Journal of Machine Learning Research 27 (2026) 1-78 Submitted 3/25; Revised 4/26; Published 7/26

Optimal Convergence Rates for Neural Operators

Mike Nguyen MIKE.NGUYEN@TU-BRAUNSCHWEIG.DE
Technical University of Braunschweig

Nicole Mücke NICOLE.MUECKE@TU-BRAUNSCHWEIG.DE
Technical University of Braunschweig

Editor: Bharath Sriperumbudur

Abstract

We introduce the neural tangent kernel (NTK) regime for two-layer neural operators and analyze their generalization properties. For early-stopped gradient descent (GD), we derive fast convergence rates that are known to be minimax optimal within the framework of non-parametric regression in reproducing kernel Hilbert spaces (RKHS). We provide bounds on the number of hidden neurons and the number of second-stage samples necessary for generalization. To justify our NTK regime, we additionally show that any operator approximable by a neural operator can also be approximated by an operator from the RKHS. A key application of neural operators is learning surrogate maps for the solution operators of partial differential equations (PDEs). We consider the standard Poisson equation to illustrate our theoretical findings with simulations.

Keywords: neural operators, early stopping, gradient descent

1. Introduction

Operator learning has emerged as a transformative methodology in machine learning, focusing on learning mappings between spaces of functions. This approach is particularly powerful for surrogate modeling tasks in fields like uncertainty quantification, inverse problems, and design optimization, where the goal often involves approximating complex operators, such as the solution operator of partial differential equations (PDEs).

Learning linear operators. A classical field where learning linear operators arises is functional data analysis (FDA), particularly in functional linear regression with functional responses, also known as function-to-function regression. FDA focuses on understanding and modeling data best described as functions rather than discrete points. Unlike traditional data, where observations might be scalar or vector-valued, functional data are represented by smooth curves, surfaces, or other continuous objects. These functions are typically observed over a continuum, such as time, space, or wavelength, and are inherently infinite-dimensional (Febrero-Bande et al., 2024; Morris, 2015). The model can be seen as a generalization of multivariate regression, where the regression coefficient is now an unknown operator. A standard approach to this problem involves applying Tikhonov regularization (see Benatia et al. (2017)), with classical rates of convergence resembling a typical bias-variance trade-off.

The framework of learning operators extends naturally to nonparametric approaches, such as conditional mean embeddings (CMEs) in reproducing kernel Hilbert spaces (RKHS). In these settings, the goal is to represent conditional distributions as operators acting on functions, enabling the integration of probabilistic modeling with functional data analysis (Klebanov et al., 2020; Park and Muandet, 2020; Grünewälder et al., 2012). Conditional mean embeddings generalize the idea of representing conditional expectations as operators in RKHS, bridging the gap between statistical learning of distributions and operator learning. In Li et al. (2022) classical Tikhonov regularization has also been analyzed in this setting, and optimal convergence rates have been established.
These approaches have inspired further research, as highlighted in Mollenhauer et al. (2022), where the problem of learning linear operators is placed into the broader framework of supervised learning in random design and analyzed through the family of spectral regularization methods, encompassing the earlier examples as special cases. This line of work is advanced in Meunier et al. (2024) and Li et al. (2024), which establish upper bounds for the finite-sample risk of general vector-valued spectral algorithms. These results are applicable to both well-specified and misspecified scenarios, where the true regression function may lie outside the hypothesis space, and achieve minimax optimality across various regimes.

Learning non-linear operators. The extension of linear techniques for operator learning into non-linear contexts has been realized through innovative approaches incorporating neural networks, enabling greater flexibility and efficiency in handling complex functional relationships. In Shimizu et al. (2024), the authors introduce a method that integrates deep learning into CME by leveraging an end-to-end neural network (NN) optimization framework. The approach replaces the computationally expensive Gram matrix inversion required in traditional CME methods with a kernel-based objective, significantly improving scalability and computational efficiency. For non-linear function-on-function regression, Rao and Reimherr (2023) proposes a novel framework featuring a continuous hidden layer with continuous neurons to model functional responses. Two strategies, function-on-function direct neural networks and function-on-function basis neural networks, are developed to enhance flexibility in modeling complex functional relationships. While demonstrating strong empirical performance through simulations and real data applications, this method lacks theoretical learning guarantees. Shi et al. (2025) presents a functional deep neural network architecture with a fully data-dependent dimension reduction strategy. The architecture comprises three steps: a kernel embedding for integral transformation using a smooth, data-driven kernel; a projection step using eigenfunction bases for dimension reduction; and a deep ReLU network for prediction. This approach is accompanied by theoretical analyses, including approximation and generalization error bounds, providing robust theoretical support for the method. These developments highlight the significant advancements in non-linear operator learning, driven by neural network-based frameworks that blend theoretical rigor with practical flexibility and scalability. Another key application area for learning non-linear operators mapping between function spaces is solving partial differential equations (PDEs), where the solution operators that connect initial or boundary conditions to the PDE's solution are of particular interest (Boullé and Townsend, 2023). Understanding and approximating these operators is essential, as they encapsulate the underlying dynamics of the system being studied (Gonon et al., 2024; Tarantola, 2005).

Traditional approaches for approximating operators, especially for solving PDEs, have been well-established for decades. Techniques such as the Finite Element Method (FEM), Finite Difference Method (FDM), and Finite Volume Method (FVM) are classical methods employed to estimate these operators (Hughes, 2003). These methods discretize the problem space into manageable subdomains, enabling numerical approximations of the solution. While highly effective, they have notable limitations. A primary drawback is their computational cost, which increases exponentially with the problem's dimensionality, a phenomenon known as the curse of dimensionality. Additionally, these methods require detailed discretization of the underlying function space, which can be challenging to achieve for high-dimensional or complex geometries. Another significant challenge arises when dealing with inverse problems or noisy data, where classical methods often struggle with stability and robustness (Strang and Fix, 1973).
As a result, there has been a growing interest in novel operator estimation techniques that incorporate machine learning, and more specifically, neural networks. Neural networks offer the capability to learn complex patterns and relationships from data, making them well-suited for operator learning tasks that might be intractable via traditional methods Goodfellow et al. (2016). A particularly intriguing development is the emergence of neural network-based operator learning techniques, which can be broadly categorized into "encoder/decoder" architectures and neural operators. Both categories of methods have demonstrated remarkable success in operator learning, showcasing their ability to handle high-dimensional inputs, complex geometries, and noisy data.

Encoder/Decoder Operators. This framework typically involves three key stages: encoding the input function into a finite-dimensional parameter space, mapping these parameters using neural network architectures, and subsequently decoding the parameters back into a functional representation. The encoding phase often employs parameterization techniques where the function is projected onto a lower-dimensional basis, allowing for computationally feasible manipulation. Spectral Neural Operators represent a prominent example utilize transformations such as Fourier or wavelet transforms to convert input functions into spectral coefficients, delivering notable robustness and efficiency, particularly in scenarios featuring periodic behaviors Fanaskov and Oseledets (2024). In parallel, PCA Operators utilize Principal Component Analysis to transform input functions into scores corresponding to principal components. This transformation highlights significant variance patterns within the data, thereby facilitating effective learning and retention of crucial functional characteristics Lanthaler (2023); Kovachki et al. (2024c). Further notable examples include the DeepONet architecture, which comprises a branch network that encodes the discrete input function space and a trunk network that encodes the domain of the output functions Lu et al. (2021). Additionally, the application of Variational Autoencoders Kingma and Welling (2022) and Convolutional Neural Networks (CNNs) is prominent; the latter utilizes encoder-decoder architectures for addressing image-to-image translation problems, where the translation can be interpreted as an operator task Badrinarayanan et al. (2017).

Neural Operators. On the other hand, neural operators constitute an innovative extension of classical neural networks designed to directly learn mappings between function spaces. The architecture of neural operators resembles that of traditional neural networks

but with a crucial distinction: each layer incorporates a kernel integral operator to leverage the entire functional information. This design enables them to effectively process and map the complex relationships inherent in function spaces. One prominent example is the Fourier Neural Operator (FNO), which employs fast Fourier transforms to model the kernel integral transformations central to many operators related to partial differential equations (PDEs). This approach offers a robust framework capable of efficiently managing high-dimensional input spaces with reduced computational complexity (Li et al., 2021; Qin et al., 2024; Kovachki et al., 2023). Another example is given by Low-Rank Neural Operators, which simplify computations by approximating the kernel integral operators with low-rank representations, thus reducing the computational complexity while retaining essential information (Kovachki et al., 2023). Physics-Informed Neural Operators (PINOs) integrate physical laws and constraints directly into the learning process via modified loss functions, which helps in ensuring that the learned mappings adhere to known physical principles. This hybrid approach combines data-driven methodologies with theoretical insights, enhancing the model's generalization capabilities across diverse scenarios from fluid dynamics to material sciences and many others (Wang et al., 2021; Raissi et al., 2019). Graph Neural Operators extend the neural operator framework to graph-structured data, efficiently capturing spatial and topological information intrinsic to many real-world problems. By adapting neural operators to graph settings, these models skillfully navigate complex data relationships, making them suitable for applications in network analysis and beyond, see e.g. (Li et al., 2020; Kovachki et al., 2023; Sharma et al., 2024).

Contribution. We aim to learn potential non-linear operators, mapping between normed function spaces. We cast this problem into the framework of supervised learning and focus on neural operators that encompass all the aforementioned examples. We assume our first-stage data are functions from some possibly infinite dimensional normed space, drawn independently and identically from some unknown probability distribution. Since we deal with infinite dimensional objects, further discretization is necessary for any practical implementation. We suppose that some evaluations of the data functions are available, which we call the second-stage samples, see Section 2. Our primary objective is to establish fast convergence rates for the mean squared error. Following the paradigm in Kovachki et al. (2024b), we investigate the data complexity of learning non-linear operators within the framework of the vector-valued neural tangent kernel (vvNTK). In doing so, we meticulously account for the precise number of neurons and the second-stage samples required to achieve these rates. To attain fast convergence rates, however, it is necessary to restrict the class of operators to be estimated. More precisely, we impose an a priori smoothness assumption defined in terms of powers of the integral operator associated with the vvNTK, commonly referred to as a Hölder source condition in classical RKHS methods. Specifically, we limit our scope to learning operators that reside in the range of powers of the integral operator, with the well-specified case included. To justify this assumption, we demonstrate that any operator approximable by a neural operator can also be approximated by an operator from the vvRKHS. Additionally, we derive convergence rates that account for the assumed eigenvalue decay of the associated vvNTK integral operator. Our bounds match those from traditional nonparametric methods in RKHS, which are known to be minimax optimal; see also Nguyen and Mücke (2023).

Organization. The rest of the paper is organized as follows. In Section 2, we present our setting and review relevant results on learning with kernels, and learning with random features. In Section 3, we present and discuss our main results. Finally, numerical experiments are presented in Section 4. A conclusion in Section 5 summarizes our results and highlights open research directions. All proofs are deferred to the Appendices.

2. Setup

In this section we provide the mathematical framework for our analysis. We let $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ be the input space of our function spaces. The unknown data probability measure on the input data space $\mathcal{X}$ is denoted by ρ_x. Further we let $\mathcal{U}$ be the function input space, mapping from $\mathcal{X} \to \mathcal{Y} \subset \mathbb{R}^{d_y}$ and $\mathcal{V}$ be the function target space, mapping from $\mathcal{X} \to \tilde{\mathcal{Y}} \subset \mathbb{R}$, to be specified in Assumption 4. The unknown data distribution on the data space $\mathcal{W} = \mathcal{U} \times \mathcal{V}$ is denoted by μ, while the marginal distribution on $\mathcal{U}$ is denoted as μ_u and the regular conditional distribution on $\mathcal{V}$ given $u \in \mathcal{U}$ is denoted by $\mu(\cdot|u)$, see e.g. Shao (2003). Given an operator $G : \mathcal{U} \to \mathcal{V}$ we further define the expected risk as

$$\mathcal{E}(G) := \mathbb{E}_\mu[\|G(u) - v\|^2_{L^2(\rho_x)}] \, . \tag{1}$$

It is known that the global minimizer of $\mathcal{E}$ over the set of all measurable operators is given by the regression operator G^*, defined as

$$G^*(u) := \int_{\mathcal{V}} v(.)\mu(dv|u). \tag{2}$$

2.1 Two-Layer Neural Operator

Hypotheses class. The hypothesis class considered in this paper is given by the following set of two-layer neural operators $\mathcal{F}_M$: Let $M \in \mathbb{N}$ be the network width. Given an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ acting componentwise[1], some continuous operator $A : \mathcal{U} \to L^2(\mathcal{X}, \rho_x; \mathbb{R}^{d_k})$, and a bias function $c : \mathcal{X} \to \mathbb{R}^{d_b}$, we consider the class $\mathcal{F}_M = \mathcal{F}_M(\sigma, A, c)$, defined by

$$\begin{aligned} \mathcal{F}_M := \Big\{ G_\theta : \mathcal{U} \to \mathcal{V} \mid \; (G_\theta u)(x) = \frac{1}{\sqrt{M}} \langle a, \sigma \left(B_1(Au)(x) + B_2 u(x) + B_3 c(x) \right) \rangle_2 \; , \\ \theta = (a, B_1, B_2, B_3) \in \mathbb{R}^M \times \mathbb{R}^{M \times d_k} \times \mathbb{R}^{M \times d_y} \times \mathbb{R}^{M \times d_b} \Big\} \; , \end{aligned} \tag{3}$$

where we denote with $\langle \, . \, , \, . \, \rangle_2$ the euclidean inner product. Setting $\tilde{d} = d_k + d_y + d_b$ and $\Theta = \mathbb{R}^M \times \mathbb{R}^{M \times \tilde{d}}$, we condense all parameters in

$$\begin{aligned} \theta = (a, B_1, B_2, B_3) = (a, B) \in \Theta \\ B = (B_1, B_2, B_3) \in \mathbb{R}^{M \times \tilde{d}}, \end{aligned}$$

1. If $v \in \mathbb{R}^k$, then $\sigma(v) = (\sigma(v_1), ..., \sigma(v_k))^T$.

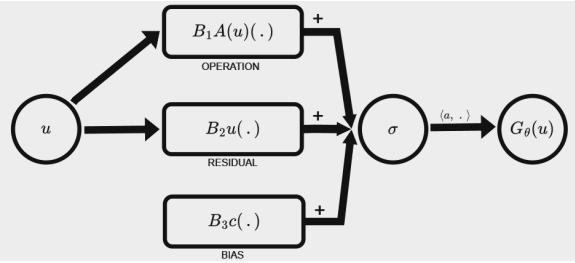

Figure 1: Depiction of the architecture of our operator class.

and equip the parameter space Θ with the norm $\|\cdot\|_\Theta$, defined by

$$\|\theta\|_\Theta^2 = \|a\|_2^2 + \|B\|_F^2, \tag{4}$$

for any $\theta \in \Theta$.

In our operator class (3), we first apply an operator transformation $B_1A(u)$, where B_1 defines a matrix and typically, A defines a kernel integral operator. Depending on the choice of A, our operator class covers a wide range of practically well-studied neural operators, such as Fourier Neural Operators, Graph Neural Operators, and Low-rank Neural Operators (Kovachki et al., 2023; Huang et al., 2024; Kovachki et al., 2024c). The kernel integral operation $A(u)$, can be seen as an intuitive extension of matrix multiplication in traditional neural networks to an infinite-dimensional setting. Inspired by Residual Neural Networks, the second transformation, $B_2u(.)$ with $B_2 \in \mathbb{R}^{M\times d_y}$ is just a matrix multiplication and keeps information of the original input $u(.)$. The final component, $B_3c(.)$ with $c : \mathcal{X} \to \mathbb{R}^{d_b}$ defines the bias. Unlike standard neural networks, the bias here is typically not a constant but a function. This function can be a linear transformation or a shallow neural network with trainable parameters. However, to keep our proofs succinct, we assume that c is given and instead multiply the bias with a trainable matrix $B_3 \in \mathbb{R}^{M\times d_b}$, see Figure 1. Examples are provided in Example 2.

In practice, one typically wraps the neural operator between a lifting operator L and a projection operator P, forming $P \circ G_\theta \circ L$. Here L and P act pointwise, meaning $L(u)(x) = L(u(x))$, and they are typically defined via linear transformations or shallow neural networks (Kovachki et al., 2023). However, to keep the proofs concise, we assume in our work that L and P are the identity operators.

Note that our operator class requires the function spaces $\mathcal{U}$ and $\mathcal{V}$ to have the same input space $\mathcal{X} \subset \mathbb{R}^{d_x}$. However, if their input spaces differ, they are typically uplifted to a common higher-dimensional space (Kovachki et al., 2023). For example, let $\mathcal{U}$ contain functions mapping from $\tilde{\mathcal{X}}$ and $\mathcal{V}$ contain functions mapping from $\mathcal{X}$, for $\dim(\tilde{\mathcal{X}}) \neq \dim(\mathcal{X})$. In this case, one could consider the operator class

$$\mathcal{F}_M \quad := \left\{ G_\theta : \mathcal{U} \to \mathcal{V} \mid \ (G_\theta u)(x) = \tfrac{1}{\sqrt{M}} \left\langle a, \sigma \left(B_1A(u)(x) + B_2u(\xi x) + B_3c(x) \right) \right\rangle \right\},$$

for a lifting or projection map $\xi : \mathcal{X} \to \tilde{\mathcal{X}}$. This generalization can be easily adapted to our analytical results.

Activation Function. For the activation function σ, we impose the following assumptions.

Assumption 1 (Activation Function) *1. **Smoothness:** The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is at least two times continuously differentiable with $\|\sigma'\|_\infty \leq C_{\sigma'}$, $\|\sigma''\|_\infty \leq C_{\sigma''}$, for some $C_{\sigma'}, C_{\sigma''} > 0$. In addition, we assume that the second derivative σ'' is $L_{\sigma''}-$ Lipschitz continuous.*

*2. **Linear growth:** For any $u \in \mathbb{R}$ we assume that $|\sigma(u)| \leq 1 + |u|$.*

The first condition requires the activation function to be sufficiently smooth: it must be twice continuously differentiable, with both the first and second derivatives uniformly bounded. In addition, we assume that the second derivative is Lipschitz continuous. Together, these assumptions ensure that the network output depends smoothly and stably on both its inputs and parameters.

The second condition restricts the growth of the activation function. Requiring at most linear growth prevents the network outputs from exploding for large inputs and guarantees that all quantities appearing in our analysis remain integrable. This condition is mild and is satisfied by many commonly used smooth activations or their standard smooth approximations, see Example 1.

The regularity assumptions on the activation function σ are satisfied by many standard smooth activations, including tanh, sigmoid, softplus, and smoothed ReLU variants. The standard $\mathrm{ReLU}(u) = \max\{u, 0\}$ and $\mathrm{ReLU}^k(u) = \max\{u, 0\}^k$ activations do not satisfy Assumption 1 due to their lack of second-order smoothness. However, they can be approximated arbitrarily well by smooth activations within our admissible class. Such approximations are commonly used in theoretical analyses of neural networks (Pinkus, 1999).

Example 1 (Activations) *1. **Smooth bounded activation.** An admissible activation is the hyperbolic tangent $\sigma(u) = \tanh(u)$. It is real analytic on $\mathbb{R}$ and thus $C^\infty(\mathbb{R})$, with bounded derivatives of all orders. Moreover, it satisfies $|\sigma(u)| \leq 1 \leq 1 + |u|$ for all $u \in \mathbb{R}$.*

2. The **softplus activation** *$\sigma_\beta(u) = \beta^{-1} \log(1 + e^{\beta u})$ has bounded derivative and the second derivative is $\sqrt{3}/18 \cdot \beta^2$-Lipschitz continuous for any $\beta > 0$.*

We show that the linear growth condition is also satisfied: Since $\sigma_\beta(u) \geq 0$, it suffices to upper bound $\sigma_\beta(u)$. For $u \geq 0$, we use $1 + e^{\beta u} \leq 2e^{\beta u}$ to obtain

$$\sigma_\beta(u) = \frac{1}{\beta} \log\big(1 + e^{\beta u}\big) \leq \frac{1}{\beta} \log\big(2e^{\beta u}\big) = u + \frac{\log 2}{\beta}.$$

For $u \leq 0$, we have $e^{\beta u} \leq 1$, hence

$$\sigma_\beta(u) \leq \frac{1}{\beta} \log(1 + 1) = \frac{\log 2}{\beta}.$$

If $\beta \geq \log 2$, *then* $\frac{\log 2}{\beta} \leq 1$, *and therefore*

$$\sigma_\beta(u) \leq \begin{cases} u+1, & u \geq 0, \\ 1, & u \leq 0, \end{cases}$$

so that $|\sigma_\beta(u)| \leq 1 + |u|$ *for all* $u \in \mathbb{R}$. *Finally, softplus converges uniformly on* $\mathbb{R}$ *to* ReLU *as* $\beta \to \infty$ *with rate* $\mathcal{O}(1/\beta)$. *One can also show that*

$$\|\sigma_\beta - ReLU\|_{L^p(\mathbb{R})} \to 0,$$

as $\beta \to \infty$, *for any* $1 \leq p \leq \infty$.

3. *Another example is a* **quadratic (Huber-type) smoothed** ReLU *defined by*

$$\sigma_\varepsilon(t) = \begin{cases} 0, & t \leq -\varepsilon, \\ \frac{(t+\varepsilon)^2}{4\varepsilon}, & |t| \leq \varepsilon, \\ t, & t \geq \varepsilon, \end{cases}$$

which belongs to $C^2(\mathbb{R})$, *has bounded first derivative and Lipschitz continuous second derivative. The Huber-type smoothed* ReLU *converges pointwise and uniformly on compact sets to the* ReLU *activation as the smoothing parameter tends to zero.*

4. ***Mollified ReLU.*** *Let* $\rho \in C_c^\infty(\mathbb{R})$ *be a standard mollifier, i.e.,* $\rho \geq 0$, $\int_{\mathbb{R}} \rho(t)\,dt = 1$, *and* $\operatorname{supp}(\rho) \subset [-1, 1]$. *For* $\varepsilon > 0$, *define the scaled mollifier*

$$\rho_\varepsilon(t) := \frac{1}{\varepsilon}\rho\left(\frac{t}{\varepsilon}\right),$$

and the mollified ReLU activation by

$$\sigma_\varepsilon(u) \;:=\; (\mathrm{ReLU} * \rho_\varepsilon)(u) \;=\; \int_{\mathbb{R}} \max(u-s, 0)\,\rho_\varepsilon(s)\,ds.$$

Then $\sigma_\varepsilon \in C^\infty(\mathbb{R})$, *has bounded first derivative and Lipschitz continuous second derivative, satisfies the linear growth bound* $|\sigma_\varepsilon(u)| \leq 1 + |u|$, *and converges uniformly on compact sets to the ReLU activation as* $\varepsilon \to 0$.

Example 2 (Operator classes) *A common choice for the transformation* $B_1 \cdot Au$ *in our neural operator architecture is a kernel integral operator of the form*

$$B_1 \cdot (Au)(x) = \int_D k(\tilde{x}, x) u(\tilde{x})\,d\tilde{x}, \tag{5}$$

where $k : D \times D \to \mathbb{R}^{d_M \times d_y}$ *and* $D \subseteq \mathcal{X}$. *Different variants of neural operators are distinguished by how the integral operator* (5) *is approximated.*

In the literature, various parameterizations of $k(y, \cdot)$ *have been proposed, ranging from neural networks to low-rank tensor approximations and spectral representations Kovachki et al. (2024a); Li et al. (2020). In the following, we focus on two representative instances introduced in Kovachki et al. (2024a).*

1. **Graph Neural Operators (GNO).** *In the GNO framework, the integral in* (5) *is approximated by a Monte Carlo-type quadrature using a finite set of sampling points* $B_x \subset D$*. Specifically, for a given bounded kernel* $k : D \times D \to \mathbb{R}^{d_k \times d_y}$*, the operator is approximated as*

$$(Au)(x) = \frac{1}{|B_x|} \sum_{x_i \in B_x} k(x_i, x) u(x_i),$$

where B_x *is typically selected either randomly from the second-stage sampling points or based on a local neighborhood of* x*—for instance, a ball* $B(x, r)$ *or a fixed number of nearest neighbors. This discretization reduces the continuous integral to a sum over local graph edges and can be naturally implemented using message-passing graph neural networks (GNNs); see (Kovachki et al., 2024a, Section 4.1) and Li et al. (2020).*

2. **Fourier Neural Operators (FNO).** *Arguably the most widely used class of neural operators is Fourier Neural Operators (FNOs), due to their computational efficiency and the ease with which the action of the kernel* k *can be learned in frequency space. This is made possible under the assumptions that the domain* D *is a periodic rectangular box, typically* $D = [0, 1]^{d_x}$*, and that the kernel depends only on the relative position, i.e.,* $k(x, y) = \kappa(x - y)$*.*

   Under these assumptions, the operator (5) *becomes a convolution and can be expressed via the Fourier transform:*

$$\int_D k(\tilde{x}, x) u(\tilde{x})\, d\tilde{x} = \mathcal{F}^{-1} \left(\mathcal{F}(k) \cdot \mathcal{F}(u) \right)(x), \tag{6}$$

   where $\mathcal{F} : L^2(D; \mathbb{C}^d) \to \ell(\mathbb{Z}^{d_x}; \mathbb{C}^d)$ *and* $\mathcal{F}^{-1}$ *denote the Fourier transform and its inverse. The kernel is thus fully represented in frequency space, allowing its action to be encoded via a collection of learned spectral multipliers.*

   Concretely, the Fourier transform of u *is approximated as*

$$(\hat{\mathcal{F}}(u))_l = \sum_{\tilde{x} \in \tilde{D}} u(\tilde{x})\, e^{-2\pi i\, l \cdot \tilde{x}} \in \mathbb{C}^{d_y},$$

   and the integral in (6) *is then approximated by*

$$\int_D k(\tilde{x}, x) u(\tilde{x})\, d\tilde{x} \approx \sum_{l \in Z} R_l \cdot \left(\sum_{\tilde{x} \in \tilde{D}} u(\tilde{x})\, e^{2\pi i\, l \cdot (x - \tilde{x})} \right), \tag{7}$$

   where:

   - $Z \subset \mathbb{Z}^{d_x}$ *is a fixed set of discrete spatial frequencies (Fourier modes),*
   - $\tilde{D}$ *is a uniform discretization of* D*,*
   - $R_l = (\mathcal{F}(k))_l \in \mathbb{C}^{M \times d_y}$ *are trainable spectral weights for each frequency* l*, capturing the action of the kernel* k *in Fourier space.*

We refer to (Kovachki et al., 2024a, Section 4.4) for a detailed discussion on the selection of Z and $\tilde{D}$ and for numerical illustrations. Approximation theory and further insights into FNOs can be found in Kovachki et al. (2021).

The spectral approximation in (7) *is directly captured by our model by setting*

$$B_1 = R := (R_1, \dots, R_{d_k}) \in \mathbb{C}^{M \times d_k}, \quad d_k := d_y \cdot |Z|,$$

and defining the lifted representation

$$((Au)(x)) := (V_{1,1}(x), \dots, V_{d_y,1}(x), V_{1,2}(x), \dots, V_{d_y,|Z|}(x)) \in \mathbb{C}^{d_k},$$

where

$$V_{k,l}(x) := \sum_{\tilde{x} \in \tilde{D}} u_k(\tilde{x})\, e^{2\pi i\, l \cdot (x - \tilde{x})}.$$

3. **CNN-/local-convolution-type example.** *Let $X \subset \mathbb{R}^{d_x}$ and let $\psi_1, \dots, \psi_{d_k} \in L^1(\mathbb{R}^{d_x})$ be compactly supported filters. Define*

$$(Au)(x) := \big((\psi_1 * u)(x), \dots, (\psi_{d_k} * u)(x)\big) \in \mathbb{R}^{d_k},$$

where $*$ denotes convolution, or on a discrete grid the corresponding local convolution sum. Further, let

$$c(x) := (1, x) \in \mathbb{R}^{1+d_x}$$

encode a constant channel together with positional information. Then A represents a bank of local convolutional feature extractors, as commonly used in CNN-based encoder/decoder architectures for operator learning and image-to-image maps.

Empirical risk minimization. Our goal is to minimize the expected risk (1) over the set $\mathcal{F}_M$, i.e.

$$\min_{G_\theta \in \mathcal{F}_M} \mathcal{E}(G_\theta)\,.$$

Here, the distributions μ, $\mu_{\mathcal{U}}$ and ρ_x are known only through n_u i.i.d. samples $((u_1, v_1), ..., (u_{n_u}, v_{n_u})) \in (\mathcal{U} \times \mathcal{V})^{n_u}$, evaluated at n_x points $(x_1, \dots x_{n_x}) \in \mathcal{X}^{n_x}$. Hence, we seek a solution for the empirical risk minimization problem

$$\min_{G_\theta \in \mathcal{F}_M} \hat{\mathcal{E}}(G_\theta)\,, \quad \hat{\mathcal{E}}(G_\theta) = \frac{1}{n_u} \sum_{j=1}^{n_u} \|G_\theta(u_j) - v_j\|_{n_x}^2\,,$$

where $\|\,.\|_{n_x}$ denotes the empirical norm, defined as $\|f\|_{n_x}^2 := \frac{1}{n_x} \sum_{j=1}^{n_x} |f(x_j)|^2$.

Gradient Descent and Initialization. We aim at analyzing the generalization properties of gradient descent, whose basic iterations are given by

$$\begin{aligned} \theta_{t+1}^j &= \theta_t^j - \alpha \partial_{\theta^j} \hat{\mathcal{E}}(G_{\theta_t}) \\ &= \theta_t^j - \frac{\alpha}{n_u} \sum_{i=1}^{n_u} \langle (G_{\theta_t}(u_i) - v_i), \partial_{\theta^j} G_{\theta_t}(u_i) \rangle_{n_x}\,, \end{aligned} \tag{8}$$

with $\alpha > 0$ being the stepsize, for some initialization $\theta_0 \in \Theta$ and
$\langle u, u' \rangle_{n_x} := \frac{1}{n_x} \sum_{i=1}^{n_x} u(x_i)u'(x_i)$ denotes the empirical inner product.

As in Nguyen and Mücke (2023); Nitanda and Suzuki (2020), we assume a symmetric initialization. The parameters for the output layer are initialized as $a_m^{(0)} = \tau$ for $m \in \{1, \ldots, M/2\}$ and $a_m^{(0)} = -\tau$ for $m \in \{M/2+1, \ldots, M\}$, for some $\tau > 0$. Let π_0 be a uniform distribution on the sphere $\mathbb{S}_1^{\tilde{d}-1} = \{b \in \mathbb{R}^{\tilde{d}} \mid \|b\|_2 = 1\} \subset \mathbb{R}^{\tilde{d}}$. The parameters for the input layer are initialized as $b_m^{(0)} = b_{m+M/2}^{(0)}$ for $m \in \{1, \ldots, M/2\}$, where $(b_m^{(0)})_{m=1}^{M/2}$ are independently drawn from the distribution π_0. The aim of the symmetric initialization is to make an initial operator $G_{\theta_0} \equiv 0$, where $\theta_0 = \left(a^{(0)}, B^{(0)}\right)$. Note that this symmetric trick does not have an impact on the limiting NTK, see Zhang et al. (2020), and is just for theoretical simplicity. Indeed, we can relax the symmetric initialization by slightly changing our source condition in Assumption 5 to

$$G^* - G_{\theta_0} = \mathcal{L}_\infty^r H^* \,, \tag{9}$$

for some $H^* \in L^2(\mathcal{U}, \mu_u)$, satisfying $\|H^*\|_{L^2(\mu_u)} \leq R$. The proofs can then be easily adapted.

2.2 Operator valued Neural Tangent Kernel (NTK)

In the more classical setting of nonparametric regression with neural networks, NTK-based techniques have been studied extensively (Nguyen and Mücke, 2023; Bietti and Mairal, 2019; Chen et al., 2020; Golikov et al., 2022; Xu et al., 2024). Complementary to this line of work, recent results have also analyzed the generalization of gradient descent via algorithmic stability for shallow neural networks in Euclidean domains; see, e.g.,Wang et al. (2025). We extend these perspectives to the setting of operator learning, which naturally leads to a vector-valued kernel. We therefore recall the definitions and properties of vector-valued kernels. For further background on vector-valued kernels and their RKHSs, we refer to Carmeli et al. (2005, 2008).

Definition 2 *1. For two function spaces $\mathcal{U}, \mathcal{V}$ we denote by $\mathcal{F}(\mathcal{U}, \mathcal{V})$ the vector space of operators $F : \mathcal{U} \to \mathcal{V}$ and by $\mathcal{C}(\mathcal{U}, \mathcal{V})$ the subspace of continuous operators. $\mathcal{L}(\mathcal{U}, \mathcal{V})$ denotes the space of linear operators from $\mathcal{U} \to \mathcal{V}$ and $\mathcal{L}(\mathcal{V})$ denotes the space of bounded linear operators from $\mathcal{V} \to \mathcal{V}$.*

2. Given a topological space $\mathcal{U}$ and a separable Hilbert space $\mathcal{V}$, a map $K : \mathcal{U} \times \mathcal{U} \longrightarrow \mathcal{L}(\mathcal{V})$ is called a $\mathcal{V}$-reproducing kernel on $\mathcal{U}$ if

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

for any $u_1, \ldots, u_N$ in $\mathcal{U}, v_1, \ldots, v_N$ in $\mathcal{V}$ and $N \geq 1$.

3. Given $u \in \mathcal{U}, K_u : \mathcal{V} \to \mathcal{F}(\mathcal{U}; \mathcal{V})$ denotes the linear operator whose action on a vector $v \in \mathcal{V}$ is the function $K_u v \in \mathcal{F}(\mathcal{U}, \mathcal{V})$ defined by

$$(K_u v)(u') = K(u', u)v \quad u' \in \mathcal{U}.$$

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

$$\begin{aligned} &K_u \in \mathcal{L}(\mathcal{V}, \mathcal{H}_K) \quad \forall\ u \in \mathcal{U} \\ &F(u) = K_u^* F \qquad \forall\ u \in \mathcal{U}, F \in \mathcal{H}_K, \end{aligned}$$

where $K_u^* : \mathcal{H}_K \to \mathcal{V}$ is the adjoint of K_u. Note that from the second property we have that $K(u,t) = K_u^* K_t$. The space $\mathcal{H}_K$ is called the reproducing kernel Hilbert space associated with K, given by

$$\mathcal{H}_K = \overline{\operatorname{span}}\left\{K_u v \mid u \in \mathcal{U}, v \in \mathcal{V}\right\}.$$

4. *A reproducing kernel $K : \mathcal{U} \times \mathcal{U} \to \mathcal{L}(\mathcal{V})$ is called Mercer if $\mathcal{H}_K$ is a subspace of $\mathcal{C}(\mathcal{U};\mathcal{V})$.*

The connection between kernel methods and standard neural networks is established via the NTK, see Jacot et al. (2018); Lee et al. (2019). For standard fully connected neural networks, the feature map of the NTK is defined via the gradient of the neural network at initialization. Similarly, we can define a feature map for neural operators by

$$\begin{aligned} \Phi^M : \mathcal{U} &\to \mathcal{L}(L^2(\mathcal{X}, \rho_x), \Theta) \\ u &\mapsto \Phi^M(u) := \Phi_u^M \end{aligned}$$

with

$$\Phi_u^M(v) := \nabla_\theta \left\langle G_{\theta_0}(u),\, v\right\rangle_{L^2(\rho_x)}.$$

For fixed $u \in \mathcal{U}$, the map $\Phi_u^M : L^2(\mathcal{X}, \rho_x) \to \Theta \simeq \mathbb{R}^P$ satisfies

$$\Phi_u^M(v) = \nabla_\theta \langle G_{\theta_0}(u), v\rangle_{L^2(\rho_x)} = \left(\langle \partial_{\theta_p} G_{\theta_0}(u), v\rangle_{L^2(\rho_x)}\right)_{p=1}^P.$$

With the Euclidean inner product on $\mathbb{R}^P$, its adjoint $(\Phi_u^M)^* : \Theta \to L^2(\mathcal{X}, \rho_x)$ is

$$(\Phi_u^M)^*(\theta) = \sum_{p=1}^P \theta_p\, \partial_p G_{\theta_0}(u), \qquad \theta = (\theta_1, \ldots, \theta_P)^\top \in \mathbb{R}^P.$$

This allows us to define an $L^2(\mathcal{X}, \rho_x)$-reproducing kernel in the sense of Definition 2, for any $u, u' \in \mathcal{U}$ via

$$\begin{aligned} K_M(u,u') = (\Phi_u^M)^* \Phi_{u'}^M &= \sum_{p=1}^P \partial_p G_{\theta_0}(u) \otimes \partial_p G_{\theta_0}(u') \\ &= \frac{1}{M} \sum_{m=1}^M \sigma\left(\left\langle b_m^{(0)}, J(u)\right\rangle\right) \otimes \sigma\left(\left\langle b_m^{(0)}, J(u')\right\rangle\right) + \\ &\quad \frac{\tau^2}{M} \sum_{m=1}^M \sum_{j=1}^{\tilde{d}} \sigma'\left(\left\langle b_m^{(0)}, J(u)\right\rangle\right) J(u)^{(j)} \otimes \sigma'\left(\left\langle b_m^{(0)}, J(u')\right\rangle\right) J(u')^{(j)}, \end{aligned}$$

where we use for the parameter matrix at initialization the notation $B^{(0)} = (b_1^{(0)}, \ldots, b_M^{(0)})^\top \in \mathbb{R}^{M \times \tilde{d}}$, $P = M \cdot (\tilde{d} + 1)$ denotes the number of parameters and for any $u \in \mathcal{U}$ we set $J(u) \in$

$\mathcal{F}(\mathcal{X}, \mathbb{R}^{d_k+d_y+d_b})$ with $J(u)(x) := (A(u)(x), u(x), c(x))^\top$. Further we define the tensor product $\otimes$ with respect to the $L^2(\rho_x)$ inner product $\langle .\,, .\rangle_{L^2(\rho_x)}$ by $(u \otimes v)(f) := \langle v\,,\, f\rangle_{L^2(\rho_x)} u$.

From Example 4 in Carmeli et al. (2008), we deduce that the corresponding vector-valued RKHS is

$$\begin{aligned}\mathcal{H}_M &= \left\{ H \in \mathcal{F}(\mathcal{U}; L^2(\mathcal{X}, \rho_x)) \;\mid\; Hu = \sum_{p=1}^{P} \theta_p \cdot \partial_{\theta_p} G_{\theta_0}(u),\ \|\theta\|_2 < \infty \right\} \\ &= \left\{ H \in \mathcal{F}(\mathcal{U}; L^2(\mathcal{X}, \rho_x)) \;\mid\; (Hu)(x) = \langle \nabla_\theta (G_{\theta_0} u)(x), \theta \rangle_2,\ \theta \in \mathbb{R}^{\mathbb{P}},\ x \in \mathcal{X} \right\},\end{aligned}$$

where we set $\nabla_\theta (G_{\theta_0} u)(x) = (\partial_{\theta_1}(G_{\theta_0} u)(x), ..., \partial_{\theta_P}(G_{\theta_0} u)(x)) \in \mathbb{R}^P$.

Given the Assumptions 1 and 4 we have for all $u \in \mathcal{U}$,

$$\|J(u)(x)\|_1 := \sum_{j=1}^{\tilde{d}} |(J(u)(x))^{(j)}| \leq 1$$

and for all M,

$$\sup_{u,u' \in \mathcal{U}} \|K_M(u, u')\| \leq 4 + \tau^2 \|\sigma'\|_\infty^2 := \kappa^2. \tag{10}$$

Further we have that $\nabla_\theta G_{\theta_0}(.) : \mathcal{U} \to L^2(\mathcal{X}, \rho_x; \mathbb{R}^P)$ is continuous and therefore K_M is Mercer.

In the following result we show that for all $u, u' \in \mathcal{U}$, $K_M(u, u')$ converges for $M \to \infty$ to $K_\infty(u, u') \in \mathcal{C}(L^2(\rho_x), L^2(\rho_x))$ in Hilbert-Schmidt norm, with

$$\begin{aligned}K_\infty(u, u') = \mathbb{E}_{\theta_0}\left[\sigma\left(\left\langle b^{(0)}, J(u)\right\rangle\right) \otimes \sigma\left(\left\langle b^{(0)}, J(u')\right\rangle\right)\right] + \\ \tau^2 \sum_{j=1}^{\tilde{d}} \mathbb{E}_{\theta_0}\left[\sigma'\left(\left\langle b^{(0)}, J(u)\right\rangle\right) J(u)^{(j)} \otimes \sigma'\left(\left\langle b^{(0)}, J(u')\right\rangle\right) J(u')^{(j)}\right].\end{aligned}$$

The proof is provided in Appendix E.2.

Proposition 3 *Given the Assumptions 1, 4 we have for any $u, u' \in \mathcal{U}$ with probability at least $1 - \delta$, where $\delta \in (0, 1)$,*

$$\|K_M(u, u') - K_\infty(u, u')\|_{HS} \leq \frac{4\kappa^2}{\sqrt{M}} \log \frac{2}{\delta}.$$

3. Main Results

In this section we state all of our results and give an outline of the proof.

3.1 Assumptions and Main Results

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

Assumption 4 (Data distribution and Operator class) *1. We assume that the input space $\mathcal{U}$ is a Banach space of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$, and is continuously embedded*[2] *into $L^2(\mathcal{X}, \rho_x; \mathbb{R}^{d_y})$.*

2. The output space $\mathcal{V}$ is a separable Hilbert space of functions that is continuously embedded in $L^2(\mathcal{X}, \rho_x; \mathbb{R})$.

3. Furthermore, we assume that $\mathcal{U}$, $\mathcal{V}$, the operator $A : \mathcal{U} \to \mathcal{F}(\mathcal{X}, \mathbb{R}^{d_k})$, and the bias function $c : \mathcal{X} \to \mathbb{R}^{d_b}$ are bounded. That is, for any $(u, v) \in \operatorname{supp}(\mu)$ and any $x \in \operatorname{supp}(\rho_x)$, we have $|v(x)| \leq 1$ and

$$\|J(u)(x)\|_1 \leq 1,$$

where

$$J(u)(x) := (A(u)(x), u(x), c(x))^\top \in \mathbb{R}^{\tilde{d}}.$$

This assumption fixes the functional framework in which the operator learning problem and its associated NTK are well posed. Inputs are modeled as elements of a Banach space $\mathcal{U}$ of functions on the domain $\mathcal{X}$ (for instance $\mathcal{C}(\mathcal{X}, \mathcal{Y})$), reflecting the infinite-dimensional nature of operator learning problems arising, for example, in PDE-based applications. The output space $\mathcal{V}$ is assumed to be a separable Hilbert space that is continuously embedded into $L^2(\mathcal{X}, \rho_x)$, which ensures that inner products, losses, and kernel evaluations are well defined with respect to the data distribution.

The uniform boundedness assumptions on the input and output functions, as well as on the auxiliary operator A and the bias function c, play a crucial role in guaranteeing stability of the model and of the induced NTK. In particular, requiring uniform bounds on the combined feature vector $J(u)(x)$, which aggregates operator evaluations, input values, and bias terms, ensures that the network outputs, their parameter derivatives, and the resulting kernel remain uniformly controlled over the support of the data distribution. These conditions prevent blow-up phenomena and ensure that the NTK limit exists as a well-defined, bounded operator-valued kernel on $\mathcal{U}$. Note that, if the input is uniformly bounded, that is,

$$\|u(x)\|_1 \leq C_u \qquad \text{for all } x \in \mathcal{X},$$

then the operator classes discussed in Example 2 are uniformly bounded as well, in the sense that

$$\|A(u)(x)\|_1 \leq C_A \qquad \text{for all } x \in \mathcal{X}.$$

Consequently, $J(u)(x)$ remains bounded, provided that the bias term is bounded.

2. Continuous embedding means that there exists a continuous injection $\iota : \mathcal{U} \hookrightarrow L^2(\mathcal{X}, \rho_x)$ such that each $u \in \mathcal{U}$ can be identified with an element of $L^2(\mathcal{X}, \rho_x)$, and there exists a constant $C > 0$ such that $\|u\|_{L^2(\rho_x)} \leq C\|u\|_{\mathcal{U}}$ for all $u \in \mathcal{U}$.

Finally, we impose these assumptions on the input and output spaces to facilitate the application of the theory of random feature approximation for vector-valued kernels as developed in Nguyen and Mücke (2023).

Rates of convergence. In the following, we establish fast convergence rates for the excess risk and thus need to impose refined a priori assumptions in terms of a source condition and eigenvalue decay.

We let

$$\begin{aligned}\mathcal{L}_\infty : L^2(\mathcal{U}, \mu_u; \mathcal{V}) &\to L^2(\mathcal{U}, \mu_u; \mathcal{V}) \\ G &\mapsto \int_{\mathcal{U}} K_\infty(u, \cdot) G(u) \mu_u(du)\end{aligned}$$

denote the kernel integral operator associated to the NTK K_∞. Here the integral is defined in the sense of Bochner integration (see, e.g., Dashti and Stuart (2015), Sec. A.2). Note that $\mathcal{L}_\infty$ is bounded and self-adjoint. Moreover, it is compact and thus has discrete spectrum $\{\mu_j\}_j$, with $\mu_j \to 0$ as $j \to \infty$. By our assumptions, $\mathcal{L}_\infty$ is of trace-class, i.e., has summable eigenvalues (see e.g. Steinwart and Christmann (2008)).

Assumption 5 (Source Condition) *We assume* [3]

$$G^* = \mathcal{L}_\infty^r H^* \,, \tag{11}$$

for some $H^* \in L^2(\mathcal{U}, \mu_u)$*, satisfying* $\|H^*\|_{L^2(\mu_u)} \leq R$ *with* $R > 0$*,* $r \geq 0$*.*

This assumption characterizes the hypothesis space and relates to the regularity of the regression operator G^*. The bigger r is, the smaller the hypothesis space is, the stronger the assumption is, and the easier the learning problem is, as $\mathcal{L}_\infty^{r_1}\left(L^2(\mathcal{U}, \mu_u; \mathcal{V})\right) \subseteq \mathcal{L}_\infty^{r_2}\left(L^2(\mathcal{U}, \mu_u; \mathcal{V})\right)$ if $r_1 \geq r_2$. Note that $G^* \in \mathcal{H}_\infty$ holds for all $r \geq \frac{1}{2}$, in fact we have $ran(\mathcal{L}_\infty^{\frac{1}{2}}) = \mathcal{H}_\infty$ (see e.g. Steinwart and Christmann (2008)). The next assumption relates to the capacity of the hypothesis space.

Assumption 6 (Effective Dimension) *For any* $\lambda > 0$ *we assume*

$$\mathcal{N}_{\mathcal{L}_\infty}(\lambda) := \operatorname{tr}\left(\mathcal{L}_\infty(\mathcal{L}_\infty + \lambda I)^{-1}\right) \leq c_b \lambda^{-b}, \tag{12}$$

for some $b \in [0, 1]$ *and* $c_b > 0$*.*

The number $\mathcal{N}_{\mathcal{L}_\infty}(\lambda)$ is called *effective dimension* or *degrees of freedom* Caponnetto and De Vito (2007). It is related to covering/entropy number conditions, see Steinwart and Christmann (2008). The condition (12) is naturally satisfied with $b = 1$, since $\mathcal{L}_\infty$ is a trace class operator which implies that its eigenvalues $\{\mu_i\}_i$ satisfy $\mu_i \lesssim i^{-1}$. Moreover, if the eigenvalues of $\mathcal{L}_\infty$ satisfy a polynomial decaying condition $\mu_i \sim i^{-c}$ for some $c > 1$, or if $\mathcal{L}_\infty$ is of finite rank, then the condition (12) holds with $b = 1/c$, or with $b = 0$. The case

3. Note that the power of $\mathcal{L}_\infty$ is defined via its SVD: $\mathcal{L}_\infty^r := \sum_{j=1}^\infty \mu_j^r \langle \phi_j, \cdot \rangle \phi_j$.

$b = 1$ is referred to as the capacity independent case. A smaller b allows deriving faster convergence rates for the studied algorithms.

Our next general result establishes an upper bound for the excess risk in terms of the stopping time T and the number of neurons M of our gradient descent algorithm (8), under the assumption that the weights remain in a vicinity of the initialization. The proof is outlined in Section 3.2 and further detailed in Appendix B.

Dependence of r and b on the architecture. The parameters r in the source condition and b in the effective-dimension assumption are not free constants independent of the architecture. They are determined by the spectral properties of the kernel integral operator $\mathcal{L}_\infty$, and therefore depend implicitly on the choice of activation σ, the operator A, the bias function c, and the input distribution μ_u. In the present paper we work at this abstract level, as is standard in RKHS learning theory. A more explicit characterization of how architectural choices translate into concrete values of r and b for general neural operators appears to be open and is left for future work, see also Remark 12.

Theorem 7 *Suppose Assumptions 1, 4, 5 and 6 are satisfied. Assume further that $\alpha \in (0, \kappa^{-2})$, $2r + b > 1$, $n_u \geq n_0 := e^{\frac{2r+b}{2r+b-1}}$,*

$$3 \leq T \leq C n_u^{\frac{1}{2r+b}}, \quad M \geq \tilde{C} B_\tau^6 \log^2(n_u) T^{2r \vee 1}, \quad n_x \geq \tilde{C} B_\tau^2 T^{2r} \log^2(T),$$

and

$$\forall \ t \in [T] \ : \quad \|\theta_t - \theta_0\|_\Theta \leq B_\tau \ , \tag{13}$$

for some $B_\tau \geq 1 + \tau$. Then we have with probability at least $1 - \delta$,

$$\|G_{\theta_T} - G^*\|_{L_2(\mu_u)} \leq \ \bar{C} \ T^{-r} \ \log^3(2/\delta) \ , \tag{14}$$

with $C, \tilde{C}, \bar{C} > 0$ independent of n_u, n_x, M, T, B_τ.

Finally, the above result directly leads to rates of convergence for operator learning in the vvNTK regime.

Corollary 8 *Suppose all assumptions of Theorem 7 are satisfied. Choosing $T_{n_u} = n_u^{\frac{1}{2r+b}}$ gives with probability at least $1 - \delta$*

$$\|G_{\theta_{T_{n_u}}} - G^*\|_{L_2(\mu_u)} \leq \ \bar{C} \ \log^3(2/\delta) \cdot \left(\frac{1}{n_u}\right)^{\frac{r}{2r+b}} \ ,$$

for some $\bar{C} > 0$.

We comment on the above result. Theorem 7 provides an upper bound for the excess risk in estimating nonlinear operators using a shallow neural operator network trained with gradient descent. The risk bound is expressed in terms of the stopping time and

demonstrates that gradient descent is, in fact, a descent algorithm, decreasing the risk at each step with high probability, provided the algorithm is stopped early. We also specify the minimum number of neurons required to achieve this bound, which depends on our a priori assumptions about the model (source condition and eigenvalue decay). Notably, in the misspecified case $r \in (0, 1/2)$, this minimum number increases and matches the results found in Nguyen and Mücke (2023) for nonparametric regression in the NTK regime. In addition, our result provides a minimum number of second stage samples from the discretization to achieve this bound.

We can reformulate Theorem 7 to obtain the sample complexity required to achieve a certain ε-accuracy. On order for the excess risk to have accuracy of $\varepsilon > 0$, the algorithm requires $\mathcal{O}(\varepsilon^{-\frac{2r+b}{r}})$ first stage samples and $\mathcal{O}(\varepsilon^{-2}\log^2(\varepsilon^{-1/r}))$ second stage samples.

Choosing $T = O(n_u^{\frac{1}{2r+b}})$ in Theorem 7 leads to the convergence rate $O(n_u^{-\frac{r}{2r+b}})$, which is known to be minimax optimal in the RKHS framework Caponnetto and De Vito (2007); Blanchard and Mücke (2017).

For the well-specified case $b = 1$, $r = \frac{1}{2}$, where we can only assume that $G^* \in \mathcal{H}_\infty$, we obtain from Theorem 7:

$$\|G_{\theta_{T_n}} - G^*\|_{L_2(\mu_u)} = \mathcal{O}\left(n^{-\frac{1}{4}}\right).$$

Note that in this case we only need $M \geq \mathcal{O}\left(\sqrt{n_u}\log^2 n_u\right)$ many neurons and $n_x \geq \mathcal{O}\left(\sqrt{n_u}\log^2 n_u\right)$ many second stage samples.

In the kernel learning framework, the assumption $2r + b > 1$ refers to easy learning problems and if $2r + b \leq 1$ one speaks of hard learning problems Pillaud-Vivien et al. (2018). In this paper we only investigate easy learning problems and leave the question, how many neurons M and second stage samples are needed to obtain optimal rates in hard learning problems Lin et al. (2020), open for future work.

We finally want to point out that all bounds given in Theorem 7 and Corollary 8 require the iterates θ_t to stay close to the initialization. We show below in Theorem 9 that this assumption is indeed satisfied.

The weights barely move. Our next result shows that the assumption (13) is indeed satisfied and the weights remain in a vicinity of the initialization θ_0. The proof is provided in Appendix C.

Theorem 9 (Bound for the Weights) *Suppose the Assumptions 1, 4, 5 and 6 are satisfied. Let $\delta \in (0, 1]$ and $T \geq 3$ and $M \geq M_0$, $n_u, n_x \geq n_0$, where M_0, n_0 are defined in* (69). *With probability at least $1 - 10\delta$ it holds*

$$\forall\ t \in [T]\ :\quad \|\theta_t - \theta_0\|_\Theta \leq B_\tau\ ,$$

where

$$B_\tau := \max\left\{1 + \tau, C_\diamond \log(T) T^{\max\{0, \frac{1}{2} - r\}} \cdot \mathcal{B}_\delta\left(\frac{1}{\alpha T}\right)\right\},$$

where $C_\diamond > 0$ depends on r, κ, α and is defined in (60), and with

$$\mathcal{B}_\delta\left(\frac{1}{\alpha T}\right) := 1 + 10\kappa\sqrt{\mathcal{N}_{\mathcal{L}_\infty}((\alpha T)^{-1})}\left(\sqrt{\frac{\alpha T}{n_u}} + \sqrt{\frac{\alpha T}{n_x}}\right)\log^{3/2}\left(\frac{12}{\delta}\right).$$

Corollary 10 (Refined Bounds) *Suppose the assumptions of Theorem 9 are satisfied. Let $T = c n_u^{\frac{1}{2r+b}}$, $2r + b > 1$.*

1. *Let $r \geq \frac{1}{2}$ and $n_u \geq n_0$, for some $n_0 \in \mathbb{N}$ depending on δ, r, b. With probability at least $1 - \delta$*

$$B_\tau(\delta, T) \leq C_{R,r,\kappa,\alpha} \log(T)\,, \tag{15}$$

if

$$M \geq C_{\kappa,\alpha,r,\sigma,\|\mathcal{L}_\infty\|}\, \tilde{d}^2 \log^4(T) T^{2r} \log^2(\tilde{d}/\delta)\,,$$
$$n_x \geq C_{\alpha,b,\kappa,\|\mathcal{L}_\infty\|} T^{1+b} \log^3(1/\delta)$$

2. *Let $r < \frac{1}{2}$. With probability at least $1 - \delta$,*

$$B_\tau(\delta, T) \leq C_{R,r,\kappa,\alpha,b} \log(T) T^{1-2r}.$$

This holds if we choose

$$M \geq C_{\kappa,\alpha,r,\sigma,\|\mathcal{L}_\infty\|}\, \tilde{d}^2 \log^4(T) T^{5-8r} \log^2(\tilde{d}/\delta),$$
$$n_x \geq C_{\alpha,b,\kappa,\|\mathcal{L}_\infty\|} T^{2r+b} \log^3(1/\delta).$$

For some constants c, $C_{R,r,\kappa,\alpha,b}$, $C_{\kappa,\alpha,r,\sigma,\|\mathcal{L}_\infty\|}, C_{\alpha,b,\kappa,\|\mathcal{L}_\infty\|} > 0$.

If we incorporate the above worst-case bound on the weights into our main result, Theorem 7, and combine the resulting constraints on M and n_x, we obtain the following corollary.

Corollary 11 *Suppose that the assumptions of Theorem 9 are satisfied. Let $\alpha \in (0, \kappa^{-2})$, assume that $2r + b > 1$, and let $n_u \geq n_0 := e^{\frac{2r+b}{2r+b-1}}$. Set*

$$T = c\, n_u^{\frac{1}{2r+b}}.$$

Then, with probability at least $1 - \delta$, it holds that

$$\|G_{\theta_T} - G^*\|_{L_2(\mu_u)} \;\leq\; C\, n_u^{-\frac{r}{2r+b}}\, \log^3\left(\frac{2}{\delta}\right). \tag{16}$$

This bound is valid provided that the following conditions hold:

1. ***Case*** $r \geq \frac{1}{2}$***:***

$$M \geq \tilde{C}\, \tilde{d}^2 \log^6(T)\, T^{2r} \log^2\left(\frac{\tilde{d}}{\delta}\right), \quad n_x \geq \tilde{C}\, \log^2(T)\, T^{(1+b)\vee 2r} \log^3\left(\frac{1}{\delta}\right).$$

2. ***Case*** $r < \frac{1}{2}$***:***

$$M \geq \tilde{C}\, \tilde{d}^2 \log^6(T)\, T^{7-12r} \log^2\left(\frac{\tilde{d}}{\delta}\right), \quad n_x \geq \tilde{C}\, \log^2(T)\, T^{(2r+b)\vee(2-2r)} \log^3\left(\frac{1}{\delta}\right).$$

Here $c, C, \tilde{C} > 0$ are constants independent of n_u, n_x, M, T, B_τ, and $\tilde{d}$.

Remark 12 (Analysis of NTK Spectrum.) *It is known that a certain eigenvalue decay of the kernel integral operator $\mathcal{L}_\infty$ implies a bound on the effective dimension. Thus, assumptions on the decay of the effective dimension directly relate to the approximation ability of the underlying RKHS, induced by the NTK.*
Meanwhile, there are some results known that shed light on the RKHS of standard neural networks, that are induced by specific NTKs and activation functions (Bietti and Mairal, 2019), Geifman et al. (2020), Chen and Xu (2020), Bietti and Bach (2020),Fan and Wang (2020). For neural operators, however, the RKHS has not yet been explored, at least to the best of our knowledge. We hope to build on the existing results for neural networks in future research to gain a deeper understanding of the RKHS in the context of neural operators.
However, the approximation properties of neural operators are well-studied (Huang et al., 2024; Schwab and Zech, 2021; Marcati and Schwab, 2023; Kovachki et al., 2021, 2024c). For example, Kovachki et al. (2024c) show that all continuous operators can be arbitrarily well approximated by neural operators with a non-polynomial and continuous activation function.

3.2 Outline of Proof

Our proof is based on a suitable error decomposition. To this end, we further introduce additional linearized iterates in $\mathcal{H}_M$:

$$F_{t+1}^M(u) = F_t^M(u) - \frac{\alpha}{n_u} \sum_{j=1}^{n_u} \ell'(F_t^M(u_j), v_j) K_M(u_j, u)\,, \tag{17}$$

$$H_t(u)(x) = \langle \nabla(G_{\theta_0}(u)(x)), \theta_t - \theta_0 \rangle_\Theta\,, \tag{18}$$

with initialization $F_0^M = H_0 = 0$. The iterative algorithm of F_t^M is known as kernel gradient descent (KGD) with respect to the kernel K_M and H_t describes a linear Taylor approximation of the neural Operator around θ_0. We may split

$$\|G_{\theta_T} - G^*\|_{L^2(\mu_u)} \leq \|G_{\theta_T} - \mathcal{S}_M H_T\|_{L^2(\mu_u)} + \|\mathcal{S}_M(H_T - F_T^M)\|_{L^2(\mu_u)} + \|\mathcal{S}_M F_T^M - G^*\|_{L^2(\mu_u)}\,, \tag{19}$$

where $\mathcal{S}_M : \mathcal{H}_M \hookrightarrow L^2(\mathcal{U}, \mu_u)$ is the inclusion of $\mathcal{H}_M$ into $L^2(\mathcal{U}, \mu_u)$. The first error in (19) describes an Taylor approximation error. More precisely we use a Taylor expansion in θ_t around the initialization θ_0. For any $x \in \mathcal{X}$ and $t \in [T]$, we have

$$\begin{aligned} G_{\theta_t}(u)(x) &= G_{\theta_0}(u)(x) + \mathcal{S}_M \langle \nabla(G_{\theta_0}(u)(x)), \theta_t - \theta_0 \rangle_\Theta + r_{(\theta_t,\theta_0)}(u)(x) \\ &= \mathcal{S}_M H_t(u)(x) + r_{(\theta_t,\theta_0)}(u)(x) \ . \end{aligned} \tag{20}$$

Here, $r_{(\theta_t,\theta_0)}(x)$ denotes the Taylor remainder and can be uniformly bounded by

$$\|G_{\theta_T} - \mathcal{S}_M H_T\|_{L^2(\mu_u)} \leq \|r_{(\theta_t,\theta_0)}\|_\infty \lesssim B_\tau \ \frac{\|\theta_t - \theta_0\|_\Theta^2}{\sqrt{M}} \ ,$$

as Proposition 25 shows. This requires the iterates $\{\theta_t\}_{t\in[T]}$ to stay close to the initialization θ_0, i.e.

$$\sup_{t\in[T]} \|\theta_t - \theta_0\|_\Theta \leq B_\tau \ ,$$

with high probability, for some $B_\tau < \infty$. We show in Theorem 9 that this is satisfied for sufficiently many neurons. The second error term in (19) depends on the number of neurons M and on the number of second stage samples n_x, see Theorem 18. More precisely, we obtained

$$\|\mathcal{S}_M(H_T - F_T^M)\|_{L^2(\mu_u)} \lesssim \log(T) B_\tau^3 \left(\frac{1}{\sqrt{M}} + \frac{1}{\sqrt{n_x}} \right) \ ,$$

Nguyen and Mücke (2023) considered a similar error term, but for neural networks instead of operators. We used improved spectral regularisation inequalities to obtain rates for second stage samples and a shorter, more concise proof than in Nguyen and Mücke (2023). The last error term in (19) describes the generalization error of KGD. Applying the results of Nguyen and Mücke (2023), we obtain that, with high probability,

$$\|\mathcal{S}_M F_T^M - G^*\|_{L^2(\mu_u)} \ \lesssim \ T^{-r},$$

provided that the number of random features satisfies

$$M \ \gtrsim \ \log(n_u)\, T^{2r}.$$

See Proposition 19 for a precise statement. As a result, we arrive at an overall bound of Theorem 7

$$\|G_{\theta_T} - G^*\|_{L^2(\mu_u)} \lesssim \log(T) B_\tau^3 \left(\frac{1}{\sqrt{M}} + \frac{1}{\sqrt{n_x}} \right) + T^{-r} \ . \tag{21}$$

Remark 13 *As shown in* (21)*, the number of neurons M and the sample size $n_\mathcal{U}$ must be chosen of order larger than*

$$O(T^{2r}) \ = \ O\big(n_\mathcal{U}^{\frac{2r}{2r+b}}\big)$$

in order to ensure that the approximation error between the neural operator G_{θ_T} and the random feature method F_T^M attains the optimal learning rate of kernel estimators. At first glance, it may seem nonintuitive that smoother target functions (i.e., larger values of the

smoothness parameter r) require a larger number of neurons, or equivalently random features M, to achieve optimal rates, since smoothness is often associated with easier learning problems. The key reason is that the optimal learning rate itself depends explicitly on r. To make this effect explicit, note that the kernel estimator admits the closed-form representation

$$F_T^M = \widehat{\mathcal{S}}_M^* \, \phi_T\big(\widehat{\mathcal{S}}_M \widehat{\mathcal{S}}_M^*\big) \, \mathbf{v},$$

where

$$\phi_T(x) := \alpha \sum_{j=0}^{T-1} (1 - \alpha x)^j$$

denotes the spectral filter associated with gradient descent (cf. Nguyen and Mücke (2023)) and $\mathbf{v} = (v_1, \dots v_{n_u}) \in \mathcal{V}^{n_u}$. Moreover, for each fixed $x \in (0, \alpha^{-1})$, one has $\phi_T(x) \to x^{-1}$ as $T \to \infty$. Consequently, for large n and large T, and in a low-noise regime in which $v_i \approx G^*(u_i)$, the estimator satisfies

$$\mathcal{S}_M F_T^M \;\approx\; \mathcal{L}_M \phi_T(\mathcal{L}_M) \, G^*,$$

and can therefore be interpreted as a (regularized) projection of G^* onto the image of $\mathcal{L}_M$. In order to achieve the optimal rate, this projection must be sufficiently close to G^* in the L^2-norm, that is, it must intersect the ball

$$B_r \;:=\; \big\{ F \in L^2(\mathcal{U}, \mu_u) : \|G^* - F\|^2 \le n_u^{-\frac{2r}{2r+b}} \big\}.$$

As r increases, the radius of this ball shrinks, making the approximation requirement increasingly restrictive. Consequently, the image $\mathrm{Im}(\mathcal{L}_M)$—and hence the random feature RKHS $\mathcal{H}_M$— must grow accordingly to ensure that this intersection remains nonempty. Since the size of $\mathcal{H}_M$ increases with M, this provides a natural explanation for why a larger number of random features is required to learn smoother target functions at the optimal rate; see also Figure 2.

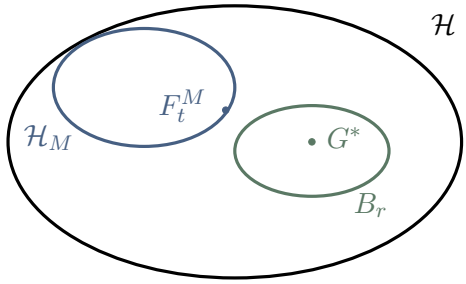

Figure 2: Geometric illustration of the approximation requirement for random feature methods. The ball B_r around the target function G^* shrinks as the smoothness parameter r increases. To attain the optimal learning rate, the RKHS $\mathcal{H}_M$ must be sufficiently rich to intersect B_r.

4. Numerical Illustration

In this section we support our theoretical investigation that in the well-specified case, the optimal generalization properties can be achieved with a neural operator comprising only $M = O(\sqrt{n_u})$ neurons and $n_x = O(\sqrt{n_u})$ second stage samples (see Theorem 7). Remarkably, the empirical evidence presented in Figure 4 and 5 aligns with this theoretical prediction. More precisely we considered the one dimensional Poisson problem,

$$\begin{aligned} -\Delta v &= u && \text{in } (0,1) \\ v &= 0 && \text{on } \{0,1\}. \end{aligned}$$

We aim to estimate the solution operator that maps the input function u to the solution v. In the one-dimensional case with zero boundary conditions, the analytical form of the solution operator is straightforward to obtain. Specifically, we have:

$$v(x) = \int_0^1 G(x,y)u(y)dy,$$

with $G(x,y) = \frac{1}{2}(x + y - |x - y|) - xy$. Therefore, we randomly sampled 800 polynomials u_i and computed their analytical solutions v_i. These functions $\{u_i, v_i\}_{i=1}^{400}$ were then used as training data to train our neural operator using GD, and a symmetric initialization (2.1). The test set $\{u_i, v_i\}_{i=401}^{800}$ was used for evaluation. As shown in Figures 6 and 7, after reaching $n_x = M = T = \sqrt{n_u} = 20$, there is no further significant improvement in the neural operator's performance.

The empirical efficiency of neural operators (NOs) has been demonstrated in numerous works through extensive numerical simulations. Importantly, these studies consider significantly more complex problem settings than the one-dimensional linear example investigated in this paper.

A setting closely related to our NTK-based perspective is presented in Nelsen and Stuart (2024). The authors provide detailed simulation studies on the direct training of operator-valued random feature kernel estimators. In particular, their experiments also analyze the influence of the number of random features on the approximation accuracy. The learned object is the solution operator associated with PDEs such as the Burgers equation and the Darcy flow equation. These simulations offer valuable insight into the statistical–computational trade-offs in random feature–based operator learning.

A broader comparative overview is provided in Gonon et al. (2024), where extensive numerical experiments benchmark different learning-based approaches for approximating solution operators. The comparison includes Fourier Neural Operators (FNOs), convolutional neural networks (CNNs), and DeepONets, as well as classical numerical schemes such as finite difference methods (FDM) and finite element methods (FEM). The experiments span a variety of nonlinear PDEs, including the Allen–Cahn equation and reaction–diffusion systems, highlighting both accuracy and computational efficiency across methods.

In Kovachki et al. (2024a), multiple classes of neural operators—such as Fourier Neural Operators (FNOs), Graph Neural Operators (GNOs), and low-rank neural operators—are compared across a broad range of experimental settings. These experiments investigate

multi-layer architectures and analyze both predictive performance and computational complexity. The study provides a systematic comparison of architectural choices within the neural operator framework.

Further large-scale numerical experiments with Fourier Neural Operators (FNOs) have demonstrated their capability to handle complex and high-dimensional tasks. Notably, FNOs have been successfully applied to global weather forecasting at high spatial resolution (Pathak et al., 2022), as well as to more challenging differential equations with complex geometries (Lowery et al., 2024; Li et al., 2021). These studies highlight the scalability and robustness of neural operators in multidimensional and real-world scientific computing settings, reinforcing their practical relevance for learning solution operators across diverse applications.

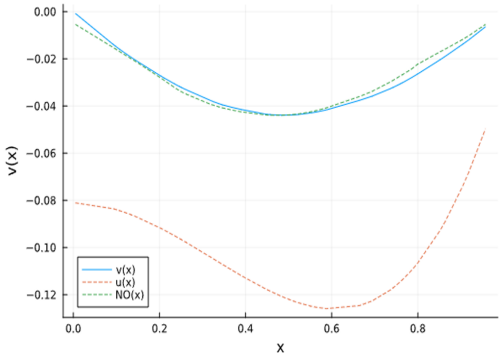

Figure 3: A random realization of a polynomial u, its solution v and the estimator $NO(u)$.

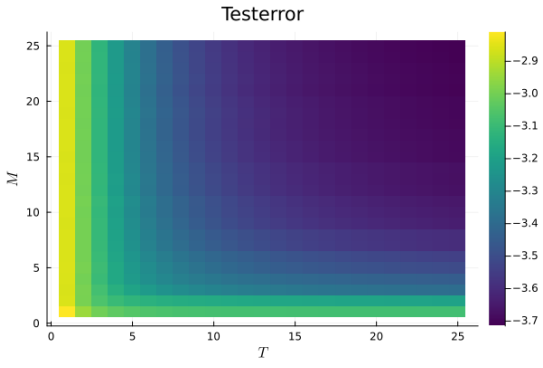

Figure 4: The logarithmic test-error for different choices of neurons M and iterations T and fixed $n_x = 50$.

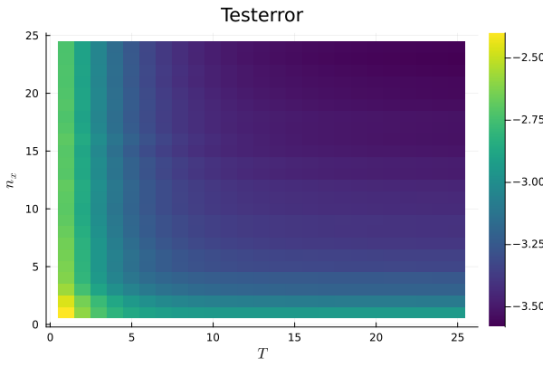

Figure 5: The logarithmic test-error for different choices of n_x and iterations T and fixed $M = 50$.

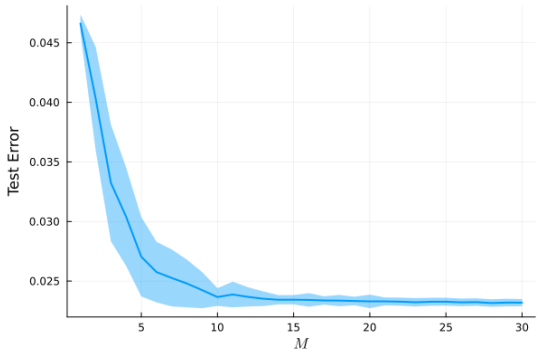

Figure 6: The test-error for different choices of M and fixed $T = 50$ and $n_x = 50$.

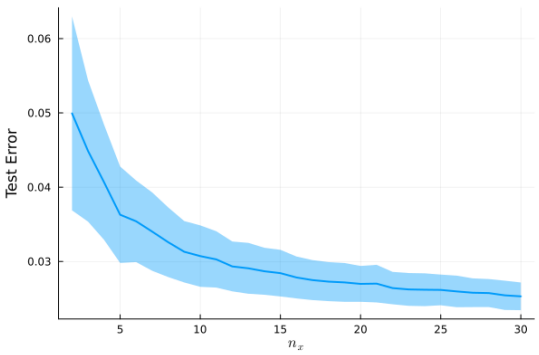

Figure 7: The test-error for different choices of n_x and fixed $T = 50$ and $M = 50$.

5. Conclusion

In this work, we studied two-layer neural operators in the neural tangent kernel regime and established optimal convergence rates for operator learning under standard source-condition and effective-dimension assumptions. Our analysis combines tools from vector-valued RKHS theory, random feature approximation, and the analysis of shallow neural networks, and extends these techniques to the setting of learning operators between function spaces. In particular, we derived high-probability excess-risk bounds for early-stopped gradient descent, together with explicit requirements on the number of neurons and second-stage samples needed to achieve these rates.

A central motivation for this framework is the learning of solution operators of partial differential equations. Neural operators provide a flexible and scalable alternative to classical discretization-based methods, especially in settings where repeated evaluations of a solution map are required. Our results give theoretical support for this approach in the NTK regime and show that, under suitable assumptions, neural operators achieve the same minimax-optimal rates known from classical nonparametric learning in RKHS.

At the same time, several important questions remain open. First, our analysis is restricted to easy learning problems, that is, to the regime $2r + b > 1$. Extending the theory to hard learning problems and understanding the corresponding requirements on the number of neurons M and second-stage samples is an interesting direction for future work. Second, while our rates are expressed in terms of the source-condition parameter r and the effective-dimension parameter b, a more explicit understanding of these quantities requires a deeper analysis of the limiting neural tangent kernel K_∞ and its associated kernel integral operator. In particular, the spectral properties of the NTK and the resulting structure of the induced vector-valued RKHS for neural operators are, to the best of our knowledge, still largely open.

We hope that the present work provides a useful first step toward a learning-theoretic understanding of neural operators and that it motivates further research on their approximation properties, induced RKHS structure, and statistical-computational trade-offs in more general settings.

Glossary

$\mathcal{U}$	The function input space, mapping from $\mathcal{X} \to \mathcal{Y} \subset \mathbb{R}^{d_y}$.	5
$\mathcal{V}$	The function target space, mapping from $\mathcal{X} \to \tilde{\mathcal{Y}} \subset \mathbb{R}$.	5
μ	Probability measure on $\mathcal{U} \times \mathcal{V}$.	5
$\mathcal{X}$	Input space of our function spaces.	5
ρ_x	Probability measure on the input data space $\mathcal{X}$.	5
$\mathcal{E}$	Mean squared error, $\mathcal{E}(G) := \mathbb{E}_\mu[\|G(u) - v\|^2_{L^2(\rho_x)}$.	5
G^*	Target operator, $G^* := \int_{\mathcal{V}} v(.)\mu(dv\|u)$.	5
M	The width of the neural operators.	5
P	Number of all trainable parameters.	12
σ	The activation function acting point-wise from $\mathbb{R} \to \mathbb{R}$.	7
$\mathcal{F}_M$	Class of neural operators.	5
n_u	Number of samples drawn from μ.	10
n_x	Number of samples drawn from ρ_x.	10
$\|\,.\|_{n_x}$	Empirical norm, $\|f\|^2_{n_x} := \frac{1}{n_x}\sum_{j=1}^{n_x} \|f(x_j)\|^2$.	10
$\langle u, u' \rangle_{n_x}$	Empirical inner product $\langle u, u' \rangle_{n_x} := \frac{1}{n_x}\sum_{i=1}^{n_x} u(x_i)u'(x_i)$.	11
$\langle .\,, .\rangle_2$	Euclidean inner product.	5
$\|\cdot\|_\Theta$	Norm of the parameter space Θ, $\|\theta\|^2_\Theta = \|a\|_2^2 + \|B\|_F^2$.	6
$J(u) \in \mathcal{F}(\mathcal{X}, \mathbb{R}^{d_k+d_y+d_b})$	$J(u)(x) := (A(u)(x), u(x), c(x))^\top$.	12
$\langle .\,, .\rangle_{L^2(\rho_x)}$	$L^2(\mathcal{X}, \rho_x)-$ inner product.	13
$\otimes$	Tensor product $(u \otimes v)(f) := u \cdot \langle v\,,\, f\rangle_{L^2(\rho_x)}$.	13

Appendix A. Preliminaries

For our analysis we need some further notation. We define

$$L^2(\mathcal{U},\mu_u) \;:=\; L^2\big(\mathcal{U},\mu_u;L^2(\mathcal{X},\rho_x)\big),$$

equipped with the norm

$$\|G\|^2_{L^2(\mu_u)} \;:=\; \int_{\mathcal{U}} \|G(u)\|^2_{L^2(\rho_x)}\,\mu_u(du).$$

For each $M\in\mathbb{N}\cup\{\infty\}$, we denote by

$$\mathcal{S}_M : \mathcal{H}_M \hookrightarrow L^2(\mathcal{U},\mu_u)$$

the canonical inclusion operator of $\mathcal{H}_M$ into $L^2(\mathcal{U},\mu_u)$.

The adjoint operator $\mathcal{S}_M^* : L^2(\mathcal{U},\mu_u)\longrightarrow\mathcal{H}_M$ is identified as

$$\mathcal{S}_M^* G = \int_{\mathcal{U}} K_{M,u}G(u)\mu_u(du)\;,$$

where $K_{M,u}$ denotes the element of $\mathcal{H}_M$ equal to the function $t\mapsto K_M(u,t)$. The covariance operator $\Sigma_M : \mathcal{H}_M\longrightarrow\mathcal{H}_M$ and the kernel integral operator $\mathcal{L}_M : L^2(\mathcal{U},\mu_u)\to L^2(\mathcal{U},\mu_u)$ are given by

$$\Sigma_M G := \mathcal{S}_M^*\mathcal{S}_M G = \int_{\mathcal{U}} K_{M,u}G(u)\mu_u(du)\;, \tag{22}$$

$$\mathcal{L}_M G := \mathcal{S}_M\mathcal{S}_M^* G = \int_{\mathcal{U}} K_{M,u}G(u)\mu_u(du)\;, \tag{23}$$

which can be shown to be positive, self-adjoint, trace class (and hence is compact). The empirical versions of these operators are given by

$$\widehat{\mathcal{S}}_M : \mathcal{H}_M\longrightarrow\mathcal{V}^{n_u}, \qquad \Big(\widehat{\mathcal{S}}_M G\Big)_j = K^*_{M,u_j}G,$$

$$\widehat{\mathcal{S}}_M^* : \mathcal{V}^{n_u}\longrightarrow\mathcal{H}_M, \qquad \widehat{\mathcal{S}}_M^*\mathbf{v} = \frac{1}{n_u}\sum_{j=1}^{n_u} K_{M,u_j}v_j,$$

$$\widehat{\Sigma}_M := \widehat{\mathcal{S}}_M^*\widehat{\mathcal{S}}_M : \mathcal{H}_M\longrightarrow\mathcal{H}_M, \qquad \widehat{\Sigma}_M = \frac{1}{n_u}\sum_{j=1}^{n_u} K_{M,u_j}K^*_{M,u_j},$$

where $\mathcal{V}^{n_u}$ is equipped with the norm $\|\mathbf{v}\|^2_{\mathcal{V}^{n_u}} := \frac{1}{n_u}\sum_{i=1}^{n_u}\|v_i\|^2_{L^2(\rho_x)}$. We introduce the following definitions similar to Definition 2 of Rudi and Rosasco (2017):

$$\mathcal{Z}_M : \Theta\to\mathcal{H}_M, \qquad (\mathcal{Z}_M\theta)(\cdot) = \nabla G_{\theta_0}(\cdot)^\top\theta,$$

$$\mathcal{Z}_M^* : \mathcal{H}_M\to\Theta, \qquad \mathcal{Z}_M^* G = \int_{\mathcal{U}}\int_{\mathcal{X}} \nabla G_{\theta_0}(u)(x)G(u)(x)d\rho_x(x)d\mu_u(u),$$

$$\widehat{\mathcal{Z}}_M : \Theta \to \left(\mathbb{R}^{n_u \times n_x}, \frac{1}{\sqrt{n_u n_x}}\|\cdot\|_F\right), \qquad \left(\widehat{\mathcal{Z}}_M \theta\right)_{i,j} = \nabla G_{\theta_0}(u_i)(x_j)^\top \theta,$$
$$\widehat{\mathcal{Z}}_M^* : \left(\mathbb{R}^{n_u \times n_x}, \frac{1}{\sqrt{n_u n_x}}\|\cdot\|_F\right) \to \Theta, \qquad \widehat{\mathcal{Z}}_M^* a = \frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x} \nabla G_{\theta_0}(u_i)(x_j) a_{i,j},$$

$$\mathcal{C}_M : \Theta \to \Theta, \qquad \mathcal{C}_M = \int_{\mathcal{U}}\int_{\mathcal{X}} \nabla G_{\theta_0}(u)(x) \nabla G_{\theta_0}(u)(x)^\top d\rho_x(x) d\mu_u(u), \tag{24}$$
$$\widehat{\mathcal{C}}_M : \Theta \to \Theta, \qquad \widehat{\mathcal{C}}_M = \frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x} \nabla G_{\theta_0}(u_i)(x_j) \nabla G_{\theta_0}(u_i)(x_j)^\top. \tag{25}$$

Remark 14 *Note that $\mathcal{C}_M$ and $\widehat{\mathcal{C}}_M$ are self-adjoint and positive operators, with spectrum in $[0, \kappa^2]$ and we further have $\mathcal{C}_M = \mathcal{Z}_M^* \mathcal{Z}_M$, $\widehat{\mathcal{C}}_M = \widehat{\mathcal{Z}}_M^* \widehat{\mathcal{Z}}_M$, $\Sigma_M = \mathcal{Z}_M \mathcal{Z}_M^*$.*

Appendix B. Error Bounds

B.1 Error Decomposition

Recall the error decomposition from Section 3.2:

$$\|G_{\theta_T} - G^*\|_{L^2(\mu_u)} \le \underbrace{\|G_{\theta_T} - \mathcal{S}_M H_T\|_{L^2(\mu_u)}}_{\mathcal{I}} + \underbrace{\|\mathcal{S}_M (H_T - F_T^M)\|_{L^2(\mu_u)}}_{\mathcal{II}} + \underbrace{\|\mathcal{S}_M F_T^M - G^*\|_{L^2(\mu_u)}}_{\mathcal{III}}. \tag{26}$$

This section is devoted to bounding each term on the right hand side of (26). To this end, we need the following assumption:

Assumption 15 *Let $T \in \mathbb{N}$. Assume we have*

$$\forall\ t \in [T]\ :\quad \|\theta_t - \theta_0\|_\Theta \le B_\tau\ , \tag{27}$$

for some constant $B_\tau \ge 1 + \tau$.

We will show in Section C that this assumption is satisfied with high probability, if M is sufficiently large.

B.2 Bounding $\mathcal{I}$

In this section we provide an estimate of the first error term $\|G_{\theta_T} - \mathcal{S}_M H_T\|_{L^2(\mu_u)}$ in (26).

Proposition 16 *Suppose that Assumption 15 is satisfied.*

$$\|G_{\theta_T} - \mathcal{S}_M H_T\|_{L^2(\mu_u)} \leq \frac{\tilde{C}_{\nabla G}(B_\tau) B_\tau^2}{\sqrt{M}} \,,$$

with $\tilde{C}_{\nabla G}(B_\tau) := 4 \max\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\} B_\tau$.

Proof [Proposition 16] From Assumption 15 and Proposition 25 we immediately obtain for all $x \in \mathcal{X}, u \in \mathcal{U}$

$$\begin{aligned} |G_{\theta_T}(u)(x) - \mathcal{S}_M H_T(u)(x)| &= |r_{(\theta_0,\theta_T)}(u)(x)| \\ &\leq \frac{C_{\nabla G}(B_\tau)}{\sqrt{M}} \, \|\theta_T - \theta_0\|_\Theta^2 \\ &\leq \frac{C_{\nabla G}(B_\tau) B_\tau^2}{\sqrt{M}} \,, \end{aligned} \tag{28}$$

with $C_{\nabla G}(B_\tau) := 2 \max\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\} \max\{1, B_\tau + \tau\}$. Moreover, since $B_\tau \geq \tau + 1$, we obtain

$$C_{\nabla G}(B_\tau) = 2 \max\left\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\right\} (B_\tau + \tau) \leq 4 \max\left\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\right\} B_\tau =: \tilde{C}_{\nabla G}(B_\tau).$$

■

B.3 Bounding $\mathcal{II}$

In this section we estimate the second term $\|\mathcal{S}_M(H_T - F_T^M)\|_{L^2(\mu_u)}$ in (26). A short calculation proves the following recursion:

Lemma 17 *Let* $t \in \mathbb{N}$, $M \in \mathbb{N}$. *Define* $\hat{u}_t := H_t - F_t^M \in \mathcal{H}_M$. *Then* $(\hat{u}_t)_t$ *follows the recursion* $\hat{u}_0 = 0$ *and*

$$\begin{aligned} \hat{u}_{t+1} &= (Id - \alpha \widehat{\Sigma}_M)\hat{u}_t - \alpha \mathcal{Z}_M \hat{\xi}_t^{(1)} - \alpha \mathcal{Z}_M \hat{\xi}_t^{(2)} - \alpha \widehat{\mathcal{S}}_M^* \hat{\xi}_t^{(3)} \\ &= \alpha \sum_{s=0}^{t-1} (Id - \alpha \widehat{\Sigma}_M)^s \left(\mathcal{Z}_M \hat{\xi}_{t-s}^{(1)} + \mathcal{Z}_M \hat{\xi}_{t-s}^{(2)} + \widehat{\mathcal{S}}_M^* \hat{\xi}_{t-s}^{(3)} \right) \,, \end{aligned}$$

where

$$\begin{aligned} \hat{\xi}_t^{(1)} &= \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} (G_{\theta_t}(u_i)(x_j) - v_i(x_j)) \left(\nabla G_{\theta_t}(u_i)(x_j) - \nabla G_{\theta_0}(u_i)(x_j)\right) \in \Theta \,, \\ \hat{\xi}_t^{(2)} &= \frac{1}{n_u} \sum_{i=1}^{n_u} (\hat{\mathbf{s}}_i - \mathbf{s_i}) \in \Theta \,, \\ \hat{\xi}_t^{(3)} &= \bar{r}_{(\theta_0,\theta_t)} \in \mathcal{V}^{n_u} \,, \end{aligned}$$

with $\bar{r}_{(\theta_0,\theta_t)} = \left(r_{(\theta_0,\theta_t)}(u_1), \ldots, r_{(\theta_0,\theta_t)}(u_{n_u})\right)$ *and*
$\hat{\mathbf{s}}_i \,, \mathbf{s}_i \in \Theta$, $\hat{\mathbf{s}}_i^{(m)} := \langle (G_{\theta_t}(u_i) - v_i), \partial_m G_{\theta_0}(u_i) \rangle_{n_x}$, $\mathbf{s}_i^{(m)} := \langle (G_{\theta_t}(u_i) - v_i), \partial_m G_{\theta_0}(u_i) \rangle_{L^2(\rho_x)}$
.

Proof [Lemma 17] Plugging in H_t the recursive definition of θ_t leads to

$$\begin{aligned} H_{t+1} &= \langle \nabla G_{\theta_0}(u), \theta_{t+1} - \theta_0 \rangle_\Theta \\ &= H_t - \alpha \mathcal{Z}_M \hat{\xi}_t^{(1)} - \alpha \mathcal{Z}_M \hat{\xi}_t^{(2)} - \alpha \widehat{\mathcal{S}}_M^* \hat{\xi}_t^{(3)} - \frac{\alpha}{n} \sum_{i=1}^n K_M(u_i, .)(H_t(u_i) - v_i). \end{aligned} \tag{29}$$

Here we used the fact that $G_{\theta_0} \equiv 0$. Therefore, applying a Taylor expansion again yields $G_{\theta_t} = H_t + r_{(\theta_0, \theta_t)}$. Combining the above recursive formula of H_t with the recursive definition of F_t^M results in the first equation of the statement. The second equation can be derived by iterating the first equation t times. ■

Theorem 18 *Let $\delta \in (0, 1]$, $\alpha < 1/\kappa^2$. Suppose Assumptions 1, 15, are satisfied. Let $M \geq M_0 := (8\tilde{d}\kappa^2 \beta_\infty \alpha T) \vee 8\kappa^4 \|\mathcal{L}_\infty\|^{-1} \log^2 \frac{2}{\delta}$ with $\beta_\infty = \log \frac{4\kappa^2(\mathcal{N}_{\mathcal{L}_\infty}(\alpha T)+1)}{\delta \|\mathcal{L}_\infty\|}$. Then we have with probability at least $1 - \delta$,*

$$\forall t \in [T]: \quad \|\mathcal{S}_M \hat{u}_t\|_{L^2(\mu_u)} \leq 4\xi(M, n_x, B_\tau, \kappa) \left(\log(T) + 6\right) ,$$

where $u_t := H_t - F_t^M$ and

$$\begin{aligned} &\xi(M, n_x, B_\tau, \kappa, \delta) \\ &:= \left(\frac{C_{\nabla G}(B_\tau) B_\tau}{\sqrt{M}} \left(\kappa B_\tau + \frac{C_{\nabla G}(B_\tau)}{\sqrt{M}} B_\tau^2 + 1 \right) + \frac{4\kappa^2 (B_\tau + 1)}{\sqrt{n_x}} \log\left(\frac{16}{\delta}\right) + \frac{(2\kappa + 1) C_{\nabla G}(B_\tau) B_\tau^2}{\sqrt{M}} \right). \end{aligned}$$

Here, the constant $C_{\nabla G}(B_\tau)$ is defined in Proposition 26. If we further assume

$$M \geq M_{II} := \max\{M_0, M_1\}$$

with

$$M_1 := 48 C_{\nabla G}(B_\tau)^2 B_\tau^4 \kappa T^{2r} \left(\log(T) + 6\right)^2 \text{ and } n_x \geq 32\kappa^4 \left(B_\tau + 1\right)^2 T^{2r} \left(\log(T) + 6\right)^2 \log\left(\frac{16}{\delta}\right),$$

we have

$$\forall t \in [T]: \quad \|\mathcal{S}_M \hat{u}_t\|_{L^2(\mu_u)} \leq \frac{1}{T^r} .$$

Proof [Theorem 18] Applying (Nguyen and Mücke, 2023, Proposition A.15) gives (for $\lambda := (\alpha T)^{-1}$ and $M \geq (8\tilde{d}\kappa^2 \beta_\infty \alpha T) \vee 8\kappa^4 \|\mathcal{L}_\infty\|^{-1} \log^2 \frac{2}{\delta}$) with probability at least $1 - \delta$

$$\|\mathcal{S}_M \hat{u}_t\|_{L^2(\mu_u)} = \|\Sigma_M^{\frac{1}{2}} \hat{u}_t\|_{\mathcal{H}_M} \leq 2 \|\widehat{\Sigma}_M^{\frac{1}{2}} \hat{u}_t\|_{\mathcal{H}_M} + 2 \frac{\|\hat{u}_t\|_{\mathcal{H}_M}}{\sqrt{\alpha T}} , \tag{30}$$

Let $a \in \{0, 1/2\}$. Using Lemma 17, we find for any $t \in [T]$,

$$\begin{aligned}
\|\widehat{\Sigma}_M^a \hat{u}_{t+1}\|_{\mathcal{H}_M} &= \alpha \left\| \sum_{s=0}^{t-1} \widehat{\Sigma}_M^a (Id - \alpha\widehat{\Sigma}_M)^s \left(\mathcal{Z}_M \hat{\xi}_{t-s}^{(1)} + \mathcal{Z}_M \hat{\xi}_{t-s}^{(2)} + \widehat{\mathcal{S}}_M^* \hat{\xi}_{t-s}^{(3)} \right) \right\|_{\mathcal{H}_M} \\
&\leq \alpha^{1-a} \sum_{s=0}^{T-1} \|(\alpha\widehat{\Sigma}_M)^a (Id - \alpha\widehat{\Sigma}_M)^s \mathcal{Z}_M \hat{\xi}_{T-s}^{(1)}\|_{\mathcal{H}_M} + \\
&\quad \alpha^{1-a} \sum_{s=0}^{T-1} \|(\alpha\widehat{\Sigma}_M)^a (Id - \alpha\widehat{\Sigma}_M)^s \mathcal{Z}_M \hat{\xi}_{T-s}^{(2)}\|_{\mathcal{H}_M} + \\
&\quad \alpha^{1-a} \sum_{s=0}^{T-1} \|(\alpha\widehat{\Sigma}_M)^a (Id - \alpha\widehat{\Sigma}_M)^s \widehat{\mathcal{S}}_M^* \hat{\xi}_{T-s}^{(3)}\|_{\mathcal{H}_M} \,. \qquad (31)
\end{aligned}$$

From Proposition 33 we further obtain with probability at least $1 - \delta$,

$$\|\widehat{\Sigma}_M^a \hat{u}_{t+1}\|_{\mathcal{H}_M} \leq \alpha^{1-a} \left(\frac{2\eta_{a+\frac{1}{2}}(T)}{\sqrt{\alpha}} + \frac{2\eta_a(T)}{\sqrt{\alpha T}} \right) \max_{s \leq T} \left(\|\hat{\xi}_s^{(1)}\|_\Theta + \|\hat{\xi}_s^{(2)}\|_\Theta + \|\xi_s^{(3)}\|_{\mathcal{V}^{n_u}} \right), \quad (32)$$

where η_a was defined in Lemma 31.

Bounding $\|\hat{\xi}_t^{(1)}\|_\Theta$: We have

$$\begin{aligned}
\|\hat{\xi}_t^{(1)}\|_\Theta &= \left\| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} (G_{\theta_t}(u_i)(x_j) - v_i(x_j)) \left(\nabla G_{\theta_t}(u_i)(x_j) - \nabla G_{\theta_0}(u_i)(x_j) \right) \right\|_\Theta \\
&\leq \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} |G_{\theta_t}(u_i)(x_j) - v_i(x_j)| \, \|\nabla G_{\theta_t}(u_i)(x_j) - \nabla G_{\theta_0}(u_i)(x_j)\|_\Theta \,.
\end{aligned}$$

From the Assumption 4 we have $|v_i(x_j)| \leq 1$ and from Proposition 26 we have

$$|G_{\theta_t}(u)(x)| \leq \kappa \|\theta_t - \theta_0\|_\Theta + \frac{C_{\nabla G}(B_\tau)}{\sqrt{M}} \|\theta_t - \theta_0\|_\Theta^2,$$

with $C_{\nabla G}(B_\tau) := 2 \max\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\} \max\{1, B_\tau + \tau\}$. Using these bounds we obtain

$$\begin{aligned}
\|\hat{\xi}_t^{(1)}\|_\Theta &\leq \left(\kappa \|\theta_t - \theta_0\|_\Theta + \frac{C_{\nabla G}(B_\tau)}{\sqrt{M}} \|\theta_t - \theta_0\|_\Theta^2 + 1 \right) \sup_{x \in \mathcal{X}, u \in \mathcal{U}} \|\nabla G_{\theta_t}(u)(x) - \nabla G_{\theta_0}(u)(x)\|_\Theta \\
&\leq \left(\kappa \|\theta_t - \theta_0\|_\Theta + \frac{C_{\nabla G}(B_\tau)}{\sqrt{M}} \|\theta_t - \theta_0\|_\Theta^2 + 1 \right) \left(\frac{C_{\nabla G}(B_\tau)}{\sqrt{M}} \|\theta_t - \theta_0\|_\Theta \right) \\
&\leq \frac{C_{\nabla G}(B_\tau) B_\tau}{\sqrt{M}} \left(\kappa B_\tau + \frac{C_{\nabla G}(B_\tau)}{\sqrt{M}} B_\tau^2 + 1 \right),
\end{aligned}$$

where we used Proposition 24 for the second inequality.

Bounding $\|\hat{\xi}_t^{(2)}\|_\Theta$: By definition of $\hat{\xi}_t^{(2)}$ we have

$$\left\|\hat{\xi}_t^{(2)}\right\|_\Theta = \Bigg\| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} \nabla G_{\theta_0}(u_i)(x_j)(G_{\theta_t}(u_i)(x_j) - v_i(x_j)) \tag{33}$$

$$- \mathbb{E}_{\rho_x} \nabla G_{\theta_0}(u_i)(x)(G_{\theta_t}(u_i)(x) - v_i(x)) \Bigg\|_\Theta, \tag{34}$$

$$\leq \left\| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} \nabla G_{\theta_0}(u_i)(x_j) v_i(x_j) - \mathbb{E}_{\rho_x} \nabla G_{\theta_0}(u_i)(x) v_i(x) \right\|_\Theta + , \tag{35}$$

$$\left\| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} \nabla G_{\theta_0}(u_i)(x_j) G_{\theta_t}(u_i)(x_j) - \mathbb{E}_{\rho_x} \nabla G_{\theta_0}(u_i)(x) G_{\theta_t}(u_i)(x) \right\|_\Theta \tag{36}$$

$$:= I + II. \tag{37}$$

For I we set $W_j := \frac{1}{n_u} \sum_{i=1}^{n_u} \nabla G_{\theta_0}(u_i)(x_j) v_i(x_j)$. Note that $\|W_j\|_2 \leq \kappa^2 := B$, $\mathbb{E}\|W_j\|_2^2 \leq \kappa^4 := V^2$. Therefore we have from Proposition 41 with probability at least $1 - \delta$,

$$I \leq \frac{4\kappa^2}{\sqrt{n_x}} \log\left(\frac{4}{\delta}\right).$$

For II, we cannot directly apply Proposition 41, since the summands involve θ_t, which means they are no longer independent. However from Proposition 49 we have that with probability at least $1 - \delta$,

$$II \leq \frac{4\kappa^2}{\sqrt{n_x}} \log\left(\frac{4}{\delta}\right) \|\theta_t - \theta_0\|_\Theta + \frac{2\kappa C_{\nabla G}(B_\tau)}{\sqrt{M}} \|\theta_t - \theta_0\|_\Theta^2.$$

Therefore we have,

$$\left\|\hat{\xi}_t^{(2)}\right\|_\Theta \leq \frac{4\kappa^2}{\sqrt{n_x}} \log\left(\frac{4}{\delta}\right) (\|\theta_t - \theta_0\|_\Theta + 1) + \frac{2\kappa C_{\nabla G}(B_\tau)}{\sqrt{M}} \|\theta_t - \theta_0\|_\Theta^2 \tag{38}$$

$$\leq \frac{4\kappa^2 (B_\tau + 1)}{\sqrt{n_x}} \log\left(\frac{4}{\delta}\right) + \frac{2\kappa C_{\nabla G}(B_\tau) B_\tau^2}{\sqrt{M}}. \tag{39}$$

Bounding $\|\xi_t^{(3)}\|_{\mathcal{V}^{n_u}}$:
Using Proposition 25 we have

$$\left\|\xi_t^{(3)}\right\|_{\mathcal{V}^{n_u}} = \sqrt{\frac{1}{n_u} \sum_{i=1}^{n_u} \|r_{(\theta_t, \theta_0)}(u)\|_{L^2(\rho_x)}^2} \leq \frac{C_{\nabla G}(B_\tau) B_\tau^2}{\sqrt{M}}.$$

To sum up:

Plugging the bounds of the errors $\xi_t^{(i)}$, $i \in \{1,2,3\}$ into 32 leads to

$$\|\widehat{\Sigma}_M^a \hat{u}_{T+1}\|_{\mathcal{H}_M} \leq \alpha^{1-a}\left(\frac{2\eta_{a+\frac{1}{2}}(T)}{\sqrt{\alpha}} + \frac{2\eta_a(T)}{\sqrt{\alpha T}}\right) \max_{s\leq T}\left(\|\hat{\xi}_s^{(1)}\|_\Theta + \|\hat{\xi}_s^{(2)}\|_\Theta + \|\xi_s^{(3)}\|_{\mathcal{V}^{n_u}}\right) \quad (40)$$

$$\leq \alpha^{1-a}\left(\frac{2\eta_{a+\frac{1}{2}}(T)}{\sqrt{\alpha}} + \frac{2\eta_a(T)}{\sqrt{\alpha T}}\right) \cdot \xi(M, n_x, B_\tau, \kappa, \delta), \quad (41)$$

with

$$\xi(M, n_x, B_\tau, \kappa, \delta) := \left(\frac{C_{\nabla G}(B_\tau)B_\tau}{\sqrt{M}}\left(\kappa B_\tau + \frac{C_{\nabla G}(B_\tau)}{\sqrt{M}}B_\tau^2 + 1\right)\right.$$
$$\left. + \frac{4\kappa^2\,(B_\tau+1)}{\sqrt{n_x}}\log\left(\frac{4}{\delta}\right) + \frac{(2\kappa+1)C_{\nabla G}(B_\tau)B_\tau^2}{\sqrt{M}}\right).$$

Plugging (40) into (30) gives,

$$\|\mathcal{S}_M \hat{u}_T\|_{L^2(\mu_u)} \leq 2\,\|\widehat{\Sigma}_M^{\frac{1}{2}}\hat{u}_T\|_{\mathcal{H}_M} + 2\,\frac{\|\hat{u}_T\|_{\mathcal{H}_M}}{\sqrt{\alpha T}}$$
$$\leq 4\xi(M, n_x, B_\tau, \kappa, \delta)\left(\eta_1(T) + \frac{\eta_{1/2}(T)}{\sqrt{T}} + \frac{\eta_{1/2}(T)}{\sqrt{T}} + \frac{\eta_0(T)}{T}\right)$$
$$= 4\xi(M, n_x, B_\tau, \kappa, \delta)\,(\log(T) + 6)\,.$$

Note that, to derive this final bound we conditioned on four events, where each of the events holds true with probability at least $1-\delta$. By Proposition 34, we therefore obtain the last inequality holds with probability at least $1-4\delta$. Redefining $\delta = 4\delta$ proves the claim. ■

B.4 Bounding $\mathcal{III}$

For bounding the last term in our error decomposition we use the results obtained in Nguyen and Mücke (2023).

Proposition 19 (Theorem A.4. in Nguyen and Mücke (2023)) *Suppose Assumptions 1, 5, 4 and 6 are satisfied. Let $\delta \in (0,1]$, $\alpha < 1/\kappa^2$, $T \leq C_1 n_u^{\frac{1}{2r+b}}$, $2r+b > 1$ and $n_u \geq n_0 := e^{\frac{2r+b}{2r+b-1}}$. With probability at least $1-\delta$, we have*

$$\|\mathcal{S}_M F_T^M - G^*\|_{L^2(\mu_u)} \leq C_2\,\log^3(2/\delta)\,T^{-r}\,,$$

provided that

$$M \geq M_{III}(T,r) := C_3\log(n_u)\cdot\begin{cases} T & : r \in \left(0, \frac{1}{2}\right) \\ T^{1+b(2r-1)} & : r \in \left[\frac{1}{2}, 1\right] \\ T^{2r} & : r \in (1,\infty) \end{cases}, \quad (42)$$

where the constants $C_1, C_2, C_3 > 0$ do not depend on n_u, n_x, δ, M, T.

Remark 20 *In Theorem A.4. of Nguyen and Mücke (2023), different assumptions on n_u were proposed, however in the proof of Theorem 3.5 from Nguyen and Mücke (2023), it was shown that $T \leq C_1 n_u^{\frac{1}{2r+b}}$ is enough to guarantee the bound of 19.*

B.5 Proof of Main Result

Proof [Theorem 7] Following the error decomposition of (26) and using the bounds of Proposition 16, 18, 19 we obtain with probability at least $1-\delta$,

$$\|G_{\theta_T} - G^*\|_{L_2(\mu_u)} \leq \frac{C_{\nabla G}(B_\tau) B_\tau^2}{\sqrt{M}} + T^{-r} + C_2 \ \log^3(2/\delta)\ T^{-r} \tag{43}$$

$$\leq C \ \log^3(2/\delta)\ T^{-r}\ . \tag{44}$$

Collecting the restrictions on the number of neurons and samples from the above Propositions, we obtain the final lower bounds $T \leq C_1 n_u^{\frac{1}{2r+b}}$, $2r+b>1$ and $n_u \geq n_0 := e^{\frac{2r+b}{2r+b-1}}$, $n_x \geq 32\kappa^2\,(B_\tau+1)\,T^{2r}\,(\log(T)+6)\log\left(\frac{16}{\delta}\right)$.

$$M \geq M_0(T,\delta,\tilde{d},r,b,\alpha,B_\tau) := \max\left\{M_{II}, M_{III}\right\}, \tag{45}$$

with M_{II}, M_{III} defined in 18 and 42. ■

Appendix C. Bounds on the Weights

C.1 Proof of main result: The weights barely move

Proof [Theorem 9] We prove this by induction over $T \in \mathbb{N}$. Assume that the claim holds for $T \in \mathbb{N}$. We show that this implies the claim for $T+1$, for all $T \in \mathbb{N}$.

A short calculation, similar to (29), shows that the weights $(\theta_t)_t$ follow the recursion

$$\theta_{T+1} - \theta_0 = \alpha \sum_{t=0}^{T} (Id - \alpha \widehat{\mathcal{C}}_M)^t (\zeta_{T-t}^{(1)} + \zeta_{T-t}^{(2)} + \zeta_{T-t}^{(3)} + \zeta_{T-t}^{(4)} + \zeta_{T-t}^{(5)})\ ,$$

with

$$\zeta_s^{(1)} = \frac{1}{n_u n_x} \sum_{i=1}^{n_u}\sum_{j=1}^{n_x} \left(G_{\theta_s}(u_i)(x_j) - G^*(u_i)(x_j)\right)\left(\nabla G_{\theta_s}(u_i)(x_j) - \nabla G_{\theta_0}(u_i)(x_j)\right)\ ,$$

$$\zeta_s^{(2)} = \frac{1}{n_u n_x} \sum_{i=1}^{n_u}\sum_{j=1}^{n_x} (G^*(u_i)(x_j) - v_i(x_j))(\nabla G_{\theta_s}(u_i)(x_j) - \nabla G_{\theta_0}(u_i)(x_j))\ ,$$

$$\zeta_s^{(3)} = \widehat{\mathcal{Z}}_M^* \bar{r}_{(\theta_0,\theta_s)}, \qquad (\bar{r}_{(\theta_0,\theta_s)})_{i,j} := r_{(\theta_0,\theta_s)}(u_i)(x_j)$$

$$\zeta_s^{(4)} = \widehat{\mathcal{Z}}_M^* \mathbf{v} - \mathcal{Z}_M^* G^*, \qquad \mathbf{v}_{i,j} := v_i(x_j)$$

$$\zeta_s^{(5)} = \mathcal{Z}_M^* G^*\ .$$

Hence,

$$\begin{aligned}\|\theta_{T+1}-\theta_0\|_\Theta \le \alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\zeta_{T-t}^{(1)}\right\|_\Theta + \alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\zeta_{T-t}^{(2)}\right\|_\Theta \\ + \alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\zeta_{T-t}^{(3)}\right\|_\Theta + \alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\zeta_{T-t}^{(4)}\right\|_\Theta \\ + \alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\zeta_{T-t}^{(5)}\right\|_\Theta .\end{aligned}$$

Bounding $\alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\zeta_{T-t}^{(1)}\right\|_\Theta$.
First, note that

$$\begin{aligned}\alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\zeta_{T-t}^{(1)}\right\|_\Theta &\le \alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^{T-t}\right\|\cdot\|\zeta_t^{(1)}\|_\Theta \\ &\le \alpha\sum_{t=0}^{T}\|\zeta_t^{(1)}\|_\Theta .\end{aligned}$$

Note that we already bounded $\zeta_t^{(1)}$ in the proof of Theorem 18. However, to bound the weights, we need a more refined bound. To this end, let

$$T \le C_{\alpha,r,\kappa}n_u^{\frac{1}{2r+b}}, \quad n_x \ge T^{2\max\{0,\frac{1}{2}-r\}}, \quad M \ge M_{III},$$

where M_{III} is defined in Proposition 19. From Proposition 22, we obtain with probability of at least $1-\delta$,

$$\begin{aligned}&\alpha\sum_{t=0}^{T}\|\zeta_t^{(1)}\|_\Theta \\ &\le C_{\alpha,\kappa,r}\cdot C_{\sigma'}\cdot\left(\frac{TB_\tau^5+T^{1+\tilde\nu}B_\tau^3}{M} + \frac{B_\tau^2}{\sqrt{M}}\cdot\left(\eta_r(T)+\left(\frac{T\log(T)}{\sqrt{n_u}}+\frac{T\log(T)}{\sqrt{n_x}}\right)\log^{\frac{3}{2}}\left(\frac{12}{\delta}\right)\right)\right) \\ &\le 2\,C_{\alpha,\kappa,r}\cdot C_{\sigma'}\cdot\left(\frac{TB_\tau^5}{M} + \frac{B_\tau^2}{\sqrt{M}}\cdot\left(\eta_r(T)+\left(\frac{T\log(T)}{\sqrt{n_u}}+\frac{T\log(T)}{\sqrt{n_x}}\right)\log^{\frac{3}{2}}\left(\frac{12}{\delta}\right)\right)\right),\end{aligned}$$

for some $C_{\alpha,\kappa,r}>0$ and where we set $\tilde\nu=\max\{0,\frac{1}{2}-r\}$, $C_{\sigma'}:=\max\{\|\sigma'\|_\infty,\|\sigma'\|_\infty^2,\|\sigma''\|_\infty,\|\sigma''\|_\infty^2\}$. For the last inequality we used $T^{\tilde\nu}\le B_\tau$ by definition of B_τ and the quantity $\eta_r(T)$ is defined in Lemma 31.

If we let

$$n_u \ge C_{\alpha,r,\kappa,b}T^{2r+b}, \tag{46}$$
$$n_x \ge T^{2\max\{0,\frac{1}{2}-r\}}, \tag{47}$$
$$M \ge \max\{M_{III}, M_{\alpha,\sigma',r}\}, \tag{48}$$
$$M\cdot n_u \ge \tilde C_{\alpha,\kappa,r}\cdot\tilde C_{\sigma'}\cdot B_\tau^2\,T^2\,\log^2(T)\,\log^3(12/\delta), \tag{49}$$
$$M\cdot n_x \ge \tilde C_{\alpha,\kappa,r}\cdot\tilde C_{\sigma'}\cdot B_\tau^2\,T^2\,\log^2(T)\,\log^3(12/\delta), \tag{50}$$

for some $\tilde{C}_{\alpha,\kappa,r} > 0$, and where

$C_{\sigma'} := \max\{\|\sigma'\|_\infty, \|\sigma'\|_\infty^2, \|\sigma''\|_\infty, \|\sigma''\|_\infty^2\}, \quad M_{\alpha,\sigma',r} := \tilde{C}_{\alpha,\kappa,r} \cdot C_{\sigma'} \cdot \max\{TB_\tau^4, \eta_r(T)^2 B_\tau\},$

we obtain

$$\alpha \sum_{t=0}^{T} \|\zeta_t^{(1)}\|_\Theta \leq B_\tau/5\,.$$

Bounding $\alpha \sum_{t=0}^{T} \left\|(Id - \alpha\widehat{\mathcal{C}}_M)^t \zeta_{T-t}^{(2)}\right\|_\Theta$.

From Lemma 23, we obtain with probability at least $1-\delta$,

$$\max_{t\in[T]} \|\zeta_t^{(2)}\|_\Theta \leq \frac{\tilde{d}\, B_\tau^2 \cdot C_{\sigma',\sigma''}}{\sqrt{n_u \cdot M}} \sqrt{\log(4\tilde{d}^2/\delta)},$$

with $C_{\sigma',\sigma''} = 334 \cdot \max\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\} \cdot \max\{L_{\sigma'}, L_{\sigma''}\}$. Hence,

$$\begin{aligned}
\alpha \sum_{t=0}^{T} \left\|(Id - \alpha\widehat{\mathcal{C}}_M)^t \zeta_{T-t}^{(2)}\right\|_\Theta &\leq \alpha \sum_{t=0}^{T} \left\|(Id - \alpha\widehat{\mathcal{C}}_M)^{T-t}\right\| \cdot \|\zeta_t^{(2)}\|_\Theta \\
&\leq \frac{\alpha T\, \tilde{d}\, B_\tau^2 \cdot C_{\sigma',\sigma''}}{\sqrt{n_u \cdot M}} \sqrt{\log(4\tilde{d}^2/\delta)} \\
&\leq \frac{1}{5} B_\tau, \qquad (51)
\end{aligned}$$

provided that

$$M \cdot n_u \geq 25\alpha^2\, B_\tau^2\, \tilde{d}^2\, C_{\sigma',\sigma''}^2\, T^2\, \log(4\tilde{d}^2/\delta). \qquad (52)$$

Bounding $\alpha \sum_{t=0}^{T} \left\|(Id - \alpha\widehat{\mathcal{C}}_M)^t \zeta_{T-t}^{(3)}\right\|_\Theta$.

Recall that with probability at least $1-\delta/5$, for all $t \in [T]$,

$$\|\theta_t - \theta_0\|_\Theta \leq B_\tau\ .$$

Since $C_{\nabla G}(B_\tau) \leq 4\max\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\} B_\tau$, we obtain from Proposition 33 and Proposition 25,

$$\begin{aligned}
\alpha \sum_{t=0}^{T} \left\|(Id - \alpha\widehat{\mathcal{C}}_M)^t \zeta_{T-t}^{(3)}\right\|_\Theta &= \alpha \sum_{t=0}^{T} \left\|(Id - \alpha\widehat{\mathcal{C}}_M)^{T-t} \zeta_t^{(3)}\right\|_\Theta \\
&\leq \frac{\sqrt{\alpha}}{\sqrt{n_u n_x}} \sum_{t=0}^{T} \left\|(Id - \alpha\widehat{\mathcal{C}}_M)^{T-t} (\sqrt{\alpha}\widehat{\mathcal{Z}}_M^*)\right\| \cdot \|\bar{r}_{(\theta_0,\theta_s)}\|_F \\
&\leq \frac{2\sqrt{\alpha T}}{\sqrt{n_u n_x}} \cdot \max_{s\in[T]} \|\bar{r}_{(\theta_0,\theta_s)}\|_F \\
&\leq 8 \cdot \max\left\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\right\} \cdot B_\tau^3 \cdot \sqrt{\frac{\alpha T}{M}} \\
&\leq \frac{1}{5}\, B_\tau\ , \qquad (53)
\end{aligned}$$

provided that $M \geq M(\zeta^{(3)})$, with

$$M(\zeta^{(3)}) = 1600 \cdot \alpha \cdot \max\left\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\right\} \cdot B_\tau^4 T \,. \tag{54}$$

Bounding $\alpha \sum_{t=0}^T \left\|(Id - \alpha\widehat{\mathcal{C}}_M)^t \zeta_{T-t}^{(4)}\right\|_\Theta$. We write

$$\begin{aligned}
\alpha \sum_{t=0}^T \left\|(Id - \alpha\widehat{\mathcal{C}}_M)^t \zeta_{T-t}^{(4)}\right\|_\Theta &= \alpha \sum_{t=0}^T \left\|(Id - \alpha\widehat{\mathcal{C}}_M)^t (\widehat{\mathcal{Z}}_M^* \mathbf{v} - \mathcal{Z}_M^* G^*)\right\|_\Theta \\
&\leq \alpha \sum_{t=0}^T \left\|(Id - \alpha\widehat{\mathcal{C}}_M)^t \widehat{\mathcal{C}}_{M,\lambda}^{1/2}\right\| \cdot \|\widehat{\mathcal{C}}_{M,\lambda}^{-1/2} \mathcal{C}_{M,\lambda}^{1/2}\| \\
&\quad \cdot \left\|\mathcal{C}_{M,\lambda}^{-1/2}\left(\widehat{\mathcal{Z}}_M^* y - \mathcal{Z}_M^* G^*\right)\right\|_\Theta \,.
\end{aligned} \tag{55}$$

For the first term, we obtain from (94),

$$\begin{aligned}
\alpha \sum_{t=0}^T \left\|(Id - \alpha\widehat{\mathcal{C}}_M)^t \widehat{\mathcal{C}}_{M,\lambda}^{1/2}\right\| &\leq \sqrt{\alpha} \sum_{t=0}^T \left\|(Id - \alpha\widehat{\mathcal{C}}_M)^t (\alpha\widehat{\mathcal{C}}_M)^{1/2}\right\| + \alpha\sqrt{\lambda} \sum_{t=0}^T \left\|(Id - \alpha\widehat{\mathcal{C}}_M)^t\right\| \\
&\leq 2\sqrt{\alpha T} + \sqrt{\lambda} \cdot (\alpha T) \\
&= 4\sqrt{\alpha T} \,,
\end{aligned} \tag{56}$$

if we take $\lambda = 1/(\alpha T)$.

For the second term, we obtain from Proposition 45, with probability at least $1 - \delta$ with $\lambda = 1/(\alpha T)$,

$$\|\widehat{\mathcal{C}}_{M,\lambda}^{-1/2} \mathcal{C}_{M,\lambda}^{1/2}\| \leq 2 \,, \tag{57}$$

if

$$M \geq C_\kappa \log^2(1/\delta), \quad n_u \geq C_\kappa T \log(T) \log^2(1/\delta) \,,$$
$$n_x \geq C_\kappa T \log(T) \log^2(1/\delta) \,,$$

for some $C_\kappa > 0$.

For the third term, we apply Proposition 48 with $\lambda = 1/(\alpha T)$ with probability of at least $1 - \delta$,

$$\begin{aligned}
&\left\|\mathcal{C}_{M,\lambda}^{-1/2}\left(\widehat{\mathcal{Z}}_M^* \mathbf{v} - \mathcal{Z}_M^* G^*\right)\right\| \\
&\quad \leq 2\left(\kappa\sqrt{\alpha T}\left(\frac{1}{n_u} + \frac{1}{n_x}\right) + 5\sqrt{\mathcal{N}_{\mathcal{L}_\infty}((\alpha T)^{-1})}\left(\frac{1}{\sqrt{n_u}} + \frac{1}{\sqrt{n_x}}\right)\right) \log^{3/2}\left(\frac{12}{\delta}\right),
\end{aligned}$$

if

$$M \geq C_{\alpha,\kappa} \tilde{d} T \log(T) \log(1/\delta),$$

and

$$n_u \geq C_{\alpha,\kappa} T \log^2(12/\delta) \,, \quad n_x \geq C_{\alpha,\kappa} T \log^2(12/\delta) \,,$$

for some $C_{\alpha,\kappa} > 0$.

The above choices for n_u and n_x imply

$$2\kappa\sqrt{\alpha T}\left(\frac{1}{n_u}+\frac{1}{n_x}\right)\log^{3/2}\left(\frac{12}{\delta}\right) \leq \frac{4\alpha\kappa}{C_{\alpha,\kappa}\sqrt{\alpha T}}.$$

Hence,

$$\begin{aligned}&\left\|\mathcal{C}_{M,\lambda}^{-1/2}\left(\widehat{\mathcal{Z}}_M^*\mathbf{v}-\mathcal{Z}_M^*G^*\right)\right\| \\ &\quad\leq \frac{4\alpha\kappa}{C_{\alpha,\kappa}\sqrt{\alpha T}} + 10\sqrt{\mathcal{N}_{\mathcal{L}_\infty}((\alpha T)^{-1})}\left(\frac{1}{\sqrt{n_u}}+\frac{1}{\sqrt{n_x}}\right)\cdot\log^{3/2}\left(\frac{12}{\delta}\right). \end{aligned} \tag{58}$$

Combining (56), (57), and (58) with (55) and recalling the definitions of B_τ and $\mathcal{B}_\delta\left(\frac{1}{\alpha T}\right)$ from Theorem 9, we find with probability of at least $1-3\delta$

$$\begin{aligned}&\alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\zeta_{T-t}^{(4)}\right\|_\Theta \\ &\leq 8\sqrt{\alpha T}\left(\frac{4\alpha\kappa}{C_{\alpha,\kappa}\sqrt{\alpha T}} + 10\sqrt{\mathcal{N}_{\mathcal{L}_\infty}((\alpha T)^{-1})}\left(\frac{1}{\sqrt{n_u}}+\frac{1}{\sqrt{n_x}}\right)\log^{3/2}\left(\frac{12}{\delta}\right)\right) \\ &\leq \frac{256\alpha\max\{1,\kappa\}}{C_{\alpha,\kappa}}\cdot\mathcal{B}_\delta\left(\frac{1}{\alpha T}\right) \\ &\leq \frac{256\alpha\max\{1,\kappa\}}{C_{\alpha,\kappa}}\cdot\log(T)\cdot T^{\max\{0,\frac{1}{2}-r\}}\cdot\mathcal{B}_\delta\left(\frac{1}{\alpha T}\right) \\ &\leq B_\tau/5\,, \end{aligned} \tag{59}$$

where the constant in the definition of B_τ is set to

$$C_\diamond := \max\left\{C_{R,r,\kappa,\alpha}, \frac{1280\alpha\max\{1,\kappa\}}{C_{\alpha,\kappa}}, 1\right\}, \tag{60}$$

with $C_{R,r,\kappa,\alpha} > 0$ defined in (68).

For the bound (59) to hold, we need to assume

$$M \geq M(\zeta^{(4)}) := C_{\alpha,\kappa}\,\tilde{d}\,T\log(T)\,\log^2(1/\delta), \tag{61}$$

$$n_u \geq n_u(\zeta^{(4)}) := C_{\alpha,\kappa}\,T\log(T)\,\log^2(12/\delta)\,, \tag{62}$$

$$n_x \geq n_x(\zeta^{(4)}) := C_{\alpha,\kappa}\,T\log(T)\,\log^2(12/\delta)\,, \tag{63}$$

for some for some $C_{\alpha,\kappa} > 0$.

Bounding $\alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\zeta_{T-t}^{(5)}\right\|_\Theta$.

We have

$$\begin{aligned}\alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\zeta^{(5)}_{T-t}\right\|_\Theta &= \alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\mathcal{Z}_M^*G^*\right\|_\Theta \\ &\le \alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\widehat{\mathcal{C}}_{M,\lambda}\right\|\cdot\left\|\widehat{\mathcal{C}}^{-1}_{M,\lambda}\mathcal{C}_{M,\lambda}\right\| \\ &\quad\cdot\left\|\mathcal{C}^{-1}_{M,\lambda}\mathcal{Z}_M^*\mathcal{L}_{M,\lambda}^{\frac{1}{2}}\right\|\cdot\left\|\mathcal{L}_{M,\lambda}^{-\frac{1}{2}}\mathcal{L}_{\infty,\lambda}^{\frac{1}{2}}\right\|\cdot\left\|\mathcal{L}_{\infty,\lambda}^{-\frac{1}{2}}G^*\right\|_{L^2(\mu_u)}\,.\end{aligned}$$

For the first term we obtain

$$\begin{aligned}\alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\widehat{\mathcal{C}}_{M,\lambda}\right\|_\Theta &\le \sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t(\alpha\widehat{\mathcal{C}}_M)\right\|_\Theta+\alpha\lambda\cdot\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\right\|_\Theta \\ &\le 2\log(T)+\lambda(\alpha T) \\ &\le 3\log(T)\,,\end{aligned}$$

if we again let $\lambda=1/(\alpha T)$ and since $T\ge 3$ by assumption.

By Proposition 47, with probability at least $1-4\delta$

$$\begin{aligned}&\left\|\widehat{\mathcal{C}}^{-1}_{M,\lambda}\mathcal{C}_{M,\lambda}\right\| \\ &\quad\le 2\left(\kappa^2\left(\frac{\alpha T}{n_u}+\frac{\alpha T}{n_x}\right)+5\kappa\sqrt{\mathcal{N}_{\mathcal{L}_\infty}((\alpha T)^{-1})}\left(\sqrt{\frac{\alpha T}{n_u}}+\sqrt{\frac{\alpha T}{n_x}}\right)\right)\log^{3/2}\left(\frac{12}{\delta}\right).\end{aligned}$$

if

$$M\ge M(\zeta^{(4)}), \tag{64}$$

$$n_u\ge\max\{4\alpha\kappa^2,C_{\alpha,\kappa}\}T\log^2(12/\delta), \tag{65}$$

$$n_x\ge\max\{4\alpha\kappa^2,C_{\alpha,\kappa}\}T\log^2(12/\delta), \tag{66}$$

where $M(\zeta^{(4)})$ is defined in (61) and where $C_{\alpha,\kappa}>0$ comes from Proposition 47. The given choices of n_u and n_x imply

$$\left(\frac{\alpha T}{n_u}+\frac{\alpha T}{n_x}\right)\cdot\log^{3/2}\left(\frac{12}{\delta}\right)\le\frac{1}{2\kappa^2}\,.$$

Hence,

$$\left\|\widehat{\mathcal{C}}^{-1}_{M,\lambda}\mathcal{C}_{M,\lambda}\right\|\le\mathcal{B}_\delta\left(\frac{1}{\alpha T}\right),$$

where $\mathcal{B}_\delta\left(\frac{1}{\alpha T}\right)$ is defined in Theorem 9.

From Proposition 35 we obtain almost surely

$$\|\mathcal{C}^{-1}_{M,\lambda}\mathcal{Z}_M^*\mathcal{L}_{M,\lambda}^{\frac{1}{2}}\|\le 2\,.$$

Furthermore, (Nguyen and Mücke, 2023, Proposition A.14) gives with Probability of at least $1-\delta$

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

By Assumption 5 and Proposition 37,

$$\left\|\mathcal{L}_{\infty,\lambda}^{-\frac{1}{2}}G^*\right\|_{L^2(\mu_u)} \leq R\kappa^r(\alpha T)^{\max\{0,\frac{1}{2}-r\}}.$$

Combining the previous bounds gives for $\lambda = 1/(\alpha T)$ with probability at least $1-5\delta$

$$\begin{aligned}\alpha\sum_{t=0}^{T}\left\|(Id-\alpha\widehat{\mathcal{C}}_M)^t\zeta_{T-t}^{(5)}\right\|_\Theta &\leq 12R\kappa^r\log(T)(\alpha T)^{\max\{0,\frac{1}{2}-r\}}\cdot\mathcal{B}_\delta\left(\frac{1}{\alpha T}\right)\\ &= \frac{1}{5}C_{R,r,\kappa,\alpha}\log(T)T^{\max\{0,\frac{1}{2}-r\}}\cdot\mathcal{B}_\delta\left(\frac{1}{\alpha T}\right)\\ &\leq \frac{B_\tau}{5},\end{aligned} \tag{67}$$

where we set

$$C_{R,r,\kappa,\alpha} := 60R\kappa^r\alpha^{\max\{0,\frac{1}{2}-r\}}. \tag{68}$$

Collecting everything. Finally, combining the bounds of all error terms and plugging in $\lambda = 1/(\alpha T)$ gives with probability at least $1-10\delta$, for all $t\in[T+1]$

$$\|\theta_t-\theta_0\|_\Theta \leq\ B_\tau,$$

provided that

$$M\cdot n_u \geq \tilde{C}_{\alpha,\kappa,r}B_\tau^2\,T^2\cdot\max\{\tilde{C}_{\sigma'}\ \log^2(T)\ \log^3(12/\delta), \tilde{d}^2C_{\sigma',\sigma''}^2\log(4\tilde{d}^2/\delta)\}, \tag{69}$$

$$M\cdot n_x \geq \tilde{C}_{\alpha,\kappa,r}\cdot\tilde{C}_{\sigma'}\cdot B_\tau^2\,T^2\ \log^2(T)\ \log^3(12/\delta), \tag{70}$$

with

$$\begin{aligned}\tilde{C}_{\sigma'} &:= \max\{C_{\sigma'}, C_{\sigma'}^2\},\\ C_{\sigma'} &:= \max\{\|\sigma'\|_\infty, \|\sigma'\|_\infty^2, \|\sigma''\|_\infty, \|\sigma''\|_\infty^2\},\\ C_{\sigma',\sigma''} &:= 334\cdot\max\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\}\cdot\max\{L_{\sigma'}, L_{\sigma''}\},\end{aligned}$$

and

$$n_u \geq \max\left\{\tilde{C}_{\alpha,\kappa}\,T\log(T)\ \log^2\left(12/\delta\right),\ C_{\alpha,r,\kappa,b}T^{2r+b}\right\}, \tag{71}$$

$$n_x \geq \tilde{C}_{\alpha,\kappa}\,T\log(T)\ \log^2\left(12/\delta\right). \tag{72}$$

Furthermore, we need

$$M \geq \max\{M_{III}, M_{\alpha,\sigma',r}, M(\zeta^{(3)}), M(\zeta^{(4)})\}, \tag{73}$$

where M_{III} is defined in Proposition 19 and

$$\begin{aligned} M_{\alpha,\sigma',r} &:= \tilde{C}_{\alpha,\kappa,r} \cdot \tilde{C}_{\sigma'} \cdot \max\{T B_\tau^4, \eta_r(T)^2 B_\tau\}, \\ M(\zeta^{(3)}) &= 1600 \cdot \alpha \cdot \max\left\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\right\} \cdot T\, B_\tau^4, \\ M(\zeta^{(4)}) &:= C_{\alpha,\kappa}\, \tilde{d}\, T \log(T)\ \log^2(1/\delta). \end{aligned}$$

■

Proof [Corollary 10]

Recall that we assume

$$T = c n_u^{\frac{1}{2r+b}}.$$

We first verify that the assumptions of Theorem 9 are satisfied for this choice of T. According to the constraints in (69) and (71), it is required that

$$n_u \geq C_{\alpha,\kappa,\|\mathcal{L}_\infty\|,r,b,\sigma',\sigma''} \cdot \max\left\{T^{2r+b},\, T\log(T)\log^2(1/\delta)\right\},$$

for some constant $C_{\alpha,\kappa,\|\mathcal{L}_\infty\|,r,b,\sigma',\sigma''} > 0$.

It is therefore sufficient that

$$n_u \geq C_{\alpha,\kappa,\|\mathcal{L}_\infty\|,r,b,\sigma',\sigma''} T^{2r+b},$$

provided that $n_u \geq n_1(\delta)$ is large enough so that

$$T_{n_1} \log(T_{n_1}) \log^2\left(\frac{1}{\delta}\right) \leq T_{n_1}^{2r+b}.$$

Hence, if

$$c \leq C^{-1}_{\alpha,\kappa,\|\mathcal{L}_\infty\|,r,b,\sigma',\sigma''}$$

our choice of T satisfies the required assumption on n_u. Now we can prove the statement.

1. Let $r > \frac{1}{2}$. A short calculation shows that

$$\begin{aligned} \mathcal{B}_\delta\left(\frac{1}{\alpha T}\right) &= 1 + 10\kappa\sqrt{\mathcal{N}_{\mathcal{L}_\infty}((\alpha T)^{-1})}\left(\sqrt{\frac{\alpha T}{n_u}} + \sqrt{\frac{\alpha T}{n_x}}\right)\log^{3/2}\left(\frac{12}{\delta}\right) \\ &\leq 1 + C_{\alpha,b,\kappa}\left(\sqrt{\frac{T^{1+b}}{n_u}} + \sqrt{\frac{T^{1+b}}{n_x}}\right)\log^{3/2}\left(\frac{12}{\delta}\right) \\ &\leq 1 + C_{\alpha,b,\kappa}\left(n_u^{\frac{1-2r}{4r+2b}} + \sqrt{\frac{T^{1+b}}{n_x}}\right)\log^{3/2}\left(\frac{12}{\delta}\right) \leq 2 \end{aligned}$$

if we let $c \leq 1$ and

$$\begin{aligned} n_u &\geq n_2 := \left(2C_{\alpha,b,\kappa}\log^{3/2}\left(\frac{12}{\delta}\right)\right)^{\frac{4r+2b}{2r-1}}, \\ n_x &\geq \left(2C_{\alpha,b,\kappa}\log^{3/2}\left(\frac{12}{\delta}\right)\right)^2 T^{1+b}. \end{aligned}$$

Hence, $B_\tau(\delta, T) \leq C_{R,r,\kappa,\alpha} \log(T)$.

For $r = \frac{1}{2}$ we similarly obtain $\mathcal{B}_\delta(T) \leq 2$ if we let $c \leq C_{\alpha,\kappa,b,\delta}$ be small enough.

We finally bound the number of neurons that are required for achieving this rate from Theorem 9. According to the constraints in (69) and (71), it is required that

$$M \geq C_{\kappa,\alpha,r,\sigma',\sigma'',\|\mathcal{L}_\infty\|} \cdot \max\left\{ \log(T) T^{2r},\, B_\tau^4 T,\, B_\tau^2\, T\, \tilde{d}^2\, \log(\tilde{d}/\delta),\, \tilde{d} T \log(T) \log^2(1/\delta) \right\}.$$

Bounding the maximum gives that it is enough if

$$M \geq C_{R,\kappa,\alpha,r,\sigma',\sigma'',\|\mathcal{L}_\infty\|}\, \tilde{d}^2 \log^4(T) T^{2r} \log^2(\tilde{d}/\delta)\ ,$$

for some $C_{R,\kappa,\alpha,r,\sigma',\sigma'',\|\mathcal{L}_\infty\|} > 0$. Similar we obtain for number of second stage samples that are required,

$$n_x \geq C_{\alpha,b,\kappa,\|\mathcal{L}_\infty\|} \cdot \max\left\{ T \log(T)^2 \log^2(1/\delta),\ T^{1+b} \log^3(1/\delta) \right\}$$

Bounding the maximum gives that it is enough if

$$n_x \geq C_{\alpha,b,\kappa,\|\mathcal{L}_\infty\|} T^{1+b} \log^3(1/\delta),$$

if $n_u \geq n_3(\delta)$ is large enough so that

$$\log^2(T_{n_3}) T_{n_3} \leq T_{n_3}^{1+b}$$

holds true.

2. If $r < \frac{1}{2}$, then similar as before we have

$$\mathcal{B}_\delta(\alpha T)\ \leq C_{\alpha,b,\kappa} T^{\frac{1}{2}-r}$$

for some $C_{\alpha,b,\kappa} > 0$, if $c \leq \log^{-3}(1/\delta)$ and $c^{-1} T^{2r+b} \leq n_x$. Hence, $B_\tau(T) \leq C_{R,r,\kappa,\alpha,b} \log(T) T^{1-2r}$. A lower bound for the number of neurons follows the same way as in the other case:

$$\begin{aligned} M &\geq C_{\kappa,\alpha,r,\sigma,\|\mathcal{L}_\infty\|}\, \tilde{d}^2 \log^4(T) T^{5-8r} \log^2(\tilde{d}/\delta), \\ n_x &\geq C_{\alpha,b,\kappa,\|\mathcal{L}_\infty\|} T^{2r+b} \log^3(1/\delta). \end{aligned}$$

To sum up the statement holds true if we set

$$\begin{aligned} c &:= \min\left\{1,\, C^{-1}_{\alpha,\kappa,\|\mathcal{L}_\infty\|,r,b,\sigma',\sigma''},\, C_{\alpha,\kappa,b,\delta}\right\}, \\ n &\geq n_0 := \max\{n_1, n_2, n_3\}. \end{aligned}$$

■

C.2 Error bounds for the weights

In this section, we derive some error bounds that are necessary to prove our main results about the weights in Section C.1.

Lemma 21 (Nguyen and Mücke (2023)) *Let*

$$F_t^* := \mathcal{S}_M^* \phi_t(\mathcal{L}_M) G^* \in \mathcal{H}_M \,, \tag{74}$$

where ϕ_t denotes the spectral regularization function associated to gradient descent , see e.g. Blanchard and Mücke (2017). Then

$$\|F_t^*\|_{\mathcal{H}_\mathcal{M}} \leq C'_{\kappa,\alpha}\, t^{\max\{0,\frac{1}{2}-r\}} \,,$$

for some $C'_{\kappa,\alpha} < \infty$.

Proposition 22 *For all $t \in [T]$, let $\|\theta_t - \theta_0\| \leq B_\tau$, and*

$$T \leq C_{\alpha,r,\kappa} n_u^{\frac{1}{2r+b}}, \quad n_x \geq T^{2\max\{0,\frac{1}{2}-r\}}, \quad M \geq M_{III},$$

where M_{III} is defined in Proposition 19. Set $\tilde{\nu} = \max\{0, \frac{1}{2} - r\}$. We have with probability of at least $1-\delta$,

$$\alpha \sum_{t=0}^{T} \|\zeta_t^{(1)}\|_\Theta \leq$$
$$C_{\alpha,\kappa,r} \cdot C_{\sigma'} \cdot \left(\frac{T B_\tau^5 + T^{1+\tilde{\nu}} B_\tau^3}{M} + \frac{B_\tau^2}{\sqrt{M}} \cdot \left(\eta_r(T) + \left(\frac{T\log(T)}{\sqrt{n_u}} + \frac{T\log(T)}{\sqrt{n_x}} \right) \log^{\frac{3}{2}}\left(\frac{12}{\delta}\right) \right) \right) ,$$

for some $C_{\alpha,\kappa,r} > 0$ and where we set $C_{\sigma'} := \max\{\|\sigma'\|_\infty, \|\sigma'\|_\infty^2, \|\sigma''\|_\infty, \|\sigma''\|_\infty^2\}$. The quantity $\eta_r(T)$ is defined in Lemma 31.

Proof [Proposition 22] By using once again a Taylor expansion we obtain for $t \in [T]$,

$$\begin{aligned}
\left\|\zeta_t^{(1)}\right\|_\Theta &= \left\| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} \left(G_{\theta_t}(u_i)(x_j) - G^*(u_i)(x_j)\right) \left(\nabla G_{\theta_t}(u_i)(x_j) - \nabla G_{\theta_0}(u_i)(x_j)\right) \right\|_\Theta \\
&\leq \left\| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} r_{(\theta_t,\theta_0)}(u_i)(x_j) \left(\nabla G_{\theta_t}(u_i)(x_j) - \nabla G_{\theta_0}(u_i)(x_j)\right) \right\|_\Theta + \\
&\quad \left\| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} \left(H_t(u_i)(x_j) - F_t^*(u_i)(x_j)\right) \left(\nabla G_{\theta_t}(u_i)(x_j) - \nabla G_{\theta_0}(u_i)(x_j)\right) \right\|_\Theta + \\
&\quad \left\| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} \left(F_t^*(u_i)(x_j) - G^*(u_i)(x_j)\right) \left(\nabla G_{\theta_t}(u_i)(x_j) - \nabla G_{\theta_0}(u_i)(x_j)\right) \right\|_\Theta \\
&=: I_t + II_t + III_t \,,
\end{aligned} \tag{75}$$

where F_t^* was defined in (74).

Bound I_t : By using Proposition 24 and Proposition 25, we obtain for $t \in [T]$

$$I_t \leq \frac{C_{\nabla G}(B_\tau)^2}{M}\|\theta_t - \theta_0\|_\Theta^3 \leq \frac{C_{\nabla G}(B_\tau)^2 B_\tau^3}{M} \leq \frac{16 \cdot \max\left\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\right\}^2 B_\tau^5}{M}.$$

Here, we use that

$$C_{\nabla G}(B_\tau) = 2 \cdot \max\left\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\right\} \max\{1, B_\tau + \tau\} \leq 4 \cdot \max\left\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\right\} B_\tau,$$

following from Assumption 15.

Bound II_t : Note that $F_t^* \in \mathcal{H}_M$, therefore there exist some θ_t^* such that $F_t^*(u)(x) = \langle \nabla G_{\theta_0}(u)(x), \theta_t^* \rangle$ and $\|F_t^*\|_{\mathcal{H}_M} = \|\theta_t^*\|_\Theta$ (see for example Steinwart and Christmann (2008) Theorem 4.21). Using these equations we obtain

$$II_t \leq \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} \Delta(u_i, x_j) \cdot \left(\|\theta_t - \theta_0\|_\Theta + \|F_t^*\|_{\mathcal{H}_M}\right), \tag{76}$$

with

$$\Delta(u_i, x_j) := \left\| \left(\nabla G_{\theta_t}(u_i)(x_j) - \nabla G_{\theta_0}(u_i)(x_j)\right) \left(\nabla G_{\theta_0}(u_i)(x_j)\right)^T \right\|.$$

By definition of the operator norm and Proposition 24, we have

$$\begin{aligned}
\Delta(u_i, x_j) &= \max_k \left\| \left(\nabla G_{\theta_t}(u_i)(x_j) - \nabla G_{\theta_0}(u_i)(x_j)\right) \left(\nabla G_{\theta_0}(u_i)(x_j)\right)^T e_k \right\|_2 \\
&\leq \|\nabla G_{\theta_t}(u_i)(x_j) - \nabla G_{\theta_0}(u_i)(x_j)\|_2 \ \frac{\kappa}{\sqrt{M}} \\
&\leq \frac{C_{\nabla G}(B_\tau)\kappa}{M} \ \|\theta_t - \theta_0\|_\Theta,
\end{aligned} \tag{77}$$

where e_k denotes the k-th unit vector in $\mathbb{R}^P$. Plugging (77) into (76) and applying Lemma 21 leads to

$$\begin{aligned}
II_t &\leq \frac{C_{\nabla G}(B_\tau)\kappa}{M}\|\theta_t - \theta_0\|_\Theta \left(\|\theta_t - \theta_0\|_\Theta + \|F_t^*\|_{\mathcal{H}_M}\right) \\
&\leq \frac{C_{\nabla G}(B_\tau) B_\tau \kappa}{M} \left(B_\tau + C'_{\kappa,\alpha}\, t^{\max\{0, \frac{1}{2} - r\}}\right) \\
&\leq \frac{4\kappa \ \max\left\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\right\} B_\tau^2}{M} \ \left(B_\tau + C'_{\kappa,\alpha}\, t^{\max\{0, \frac{1}{2} - r\}}\right).
\end{aligned} \tag{78}$$

Bound III_t : By Proposition 24 we have

$$\begin{aligned}
III_t &= \left\| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} \left(F_t^*(u_i)(x_j) - G^*(u_i)(x_j)\right) \left(\nabla G_{\theta_t}(u_i)(x_j) - \nabla G_{\theta_0}(u_i)(x_j)\right) \right\|_\Theta \\
&\leq \frac{C_{\nabla G}(B_\tau) B_\tau}{\sqrt{M}} \cdot \left(III_t^{(i)} + III_t^{(ii)}\right) \\
&\leq \frac{4 \max\left\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\right\} B_\tau^2}{\sqrt{M}} \cdot \left(III_t^{(i)} + III_t^{(ii)}\right),
\end{aligned} \tag{79}$$

where

$$III_t^{(i)} = \left| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} |F_t^*(u_i)(x_j) - G^*(u_i)(x_j)| - \mathbb{E}\,|F_t^*(u)(x) - G^*(u)(x)| \right| ,$$
$$III_t^{(ii)} = \mathbb{E}\,|F_t^*(u)(x) - G^*(u)(x)| \ .$$

To sum up: If we plug the bounds of I_t, II_t, III_t into (75) we get for some $C_{\alpha,\kappa} > 0$ and by setting $\tilde{\nu} = \max\{0, \frac{1}{2} - r\}$,

$$\alpha \sum_{t=0}^{T} \|\zeta_t^{(1)}\|_\Theta \leq$$
$$C_{\alpha,\kappa} \max\{\|\sigma'\|_\infty, \|\sigma'\|_\infty^2, \|\sigma''\|_\infty, \|\sigma''\|_\infty^2\} \cdot \left(\frac{T B_\tau^5 + T^{1+\tilde{\nu}} B_\tau^3}{M} \ + \ \frac{B_\tau^2}{\sqrt{M}} \cdot \left(\sum_{t=0}^{T} III_t^{(i)} + III_t^{(ii)} \right) \right) . \tag{80}$$

Assume mow that

$$T \leq C_{\alpha,r,\kappa} n_u^{\frac{1}{2r+b}}, \quad n_x \geq t^{2\max\{0,\frac{1}{2}-r\}}, \quad M \geq M_{III},$$

where M_{III} is defined in Proposition 19. Applying Proposition 50 gives then for all $t \in [T]$ with probability of at least $1 - \delta$,

$$\begin{aligned} III_t^{(i)} &= \left| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} |F_t^*(u_i)(x_j) - G^*(u_i)(x_j)| - \mathbb{E}\,|F_t^*(u)(x) - G^*(u)(x)| \right| \\ &\leq \tilde{C}_{\alpha,\kappa} \left(\frac{1}{\sqrt{n_u}} + \frac{1}{\sqrt{n_x}} \right) \log^{3/2}\left(\frac{12}{\delta} \right), \end{aligned} \tag{81}$$

for some $\tilde{C}_{\alpha,\kappa} > 0$. Combining this bound with Proposition 34, we obtain,

$$\sum_{t=0}^{T} III_t^{(i)} \leq \tilde{C}_{\alpha,\kappa} \left(\frac{T \log(T)}{\sqrt{n_u}} + \frac{T \log(T)}{\sqrt{n_x}} \right) \log\left(\frac{12}{\delta} \right). \tag{82}$$

With the same assumptions as for $III_t^{(i)}$ we have from (Nguyen and Mücke, 2023, Proposition A.1) (with $\lambda = (\alpha t)^{-1}$)

$$\begin{aligned} III_t^{(ii)} = \mathbb{E}\,|F_t^*(u)(x) - G^*(u)(x)| \leq \|\mathcal{S}_M F_t^* - G^*\|_{L^2(\mu_u)} \leq C_{\alpha,r}\ t^{-r} \ , \\ \sum_{t=0}^{T} III_t^{(ii)} \leq C_{\alpha,r}\ \eta_r(T) \ , \end{aligned} \tag{83}$$

for some $C_{\alpha,r} > 0$ and where we used Jensen inequality in the first equation. Plugging (83) and (82) into (80), implies

$$\alpha \sum_{t=0}^{T} \|\zeta_t^{(1)}\|_\Theta \leq$$
$$C_{\alpha,\kappa,r} \cdot C_{\sigma'} \cdot \left(\frac{T B_\tau^5 + T^{1+\tilde{\nu}} B_\tau^3}{M} + \frac{B_\tau^2}{\sqrt{M}} \cdot \left(\eta_r(T) + \left(\frac{T \log(T)}{\sqrt{n_u}} + \frac{T \log(T)}{\sqrt{n_x}} \right) \log^{\frac{3}{2}} \left(\frac{12}{\delta} \right) \right) \right),$$

for some $C_{\alpha,\kappa,r} > 0$ and where we set $C_{\sigma'} := \max\{\|\sigma'\|_\infty, \|\sigma'\|_\infty^2, \|\sigma''\|_\infty, \|\sigma''\|_\infty^2\}$. ■

Proposition 23 *For all $t \in [T]$ and some $T \in \mathbb{N}$, let $\|\theta_t - \theta_0\| \leq B_\tau$. Then, with probability of at least $1 - \delta$,*

$$\max_{t \in [T]} \|\zeta_t^{(2)}\|_\Theta \leq \frac{\tilde{d}\, B_\tau^2 \cdot C_{\sigma',\sigma''}}{\sqrt{n_u \cdot M}} \sqrt{\log(4\tilde{d}^2/\delta)},$$

with $C_{\sigma',\sigma''} = 334 \cdot \max\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\} \cdot \max\{L_{\sigma'}, L_{\sigma''}\}$. Here, $L_{\sigma'}$ and $L_{\sigma''}$ denote the Lipschitz constants of σ' and σ'', respectively.

Proof [Proposition 23] We start with the following decomposition for any $t \in [T]$,

$$\|\zeta_t^{(2)}\|_\Theta = \left\| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} (G^*(u_i)(x_j) - v_i(x_j))(\nabla G_{\theta_s}(u_i)(x_j) - \nabla G_{\theta_0}(u_i)(x_j)) \right\|_\Theta$$
$$= \left[\frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} (G^*(u_i)(x_j) - v_i(x_j)) \left(\sigma((b_m^{(t)})^\top J(u_i)(x_j)) - \sigma((b_m^{(0)})^\top J(u_i)(x_j)) \right) \right)^2 + \right.$$
$$\frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{\tilde{d}} \left(\frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} (G^*(u_i)(x_j) - v_i(x_j)) \right.$$
$$\left. \left. \cdot \left(a_m^{(t)} \sigma'((b_m^{(t)})^\top J(u_i)(x_j)) J(u_i)(x_j)^{(k)} - a_m^{(0)} \sigma'((b_m^{(0)})^\top J(u_i)(x_j)) J(u_i)(x_j)^{(k)} \right) \right)^2 \right]^{\frac{1}{2}}$$
$$=: \sqrt{I + II} \tag{84}$$

We will now bound I and II separately.

I) Using the mean-value theorem for $b_m^{(t)}$ we obtain for some $\tilde{b}_m$ on the line between $b_m^{(t)}$ and $b_m^{(0)}$:

$$
\begin{aligned}
I =& \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\left(\sigma((b_m^{(t)})^\top J(u_i)(x_j))-\sigma((b_m^{(0)})^\top J(u_i)(x_j))\right)\right)^2\\
=& \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\sigma'(\tilde{b}_m^\top J(u_i)(x_j))\left\langle b_m^{(t)}-b_m^{(0)},\, J(u_i)(x_j)\right\rangle\right)^2\\
\leq& \frac{1}{M}\sum_{m=1}^{M}\left\|b_m^{(t)}-b_m^{(0)}\right\|_2^2\left\|\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\sigma'(\tilde{b}_m^\top J(u_i)(x_j))J(u_i)(x_j)\right\|_2^2\\
\leq& \frac{1}{M}\left\|\theta_t-\theta_0\right\|_\Theta^2\sup_{\|b\|_2\leq C_{B_\tau}}\left\|\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\sigma'(b^\top J(u_i)(x_j))J(u_i)(x_j)\right\|_2^2\\
\leq& \frac{B_\tau^2}{M}\sum_{k=1}^{\tilde{d}}\sup_{\|b\|_2\leq C_{B_\tau}}\left(\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\sigma'(b^\top J(u_i)(x_j))J(u_i)(x_j)^{(k)}\right)^2,
\end{aligned}
$$

with $C_{B_\tau} := B_\tau + 1$. For the penultimate inequality we used that $max_{i\in[M],t\in[T]}\{|a_t^{(i)} - a_0^{(i)}|, \|b_t^{(i)} - b_0^{(i)}\|_2\} \leq B_\tau$ and therefore

$$
\max_{i\in[M]}\{|\tilde{a}^{(i)}|, \|\tilde{b}^{(i)}\|_2\} \leq \max_{i\in[M]}\{|a_t^{(i)}-a_0^{(i)}|+|a_0^{(i)}|, \|b_t^{(i)}-b_0^{(i)}\|_2+\|b_0^{(i)}\|_2\} \leq B_\tau + 1. \tag{85}
$$

II) Similarly, we obtain with (85) and the Cauchy-Schwarz inequality

$$
\begin{aligned}
II = \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{\tilde{d}}\Bigg(&\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\\
&\cdot\left(a_m^{(t)}\sigma'((b_m^{(t)})^\top J(u_i)(x_j))J(u_i)(x_j)^{(k)}-a_m^{(0)}\sigma'((b_m^{(0)})^\top J(u_i)(x_j))J(u_i)(x_j)^{(k)}\right)\Bigg)^2\\
\leq&\frac{2}{M}\sum_{m=1}^{M}\sum_{k=1}^{\tilde{d}}\left(\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\left(a_m^{(t)}-a_m^{(0)}\right)\sigma'((b_m^{(t)})^\top J(u_i)(x_j))J(u_i)(x_j)^{(k)}\right)^2+\\
&\frac{2}{M}\sum_{m=1}^{M}\sum_{k=1}^{\tilde{d}}\Bigg(\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))a_m^{(0)}\\
&\cdot\left(\sigma'((b_m^{(t)})^\top J(u_i)(x_j))J(u_i)(x_j)^{(k)}-\sigma'((b_m^{(0)})^\top J(u_i)(x_j))J(u_i)(x_j)^{(k)}\right)\Bigg)^2\\
\leq&\frac{2B_\tau^2}{M}\sum_{k=1}^{\tilde{d}}\sup_{\|b\|_2\leq C_{B_\tau}}\left(\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\sigma'(b^\top J(u_i)(x_j))J(u_i)(x_j)^{(k)}\right)^2+\\
&\frac{2}{M}\sum_{m=1}^{M}\sum_{k=1}^{\tilde{d}}\Bigg(\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\\
&\cdot\left(\sigma'((b_m^{(t)})^\top J(u_i)(x_j))J(u_i)(x_j)^{(k)}-\sigma'((b_m^{(0)})^\top J(u_i)(x_j))J(u_i)(x_j)^{(k)}\right)\Bigg)^2.
\end{aligned}
\tag{86}
$$

For the second sum (86) we use again the mean-value theorem to obtain

$$
\begin{aligned}
&\frac{2}{M}\sum_{m=1}^{M}\sum_{k=1}^{\tilde{d}}\Bigg(\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\\
&\qquad\cdot\left(\sigma'((b_m^{(t)})^\top J(u_i)(x_j))J(u_i)(x_j)^{(k)}-\sigma'((b_m^{(0)})^\top J(u_i)(x_j))J(u_i)(x_j)^{(k)}\right)\Bigg)^2\\
&\leq\frac{2}{M}\sum_{m=1}^{M}\sum_{k=1}^{\tilde{d}}\left(\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\sigma''(\tilde{b}_m^\top J(u_i)(x_j))\left\langle b_m^{(t)}-b_m^{(0)},\,J(u_i)(x_j)\right\rangle J(u_i)(x_j)^{(k)}\right)^2\\
&\leq\frac{2}{M}\sum_{m=1}^{M}\left\|b_m^{(t)}-b_m^{(0)}\right\|_2^2\sum_{k=1}^{\tilde{d}}\left\|\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\sigma''(\tilde{b}_m^\top J(u_i)(x_j))\,J(u_i)(x_j)J(u_i)(x_j)^{(k)}\right\|_2^2\\
&\leq\frac{2B_\tau^2}{M}\sum_{k,l=1}^{\tilde{d}}\sup_{\|b\|_2\leq C_{B_\tau}}\left(\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\sigma''(b^\top J(u_i)(x_j))J(u_i)(x_j)^{(k)}J(u_i)(x_j)^{(l)}\right)^2.
\end{aligned}
$$

To sum up, we obtain by plugging the bounds of I and II into (84),

$$\max_{t\in[T]}\|\zeta_t^{(2)}\|_\Theta \leq \frac{B_\tau}{\sqrt{M}}\Bigg[3\sum_{k=1}^{\tilde{d}}\sup_{\|b\|_2\leq C_{B_\tau}}\left(\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\sigma'(b^\top J(u_i)(x_j))J(u_i)(x_j)^{(k)}\right)^2 + 2\sum_{k,l=1}^{\tilde{d}}\sup_{\|b\|_2\leq C_{B_\tau}}\left(\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}(G^*(u_i)(x_j)-v_i(x_j))\sigma''(b^\top J(u_i)(x_j))J(u_i)(x_j)^{(k)}J(u_i)(x_j)^{(l)}\right)^2\Bigg]^{\frac{1}{2}}. \tag{87}$$

Setting $z=(u,v)\in\mathcal{Z}:=\mathcal{U}\times\mathcal{V}$ we define for all $k,l\in[\tilde{d}]$,

$$h_{k,j}(z) = (G^*(u)(x_j)-v(x_j))J(u)(x_j)^{(k)},$$
$$h_{k,l,j}(z) = (G^*(u)(x_j)-v(x_j))J(u)(x_j)^{(k)}J(u)(x_j)^{(l)},$$

and the following events,

$$E_k := \left\{\sup_{\|b\|_2\leq C_{B_\tau}}\left|\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}h_{k,j}(z_i)\sigma'(b^\top J(u_i)(x_j))\right| \leq \frac{24\|\sigma'\|_\infty}{\sqrt{n_u}}\left(2C_{B_\tau}L_{\sigma'}+\sqrt{\log\frac{2}{\delta}}\right)\right\},$$
$$E_{k,l} := \left\{\sup_{\|b\|_2\leq C_{B_\tau}}\left|\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}h_{k,l,j}(z_i)\sigma''(b^\top J(u_i)(x_j))\right| \leq \frac{24\|\sigma''\|_\infty}{\sqrt{n_u}}\left(2C_{B_\tau}L_{\sigma''}+\sqrt{\log\frac{2}{\delta}}\right)\right\}.$$

Here, $L_{\sigma'}, L_{\sigma''}$ denote the Lipschitz constants of σ' and σ'', respectively, and $C_{B_\tau}=B_\tau+1$. We aim at applying Proposition 29 for bounding the probabilities. To this end, note that $\|h_{k,j}\|_\infty\leq 2$ and $\|h_{k,l,j}\|_\infty\leq 2$, so $C_h=2$. Moreover,

$$\mathbb{E}_\mu[h_{k,j}(z)\sigma'(b^\top x)] = \mathbb{E}_\mu[h_{k,l,j}(z)\sigma''(b^\top x)] = 0.$$

As a result, the events $E_k, E_{k,l}$ occur with probability of at least $1-\delta$.

Finally, we obtain from Proposition 34 together with (87), that with probability of at least $1-\delta$

$$\max_{t\in[T]}\|\zeta_t^{(2)}\|_\Theta \leq \frac{\sqrt{3}\tilde{d}B_\tau}{\sqrt{M}}\,\frac{24\max\{\|\sigma'\|_\infty,\|\sigma''\|_\infty\}}{\sqrt{n_u}}\left(4C_{B_\tau}\max\{L_{\sigma'},L_{\sigma''}\}+2\sqrt{\log\frac{2(\tilde{d}^2+\tilde{d})}{\delta}}\right)$$
$$\leq \frac{\tilde{d}\,B_\tau^2\cdot C_{\sigma',\sigma''}}{\sqrt{n_u\cdot M}}\;\sqrt{\log(4\tilde{d}^2/\delta)},$$

with $C_{\sigma',\sigma''}=334\cdot\max\{\|\sigma'\|_\infty,\|\sigma''\|_\infty\}\cdot\max\{L_{\sigma'},L_{\sigma''}\}$. This proves the claim.

■

Appendix D. Technical Inequalities

D.1 Lipschitz Bounds

In this section we provide some technical inequalities. We start proving that the gradient of our neural operator is pointewisely Lipschitz-continuous on a ball around θ_0 which we define as

$$Q_{\theta_0}(R) := \{\theta = (a, B) \in \Theta : \|\theta - \theta_0\|_\Theta \leq R\}.$$

Moreover, for $u \in \mathcal{U}$ we define

$$\|u\|_\infty := \sup_{x \in \text{supp}(\rho_X)} \|u(x)\|_2.$$

Proposition 24 *Suppose that Assumptions 1 and 4 are satisfied. Let $u \in \mathcal{U}$ with $\|u\|_\infty \leq 1$. Then, for any $x \in \mathcal{X}$ and any $\theta, \tilde{\theta} \in Q_{\theta_0}(R)$,*

$$\left\|\nabla G_\theta(u)(x) - \nabla G_{\tilde{\theta}}(u)(x)\right\|_\Theta \leq \frac{C_{\nabla G}(R)}{\sqrt{M}} \|\theta - \tilde{\theta}\|_\Theta.$$

Here, the constant $C_{\nabla G}(R)$ is given by

$$C_{\nabla G}(R) := 2 \max\left\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\right\} \max\{1, R + \tau\}. \tag{88}$$

Proof [Proposition 24] Recall the definition of our operator class

$$\mathcal{F}_M := \Big\{G_\theta : \mathcal{U} \to \mathcal{V} \mid \ G_\theta(u) = \tfrac{1}{\sqrt{M}} \textstyle\sum_{m=1}^M a_m \sigma\left(\langle b_m, J(u)\rangle\right), \\ \theta = (a, B) \in \mathbb{R}^M \times \mathbb{R}^{M \times (d_k + d_y + d_b)}\Big\}.$$

The partial derivatives are given by

$$\partial_{a_m} G_\theta(u)(x) = \frac{1}{\sqrt{M}} \sigma\left(\langle b_m, J(u)(x)\rangle\right), \quad \partial_{b_m^j} G_\theta(u)(x) = \frac{a_m}{\sqrt{M}} \sigma'\left(\langle b_m, J(u)(x)\rangle\right) (J(u)(x))^{(j)}).$$

Denoting further $\nabla G_\theta(u) = (\partial_a G_\theta(u), \partial_B G_\theta(u))$

$$\partial_a G_\theta(u) = (\partial_{a_1} G_\theta(u), \ldots, \partial_{a_M} G_\theta(u)) \in \mathbb{R}^M, \quad \partial_B G_\theta(u) = \left(\partial_{b_m^j} G_\theta(u)(x)\right)_{m,j} \in \mathbb{R}^{M \times (d_k + d_y + d_b)}$$

we then have for any $u \in \mathcal{U}$, $x \in supp(\rho_X)$,

$$\begin{aligned}
\left\|\partial_a G_\theta(u)(x) - \partial_{\tilde{a}} G_{\tilde{\theta}}(u)(x)\right\|_2^2 &= \sum_{m=1}^M \left(\partial_{a_m} G_\theta(u)(x) - \partial_{\tilde{a}_m} G_{\tilde{\theta}}(u)(x)\right)^2 \\
&= \frac{1}{M} \sum_{m=1}^M \left(\sigma(\langle b_m, J(u)(x)\rangle) - \sigma(\langle \tilde{b}_m, J(u)(x)\rangle)\right)^2 \\
&\leq \frac{\|\sigma'\|_\infty^2}{M} \sum_{m=1}^M \left\|b_m - \tilde{b}_m\right\|_2^2 \|J(u)(x)\|_2^2 \\
&\leq \frac{\|\sigma'\|_\infty^2}{M} \|B - \tilde{B}\|_F^2,
\end{aligned}$$

where we used the assumption $\|J(u)(x)\|_2 \leq \|J(u)(x)\|_1 \leq 1$. In addition,

$$\begin{aligned}
&\left\|\partial_B G_\theta(u)(x) - \partial_{\tilde{B}} G_{\tilde{\theta}}(u)(x)\right\|_2^2 = \sum_{m=1}^{M} \sum_{j=1}^{d_k+d_y+d_b} \left(\partial_{b_m} G_\theta(u)(x) - \partial_{\tilde{b}_m} G_{\tilde{\theta}}(u)(x)\right)^2 \\
&= \frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{d_k+d_y+d_b} \left(a_m \sigma'(\langle b_m, J(u)(x)\rangle)(J(u)(x))^{(j)}) - \tilde{a}_M \sigma'(\langle \tilde{b}_m, J(u)(x)\rangle)(J(u)(x))^{(j)})\right)^2 \\
&\leq (I) + (II)
\end{aligned}$$

where

$$(I) = \frac{2}{M} \sum_{m=1}^{M} \sum_{j=1}^{d_k+d_y+d_b} (a_m - \tilde{a}_m)^2 \left(\sigma'\left\langle b_m, J(u)(x)\right\rangle\right)^2 ((J(u)(x))^{(j)})^2 \leq \frac{2\|\sigma'\|_\infty^2}{M} \|a - \tilde{a}\|_2^2$$

and where

$$\begin{aligned}
(II) &= \frac{2}{M} \sum_{m=1}^{M} \sum_{j=1}^{d_k+d_y+d_b} (\tilde{a}_m - a_{0,m})^2 \left(\sigma'(\langle b_m, J(u)(x)\rangle) - \sigma'(\langle \tilde{b}_m, J(u)(x)\rangle)\right)^2 ((J(u)(x))^{(j)}))^2 \\
&\quad + \frac{2}{M} \sum_{m=1}^{M} \sum_{j=1}^{d_k+d_y+d_b} a_{0,m}^2 \left(\sigma'(\langle b_m, J(u)(x)\rangle) - \sigma'(\langle \tilde{b}_m, J(u)(x)\rangle)\right)^2 ((J(u)(x))^{(j)}))^2 \\
&\leq \frac{2R^2\|\sigma''\|_\infty^2}{M} \sum_{m=1}^{M} \sum_{j=1}^{d_k+d_y+d_b} \left(\langle b_m - \tilde{b}_m, J(u)(x)\rangle\right)^2 ((J(u)(x))^{(j)}))^2 \\
&\quad + \frac{2\tau^2\|\sigma''\|_\infty^2}{M} \sum_{m=1}^{M} \sum_{j=1}^{d_k+d_y+d_b} \left(\langle b_m - \tilde{b}_m, J(u)(x)\rangle\right)^2 ((J(u)(x))^{(j)}))^2 \\
&\leq \frac{2\|\sigma''\|_\infty^2 (R^2+\tau^2)}{M} \|B - \tilde{B}\|_F^2.
\end{aligned}$$

Finally,

$$\begin{aligned}
\left\|\nabla G_\theta(u)(x) - \nabla g_{\tilde{\theta}}(u)(x)\right\|_\Theta^2 &= \left\|\partial_a G_\theta(u)(x) - \partial_{\tilde{a}} G_{\tilde{\theta}}(u)(x)\right\|_2^2 + \left\|\partial_B G_\theta(u)(x) - \partial_{\tilde{B}} G_{\tilde{\theta}}(u)(x)\right\|_F^2 \\
&\leq \frac{\|\sigma'\|_\infty^2}{M} \|B - \tilde{B}\|_F^2 + \frac{2\|\sigma'\|_\infty^2}{M} \|a - \tilde{a}\|_2^2 + \frac{2\|\sigma''\|_\infty^2 (R^2+\tau^2)}{M} \|B - \tilde{B}\|_F^2 \\
&\leq \frac{4\|\sigma'\|_\infty^2}{M} \max\left\{1, R^2+\tau^2\right\} \left(\|a - \tilde{a}\|_2^2 + \|B - \tilde{B}\|_F^2\right) \\
&\leq \frac{C_{\nabla G}(R)^2}{M} \|\theta - \tilde{\theta}\|_\Theta^2,
\end{aligned}$$

where we set $C_{\nabla G}(R) = 2\max\left\{\|\sigma'\|_\infty, \|\sigma''\|_\infty\right\} \max\left\{1, R+\tau\right\}$. ■

Proposition 25 *Suppose Assumption 1 holds and that $\theta, \bar{\theta} \in Q_{\theta_0}(R)$. For all $x \in \mathcal{X}$ and $u \in \mathcal{U}$, denote by $r_{(\theta,\bar{\theta})}(u)(x)$ the remainder term of the Taylor expansion from*

$$G_\theta(u)(x) - G_{\bar{\theta}}(u)(x) = \left\langle \nabla G_{\bar{\theta}}(u)(x), \theta - \bar{\theta} \right\rangle_\Theta + r_{(\theta,\bar{\theta})}(u)(x).$$

Then,

1. $|r_{(\theta,\bar{\theta})}(u)(x)| \leq \frac{C_{\nabla G}(R)}{\sqrt{M}} \|\theta - \bar{\theta}\|_\Theta^2$,
2. $\left\langle \nabla G_\theta(u)(x) - \nabla G_{\bar{\theta}}(u)(x), \theta - \bar{\theta} \right\rangle_\Theta \leq \frac{C_{\nabla G}(R)}{\sqrt{M}} \|\theta - \bar{\theta}\|_\Theta^2$,

where $C_{\nabla G}(R)$ is defined in Proposition 24.

Proof [Proposition 25]

1. Recall that the remainder term $r_{(\theta,\bar{\theta})}$ of the Taylor expansion is given by

$$r_{(\theta,\bar{\theta})}(u)(x) = \left(\theta - \bar{\theta}\right)^T \nabla^2 G_{\tilde{\theta}}(u)(x) \left(\theta - \bar{\theta}\right),$$

where $\nabla^2 G_{\tilde{\theta}}(u)(x)$ denotes the Hessian matrix evaluated at some $\tilde{\theta}$ on the line between θ and $\bar{\theta}$. Note that the remainder term can be bounded by the Lipschitz constant L of the gradients i.e. for any $\mathbf{v} \in \Theta$ we have,

$$\begin{aligned}
\nabla^2 G_{\tilde{\theta}}(u)(x)\mathbf{v} &= \lim_{h \to 0} \frac{\nabla G_{\tilde{\theta}+h\mathbf{v}}(u)(x) - \nabla G_{\tilde{\theta}}(u)(x)}{h} \\
&\leq \lim_{h \to 0} \frac{\|\nabla G_{\tilde{\theta}+h\mathbf{v}}(x) - \nabla G_{\tilde{\theta}}(u)(x)\|_2}{|h|} \\
&\leq \lim_{h \to 0} L \, \frac{|h| \|\mathbf{v}\|_2}{|h|} \\
&\leq L \, \|\mathbf{v}\|.
\end{aligned}$$

Proposition 24 shows that for the set $Q_{\theta_0}(R)$ this Lipschitz constant is given by $L = C_{\nabla G}/\sqrt{M}$. Therefore by setting $\mathbf{v} = \theta - \bar{\theta}$ we obtain for the remainder term,

$$|r_{(\theta,\bar{\theta})}(u)(x)| \leq \frac{C_{\nabla G}(R)}{\sqrt{M}} \|\theta - \bar{\theta}\|_\Theta^2.$$

2. Using again Proposition 24 together with Cauchy - Schwarz inequality, we obtain for the second inequality,

$$\left\langle \nabla G_\theta(u)(x) - \nabla G_{\bar{\theta}}(u)(x), \theta - \bar{\theta} \right\rangle_\Theta \leq \frac{C_{\nabla G}(R)}{\sqrt{M}} \|\theta - \bar{\theta}\|_\Theta^2.$$

■

Proposition 26 *Suppose the Assumptions 1, 4 hold and that $\theta \in Q_{\theta_0}(R)$. Then, for any $u \in \mathcal{U}, x \in \mathcal{X}$, we have*

$$|G_\theta(u)(x) - G_{\theta_0}(u)(x)| \leq \kappa \cdot \|\theta - \theta_0\|_\Theta + \frac{C_{\nabla G}(R)}{\sqrt{M}} \|\theta - \theta_0\|_\Theta^2,$$

where $\kappa > 0$ is the constant from (10).

Proof [Proposition 26]

By Assumption 4, for any $(u, v) \in \operatorname{supp}(\mu)$ and any $x \in \operatorname{supp}(\rho_x)$, we have

$$\|J(u)(x)\|_1 \leq 1.$$

We compute the squared gradient norm at initialization θ_0:

$$\begin{aligned}
\|\nabla G_{\theta_0}(u)(x)\|_\Theta^2 &= \sum_{p=1}^{P} \big(\partial_p (G_{\theta_0}(u)(x))\big)^2 \\
&= \frac{1}{M} \sum_{m=1}^{M} \sigma\Big(\Big\langle b_m^{(0)}, J(u)(x) \Big\rangle\Big)^2 \\
&\quad + \frac{\tau^2}{M} \sum_{m=1}^{M} \sum_{j=1}^{\tilde{d}} \sigma'\Big(\Big\langle b_m^{(0)}, J(u)(x) \Big\rangle\Big)^2 \big(J(u)(x)^{(j)}\big)^2. \qquad (89)
\end{aligned}$$

Using $\|J(u)(x)\|_1 \leq 1$ and the Assumption 1, we obtain

$$\|\nabla G_{\theta_0}(u)(x)\|_\Theta^2 \leq 4 + \tau^2 \|\sigma'\|_\infty^2 = \kappa^2.$$

Using this bound together with Proposition 25, we obtain

$$\begin{aligned}
G_\theta(u)(x) - G_{\theta_0}(u)(x) &= \langle \nabla G_{\theta_0}(u)(x), \theta - \theta_0 \rangle_\Theta + r_{(\theta,\theta_0)}(u)(x) \\
&\leq \kappa \|\theta - \theta_0\|_\Theta + \frac{C_{\nabla G}(R)}{\sqrt{M}} \|\theta - \theta_0\|_\Theta^2.
\end{aligned}$$

■

D.2 Uniform Bounds in Hilbert Spaces

In this section we recall uniform convergence bounds for functions with values in a real separable Hilbert space.

Proposition 27 *Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z} := \mathcal{U} \times \mathcal{V}$ to $[-C, C]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size n_u, the following holds*

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{n_u}\sum_{i=1}^{n} g\left(z_i\right)-\mathbb{E}[g(z)]\right| \leq 8C\widehat{\mathfrak{R}}_S(\mathcal{G})+12C\sqrt{\frac{\log\frac{2}{\delta}}{2n_u}},$$

where

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) := \mathbb{E}_r \sup_{g\in\mathcal{G}}\frac{1}{n_u}\sum_{i=1}^{n_u} g(x_i)r_i \tag{90}$$

denotes the empirical Rademacher complexity of the class $\mathcal{G}$ with r_i being iid rademacher variables.

Proof [Proposition 27] The proof can be found in Mohri et al. (2018). ■

Proposition 28 *Let $\tilde{\sigma}:\mathbb{R}\to\mathbb{R}$ be Lipschitz continuous with Lipschitz constant $c_{\tilde{\sigma}}$. Let further $h_j:\mathcal{Z}=\mathcal{U}\times\mathcal{V}\to\mathbb{R}$, $j=1,...,n_x$, and define*

$$\mathcal{G}_C=\{g: B_{\tilde{d}}(C)\times\mathcal{Z}\to\mathbb{R} : g(b,z)=\frac{1}{n_x}\sum_{j=1}^{n_x} h_j(z)\tilde{\sigma}(\langle b, J(u)(x_j)\rangle)\},$$

where $B_{\tilde{d}}(C)$ denotes the closed ball of radius C in $\mathbb{R}^{\tilde{d}}$. Provided the Assumptions 4 and 1 we have

$$\widehat{\mathfrak{R}}_S(\mathcal{G}_C)\leq\frac{CC_hc_{\tilde{\sigma}}}{\sqrt{n_u}},$$

with $C_h:=\max_{j\in[n_x]}\|h_j\|_\infty$.

Proof [Proposition 28] By definition of $\widehat{\mathfrak{R}}_S(\mathcal{G}_C)$ we have,

$$\widehat{\mathfrak{R}}_S(\mathcal{G}_C)=\frac{1}{n_u}\mathbb{E}_r\left[\sup_{\|b\|_2\leq C}\sum_{i=1}^{n_u} g\left(b,z_i\right)r_i\right]=\frac{1}{n_u}\mathbb{E}_{r_1,\ldots,r_{n_u-1}}\left[\mathbb{E}_{r_{n_u}}\left[\sup_{\|a\|_2\leq C} q_{n_u-1}(b)+g\left(b,z_{n_u}\right)r_{n_u}\right]\right],$$

where $q_{n_u-1}(b)=\sum_{i=1}^{n_u-1} g\left(b,z_i\right)r_i$. By definition of the supremum, for any $\epsilon>0$, there exist $\|b_1\|_2,\|b_2\|_2\leq C$ such that

$$q_{n_u-1}\left(b_1\right)+g(b_1,z_{n_u})\geq(1-\epsilon)\left[\sup_{\|b\|_2\leq C} q_{n_u-1}(b)+g\left(b,z_{n_u}\right)\right]$$
$$\text{and } q_{n_u-1}\left(b_2\right)-g\left(b_2,z_{n_u}\right)\geq(1-\epsilon)\left[\sup_{\|b\|_2\leq C} q_{n_u-1}(b)-g\left(b,z_{n_u}\right)\right].$$

Thus by definition of $\mathbb{E}_{r_{n_u}}$,

$$\begin{aligned}
&(1-\epsilon)\mathbb{E}_{r_{n_u}}\left[\sup_{\|b\|_2\leq C} q_{n_u-1}(b) + r_{n_u} g(b, z_{n_u})\right] \\
=&(1-\epsilon)\left[\frac{1}{2}\sup_{\|b\|_2\leq C}\left[q_{n_u-1}(b) + g(b, z_{n_u})\right] + \frac{1}{2}\sup_{\|b\|_2\leq C}\left[q_{n_u-1}(b) - g(b, z_{n_u})\right]\right] \\
\leq&\frac{1}{2}\left[q_{n_u-1}(b_1) + g(b_1, z_{n_u})\right] + \frac{1}{2}\left[q_{n_u-1}(b_2) - g(b_2, z_{n_u})\right].
\end{aligned}$$

Now set $s^{n_u} = \operatorname{sgn}\tilde{g}(z_{n_u}, b_1 - b_2)$ with $\tilde{g}(z, b) := \frac{1}{n_x}\sum_{j=1}^{n_x} h_j(z)\langle b, J(u)(x_j)\rangle$ (note that s^{n_u} is independent of r_{n_u}). Then, the previous inequality implies together with the Lipschitz property:

$$\begin{aligned}
&(1-\epsilon)\mathbb{E}_{r_{n_u}}\left[\sup_{\|b\|_2\leq C} q_{n_u-1}(b) + r_{n_u} g(b, z_{n_u})\right] \\
&\leq \frac{1}{2}\left[q_{n_u-1}(b_1) + q_{n_u-1}(b_2) + c_\sigma s^{n_u}\tilde{g}(z_{n_u}, b_1 - b_2)\right] \\
&\leq \frac{1}{2}\sup_{\|b\|_2\leq C}\left[q_{n_u-1}(b) + c_\sigma s^{n_u}\tilde{g}(z_{n_u}, b)\right] + \frac{1}{2}\sup_{\|b\|_2\leq C}\left[q_{n_u-1}(b) - c_\sigma s^{n_u}\tilde{g}(z_{n_u}, b)\right] \\
&= \frac{1}{2}\sup_{\|b\|_2\leq C}\left[q_{n_u-1}(b) + c_\sigma\tilde{g}(z_{n_u}, b)\right] + \frac{1}{2}\sup_{\|b\|_2\leq C}\left[q_{n_u-1}(b) - c_\sigma\tilde{g}(z_{n_u}, b)\right] \\
&= \mathbb{E}_{r_n}\left[\sup_{\|b\|_2\leq C} q_{n_u-1}(b) + r_{n_u} c_\sigma\tilde{g}(z_{n_u}, b)\right].
\end{aligned}$$

Therefore, we proved

$$(1-\epsilon)\mathbb{E}_r\left[\sup_{\|b\|_2\leq C} q_{n_u}(b)\right] \leq \mathbb{E}_{r_{n_u}}\left[\sup_{\|b\|_2\leq C} q_{n_u-1}(b) + r_{n_u} c_\sigma\tilde{g}(z_{n_u}, b)\right]. \tag{91}$$

Proceeding in the same way for all other q_i with $i \leq n_u - 1$ leads to

$$\begin{aligned}
(1-\epsilon)^{n_u}\frac{1}{n_u}\mathbb{E}_r\left[\sup_{\|b\|_2\leq C}\sum_{i=1}^{n_u} g\left(b, z_i\right) r_i\right] &\leq \frac{1}{n_u}\mathbb{E}_r\left[\sup_{\|b\|_2\leq C}\sum_{i=1}^{n_u} r_i c_\sigma \tilde{g}(z_i, b)\right] \\
&= \frac{c_{\tilde{\sigma}}}{n_u}\mathbb{E}_r\left[\sup_{\|b\|_2\leq C}\left\langle b, \sum_{i=1}^{n_u} r_i \frac{1}{n_x}\sum_{j=1}^{n_x} h_j(z) J(u)(x_j)\right\rangle\right] \\
&\leq \frac{C c_{\tilde{\sigma}}}{n_u}\mathbb{E}_r\left[\left\|\sum_{i=1}^{n_u} r_i \frac{1}{n_x}\sum_{j=1}^{n_x} h_j(z) J(u)(x_j)\right\|_2\right] \\
&\leq \frac{C c_{\tilde{\sigma}}}{n_u}\sqrt{\sum_{k=1}^{\tilde{d}}\mathbb{E}_r\left(\sum_{i=1}^{n_u} r_i \frac{1}{n_x}\sum_{j=1}^{n_x} h_j(z) J(u)(x_j)^{(k)}\right)^2} \\
&= \frac{C c_{\tilde{\sigma}}}{n_u}\sqrt{\sum_{k=1}^{\tilde{d}}\sum_{i=1}^{n_u}\left(\frac{1}{n_x}\sum_{j=1}^{n_x} h_j(z) J(u)(x_j)^{(k)}\right)^2} \\
&\leq \frac{C C_h c_{\tilde{\sigma}}}{n_u}\sqrt{\frac{1}{n_x}\sum_{j=1}^{n_x}\sum_{i=1}^{n_u}\sum_{k=1}^{\tilde{d}}\left(J(u)(x_j)^{(k)}\right)^2} \\
&\leq \frac{C C_h c_{\tilde{\sigma}}}{\sqrt{n_u}}.
\end{aligned}$$

Hence, we obtain for $\epsilon \to 0$

$$\frac{1}{n_u}\mathbb{E}_r\left[\sup_{\|b\|_2\leq C}\sum_{i=1}^{n_u} g\left(b, z_i\right) r_i\right] \leq \frac{C C_h c_{\tilde{\sigma}}}{\sqrt{n_u}}.$$

■

Proposition 29 *Let $\tilde{\sigma} : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous with Lipschitz constant $c_{\tilde{\sigma}}$. Let further $h_j : \mathcal{Z} = \mathcal{U} \times \mathcal{V} \to \mathbb{R}$, $j = 1, ..., n_x$, such that*

$$\mathbb{E}[h_j(z)\tilde{\sigma}(b^\top J(u)(x))] = 0$$

for any $\|b\|_2 \leq C_B$ and $\max_{j\in[n_x]}\|h_j\|_\infty \leq C_h$. Then, we have with probability of at least $1-\delta$,

$$\sup_{\|b\|_2\leq C_B}\left|\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x} h_j(z_i)\tilde{\sigma}(b^\top J(u_i)(x_j))\right| \leq \frac{12 C_h \|\tilde{\sigma}\|_\infty}{\sqrt{n_u}}\left(C_B c_{\tilde{\sigma}} C_h + \sqrt{\log\frac{2}{\delta}}\right).$$

Proof [Proposition 29] From Proposition 27 we have with probability $1-\delta$

$$\sup_{\|b\|_2 \leq C_B} \left| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} h_j(z_i) \tilde{\sigma}(b^\top J(u_i)(x_j)) \right| \\ \leq 8C_h \|\tilde{\sigma}\|_\infty \widehat{\mathfrak{R}}_S(\mathcal{G}_{C_B}) + 12 C_h \|\tilde{\sigma}\|_\infty \sqrt{\frac{\log \frac{2}{\delta}}{2n_u}}, \tag{92}$$

where $\widehat{\mathfrak{R}}_S(\mathcal{G}_{C_B})$ denotes the empirical Rademacher complexity defined in (90) and the function class $\mathcal{G}_{C_B}$ is defined as

$$\mathcal{G}_{C_B} = \left\{ g : B_{\tilde{d}}(C_B) \times \mathcal{Z} \to \mathbb{R} : g(b,z) = \sum_{j=1}^{n_x} h_j(z_i) \tilde{\sigma}(b^\top J(u_i)(x_j)) \right\}.$$

From Proposition 28 we have

$$\widehat{\mathfrak{R}}_S(\mathcal{G}_{C_B}) \leq \frac{C_B c_{\tilde{\sigma}} C_h}{\sqrt{n_u}}.$$

Plugging this inequality into (92) proves the claim. ■

D.3 Elementary Inequalities

Lemma 30 *The function $f(x) = (1-x)^j x^r$ for $j, r > 0$ has a global maximum on $[0,1]$ at $\frac{r}{r+j}$. Therefore we have $\sup_{x \in [0,1]} f(x) \leq \left(\frac{r}{r+j}\right)^r$.*

Proof [Lemma 30] If $f(x)$ is continuous and has a maximum, then $\log f(x)$ will at the same point because $\log x$ is a monotonically increasing function.

$$\ln f(x) = r \log x + j \log(1-x) \\ \Rightarrow \frac{\mathrm{d} \ln f(x)}{\mathrm{d}\, x} = \frac{r}{x} - \frac{j}{1-x} \\ \Rightarrow \frac{\mathrm{d}^2 \ln f(x)}{\mathrm{d}\, x^2} = -\frac{r}{x^2} - \frac{j}{(1-x)^2}.$$

The second derivative is obviously less than zero everywhere, and the first derivative is equal zero at $\frac{r}{r+j}$. Therefore it is a global maximum. ■

Lemma 31 *Let $v \geq 0, t \in \mathbb{N}$. Then*

$$\sum_{i=1}^{t} i^{-v} \leq \eta_v(t) \quad \text{with} \quad \eta_v(t) := \begin{cases} \frac{v}{v-1} & v > 1 \\ 1 + \log(t) & v = 1 \\ \frac{t^{1-v}}{1-v} & v \in [0,1) \end{cases}$$

Proof [Lemma 31] We will only prove the first case $v > 1$. The other cases can be bounded with the same arguments. Since

$$i^{-v} \leq \int_{i-1}^{i} x^{-v} dx$$

we have for $v > 1$,

$$\sum_{i=1}^{t} i^{-v} \leq 1 + \int_{1}^{t} x^{-v} dx \leq \frac{v}{v-1}.$$

■

Lemma 32 *Let $a \geq b \geq 0$. Then we have*

$$\sum_{s=1}^{T-1} \frac{1}{s^a} \frac{1}{(T-s)^b} \leq \frac{2^a}{T^a} \eta_b(T-1) + \frac{2^b}{T^b} \eta_a(T-1).$$

Proof [Lemma 32] For any $t \in [T]$ we have

$$\begin{aligned}
\sum_{s=1}^{T-1} \frac{1}{s^a} \frac{1}{(T-s)^b} &= \sum_{s=t+1}^{T-1} \frac{1}{s^a} \frac{1}{(T-s)^b} + \sum_{s=1}^{t} \frac{1}{s^a} \frac{1}{(T-s)^b} \\
&\leq \frac{1}{t^a} \sum_{s=t+1}^{T-1} \frac{1}{(T-s)^b} + \frac{1}{(T-t)^b} \sum_{s=1}^{t} \frac{1}{s^a} \\
&\leq \frac{1}{t^a} \sum_{s=1}^{T-1} \frac{1}{s^b} + \frac{1}{(T-t)^b} \sum_{s=1}^{T-1} \frac{1}{s^a} \\
&\leq \frac{1}{t^a} \eta_b(T-1) + \frac{1}{(T-t)^b} \eta_a(T-1).
\end{aligned}$$

Setting $t = T/2$ the result follows. ■

Proposition 33 *For $T \in \mathbb{N}$, $a \in \{0, \frac{1}{2}\}$, $\alpha \in (0, \frac{1}{\kappa^2}]$ we have,*

$$\begin{aligned}
&i) \quad \sum_{s=0}^{T-1} \|(\alpha \widehat{\Sigma}_M)^a (Id - \alpha \widehat{\Sigma}_M)^s\| \leq \eta_a(T), \\
&ii) \quad \sum_{s=0}^{T-1} \|(\alpha \widehat{\Sigma}_M)^a (Id - \alpha \widehat{\Sigma}_M)^s \widehat{\mathcal{S}}_M^*\| \leq \eta_{a+\frac{1}{2}}(T)/\sqrt{\alpha}, \\
&iii) \quad \sum_{s=0}^{T-1} \left\| (Id - \alpha \widehat{\mathcal{C}}_M)^s (\sqrt{\alpha} \widehat{\mathcal{Z}}_M^*) \right\| \leq \eta_{\frac{1}{2}}(T) = 2\sqrt{T}.
\end{aligned}$$

Further we have for any $M \geq 8\,\tilde{d}\,\kappa^2\beta_\infty T \vee 8\kappa^4\|\mathcal{L}_\infty\|^{-1}\log^2\frac{2}{\delta}$ *with* $\beta_\infty = \log\frac{4\kappa^2(\mathcal{N}_{\mathcal{L}_\infty}(T)+1)}{\delta\|\mathcal{L}_\infty\|}$, $\tilde{d} := d_k + d_y + d_b$ *that with probability at least* $1-\delta/2$,

$$iv)\ \sum_{s=0}^{T-1}\|(\alpha\widehat{\Sigma}_M)^a(Id-\alpha\widehat{\Sigma}_M)^s\mathcal{Z}_M\| \leq \frac{2\eta_{a+\frac{1}{2}}(T)}{\sqrt{\alpha}} + \frac{2\eta_a(T)}{\sqrt{\alpha T}},$$

Proof [Proposition 33] Note that the spectrum of $\alpha\widehat{\Sigma}_M$ is contained in $[0,1]$, therefore by Lemma 30 we have for any $r \geq 0$,

$$\|(\alpha\widehat{\Sigma}_M)^r(Id-\alpha\widehat{\Sigma}_M)^s\| = \sup_{x\in[0,1]}(1-x)^s x^r \leq \left(\frac{r}{r+s}\right)^r, \tag{93}$$

where we use the convention $(0/0)^0 := 1$.
From (93) and Lemma 31 we further obtain for any $r \in [0,1]$

$$\sum_{s=0}^{T-1}\|(\alpha\widehat{\Sigma}_M)^r(Id-\alpha\widehat{\Sigma}_M)^s\| \leq \sum_{s=0}^{T-1}\left(\frac{r}{r+s}\right)^r \tag{94}$$

$$\leq \sum_{s=1}^{T}\left(\frac{1}{s}\right)^r \leq \eta_r(T). \tag{95}$$

This already proves part *i*). For part *ii*) we have from (93) Lemma 31

$$\begin{aligned}
&\sum_{s=0}^{T-1}\|(Id-\alpha\widehat{\Sigma}_M)^s\widehat{\mathcal{S}}_M^*\| \\
&= \sum_{s=0}^{T-1}\sqrt{\|(\alpha\widehat{\Sigma}_M)^{2a+1}(Id-\alpha\widehat{\Sigma}_M)^{2s}\|/\alpha} \\
&\leq \frac{1}{\sqrt{\alpha}}\sum_{s=0}^{T-1}\sqrt{\left(\frac{2a+1}{2a+1+2s}\right)^{2a+1}} \leq \frac{1}{\sqrt{\alpha}}\sum_{s=1}^{T}\left(\frac{1}{s}\right)^{a+\frac{1}{2}} \leq \eta_{a+\frac{1}{2}}(T)/\sqrt{\alpha}.
\end{aligned}$$

Similar as in part *ii*) we have for *iii*) we have from (93) Lemma 31

$$\begin{aligned}
&\sum_{s=0}^{T-1}\|(Id-\alpha\widehat{\mathcal{C}}_M)^s(\sqrt{\alpha}\widehat{\mathcal{Z}}_M^*)\| \\
&= \sum_{s=0}^{T-1}\sqrt{\|(Id-\alpha\widehat{\mathcal{C}}_M)^{2s}(\sqrt{\alpha}\widehat{\mathcal{C}}_M)\|} \\
&\leq \frac{1}{\sqrt{\alpha}}\sum_{s=0}^{T-1}\sqrt{\left(\frac{1}{1+2s}\right)} \leq \sum_{s=0}^{T}\left(\frac{1}{s}\right)^{\frac{1}{2}} \leq \eta_{\frac{1}{2}}(T).
\end{aligned}$$

For part *iv*) we have with probability at least $1-\delta/2$

$$\begin{aligned}
&\sum_{s=0}^{T-1} \|(\alpha\widehat{\Sigma}_M)^a(Id-\alpha\widehat{\Sigma}_M)^s\mathcal{Z}_M\| \\
&\leq \frac{1}{\sqrt{\alpha}}\sum_{s=0}^{T-1} \|(\alpha\widehat{\Sigma}_M)^a(Id-\alpha\widehat{\Sigma}_M)^s(\alpha\widehat{\Sigma}_{M,\lambda})^{\frac{1}{2}}\|\|\widehat{\Sigma}_{M,\lambda}^{-\frac{1}{2}}\Sigma_{M,\lambda}^{\frac{1}{2}}\|\|\Sigma_{M,\lambda}^{-\frac{1}{2}}\mathcal{Z}_M\| \\
&\leq \frac{2\eta_{a+\frac{1}{2}}(T)}{\sqrt{\alpha}} + \frac{2\eta_a(T)}{\sqrt{\alpha T}},
\end{aligned}$$

where we used in the second inequality (94), (Nguyen and Mücke, 2023, Proposition A.15) (for $\lambda := (\alpha T)^{-1}$) and 35. ■

Proposition 34 *For $i = 1, ..., k$, let $\delta_i \in (0,1)$ and let E_i be events with probability of at least $1-\delta_i$. Set*

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

If $\mathbb{P}(A|E) \geq 1-\delta$ for some event A, then

$$\mathbb{P}(A) \geq (1-\delta)\left(1-\sum_{i=1}^{k}\delta_i\right).$$

Proof [Proposition 34] We write

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

■

Appendix E. Operator Bounds and Concentration Inequalities

E.1 Operator Bounds

Proposition 35

a) For any $\lambda > 0$ we have

$$\left\|\widehat{\mathcal{C}}_{M,\lambda}^{-\frac{1}{2}}\widehat{\mathcal{Z}}_M^*\right\| \leq 1, \quad \left\|\Sigma_{M,\lambda}^{-\frac{1}{2}}\mathcal{Z}_M\right\| \leq 1,$$

b) For any $\lambda > 0$ we have

$$\|\mathcal{C}_{M,\lambda}^{-1}\mathcal{Z}_M^*\mathcal{L}_{M,\lambda}^{\frac{1}{2}}\| \leq 2.$$

Proof The bound of b) can be found in Rudi and Rosasco (2017) (see proof of lemma 3). For a) we have

$$\begin{aligned}\left\|\widehat{\mathcal{C}}_{M,\lambda}^{-\frac{1}{2}}\widehat{\mathcal{Z}}_M^*\right\|^2 &= \left\|(\widehat{\mathcal{Z}}_M^*\widehat{\mathcal{Z}}_M + \lambda)^{-1/2}\widehat{\mathcal{Z}}_M^*\right\|^2 \\ &= \left\|(\widehat{\mathcal{Z}}_M^*\widehat{\mathcal{Z}}_M + \lambda)^{-1/2}\widehat{\mathcal{Z}}_M^*\widehat{\mathcal{Z}}_M(\widehat{\mathcal{Z}}_M^*\widehat{\mathcal{Z}}_M + \lambda)^{-1/2}\right\| \\ &= \left\|\widehat{\mathcal{Z}}_M^*\widehat{\mathcal{Z}}_M(\widehat{\mathcal{Z}}_M^*\widehat{\mathcal{Z}}_M + \lambda)^{-1}\right\| \leq 1.\end{aligned}$$

The second inequality follows the same way where we used that $\Sigma_M = \mathcal{Z}_M\mathcal{Z}_M^*$. ■

Proposition 36 *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*

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

with

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

Proposition 37 *For $r \geq 0$ we have*

$$\left\|\mathcal{L}_{\infty,\lambda}^{-\frac{1}{2}}\mathcal{L}_\infty^r\right\| \leq \kappa^r\lambda^{-(\frac{1}{2}-r)^+}.$$

Proof For $r \geq 1/2$ we have $\left\|\mathcal{L}_{\infty,\lambda}^{-\frac{1}{2}}\mathcal{L}_\infty^r\right\| \leq \kappa^{r-1/2}$. For $r < 1/2$ we have

$$\left\|\mathcal{L}_{\infty,\lambda}^{-\frac{1}{2}}\mathcal{L}_\infty^r\right\| = \max_{i\in\mathbb{N}}\left|\frac{\mu_i^r}{(\mu_i+\lambda)^{1/2}}\right| \leq \lambda^{r-1/2}$$

since $\left|\frac{\mu_i^r}{(\mu_i+\lambda)^{1/2}}\right| \leq \begin{cases} \lambda^{r-1/2} & \lambda \geq \mu_i \\ \mu_i^{r-1/2} & \mu_i \geq \lambda \end{cases}$
$\leq \lambda^{r-1/2}$. ■

Proposition 38 *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

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

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

Proof See for example Aleksandrov and Peller (2009), Blanchard and Mücke (2017) (Proposition B.1.) ■

E.2 Concentration Inequalities

In this subsection we generalize Bernstein type inequalities for double sums and apply them to our operators from section A. The first inequality 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 39 *Let $\xi_1, \cdots, \xi_m$ be a sequence of independently and identically distributed selfadjoint Hilbert-Schmidt operators on a separable Hilbert space. Assume that $\|\xi_1\| \leq B/2$ almost surely for some $B > 0$. Let $\mathcal{V}$ be a positive trace-class operator such that $\mathbb{E}\left[\xi_1^2\right] \preccurlyeq \mathcal{V}$. Then with probability at least $1 - \delta, (\delta \in]0, 1[)$, there holds*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \xi_i - \mathbb{E}[\xi] \right\| \leq \frac{2B\beta}{3m} + \sqrt{\frac{2\|\mathcal{V}\|\beta}{m}}, \quad \beta = \log \frac{4 \operatorname{tr} \mathcal{V}}{\|\mathcal{V}\|\delta} .$$

Corollary 40 *Let $X_1, \cdots, X_m$ and $Y_1, \ldots Y_n$ be independent sequences of independently and identically distributed random variables such that $f(X_i, Y_j)$ defines a selfadjoint Hilbert-Schmidt operator on a separable Hilbert space $\mathcal{H}$, for some function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{F}(\mathcal{H}, \mathcal{H})$. Assume that $\|f(X_1, Y_1)\| \leq B/2$ almost surely for some $B > 0$. Let for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $f(x, y)^2$ and $\mathcal{V}$ be positive trace-class operators such that $\mathbb{E}\left[f(X, Y)^2\right] \preccurlyeq \mathcal{V}$. Then with probability at least $1 - 2\delta$, there holds*

$$\left\| \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} f(X_i, Y_j) - \mathbb{E} f(X, Y) \right\| \leq \frac{2B\tilde{\beta}}{3n} + \frac{2B\beta}{3m} + \sqrt{\frac{2B\|\tilde{\mathcal{V}}\|\tilde{\beta}}{n}} + \sqrt{\frac{2\|\mathcal{V}\|\beta}{m}} ,$$

with $\beta := \log \frac{4 \operatorname{tr} \mathcal{V}}{\|\mathcal{V}\|\delta}$, $\tilde{\beta} := \log \frac{4 \operatorname{tr} \tilde{\mathcal{V}}}{\|\tilde{\mathcal{V}}\|\delta}$ and $\tilde{\mathcal{V}} := \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_Y f(X_i, Y)$.

Proof We start by proving that the following events hold true with probability $1-\delta$,

$$A := \left\{ \left\| \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} f(X_i, Y_j) - \mathbb{E}_Y f(X_i, Y) \right\| \leq \frac{2B\tilde{\beta}}{3n} + \sqrt{\frac{2B\|\tilde{\mathcal{V}}\|\tilde{\beta}}{n}} \right\}, \tag{96}$$

$$B := \left\{ \left\| \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_Y f(X_i, Y) - \mathbb{E} f(X, Y) \right\| \leq \frac{2B\beta}{3m} + \sqrt{\frac{2\|\mathcal{V}\|\beta}{m}} \right\}. \tag{97}$$

Bounding A: Using that X_i, Y_j are independent, we can condition on $\{X_i = x_i\}_{i=1}^m$ as follow,

$$\mathbb{P}(A) = \mathbb{E}_{\mu_x^{\otimes m}} \left[\mathbb{P}(A | x_1, \ldots, x_m) \right]. \tag{98}$$

We set $\xi_j := \frac{1}{m} \sum_{i=1}^m f(x_i, Y_j)$. Note that conditioned on $\{X_i = x_i\}_{i=1}^m$, where x_i is contained in the support of μ_X, we have that ξ_j are iid w.r.t. μ_Y. By assumption we have $\|\xi_j\| \leq B/2$ and for the second moment we have by Jensen inequality and by assumption, $\mathbb{E}_Y[\xi^2] \preccurlyeq \frac{1}{m} \sum_{i=1}^m \mathbb{E}_Y f(x_i, Y)^2 \preccurlyeq B \frac{1}{m} \sum_{i=1}^m \mathbb{E}_Y f(x_i, Y) := B\tilde{\mathcal{V}}$.

From Proposition 39 we therefore have, that with probability (over $\mu_Y^{\otimes n}$) at least $1-\delta$,

$$\left\| \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} f(x_i, Y_j) - \mathbb{E}_Y f(x_i, Y) \right\| \leq \frac{2B\tilde{\beta}}{3n} + \sqrt{\frac{2B\|\tilde{\mathcal{V}}\|\tilde{\beta}}{n}}.$$

From (98) we therefore have that $P(A) \geq 1-\delta$.

Bounding B: Again we set $\xi_i = \mathbb{E}_Y[f(X_i, Y)]$. By assumption we have $\|\xi\| \leq B/2$ and $\mathbb{E}_X \xi^2 \preccurlyeq \mathcal{V}$. From Proposition 39 we therefore have, that with probability at least $1-\delta$, the event B holds true.

To sum up: Using the bounds of event A and B, we therefore have with probability at least $1-2\delta$,

$$\left\| \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} f(X_i, Y_j) - \mathbb{E} f(X, Y) \right\|_{\mathcal{H}} \leq \left\| \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} f(X_i, Y_j) - \mathbb{E}_Y f(X_i, Y) \right\| \tag{99}$$

$$+ \left\| \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_Y f(X_i, Y) - \mathbb{E} f(X, Y) \right\| \tag{100}$$

$$\leq \frac{2B\tilde{\beta}}{3n} + \frac{2B\beta}{3m} + \sqrt{\frac{2B\|\tilde{\mathcal{V}}\|\tilde{\beta}}{n}} + \sqrt{\frac{2\|\mathcal{V}\|\beta}{m}}. \tag{101}$$

This proves the claim. ■

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

Proposition 41 (Bernstein Inequality) *Let $W_1, \cdots, W_n$ be i.i.d random variables in a separable Hilbert space $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}}$. Suppose that there are two positive constants B and V such that*

$$\mathbb{E}\left[\|W_1 - \mathbb{E}[W_1]\|_{\mathcal{H}}^l\right] \leq \frac{1}{2} l! B^{l-2} V^2, \quad \forall l \geq 2 \,. \tag{102}$$

Then for any $\delta \in (0,1]$, the following holds with probability at least $1-\delta$:

$$\left\| \frac{1}{n} \sum_{k=1}^{n} W_k - \mathbb{E}[W_1] \right\|_{\mathcal{H}} \leq 2 \left(\frac{B}{n} + \frac{V}{\sqrt{n}} \right) \log\left(\frac{4}{\delta}\right) \,.$$

In particular, (102) *holds if*

$$\|W_1\|_{\mathcal{H}} \leq B/2 \quad \text{a.s.,} \quad \text{and} \quad \mathbb{E}\left[\|W_1\|_{\mathcal{H}}^2\right] \leq V^2 \,.$$

Corollary 42 *Let $X_1, \cdots, X_m$ and $Y_1, \ldots Y_n$ be independent sequences of independently and identically distributed random variables such that $f(X_i, Y_j)$ is contained in a separable Hilbert space $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}}$, for some function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}$. Suppose that there are two positive constants B and V such that*

$$\|f(X_1, Y_1)\|_{\mathcal{H}} \leq B/2 \quad \text{a.s.,} \quad \text{and} \quad \mathbb{E}\left[\|f(X_1, Y_1)\|_{\mathcal{H}}^2\right] \leq V^2 \,.$$

Then for any $\delta \in (0,1]$, the following holds with probability at least $1-\delta$:

$$\left\| \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} f(X_i, Y_j) - \mathbb{E} f(X, Y) \right\|_{\mathcal{H}} \leq 2 \left(\frac{B}{m} + \frac{V}{\sqrt{m}} + \frac{B}{n} + \sqrt{\frac{Z}{n}} \right) \log\left(\frac{12}{\delta}\right) \,,$$

with $Z := \left(\frac{B^2}{m} + \frac{BV}{\sqrt{m}}\right) \log\left(\frac{12}{\delta}\right) + V^2$.

Proof We start by proving that the following events hold true with high probability,

$$A := \left\{ \left| \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_Y[\|f(X_i, Y)\|_{\mathcal{H}}^2] - \mathbb{E}[\|f(X, Y)\|_{\mathcal{H}}^2] \right| \leq \left(\frac{B^2}{m} + \frac{BV}{\sqrt{m}} \right) \log\left(\frac{4}{\delta}\right) \right\}, \tag{103}$$

$$B := \left\{ \left\| \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} f(X_i, Y_j) - \mathbb{E}_Y f(X_i, Y) \right\|_{\mathcal{H}} \leq 2 \left(\frac{B}{n} + \sqrt{\frac{\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_Y[\|f(X_i, Y)\|_{\mathcal{H}}^2]}{n}} \right) \log\left(\frac{4}{\delta}\right) \right\}, \tag{104}$$

$$C := \left\{ \left\| \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} f(X_i, Y_j) - \mathbb{E}_Y f(X_i, Y) \right\|_{\mathcal{H}} \leq 2 \left(\frac{B}{n} + \sqrt{\frac{Z}{n}} \right) \log\left(\frac{4}{\delta}\right) \right\}, \tag{105}$$

$$D := \left\{ \left\| \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_Y f(X_i, Y) - \mathbb{E} f(X, Y) \right\|_{\mathcal{H}} \leq 2 \left(\frac{B}{m} + \frac{V}{\sqrt{m}} \right) \log\left(\frac{4}{\delta}\right) \right\}, \tag{106}$$

with $Z := \left(\frac{B^2}{m} + \frac{BV}{\sqrt{m}}\right)\log\left(\frac{4}{\delta}\right) + V^2$.

Bounding A: Set $W_i = \mathbb{E}_Y[\|f(X_i,Y)\|_{\mathcal{H}}^2]$. Then by assumption $|W_i| \leq B^2/4$ and $\mathbb{E}[|W_i|^2] \leq \frac{B^2V^2}{4}$. From Proposition 41 we therefore have with probability at least $1-\delta$,

$$\left|\frac{1}{m}\sum_{i=1}^m \mathbb{E}_Y[\|f(X_i,Y)\|_{\mathcal{H}}^2] - \mathbb{E}[\|f(X,Y)\|_{\mathcal{H}}^2]\right| \leq \left(\frac{B^2}{m} + \frac{BV}{\sqrt{m}}\right)\log\left(\frac{4}{\delta}\right).$$

Bounding B: Using that X_i, Y_j are independent, we can condition on $\{X_i = x_i\}_{i=1}^m$ as follow,

$$\mathbb{P}(B) = \mathbb{E}_{\mu_x^{\otimes m}}\left[\mathbb{P}(B|x_1,\ldots,x_m)\right]. \tag{107}$$

We set $W_j := \frac{1}{m}\sum_{i=1}^m f(x_i,Y_j)$. Note that conditioned on $\{X_i = x_i\}_{i=1}^m$ we have that W_j are iid w.r.t. μ_Y. By assumption we have $\|W_j\|_{\mathcal{H}} \leq B/2$ and for the second moment we have

$$\mathbb{E}_Y[\|W\|_{\mathcal{H}}^2] \leq \frac{1}{m}\sum_{i=1}^m \mathbb{E}_Y[\|f(x_i,Y)\|_{\mathcal{H}}^2].$$

From Proposition 41 we therefore have, that with probability (over $\mu_y^{\otimes n}$) at least $1-\delta$,

$$\left\|\frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n f(x_i,Y_j) - \mathbb{E}_Y f(x_i,Y)\right\|_{\mathcal{H}} \leq 2\left(\frac{B}{n} + \sqrt{\frac{\frac{1}{m}\sum_{i=1}^m \mathbb{E}_Y[\|f(x_i,Y)\|_{\mathcal{H}}^2]}{n}}\right)\log\left(\frac{4}{\delta}\right).$$

From (107) we therefore have that $P(B) \geq 1-\delta$.

Bounding C: Note that $P(C|A,B) = 1$, since conditioned on A we have by assumption,

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m \mathbb{E}_Y[\|f(X_i,Y)\|_{\mathcal{H}}^2] &\leq \left(\frac{B^2}{m} + \frac{BV}{\sqrt{m}}\right)\log\left(\frac{4}{\delta}\right) + \mathbb{E}[\|f(X,Y)\|_{\mathcal{H}}^2] \\ &\leq \left(\frac{B^2}{m} + \frac{BV}{\sqrt{m}}\right)\log\left(\frac{4}{\delta}\right) + V^2,\end{aligned}$$

and plugging this inequality into the bound of B, leads to the bound of C. By Proposition 34 we therefore have, $P(C) \geq 1-2\delta$.

Bounding D: Again we set $W_i = \mathbb{E}_Y[\|f(X_i,Y)\|_{\mathcal{H}}]$ and obtain by Proposition 41 with probability at least $1-\delta$,

$$\left\|\frac{1}{m}\sum_{i=1}^m \mathbb{E}_Y f(X_i,Y) - \mathbb{E}f(X,Y)\right\|_{\mathcal{H}} \leq 2\left(\frac{B}{m} + \frac{V}{\sqrt{m}}\right)\log\left(\frac{4}{\delta}\right).$$

To sum up: Using the bounds of event C and D, we therefore have with probability at least $1-3\delta$,

$$\left\|\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} f(X_i,Y_j)-\mathbb{E}f(X,Y)\right\|_{\mathcal{H}} \leq \left\|\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} f(X_i,Y_j)-\mathbb{E}_Y f(X_i,Y)\right\|_{\mathcal{H}} \tag{108}$$

$$+\left\|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_Y f(X_i,Y)-\mathbb{E}f(X,Y)\right\|_{\mathcal{H}} \tag{109}$$

$$\leq 2\left(\frac{B}{m}+\frac{V}{\sqrt{m}}+\frac{B}{n}+\sqrt{\frac{Z}{n}}\right)\log\left(\frac{4}{\delta}\right). \tag{110}$$

This proves the claim. ■

Proposition 43 (Rudi and Rosasco (2017) (Lemma 9)) *For any $M \geq 8\kappa^4\|\mathcal{L}_\infty\|^{-1}\log^2\frac{2}{\delta}$ we have with probability at least $1-\delta$*

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

Proposition 44 (Nguyen and Mücke (2023) Proposition A.18.) *For any $\lambda > 0$ with $M \geq \frac{8\tilde{d}\kappa^2 C_{\lambda,\delta,\kappa}}{\lambda}$ with $C_{\lambda,\delta,\kappa} := \log\frac{80\kappa^4}{\|\mathcal{L}_\infty\|\lambda\delta}$ we have with probability at least $1-\delta$*

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

Proposition 45 *For any $\lambda \in (0,1]$, any $M \geq 8\kappa^4\|\mathcal{L}_\infty\|^{-1}\log^2\frac{2}{\delta}$ and n_u, $n_x \geq \frac{72\kappa^4 C_{\lambda,\delta,\kappa}(1+\|\mathcal{L}_\infty\|)}{\lambda}$ with $C_{\lambda,\delta,\kappa} := \log\frac{80\kappa^4}{\|\mathcal{L}_\infty\|\lambda\delta}$, define the following events,*

$$E_1 = \left\{\left\|\mathcal{C}_{M,\lambda}^{-\frac{1}{2}}\left(\widehat{\mathcal{C}}_M-\mathcal{C}_M\right)\mathcal{C}_{M,\lambda}^{-\frac{1}{2}}\right\| \leq 3/4\right\}, \tag{111}$$

$$E_2 = \left\{\left\|\widehat{\mathcal{C}}_{M,\lambda}^{-\frac{1}{2}}\mathcal{C}_{M,\lambda}^{\frac{1}{2}}\right\| \leq 2, \quad \left\|\widehat{\mathcal{C}}_{M,\lambda}^{\frac{1}{2}}\mathcal{C}_{M,\lambda}^{-\frac{1}{2}}\right\| \leq 2\right\}. \tag{112}$$

Providing Assumption 4 we have that both events holds true with probability at least $1-3\delta$.

Proof E_1) Set $f(u,x) := \mathcal{C}_{M,\lambda}^{-\frac{1}{2}}\nabla G_{\theta_0}(u)(x)\nabla G_{\theta_0}(u)(x)^\top\mathcal{C}_{M,\lambda}^{-\frac{1}{2}}$. We have $\|f(u,x)\| \leq \kappa^2/\lambda := B$ and

$$\mathbb{E}_{\rho_x}[f(u,x)^2] \preccurlyeq \frac{\kappa^2}{\lambda}\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1} := \mathcal{V}.$$

From Proposition 40 we have with probability at least $1-\delta$,

$$\left\|\mathcal{C}_{M,\lambda}^{-\frac{1}{2}}\left(\widehat{\mathcal{C}}_M-\mathcal{C}_M\right)\mathcal{C}_{M,\lambda}^{-\frac{1}{2}}\right\| \leq \frac{2\kappa^2\tilde{\beta}}{3\lambda n_x}+\frac{2\kappa^2\beta}{3\lambda n_u}+\sqrt{\frac{2\kappa^2\|\tilde{\mathcal{V}}\|\tilde{\beta}}{\lambda n_x}}+\sqrt{\frac{2\kappa^2\beta}{\lambda n_u}}, \tag{113}$$

with $\beta := \log\frac{4\operatorname{tr}\mathcal{V}}{\|\mathcal{V}\|\delta}$, $\tilde{\beta} := \log\frac{4\operatorname{tr}\tilde{\mathcal{V}}}{\|\tilde{\mathcal{V}}\|\delta}$ and $\tilde{\mathcal{V}} := \frac{1}{n_u}\sum_{i=1}^{n_u}\mathbb{E}_{\rho_x}f(u_i,x)$ and where we used $\|\mathcal{V}\| \leq \kappa^2/\lambda$.

For β we have

$$\begin{aligned}
\beta &= \log\frac{4\operatorname{tr}\mathcal{V}}{\|\mathcal{V}\|\delta} = \log\frac{4\operatorname{Tr}\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}}{\|\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}\|\delta}\\
&= \log\frac{4\operatorname{Tr}\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}(\|\mathcal{C}_M\|+\lambda)}{\|\mathcal{C}_M\|\delta}\\
&\leq \log\frac{4\frac{\kappa^2}{\lambda}\|\mathcal{C}_M\|+4Tr(\mathcal{C}_M)}{\|\mathcal{C}_M\|\delta}\\
&\leq \log\frac{8\kappa^2}{\lambda\|\mathcal{C}_M\|\delta} \qquad (114)\\
&\leq \log\frac{16\kappa^4}{\lambda\|\mathcal{L}_\infty\|\delta} := C_{\lambda,\delta}, \qquad (115)
\end{aligned}$$

where we used for the last equation, that we have from Proposition 43,

$$\|\mathcal{C}_M\| = \|\mathcal{L}_M\| \geq \frac{1}{2}\|\mathcal{L}_\infty\|. \tag{116}$$

Now we want to further bound $\tilde{\mathcal{V}}$ and $\tilde{\beta}$.
First we set $\xi_i = \mathbb{E}_{\rho_x}f(u_i,x)$. Note that $\|\xi\| \leq \kappa^2/\lambda$ and

$$\mathbb{E}\xi^2 \preccurlyeq \mathbb{E}f(u,x)^2 \preccurlyeq \mathcal{V}.$$

From Proposition 39 and (114) we therefore have with probability at least $1-\delta$,

$$\left\|\tilde{\mathcal{V}}-\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}\right\| \leq \frac{2\kappa^2 C_{\lambda,\delta}}{3\lambda n_u}+\kappa\sqrt{\frac{2C_{\lambda,\delta}}{\lambda n_u}}.$$

Now let $n_u \geq \frac{8C_{\lambda,\delta}\kappa^2}{\lambda}$, then we have $\left\|\tilde{\mathcal{V}}-\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}\right\| \leq 1$ and therefore

$$\left\|\tilde{\mathcal{V}}\right\| \leq \left\|\tilde{\mathcal{V}}-\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}\right\| + \|\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}\| \leq 2. \tag{117}$$

If we assume $n_u \geq \frac{72\kappa^4 C_{\lambda,\delta}\|\mathcal{L}_\infty\|}{\lambda}$, we instead have

$$\left\|\tilde{\mathcal{V}}-\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}\right\| \leq \frac{\|\mathcal{L}_\infty\|}{5\kappa^2}.$$

This implies together with the reversed triangle inequality,

$$\begin{aligned}\|\tilde{\mathcal{V}}\| &\geq \|\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}\| - \left\|\tilde{\mathcal{V}} - \mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}\right\| \\ &\geq \|\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}\| - \frac{\|\mathcal{L}_\infty\|}{5\kappa^2} = \frac{\|\mathcal{C}_M\|}{\|\mathcal{C}_M\|+\lambda} - \frac{\|\mathcal{L}_\infty\|}{5\kappa^2} \\ &\geq \frac{\|\mathcal{L}_\infty\|}{4\kappa^2} - \frac{\|\mathcal{L}_\infty\|}{5\kappa^2} \geq \frac{\|\mathcal{L}_\infty\|}{20\kappa^2}, \end{aligned} \tag{118}$$

where we used (116) and $\lambda \leq 1 \leq \kappa^2$ in the last step.

Plugging (118) into $\tilde{\beta}$ gives,

$$\tilde{\beta} = \log \frac{4 \operatorname{tr} \tilde{\mathcal{V}}}{\|\tilde{\mathcal{V}}\|\delta} \leq \log \frac{80\kappa^4}{\|\mathcal{L}_\infty\|\lambda\delta} := C_{\lambda,\delta,\kappa}. \tag{119}$$

To sum up we have $\left\|\tilde{\mathcal{V}}\right\| \leq 2$ and $\tilde{\beta},\, \beta \leq C_{\lambda,\delta,\kappa}$. Plugging these bounds into (113) leads to

$$\left\|\mathcal{C}_{M,\lambda}^{-\frac{1}{2}}\left(\widehat{\mathcal{C}}_M - \mathcal{C}_M\right)\mathcal{C}_{M,\lambda}^{-\frac{1}{2}}\right\| \leq \frac{2\kappa^2 C_{\lambda,\delta,\kappa}}{3\lambda n_x} + \frac{2\kappa^2 C_{\lambda,\delta,\kappa}}{3\lambda n_u} + \sqrt{\frac{4\kappa^2 C_{\lambda,\delta,\kappa}}{\lambda n_x}} + \sqrt{\frac{2\kappa^2 C_{\lambda,\delta,\kappa}}{\lambda n_u}}. \tag{120}$$

using the assumption $n_u,\, n_x \geq \frac{72\kappa^4 C_{\lambda,\delta,\kappa}(1+\|\mathcal{L}_\infty\|)}{\lambda}$, proves the result.

E_2) The result now follows from Proposition 36 and the event E_1.

■

Related operator-theoretic techniques for regularized least-squares learning in RKHSs were developed by Lin, Guo, and Zhou (Lin et al., 2017).

Proposition 46 *For any $\lambda > 0$ define the following events,*

$$E_1 = \left\{ \left\|\mathcal{C}_{M,\lambda}^{-\frac{1}{2}}\left(\widehat{\mathcal{C}}_M - \mathcal{C}_M\right)\right\|_{HS} \right. \tag{121}$$

$$\left. \leq 2\left(\frac{\kappa^2}{\sqrt{\lambda}}\left(\frac{1}{n_u}+\frac{1}{n_x}\right) + \sqrt{25\kappa^2 \mathcal{N}_{\mathcal{L}_\infty}(\mathcal{L})}\left(\frac{1}{\sqrt{n_u}}+\frac{1}{\sqrt{n_x}}\right)\right)\log^{3/2}\left(\frac{12}{\delta}\right) \right\}, \tag{122}$$

$$E_2 = \left\{ \left\|\widehat{\mathcal{C}}_M - \mathcal{C}_M\right\|_{HS} \leq \left(2\kappa^2\left(\frac{1}{n_u}+\frac{1}{n_x}\right) + 4\kappa^2\left(\frac{1}{\sqrt{n_u}}+\frac{1}{\sqrt{n_x}}\right)\right)\log^{3/2}\left(\frac{12}{\delta}\right)\right\}. \tag{123}$$

Providing Assumption 4 we have for any $\delta \in (0,1)$ that each of the above events holds true with probability at least $1-\delta$.

Proof

E_1) Set $f(u_i, x_j) := \mathcal{C}_{M,\lambda}^{-\frac{1}{2}} \nabla G_{\theta_0}(u_i)(x_j) \nabla G_{\theta_0}(u_i)(x_j)^\top$. Note that

$$\|f(u_i, x_j)\|_{HS} \leq \kappa^2/\sqrt{\lambda} := B$$

For the second moment we have,

$$\begin{aligned}\mathbb{E}\,\|f(u,x)\|_{HS}^2 &\leq \kappa^2 Tr(\mathcal{C}_{M,\lambda}^{-1}\mathbb{E}\left[\nabla G_{\theta_0}(u)(x)\nabla G_{\theta_0}(u)(x)^\top\right])\\ &= \kappa^2 Tr(\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}) = \kappa^2\mathcal{N}_M(\lambda),\end{aligned}$$

where we used for the last equation, that $\Sigma_M = Z_M Z_M^*$ and $\mathcal{C}_M = \mathcal{Z}_M^*\mathcal{Z}_M$, so therefore

$$\mathcal{N}_M(\lambda) = \operatorname{Tr}\Sigma_M\Sigma_{M,\lambda}^{-1} = \operatorname{Tr}\mathcal{Z}_M^*\Sigma_{M,\lambda}^{-1}\mathcal{Z}_M = \operatorname{Tr}\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}. \tag{124}$$

Further we have by Proposition 44:

$$\mathcal{N}_{\mathcal{L}_M}(\lambda) \leq \left(1 + 2\log\frac{4}{\delta}\right) 4\mathcal{N}_{\mathcal{L}_\infty}(\lambda) \leq 12\log(4/\delta)\,\mathcal{N}_{\mathcal{L}_\infty}(\lambda)\,. \tag{125}$$

From Proposition 42 we therefore have for $V^2 := 12\kappa^2\log(4/\delta)\,\mathcal{N}_{\mathcal{L}_\infty}(\lambda)$

$$\left\|\mathcal{C}_{M,\lambda}^{-\frac{1}{2}}\left(\widehat{\mathcal{C}}_M - \mathcal{C}_M\right)\right\|_{HS} \leq 2\left(\frac{B}{n_u} + \frac{V}{\sqrt{n_u}} + \frac{B}{n_x} + \sqrt{\frac{Z}{n_x}}\right)\log\left(\frac{12}{\delta}\right),$$

with $Z := \left(\frac{B^2}{n_u} + \frac{BV}{\sqrt{n_u}}\right)\log\left(\frac{12}{\delta}\right) + V^2$.

Note that if we let $n_u \geq \frac{\kappa^2}{\lambda}\log^2(12/\delta)$ we further obtain $Z \leq 25\kappa^2\log(4/\delta)\,\mathcal{N}_{\mathcal{L}_\infty}(\lambda)$, where we used that $\log(4/\delta)\,\mathcal{N}_{\mathcal{L}_\infty}(\lambda) \geq 1$. Therefore we have

$$\begin{aligned}&\left\|\mathcal{C}_{M,\lambda}^{-\frac{1}{2}}\left(\widehat{\mathcal{C}}_M - \mathcal{C}_M\right)\right\|_{HS}\\ &\leq 2\left(\frac{\kappa^2}{\sqrt{\lambda}}\left(\frac{1}{n_u} + \frac{1}{n_x}\right) + \sqrt{25\kappa^2\mathcal{N}_{\mathcal{L}_\infty}(\lambda)}\left(\frac{1}{\sqrt{n_u}} + \frac{1}{\sqrt{n_x}}\right)\right)\log^{3/2}\left(\frac{12}{\delta}\right)\,.\end{aligned}$$

E_2) Set $f(u_i, x_j) := \nabla G_{\theta_0}(u_i)(x_j)\nabla G_{\theta_0}(u_i)(x_j)^\top$. Note that

$$\|f(u_i,x_j)\|_{HS} \leq \kappa^2 =: B$$

For the second moment we have,

$$\mathbb{E}\,\|f(u,x)\|_{HS}^2 \leq \kappa^4 =: V^2$$

From Proposition 42 we therefore have

$$\begin{aligned}\left\|\widehat{\mathcal{C}}_M - \mathcal{C}_M\right\|_{HS} &\leq 2\left(\frac{\kappa^2}{n_u} + \frac{\kappa^2}{\sqrt{n_u}} + \frac{\kappa^2}{n_x} + \sqrt{\frac{Z}{n_x}}\right)\log\left(\frac{12}{\delta}\right)\\ &\leq \left(2\kappa^2\left(\frac{1}{n_u} + \frac{1}{n_x}\right) + 4\kappa^2\left(\frac{1}{\sqrt{n_u}} + \frac{1}{\sqrt{n_x}}\right)\right)\log^{3/2}\left(\frac{12}{\delta}\right),\end{aligned}$$

with $Z := \left(\frac{\kappa^4}{n_u} + \frac{\kappa^4}{\sqrt{n_u}}\right)\log\left(\frac{12}{\delta}\right) + \kappa^4$.

■

Proposition 47 *For any $\lambda \in (0,1]$, any $M \geq 8\kappa^4 \|\mathcal{L}_\infty\|^{-1} \log^2 \frac{2}{\delta}$ and n_u, $n_x \geq \frac{72\kappa^4 C_{\lambda,\delta,\kappa}(1+\|\mathcal{L}_\infty\|)}{\lambda}$ with $C_{\lambda,\delta,\kappa} := \log \frac{80\kappa^4}{\|\mathcal{L}_\infty\|\lambda\delta}$, we have that with probability at least $1-4\delta$,*

$$\left\|\widehat{\mathcal{C}}_{M,\lambda}^{-1}\mathcal{C}_{M,\lambda}^{1}\right\| \leq 2\left(\frac{\kappa^2}{\lambda}\left(\frac{1}{n_u}+\frac{1}{n_x}\right)+5\kappa\sqrt{\mathcal{N}_{\mathcal{L}_\infty}(\lambda)}\left(\frac{1}{\sqrt{\lambda n_u}}+\frac{1}{\sqrt{\lambda n_x}}\right)\right)\log^{3/2}\left(\frac{12}{\delta}\right). \quad (126)$$

If in addition n_u, $n_x \geq \frac{400\kappa^2\mathcal{N}_{\mathcal{L}_\infty}(\lambda)}{\lambda}\log^3\left(\frac{12}{\delta}\right)$

$$\left\|\widehat{\mathcal{C}}_{M,\lambda}^{-1}\mathcal{C}_{M,\lambda}\right\| \leq 2.$$

Proof Using Proposition 46 we obtain with probability at least $1-3\delta$,

$$\begin{aligned}
&\left\|\widehat{\mathcal{C}}_{M,\lambda}^{-1}\mathcal{C}_{M,\lambda}\right\| \\
&\leq \left\|\widehat{\mathcal{C}}_{M,\lambda}^{-1}\left(\widehat{\mathcal{C}}_M-\mathcal{C}_M\right)\right\|_{HS}+1 \\
&\leq \frac{1}{\sqrt{\lambda}}\left\|\widehat{\mathcal{C}}_{M,\lambda}^{-\frac{1}{2}}\mathcal{C}_{M,\lambda}^{\frac{1}{2}}\right\|\left\|\mathcal{C}_{M,\lambda}^{-\frac{1}{2}}\left(\widehat{\mathcal{C}}_M-\mathcal{C}_M\right)\right\|_{HS}+1 \\
&\leq \left\|\widehat{\mathcal{C}}_{M,\lambda}^{-\frac{1}{2}}\mathcal{C}_{M,\lambda}^{\frac{1}{2}}\right\|\left(\frac{\kappa^2}{\sqrt{\lambda}}\left(\frac{1}{n_u}+\frac{1}{n_x}\right)+5\sqrt{\kappa^2\mathcal{N}_{\mathcal{L}_\infty}(\lambda)}\left(\frac{1}{\sqrt{\lambda n_u}}+\frac{1}{\sqrt{\lambda n_x}}\right)\right)\log^{3/2}\left(\frac{12}{\delta}\right)+1.
\end{aligned} \quad (127)$$

From Proposition 45 we have with probability at least $1-\delta$,

$$\left\|\widehat{\mathcal{C}}_{M,\lambda}^{-\frac{1}{2}}\mathcal{C}_{M,\lambda}^{\frac{1}{2}}\right\| \leq 2 \quad (128)$$

and therefore

$$\left\|\widehat{\mathcal{C}}_{M,\lambda}^{-1}\mathcal{C}_{M,\lambda}^{1}\right\| \leq 2\left(\frac{\kappa^2}{\lambda}\left(\frac{1}{n_u}+\frac{1}{n_x}\right)+5\sqrt{\kappa^2\mathcal{N}_{\mathcal{L}_\infty}(\lambda)}\left(\frac{1}{\sqrt{\lambda n_u}}+\frac{1}{\sqrt{\lambda n_x}}\right)\right)\log^{3/2}\left(\frac{12}{\delta}\right). \quad (129)$$

The above inequality therefore holds if we condition on the events ((127), (128)). Using Proposition 34 now shows that the above inequality holds with probability at least $1-4\delta$.

If in addition n_u, $n_x \geq \frac{400\kappa^2\mathcal{N}_{\mathcal{L}_\infty}(\lambda)}{\lambda}\log^3\left(\frac{12}{\delta}\right)$ we further obtain

$$\left\|\widehat{\mathcal{C}}_{M,\lambda}^{-1}\mathcal{C}_{M,\lambda}\right\| \leq 2.$$

■

Proposition 48 *Let $M \geq \frac{8\tilde{d}\kappa^2 C_{\lambda,\delta,\kappa}}{\lambda}$, $n_u \geq \frac{\kappa^2}{\lambda}\log^2(12/\delta)$ with $C_{\lambda,\delta,\kappa} := \log\frac{80\kappa^4}{\|\mathcal{L}_\infty\|\lambda\delta}$. We have with probability at least $1-\delta$ that the following event holds true:*

$$\begin{aligned}
&\left\|\mathcal{C}_{M,\lambda}^{-1/2}\left(\widehat{\mathcal{Z}}_M^*\mathbf{v}-\mathcal{Z}_M^*G^*\right)\right\| \\
&\leq 2\left(\frac{\kappa}{\sqrt{\lambda}}\left(\frac{1}{n_u}+\frac{1}{n_x}\right)+5\sqrt{\mathcal{N}_{\mathcal{L}_\infty}(\lambda)}\left(\frac{1}{\sqrt{n_u}}+\frac{1}{\sqrt{n_x}}\right)\right)\log^{3/2}\left(\frac{12}{\delta}\right).
\end{aligned}$$

Proof [Proposition 48]
By (24) we have

$$\left\|\mathcal{C}_{M,\lambda}^{-1/2}\left(\widehat{\mathcal{Z}}_M^*\mathbf{v}-\mathcal{Z}_M^*G^*\right)\right\|_2=\left\|\frac{1}{n_u n_x}\sum_{i=1}^{n_u}\sum_{j=1}^{n_x}f(z_i,x_i)-\mathbb{E}f(z,x)\right\|_2$$

where $f(z_i,x_j):=\mathcal{C}_{M,\lambda}^{-1/2}\nabla G_{\theta_0}(u_i)(x_j)v_i(x_j)$ and $z_i=(u_i,v_i)$.
Note that by Assumption 4 we have

$$\|f(z,x)\|_2\leq\left\|\mathcal{C}_{M,\lambda}^{-1/2}\nabla G_{\theta_0}(u)(x)\right\|_2\leq\frac{1}{\sqrt{\lambda}}\sup_{x\in\mathcal{X},u\in\mathcal{U}}\|\nabla G_{\theta_0}(u)(x)\|_2\leq\frac{\kappa}{\sqrt{\lambda}}$$

and

$$\mathbb{E}\|f(z,x)\|_2^2\leq\int_{\mathcal{U}\times\mathcal{X}}\left\|\mathcal{C}_{M,\lambda}^{-1/2}\nabla G_{\theta_0}(u)(x)\right\|_2^2:=\nu.$$

Now we need to prove that $\nu=\mathcal{N}_M(\lambda)$. Note that we have $\Sigma_M=Z_MZ_M^*$ and $\mathcal{C}_M=\mathcal{Z}_M^*\mathcal{Z}_M$, so

$$\mathcal{N}_M(\lambda)=\operatorname{Tr}\Sigma_M\Sigma_{M,\lambda}^{-1}=\operatorname{Tr}\mathcal{Z}_M^*\Sigma_{M,\lambda}^{-1}\mathcal{Z}_M=\operatorname{Tr}\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1},$$

since $\mathcal{Z}_M^*\Sigma_{M,\lambda}^{-1}\mathcal{Z}_M=\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}$. By the the ciclicity of the trace and the definition of $\mathcal{C}_M$, we have

$$\operatorname{Tr}\mathcal{C}_M\mathcal{C}_{M,\lambda}^{-1}=\int_{\mathcal{X}\times\mathcal{U}}\operatorname{Tr}\left(\nabla G_{\theta_0}(u)(x)\nabla G_{\theta_0}(u)(x)^\top\mathcal{C}_{M,\lambda}^{-1}\right)=\int_{\mathcal{X}\times\mathcal{U}}\left\|\mathcal{C}_{M,\lambda}^{-1/2}\nabla G_{\theta_0}(u)(x)\right\|_2^2=\nu.$$

Further we have by Proposition 44:

$$\mathcal{N}_{\mathcal{L}_M}(\lambda)\leq\left(1+2\log\frac{4}{\delta}\right)4\mathcal{N}_{\mathcal{L}_\infty}(\lambda)\leq12\log(4/\delta)\,\mathcal{N}_{\mathcal{L}_\infty}(\lambda)\,. \tag{130}$$

So by setting $B=\frac{\kappa}{\sqrt{\lambda}}$ and $V^2=12\log(4/\delta)\,\mathcal{N}_{\mathcal{L}_\infty}(\lambda)$, it follows from Proposition 42, that with probability at least $1-\delta$,

$$\begin{aligned}&\left\|\mathcal{C}_{M,\lambda}^{-1/2}\left(\widehat{\mathcal{Z}}_M^*\mathbf{v}-\mathcal{Z}_M^*G^*\right)\right\|\\&\leq2\left(\frac{B}{n_u}+\frac{V}{\sqrt{n_u}}+\frac{B}{n_x}+\sqrt{\frac{Z}{n_x}}\right)\log\left(\frac{12}{\delta}\right)\,,\end{aligned}$$

with $Z:=\left(\frac{B^2}{n_u}+\frac{BV}{\sqrt{n_u}}\right)\log\left(\frac{12}{\delta}\right)+V^2$. Note that if we let $n_u\geq\frac{\kappa^2}{\lambda}\log^2(12/\delta)$ we further obtain $Z\leq25\log(4/\delta)\,\mathcal{N}_{\mathcal{L}_\infty}(\lambda)$, where we used that $\log(4/\delta)\,\mathcal{N}_{\mathcal{L}_\infty}(\lambda)\geq1$. Therefore we have

$$\begin{aligned}&\left\|\mathcal{C}_{M,\lambda}^{-1/2}\left(\widehat{\mathcal{Z}}_M^*\mathbf{v}-\mathcal{Z}_M^*G^*\right)\right\|\\&\leq2\left(\frac{\kappa}{\sqrt{\lambda}}\left(\frac{1}{n_u}+\frac{1}{n_x}\right)+\sqrt{25\mathcal{N}_{\mathcal{L}_\infty}(\lambda)}\left(\frac{1}{\sqrt{n_u}}+\frac{1}{\sqrt{n_x}}\right)\right)\log^{3/2}\left(\frac{12}{\delta}\right)\,.\end{aligned}$$

■

Proposition 49 *Given that $\theta_t \in Q_{\theta_0}(R)$ and Assumption 1, we have with probability at least $1-\delta$,*

$$\left\| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} \nabla G_{\theta_0}(u_i)(x_j) G_{\theta_t}(u_i)(x_j) - \mathbb{E}_{\rho_x} \nabla G_{\theta_0}(u_i)(x) G_{\theta_t}(u_i)(x) \right\|_\Theta$$
$$\leq \frac{4\kappa^2}{\sqrt{n_x}} \log\left(\frac{4}{\delta}\right) \|\theta_t - \theta_0\|_\Theta + \frac{2\kappa C_{\nabla g}(R)}{\sqrt{M}} \|\theta_t - \theta_0\|_\Theta^2.$$

Proof [Proposition 49] Using a Taylor expansion in θ_0 we have

$$G_{\theta_t}(u)(x)(u)(x) = \langle \nabla G_{\theta_0}(u)(x), \theta_t - \theta_0 \rangle_\Theta + r_{(\theta_t,\theta_0)}(u)(x).$$

Therefore

$$\left\| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} \nabla G_{\theta_0}(u_i)(x_j) G_{\theta_t}(u_i)(x_j) - \mathbb{E}_{\rho_x} \nabla G_{\theta_0}(u_i)(x) G_{\theta_t}(u_i)(x) \right\|_\Theta \tag{131}$$
$$\leq \left\| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} \nabla G_{\theta_0}(u_i)(x_j) \langle \nabla G_{\theta_0}(u_i)(x_j), \theta_t - \theta_0 \rangle_\Theta - \mathbb{E}_{\rho_x} \nabla G_{\theta_0}(u_i)(x) \langle \nabla G_{\theta_0}(u_i)(x), \theta_t - \theta_0 \rangle_\Theta \right\|_\Theta \tag{132}$$
$$+ \left\| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} \nabla G_{\theta_0}(u_i)(x_j) r_{(\theta_t,\theta_0)}(u_i)(x_j) - \mathbb{E}_{\rho_x} \nabla G_{\theta_0}(u_i)(x) r_{(\theta_t,\theta_0)}(u_i)(x) \right\|_\Theta \tag{133}$$
$$:= I + II. \tag{134}$$

For I we have

$$I \leq \left\| \frac{1}{n_x} \sum_{j=1}^{n_x} W_j - \mathbb{E}_{\rho_x} W_j \right\|_{HS} \|\theta_t - \theta_0\|_\Theta$$

with $W_j := \frac{1}{n_u} \sum_{i=1}^{n_u} \nabla G_{\theta_0}(u_i)(x_j) \nabla G_{\theta_0}(u_i)(x_j)^T$. Note that $\|W\|_{HS} \leq \kappa^2$, $\|W\|_{HS}^2 \leq \kappa^4$. Therefore we have from Proposition 41 with probability at least $1-\delta$,

$$I \leq \frac{4\kappa^2}{\sqrt{n_x}} \log\left(\frac{4}{\delta}\right) \|\theta_t - \theta_0\|_\Theta .$$

For II we have from Proposition 25,

$$II \leq 2\kappa \frac{C_{\nabla g}(R)}{\sqrt{M}} \|\theta_t - \theta_0\|_\Theta^2.$$

■

Proposition 50 *Let $T \leq C n_u^{\frac{1}{2r+b}}$, $2r+b>1$, $n_x \geq t^{2\max\{0,\frac{1}{2}-r\}}$ and $M \geq M_{III}$ with M_{III} from Proposition 19 and some $C>0$, defined in the proof. With probability at least $1-\delta$, we have for $t \in [T]$,*

$$\left| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} |F_t^*(u_i)(x_j) - G^*(u_i)(x_j)| - \mathbb{E}\,|F_t^*(u)(x) - G^*(u)(x)| \right| \\ \leq \tilde{C}_{\alpha,\kappa} \left(\frac{1}{\sqrt{n_u}} + \frac{1}{\sqrt{n_x}} \right) \log^{3/2}\left(\frac{12}{\delta}\right).$$

Proof [Proposition 50] First we start with bounding F_t^* and G^*. Note that $F_t^* \in \mathcal{H}_M$, therefore there exist some θ_t^* such that $F_t(u)(x) = \langle \nabla G_{\theta_0}(u)(x), \theta_t^* \rangle$ and $\|F_t^*\|_{\mathcal{H}_M} = \|\theta_t^*\|_\Theta$ (see for example Steinwart and Christmann (2008) Theorem 4.21). Therefore we have from Lemma 21

$$|F_t(u)(x)| = |\langle \nabla G_{\theta_0}(u)(x), \theta_t^* \rangle| \leq \kappa \|F_t^*\|_{\mathcal{H}_M} \leq C'_{\kappa,\alpha}\, t^{\max\{0,\frac{1}{2}-r\}}.$$

For G^* we have from Assumption 4

$$|G^*(u)(x)| \leq \int_{\mathcal{V}} |v(x)| \rho(dv|u) \leq 1.$$

Now set $f(u_i, x_j) := |F_t^*(u_i)(x_j) - G^*(u_i)(x_j)|$.

Then we have $|f(u_i,x_j)| \leq C'_{\kappa,\alpha}\, t^{\max\{0,\frac{1}{2}-r\}} + 1 := C_{\kappa,\alpha}\, t^{\max\{0,\frac{1}{2}-r\}} := B$ and from (Nguyen and Mücke, 2023, Proposition A.1) we have (with $\lambda = (\alpha t)^{-1}$)

$$\mathbb{E} f(u,x)^2 = \|\mathcal{S}_M F_t^* - G^*\|^2_{L^2(\mu_u)} \leq C_{\alpha,r}\, t^{-2r} := V^2\,, \tag{135}$$

for some $C_{r,\alpha} < \infty$ as long as $M \geq M_{III}$, $T \leq C n_u^{\frac{1}{2r+b}}$, with C from Theorem 19 .

From Proposition 42 we therefore have,

$$\left| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} |F_t^*(u_i)(x_j) - G^*(u_i)(x_j)| - \mathbb{E}\,|F_t^*(u)(x) - G^*(u)(x)| \right| \\ \leq 2 \left(\frac{B}{n_u} + \frac{V}{\sqrt{n_u}} + \frac{B}{n_x} + \sqrt{\frac{Z}{n_x}} \right) \log\left(\frac{12}{\delta}\right)$$

with $Z := \left(\frac{B^2}{n_u} + \frac{BV}{\sqrt{n_u}} \right) \log\left(\frac{12}{\delta}\right) + V^2$. From $T \leq C_{\alpha,r} n_u^{\frac{1}{2r+b}}$ for some $C_{\alpha,r} > 0$ we further have,

$$\left| \frac{1}{n_u n_x} \sum_{i=1}^{n_u} \sum_{j=1}^{n_x} |F_t^*(u_i)(x_j) - G^*(u_i)(x_j)| - \mathbb{E}\,|F_t^*(u)(x) - G^*(u)(x)| \right| \\ \leq \hat{C}_{\alpha,\kappa} \left(\frac{t^{\max\{0,\frac{1}{2}-r\}}}{n_u} + \frac{1}{t^r \sqrt{n_u}} + \frac{t^{\max\{0,\frac{1}{2}-r\}}}{n_x} + \frac{1}{\sqrt{n_x}} \right) \log^{3/2}\left(\frac{12}{\delta}\right) \\ \leq \tilde{C}_{\alpha,\kappa} \left(\frac{1}{\sqrt{n_u}} + \frac{1}{\sqrt{n_x}} \right) \log^{3/2}\left(\frac{12}{\delta}\right).$$

■

Proof [Proposition 3]
Recall that

$$K_M(u,u') = (\Phi_u^M)^*\Phi_{u'}^M = \sum_{m=1}^{3M} \partial_m G_{\theta_0}(u) \otimes \partial_m G_{\theta_0}(u')$$
$$= \frac{1}{M}\sum_{m=1}^{M} \sigma\left(\left\langle b_m^{(0)}, J(u)\right\rangle\right) \otimes \sigma\left(\left\langle b_m^{(0)}, J(u')\right\rangle\right) +$$
$$\frac{1}{M}\sum_{m=1}^{M} \tau^2\sigma'\left(\left\langle b_m^{(0)}, J(u)\right\rangle\right) J(u) \otimes \sigma'\left(\left\langle b_m^{(0)}, J(u')\right\rangle\right) J(u'),$$

We now set

$$w_m = \sigma\left(\left\langle b_m^{(0)}, J(u)\right\rangle\right) \otimes \sigma\left(\left\langle b_m^{(0)}, J(u')\right\rangle\right) +$$
$$\tau^2\sigma'\left(\left\langle b_m^{(0)}, J(u)\right\rangle\right) J(u) \otimes \sigma'\left(\left\langle b_m^{(0)}, J(u')\right\rangle\right) J(u')\,.$$

Note that $\|w_1\|_{HS} \leq \kappa^2$and $\mathbb{E}\left[\|w_1\|_{HS}^2\right] \leq \kappa^4$. The result now follows from 41. ■

References

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

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. doi: 10.1109/TPAMI.2016.2644615.

David Benatia, Marine Carrasco, and Jean-Pierre Florens. Functional linear regression with functional response. *Journal of econometrics*, 201(2):269–291, 2017.

Alberto Bietti and Francis Bach. Deep equals shallow for relu networks in kernel regimes. *arXiv preprint arXiv:2009.14397*, 2020.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Nicolas Boullé and Alex Townsend. A mathematical guide to operator learning, 2023. URL `https://arxiv.org/abs/2312.14688`.

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

C. Carmeli, E. De Vito, and V. Umanità A. Toigo. Vector valued reproducing kernel Hilbert spaces and universality, 2008.

Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Reproducing kernel Hilbert spaces and mercer theorem, 2005.

Lin Chen and Sheng Xu. Deep neural tangent kernel and Laplace kernel have the same rkhs. *arXiv preprint arXiv:2009.10683*, 2020.

Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang. A generalized neural tangent kernel analysis for two-layer neural networks. *Advances in Neural Information Processing Systems*, 33:13363–13373, 2020.

Masoumeh Dashti and Andrew M. Stuart. The bayesian approach to inverse problems, 2015. URL `https://arxiv.org/abs/1302.6989`.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in neural information processing systems*, 33: 7710–7721, 2020.

V. Fanaskov and I. Oseledets. Spectral neural operators, 2024. URL `https://arxiv.org/abs/2205.10573`.

Manuel Febrero-Bande, Manuel Oviedo-de la Fuente, Mohammad Darbalaei, and Morteza Amini. Functional regression models with functional response: a new approach and a comparative study. *Computational Statistics*, pages 1–27, 2024.

Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Basri Ronen. On the similarity between the Laplace and neural tangent kernels. *Advances in Neural Information Processing Systems*, 33:1451–1461, 2020.

Eugene Golikov, Eduard Pokonechnyy, and Vladimir Korviakov. Neural tangent kernel: A survey. *arXiv preprint arXiv:2208.13614*, 2022.

Lukas Gonon, Arnulf Jentzen, Benno Kuckuck, Siyu Liang, Adrian Riekert, and Philippe von Wurstemberger. An overview on machine learning methods for partial differential equations: from physics informed neural networks to deep operator learning, 2024. URL `https://arxiv.org/abs/2408.13222`.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pages 1803–1810, 2012.

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

Thomas JR Hughes. *The finite element method: linear static and dynamic finite element analysis*. Courier Corporation, 2003.

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

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL `https://arxiv.org/abs/1312.6114`.

Ilja Klebanov, Ingmar Schuster, and Timothy John Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2(3):583–606, 2020.

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`.

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, 2024a. ISSN 1532-4435.

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, 2023. URL `https://api.semanticscholar.org/CorpusID:259149906`.

Nikola B Kovachki, Samuel Lanthaler, and Hrushikesh Mhaskar. Data complexity estimates for operator learning. *arXiv preprint arXiv:2405.15992*, 2024b.

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

Samuel Lanthaler. Operator learning with pca-net: upper and lower complexity bounds. *Journal of Machine Learning Research*, 24(318):1–67, 2023. URL `http://jmlr.org/papers/v24/23-0478.html`.

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.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal rates for regularized conditional mean embedding learning. *Advances in Neural Information Processing Systems*, 35:4433–4445, 2022.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Towards optimal sobolev norm rates for the vector-valued regularized least-squares algorithm. *Journal of Machine Learning Research*, 25(181):1–51, 2024.

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.

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*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

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.

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, 2020.

Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18(92):1–31, 2017.

Matthew Lowery, John Turnage, Zachary Morrow, John D. Jakeman, Akil Narayan, Shandian Zhe, and Varun Shankar. Kernel neural operators (knos) for scalable, memory-efficient, geometrically-flexible operator learning, 2024. URL https://arxiv.org/abs/2407.00809.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 03 2021. doi: 10.1038/s42256-021-00302-5.

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.

Dimitri Meunier, Zikai Shen, Mattes Mollenhauer, Arthur Gretton, and Zhu Li. Optimal rates for vector-valued spectral regularization learning algorithms. *arXiv preprint arXiv:2405.14778*, 2024.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, Cambridge, MA, 2 edition, 2018.

Mattes Mollenhauer, Nicole Mücke, and TJ Sullivan. Learning linear operators: Infinite-dimensional regression as a well-behaved non-compact inverse problem. *arXiv preprint arXiv:2211.08875*, 2022.

Jeffrey S Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2 (1):321–359, 2015.

Nicholas H. Nelsen and Andrew M. Stuart. Operator learning using random features: A tool for scientific computing. *SIAM Review*, 66(3):535–571, May 2024. ISSN 1095-7200. doi: 10.1137/24m1648703. URL http://dx.doi.org/10.1137/24M1648703.

Mike Nguyen and Nicole Mücke. Random feature approximation for general spectral methods, 2023.

Mike Nguyen and Nicole Mücke. How many neurons do we need? A refined analysis for shallow networks trained with gradient descent, 2023. URL https://arxiv.org/abs/2309.08044.

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.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in neural information processing systems*, 33:21247–21259, 2020.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes, 2018.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta numerica*, 8:143–195, 1999. URL https://doi.org/10.1017/S0962492900002919.

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.

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.

Aniruddha Rajendra Rao and Matthew Reimherr. Modern non-linear function-on-function regression. *Statistics and Computing*, 33(6):130, 2023.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017.

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.

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

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.

Zhongjie Shi, Jun Fan, Linhao Song, Ding-Xuan Zhou, and Johan A.K. Suykens. Nonlinear functional regression by functional deep neural network with kernel embedding. *Journal of Machine Learning Research*, 26(284):1–49, 2025. URL http://jmlr.org/papers/v26/23-0659.html.

Eiki Shimizu, Kenji Fukumizu, and Dino Sejdinovic. Neural-kernel conditional mean embeddings. *arXiv preprint arXiv:2403.10859*, 2024.

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

Gilbert Strang and George J. Fix. *An Analysis of the Finite Element Method*. Prentice Hall, 1973.

Albert Tarantola. Inverse problem theory and methods for model parameter estimation, 01 2005.

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

Puyu Wang, Yunwen Lei, Di Wang, Yiming Ying, and Ding-Xuan Zhou. Generalization guarantees of gradient descent for shallow neural networks. *Neural Computation*, 37(2):344–402, 2025.

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.

Xianliang Xu, Ye Li, and Zhongyi Huang. Convergence analysis of wide shallow neural operators within the framework of neural tangent kernel. *arXiv preprint arXiv:2412.05545*, 2024.

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.