concave likelihood $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, Y)$ for $\boldsymbol{\theta}$ being an infinite-dimensional parameter and Y denoting the random data. The following result quantifies the mis-tie between the penalized sample-level estimate

$$
\hat{\boldsymbol{\theta}} \coloneqq \arg\max_{\boldsymbol{\theta}} \underbrace{L(\boldsymbol{\theta}) - \rho\left\|\boldsymbol{\theta}\right\|^2}_{L_\rho(\boldsymbol{\theta})},
$$

penalized population-level estimate

$$
\boldsymbol{\theta}^*_{\boldsymbol{\rho}} \coloneqq \arg\max_{\boldsymbol{\theta}} \mathbb{E}\left[L(\boldsymbol{\theta})\right] - \rho\left\|\boldsymbol{\theta}\right\|^2
$$

and non-penalized population-level estimate

$$
\boldsymbol{\theta}^* \coloneqq \arg\max_{\boldsymbol{\theta}} \mathbb{E}\left[L(\boldsymbol{\theta})\right].
$$

Also define

$$
D^2_\rho \coloneqq -\nabla^2\left(\mathbb{E}\left[L(\boldsymbol{\theta})\right] - \rho\left\|\boldsymbol{\theta}\right\|^2\right) \text{ and } \nabla\zeta \coloneqq \nabla\zeta(\boldsymbol{\theta}) = \nabla\left(L(\boldsymbol{\theta}) - \mathbb{E}\left[L(\boldsymbol{\theta})\right]\right).
$$

Lemma E.1. *Fisher expansion holds:*

$$D_\rho \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*_{\boldsymbol{\rho}} \right) = D_\rho^{-1} \nabla \zeta.$$

Proof. Using Taylor expansion around the stationary point $\hat{\boldsymbol{\theta}}$ we have

$$\nabla L_\rho(\boldsymbol{\theta}) = -D_\rho^2 \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right).$$

Now we notice $\nabla \mathbb{E}\left[L_\rho(\boldsymbol{\theta}^*_{\boldsymbol{\rho}}) \right] = 0$ and obtain

$$\nabla L_\rho(\boldsymbol{\theta}^*_{\boldsymbol{\rho}}) = \nabla \zeta = -D_\rho^2 \left(\boldsymbol{\theta}^*_{\boldsymbol{\rho}} - \hat{\boldsymbol{\theta}} \right).$$

Finally, relying on the fact that $\rho > 0$, thus D_ρ^2 is invertible, we multiply both sides of the last equation by D_ρ^{-1}. □

This result has been generalized in [39] (Theorem 2.2).

E.2 Gaussian Comparison

Lemma E.2 (Theorem 2.1 by [17])**.** *Consider two centered Gaussian elements of a Hilbert space ζ and η with covariance operators Σ_1 and Σ_2 and a deterministic Hilbert element a. Denote*

$$\kappa(\Sigma) = \left(\sqrt{\sum_{j>1} \lambda_j^2(\Sigma)} \sqrt{\sum_{j \geq 1} \lambda_j^2(\Sigma)} \right)^{-1/2}.$$

Then

$$\sup_{r>0} \left| \mathbb{P}\left\{ \|\zeta - a\| < r \right\} - \mathbb{P}\left\{ \|\eta\| < r \right\} \right| \leq \mathrm{C} \left(\kappa(\Sigma_1) + \kappa(\Sigma_2) \right) \left(\|\Sigma_1 - \Sigma_2\|_1 + \|a\|^2 \right).$$

E.3 Central Limit Theorem

Lemma E.3 (Corollary of the main theorem by [51])**.** *Consider centered X_1, X_2, .., X_n being i.i.d. elements of Hilbert space with covariance operator V and a centered Gaussian element Y with the same covariance operator. Then*

$$\sup_{r>0} \left| \mathbb{P}\left\{ \left\| n^{-1/2} \sum_{i=1}^n X_i \right\| < r \right\} - \mathbb{P}\left\{ \|Y\| < r \right\} \right| \leq R,$$

where

$$R := \mathrm{C} \left(\prod_{i=1}^6 \lambda_i(V) \right)^{-1} \mathbb{E}\left[\|X_1\|^2 \right]^{-3/2} \mathbb{E}\left[\|X_1\|^3 \right] n^{-1/2}.$$

E.4 Bernstein matrix inequality

Lemma E.4 (Corollary of Theorem 3.3 [21]). *Consider centered i.i.d. matrices X_1, X_2, .., X_n. Let for some positive A and B*

$$\lambda_1(X_i) \leq A,$$
$$\lambda_1\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right]\right) \leq B,$$
$$\operatorname{tr}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right]\right) \leq B.$$

Then for any $\mathrm{t} > 0$

$$\mathbb{P}\left\{\lambda_1\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) > \sqrt{\frac{2B\mathrm{t}}{n}} + \frac{At}{3n}\right\} \leq \mathrm{t}\left(e^{\mathrm{t}} - \mathrm{t} - 1\right)^{-1}.$$

Note, for positive $\mathrm{t} \geq 2.6$

$$\mathrm{t}\left(e^{\mathrm{t}} - \mathrm{t} - 1\right)^{-1} \leq e^{-\mathrm{t}/2}.$$