Random Features for Operator-Valued Kernels: Bridging Kernel Methods and Neural Operators

Mike Nguyen
Technical University of Braunschweig

Nicole Mücke
Technical University of Braunschweig

Abstract

In this work, we investigate the generalization properties of random feature methods. Our analysis extends prior results for Tikhonov regularization to a broad class of spectral regularization techniques and further generalizes the setting to operator-valued kernels. This unified framework enables a rigorous theoretical analysis of neural operators and neural networks through the lens of the Neural Tangent Kernel (NTK). In particular, it allows us to establish optimal learning rates and provides a good understanding of how many neurons are required to achieve a given accuracy. Furthermore, we establish minimax rates in the well-specified case and also in the misspecified case, where the target is not contained in the reproducing kernel Hilbert space. These results sharpen and complete earlier findings for specific kernel algorithms.

1 INTRODUCTION

Operator learning has become a powerful paradigm in machine learning. It is particularly well suited for surrogate modeling in areas such as uncertainty quantification, inverse problems, and design optimization. The central objective in these applications is to approximate potentially nonlinear operators, for instance solution operators of partial differential equations (PDEs), see Kovachki et al. (2023a). Among the most widely used approaches in this context are *Neural Operators* (NOs), which generalize classical neural networks to the broader task of learning operators, i.e., mappings between (potentially) infinite-dimensional function spaces (Kovachki et al., 2023b; Li et al., 2021c; Qin et al., 2024; Wang et al., 2021; Raissi et al., 2019; Li et al., 2020; Sharma et al., 2024).

Despite their practical success in scientific computing, the theoretical understanding of Neural Operators (NOs) is still limited. Existing work has primarily focused on approximation properties (Huang et al., 2024; Schwab and Zech, 2021; Marcati and Schwab, 2023; Kovachki et al., 2021, 2024), while generalization results are comparatively scarce (see, e.g., Kim and Kang (2022); Lara Benitez et al. (2024)). Recent progress has been made in the neural tangent kernel (NTK) regime, where Nguyen and Mücke (2024) established minimax rates for operator learning. At the same time, gradient descent (GD) dynamics suggest a connection to random feature approximations of vector-valued reproducing kernels, but convergence guarantees for such methods are not yet available. To date, only Tikhonov regularization has been analyzed in this setting (Lanthaler and Nelsen, 2023). In this work, we contribute by deriving minimax rates for a broad class of spectral regularization schemes, including GD, thereby enabling the derivation of generalization bounds for NOs.

In addition to bridging this theoretical gap, random feature approximation (RFA) is of independent interest. Kernel methods remain state-of-the-art in many non-parametric statistical applications and provide an elegant framework for developing new theoretical insights (Lin and Cevher, 2018; Lin et al., 2020a; Zhang et al., 2024). However, their benefits come at a substantial computational cost, making them infeasible for large-scale datasets. Classical kernelized algorithms require storing the kernel Gram matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, with entries $\mathbf{K}_{i,j} = K(u_i, u_j)$ for kernel function $K(\cdot,\cdot)$ and data points u_i, u_j. This entails a memory cost of $O(n^2)$ and a time cost of up to $O(n^3)$, where n denotes the dataset size (Schoelkopf and Smola, 2002).

RFA alleviates these costs by exploiting integral representations of kernels that can be approximated via

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

finite sums of random features. For kernel ridge regression (KRR), this reduces memory and computational costs to $O(nM)$ and $O(nM^2)$, respectively, where M is the number of random features (Rudi and Rosasco, 2016). For gradient descent (GD), the cost becomes $O(nMt)$, with t denoting the number of iterations, and can be further reduced by acceleration methods such as Heavy-Ball or Nesterov, which achieve the same generalization error as GD in only $\sqrt{t}$ iterations (Pagliana and Rosasco, 2019). Our analysis further suggests that M should typically exceed t, which explains why random feature methods often perform best when combined with iterative regularization schemes. This observation highlights the need for rigorous theoretical guarantees in such settings.

For kernel ridge regression (KRR), the central question of how many random features are required to achieve optimal convergence rates has been studied extensively (Rahimi and Recht, 2007; Li et al., 2021b; Zhen et al., 2020; Rudi and Rosasco, 2016; Lanthaler and Nelsen, 2023). Rahimi and Recht (2007) first established optimal rates for $M = O(n)$ random features in the case of real-valued kernels (rvk). This was later improved by Rudi and Rosasco (2016) to $M = O(\sqrt{n}\log n)$, and further extended to stochastic kernel ridge regression in Carratino et al. (2019). Most recently, Lanthaler and Nelsen (2023) removed the logarithmic factor, showing that $M = O(\sqrt{n})$ suffices in the more general setting of vector-valued kernels (vvk). A detailed comparison is provided in Table 1. These results for M hold in the well-specified case where the target function belongs to the RKHS. The question of how many random features are required for broader smoothness classes, where the target function may lie outside the RKHS, remains open.

Contribution. Our main motivation for studying vector-valued kernels and random feature approximation (RFA) is to derive generalization guarantees for Neural Operators (NOs).

To the best of our knowledge, RFA with vector-valued kernels has previously been analyzed mainly in the setting of KRR. In this work, we develop a unified framework based on spectral filtering (Caponnetto and De Vito, 2007), which yields optimal convergence rates for a broad class of learning methods with either explicit or implicit regularization. This includes gradient descent and acceleration techniques, and it recovers as special cases the KRR results of Rudi and Rosasco (2016) for real-valued kernels and of Lanthaler and Nelsen (2023) for vector-valued kernels.

Our framework further accommodates kernels represented as sums of integral kernels, covering in particular operator-valued neural tangent kernels. This extension provides, for the first time, convergence rates for random feature methods with vector-valued kernels beyond KRR, thereby opening the door to rigorous statistical guarantees for NOs in the NTK regime.

A key advantage of our approach is that both the convergence rates and the number of random features required for optimality are independent of the dimension of the input space. This makes the results directly applicable to NOs, where inputs are functions rather than finite-dimensional vectors. At the same time, the number of random features scales only quadratically with the feature dimension of the combined input representation per neuron, yielding generalization guarantees for NOs that combine minimax-optimal statistical rates with computational efficiency.

Table 1: Comparison of random feature requirements (M) for achieving generalization error of order $O(n^{-\frac{1}{2}})$. The last column indicates the smoothness classes where optimal rates are known; see Assumptions 3.2, 3.3 for the meaning of $r, b > 0$. [1]=Rahimi and Recht (2007), [2]=Rudi and Rosasco (2016), [3]=Lanthaler and Nelsen (2023)

References	M	Method	Smoothness
[1]	$O(n)$	KRR (rvk)	$r \in [0.5, 1]$
[2]	$O(\sqrt{n}\log n)$	KRR (rvk)	$r \in [0.5, 1]$
[3]	$O(\sqrt{n})$	KRR (vvk)	$r = 0.5$
Our	$O(\sqrt{n}\log n)$	Spec. (vvk)	$2r + b > 1$

The rest of the paper is organized as follows. Section 2 introduces the setting, recalls key definitions in the context of random feature methods, and motivates the framework by linking it to Neural Operators. Section 3 presents and discusses our main results. Section 4 concludes with a summary of our findings. Appendix A provides further details on learning with Neural Operators and includes numerical illustrations that support our theoretical results. All proofs are deferred to Appendix B.

Notation. By $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ we denote the space of bounded linear operators between real separable Hilbert spaces $\mathcal{H}_1$, $\mathcal{H}_2$. We write $\mathcal{L}(\mathcal{H}, \mathcal{H}) = \mathcal{L}(\mathcal{H})$. For $\Gamma \in \mathcal{L}(\mathcal{H})$, we denote by Γ^* the adjoint operator. If $h \in \mathcal{H}$, we write $h \otimes h := \langle \cdot, h\rangle h$. For a compact operator $\Gamma \in \mathcal{L}(\mathcal{H})$, the trace is defined by $\mathrm{tr}(\Gamma) = \sum_{k=1}^{\infty}\langle \Gamma e_k, e_k\rangle$, where $\{e_k\}_{k=1}^{\infty}$ is any orthonormal basis of $\mathcal{H}$. We denote by $\mathcal{F}(\mathcal{U}, \mathcal{V})$ the space of measurable operators from $\mathcal{U}$ to $\mathcal{V}$. We write $L^2(\mathcal{U}, \rho_{\mathcal{U}}) := L^2(\mathcal{U}, \rho_{\mathcal{U}}; \mathcal{V})$ for the L^2 space equipped with the norm $\|f\|^2_{L^2(\rho_{\mathcal{U}})} := \int_{\mathcal{U}} \|f(u)\|^2_{\mathcal{V}}\, d\rho_{\mathcal{U}}(u)$. We let $[n] := \{1, \dots, n\}$ and denote the output vector by $\mathbf{v} = (v_1, \dots, v_n) \in \mathcal{V}^n$ with norm $\|\mathbf{v}\|_2^2 := \sum_{i=1}^{n} \|v_i\|^2_{\mathcal{V}}$.

2 MATHEMATICAL FRAMEWORK

We consider an input space $(\mathcal{U}, ||\cdot||_{\mathcal{U}})$, where $\mathcal{U}$ is a Banach space, and an output space $(\mathcal{V}, ||\cdot||_{\mathcal{V}})$, where $\mathcal{V}$ is a separable Hilbert space. These assumptions facilitate the use of the theory of vector-valued kernels (Carmeli et al., 2005, 2008). The data space is given by $\mathcal{Z} = \mathcal{U} \times \mathcal{V}$, equipped with an unknown distribution ρ. We denote by $\rho_{\mathcal{U}}$ the marginal distribution on $\mathcal{U}$, and by $\rho(\cdot \mid u)$ the regular conditional distribution on $\mathcal{V}$ given $u \in \mathcal{U}$; see Shao (2003).

Given a measurable operator $G : \mathcal{U} \to \mathcal{V}$ we further define the expected risk as

$$\mathcal{E}(G) := \mathbb{E}[\ell(G(u), v)] \,, \tag{2.1}$$

where the expectation is taken w.r.t. the distribution ρ and $\ell : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_+$ is the least-square loss $\ell(v, v') = \frac{1}{2}\|v - v'\|_{\mathcal{V}}^2$. It is known that the global minimizer of $\mathcal{E}$ over the set of all measurable functions is given by the regression operator $G_\rho(u) = \int_{\mathcal{V}} v\rho(dv|u)$.

We consider a standard statistical learning setting where we are given data $(u_j, v_j)_{j=1}^n$, sampled identically and independently with respect to ρ on $\mathcal{U} \times \mathcal{V}$.

2.1 Motivation: Generalization Bounds for Neural Operators

Shallow NOs. To connect shallow Neural Operators (NOs) with vector-valued kernels, it is useful to recall their definition and highlight how their training dynamics give rise to neural tangent kernels (NTKs). Following Nguyen and Mücke (2024), the class of two-layer NOs is defined as follows. Let $\mathcal{U}$ denote the function input space, mapping from the measure space $(\mathcal{X}, \rho_x)$ to $\mathcal{Y} \subset \mathbb{R}^{d_y}$, and let $\mathcal{V}$ be the target function space, containing measurable functions mapping from $(\mathcal{X}, \rho_x)$ to $\tilde{\mathcal{Y}} \subset \mathbb{R}$. For a network of width $M \in \mathbb{N}$, let $\sigma : \mathbb{R} \to \mathbb{R}$ be an activation function acting pointwise, and let $A : \mathcal{U} \to \mathcal{F}(\mathcal{X}, \mathbb{R}^{d_k})$ be a continuous operator. We then define the class of shallow NOs by

$$\begin{aligned}\mathcal{F}_M := \Big\{ G_\theta : \mathcal{U} \to \mathcal{V} \,\Big|\, & G_\theta(u)(x) \\ &= \frac{\big\langle a,\ \sigma\big(B_1 A(u)(x) + B_2 u(x) + B_3 c(x)\big)\big\rangle}{\sqrt{M}}, \\ & \theta = (a, B_1, B_2, B_3) \\ & \in \mathbb{R}^M \times \mathbb{R}^{M\times d_k} \times \mathbb{R}^{M\times d_y} \times \mathbb{R}^{M\times d_b}, \\ & c : \mathcal{X} \to \mathbb{R}^{d_b} \Big\}.\end{aligned}$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product. We collect all parameters in $\theta = (a, B_1, B_2, B_3) = (a, B)$, with $B = (B_1, B_2, B_3) \in \mathbb{R}^{M\times\tilde{d}}$, and *feature dimension* $\tilde{d} := d_k + d_y + d_b$. The goal is to minimize the expected risk (2.1) over the set $\mathcal{F}_M$, i.e.,

$$\min_{G_\theta \in \mathcal{F}_M} \mathbb{E}[\ell(G_\theta(u), v)] \,.$$

Relation to vector-valued kernels via the NTK. The connection between kernel methods and standard neural networks is established via the neural tangent kernel (NTK), see Jacot et al. (2018); Lee et al. (2019). For fully connected neural networks, it is known that under gradient descent (GD) training and in the infinite-width limit, the network function linearizes around initialization, and the dynamics are governed by the NTK.

In complete analogy, the training dynamics of shallow NOs correspond to kernel gradient descent in the reproducing kernel Hilbert space (RKHS) induced by a vector-valued kernel. We define for neural operators the NTK feature map

$$\Phi^M : \mathcal{U} \to \mathcal{F}(L^2(\mathcal{X}, \rho_x), \Theta), \qquad \Phi^M(u) := \Phi_u^M,$$

with random initialization $\theta_0 \in \Theta$ and

$$\Phi_u^M(v) := \nabla_\theta \big\langle G_{\theta_0}(u), v\big\rangle_{L^2(\rho_x)} \,, \qquad v \in L^2(\mathcal{X}, \rho_x).$$

For any $u, \tilde{u} \in \mathcal{U}$, the vector-valued kernel $K_M : L^2(\mathcal{X}, \rho_x) \to L^2(\mathcal{X}, \rho_x)$ is defined by

$$K_M(u, \tilde{u}) := (\Phi_u^M)^* \Phi_{\tilde{u}}^M$$

and expands as

$$\begin{aligned} K_M(u, \tilde{u}) = \frac{1}{M}\sum_{m=1}^{M} \psi_m(u) \otimes \psi_m(\tilde{u}) \\ + \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{\tilde{d}} \psi'_{m,j}(u) \otimes \psi'_{m,j}(\tilde{u}). \end{aligned}$$

Here, for $u \in \mathcal{U}$ and $x \in \mathcal{X}$, we set

$$\begin{aligned} J(u)(x) &:= \big(A(u)(x),\, u(x),\, c(x)\big)^\top, \\ \psi_m(u) &:= \sigma\big(\langle b_m^{(0)}, J(u)\rangle\big), \\ \psi'_{m,j}(u) &:= \sigma'\big(\langle b_m^{(0)}, J(u)\rangle\big) J(u)^{(j)} \end{aligned}$$

and denote the collection of initial weight vectors by $B^{(0)} = (b_1^{(0)}, \ldots, b_M^{(0)})^\top \in \mathbb{R}^{M\times\tilde{d}}$, see Nguyen and Mücke (2024).

Relation to Random Feature Approximation. The representation of K_M shows that the NTK of shallow NOs admits the form of a random feature approximation. Indeed, each term $\psi_m(u) \otimes \psi_m(\tilde{u})$ and $\psi'_{m,j}(u) \otimes \psi'_{m,j}(\tilde{u})$ constitutes a nonlinear random feature, sampled through the random initialization $b_m^{(0)}$.

Thus, K_M is a Monte–Carlo approximation of the limiting kernel

$$K(u,\tilde u) := \mathbb{E}_{\theta_0}\Big[\psi(u)\otimes\psi(\tilde u) + \sum_{j=1}^{\tilde d}\psi'_j(u)\otimes\psi'_j(\tilde u)\Big], \tag{2.2}$$

where $\psi(u) = \sigma\big(\langle b^{(0)}, J(u)\rangle\big)$, $\psi'_j(u) = \sigma'\big(\langle b^{(0)}, J(u)\rangle\big)J(u)^{(j)}$ and the expectation is taken w.r.t. the initialization distribution of θ_0, see Nguyen and Mücke (2024), Proposition 2.3.

Consequently, training shallow NOs with gradient descent in the NTK regime is equivalent to performing kernel gradient descent in the vector-valued RKHS associated with K, using K_M as a random feature approximation. This interpretation provides the bridge between Neural Operators and the statistical theory of random features for vector-valued kernels.

Generalization Bound. Assume a NO is trained with gradient descent by empirical risk minimization over the class $\mathcal{F}_M$ and let θ_t denote the parameter update after $t \in \mathbb{N}$ iterations. The excess risk of the corresponding neural operator $G_{\theta_t} \in \mathcal{F}_M$ can be decomposed as

$$\begin{aligned}&\|G_{\theta_t} - G_\rho\|^2_{L^2(\rho_{\mathcal{U}})} \\ &\lesssim \|G_{\theta_t} - F_t^M\|^2_{L^2(\rho_{\mathcal{U}})} + \|F_t^M - G_\rho\|^2_{L^2(\rho_{\mathcal{U}})}\ ,\end{aligned} \tag{2.3}$$

where F_t^M denotes the random-feature estimator associated with the vector-valued kernel K_M, trained with the same GD dynamics. The first term captures the discrepancy between the finite-width NO and its NTK-based random feature approximation, while the second term measures the generalization error of the random feature method itself. **The central objective of this paper is to establish optimal learning guarantees for the second term, and to extend the analysis to a broader class of regularization schemes and kernels.**

2.2 Kernel Methods, Random Features, and Regularization: General Setup

Vector-valued Kernels. Let $K : \mathcal{U}\times\mathcal{U} \longrightarrow \mathcal{L}(\mathcal{V})$ denote a reproducing *$\mathcal{V}$-valued kernel of positive type*, where $\mathcal{L}(\mathcal{V})$ denotes the Banach space of bounded linear operators on $\mathcal{V}$. For $u \in \mathcal{U}$, we write $K_u : \mathcal{V} \to \mathcal{F}(\mathcal{U},\mathcal{V})$ for the operator that maps $v \in \mathcal{V}$ to the function $K_u v \in \mathcal{F}(\mathcal{U},\mathcal{V})$ given by

$$(K_u v)(\tilde u) = K(\tilde u, u)v, \qquad \tilde u \in \mathcal{U}. \tag{2.4}$$

By $\mathcal{H}$ we denote the associated unique $\mathcal{V}$-valued RKHS on $\mathcal{U}$, which can be continuously included into $\mathcal{F}(\mathcal{U},\mathcal{V})$. We assume that K is *2-bounded*, implying that the inclusion $S : \mathcal{H} \hookrightarrow L^2(\mathcal{U},\rho_{\mathcal{U}};\mathcal{V})$ is a bounded linear map. A brief review of the key definitions related to vector-valued kernels is provided in Appendix A, see also Carmeli et al. (2005, 2008).

Assumption 2.1 (Kernel)**.** *Assume that the kernel K admits an integral representation of the form*

$$K(u,\tilde u) = \sum_{i=1}^{p}\int_\Omega \varphi_i(u,\omega)\otimes\varphi_i(\tilde u,\omega)\,d\pi(\omega), \tag{2.5}$$

where $\varphi_i : \mathcal{U}\times\Omega \to \mathcal{V}$ for $i = 1,\dots,p$, and (Ω,π) is a probability space. Moreover, assume that for all $u \in \mathcal{U}$,

$$\sum_{i=1}^{p}\|\varphi_i(u,\omega)\|^2_{\mathcal{V}} \le \kappa^2 \quad \pi - \textit{almost surely}\ . \tag{2.6}$$

Note that Assumption 2.1 implies the Hilbert–Schmidt bound $\|K(u,\tilde u)\|_{HS} \le \kappa^2$ $\rho_{\mathcal{U}}$-almost surely. Examples of (2.5) include the *Gaussian kernel* and *Random Fourier features* (Rahimi and Recht, 2007; Rudi and Rosasco, 2016). In contrast to these works, we allow for the additional sum over i, which enables us to cover special cases of Neural Tangent Kernels (2.2), see also Jacot et al. (2018); Nitanda and Suzuki (2020); Li et al. (2021a); Munteanu et al. (2022); Oymak and Soltanolkotabi (2019); Nguyen and Mücke (2023).

Random Feature Approximations. The idea of RFA is to approximate kernels that admit an integral representation (2.5) by a kernel represented by a finite sum, i.e., $K(u,\tilde u) \approx K_M(u,\tilde u)$ for $u,\tilde u \in \mathcal{U}$ and $M \in \mathbb{N}$, where

$$K_M(u,\tilde u) := \sum_{i=1}^{p}\frac{1}{M}\sum_{m=1}^{M}\varphi_i(u,\omega_m)\otimes\varphi_i(\tilde u,\omega_m).$$

Here, $\varphi_i : \mathcal{U}\times\Omega \to \mathcal{V}$ for some probability space (Ω,π), and $\{\omega_m\}_{m=1}^M$ are drawn i.i.d. from π.

The associated RKHS is denoted by $\mathcal{H}_M$. Under Assumption (2.6), the inclusion $\mathcal{S}_M : \mathcal{H}_M \hookrightarrow L^2(\mathcal{U},\rho_{\mathcal{U}};\mathcal{V})$ is a bounded linear map.

The main benefit of RFA is that it renders kernel methods computationally feasible on large datasets while preserving their statistical guarantees. Classical kernel methods require storing the full $n\times n$ Gram matrix and solving linear systems at cost $O(n^2)$ in memory and $O(n^3)$ in time. In contrast, RFA replaces the kernel with an explicit M-dimensional feature map, reducing the computational cost to $O(nM^2)$ (for ridge regression) or $O(nMt)$ (for t gradient descent iterations), with only $O(nM)$ memory. Statistically, it has been shown that for many kernels $M = O(\sqrt{n})$ features suffice to achieve minimax-optimal learning rates. Thus, RFA provides a scalable and flexible framework for kernel methods.

Regularization. Since the expected risk (2.1) cannot be minimized directly, the standard procedure is empirical risk minimization (ERM) over the hypothesis space $\mathcal{H}$,

$$\min_{F\in\mathcal{H}} \widehat{\mathcal{E}}(F), \qquad \widehat{\mathcal{E}}(F) = \frac{1}{n}\sum_{j=1}^{n} \ell\big(F(u_j), v_j\big).$$

However, direct ERM in an RKHS setting is typically ill-posed. To avoid overfitting and to obtain consistent estimators, *regularization* is introduced.

Definition 2.2 (Regularization function)**.** *Let* $\phi : (0,1]\times[0,1]\to\mathbb{R}$ *and set* $\phi_\lambda(t)=\phi(\lambda,t)$*. The family* $\{\phi_\lambda\}_\lambda$ *is called a family of* regularization functions *if there exist constants* $D, E, c_0 > 0$ *such that for all* $0<\lambda\leq 1$*:*

$$\text{i)}\quad \sup_{0<t\leq 1} |t\phi_\lambda(t)| \leq D, \tag{2.7}$$

$$\text{ii)}\quad \sup_{0<t\leq 1} |\phi_\lambda(t)| \leq \tfrac{E}{\lambda}, \tag{2.8}$$

$$\text{iii)}\quad \sup_{0<t\leq 1} |r_\lambda(t)| \leq c_0,\quad r_\lambda(t) := 1 - t\phi_\lambda(t). \tag{2.9}$$

This family of methods, known as *spectral regularization*, encompasses both explicit regularization, such as Tikhonov regularization, and implicit regularization through iterative schemes, including gradient descent and accelerated methods. Originally developed for (statistical) inverse problems (Engl et al., 1996), these techniques have since been applied in machine learning, in particular to non-parametric least-squares regression (Caponnetto and De Vito, 2007; Bauer et al., 2007; Blanchard and Mücke, 2017; Lin et al., 2020b).

It has been shown in Gerfo et al. (2008); Blanchard and Mücke (2017) that attainable learning rates are essentially determined by the *qualification* of the regularization $\{\phi_\lambda\}_\lambda$, i.e., the largest $\nu>0$ such that for all $q\in[0,\nu]$ and $0<\lambda\leq 1$:

$$\sup_{0<t\leq 1} |r_\lambda(t)|\, t^q \leq c_q\, \lambda^q, \tag{2.10}$$

for some constant $c_q > 0$.

A principled approach is to exploit the spectral structure of the empirical operators $\widehat{\Sigma}_M : \mathcal{H}_M \to \mathcal{H}_M$ and $\widehat{\mathcal{S}}_M^* : \mathcal{V}^n \to \mathcal{H}_M$, defined as

$$\widehat{\Sigma}_M = \frac{1}{n}\sum_{j=1}^{n} K_{M,u_j} K_{M,u_j}^*,$$

$$\widehat{\mathcal{S}}_M^* \mathbf{v} = \frac{1}{n}\sum_{j=1}^{n} K_{M,u_j} v_j.$$

With these operators, spectral regularization estimators combined with RFA take the form

$$F_\lambda^M = \phi_\lambda(\widehat{\Sigma}_M)\,\widehat{\mathcal{S}}_M^* \mathbf{v} \;\in\; \mathcal{H}_M. \tag{2.11}$$

3 MAIN RESULTS

3.1 Assumptions and Main Results

In this section we formulate our assumptions and state our main results.

Assumption 3.1 (Data Distribution)**.** *There exists positive constants* Q *and* Z *such that for all* $l\geq 2$ *with* $l\in\mathbb{N}$*,*

$$\int_{\mathcal{V}} \|v\|_{\mathcal{V}}^l \, d\rho(v\mid u) \leq \frac{1}{2} l! Z^{l-2} Q^2$$

$\rho_{\mathcal{U}}$*-almost surely.*

This assumption is satisfied, for example, if v is bounded almost surely. It further implies that the regression operator G_ρ is bounded almost surely, since

$$\|G_\rho(u)\|_{\mathcal{V}} \leq \int_{\mathcal{V}} \|v\|_{\mathcal{V}}\, d\rho(v\mid u) \leq \left(\int_{\mathcal{V}} \|v\|_{\mathcal{V}}^2\, d\rho(v\mid u)\right)^{\frac{1}{2}} \leq Q\,.$$

To characterize the smoothness of G_ρ relative to the kernel, we impose a so-called *source condition*. This condition links G_ρ to the spectral properties of the kernel integral operator and plays a central role in determining the attainable learning rates.

Denote by $\mathcal{L} : L^2(\mathcal{U},\rho_{\mathcal{U}}) \to L^2(\mathcal{U},\rho_{\mathcal{U}})$ the kernel integral operator associated to K, i.e.

$$\mathcal{L}G = \int_{\mathcal{U}} K_u G(u)\, \rho_{\mathcal{U}}(du).$$

Assumption 3.2 (Source Condition)**.** *Let* $R>0$*,* $r>0$*. We assume* $G_\rho = \mathcal{L}^r H$*, for some* $H\in L^2(\mathcal{U},\rho_{\mathcal{U}})$*, satisfying* $\|H\|_{L^2(\rho_{\mathcal{U}})} \leq R$ *.*

This condition links G_ρ to the spectral properties of the kernel integral operator and plays a central role in determining the attainable learning rates. The parameter $r>0$ quantifies the degree of smoothness of G_ρ: larger values of r correspond to higher regularity. In particular, the case $r=\frac{1}{2}$ corresponds to the well-specified setting where G_ρ lies in $\mathcal{H}$, while $r>\frac{1}{2}$ reflects additional smoothness, and $r<\frac{1}{2}$ corresponds to a misspecified setting. For more details, we refer to e.g. Bauer et al. (2007); Lin et al. (2020b).

While the source condition controls the regularity of the target function, a complementary assumption is required to capture the complexity of the hypothesis space. This is typically expressed in terms of the *effective dimension*, which measures the capacity of the RKHS relative to the kernel operator spectrum.

Assumption 3.3 (Effective Dimension)**.** *For some $b \in [0,1]$ and $c_b > 0$, assume that for all $\lambda > 0$ the operator $\mathcal{L}$ satisfies*

$$\mathcal{N}(\lambda) := \operatorname{tr}\left(\mathcal{L}(\mathcal{L}+\lambda I)^{-1}\right) \le c_b \lambda^{-b}. \tag{3.1}$$

Moreover, we assume that $2r + b > 1$.

Here, $\mathcal{N}(\lambda)$ is the *effective dimension* of the kernel, which quantifies the number of effective degrees of freedom of the hypothesis space. Intuitively, it reflects the decay of the eigenvalues of $\mathcal{L}$: smaller values of b correspond to faster decay (lower capacity), while larger values of b indicate slower decay and hence higher complexity. The condition (3.1) is always satisfied with $b = 1$, since $\mathcal{L}$ is trace class and its eigenvalues $\{\mu_i\}$ satisfy $\mu_i \lesssim i^{-1}$. If, more generally, the eigenvalues decay polynomially as $\mu_i \sim i^{-c}$ with $c > 1$, then (3.1) holds with $b = 1/c$; if $\mathcal{L}$ has finite rank, then $b = 0$. The case $b = 1$ is often called the *capacity-independent* case. Smaller values of b allow for faster convergence rates of the learning algorithms.

The choice of the regularization parameter λ and the number of random features M in (2.11) is crucial for balancing approximation, estimation, and optimization errors (see Appendix B). In practice, both parameters are determined as functions of the sample size n, in order to guarantee optimal statistical performance. The following result establishes that, under the source and capacity assumptions, our RF estimator achieves the minimax-optimal convergence rate. Moreover, it provides explicit conditions on λ_n and M_n that ensure the desired statistical guarantees. The proof is provided in Appendix B.

Theorem 3.4. *Suppose Assumptions 3.1–3.3 hold. Let $\{\phi_\lambda\}_\lambda$ be a family of regularization functions with qualification $\nu > 0$. Let $\delta \in (0,1)$ and choose*

$$\lambda_n = C\, n^{-\frac{1}{2r+b}} \log^3\!\left(\tfrac{2}{\delta}\right).$$

Then, with probability of at least $1-\delta$, the RF estimator (2.11) satisfies

$$\|G_\rho - \mathcal{S}_{M_n} F_{\lambda_n}^{M_n}\|_{L^2(\rho_{\mathcal{U}})} \;\le\; \bar{C}\, n^{-\frac{r}{2r+b}} \log^{3r+1}\!\left(\tfrac{1}{\delta}\right),$$

provided that $\nu \ge r \vee 1$, $n \ge n_0 := \exp\!\left(\frac{2r+b}{2r+b-1}\right)$, and the number of random features satisfies

$$M_n \;\ge\; p \cdot \tilde{C} \cdot \log(n) \cdot \begin{cases} n^{\frac{1}{2r+b}}, & r \in (0, \frac{1}{2}), \\ n^{\frac{1+b(2r-1)}{2r+b}}, & r \in [\frac{1}{2}, 1], \\ n^{\frac{2r}{2r+b}}, & r \in (1, \infty). \end{cases}$$

Here $C, \tilde{C}, \bar{C}$ are constants independent of n, M, λ.

Since our framework encompasses operator-valued RFA trained via GD, Theorem 3.4 can be directly applied to derive generalization bounds for NOs. Recall the excess-risk decomposition for shallow NOs in (2.3). Nguyen and Mücke (2024) showed that the first term is bounded by $O(\log n / M_n)$. Hence, when the number of neurons scales with the number of random features, i.e., $M_n = O\!\left(n^{\frac{2r}{2r+b}} \log n\right)$, Theorem 3.4 implies that NOs achieve the same minimax rates as nonparametric kernel methods.

Corollary 3.5 (Nguyen and Mücke (2024), Theorem. 3.5)**.** *Suppose the assumptions of Theorem 3.4 hold. Let $G_{\theta_{T_n}}$ denote the NO as defined in Section 2.1, where T_n is the number of GD iterations. Let M_n denote the network width. Assume*

$$\lambda_n \;=\; T_n^{-1} \;=\; C\, n^{-\frac{1}{2r+b}},$$
$$M_n \;\ge\; \tilde{d}^2 \cdot \tilde{C}\, B_{T_n}^6 \left(T_n^{2r} \vee T_n\right) \log^2 n\,,$$

where $B_{T_n} > 0$ bounds the parameter drift,

$$\|\theta_t - \theta_0\|_2 \;\le\; B_{T_n} \quad \textit{for all } t \in [T_n],$$

and $C, \tilde{C}$ are positive constants independent of n, M_n, T_n, B_{T_n}. Then, with probability of at least $1-\delta$,

$$\|G_{\theta_{T_n}} - G_\rho\|_{L^2(\rho_{\mathcal{U}})} \;\le\; \bar{C}\, n^{-\frac{r}{2r+b}} \log^3(2/\delta),$$

for some constant $\bar{C} > 0$ independent of n, M_n, T_n, B_{T_n}.

Further details on the training, the initialization of θ_0, and the above corollary are provided in Appendix A.

3.2 Discussion

Discussion of Theorem 3.4. Theorem 3.4 establishes minimax rates for a broad class of spectral filtering methods. In the well-specified case ($r = \frac{1}{2}$, $b = 1$), achieving a squared L^2-error bound of order $O(1/\sqrt{n})$ requires $t_n = 1/\lambda_n = O(\sqrt{n})$ iterations and $M_n = O(\sqrt{n} \log n)$ random features. For smoother target functions with regularity $r \ge 1$, the optimal convergence rate is attained with $t_n = O(n^{\frac{1}{2r+1}})$ iterations, but requires $M_n = O(t_n^{2r} \log n)$ random features. This highlights an interesting trade-off: higher smoothness reduces the number of necessary iterations but increases the number of random features required for optimal generalization.

In contrast, in the misspecified case $r < \frac{1}{2}$, the attainable rate of order $O(n^{-\frac{r}{2r+1}})$ is slower, reflecting the limited regularity of the target function. In this regime, the required number of random features is only $M_n = O(n^{\frac{1}{2r+1}} \log n)$, which is significantly smaller than in the well-specified or smooth cases. Overall, Theorem 3.4 shows that random feature methods with spectral regularization achieve the same minimax-optimal rates as exact kernel methods (Caponnetto and De Vito, 2007; Blanchard and Mücke, 2017), while offering improved computational scalability.

Comparison with prior work. Compared to the results of Rudi and Rosasco (2016); Lanthaler and Nelsen (2023), our work extends the analysis from KRR to general spectral filtering methods and establishes optimal convergence rates for all smoothness levels $r < \frac{1}{2}$ satisfying $2r + b > 1$ (the *easy learning regime*). With respect to the number of required random features, we recover the same order as Rudi and Rosasco (2016), namely $M_n = O(\sqrt{n} \log n)$. The analysis in Lanthaler and Nelsen (2023) is based on slightly different source assumptions, which coincide with ours in the well-specified case. Using a *random kitchen sinks* approach (Rahimi and Recht, 2008), they further showed that the logarithmic factor can be removed, proving that $M_n = O(\sqrt{n})$ random features suffice to achieve optimal rates. However, their results do not exploit prior knowledge about the effective dimension and therefore only establish optimal rates in the well-specified setting $b = 1$, $r = \frac{1}{2}$.

NNs and NOs. A connection between learning with neural networks and random feature approximation (RFA) was already observed in Yehudai and Shamir (2019). Roughly speaking, they noted that learning with neural networks is possible whenever learning with random features is possible. At the same time, they showed that neural networks cannot be used to learn even a single ReLU neuron under Gaussian inputs in $\mathbb{R}^d$ with poly(d) weights, unless the network size (or the magnitude of its weights) is exponentially large in d. For smoother activations in the NTK regime, Nguyen and Mücke (2023) improved on this result by showing that optimality can be achieved with only a polynomial number of random features in both the input dimension d and the sample complexity n.

Our approach extends these insights to the operator-valued setting relevant for Neural Operators (NOs). In contrast to the vector-input case, our rates are *dimension-free in the input space* $\mathcal{U}$. However, the sum structure of the kernel representation (2.5) introduces a linear dependence on the number of summands p, as reflected in Theorem 3.4. For NOs, the input space $\mathcal{U}$ is typically a function space, e.g., continuous mappings from $\mathcal{X}$ to $\mathbb{R}^{d_y}$. In this case, the output dimension d_y of the input functions enters the feature dimension $\tilde{d} = d_k + d_y + d_b$, which directly determines the number of required random features. This results in an overall dependence of order $\tilde{d}^2$ (see Section 2.1) in our bounds, as stated in Corollary 3.5. Thus, our results reveal a clear trade-off: generalization rates are independent of the (possibly infinite) dimension of $\mathcal{U}$, but the computational cost scales quadratically with the feature dimension $\tilde{d}$.

4 CONCLUSION

We developed a unified spectral filtering framework for RFA with vector-valued kernels, motivated by the goal of deriving generalization guarantees for NOs. Our analysis extends beyond KRR to a broad class of learning algorithms with explicit or implicit regularization, and recovers previous results as special cases. A key advantage of our approach is that both convergence rates and feature requirements are dimension-free in the (possibly infinite) input space, making the results directly applicable to NOs. At the same time, our bounds scale only quadratically with the feature dimension per neuron, providing the first minimax-optimal guarantees for NOs that combine statistical efficiency with computational tractability.

The main theoretical result, Theorem 3.4, establishes minimax rates for RFA under standard source and capacity assumptions, matching those of exact kernel methods while requiring significantly fewer resources. Our discussion highlights trade-offs between smoothness, iteration complexity, and the number of random features, as well as the contrast between well-specified, smooth, and misspecified regimes. Compared to prior work, our framework delivers optimal rates for all $r < \frac{1}{2}$ in the easy learning regime and covers operator-valued neural tangent kernels, linking neural networks, RFs, and NOs within a single theoretical setting. An interesting direction for future work is to investigate whether the quadratic dependence on the feature dimension $\tilde{d}$ can be further reduced, and to extend the analysis beyond the NTK regime to deeper architectures.

References

A. B. Aleksandrov and V. V. Peller. Operator hölder–zygmund functions, 2009. URL https://arxiv.org/abs/0907.3049. B.6

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007. 2.2, 3.1

Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18:971–1013, 2017. 2.2, 3.2, B.1, B.6

Romain Brault, Florence d'Alché Buc, and Markus Heinonen. Random fourier features for operator-valued kernels, 2016. URL https://arxiv.org/abs/1605.02536. A.1

A. Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007. 1, 2.2, 3.2, B.20

C. Carmeli, E. De Vito, and V. Umanità A. Toigo. Vector valued reproducing kernel hilbert spaces and universality, 2008. 2, 2.2, A.1

Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Reproducing kernel hilbert spaces and mercer theorem, 2005. 2, 2.2, A.1

Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with sgd and random features, 2019. 1

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996. 2.2

L. Lo Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008. ISSN 0899-7667. doi: 10.1162/neco.2008.05-07-517. URL https://doi.org/10.1162/neco.2008.05-07-517. 2.2

Daniel Zhengyu Huang, Nicholas H. Nelsen, and Margaret Trautner. An operator learning perspective on parameter-to-observable maps, 2024. 1

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018. 2.1, 2.2

Taeyoung Kim and Myungjoo Kang. Bounding the rademacher complexity of fourier neural operators, 2022. URL https://arxiv.org/abs/2209.05150. 1

Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for fourier neural operators. *Journal of Machine Learning Research*, 22(290):1–76, 2021. URL http://jmlr.org/papers/v22/21-0806.html. 1

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023a. 1

Nikola B. Kovachki, Zong-Yi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *J. Mach. Learn. Res.*, 24:89:1–89:97, 2023b. URL https://api.semanticscholar.org/CorpusID:259149906. 1

Nikola B. Kovachki, Samuel Lanthaler, and Andrew M. Stuart. Operator learning: Algorithms and analysis, 2024. 1

Samuel Lanthaler and Nicholas H. Nelsen. Error bounds for learning with vector-valued random features, 2023. 1, 1, 3.2, A.1

Jose Antonio Lara Benitez, Takashi Furuya, Florian Faucher, Anastasis Kratsios, Xavier Tricoche, and Maarten V. de Hoop. Out-of-distributional risk bounds for neural operators with applications to the helmholtz equation. *Journal of Computational Physics*, 513:113168, 2024. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2024.113168. URL https://www.sciencedirect.com/science/article/pii/S0021999124004170. 1

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019. 2.1

Mufan Bill Li, Mihai Nica, and Daniel M. Roy. The future is log-gaussian: Resnets and their infinite-depth-and-width limit at initialization, 2021a. 2.2

Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random fourier features, 2021b. 1

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations, 2020. URL https://arxiv.org/abs/2003.03485. 1

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021c. URL https://openreview.net/forum?id=c8P9NQVtmnO. 1

Junhong Lin and Volkan Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms, 2018. URL https://arxiv.org/abs/1801.07226. 1, B.10, B.6, B.6

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020a. 1

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020b. 2.2, 3.1

Carlo Marcati and Christoph Schwab. Exponential convergence of deep operator networks for elliptic partial differential equations. *SIAM Journal on Numerical Analysis*, 61(3):1513–1545, jun 2023. ISSN 1095-7170. doi: 10.1137/21m1465718. URL http://dx.doi.org/10.1137/21M1465718. 1

Ha Quang Minh. Operator-valued bochner theorem, fourier feature maps for operator-valued kernels, and vector-valued learning, 2016. URL https://arxiv.org/abs/1608.05639. A.1

Mattes Mollenhauer, Nicole Mücke, and T. J. Sullivan. Learning linear operators: Infinite-dimensional regression as a well-behaved non-compact inverse problem, 2024. URL https://arxiv.org/abs/2211.08875. A.1

Alexander Munteanu, Simon Omlor, Zhao Song, and David P. Woodruff. Bounding the width of neural networks via coupled initialization – a worst case analysis, 2022. 2.2

Mike Nguyen and Nicole Mücke. How many neurons do we need? a refined analysis for shallow networks trained with gradient descent, 2023. 2.2, 3.2

Mike Nguyen and Nicole Mücke. Optimal convergence rates for neural operators, 2024. URL https://arxiv.org/abs/2412.17518. 1, 2.1, 2.1, 3.1, 3.5, A.2, A.2, A.2, A.1, A.2

Atsushi Nitanda and Taiji Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. In *International Conference on Learning Representations*. arXiv, 2020. 2.2

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks, 2019. 2.2

Nicolò Pagliana and Lorenzo Rosasco. Implicit regularization of accelerated methods in hilbert spaces. *Advances in Neural Information Processing Systems*, 32:14481–14491, 2019. 1

Shaoxiang Qin, Fuyuan Lyu, Wenhui Peng, Dingyang Geng, Ju Wang, Xing Tang, Sylvie Leroyer, Naiping Gao, Xue Liu, and Liangzhu Leon Wang. Toward a better understanding of fourier neural operators from a spectral perspective, 2024. URL https://arxiv.org/abs/2404.07200. 1

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2007. 1, 1, 2.2

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/file/0efe32849d230d7f53049ddc4a4b0c60-Paper.pdf. 3.2

M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2018.10.045. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125. 1

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features, 2016. 1, 1, 2.2, 3.2, B.1, B.8, B.5, B.13, B.6

B. Schoelkopf and A. J. Smola. *Learning with Kernels, Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. MIT Press, 2002. 1

Christoph Schwab and Jakob Zech. Deep learning in high dimension: Neural network approximation of analytic functions in $l^2(\mathbb{R}^d, \gamma_d)$, 2021. URL https://arxiv.org/abs/2111.07080. 1

Jun Shao. *Mathematical Statistics*. Springer-Verlag New York Inc, 2nd edition, 2003. 2

Anil Sharma, Suresh Singh, and Sameer Ratna. Graph neural network operators: A review. *Multimedia Tools and Applications*, 83:23413–23436, 2024. doi: 10.1007/s11042-023-16440-4. 1

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008. A.1, B.4

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2011. B.6

Sifan Wang, Hanwen Wang, and Paris Perdikaris. Learning the solution operator of parametric partial differential equations with physics-informed deeponets. *Science Advances*, 7(40):eabi8605, 2021. doi: 10.1126/sciadv.abi8605. URL https://www.science.org/doi/abs/10.1126/sciadv.abi8605. 1

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *Advances in neural information processing systems*, 32, 2019. 3.2

Haobo Zhang, Yicheng Li, and Qian Lin. On the optimality of misspecified spectral algorithms, 2024. URL https://arxiv.org/abs/2303.14942. 1

Yaoyu Zhang, Zhi-Qin John Xu, Tao Luo, and Zheng Ma. A type of generalization error induced by initialization in deep neural networks. In *Mathematical and Scientific Machine Learning*, pages 144–164. PMLR, 2020. A.2

Xiantong Zhen, Haoliang Sun, Yingjun Du, Jun Xu, Yilong Yin, Ling Shao, and Cees Snoek. Learning to learn kernels with variational random features, 2020. 1

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Additional Material

This appendix begins by reviewing the key definitions related to vector-valued kernels. We then provide additional background on Neural Operators and present a more detailed version of Corollary 3.5, accompanied by a brief discussion of its implications. Finally, we include numerical illustrations that support our main theoretical result, Theorem 3.4.

A.1 Preliminaries on Vector-Valued Kernels

The classical theory of real-valued reproducing kernel Hilbert spaces (Steinwart and Christmann, 2008), including fundamental results such as Mercer's theorem, has been extended to the vector-valued setting in Carmeli et al. (2005, 2008). These extensions provide a rigorous mathematical foundation for analyzing operator learning problems within a kernel framework (Minh, 2016; Brault et al., 2016; Mollenhauer et al., 2024; Lanthaler and Nelsen, 2023), and have become standard tools in the operator learning literature. For completeness, we briefly recall the key definitions below.

Let $\mathcal{U}$ be a topological space and let $\mathcal{V}$ be a separable Hilbert space. A map

$$K : \mathcal{U} \times \mathcal{U} \to \mathcal{L}(\mathcal{V}),$$

where $\mathcal{L}(\mathcal{V})$ denotes the space of bounded linear operators on $\mathcal{V}$, is called a *$\mathcal{V}$-reproducing kernel* on $\mathcal{U}$ if, for any finite set of points $u_1, \ldots, u_N \in \mathcal{U}$ and vectors $v_1, \ldots, v_N \in \mathcal{V}$, it holds that

$$\sum_{i,j=1}^{N} \langle K(u_i, u_j)v_j, v_i\rangle_{\mathcal{V}} \geq 0.$$

This condition is the natural generalization of positive definiteness from the scalar- to the vector-valued setting.

For each $u \in \mathcal{U}$, we define the linear operator

$$K_u : \mathcal{V} \to \mathcal{F}(\mathcal{U}; \mathcal{V}),$$

where $\mathcal{F}(\mathcal{U}; \mathcal{V})$ denotes the space of measurable $\mathcal{V}$-valued functions on $\mathcal{U}$. Its action on $v \in \mathcal{V}$ is given by

$$(K_u v)(\tilde{u}) = K(\tilde{u}, u)v \qquad \text{for all } \tilde{u} \in \mathcal{U}.$$

Given a $\mathcal{V}$-reproducing kernel K, there exists a unique Hilbert space $\mathcal{H}_K \subset \mathcal{F}(\mathcal{U}; \mathcal{V})$ such that

$$K_u \in \mathcal{L}(\mathcal{V}, \mathcal{H}_K) \quad \text{for all } u \in \mathcal{U},$$

and for every $F \in \mathcal{H}_K$,

$$F(u) = K_u^* F \qquad \text{for all } u \in \mathcal{U},$$

where $K_u^* : \mathcal{H}_K \to \mathcal{V}$ denotes the adjoint of K_u. This is the vector-valued analogue of the classical reproducing property. In particular, it implies

$$K(u, \tilde{u}) = K_u^* K_{\tilde{u}}.$$

The space $\mathcal{H}_K$ is called the *vector-valued reproducing kernel Hilbert space (RKHS)* associated with K, and it is given by

$$\mathcal{H}_K = \overline{\operatorname{span}}\big\{K_u v \,\big|\, u \in \mathcal{U},\, v \in \mathcal{V}\big\}.$$

Finally, a reproducing kernel $K : \mathcal{U} \times \mathcal{U} \to \mathcal{L}(\mathcal{V})$ is called a *Mercer kernel* if $\mathcal{H}_K$ is a subspace of $\mathcal{C}(\mathcal{U}; \mathcal{V})$, the space of continuous $\mathcal{V}$-valued functions on $\mathcal{U}$.

A.2 Learning with Neural Operators

In this section we present a detailed version of Corollary 3.5, based on Nguyen and Mücke (2024, Theorem 3.5). For completeness, we briefly recall the neural operator learning setup considered in their work, including the training procedure via gradient descent.

Neural Operator Setting. We consider the general operator learning framework introduced in Section 2, where the input and output spaces $\mathcal{U}$ and $\mathcal{V}$ are viewed as infinite-dimensional function spaces. To enable practical training based on function-valued data, an additional discretization step is required. This involves so-called *second-stage samples* drawn from the input domain of the functions.

Specifically, let $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ denote the input domain, and let μ be an unknown probability measure on $\mathcal{X}$. The spaces $\mathcal{U}$ and $\mathcal{V}$ consist of functions mapping $\mathcal{X} \to \mathcal{Y} \subset \mathbb{R}^{d_y}$ and $\mathcal{X} \to \tilde{\mathcal{Y}} \subset \mathbb{R}$, respectively. We observe n_u i.i.d. first-stage samples

$$\big\{(u_i, v_i)\big\}_{i=1}^{n_u} \subset \mathcal{U} \times \mathcal{V},$$

each evaluated at n_x i.i.d. second-stage samples $(x_1, \ldots, x_{n_x}) \in \mathcal{X}^{n_x}$. These discretized evaluations are then used to train a shallow neural operator via gradient descent.

Gradient Descent and Initialization. Recall the class of shallow NOs introduced in Section 2.1. Following Nguyen and Mücke (2024), the parameters are trained by gradient descent on the empirical loss computed from the first-stage samples evaluated at the second-stage points:

$$\begin{aligned}\theta_{t+1}^j &= \theta_t^j - \alpha\, \partial_{\theta^j} \widehat{\mathcal{E}}\big(G_{\theta_t}\big) \\ &= \theta_t^j - \frac{\alpha}{n_u} \sum_{i=1}^{n_u} \Big\langle G_{\theta_t}(u_i) - v_i,\, \partial_{\theta^j} G_{\theta_t}(u_i) \Big\rangle_{n_x},\end{aligned} \tag{A.1}$$

where $\alpha > 0$ is the step size and

$$\langle f, g \rangle_{n_x} := \frac{1}{n_x} \sum_{k=1}^{n_x} f(x_k) g(x_k)$$

denotes the empirical inner product over the second-stage samples.

Nguyen and Mücke (2024) employ a symmetric initialization scheme for the network parameters θ_0 to ensure that $G_{\theta_0} \equiv 0$. Importantly, this symmetric trick does not affect the limiting neural tangent kernel (NTK); see Zhang et al. (2020) for details.

Specifically, the weights in the output layer are initialized symmetrically as

$$a_m^{(0)} = \tau \quad \text{for } m = 1, \ldots, M/2, \qquad a_m^{(0)} = -\tau \quad \text{for } m = M/2 + 1, \ldots, M,$$

where $\tau > 0$ is a fixed constant. The input layer parameters are initialized in a coupled manner,

$$b_m^{(0)} = b_{m+M/2}^{(0)} \quad \text{for } m \in \{1, \ldots, M/2\},$$

where the first half of the parameters $\{b_m^{(0)}\}_{m=1}^{M/2}$ are drawn independently from the initialization distribution π_0.

Now we are ready to state the original theorem from Nguyen and Mücke (2024), which provides a generalization bound for Neural Operators trained via the empirical GD algorithm.

Theorem A.1 (Nguyen and Mücke (2024), Theorem 3.5)**.** *Suppose Assumptions 3.1, 3.2, and 3.3 hold. Let $G_{\theta_{T_n}}$ denote the Neural Operator as defined in Section 2.1, where T_n is the number of GD iterations in (A.1). Let M_{n_u} denote the network width. Assume that $\alpha \in (0, \kappa^{-2})$, $n_u \geq n_0 := e^{\frac{2r+b}{2r+b-1}}$, and*

$$T_{n_u} = C n_u^{\frac{1}{2r+b}}, \qquad M_{n_u} \geq \tilde{C} B_{T_{n_u}}^6 \log^2(n_u) T_{n_u}^{2r \vee 1}, \qquad n_x \geq \tilde{C} B_{T_{n_u}}^2 T_{n_u}^{2r} \log^2 T_{n_u},$$

with

$$\|\theta_t - \theta_0\|_\Theta \leq B_{T_{n_u}} \quad \textit{for all } t \in [T_{n_u}]. \tag{A.2}$$

Then, with probability at least $1 - \delta$,

$$\|G_{\theta_{T_{n_u}}} - G_\rho\|_{L^2(\rho_{\mathcal{U}})} \;\leq\; \bar{C}\, n_u^{-\frac{r}{2r+b}} \log^3 \frac{2}{\delta}, \tag{A.3}$$

where $C, \tilde{C}, \bar{C} > 0$ are independent of $n_u, n_x, M_{n_u}, T_{n_u}, B_{T_{n_u}}$.

Discussion. Nguyen and Mücke (2024, Theorem 3.7) show that (A.2) holds with high probability for $B_T = O(\log T)$. Consequently, a two-layer Neural Operator trained by gradient descent achieves the minimax-optimal learning rate $n_u^{-\frac{r}{2r+b}}$. Notably, the required network width M matches the number of random features needed for kernel gradient descent and is independent of n_x. The lower bound on n_x arises from controlling the first term in the error decomposition (2.3), which accounts for the discretization of function-valued samples. To ensure that the empirical gradient descent in (A.1) closely tracks the population gradient flow, the empirical inner product over second-stage samples must uniformly approximate the $L^2(\mathcal{X}, \mu)$ inner product. Hoeffding-type concentration inequalities yield an $O(n_x^{-1/2})$ discretization error, which must match the minimax-optimal learning rate. This requirement leads to the stated lower bound on n_x.

A.3 Numerical Illustration

We analyze the behavior of kernel gradient descent with respect to the real-valued NTK. In our simulations, we use $n = 5000$ training and test samples drawn from two datasets: (i) a standard normal distribution with input dimension $d = 1$, and (ii) a subset of the SUSY[1] classification dataset with input dimension $d = 14$. All reported results are averaged over 50 independent runs of the algorithm.

Our theoretical analysis suggests that a number of random features of order $M = O(\sqrt{n}\, p)$, where $p = d + 2$, is sufficient to achieve optimal learning performance. Indeed, Figure 1 shows that for both datasets, once M exceeds a threshold of order $O(\sqrt{n}\, p)$ and the number of GD iterations T is fixed, further increasing M does not lead to any improvement in the test error.

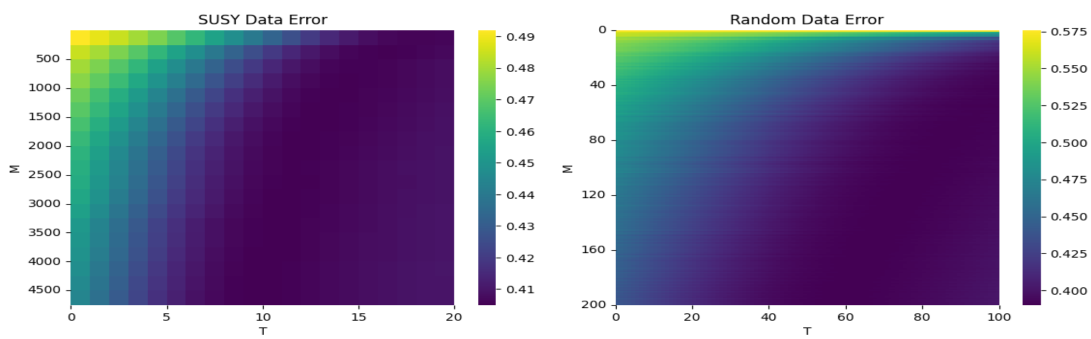

Figure 1: Heat plot of the test-error for different numbers of RF M and iterations T.
Left: Error of SUSY data set. **Right:** Error of random data set.

B Proofs

In this section, we provide the proofs of our main results.

Notation. Throughout the proofs, we use the following shorthand notation. For any bounded linear operator mapping between two Hilbert spaces A and $\lambda > 0$, we write $A_\lambda := A + \lambda I$, where I denotes the identity operator. For $G : \mathcal{U} \to \mathcal{V}$ and $(u_1, ..., u_n) \in \mathcal{U}^n$, we define the vector $\bar{G} := \big(G(u_1), \ldots, G(u_n)\big) \in \mathcal{V}^n$. Furthermore, let $C_\bullet > 0$ denote a generic constant that may change from line to line but depends only on the quantities $\kappa, r, b, c_q, c_b, E, D, Q, Z$, and not on δ, p, λ, or n.

We recall the following operator definitions. Let $\mathcal{S}_M : \mathcal{H}_M \hookrightarrow L^2(\mathcal{U}, \rho_\mathcal{U})$ denote the inclusion operator of $\mathcal{H}_M$ into $L^2(\mathcal{U}, \rho_\mathcal{U})$ for $M \in \mathbb{N}$. Its adjoint $\mathcal{S}_M^* : L^2(\mathcal{U}, \rho_\mathcal{U}) \to \mathcal{H}_M$ is given by

$$\mathcal{S}_M^* G = \int_\mathcal{U} K_{M,u}\, G(u)\, \rho_\mathcal{U}(\mathrm{d}u).$$

The covariance operator $\Sigma_M : \mathcal{H}_M \to \mathcal{H}_M$ and the kernel integral operator $\mathcal{L}_M : L^2(\mathcal{U}, \rho_\mathcal{U}) \to L^2(\mathcal{U}, \rho_\mathcal{U})$ are

[1] https://archive.ics.uci.edu/ml/datasets/SUSY

defined as

$$\Sigma_M G := \int_{\mathcal{U}} K_{M,u} K_{M,u}^* G \, \rho_{\mathcal{U}}(\mathrm{d}u),$$
$$\mathcal{L}_M G := \int_{\mathcal{U}} K_{M,u} G(u) \, \rho_{\mathcal{U}}(\mathrm{d}u).$$

The empirical counterparts of these operators, obtained by replacing $\rho_{\mathcal{U}}$ with the empirical measure, are given by

$$\widehat{\mathcal{S}}_M : \mathcal{H}_M \to \mathcal{V}^n, \qquad (\widehat{\mathcal{S}}_M G)_j = K_{M,u_j}^* G,$$
$$\widehat{\mathcal{S}}_M^* : \mathcal{V}^n \to \mathcal{H}_M, \qquad \widehat{\mathcal{S}}_M^* \mathbf{v} = \frac{1}{n} \sum_{j=1}^n K_{M,u_j} v_j,$$
$$\widehat{\Sigma}_M : \mathcal{H}_M \to \mathcal{H}_M, \qquad \widehat{\Sigma}_M = \frac{1}{n} \sum_{j=1}^n K_{M,u_j} K_{M,u_j}^*.$$

B.1 Proof Organization and Error Decomposition

In the classical kernel setting without random features, the standard analysis introduces the idealized population estimator

$$F_\lambda^* \coloneqq \mathcal{S}^* \, \phi_\lambda(\mathcal{L}) \, G_\rho,$$

which enables a decomposition of the error into bias and variance components:

$$\begin{aligned} G_\rho - \mathcal{S}\widehat{F}_\lambda &= \big(G_\rho - \mathcal{S}F_\lambda^*\big) + \big(\mathcal{S}F_\lambda^* - \mathcal{S}\widehat{F}_\lambda\big) \\ &= r_\lambda(\mathcal{L}) \, \mathcal{L}^r H + \big(\mathcal{S}F_\lambda^* - \mathcal{S}\widehat{F}_\lambda\big), \end{aligned}$$

where the bias is controlled via the residual polynomial r_λ, while the variance is handled through Hoeffding-type concentration inequalities (Blanchard and Mücke, 2017).

In the random feature setting, our analysis adapts this approach by introducing the idealized estimator

$$F_\lambda^* \coloneqq \mathcal{S}_M^* \, \phi_\lambda(\mathcal{L}_M) \, G_\rho,$$

which naturally leads to an additional *approximation error* due to the finite-dimensional random feature space. Specifically, we decompose the excess risk as

$$\begin{aligned} \|G_\rho - \mathcal{S}_M F_\lambda^M\|_{L^2(\rho_{\mathcal{U}})} &\le \|G_\rho - \mathcal{S}_M F_\lambda^*\|_{L^2(\rho_{\mathcal{U}})} + \|\mathcal{S}_M F_\lambda^* - \mathcal{S}_M F_\lambda^M\|_{L^2(\rho_{\mathcal{U}})} \\ &=: \text{Approximation Error} + \text{Estimation Error}. \end{aligned} \tag{B.1}$$

A key technical challenge lies in controlling the approximation error, which requires comparing the population operator $\mathcal{L}^r$ with its random feature counterpart $\mathcal{L}_M^r$. While prior work (Rudi and Rosasco, 2016) focuses on the case $r \in [0.5, 1]$, we develop in Section B.5 novel operator inequalities that allow us to treat arbitrary $r > 0$. For the estimation term, classical analyses rely on the specific structure of kernel ridge regression. In contrast, our approach uses a refined decomposition that exploits the polynomial structure of the residual r_λ, enabling us to obtain variance bounds *uniformly over all regularization filters*.

We bound the approximation and estimation errors separately in Sections B.2 and B.3, respectively, and then combine these bounds in Section B.4 to prove our main result, Theorem 3.4. The required operator inequalities are deferred to Section B.5, while the necessary concentration inequalities are collected in Section B.6.

B.2 Bounding the Approximation Error

Proposition B.1. *Suppose that Assumptions 3.1, 2.1, 3.2, 3.3 and $\nu \geq r \vee 1$ hold. For any $\lambda \in (0,1]$, assume that*

$$M \;\geq\; p\,C_\bullet\,\log^2\!\big(\delta^{-1}\big)\,\log\!\big(\lambda^{-1}\big)\cdot \begin{cases} \lambda^{-1}, & r \in \big(0,\tfrac{1}{2}\big),\\ \lambda^{\,b(1-2r)-1}, & r \in \big[\tfrac{1}{2},1\big],\\ \lambda^{-2r}, & r \in (1,\infty), \end{cases}$$

then the approximation term in (B.1) satisfies, with probability at least $1-\delta$,

$$\|G_\rho - \mathcal{S}_M F_\lambda^*\|_{L^2(\rho_{\mathcal{U}})} \;\leq\; C_\bullet\,\lambda^r.$$

Proof. From Assumption 3.2, we have $G_\rho = \mathcal{L}^r H$ with $\|H\|_{L^2(\rho_{\mathcal{U}})} \leq R$. Hence,

$$\begin{aligned}\|G_\rho - \mathcal{S}_M F_\lambda^*\|_{L^2(\rho_{\mathcal{U}})} &= \big\|\big(\mathcal{L}_M\phi_\lambda(\mathcal{L}_M) - I\big)\mathcal{L}^r H\big\|_{L^2(\rho_{\mathcal{U}})}\\ &\leq R\big\|r_\lambda(\mathcal{L}_M)\,\mathcal{L}^r\big\|, \end{aligned}\tag{B.2}$$

where r_λ denotes the residual polynomial defined in (2.9).

For the remaining term, we obtain

$$\begin{aligned} R\big\|r_\lambda(\mathcal{L}_M)\,\mathcal{L}^r\big\| &\leq R\big\|r_\lambda(\mathcal{L}_M)\,\mathcal{L}_{M,\lambda}^{(r\vee 1)}\big\|\big\|\mathcal{L}_{M,\lambda}^{-(r\vee 1)}\mathcal{L}_\lambda^r\big\|\\ &\leq 3R\,c_{r\vee 1}\,\lambda^r, \end{aligned}$$

where the last inequality follows from the bounds $\|r_\lambda(\mathcal{L}_M)\mathcal{L}_{M,\lambda}^{(r\vee1)}\| \leq c_{r\vee 1}\lambda^{(r\vee 1)}$ from (2.10) and $\|\mathcal{L}_{M,\lambda}^{-(r\vee1)}\mathcal{L}_\lambda^r\| \leq 3\lambda^{-(1-r)^+}$ from Proposition B.16, which holds with probability at least $1-3\delta$. □

B.3 Bounding the Estimation Error

We now turn to bounding the variance-type error in (B.1). To this end, we decompose it into two parts: (i) a classical estimation error term, which can be controlled via standard concentration inequalities, and (ii) an additional approximation-type term, whose contribution is bounded using properties of the residual polynomial.

Proposition B.2. *Suppose that Assumptions 3.1, 2.1, 3.2, and 3.3 hold, and let $\nu \geq r \vee 1$. Then, for any $s \in [0,\frac{1}{2}]$ and $\lambda \in (0,1]$, the following holds with probability at least $1-\delta$:*

$$\Big\|\Sigma_M^{\frac{1}{2}-s}(F_\lambda^M - F_\lambda^*)\Big\|_{\mathcal{H}_M} \;\leq\; C_\bullet \log\frac{1}{\delta}\,\lambda^{r-s},$$

provided that

$$M \;\geq\; p\,C_\bullet\,\log^2\!\big(\delta^{-1}\big)\,\log\!\big(\lambda^{-1}\big)\cdot \begin{cases} \lambda^{-1}, & r \in \big(0,\tfrac{1}{2}\big),\\ \lambda^{\,b(1-2r)-1}, & r \in \big[\tfrac{1}{2},1\big],\\ \lambda^{-2r}, & r \in (1,\infty), \end{cases}$$

and

$$n \;\geq\; C_\bullet\,\log^{3(2r+b)}\!\big(\delta^{-1}\big)\,\lambda^{-(2r+b)}\,, \quad n \geq n_0 := \exp\!\Big(\tfrac{2r+b}{2r+b-1}\Big).$$

Proof. We begin with the decomposition

$$\begin{aligned}\big\|\Sigma_M^{\frac{1}{2}-s}(F_\lambda^M - F_\lambda^*)\big\|_{\mathcal{H}_M} &\leq \big\|\Sigma_M^{\frac{1}{2}-s}\big(\phi_\lambda(\widehat{\Sigma}_M)\widehat{\mathcal{S}}_M^*\mathbf{v} - \phi_\lambda(\widehat{\Sigma}_M)\widehat{\Sigma}_M F_\lambda^*\big)\big\|_{\mathcal{H}_M}\\ &\quad + \big\|\Sigma_M^{\frac{1}{2}-s}\big(\phi_\lambda(\widehat{\Sigma}_M)\widehat{\Sigma}_M - I\big)F_\lambda^*\big\|_{\mathcal{H}_M}\\ &= \big\|\Sigma_M^{\frac{1}{2}-s}\phi_\lambda(\widehat{\Sigma}_M)\widehat{\mathcal{S}}_M^*\big(\mathbf{v} - \widehat{\mathcal{S}}_M F_\lambda^*\big)\big\|_{\mathcal{H}_M} + \big\|\Sigma_M^{\frac{1}{2}-s} r_\lambda(\widehat{\Sigma}_M)F_\lambda^*\big\|_{\mathcal{H}_M}\\ &=: \mathbf{(I)} + \mathbf{(II)}. \end{aligned}\tag{B.3}$$

We bound the two terms **(I)** and **(II)** separately. Specifically, by Proposition B.3, with probability at least $1-\delta$,

$$\textbf{(I)} \;\le\; C_\bullet \,\log\tfrac{1}{\delta}\,\lambda^{r-s},$$

and

$$\textbf{(II)} \;\le\; C_\bullet \,\lambda^{r-s}.$$

Combining these bounds yields the stated result. ☐

Proposition B.3. *Suppose the assumptions of Proposition B.2 hold. Then, for any $s \in [0, \frac{1}{2}]$ and $\lambda \in (0,1]$, with probability at least $1-\delta$,*

$$\textbf{\textit{(I)}} \quad \big\|\Sigma_M^{\frac{1}{2}-s}\,\phi_\lambda(\widehat{\Sigma}_M)\,\widehat{\mathcal{S}}_M^*\big(\mathbf{v}-\widehat{\mathcal{S}}_M F_\lambda^*\big)\big\|_{\mathcal{H}_M} \;\le\; C_\bullet\,\log\tfrac{1}{\delta}\,\lambda^{r-s}, \tag{B.4}$$

$$\textbf{\textit{(II)}} \quad \big\|\Sigma_M^{\frac{1}{2}-s}\,r_\lambda(\widehat{\Sigma}_M)\,F_\lambda^*\big\|_{\mathcal{H}_M} \;\le\; C_\bullet\,\lambda^{r-s}. \tag{B.5}$$

Proof. **(I)** We begin with the decomposition

$$\begin{aligned}\big\|\Sigma_M^{\frac{1}{2}-s}\,\phi_\lambda(\widehat{\Sigma}_M)\,\widehat{\mathcal{S}}_M^*\big(\mathbf{v}-\widehat{\mathcal{S}}_M F_\lambda^*\big)\big\|_{\mathcal{H}_M} &\le \big\|\Sigma_M^{\frac{1}{2}-s}\phi_\lambda(\widehat{\Sigma}_M)\Sigma_{M,\lambda}^{\frac{1}{2}}\big\|\,\big\|\Sigma_{M,\lambda}^{-\frac{1}{2}}\widehat{\mathcal{S}}_M^*\big(\mathbf{v}-\widehat{\mathcal{S}}_M F_\lambda^*\big)\big\|_{\mathcal{H}_M}\\ &=: i\cdot ii. \end{aligned} \tag{B.6}$$

Step (i). By Proposition B.15, with probability at least $1-4\delta$,

$$\|\widehat{\Sigma}_{M,\lambda}^{-\frac{1}{2}}\Sigma_{M,\lambda}^{\frac{1}{2}}\| \le 2. \tag{B.7}$$

Hence

$$\begin{aligned}\big\|\Sigma_M^{\frac{1}{2}-s}\phi_\lambda(\widehat{\Sigma}_M)\Sigma_{M,\lambda}^{\frac{1}{2}}\big\| &\le \lambda^{-s}\big\|\Sigma_{M,\lambda}^{\frac{1}{2}}\phi_\lambda(\widehat{\Sigma}_M)\Sigma_{M,\lambda}^{\frac{1}{2}}\big\|\\ &\le \lambda^{-s}\big\|\widehat{\Sigma}_{M,\lambda}\phi_\lambda(\widehat{\Sigma}_M)\big\|\,\big\|\Sigma_{M,\lambda}^{\frac{1}{2}}\widehat{\Sigma}_{M,\lambda}^{-\frac{1}{2}}\big\|^2 \;\le\; 4D\,\lambda^{-s},\end{aligned}$$

where D is defined in (2.7).

Step (ii). We decompose

$$\begin{aligned}\big\|\Sigma_{M,\lambda}^{-\frac{1}{2}}\widehat{\mathcal{S}}_M^*\big(\mathbf{v}-\widehat{\mathcal{S}}_M F_\lambda^*\big)\big\|_{\mathcal{H}_M} &\le \big\|\Sigma_{M,\lambda}^{-\frac{1}{2}}\widehat{\mathcal{S}}_M^*\big(\mathbf{v}-\bar{G}_\rho\big)\big\|_{\mathcal{H}_M} + \big\|\Sigma_{M,\lambda}^{-\frac{1}{2}}\widehat{\mathcal{S}}_M^*\big(\bar{G}_\rho-\widehat{\mathcal{S}}_M F_\lambda^*\big)\big\|_{\mathcal{H}_M}\\ &=: a+b.\end{aligned}$$

Term (a): Using Proposition B.22 together with Proposition B.18, we obtain, with probability at least $1-3\delta$,

$$\begin{aligned}\big\|\Sigma_{M,\lambda}^{-\frac{1}{2}}\widehat{\mathcal{S}}_M^*\big(\mathbf{v}-\bar{G}_\rho\big)\big\|_{\mathcal{H}_M} &\le \left(\frac{4QZ\kappa}{\sqrt{\lambda}n} + \frac{8Q\sqrt{(1+2\log\frac{2}{\delta})\mathcal{N}_{\mathcal{L}}(\lambda)}}{\sqrt{n}}\right)\log\tfrac{2}{\delta}\\ &\le C_\bullet\left(\frac{1}{\sqrt{\lambda}n} + \sqrt{\frac{\log\frac{1}{\delta}}{n\lambda^b}}\right)\log\tfrac{1}{\delta} \;\le\; C_\bullet\,\lambda^r\log\tfrac{1}{\delta},\end{aligned}$$

where the last inequality follows from the assumption on n.

Term (b): From Proposition B.15,

$$\big\|\Sigma_{M,\lambda}^{-\frac{1}{2}}\widehat{\mathcal{S}}_M^*\big\|^2 \le 2\big\|\widehat{\Sigma}_{M,\lambda}^{-\frac{1}{2}}\widehat{\mathcal{S}}_M^*\big\|^2 = 2\big\|\widehat{\mathcal{S}}_M^*\widehat{\mathcal{S}}_M(\widehat{\mathcal{S}}_M^*\widehat{\mathcal{S}}_M+\lambda)^{-1}\big\| \le 2.$$

Therefor we obtain by Proposition B.23, with probability at least $1-\delta$,

$$\begin{aligned}\big\|\Sigma_{M,\lambda}^{-\frac{1}{2}}\widehat{\mathcal{S}}_M^*(\bar{G}_\rho - \widehat{\mathcal{S}}_M F_\lambda^*)\big\|_{\mathcal{H}_M} &\leq \frac{2}{\sqrt{n}}\|\bar{G}_\rho - \widehat{\mathcal{S}}_M F_\lambda^*\|_2 \\ &\leq 2\sqrt{\left|\frac{1}{n}\|\bar{G}_\rho - \widehat{\mathcal{S}}_M F_\lambda^*\|_2^2 - \|G_\rho - \mathcal{S}_M F_\lambda^*\|_{L^2(\rho_{\mathcal{U}})}^2\right|} + \|G_\rho - \mathcal{S}_M F_\lambda^*\|_{L^2(\rho_{\mathcal{U}})} \\ &\leq C_\bullet\sqrt{\left(\frac{\lambda^{-2(\frac{1}{2}-r)^+}}{n} + \frac{\lambda^{-(\frac{1}{2}-r)^+}\|G_\rho - \mathcal{S}_M F_\lambda^*\|_{L^2(\rho_{\mathcal{U}})}}{\sqrt{n}}\right)\log\tfrac{1}{\delta}} + \|G_\rho - \mathcal{S}_M F_\lambda^*\|_{L^2(\rho_{\mathcal{U}})}.\end{aligned}$$

Applying Proposition B.1 yields, with probability at least $1-3\delta$,

$$\begin{aligned}\big\|\Sigma_{M,\lambda}^{-\frac{1}{2}}\widehat{\mathcal{S}}_M^*(\bar{G}_\rho - \widehat{\mathcal{S}}_M F_\lambda^*)\big\|_{\mathcal{H}_M} &\leq C_\bullet\sqrt{\left(\frac{\lambda^{-(1-2r)^+}}{n} + \frac{\lambda^{-(\frac{1}{2}-r)^+}\lambda^r}{\sqrt{n}} + \lambda^{2r}\right)\log\tfrac{1}{\delta}} \\ &\leq C_\bullet\,\lambda^r\sqrt{\log\tfrac{1}{\delta}},\end{aligned}$$

where the last inequality follows from the assumption on n. Therefore,

$$ii \leq C_\bullet\,\lambda^r\log\tfrac{1}{\delta}.$$

Combining (i) and (ii) in (B.6) proves (B.4). Collecting all probabilities and applying Proposition B.5 gives total probability at least $1-11\delta$. Redefining δ completes part (I).

(II) We next bound the term in (B.5). With probability at least $1-4\delta$,

$$\begin{aligned}\|\Sigma_M^{\frac{1}{2}-s}r_\lambda(\widehat{\Sigma}_M)F_\lambda^*\|_{\mathcal{H}_M} &\leq \lambda^{-s}\|\Sigma_{M,\lambda}^{\frac{1}{2}}r_\lambda(\widehat{\Sigma}_M)F_\lambda^*\|_{\mathcal{H}_M} \\ &\leq \lambda^{-s}\|\Sigma_{M,\lambda}^{\frac{1}{2}}\widehat{\Sigma}_{M,\lambda}^{-\frac{1}{2}}\|\,\|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}}r_\lambda(\widehat{\Sigma}_M)F_\lambda^*\|_{\mathcal{H}_M} \\ &\leq 2\lambda^{-s}\|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}}r_\lambda(\widehat{\Sigma}_M)F_\lambda^*\|_{\mathcal{H}_M},\end{aligned}$$

where we again used Proposition B.15. Writing out $F_\lambda^* = \mathcal{S}_M^*\phi_\lambda(\mathcal{L}_M)\mathcal{L}^r H$ gives

$$2\lambda^{-s}\|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}}r_\lambda(\widehat{\Sigma}_M)F_\lambda^*\|_{\mathcal{H}_M} \;\leq\; 2R\lambda^{-s}\|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}}r_\lambda(\widehat{\Sigma}_M)\mathcal{S}_M^*\phi_\lambda(\mathcal{L}_M)\mathcal{L}^r\|. \tag{B.8}$$

We distinguish two cases.

Case $r \leq \frac{1}{2}$. We have

$$\begin{aligned}\|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}}r_\lambda(\widehat{\Sigma}_M)\mathcal{S}_M^*\phi_\lambda(\mathcal{L}_M)\mathcal{L}^r\| &\leq \|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}}r_\lambda(\widehat{\Sigma}_M)\mathcal{S}_M^*\phi_\lambda(\mathcal{L}_M)\mathcal{L}_{M,\lambda}^r\|\|\mathcal{L}_{M,\lambda}^{-r}\mathcal{L}_\lambda^r\| \\ &= \|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}}r_\lambda(\widehat{\Sigma}_M)\Sigma_{M,\lambda}^r\mathcal{S}_M^*\phi_\lambda(\mathcal{L}_M)\|\|\mathcal{L}_{M,\lambda}^{-r}\mathcal{L}_\lambda^r\| \\ &\leq \|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}}r_\lambda(\widehat{\Sigma}_M)\Sigma_{M,\lambda}^r\|\|\mathcal{L}_M^{\frac{1}{2}}\phi_\lambda(\mathcal{L}_M)\|\|\mathcal{L}_{M,\lambda}^{-r}\mathcal{L}_\lambda^r\|.\end{aligned}$$

By Proposition B.10, $\|\mathcal{L}_M^{\frac{1}{2}}\phi_\lambda(\mathcal{L}_M)\| \leq D\lambda^{-1/2}$. Moreover, by Proposition B.7 and Proposition B.14, $\|\mathcal{L}_{M,\lambda}^{-r}\mathcal{L}_\lambda^r\| \leq 2$ with probability at least $1-4\delta$. Hence,

$$\|\Sigma_M^{\frac{1}{2}-s}r_\lambda(\widehat{\Sigma}_M)F_\lambda^*\|_{\mathcal{H}_M} \leq 4DR\lambda^{-s-\frac{1}{2}}\|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}}r_\lambda(\widehat{\Sigma}_M)\Sigma_{M,\lambda}^r\|. \tag{B.9}$$

To bound the remaining term, use Proposition B.15: $\|\widehat{\Sigma}_{M,\lambda}^{-r}\Sigma_{M,\lambda}^r\| \leq 2$. Together with (2.10), this yields

$$\|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}}r_\lambda(\widehat{\Sigma}_M)\Sigma_{M,\lambda}^r\| \leq 2c_{\frac{1}{2}+r}\lambda^{\frac{1}{2}+r}.$$

Plugging this into (B.9) gives

$$\|\Sigma_M^{\frac{1}{2}-s}r_\lambda(\widehat{\Sigma}_M)F_\lambda^*\|_{\mathcal{H}_M} \leq 8DR\,c_{\frac{1}{2}+r}\lambda^{r-s}.$$

Case $r > \frac{1}{2}$. We proceed analogously:

$$\|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}} r_\lambda(\widehat{\Sigma}_M)\mathcal{S}_M^* \phi_\lambda(\mathcal{L}_M)\mathcal{L}^r\| \leq \|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}} r_\lambda(\widehat{\Sigma}_M)\mathcal{S}_M^* \phi_\lambda(\mathcal{L}_M)\mathcal{L}_{M,\lambda}^{(r\vee 1)}\|\|\mathcal{L}_{M,\lambda}^{-(r\vee 1)}\mathcal{L}_\lambda^r\|$$
$$\leq \|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}} r_\lambda(\widehat{\Sigma}_M)\Sigma_{M,\lambda}^{(r\vee 1)-\frac{1}{2}}\|\|\mathcal{L}_M\phi_\lambda(\mathcal{L}_M)\|\|\mathcal{L}_{M,\lambda}^{-(r\vee 1)}\mathcal{L}_\lambda^r\|.$$

By the spectral method properties and Proposition B.16, $\|\mathcal{L}_M\phi_\lambda(\mathcal{L}_M)\| \leq D$ and $\|\mathcal{L}_{M,\lambda}^{-(r\vee 1)}\mathcal{L}_\lambda^r\| \leq 3\lambda^{-(1-r)^+}$. Hence,

$$\|\Sigma_M^{\frac{1}{2}-s} r_\lambda(\widehat{\Sigma}_M)F_\lambda^*\|_{\mathcal{H}_M} \leq \frac{6DR}{\lambda^{s+(1-r)^+}}\|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}} r_\lambda(\widehat{\Sigma}_M)\Sigma_{M,\lambda}^{(r\vee 1)-\frac{1}{2}}\|. \tag{B.10}$$

Using (2.10) and Proposition B.17, with probability at least $1-\delta$,

$$\|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}} r_\lambda(\widehat{\Sigma}_M)\Sigma_{M,\lambda}^{(r\vee 1)-\frac{1}{2}}\| \leq 2c_{r\vee 1}\lambda^{(r\vee 1)}. \tag{B.11}$$

Substituting this into (B.10) yields

$$\|\Sigma_M^{\frac{1}{2}-s} r_\lambda(\widehat{\Sigma}_M)F_\lambda^*\|_{\mathcal{H}_M} \leq 12DR\, c_{\frac{1}{2}+r}\lambda^{r-s}.$$

Combining both cases establishes (B.5). Collecting all concentration bounds and applying Proposition B.5 gives probability at least $1-8\delta$. Redefining δ completes the proof. □

B.4 Combining the Error Bounds

Theorem B.4. *Suppose that Assumptions 3.1, 2.1, 3.2, and 3.3 hold, and let $\nu \geq r \vee 1$. Then, for any $s \in [0, \frac{1}{2}]$ and $\lambda \in (0, 1]$, the following holds with probability at least $1-\delta$:*

$$\|G_\rho - \mathcal{S}_M F_\lambda^M\|_{L^2(\rho_\mathcal{U})} \;\leq\; C_\bullet \log\tfrac{1}{\delta}\, \lambda^r,$$

provided that

$$M \;\geq\; p\, C_\bullet \log^2\!\big(\delta^{-1}\big)\, \log\!\big(\lambda^{-1}\big)\cdot \begin{cases} \lambda^{-1}, & r \in \big(0, \frac{1}{2}\big), \\ \lambda^{b(1-2r)-1}, & r \in \big[\frac{1}{2}, 1\big], \\ \lambda^{-2r}, & r \in (1, \infty), \end{cases}$$

and

$$n \;\geq\; C_\bullet \log^{3(2r+b)}\big(\delta^{-1}\big)\,\lambda^{-(2r+b)}, \quad n \geq n_0 := \exp\!\Big(\tfrac{2r+b}{2r+b-1}\Big).$$

Proof. We begin with the following decomposition:

$$\|G_\rho - \mathcal{S}_M F_\lambda^M\|_{L^2(\rho_\mathcal{U})} \;\leq\; \|G_\rho - \mathcal{S}_M F_\lambda^*\|_{L^2(\rho_\mathcal{U})} \;+\; \|\mathcal{S}_M(F_\lambda^M - F_\lambda^*)\|_{L^2(\rho_\mathcal{U})} \;=:\; T_1 + T_2. \tag{B.12}$$

We now bound T_1 and T_2 separately.

Step 1: Bounding T_1. By Proposition B.1, with probability at least $1-\delta$,

$$\|G_\rho - \mathcal{S}_M F_\lambda^*\|_{L^2(\rho_\mathcal{U})} \;\leq\; C_\bullet\, \lambda^r. \tag{B.13}$$

Step 2: Bounding T_2. Using Mercer's theorem (see, e.g., Steinwart and Christmann (2008)) and Proposition B.2, we have, with probability at least $1-\delta$,

$$\|\mathcal{S}_M(F_\lambda^M - F_\lambda^*)\|_{L^2(\rho_\mathcal{U})} \;=\; \|\Sigma_M^{\frac{1}{2}}(F_\lambda^M - F_\lambda^*)\|_{\mathcal{H}_M} \;\leq\; C_\bullet \log\tfrac{1}{\delta}\, \lambda^r.$$

Conclusion. Combining the bounds for T_1 and T_2 in (B.12) establishes the result:

$$\|G_\rho - \mathcal{S}_M F_\lambda^M\|_{L^2(\rho_{\mathcal{U}})} \;\leq\; C_\bullet \, \log\tfrac{1}{\delta} \, \lambda^r.$$

□

We finally note that Theorem 3.4 follows directly from Theorem B.4.

B.5 Technical Inequalities

Proposition B.5. *Let E_i be events with probability at least $1-\delta_i$ and set*

$$E := \bigcap_{i=1}^{k} E_i.$$

If we can show for some event A that $\mathbb{P}(A|E) \geq 1-\delta$ then we also have

$$\begin{aligned}\mathbb{P}(A) &\geq \int_E \mathbb{P}(A|\omega)d\mathbb{P}(\omega) \geq (1-\delta)\mathbb{P}(E)\\ &= (1-\delta)\left(1-\mathbb{P}\left(\bigcup_{i=1}^{k}(\Omega/E_i)\right)\right) \geq (1-\delta)\left(1-\sum_{i=1}^{k}\delta_i\right).\end{aligned}$$

Proposition B.6 (Aleksandrov and Peller (2009), Blanchard and Mücke (2017) (Proposition B.1.))**.** *Let B_1, B_2 be two non-negative self-adjoint operators on some Hilbert space with $\|B_j\| \leq a, j=1,2$, for some non-negative a.*

(i) If $0 \leq r \leq 1$, then

$$\|B_1^r - B_2^r\| \leq C_r \|B_1 - B_2\|^r,$$

for some $C_r < \infty$.

(ii) If $r > 1$, then

$$\|B_1^r - B_2^r\| \leq C_{a,r}\|B_1 - B_2\|,$$

for some $C_{a,r} < \infty$.

Proposition B.7 (Fujii et al., 1993, Cordes inequality)**.** *Let A and B be two positive bounded linear operators on a separable Hilbert space. Then*

$$\|A^s B^s\| \leq \|AB\|^s, \qquad \text{when } 0 \leq s \leq 1.$$

Proposition B.8 (Rudi and Rosasco (2016) (Proposition 9))**.** *Let $\mathcal{H}, \mathcal{K}$ be two separable Hilbert spaces and X, A be bounded linear operators, with $A : \mathcal{H} \to \mathcal{K}$ and $B : \mathcal{H} \to \mathcal{H}$ be positive semidefinite.*

$$\|AB^\sigma\| \leq \|A\|^{1-\sigma}\|AB\|^\sigma, \quad \forall \sigma \in [0,1].$$

Proposition B.9. *Let H_1, H_2 be two separable Hilbert spaces and $\mathcal{S} : H_1 \to H_2$ a compact operator. Then for any function $f : [0, \|\mathcal{S}\|] \to [0, \infty[$,*

$$f(\mathcal{S}\mathcal{S}^*)\mathcal{S} = \mathcal{S}f(\mathcal{S}^*\mathcal{S}).$$

Proof. The result can be proved using singular value decomposition of a compact operator. □

Proposition B.10 (Lin and Cevher (2018) (Lemma 10))**.** *Let L be a compact, positive operator on a separable Hilbert space H such that $\|L\| \leq \kappa^2$. Then for any $\lambda \geq 0$,*

$$\begin{aligned}\|(L+\lambda)^\alpha \phi_\lambda(L)\| &\leq 2D\lambda^{-(1-\alpha)}, \quad \forall \alpha \in [0,1],\\ \|L^\alpha \phi_\lambda(L)\| &\leq D\lambda^{-(1-\alpha)}, \quad \forall \alpha \in [0,1],\end{aligned}$$

where D is defined in (2.7).

Proposition B.11. *With probability at least $1-\delta$, the following bounds hold:*

$$\|F_\lambda^*\|_\infty \le 2\,\kappa^{2r+1}\,R\,D\,\lambda^{-(\frac{1}{2}-r)^+},$$
$$\|F_\lambda^*\|_{\mathcal{H}_M} \le 2\,\kappa^{2r}\,R\,D\,\lambda^{-(\frac{1}{2}-r)^+},$$

provided that

$$M \ \ge\ \frac{8\,p\,\kappa^2\,\beta_\infty}{\lambda}, \quad \textit{with} \quad \beta_\infty = \log\frac{4\,\kappa^2\big(\mathcal{N}_{\mathcal{L}}(\lambda)+1\big)}{\delta\,\|\mathcal{L}\|}.$$

Proof. Since $F_\lambda^* \in \mathcal{H}_M$, we obtain from the reproducing property and the definition $G_\rho = \mathcal{L}^r H$ that, for any $x \in \mathcal{X}$,

$$\|F_\lambda^*(x)\|_{\mathcal{Y}} = \|K_{M,x}^* F_\lambda^*\|_{\mathcal{Y}} \ \le\ \kappa\,\|F_\lambda^*\|_{\mathcal{H}_M} = \kappa\,\|\mathcal{S}_M^*\phi_\lambda(\mathcal{L}_M)G_\rho\|_{\mathcal{H}_M}$$
$$= \kappa\,\|\mathcal{L}_M^{\frac{1}{2}}\phi_\lambda(\mathcal{L}_M)\mathcal{L}^r H\|_{L^2(\rho_x)}.$$

Using $\|H\|_{L^2(\rho_x)} \le R$, we find

$$\|F_\lambda^*\|_\infty \le \kappa\,R\,\|\mathcal{L}_M^{\frac{1}{2}}\phi_\lambda(\mathcal{L}_M)\mathcal{L}_{M,\lambda}^{(r\wedge\frac{1}{2})}\|\,\|\mathcal{L}_{M,\lambda}^{-(r\wedge\frac{1}{2})}\mathcal{L}^r\| = \kappa\,R\,(I)\,(II), \tag{B.14}$$
$$\|F_\lambda^*\|_{\mathcal{H}_M} \le R\,\|\mathcal{L}_M^{\frac{1}{2}}\phi_\lambda(\mathcal{L}_M)\mathcal{L}_{M,\lambda}^{(r\wedge\frac{1}{2})}\|\,\|\mathcal{L}_{M,\lambda}^{-(r\wedge\frac{1}{2})}\mathcal{L}^r\| = R\,(I)\,(II). \tag{B.15}$$

Step (I). By Proposition B.10,

$$I = \left\|\mathcal{L}_M^{\frac{1}{2}+(r\wedge\frac{1}{2})}\phi_\lambda(\mathcal{L}_M)\right\| \le \begin{cases} D, & r \ge \frac{1}{2}, \\ D\,\lambda^{r-\frac{1}{2}}, & r < \frac{1}{2}, \end{cases}$$
$$\le D\,\lambda^{-(\frac{1}{2}-r)^+}.$$

Step (II). From Propositions B.7 and B.14, with probability at least $1-\delta$, we obtain

$$II = \begin{cases} \|\mathcal{L}_{M,\lambda}^{-\frac{1}{2}}\mathcal{L}_\lambda^r\| \le \|\mathcal{L}_{M,\lambda}^{-\frac{1}{2}}\mathcal{L}_\lambda^{\frac{1}{2}}\|\|\mathcal{L}^{r-\frac{1}{2}}\| \le 2\,\kappa^{2r-1}, & r \ge \frac{1}{2}, \\ \|\mathcal{L}_{M,\lambda}^{-r}\mathcal{L}_\lambda^r\| \le \|\mathcal{L}_{M,\lambda}^{-\frac{1}{2}}\mathcal{L}_\lambda^{\frac{1}{2}}\|^{2r} \le 4^r \le 2, & r < \frac{1}{2}, \end{cases}$$
$$\le 2\,\kappa^{2r}.$$

Combining the bounds on I and II in (B.14) and (B.15) yields

$$\|F_\lambda^*\|_\infty \le 2\,\kappa^{2r+1}\,R\,D\,\lambda^{-(\frac{1}{2}-r)^+},$$
$$\|F_\lambda^*\|_{\mathcal{H}_M} \le 2\,\kappa^{2r}\,R\,D\,\lambda^{-(\frac{1}{2}-r)^+}.$$

□

Proposition B.12. *Let $\mathcal{H}$ be a separable Hilbert space and let A and B be two bounded self-adjoint positive linear operators on $\mathcal{H}$ and $\lambda > 0$. Then*

$$\left\|A_\lambda^{-\frac{1}{2}}B_\lambda^{\frac{1}{2}}\right\| \le (1-c)^{-\frac{1}{2}}, \quad \left\|A_\lambda^{\frac{1}{2}}B_\lambda^{-\frac{1}{2}}\right\| \le (1+c)^{\frac{1}{2}},$$

with

$$c = \left\|B_\lambda^{-\frac{1}{2}}(A-B)B_\lambda^{-\frac{1}{2}}\right\|.$$

Proof. The proof for the first inequality can for example be found in Rudi and Rosasco (2016) (Proposition 8). Using simple calculations the second inequality follows from

$$\left\|(A+\lambda I)^{\frac{1}{2}}(B+\lambda I)^{-\frac{1}{2}}\right\|^2 = \left\|(B+\lambda I)^{-\frac{1}{2}}(A+\lambda I)(B+\lambda I)^{-\frac{1}{2}}\right\|$$
$$\leq \left\|(B+\lambda I)^{-\frac{1}{2}}(A-B)(B+\lambda I)^{-\frac{1}{2}}\right\| + \|I\| \leq 1+c.$$

□

Proposition B.13 (Rudi and Rosasco (2016) (Lemma 9))**.** *For any $M \geq 8\kappa^4\|\mathcal{L}\|^{-1}\log^2\frac{2}{\delta}$ we have with probability at least $1-\delta$*

$$\|\mathcal{L}_M\| \geq \frac{1}{2}\|\mathcal{L}\|.$$

Proof. For $M \geq 8\kappa^4\|\mathcal{L}\|^{-1}\log^2\frac{2}{\delta}$ we have from Proposition B.21 (E_6) that with probability at least $1-\delta$, $\|\mathcal{L}-\mathcal{L}_M\|_{HS} \leq \frac{1}{2}\|\mathcal{L}\|$ and therefore

$$\|\mathcal{L}_M\| \geq \|\mathcal{L}\| - \|\mathcal{L}-\mathcal{L}_M\|_{HS} \geq \frac{1}{2}\|\mathcal{L}\|.$$

□

Proposition B.14. *Providing Assumption 2.1 we have for any $M \geq \frac{8p\kappa^2\beta_\infty}{\lambda}$ where $\beta_\infty = \log\frac{4\kappa^2(\mathcal{N}_{\mathcal{L}}(\lambda)+1)}{\delta\|\mathcal{L}\|}$ with probability at least $1-\delta$*

$$\left\|\mathcal{L}_{M,\lambda}^{-\frac{1}{2}}\mathcal{L}_{\lambda}^{\frac{1}{2}}\right\| \leq 2, \quad \left\|\mathcal{L}_{M,\lambda}^{\frac{1}{2}}\mathcal{L}_{\lambda}^{-\frac{1}{2}}\right\| \leq 2.$$

Proof. From Proposition B.21 (E_2) we have for any $\lambda > 0$,

$$\left\|\mathcal{L}_{\lambda}^{-\frac{1}{2}}(\mathcal{L}_M-\mathcal{L})\mathcal{L}_{\lambda}^{-\frac{1}{2}}\right\| \leq \frac{4\kappa^2\beta_\infty}{3M\lambda} + \sqrt{\frac{2p\kappa^2\beta_\infty}{M\lambda}}. \tag{B.16}$$

From $M \geq \frac{8p\kappa^2\beta_\infty}{\lambda}$ we therefore obtain

$$\left\|\mathcal{L}_{\lambda}^{-\frac{1}{2}}(\mathcal{L}_M-\mathcal{L})\mathcal{L}_{\lambda}^{-\frac{1}{2}}\right\| \leq \frac{3}{4}. \tag{B.17}$$

The result now follows from Proposition B.12 □

Proposition B.15. *Providing Assumption 2.1 we have for any $n \geq \frac{8\kappa^2\tilde{\beta}}{\lambda}$ with $\tilde{\beta} := \log\frac{4\kappa^2\left(\left(1+2\log\frac{2}{\delta}\right)4\mathcal{N}_{\mathcal{L}}(\lambda)+1\right)}{\delta\|\mathcal{L}\|}$ and $M \geq \frac{8p\kappa^2\beta_\infty}{\lambda} \vee 8\kappa^4\|\mathcal{L}\|^{-1}\log^2\frac{2}{\delta}$, where $\beta_\infty = \log\frac{4\kappa^2(\mathcal{N}_{\mathcal{L}}(\lambda)+1)}{\delta\|\mathcal{L}\|}$ that with probability at least $1-4\delta$*

$$\left\|\widehat{\Sigma}_{M,\lambda}^{-\frac{1}{2}}\Sigma_{M,\lambda}^{\frac{1}{2}}\right\| \leq 2, \quad \left\|\widehat{\Sigma}_{M,\lambda}^{\frac{1}{2}}\Sigma_{M,\lambda}^{-\frac{1}{2}}\right\| \leq 2.$$

Note that we use this bound for the variance term in (B.7). Furthermore, a short calculation shows that the above assumption on n is indeed satisfied under the sample size condition stated in Proposition B.2, namely,

$$n \;\geq\; C_\bullet\,\log^{3(2r+b)}\!\left(\delta^{-1}\right)\lambda^{-(2r+b)}, \qquad n \;\geq\; n_0 := \exp\!\left(\tfrac{2r+b}{2r+b-1}\right).$$

Proof. From Proposition B.21 (E_1) we have for any $\lambda > 0$ with probability at least $1-\delta$,

$$\left\|\Sigma_{M,\lambda}^{-\frac{1}{2}}\left(\widehat{\Sigma}_M-\Sigma_M\right)\Sigma_{M,\lambda}^{-\frac{1}{2}}\right\| \leq \frac{4\kappa^2\beta_M}{3n\lambda} + \sqrt{\frac{2\kappa^2\beta_M}{n\lambda}}, \tag{B.18}$$

with $\beta_M = \log \frac{4\kappa^2(\mathcal{N}_{\mathcal{L}_M}(\lambda)+1)}{\delta\|\mathcal{L}_M\|}$. For $M \geq \frac{8p\kappa^2\beta_\infty}{\lambda}$ we obtain from Proposition B.18 that with probability at least $1-2\delta$,

$$\mathcal{N}_{\mathcal{L}_M}(\lambda) \leq \left(1 + 2\log\frac{2}{\delta}\right) 4\mathcal{N}_{\mathcal{L}}(\lambda). \tag{B.19}$$

From Proposition B.13 we have with probability $1-\delta$,

$$\|\mathcal{L}_M\| \geq \frac{1}{2}\|\mathcal{L}\|. \tag{B.20}$$

Note that the bounds of (B.19) and (B.20) imply $\beta_M \leq \tilde{\beta} = \log \frac{4\kappa^2\left(\left(1+2\log\frac{2}{\delta}\right)4\mathcal{N}_{\mathcal{L}}(\lambda)+1\right)}{\delta\|\mathcal{L}\|}$. Using this together with $n \geq \frac{8\kappa^2\tilde{\beta}}{\lambda}$ we obtain for (B.18)

$$\left\|\Sigma_{M,\lambda}^{-\frac{1}{2}}\left(\widehat{\Sigma}_M - \Sigma_M\right)\Sigma_{M,\lambda}^{-\frac{1}{2}}\right\| \leq \frac{4\kappa^2\beta_M}{3n\lambda} + \sqrt{\frac{2\kappa^2\beta_M}{n\lambda}} \tag{B.21}$$

$$\leq \frac{4\kappa^2\tilde{\beta}}{3n\lambda} + \sqrt{\frac{2\kappa^2\tilde{\beta}}{n\lambda}} \leq \frac{3}{4}. \tag{B.22}$$

Note that from Proposition B.5 we have that the above inequality holds with probability at least $1-4\delta$. The result now follows from Proposition B.12 □

Proposition B.16. *Providing Assumption 2.1 we have for any*

$$M \geq \begin{cases} \frac{8p\kappa^2\beta_\infty}{\lambda} & r \in \left(0, \frac{1}{2}\right) \\ \frac{(8p\kappa^2\beta_\infty)\vee C_1^{\frac{1}{r}}}{\lambda} \vee \frac{C_2}{\lambda^{1+b(2r-1)}} & r \in \left[\frac{1}{2}, 1\right] \\ \frac{C_3}{\lambda^{2r}} & r \in (1,\infty) \end{cases}$$

with probability at least $1-3\delta$,

$$\left\|\mathcal{L}_{M,\lambda}^{-(r\vee 1)}\mathcal{L}_\lambda^r\right\| \leq \frac{3}{\lambda^{(1-r)^+}},$$

where $C_1 = 2(4\kappa\log\frac{2}{\delta})^{2r-1}(8p\kappa^2\beta_\infty)^{1-r}$, $C_2 = 4(4c_b\kappa^2\log\frac{2}{\delta})^{2r-1}(8p\kappa^2\beta_\infty)^{2-2r}$, $C_3 := 4\kappa^4 C_{\kappa,r}^2 \log^2\frac{2}{\delta}$ *and with* $C_{\kappa,r}$ *from Proposition B.6.*

Proof. For the proof we need to differ between the following three cases:

- CASE $(r \leq \frac{1}{2})$: From Proposition B.14 we have with probability at least $1-\delta$,

$$\begin{aligned}\left\|\mathcal{L}_{M,\lambda}^{-(r\vee 1)}\mathcal{L}_\lambda^r\right\| &= \left\|\mathcal{L}_{M,\lambda}^{-1}\mathcal{L}_\lambda^r\right\| \\ &\leq \lambda^{r-1}\left\|\mathcal{L}_{M,\lambda}^{-r}\mathcal{L}_\lambda^r\right\| \\ &\leq \lambda^{r-1}\left\|\mathcal{L}_{M,\lambda}^{-\frac{1}{2}}\mathcal{L}_\lambda^{\frac{1}{2}}\right\|^{2r} \leq 2^{2r}\lambda^{r-1} \leq 3\lambda^{r-1}.\end{aligned}$$

- CASE $(r \in [\frac{1}{2}, 1])$: Using $\left\|\mathcal{L}_\lambda^{-1}\mathcal{L}_\lambda^r\right\| \leq \lambda^{r-1}$ we have

$$\left\|\mathcal{L}_{M,\lambda}^{-(r\vee 1)}\mathcal{L}_\lambda^r\right\| = \left\|\mathcal{L}_{M,\lambda}^{-1}\mathcal{L}_\lambda^r\right\| \tag{B.23}$$

$$\leq \left\|\left(\mathcal{L}_{M,\lambda}^{-1} - \mathcal{L}_\lambda^{-1}\right)\mathcal{L}_\lambda^r\right\| + \lambda^{r-1}. \tag{B.24}$$

For the norm of the last inequality we have from the algebraic identity $A^{-1} - B^{-1} = A^{-1}(A - B)B^{-1}$:

$$\left\| \left(\mathcal{L}_{M,\lambda}^{-1} - \mathcal{L}_{\lambda}^{-1} \right) \mathcal{L}_{\lambda}^{r} \right\| = \left\| \mathcal{L}_{M,\lambda}^{-1} (\mathcal{L}_{M,\lambda} - \mathcal{L}_{\lambda}) \mathcal{L}_{\lambda}^{r-1} \right\|$$

and from Proposition B.14 we further have with probability at least $1 - \delta$,

$$\begin{aligned} &\left\| \mathcal{L}_{M,\lambda}^{-1} (\mathcal{L}_{M,\lambda} - \mathcal{L}_{\lambda}) \mathcal{L}_{\lambda}^{r-1} \right\| \\ &\leq \lambda^{-\frac{1}{2}} \left\| \mathcal{L}_{M,\lambda}^{-\frac{1}{2}} \mathcal{L}_{\lambda}^{\frac{1}{2}} \right\| \left\| \mathcal{L}_{\lambda}^{-\frac{1}{2}} (\mathcal{L}_{M,\lambda} - \mathcal{L}_{\lambda}) \mathcal{L}_{\lambda}^{r-1} \right\| \\ &\leq 2\lambda^{-\frac{1}{2}} \left\| \mathcal{L}_{\lambda}^{-\frac{1}{2}} (\mathcal{L}_{M,\lambda} - \mathcal{L}_{\lambda}) \mathcal{L}_{\lambda}^{r-1} \right\|. \end{aligned}$$

Since $\sigma := 2 - 2r \leq 1$ we have from Proposition B.8

$$\begin{aligned} &\left\| \mathcal{L}_{\lambda}^{-\frac{1}{2}} (\mathcal{L}_{M,\lambda} - \mathcal{L}_{\lambda}) \mathcal{L}_{\lambda}^{r-1} \right\| \\ &\leq \left\| \mathcal{L}_{\lambda}^{-\frac{1}{2}} (\mathcal{L}_{M,\lambda} - \mathcal{L}_{\lambda}) \right\|^{2r-1} \left\| \mathcal{L}_{\lambda}^{-\frac{1}{2}} (\mathcal{L}_{M,\lambda} - \mathcal{L}_{\lambda}) \mathcal{L}_{\lambda}^{-\frac{1}{2}} \right\|^{2-2r}. \end{aligned}$$

Using Proposition B.21 (E_2 and E_5) we have for the last expression with probability at least $1 - 2\delta$,

$$\leq \left[\left(\frac{2\kappa}{\sqrt{\lambda} M} + \sqrt{\frac{4\kappa^2 \mathcal{N}_{\mathcal{L}}(\lambda)}{M}} \right) \log \frac{2}{\delta} \right]^{2r-1} \left(\frac{4\kappa^2 \beta_\infty}{3M\lambda} + \sqrt{\frac{2p\kappa^2 \beta_\infty}{M\lambda}} \right)^{2-2r},$$

with $\beta_\infty = \log \frac{4\kappa^2(\mathcal{N}_{\mathcal{L}}(\lambda)+1)}{\delta \|\mathcal{L}\|}$. Using this together with $M \geq \frac{8p\kappa^2 \beta_\infty}{\lambda}$ and the simple inequality $(a+b)^{2r-1} \leq a^{2r-1} + b^{2r-1}$ we have

$$\begin{aligned} &\left\| \left(\mathcal{L}_{M,\lambda}^{-1} - \mathcal{L}_{\lambda}^{-1} \right) \mathcal{L}_{\lambda}^{r} \right\| \\ &\leq 2\lambda^{-\frac{1}{2}} \left(\frac{4\kappa \log \frac{2}{\delta}}{\sqrt{\lambda} M} + \sqrt{\frac{4\kappa^2 \mathcal{N}_{\mathcal{L}}(\lambda) \log \frac{2}{\delta}}{M}} \right)^{2r-1} \left(\frac{4\kappa^2 \beta_\infty}{3M\lambda} + \sqrt{\frac{2p\kappa^2 \beta_\infty}{M\lambda}} \right)^{2-2r} \\ &\leq 2\lambda^{-\frac{1}{2}} \left(\frac{4\kappa \log \frac{2}{\delta}}{\sqrt{\lambda} M} + \sqrt{\frac{4\kappa^2 \mathcal{N}_{\mathcal{L}}(\lambda) \log \frac{2}{\delta}}{M}} \right)^{2r-1} \left(2\sqrt{\frac{2p\kappa^2 \beta_\infty}{M\lambda}} \right)^{2-2r} \\ &\leq \frac{C_1}{\lambda M^r} + \sqrt{\frac{C_2' \mathcal{N}_{\mathcal{L}}(\lambda)^{2r-1}}{M\lambda^{3-2r}}} \leq \frac{C_1}{\lambda M^r} + \sqrt{\frac{C_2}{M\lambda^{3-2r+b(2r-1)}}}, \end{aligned}$$

where we used in the last inequality the assumption $\mathcal{N}_{\mathcal{L}}(\lambda) \leq c_b \lambda^{-b}$ and set $C_1 = 2(4\kappa \log \frac{2}{\delta})^{2r-1} (8p\kappa^2 \beta_\infty)^{1-r}$, $C_2 = 4(4c_b\kappa^2 \log^2 \frac{2}{\delta})^{2r-1} (8p\kappa^2 \beta_\infty)^{2-2r}$. From $M \geq \frac{C_1^{\frac{1}{r}}}{\lambda}$ and $M \geq \frac{C_2}{\lambda^{1+b(2r-1)}}$ we obtain

$$\left\| \left(\mathcal{L}_{M,\lambda}^{-1} - \mathcal{L}_{\lambda}^{-1} \right) \mathcal{L}_{\lambda}^{r} \right\| \leq \frac{C_1}{\lambda M^r} + \sqrt{\frac{C_2}{M\lambda^{3-2r+b(2r-1)}}} \leq 2\lambda^{r-1}.$$

Plugging this bound into (B.24) leads to

$$\left\| \mathcal{L}_{M,\lambda}^{-(r\vee 1)} \mathcal{L}_{\lambda}^{r} \right\| \leq 3\lambda^{r-1}.$$

- CASE $(r \geq 1)$:

$$\begin{aligned} \left\| \mathcal{L}_{M,\lambda}^{-(r\vee 1)} \mathcal{L}_{\lambda}^{r} \right\| &= \left\| \mathcal{L}_{M,\lambda}^{-r} \mathcal{L}_{\lambda}^{r} \right\| \\ &\leq 1 + \left\| \mathcal{L}_{M,\lambda}^{-r} (\mathcal{L}_{\lambda}^{r} - \mathcal{L}_{M,\lambda}^{r}) \right\| \\ &\leq 1 + \lambda^{-r} C_{\kappa,r} \|\mathcal{L}_{\lambda} - \mathcal{L}_{M,\lambda}\|, \end{aligned}$$

where $C_{\kappa,r}$ is defined in Proposition B.6. From the bound of Proposition B.21 (E_6) we therefore obtain

$$\left\|\mathcal{L}_{M,\lambda}^{-(r\vee 1)}\mathcal{L}_{\lambda}^{r}\right\| \le 1+\lambda^{-r}C_{1,r}\left(\frac{2\kappa^2}{M}+\frac{2\kappa^2}{\sqrt{M}}\right)\log\frac{2}{\delta}\le 3,$$

where used $M\ge C_3\lambda^{-2r}$, with $C_3:=4\kappa^4C_{1,r}^2\log^2\frac{2}{\delta}$.

□

Proposition B.17. *For any $q>0$, $n\ge\max\{8C_{\kappa,q}^2\kappa^4\lambda^{-2q}\log^2\frac{2}{\delta},\,100\kappa^2\mathcal{N}_{\mathcal{L}}(\lambda)\lambda^{-1}\log^3\frac{2}{\delta}\}$ and $M\ge\frac{8p\kappa^2\beta_\infty}{\lambda}$ we have with probability at least $1-\delta$,*

$$\left\|\widehat{\Sigma}_{M,\lambda}^{-q}\Sigma_{M,\lambda}^{q}\right\|\le 2.$$

Note that we use this bound for the variance term in (B.11), with $q=(r\vee 1)-\frac{1}{2}$. Furthermore, the above assumption on n is automatically satisfied in the case $q=(r\vee 1)-\frac{1}{2}$ (with $r\ge\frac{1}{2}$) by the sample size condition stated in Proposition B.2, namely

$$n\;\ge\;C_\bullet\,\log^{3(2r+b)}\!\left(\delta^{-1}\right)\lambda^{-(2r+b)}.$$

Proof.
- **Case $q<1$:**

From Proposition B.21(E_3) we obtain with probability at least $1-\delta$,

$$\begin{aligned}\left\|\widehat{\Sigma}_{M,\lambda}^{-q}\Sigma_{M,\lambda}^{q}\right\| &= \left\|\widehat{\Sigma}_{M,\lambda}^{-1}\Sigma_{M,\lambda}\right\|^{q}\\ &\le \left\|\widehat{\Sigma}_{M,\lambda}^{-1}\left(\widehat{\Sigma}_M-\Sigma_M\right)\right\|_{HS}+1\\ &\le \frac{1}{\sqrt{\lambda}}\left\|\widehat{\Sigma}_{M,\lambda}^{-\frac{1}{2}}\Sigma_{M,\lambda}^{\frac{1}{2}}\right\|\left\|\Sigma_{M,\lambda}^{-\frac{1}{2}}\left(\widehat{\Sigma}_M-\Sigma_M\right)\right\|_{HS}+1\\ &\le \frac{1}{\sqrt{\lambda}}\left\|\widehat{\Sigma}_{M,\lambda}^{-\frac{1}{2}}\Sigma_{M,\lambda}^{\frac{1}{2}}\right\|\left(\frac{2\kappa}{\sqrt{\lambda}n}+\sqrt{\frac{4\kappa^2\mathcal{N}_{\mathcal{L}_M}(\lambda)}{n}}\right)\log\frac{2}{\delta}+1.\end{aligned}$$

We have from Proposition B.15 with probability at least $1-\delta$,

$$\left\|\widehat{\Sigma}_{M,\lambda}^{-\frac{1}{2}}\Sigma_{M,\lambda}^{\frac{1}{2}}\right\|\le 2$$

and therefore

$$\left\|\widehat{\Sigma}_{M,\lambda}^{-q}\Sigma_{M,\lambda}^{q}\right\|\le\frac{2}{\sqrt{\lambda}}\left(\frac{2\kappa}{\sqrt{\lambda}n}+\sqrt{\frac{4\kappa^2\mathcal{N}_{\mathcal{L}_M}(\lambda)}{n}}\right)\log\frac{2}{\delta}+1. \tag{B.25}$$

From B.18 we have with probability at least $1-2\delta$,

$$\mathcal{N}_{\mathcal{L}_M}(\lambda)\le\left(1+2\log\frac{2}{\delta}\right)4\mathcal{N}_{\mathcal{L}}(\lambda).$$

Plugging this bound into (B.25) leads to

$$\left\|\widehat{\Sigma}_{M,\lambda}^{-q}\Sigma_{M,\lambda}^{q}\right\|\le\frac{2}{\sqrt{\lambda}}\left(\frac{2\kappa}{\sqrt{\lambda}n}+\sqrt{\frac{4\kappa^2\left(1+2\log\frac{2}{\delta}\right)4\mathcal{N}_{\mathcal{L}}(\lambda)}{n}}\right)\log\frac{2}{\delta}+1\le 2, \tag{B.26}$$

where we used $n\ge 100\kappa^2\mathcal{N}_{\mathcal{L}}(\lambda)\lambda^{-1}\log^3\frac{2}{\delta}$ in the last inequality.

- **Case $q \geq 1$:** From Proposition B.6 and Proposition B.21(E_7) we have with probability at least $1-\delta$,

$$\begin{aligned}\left\|\widehat{\Sigma}_{M,\lambda}^{-q}\Sigma_{M,\lambda}^{q}\right\| &\leq \lambda^{-q}\left\|\widehat{\Sigma}_{M}^{q}-\Sigma_{M}^{q}\right\|_{HS}+1\\ &\leq \lambda^{-q}C_{\kappa,q}\left\|\widehat{\Sigma}_{M}-\Sigma_{M}\right\|_{HS}+1\\ &\leq \lambda^{-q}C_{\kappa,q}\left(\frac{2\kappa^2}{n}+\frac{2\kappa^2}{\sqrt{n}}\right)\log\frac{2}{\delta}+1\leq 2,\end{aligned}$$

  where we used $n \geq 8C_{\kappa,r}^2\kappa^4\lambda^{-2q}\log^2\frac{2}{\delta}$ for the last inequality.

□

Proposition B.18. *For any $M \geq \frac{8p\kappa^2\beta_\infty}{\lambda}$ we have with probability at least $1-2\delta$,*

$$\mathcal{N}_{\mathcal{L}_M}(\lambda) \leq \left(1+2\log\frac{2}{\delta}\right)4\mathcal{N}_{\mathcal{L}}(\lambda).$$

Proof.

$$\begin{aligned}\mathcal{N}_{\mathcal{L}_M}(\lambda) \leq&\ \operatorname{Tr}[\mathcal{L}_M\mathcal{L}_\lambda^{-1}]\left\|\mathcal{L}_\lambda^{\frac{1}{2}}\mathcal{L}_{M,\lambda}^{-\frac{1}{2}}\right\|^2\\ &= \left(\mathcal{N}_{\mathcal{L}}+\ \operatorname{Tr}[(\mathcal{L}_M-\mathcal{L})\mathcal{L}_\lambda^{-1}]\right)\left\|\mathcal{L}_\lambda^{\frac{1}{2}}\mathcal{L}_{M,\lambda}^{-\frac{1}{2}}\right\|^2\\ &= (\mathcal{N}_{\mathcal{L}}+\|B\|_{HS})\left\|\mathcal{L}_\lambda^{\frac{1}{2}}\mathcal{L}_{M,\lambda}^{-\frac{1}{2}}\right\|^2,\end{aligned}$$

where $B := \mathcal{L}_\lambda^{-\frac{1}{2}}(\mathcal{L}_M-\mathcal{L})\mathcal{L}_\lambda^{-\frac{1}{2}}$. Proposition B.21($E_4$) we have with probability at least $1-\delta$,

$$\|B\|_{HS} \leq 2\left(\frac{2\kappa^2}{\lambda M}+\sqrt{\frac{\kappa^2\mathcal{N}_{\mathcal{L}}(\lambda)}{\lambda M}}\right)\log\frac{2}{\delta}.$$

Using $\lambda > 4\kappa^2 M^{-1}$ we obtain

$$\|B\|_{HS} \leq 2\mathcal{N}_{\mathcal{L}}(\lambda)\log\frac{2}{\delta}.$$

Further we have from Proposition B.14 with probability at least $1-\delta$,

$$\left\|\mathcal{L}_\lambda^{\frac{1}{2}}\mathcal{L}_{M,\lambda}^{-\frac{1}{2}}\right\|^2 \leq 4.$$

To sum up, we obtain

$$\mathcal{N}_{\mathcal{L}_M}(\lambda) \leq (\mathcal{N}_{\mathcal{L}}+\|B\|_{HS})\left\|\mathcal{L}_\lambda^{\frac{1}{2}}\mathcal{L}_{M,\lambda}^{-\frac{1}{2}}\right\|^2 \leq \left(1+2\log\frac{2}{\delta}\right)4\mathcal{N}_{\mathcal{L}}(\lambda).$$

□

B.6 Concentration Inequalities

Proposition B.19. *Let $\mathcal{X}_1,\cdots,\mathcal{X}_m$ be a sequence of independently and identically distributed selfadjoint Hilbert-Schmidt operators on a separable Hilbert space. Assume that $\mathbb{E}[\mathcal{X}_1]=0$, and $\|\mathcal{X}_1\|\leq B$ almost surely for some $B>0$. Let $\mathcal{V}$ be a positive trace-class operator such that $\mathbb{E}\left[\mathcal{X}_1^2\right]\preccurlyeq\mathcal{V}$. Then with probability at least $1-\delta$, $(\delta\in]0,1[)$, there holds*

$$\left\|\frac{1}{m}\sum_{i=1}^m\mathcal{X}_i\right\| \leq \frac{2B\beta}{3m}+\sqrt{\frac{2\|\mathcal{V}\|\beta}{m}},\quad \beta=\log\frac{4\operatorname{tr}\mathcal{V}}{\|\mathcal{V}\|\delta}.$$

Proof. The proposition was first established for matrices by Tropp (2011). For the general case including operators the proof can for example be found in Lin and Cevher (2018) (see Lemma 26). □

Proposition B.20. *The following concentration result for Hilbert space valued random variables can be found in (Caponnetto and De Vito, 2007 Caponnetto and De Vito (2007)).*

Let $w_1, \cdots, w_n$ be i.i.d random variables in a separable Hilbert space with norm $\|.\|$. Suppose that there are two positive constants B and σ^2 such that

$$\mathbb{E}\Big[\|w_1 - \mathbb{E}[w_1]\|^l\Big] \le \frac{1}{2} l! B^{l-2} V^2, \quad \forall l \ge 2. \tag{B.27}$$

Then for any $0 < \delta < 1/2$, the following holds with probability at least $1-\delta$,

$$\left\| \frac{1}{n}\sum_{k=1}^{n} w_n - \mathbb{E}[w_1] \right\| \le \left(\frac{2B}{n} + \frac{2V}{\sqrt{n}} \right) \log\frac{2}{\delta}.$$

In particular, (B.27) holds if

$$\|w_1\| \le B/2 \quad a.s., \quad \text{and} \quad \mathbb{E}\Big[\|w_1\|^2\Big] \le V^2.$$

Proposition B.21. *For any $\lambda > 0$ define the following events,*

$$E_1 = \left\{ \left\| \Sigma_{M,\lambda}^{-\frac{1}{2}} \left(\widehat{\Sigma}_M - \Sigma_M \right) \Sigma_{M,\lambda}^{-\frac{1}{2}} \right\| \le \frac{4\kappa^2 \beta_M}{3n\lambda} + \sqrt{\frac{2\kappa^2\beta_M}{n\lambda}} \right\}, \qquad \beta_M = \log\frac{4\kappa^2(\mathcal{N}_{\mathcal{L}_M}(\lambda)+1)}{\delta\|\mathcal{L}_M\|},$$

$$E_2 = \left\{ \left\| \mathcal{L}_\lambda^{-\frac{1}{2}} (\mathcal{L}_M - \mathcal{L}) \mathcal{L}_\lambda^{-\frac{1}{2}} \right\| \le \frac{4\kappa^2 \beta_\infty}{3M\lambda} + \sqrt{\frac{2p\kappa^2\beta_\infty}{M\lambda}} \right\}, \qquad \beta_\infty = \log\frac{4\kappa^2(\mathcal{N}_{\mathcal{L}}(\lambda)+1)}{\delta\|\mathcal{L}\|},$$

$$E_3 = \left\{ \left\| \Sigma_{M,\lambda}^{-\frac{1}{2}} \left(\widehat{\Sigma}_M - \Sigma_M \right) \right\|_{HS} \le \left(\frac{2\kappa}{\sqrt{\lambda}n} + \sqrt{\frac{4\kappa^2 \mathcal{N}_{\mathcal{L}_M}(\lambda)}{n}} \right) \log\frac{2}{\delta} \right\},$$

$$E_4 = \left\{ \left\| \mathcal{L}_\lambda^{-\frac{1}{2}} (\mathcal{L}_M - \mathcal{L}) \mathcal{L}_\lambda^{-\frac{1}{2}} \right\|_{HS} \le \left(\frac{4\kappa^2}{\lambda M} + \sqrt{\frac{4\kappa^2 \mathcal{N}_{\mathcal{L}}(\lambda)}{\lambda M}} \right) \log\frac{2}{\delta} \right\},$$

$$E_5 = \left\{ \left\| \mathcal{L}_\lambda^{-\frac{1}{2}} (\mathcal{L}_M - \mathcal{L}) \right\| \le \left(\frac{2\kappa}{\sqrt{\lambda}M} + \sqrt{\frac{4\kappa^2 \mathcal{N}_{\mathcal{L}}(\lambda)}{M}} \right) \log\frac{2}{\delta} \right\},$$

$$E_6 = \left\{ \|\mathcal{L} - \mathcal{L}_M\|_{HS} \le \left(\frac{2\kappa^2}{M} + \frac{2\kappa^2}{\sqrt{M}} \right) \log\frac{2}{\delta} \right\},$$

$$E_7 = \left\{ \left\| \widehat{\Sigma}_M - \Sigma_M \right\|_{HS} \le \left(\frac{2\kappa^2}{n} + \frac{2\kappa^2}{\sqrt{n}} \right) \log\frac{2}{\delta} \right\}.$$

Providing Assumption 3.1 we have for any $\delta \in (0,1)$ that each of the above events holds true with probability at least $1-\delta$.

Proof. The bound for E_1 follows exactly the same steps as in the proof of Lin and Cevher (2018) (Lemma 18). The events $E_2 - E_7$ have been bounded in Rudi and Rosasco (2016) (see Proposition 6, Lemma 8 and Proposition 10). However, due to different assumptions and a different setting we attain slightly different bounds and therefore give the proof of the events $E_2 - E_7$ for completeness.

E_2) First note that $\mathcal{L}_M$ can be expressed as

$$\mathcal{L}_M = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{p} \varphi_m^{(i)} \otimes \varphi_m^{(i)},$$

where $\varphi_m(\cdot) = \varphi(\cdot, \omega_m)$, and the tensor product is taken with respect to the $L^2(\mathcal{U}, \rho_{\mathcal{U}})$ inner product. The above identity follows by straightforward calculation:

$$\begin{aligned}(\mathcal{L}_M G)(u) &= \int K_M(u, \tilde{u})\, G(\tilde{u})\, d\rho_{\mathcal{U}}(\tilde{u}) \\ &= \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{p} \int \varphi_m^{(i)}(u) \otimes_{\mathcal{V}} \varphi_m^{(i)}(\tilde{u})\, G(\tilde{u})\, d\rho_{\mathcal{U}}(\tilde{u}) \\ &= \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{p} \varphi_m^{(i)}(u) \int \langle \varphi_m^{(i)}(\tilde{u}), G(\tilde{u}) \rangle_{\mathcal{V}} d\rho_{\mathcal{U}}(\tilde{u}) \\ &= \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{p} (\varphi_m^{(i)} \otimes \varphi_m^{(i)})(G)(u).\end{aligned}$$

Similarly, we have $\mathcal{L} = \mathbb{E}[\sum_{i=1}^{p} \varphi_i \otimes \varphi_i]$.

Now define $\mathcal{X}_m := \mathcal{L}_\lambda^{-\frac{1}{2}}(\mathcal{L}_M^{(m)} - \mathcal{L})\mathcal{L}_\lambda^{-\frac{1}{2}}$, with $\mathcal{L}_M^{(m)} := \sum_{i=1}^{p} \varphi_m^{(i)} \otimes \varphi_m^{(i)}$. We now obtain

$$\|\mathcal{X}_1\| \leq \left\| \mathcal{L}_\lambda^{-\frac{1}{2}} \mathcal{L}_M^{(m)} \mathcal{L}_\lambda^{-\frac{1}{2}} \right\| + \mathbb{E}\left\| \mathcal{L}_\lambda^{-\frac{1}{2}} \mathcal{L}_M^{(m)} \mathcal{L}_\lambda^{-\frac{1}{2}} \right\| \leq 2\frac{\kappa^2}{\lambda} := B,$$

where we used for the last inequality

$$\left\| \mathcal{L}_\lambda^{-\frac{1}{2}} \mathcal{L}_M^{(m)} \mathcal{L}_\lambda^{-\frac{1}{2}} \right\| \leq \lambda^{-1} \left\| \mathcal{L}_M^{(m)} \right\| \leq \frac{\kappa^2}{\lambda}.$$

For the second moment we have from Jensen-inequality

$$\begin{aligned}\mathbb{E}\left[\mathcal{X}^2\right] &\preccurlyeq \mathbb{E}\left[\left(\mathcal{L}_\lambda^{-\frac{1}{2}} \mathcal{L}_M^{(m)} \mathcal{L}_\lambda^{-\frac{1}{2}} \right)^2 \right] \\ &\preccurlyeq \mathbb{E}\left[p \sum_{i=1}^{p} \left(\mathcal{L}_\lambda^{-\frac{1}{2}} \varphi_m^{(i)} \otimes \varphi_m^{(i)} \mathcal{L}_\lambda^{-\frac{1}{2}} \right)^2 \right] \\ &= \mathbb{E}\left[p \sum_{i=1}^{p} \left\| \mathcal{L}_\lambda^{-\frac{1}{2}} \varphi_m^{(i)} \right\|_{L^2_{\rho_x}}^2 \mathcal{L}_\lambda^{-\frac{1}{2}} \varphi_m^{(i)} \otimes \varphi_m^{(i)} \mathcal{L}_\lambda^{-\frac{1}{2}} \right] \\ &\preccurlyeq \mathbb{E}\left[p \frac{\kappa^2}{\lambda} \mathcal{L}_\lambda^{-\frac{1}{2}} \mathcal{L}_M^{(m)} \mathcal{L}_\lambda^{-\frac{1}{2}} \right] \\ &= \frac{p\kappa^2}{\lambda} \mathcal{L}\mathcal{L}_\lambda^{-1} := \mathcal{V}\end{aligned}$$

For $\beta = \log \frac{4 \operatorname{tr} \mathcal{V}}{\|\mathcal{V}\|\delta}$ we have

$$\begin{aligned}\beta &= \log \frac{4\mathcal{N}_{\mathcal{L}}(\lambda)}{\|\mathcal{L}\mathcal{L}_\lambda^{-1}\|\delta} \\ &= \log \frac{4\mathcal{N}_{\mathcal{L}}(\lambda)(\|\mathcal{L}\| + \lambda)}{\|\mathcal{L}\|\delta} \\ &\leq \log \frac{4\mathcal{N}_{\mathcal{L}}(\lambda)\|\mathcal{L}\| + 4 \operatorname{tr} \mathcal{L}}{\|\mathcal{L}\|\delta} \leq \log \frac{4\kappa^2(\mathcal{N}_{\mathcal{L}}(\lambda) + 1)}{\|\mathcal{L}\|\delta}.\end{aligned}$$

The claim now follows from Proposition B.19.

E_3) Set $w_i := \Sigma_{M,\lambda}^{-\frac{1}{2}}\xi_i$, where $\xi_i := K_{M,u_i}K_{M,u_i}^*$. Note that $\mathbb{E}[\xi_i] = \Sigma_M$. Then, we have

$$\begin{aligned}\|w_i\|_{\mathrm{HS}} &= \big\|\Sigma_{M,\lambda}^{-\frac{1}{2}}K_{M,u_i}K_{M,u_i}^*\big\|_{\mathrm{HS}}\\ &\le \|\Sigma_{M,\lambda}^{-\frac{1}{2}}\|\,\|K_{M,u_i}K_{M,u_i}^*\|_{\mathrm{HS}}\\ &\le \lambda^{-\frac{1}{2}}\,\|K_M(u_i,u_i)\|_{\mathrm{HS}} \;\le\; \frac{\kappa^2}{\sqrt{\lambda}} \;=: B.\end{aligned}$$

For the second moment, we obtain

$$\begin{aligned}\mathbb{E}\|w_i\|_{\mathrm{HS}}^2 &= \mathbb{E}\ \mathrm{tr}\Big[\Sigma_{M,\lambda}^{-\frac{1}{2}}\xi_i\Sigma_{M,\lambda}^{-1}\xi_i\Sigma_{M,\lambda}^{-\frac{1}{2}}\Big]\\ &\le \kappa^2\ \mathbb{E}\ \mathrm{tr}\Big[\Sigma_{M,\lambda}^{-\frac{1}{2}}K_{M,u_i}K_{M,u_i}^*\Sigma_{M,\lambda}^{-\frac{1}{2}}\Big]\\ &= \kappa^2\ \mathrm{tr}\big(\Sigma_{M,\lambda}^{-1}\Sigma_M\big) = \kappa^2\,\mathcal{N}_{\mathcal{L}_M}(\lambda) \;=: V^2.\end{aligned}$$

The claim then follows directly by applying Proposition B.20.

E_4) Set $w_m := \mathcal{L}_\lambda^{-\frac{1}{2}}(\mathcal{L}_M^{(m)} - \mathcal{L})\mathcal{L}_\lambda^{-\frac{1}{2}}$. Note that we have

$$\begin{aligned}\|w_m\|_{HS} &\le \Big\|\mathcal{L}_\lambda^{-\frac{1}{2}}\mathcal{L}_M^{(m)}\mathcal{L}_\lambda^{-\frac{1}{2}}\Big\|_{HS} + \mathrm{tr}\big[\mathcal{L}\mathcal{L}_\lambda^{-1}\big]\\ &\le \left\|\mathcal{L}_\lambda^{-\frac{1}{2}}\left(\sum_{i=1}^p \varphi_m^{(i)}\otimes\varphi_m^{(i)}\right)\mathcal{L}_\lambda^{-\frac{1}{2}}\right\|_{HS} + \mathcal{N}_{\mathcal{L}}(\lambda)\\ &\le \lambda^{-1}\sum_{i=1}^p\Big\|\varphi_m^{(i)}\otimes\varphi_m^{(i)}\Big\|_{HS} + \mathcal{N}_{\mathcal{L}}(\lambda)\\ &\le \lambda^{-1}\sum_{i=1}^p\Big\|\varphi_m^{(i)}\Big\|_{L^2_{\rho_x}}^2 + \mathcal{N}_{\mathcal{L}}(\lambda) \le \frac{2\kappa^2}{\lambda} =: B.\end{aligned}$$

For the second moment we have,

$$\mathbb{E}\|w_m\|_{HS}^2 \le \mathbb{E}\,\mathrm{tr}\left[\left(\mathcal{L}_\lambda^{-\frac{1}{2}}\mathcal{L}_M^{(m)}\mathcal{L}_\lambda^{-\frac{1}{2}}\right)^2\right] \le \frac{\kappa^2}{\lambda}\mathbb{E}\,\mathrm{tr}\Big[\mathcal{L}_\lambda^{-\frac{1}{2}}\mathcal{L}_M^{(m)}\mathcal{L}_\lambda^{-\frac{1}{2}}\Big] = \frac{\kappa^2}{\lambda}\mathcal{N}_{\mathcal{L}}(\lambda) =: V^2,$$

where we used $\|\mathcal{L}_\lambda^{-\frac{1}{2}}\mathcal{L}_M^{(m)}\mathcal{L}_\lambda^{-\frac{1}{2}}\| \le \frac{\kappa^2}{\lambda}$ for the last inequality. The claim now follows from Proposition B.20.

E_5) Set $w_m := \mathcal{L}_\lambda^{-\frac{1}{2}}\mathcal{L}_M^{(m)}$. Note that we have

$$\begin{aligned}\|w_m\|_{HS} &\le \Big\|\mathcal{L}_\lambda^{-\frac{1}{2}}\mathcal{L}_M^{(m)}\Big\|_{HS}\\ &\le \left\|\mathcal{L}_\lambda^{-\frac{1}{2}}\left(\sum_{i=1}^p\varphi_m^{(i)}\otimes\varphi_m^{(i)}\right)\right\|_{HS}\\ &\le \lambda^{-1/2}\sum_{i=1}^p\Big\|\varphi_m^{(i)}\Big\|_{L^2_{\rho_x}}^2 \le \frac{\kappa^2}{\sqrt{\lambda}} =: B.\end{aligned}$$

For the second moment we have,

$$\mathbb{E}\|w_m\|_{HS}^2 \le \kappa^2\mathbb{E}\|\mathcal{L}_\lambda^{-\frac{1}{2}}\mathcal{L}_M^{(m)}\mathcal{L}_\lambda^{-\frac{1}{2}}\|_{HS} \le \kappa^2\mathbb{E}\,\mathrm{tr}\Big[\mathcal{L}_\lambda^{-\frac{1}{2}}\mathcal{L}_M^{(m)}\mathcal{L}_\lambda^{-\frac{1}{2}}\Big] = \kappa^2\mathcal{N}_{\mathcal{L}}(\lambda) =: V^2$$

The claim now follows from Proposition B.20 together with the fact that the operator norm can be bounded by the Hilbert-Schmidt norm: $\|.\| \leq \|.\|_{HS}$.

E_6) Set $w_m := \mathcal{L}_M^{(m)}$. Note that we have

$$\begin{aligned} \|w_m\|_{HS} &\leq \left\|\mathcal{L}_M^{(m)}\right\|_{HS} = \left\|\sum_{i=1}^{p} \varphi_m^{(i)} \otimes \varphi_m^{(i)}\right\|_{HS} \\ &\leq \sum_{i=1}^{p} \left\|\varphi_m^{(i)}\right\|_{L^2_{\rho_x}}^2 \leq \kappa^2 =: B. \end{aligned}$$

For the second moment we have,

$$\mathbb{E}\|w_m\|_{HS}^2 \leq \kappa^4 =: V^2.$$

The claim now follows from Proposition B.20

E_7) Set $w_i := \xi_i = K_{M,x_i} K_{M,x_i}^*$. Note that

$$\|w_i\|_{HS} = \left\|K_{M,x_i} K_{M,x_i}^*\right\|_{HS} \leq \kappa^2 =: B.$$

For the second moment we have,

$$\mathbb{E}\|w_i\|_{HS}^2 \leq \kappa^4 =: V^2.$$

The claim now follows from Proposition B.20. □

Proposition B.22. *Provided Assumption 3.1, the following event holds with probability at least* $1 - \delta$*:*

$$E_8 = \left\{ \left\|\Sigma_{M,\lambda}^{-\frac{1}{2}} \widehat{\mathcal{S}}_M^* \big(\mathbf{v} - \bar{G}_\rho\big)\right\|_{\mathcal{H}_M} \leq \left(\frac{4QZ\kappa}{\sqrt{\lambda} n} + \frac{4Q\sqrt{\mathcal{N}_{\mathcal{L}_M}(\lambda)}}{\sqrt{n}}\right) \log \frac{2}{\delta} \right\}.$$

Proof. We use Proposition B.20 to prove the statement. Define

$$w_i := \Sigma_{M,\lambda}^{-\frac{1}{2}} K_{M,u_i} \big(v_i - G_\rho(u_i)\big).$$

Note that $\mathbb{E}[w_i] = 0$ and

$$\frac{1}{n} \sum_{i=1}^{n} w_i = \Sigma_{M,\lambda}^{-\frac{1}{2}} \widehat{\mathcal{S}}_M^* \big(\mathbf{v} - \bar{G}_\rho\big).$$

Moreover, by Assumption 3.1, we have

$$\begin{aligned}
\mathbb{E}\big[\|w\|_{\mathcal{H}_M}^l\big] &= \int_{\mathcal{U}}\int_{\mathcal{V}} \|v - G_\rho(u)\|_{\mathcal{V}}^l \,\rho(dv|u)\, \|\Sigma_{M,\lambda}^{-\frac{1}{2}} K_{M,u}\|^l \rho_{\mathcal{U}}(du) \\
&\le 2^{l-1} \int_{\mathcal{U}}\int_{\mathcal{V}} \big(\|v\|_{\mathcal{V}}^l + Q^l\big)\rho(dv|u)\, \|\Sigma_{M,\lambda}^{-\frac{1}{2}} K_{M,u}\|^l \rho_{\mathcal{U}}(du) \\
&\le 2^{l-1}\big(\tfrac{1}{2} l! Z^{\,l-2} Q^2 + Q^l\big) \int_{\mathcal{U}} \|\Sigma_{M,\lambda}^{-\frac{1}{2}} K_{M,u}\|^l \rho_{\mathcal{U}}(du) \\
&\le 2^{l-1}\big(\tfrac{1}{2} l! Z^{\,l-2} Q^2 + Q^l\big) \sup_{u\in\mathcal{U}} \|\Sigma_{M,\lambda}^{-\frac{1}{2}} K_{M,u}\|^{l-2} \int_{\mathcal{U}} \mathrm{tr}\big(\Sigma_{M,\lambda}^{-1} K_{M,u} K_{M,u}^*\big)\rho_{\mathcal{U}}(du) \\
&\le 2^{l-1}\big(\tfrac{1}{2} l! Z^{\,l-2} Q^2 + Q^l\big) \Big(\tfrac{\kappa}{\sqrt{\lambda}}\Big)^{l-2} \mathrm{tr}\Big(\Sigma_{M,\lambda}^{-1} \int_{\mathcal{U}} K_{M,u} K_{M,u}^* \rho_{\mathcal{U}}(du)\Big) \\
&\le \tfrac{1}{2} l! \Big(\tfrac{2QZ\kappa}{\sqrt{\lambda}}\Big)^{l-2} \Big(2Q\sqrt{\mathcal{N}_{\mathcal{L}_M}(\lambda)}\Big)^2 \\
&= \tfrac{1}{2} l!\, B^{l-2} V^2.
\end{aligned}$$

Therefore, the statement follows directly from Proposition B.20. ☐

Proposition B.23. *Suppose that $\|G_\rho\|_\infty \le Q$ and that the bound from Proposition B.11 holds:*

$$\|F_\lambda^*\|_\infty \le C_{\kappa,R,D}\, \lambda^{-(\frac{1}{2}-r)^+}, \quad \textit{where } C_{\kappa,R,D} = 2\kappa^{2r+1} RD.$$

Then, the following event holds with probability at least $1-\delta$:

$$E_9 = \left\{ \left| \frac{1}{n}\big\|\bar{G}_\rho - \widehat{\mathcal{S}}_M F_\lambda^*\big\|_2^2 - \big\|G_\rho - \mathcal{S}_M F_\lambda^*\big\|_{L^2(\rho_{\mathcal{U}})}^2 \right| \le 2\left(\frac{B_\lambda}{n} + \frac{V_\lambda}{\sqrt{n}}\right) \log\frac{2}{\delta} \right\},$$

where

$$B_\lambda := 4\Big(Q^2 + C_{\kappa,R,D}^2\, \lambda^{-2(\frac{1}{2}-r)^+}\Big), \qquad V_\lambda := \sqrt{2}\Big(Q + C_{\kappa,R,D}\, \lambda^{-(\frac{1}{2}-r)^+}\Big)\big\|G_\rho - \mathcal{S}_M F_\lambda^*\big\|_{L^2(\rho_{\mathcal{U}})}.$$

Proof. We apply Proposition B.20. Define

$$w_i := \big\|G_\rho(u_i) - F_\lambda^*(u_i)\big\|_{\mathcal{V}}^2.$$

Then $\mathbb{E}[w_i] = \|G_\rho - \mathcal{S}_M F_\lambda^*\|_{L^2(\rho_{\mathcal{U}})}^2$, and therefore

$$\left|\frac{1}{n}\sum_{i=1}^n w_i - \mathbb{E}[w_1]\right| = \left|\frac{1}{n}\big\|\bar{G}_\rho - \widehat{\mathcal{S}}_M F_\lambda^*\big\|_2^2 - \big\|G_\rho - \mathcal{S}_M F_\lambda^*\big\|_{L^2(\rho_{\mathcal{U}})}^2\right|.$$

It remains to bound $|w_i|$ and $\mathbb{E}[w_1^2]$. Using $\|G_\rho\|_\infty := \sup_{u\in\mathcal{U}} \|G_\rho(u)\|_{\mathcal{V}} \le Q$ and Proposition B.11, we obtain

$$|w_i| \le 2\Big(Q^2 + C_{\kappa,R,D}^2\, \lambda^{-2(\frac{1}{2}-r)^+}\Big),$$

and further,

$$\begin{aligned}
\mathbb{E}\big[w_1^2\big] &\le 2\Big(Q^2 + C_{\kappa,R,D}^2\, \lambda^{-2(\frac{1}{2}-r)^+}\Big)\mathbb{E}[w_1] \\
&= 2\Big(Q^2 + C_{\kappa,R,D}^2\, \lambda^{-2(\frac{1}{2}-r)^+}\Big)\big\|G_\rho - \mathcal{S}_M F_\lambda^*\big\|_{L^2(\rho_{\mathcal{U}})}^2.
\end{aligned}$$

Hence, the claim follows directly from Proposition B.20. ☐