# UNIFIED GENERALIZATION ANALYSIS FOR PHYSICS INFORMED NEURAL NETWORKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Physics-Informed Neural Networks (PINNs) and their variational counterparts (VPINNs) are neural networks that incorporate physical laws, making them useful for scientific problems. Existing generalization analyses for PINNs and VPINNs remain limited, often requiring restrictive assumptions such as stability conditions or linear ellipticity. In this paper, we derive generalization bounds for neural networks that involve differentiation with respect to input variables, covering PINNs and VPINNs under a unified framework. We apply Taylor expansion to represent nonlinear differential operators as linear operators on a high-dimensional space, enabling the use of Koopman-based analysis and showing that high-rank networks can generalize well even in settings involving differential operators. We also show that the nonlinearity of the differential operator exponentially enlarges the bound, highlighting its significant impact on generalization.

## 1 INTRODUCTION

Physics-Informed Neural Networks (PINNs) are a class of machine learning methods designed to solve differential equations by incorporating physical laws directly into the training process of neural networks (Raissi et al., 2019). The key idea is to train a neural network to approximate the solution of a physical system while ensuring that it satisfies the underlying equations, boundary conditions, and initial conditions. PINNs have gained attention as part of the broader field of scientific machine learning (SciML), offering a flexible framework for solving forward and inverse problems, including parameter estimation and system identification (Ren et al., 2025). They are particularly useful in areas such as fluid dynamics, solid mechanics, and diffusion processes, where traditional numerical methods may be computationally expensive or difficult to apply (Cai et al., 2021; Botarelli et al., 2025). Variational Physics-Informed Neural Networks (VPINNs) are a class of PINNs that employs a weak (variational) formulation. They are recognized as one of the more accurate and stable approaches in this family of methods (Kumar & Singh, 2026).

Generalization error analysis of neural networks has been a major topic in machine learning for understanding the performance of models. Generalization error measures the gap between the training error and the true error, and characterizes how well a model fits unseen test data. A typical approach for deriving generalization bounds is bounding some complexity of networks (Bartlett & Mendelson, 2002; Mohri et al., 2018). Classically, the VC-dimension theory (Harvey et al., 2017; Anthony & Bartlett, 2009) has been applied to show a large number of parameters makes the complexity and generalization error large. For over-parameterized regimes, norm-based bounds have been investigated (Neyshabur et al., 2015; Bartlett et al., 2017; Golowich et al., 2018; Neyshabur et al., 2018; Wei & Ma, 2019; 2020; Li et al., 2021; Ju et al., 2022; Weinan E et al., 2022). These bounds do not depend on the number of parameters explicitly. Another approach to tackle over-parameterized networks is a compression-based approach (Arora et al., 2018; Suzuki et al., 2020), which explain the generalization of networks by investigating how much the networks can be compressed. These bounds get small as the ranks of the weight matrices become small. On the other hand, Hashimoto et al. (2024; 2026) proposed Koopman-based bounds, which describes that high-rank networks can also generalize well. They represent the composition structure of neural networks using linear operators called Koopman operators and derive a generalization bound by evaluating the norm of Koopman operators. Using this approach, the determinants of weight matrices appear in the denominator of the bound, which implies high-rank weight matrices such as unitary matrices can also make the generalization error small.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The analyses in the previous paragraph are valid only for standard neural networks without differential operators since the loss function depends on the derivative of neural networks, instead of the original neural networks. Generalization error analysis for PINNs has also been investigated to clarify how the number of collocation points and the architecture of the neural network affect the performance. However, the applicability of existing analyses is limited. For example, the stability assumption (Mishra & Molinaro, 2023), the linear elliptic assumption (Xu et al., 2025) are needed for deriving the bounds. There are no unified analysis with general settings. In particular, generalization error analysis for VPINNs is extremely limited. Berrone et al. (2022) analyzed VPINNs, but the linear elliptic assumption is also needed in this case.

In this paper, we derive a generalization bound for neural networks involving differentiations of neural networks with respect to input variables, including PINNs and VPINNs. Our setting is suitable for various types of nonlinear differential equations, which can be approximated by polynomials. We apply the Taylor expansion and describe the nonlinear differential operator with a linear operator on a high dimensional space. This enables us to consider the "adjoint" of the operator and separate the model from the differential operator. Building on this decoupled structure, we extend the Koopman-based framework to settings involving differential operators, thereby establishing new generalization bounds that reveal high-rank networks can generalize well even in the presence of differential operators. In addition, the nonlinearity of the differential operator exponentially makes the bound large, which implies that the nonlinearity significantly affects the generalization property. We primarily focus on VPINNs, but the model that we consider in this paper is interpreted as an approximation of the standard PINNs. Thus, the results are expected to also be effective for the standard PINNs. In addition, we can also apply our results to other neural networks involving differentia operators, such as Hamiltonian neural networks (Greydanus et al., 2019) and Lagrangian neural networks (Cranmer et al., 2020). Practically, our bound is useful to design regularization terms. By learning neural networks so that our bound becomes small, we can obtain high test performances. This paper sheds light on the generalization property of neural networks involving derivatives in a unified manner and shows the connection to the Koopman-based analysis.

**Notations** For a Lebesgue measure space $\mathcal{X}$ and $\tilde{\mathcal{X}} \subseteq \mathbb{R}^d$, $L^2(\mathcal{X}, \tilde{\mathcal{X}})$ is the space of square integrable functions from $\mathcal{X}$ to $\tilde{\mathcal{X}}$. For Banach spaces $\mathcal{E}$ and $\mathcal{F}$, we denote by $\mathcal{B}(\mathcal{F}, \mathcal{E})$ the space of bounded linear operators from $\mathcal{F}$ to $\mathcal{E}$. If $\mathcal{D} \in \mathcal{B}(\mathcal{F}, \mathcal{E})$, we denote $\mathcal{D}(u)$ by $\mathcal{D}u$.

## 2 PRELIMINARIES

We review mathematical notions and existing results related to this paper.

### 2.1 FRÉCHET DERIVATIVE AND TAYLOR'S THEOREM FOR BANACH SPACES

We introduce the derivative of a map from a Banach space to a Banach space. Let $\mathcal{E}, \mathcal{F}$ be Banach spaces, and let $\mathcal{U} \subset \mathcal{F}$ be an open subset.

**Definition 2.1.** A mapping $\mathcal{D} : \mathcal{U} \to \mathcal{E}$ is said to be *Fréchet differentiable* at a point $u \in \mathcal{U}$ if there exists a bounded linear operator $\mathcal{D}' \in \mathcal{B}(\mathcal{F}, \mathcal{E})$ such that

$$\lim_{\|h\|_{\mathcal{F}} \to 0} \frac{\|\mathcal{D}(u+h) - \mathcal{D}(u) - \mathcal{D}'(u)h\|_{\mathcal{E}}}{\|h\|_{\mathcal{F}}} = 0.$$

The operator $\mathcal{D}'(u)$ is called the *Fréchet derivative* of $\mathcal{D}$ at $u$. We denote by $\mathcal{D}^{(i)}$, the $i$th Fréchet derivative. If $\mathcal{D}$ is Fréchet differentiable at every point in $\mathcal{U}$, then $\mathcal{D}$ is said to be Fréchet differentiable on $\mathcal{U}$.

We denote by $\mathcal{C}^r(\mathcal{U}, \mathcal{E})$ the space of functions $\mathcal{D}$ where $\mathcal{D}, \mathcal{D}', \ldots, \mathcal{D}^{(r-1)}$ are Fréchet differentiable at each $v \in \mathcal{U}$, $\mathcal{D}^{(r)} : \mathcal{U} \to \mathcal{B}(\mathcal{F}^{\otimes r}, \mathcal{E})$ is continuous. Here, $\mathcal{F}^{\otimes r} = \mathcal{F} \otimes \cdots \otimes \mathcal{F}$ ($r$ copies) and $\otimes$ is the tensor product.

Similar to the case of the standard derivative, we have the following Taylor's theorem with the Fréchet derivative.

**Proposition 2.2.** *Let $\mathcal{D} \in \mathcal{C}^r(\mathcal{U}, \mathcal{E})$, $v \in \mathcal{U}$, and $u \in \mathcal{F}$ so small that $\{v + tu \mid 0 \leq t \leq 1\} \subset \mathcal{U}$. Then, we have*

$$\mathcal{D}(v+u) = \sum_{i=0}^{r} \frac{\mathcal{D}^{(i)}(v)\, u^{\otimes i}}{i!} + S(v,u)\, u^{\otimes r},$$

*where $S(v,u) = \int_0^1 \frac{(1-t)^{r-1}}{(r-1)!}(\mathcal{D}^{(r)}(v+tu) - \mathcal{D}^{(r)}(v))\mathrm{d}t$ is the remainder.*

### 2.2 Koopman operator and Koopman-based Rademacher complexity bound

Koopman operator is a linear operator that represents the composition of nonlinear functions. Since neural networks are constructed using compositions, Koopman operators play an essential role in analyzing neural networks. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a Lebesgue measure space. Koopman operators are defined as follows.

**Definition 2.3** (Koopman operator)**.** Let $\tilde{\mathcal{X}} \subseteq \mathbb{R}^{d_1}$ and $\mathcal{X} \subseteq \mathbb{R}^{d_2}$. The *Koopman operator* $K_\sigma$ with respect to a map $\sigma : \tilde{\mathcal{X}} \to \mathcal{X}$ is a linear operator from $L^2(\mathcal{X}, \mathbb{R}^{\tilde{d}})$ to $L^2(\tilde{\mathcal{X}}, \mathbb{R}^{\tilde{d}})$ that is defined as $K_\sigma h(x) = h(\sigma(x))$ for $h \in L^2(\tilde{\mathcal{X}}, \mathbb{R}^{\tilde{d}})$.

We will consider the Koopman operators with respect to activation functions. Throughout this paper, we assume these Koopman operators are bounded. Indeed, we have the following lemma regarding the sufficient condition of the boundedness of Koopman operators.

**Lemma 2.4.** *Assume $\sigma : \tilde{\mathcal{X}} \to \mathcal{X}$ is bijective, $\sigma^{-1}$ is differentiable, and the Jacobian of $\sigma^{-1}$ is bounded in $\mathcal{X}$. Then, we have $\|K_\sigma\| \leq \sup_{x \in \mathcal{X}} |J\sigma^{-1}(x)|^{1/2}$, where $J\sigma^{-1}$ is the Jacobian of $\sigma^{-1}$. In particular, the Koopman operator $K_\sigma$ is bounded.*

The following lemma is regarding the boundedness of well-known elementwise activation functions defined as $\sigma([x_1, \ldots, x_d]) = [\tilde{\sigma}(x_1), \cdots, \tilde{\sigma}(x_d)]$ for a map $\tilde{\sigma} : \mathbb{R} \to \mathbb{R}$.

**Lemma 2.5.** *Let $\tilde{\mathcal{X}} = [a_1, b_1] \times \cdots \times [a_d, b_d] \subseteq \mathbb{R}^d$ be a bounded rectangular domain, and let $\mathcal{X} = \sigma(\tilde{\mathcal{X}})$. If $\sigma$ is the elementwise hyperbolic tangent defined as $\tilde{\sigma}(x) = \tanh(x)$, then we have $\mathcal{X} \subset [-1,1]^d$ and $\|K_\sigma\| \leq (\prod_{i=1}^d \sup_{x \in \tilde{\sigma}([a_i,b_i])} 1/(1-x^2))^{1/2}$. If $\sigma$ is the elementwise sigmoid defined as $\tilde{\sigma}(x) = 1/(1+\mathrm{e}^{-x})$, then we have $\mathcal{X} \subset [-1,1]^d$ and $\|K_\sigma\| \leq (\prod_{i=1}^d \sup_{x \in \tilde{\sigma}([a_i,b_i])} 1/(x - x^2))^{1/2}$.*

Hashimoto et al. (2024; 2026) proposed applying Koopman operators to derive Rademacher complexity bounds of neural networks. Let $d_l \in \mathbb{N}$, $D > 0$, and $\Theta_l = \{W \in \mathbb{R}^{d_l \times d_{l-1}} \mid |\det W_l^* W_l|^{-1/4} \leq D\}$. Here, $^*$ represents the adjoint. Let $\mathcal{X}_0 \subset \mathbb{R}^{d_0}$ be the input space, $W_l \sigma_{l-1}(W_{l-1} \cdots W_2 \sigma_1(W_1 \mathcal{X}_0)) \subseteq \tilde{\mathcal{X}}_l \subseteq \mathbb{R}^{d_l}$, and $\sigma_l(W_l \cdots W_2 \sigma_1(W_1 \mathcal{X}_0)) \subseteq \mathcal{X}_l \subseteq \mathbb{R}^{d_l}$ that satisfy $\mu_{\mathbb{R}^{d_l}}(\mathcal{X}_l) > 0$ and $\mu_{\mathbb{R}^{d_l}}(\tilde{\mathcal{X}}_l) > 0$. Here, $\mu_{\mathbb{R}^{d_l}}$ is the Lebesgue measure on $\mathbb{R}^{d_l}$ and $\sigma_l : \tilde{\mathcal{X}}_l \to \mathcal{X}_l$ is the activation function at layer $l$. We assume the Koopman operator $K_{\sigma_l}$ is bounded. Let $v \in L^2(\tilde{\mathcal{X}}_L, \mathcal{X}_L)$ be the final nonlinear transformation. Consider the neural network

$$u_\theta(x) = v(W_L \sigma_{L-1}(W_{L-1} \sigma_{L-2}(\cdots \sigma_1(W_1 x)))) \tag{1}$$

with learnable parameters $\theta = (W_1, \ldots, W_L)$.

Let $\alpha(h) = (\int_{W_l \mathcal{X}_{l-1}} |h(x)|^2 d\mu_{\mathcal{R}(W_l)}(x) / \int_{\tilde{\mathcal{X}}_l} |h(x)|^2 \mathrm{d}\mu_{\mathbb{R}^{d_l}}(x))^{1/2}$ for $h \in L^2(\tilde{\mathcal{X}}_l, \mathcal{X}_L)$. Let $f_l = v \circ W_L \circ \sigma_{L-1} \circ \cdots \circ W_{l+1} \circ \sigma_l$. They derived the following lemma regarding the $L^2$ norm of $u_\theta$.

**Lemma 2.6.** *We have*

$$\|u_\theta\| \leq \frac{\|v\| \prod_{l=1}^{L-1} \|K_{\sigma_l}\| \alpha(f_l)}{\prod_{l=1}^{L} |\det W_l^* W_l|^{1/4}}.$$

Let $p_x \in L^2(\mathcal{X}_0, \mathbb{R}^{d_L})$ for $x_n \in \mathcal{X}_0$, and consider a regularized model $V_\theta(x) = \int_{\mathcal{X}_0} u_\theta(y) \cdot p_x(y) \mathrm{d}y$. Let $D > 0$ and $\mathcal{V}_\Theta = \{V_\theta \mid \theta \in \Theta\}$, where $\Theta = \{(W_1, \ldots, W_L) \mid W_l \in \Theta_l\}$. Using Lemma 2.6, they derive the following bound.

**Proposition 2.7.** *Assume there exists $F > 0$ such that $\|p_{x_n}\| \leq F$. Then, the empirical Rademacher complexity of $\mathcal{V}_\Theta$ and samples $x_1, \ldots, x_N \in \mathcal{X}_0$ is bounded by*

$$\sup_{\theta \in \Theta} \frac{F\|v\| \prod_{l=1}^{L-1} \|K_{\sigma_l}\| \alpha(f_l)}{\sqrt{N} \prod_{l=1}^{L} |\det W_l^* W_l|^{1/4}}. \tag{2}$$

The right hand side of (2) includes the determinants of the weight matrices. Thus, the bounds can become smaller even if $W_l$ has large singular values. This result with Koopman operators describes why networks with high-rank weight matrices generalize well.

## 3 Problem setting

We consider a model with differentiations of a neural network with respect to inputs. Typical examples are PINNs, Hamiltonian neural networks, and Lagrangian neural networks. Let $d, \tilde{d} \in \mathbb{N}$, $\mathcal{X} \subset \mathbb{R}^d$ be a bounded domain, $C > 0$, and $\mathcal{F} := \{u \in H^s(\mathcal{X}, \mathbb{R}^{\tilde{d}}) \mid \|u\|_{H^s(\mathcal{X}, \mathbb{R}^{\tilde{d}})} \leq C\}$. Here, $H^s(\mathcal{X}, \mathbb{R}^{\tilde{d}})$ is the Sobolev space of order $s$ from $\mathcal{X}$ to $\mathbb{R}^{\tilde{d}}$. Let $\mathcal{D} : \mathcal{F} \to L^2(\mathcal{X}, \mathbb{R}^{\tilde{d}})$ be a differential operator of order $s$. Note that $\mathcal{D}$ can be nonlinear. Let $\Theta$ be a parameter space and consider a neural network $u_\theta \in \mathcal{F}$ parametrized by $\theta \in \Theta$. We denote by $\mathcal{U}_\Theta$ the class of neural networks parametrized by $\Theta$. Consider the differentiation of the neural network $\mathcal{D}(u_\theta)$ and its variational form:

$$V_\theta(x_n) = \int_{\mathcal{X}} \mathcal{D}(u_\theta)(y) \cdot p_x(y) \mathrm{d}y = \langle p_{x_n}, \mathcal{D}(u_\theta) \rangle_{L^2(\mathcal{X})} \quad (n = 1, \ldots, N),$$

where $x_1, \ldots, x_N \in \mathcal{X}$ are collocation points and for $x_n$, $p_{x_n} \in H^s(\mathcal{X}, \mathbb{R}^{\tilde{d}})$ is a nonzero test function satisfying $p_{x_n}^{(r)}(x) = 0$ for $x \in \partial\mathcal{X}$, $n = 1, \ldots, N$, $r = 1, \ldots, s$ and there exists $B > 0$ such that for $n = 1, \ldots, N$, $\|p_{x_n}\| \leq B$. We denote $\mathcal{V}_\Theta = \{V_\theta \mid u_\theta \in \mathcal{U}_\Theta\}$, the class of regularized models. Consider a $L$-Lipschitz continuous loss function $\mathcal{L}_n : \mathbb{R} \to \mathbb{R}_+$ for $n = 1, \ldots, N$ and minimize the mean loss:

$$\frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_n \Big( V_\theta(x_n) \Big). \tag{3}$$

Typical examples of the test function $q_{x_n}$ and the loss function $\mathcal{L}_n$ are as follows.

**Example 3.1.** Let $q_{x_n}(y) = \mathrm{e}^{-1/(1-c^2\|x_n - y\|^2)}$ for $\|x_n - y\| \leq 1/c$ and $q_{x_n}(y) = 0$ for $\|x_n - y\| > 1/c$. Let $\tilde{d} = 1$ and $p_{x_n}(y) = q_{x_n}(y)/(\int_{\mathcal{X}} q_{x_n}(z) \mathrm{d}z)$. Then, $p_{x_n}$ is infinitely differentiable. Consider a PDE $\mathcal{D}(u) = f$ for $f \in L^2(\mathcal{X}, \mathbb{R}^{\tilde{d}})$. Let $\mathcal{L}_n(z) = |z - \int_{\mathcal{X}} f(y) p_{x_n}(y) \mathrm{d}y|^2$. Assume for some $D > 0$ and $E > 0$, $|\int_{\mathcal{X}} \mathcal{D}(u_\theta)(y) p_{x_n}(y) \mathrm{d}y| \leq D$ and $|\int_{\mathcal{X}} f(y) p_{x_n}(y) \mathrm{d}y| \leq E$. Then, $\mathcal{L}_n$ is $2(D + E)$-Lipchitz. Note that since $p_{x_n}$ goes to the delta function as $c \to \infty$, we have

$$\lim_{c \to \infty} \int_{\mathcal{X}} \mathcal{D}(u_\theta)(y) p_{x_n}(y) \mathrm{d}y = \mathcal{D}(u_\theta)(x_n). \tag{4}$$

In this case, Eq. (3) is the loss for the variational form such as VPINNs, but as $c \to \infty$, it becomes the loss for the strong form, such as PINNs for the interior collocation point $x_n$.

## 4 Generalization bounds

Our goal is to derive Koopman-based generalization bounds of the variational form $V_\theta$. Since we have the differential operator $\mathcal{D}$ in $V_\theta$ and the loss function depends on the derivative of the neural network, not the neural network itself, the derivation is not straightforward. This is the main difference between the existing setting in Proposition 2.7 and our setting. To tackle this situation, we first consider a simple case with linear differential operator $\mathcal{D}$. In this case, we can consider the adjoint of $\mathcal{D}$. We then consider nonlinear cases. Applying the Taylor expansion, we can regard $\mathcal{D}$ as a linear operator on a higher dimensional space, which enables us to consider a similar notion of the adjoint of $\mathcal{D}$.

Since the loss function $\mathcal{L}_n$ is Lipchitz continuous, we can derive a generalization bound by deriving a Rademacher complexity bound for the function class $\{\int_{\mathcal{X}} \mathcal{D}(u_\theta)(y) \cdot p_{x_n}(y) \mathrm{d}y \mid u_\theta \in \mathcal{U}_\Theta\}$ (Mohri

et al., 2018, Theorem 3.5). Let $\Omega$ be a probability space equipped with a probability measure $P$. Let $\epsilon_1, \ldots, \epsilon_N : \Omega \to \mathbb{C}$ be i.i.d. Rademacher variables (random variables following the uniform distribution on $\{-1, 1\}$). For a measurable function $\epsilon : \Omega \to \mathbb{C}$, we denote by $\mathrm{E}[\epsilon]$ the integral $\int_\Omega \epsilon(\omega) \mathrm{d}P(\omega)$. The empirical Rademacher complexity $\hat{R}(\mathcal{U}, x_1, \ldots, x_N)$ of a function class $\mathcal{U}$ is defined as $\hat{R}(\mathcal{U}, x_1, \ldots, x_N) = \mathrm{E}[\sup_{u \in \mathcal{U}} \sum_{n=1}^N u(x_n)\epsilon_n]/N$.

### 4.1 Linear case

Assume $\mathcal{D}$ is bounded and linear. Then, by the Cauchy Schwarz inequality, we have

$$\sum_{n=1}^N \langle p_{x_n}, \mathcal{D}(u_\theta)\rangle = \left\langle \sum_{n=1}^N \mathcal{D}^* p_{x_n}, u_\theta \right\rangle \leq \|u_\theta\| \left\| \sum_{n=1}^N \mathcal{D}^* p_{x_n} \right\|,$$

where $^*$ means the adjoint. This transformation enables us to force the action of $\mathcal{D}$ to $p_{x_n}$, instead of $u_\theta$, and to apply the same analysis as the neural network $u_\theta$ without $\mathcal{D}$. Indeed, we have the following theorem. The proof is documented in Appendix A

**Theorem 4.1.** *Let $\mathbb{D}_{x_n} = \mathcal{D}^* p_{x_n}$. Assume there exists $F > 0$ such that for any $n = 1, \ldots, N$, $\|\mathbb{D}_{x_n}\| \leq F$. Then, we have*

$$\hat{R}(\mathcal{V}_\Theta, x_1, \ldots, x_N) \leq \frac{F}{\sqrt{N}} \sup_{u_\theta \in \mathcal{U}_\Theta} \|u_\theta\|.$$

### 4.2 Polynomial case

When $\mathcal{D}$ is nonlinear, we do not have $\mathcal{D}^*$. However, if $\mathcal{D}$ is approximated by a polynomial, then we can construct a linear operator on a higher dimensional space. Then, we have the adjoint of the linear operator.

Let $\tilde{\mathcal{D}} \in \mathcal{C}^r(\mathcal{F}, L^2(\mathcal{X}, \mathbb{R}^d))$ be a differential operator whose remainder $S(0, u_\theta)$ of $r$-order Taylor approximation at 0 is bounded as $\|S(0, u_\theta)\| \leq \epsilon$ for some $\epsilon > 0$. Assume $\mathcal{D}$ is in the following form:

$$\mathcal{D}(u_\theta) = g \circ u_\theta + \tilde{\mathcal{D}}(u_\theta) \tag{5}$$

with $g : \mathbb{R}^{\tilde{d}} \to \mathbb{R}^{\tilde{d}}$. In many real-world physical systems, such as fluid flows and diffusion-dominated problems, terms in $\mathcal{D}$ with differentiations are polynomials and non-polynomial terms only depend on the solution $u$, not the derivatives of $u$. In this case, we can regard $\mathcal{D}$ as a linear operator on a higher dimensional space and we have

$$\begin{aligned}
\sum_{n=1}^N \langle p_{x_n}, \mathcal{D}(u_\theta)\rangle &= \sum_{n=1}^N \left( \langle p_{x_n}, g \circ u_\theta \rangle + \sum_{i=0}^r \frac{1}{i!} \langle p_{x_n}, \tilde{\mathcal{D}}^{(i)}(0) u_\theta^{\otimes i} \rangle + \langle p_{x_n}, S(0, u_\theta) u_\theta^{\otimes r} \rangle \right) \\
&= \left\langle \sum_{n=1}^N \mathbb{D}_{x_n}, \bigoplus_{i=0}^r u_\theta^{\otimes i} \oplus (g \circ u_\theta) \right\rangle + \left\langle \sum_{n=1}^N p_{x_n}, S(0, u_\theta) u_\theta^{\otimes r} \right\rangle \\
&\leq \left\| \bigoplus_{i=0}^r u_\theta^{\otimes i} \oplus (g \circ u_\theta) \right\| \left\| \sum_{n=1}^N \mathbb{D}_{x_n} \right\| + \|S(0, u_\theta) u_\theta^{\otimes r}\| \left\| \sum_{n=1}^N p_{x_n} \right\|,
\end{aligned}$$

where $\mathbb{D}_{x_n} = \bigoplus_{i=0}^r \frac{1}{i!} \mathcal{D}^{(i)}(0)^* p_{x_n} \oplus p_{x_n}$. Using this perspective, we have the following theorem in the same manner as Theorem 4.1.

**Theorem 4.2.** *Let $\mathbb{D}_{x_n} = \bigoplus_{i=0}^r \frac{1}{i!} \tilde{\mathcal{D}}^{(i)}(0)^* p_{x_n} \oplus p_{x_n}$. Assume there exists $\epsilon > 0$ such that $\|S(0, u_\theta)\| \leq \epsilon$. Assume, in addition, there exists $F > 0$ such that for any $n = 1, \ldots, N$, $\|\mathbb{D}_{x_n}\| \leq F$ and $\|p_{x_n}\| \leq F$. Then, we have*

$$\begin{aligned}
\hat{R}(\mathcal{V}_\Theta, x_1, \ldots, x_N) &\leq \frac{F}{\sqrt{N}} \left( \sup_{u_\theta \in \mathcal{U}_\Theta} \left\| \bigoplus_{i=0}^r u_\theta^{\otimes i} \oplus (g \circ u_\theta) \right\| + \epsilon \sup_{u_\theta \in \mathcal{U}_\Theta} \|u_\theta^{\otimes r}\| \right) \\
&= \frac{F}{\sqrt{N}} \left( \sup_{u_\theta \in \mathcal{U}_\Theta} \left( \sum_{i=0}^r \|u_\theta\|^{2i} + \|g \circ u_\theta\|^2 \right)^{1/2} + \epsilon \sup_{u_\theta \in \mathcal{U}_\Theta} \|u_\theta\|^{2r} \right).
\end{aligned}$$

The order $r$ represents the nonlinearlity of $\mathcal{D}$. Thus, Theorem 4.2 shows that as the nonlinearlity of $\mathcal{D}$ grows, the Rademacher complexity becomes large.

**Remark 4.3.** *Since we have*

$$\langle \mathbb{D}_{x_n}, \mathbb{D}_{x_n} \rangle = \sum_{i=0}^{r} \left\| \frac{1}{i!} \mathcal{D}^{(i)}(0)^* p_{x_n} \right\|^2 + \|p_{x_n}\|^2,$$

*the constant $F$ does not depend on $\theta$.*

**Remark 4.4.** *There is a tradeoff between the factors $r$ and $\epsilon$. When $\mathcal{D}$ is not polynomial, if we take $r$ large, then $\epsilon$ becomes small.*

**Example 4.5.** Let $d = \tilde{d}$. Consider the nonlinear convection operator

$$\mathcal{D}(u) = (u \cdot \nabla)u = \sum_{k=1}^{d} u_k \, \partial_k u_j = [1, \dots, 1] \begin{bmatrix} \partial_{t_1} & & \\ & \ddots & \\ & & \partial_{t_d} \end{bmatrix} \begin{bmatrix} \tilde{u}_{11} & \cdots & \tilde{u}_{1d} \\ \vdots & \ddots & \vdots \\ \tilde{u}_{1d} & \cdots & \tilde{u}_{dd} \end{bmatrix},$$

where $\tilde{u}_{ij}(s,t) = u_i(s)\, u_j(t) = (u_i \otimes u_j)(s,t)$ for $s = [s_1, \dots, s_d], [t_1, \dots, t_d] \in \mathbb{R}^d$. Note that this operator is the second order polynomial with respect to $u$. This operator appears, for example, in the Navier–Stokes equation.

The adjoint $\mathcal{D}''(0)^*$ acts on $v$ as

$$\mathcal{D}''(0)^* v = \begin{bmatrix} \partial_{t_1}^* & & \\ & \ddots & \\ & & \partial_{t_d}^* \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} [v_1 \ \cdots \ v_d] = \begin{bmatrix} \partial_1^* v_1 & \cdots & \partial_1^* v_d \\ \vdots & \ddots & \vdots \\ \partial_d^* v_1 & \cdots & \partial_d^* v_d \end{bmatrix} = - \begin{bmatrix} \partial_1 v_1 & \cdots & \partial_1 v_d \\ \vdots & \ddots & \vdots \\ \partial_d v_1 & \cdots & \partial_d v_d \end{bmatrix}.$$

Therefore, we have

$$\|\mathcal{D}''(0)^* v\|^2 = \sum_{i,j=1}^{d} \|\partial_i v_j\|^2,$$

which shows that $\|\mathcal{D}''(0)^* p_{x_n}\|$ does not depend on $x_n$ if the norm of the derivative of $p_{x_n}$ does not depend on $x_n$. For example, if $p_{x_n}$ is shift-invariant, i.e., $p_{x_n}(y) = \tilde{p}(y - x_n)$ for some $\tilde{p}$, then the norm of the derivative of $p_{x_n}$ does not depend on $x_n$. As a result, by Remark 4.3, the assumption $\|\mathcal{D}_{x_n}\| < F$ in Theorem 4.2 is satisfied. Note that since $\mathcal{D}$ is the second-order polynomial, the order $r$ in Theorem 4.2 is $r = 2$ in this case.

### 4.3 Nonlinear transformation of a linear differential operator

Another important nonlinear form is the nonlinear transformation of a linear differential operator. Let $\mathcal{D} : \mathcal{F} \to L^2(\mathcal{X}, \mathbb{R}^{\tilde{d}})$ be linear. Assume $\mathcal{D}$ is in the following form:

$$\mathcal{D}(u_\theta) = g \circ \tilde{\mathcal{D}} u_\theta \tag{6}$$

with $\tilde{\mathcal{D}} : \mathcal{F} \to L^2(\Omega, \mathbb{R}^{\hat{d}})$ being bounded linear and $g : \mathbb{R}^{\hat{d}} \to \mathbb{R}^{\tilde{d}}$. Typical examples are the Hamilton–Jacobi equations (Crandall & Lions, 1983) and the Monge–Ampère equations (Gong et al., 2020). Assume there exists $\epsilon > 0$, and assume $g \in \mathcal{C}^r(\mathbb{R}^{\hat{d}}, \mathbb{R}^{\tilde{d}})$ whose remainder $S(0, \tilde{\mathcal{D}} u_\theta(x))$ of the $r$-order Taylor approximation at 0 is bounded as $\|S(0, \tilde{\mathcal{D}} u_\theta)\| \leq \epsilon$. In this case, we can approximate $\mathcal{D}$ as a polynomial of $\tilde{\mathcal{D}} u_\theta$, which is represented as a linear operator on a higher dimensional space. We have

$$\begin{aligned} \sum_{n=1}^{N} \langle p_{x_n}, \mathcal{D}(u_\theta) \rangle &= \sum_{n=1}^{N} \left( \sum_{i=0}^{r} \frac{1}{i!} \langle p_{x_n}, g^{(i)}(0) P^i \tilde{\mathcal{D}}^{\otimes i} u_\theta^{\otimes i} \rangle + \langle p_{x_n}, S(0, \tilde{\mathcal{D}} u_\theta(\cdot))(\tilde{\mathcal{D}} u_\theta(\cdot))^{\otimes r} \rangle \right) \\ &= \left\langle \sum_{n=1}^{N} \mathbb{D}_{x_n}, \bigoplus_{i=0}^{r} u_\theta^{\otimes i} \right\rangle + \left\langle \sum_{n=1}^{N} p_{x_n}, S(0, \tilde{\mathcal{D}} u_\theta(\cdot))(\tilde{\mathcal{D}} u_\theta(\cdot))^{\otimes r} \right\rangle \end{aligned}$$

$$\leq \left\| \bigoplus_{i=0}^{r} u_\theta^{\otimes i} \right\| \left\| \sum_{n=1}^{N} \mathbb{D}_{x_n} \right\| + \left\| S(0, \tilde{\mathcal{D}} u_\theta(\cdot))(\tilde{\mathcal{D}} u_\theta(\cdot))^{\otimes r} \right\| \left\| \sum_{n=1}^{N} p_{x_n} \right\|,$$

where $\mathbb{D}_{x_n} = \bigotimes_{i=0}^{r} 1/(i!)(\tilde{\mathcal{D}}^{\otimes i})^*(P^i)^* g^{(i)}(0)^* p_{x_n}$, $P^i : L^2(\mathcal{X}, \mathbb{R}^{\tilde{d}})^{\otimes i} \to L^2(\mathcal{X}, (\mathbb{R}^{\tilde{d}})^{\otimes i})$ is defined as $P^i u^{\otimes i}(x) = u^{\otimes i}(x, \dots, x) = u(x)^{\otimes i}$, and for functions $u$ and $v$ of $x$, $u(\cdot)v(\cdot)$ means $x \mapsto u(x)v(x)$. In the same manner as Theorem 4.1, we have the following theorem.

**Theorem 4.6.** *Let $\mathbb{D}_{x_n} = \bigoplus_{i=0}^{r} \frac{1}{i!}(\tilde{\mathcal{D}}^{\otimes i})^*(P^i)^* g^{(i)}(0)^* p_{x_n}$. Assume there exists $\epsilon > 0$ such that $\|S(0, \tilde{\mathcal{D}} u_\theta)\| \leq \epsilon$. Assume, in addition, there exists $F > 0$ such that for any $n = 1, \dots, N$, $\|\mathbb{D}_{x_n}\| \leq F$ and $\|p_{x_n}\| \leq F$. Then, we have*

$$\begin{aligned}\hat{R}(\mathcal{V}_\Theta, x_1, \dots, x_N) &\leq \frac{F}{\sqrt{N}} \left( \sup_{u_\theta \in \mathcal{U}_\Theta} \left\| \bigoplus_{i=0}^{r} u_\theta^{\otimes i} \right\| + \epsilon \sup_{u_\theta \in \mathcal{U}_\Theta} \|(\tilde{\mathcal{D}} u_\theta)^{\otimes r}\| \right) \\ &= \frac{F}{\sqrt{N}} \left( \sup_{u_\theta \in \mathcal{U}_\Theta} \left( \sum_{i=0}^{r} \|u_\theta\|^{2i} \right)^{1/2} + \epsilon \sup_{u_\theta \in \mathcal{U}_\Theta} \|(\tilde{\mathcal{D}} u_\theta)\|^r \right).\end{aligned}$$

The order $r$ represents the nonlinearity of $g$ in this case. Thus, Theorem 4.6 shows that as the nonlinearity of $g$ grows, the Rademacher complexity becomes large. Note that in the same manner as Remark 4.3, the factor $F$ does not depend on $\theta$.

### 4.4 Application of Koopman-based bounds

Since Theorems 4.1, 4.2, and 4.6 includes the term $\|u_\theta\|$, we can apply Lemma 2.6 to understand the relationship between the Rademacher complexity and the weight matrices. We set $\mathcal{X}_0 = \mathcal{X}$, $d_L = \tilde{d}$. Consider the neural network (1). Then, we obtain the following bound.

Let $\alpha(h) = (\int_{W_l \mathcal{X}_{l-1}} |h(x)|^2 d\mu_{\mathcal{R}(W_l)}(x) / \int_{\tilde{\mathcal{X}}_l} |h(x)|^2 \mathrm{d}\mu_{\mathbb{R}^{d_l}}(x))^{1/2}$ for $h \in \tilde{\mathcal{H}}_l$. Let $D > 0$, $\Theta = \{(W_1, \dots, W_L) \mid |\det W_1^* W_1|^{-1/4}, \dots, |\det W_L^* W_L|^{-1/4} \leq D\}$.

**Theorem 4.7.** *Assume $\mathcal{D}$ is linear. Let $\mathbb{D}_{x_n} = \mathcal{D}^* p_{x_n}$. Assume there exists $F > 0$ such that for any $n = 1, \dots, N$, $\|\mathbb{D}_{x_n}\| \leq F$. Then, we have*

$$\hat{R}(\mathcal{V}_\Theta, x_1, \dots, x_N) \leq \frac{F}{\sqrt{N}} \sup_{\theta \in \Theta} \frac{\|v\| \prod_{l=1}^{L-1} \|K_{\sigma_l}\| \alpha(f_l)}{\prod_{l=1}^{L} |\det W_l^* W_l|^{1/4}}.$$

**Theorem 4.8.** *Assume $\mathcal{D}$ is in the form (5). $\mathbb{D}_{x_n} = \bigoplus_{i=0}^{r} \frac{1}{i!} \tilde{\mathcal{D}}^{(i)}(0)^* p_{x_n} \oplus p_{x_n}$. Assume there exists $\epsilon > 0$ such that $\|S(0, u_\theta)\| \leq \epsilon$. Assume, in addition, there exists $F > 0$ such that for any $n = 1, \dots, N$, $\|\mathbb{D}_{x_n}\| \leq F$ and $\|p_{x_n}\| \leq F$. Then, we have*

$$\begin{aligned}&\hat{R}(\mathcal{V}_\Theta, x_1, \dots, x_N) \\ &\leq \frac{F}{\sqrt{N}} \sup_{\theta \in \Theta} \left( \sum_{i=0}^{r} \left( \frac{\|v\| \prod_{l=1}^{L-1} \|K_{\sigma_l}\| \alpha(f_l)}{\prod_{l=1}^{L} |\det W_l^* W_l|^{1/4}} \right)^{2i} + \left( \frac{\|g \circ v\| \prod_{l=1}^{L-1} \|K_{\sigma_l}\| \alpha(f_l)}{\prod_{l=1}^{L} |\det W_l^* W_l|^{1/4}} \right)^2 \right)^{1/2} \\ &\qquad + \frac{F\epsilon}{\sqrt{N}} \sup_{\theta \in \Theta} \left( \frac{\|v\| \prod_{l=1}^{L-1} \|K_{\sigma_l}\| \alpha(f_l)}{\prod_{l=1}^{L} |\det W_l^* W_l|^{1/4}} \right)^{2r}.\end{aligned}$$

**Theorem 4.9.** *Assume $\mathcal{D}$ is in the form (6). Let $\mathbb{D}_{x_n} = \bigoplus_{i=0}^{r} \frac{1}{i!}(\tilde{\mathcal{D}}^{\otimes i})^*(P^i)^* g^{(i)}(0)^* p_{x_n}$. Assume there exists $\epsilon > 0$ such that $\|S(0, \tilde{\mathcal{D}} u_\theta)\| \leq \epsilon$. Assume, in addition, there exists $F > 0$ such that for any $n = 1, \dots, N$, $\|\mathbb{D}_{x_n}\| \leq F$ and $\|p_{x_n}\| \leq F$. Then, we have*

$$\hat{R}(\mathcal{V}_\Theta, x_1, \dots, x_N) \leq \frac{F}{\sqrt{N}} \sup_{\theta \in \Theta} \left( \sum_{i=0}^{r} \left( \frac{\|v\| \prod_{l=1}^{L-1} \|K_{\sigma_l}\| \alpha(f_l)}{\prod_{l=1}^{L} |\det W_l^* W_l|^{1/4}} \right)^{2i} \right)^{1/2} + \frac{F\epsilon}{\sqrt{N}} \sup_{\theta \in \Theta} \|\tilde{\mathcal{D}} u_\theta\|^{2r}.$$

Since the determinants of weight matrices appear in the denominator of the bound, the bound implies that similar to the standard neural networks, for neural networks with differential operators, high-rank weight matrices such as unitary matrices can also make the generalization error small. Regarding the factor $\|K_{\sigma_l}\|$, we can apply Lemmas 2.4 and 2.5 to obtain an upper bound.

**Remark 4.10.** *Let $p_x$ be the same function as in Example 3.1. Then, we have $\|p_{x_n}\| \to \infty$ as $c \to \infty$. Thus, the constant $F$ depends on $c$ and $F \to \infty$ as $c \to \infty$. Note that $F$ does not depend on the weight matrices (see Remark 4.3). Thus, we can expect that the above bounds also captures the behavior of the standard PINN. We numerically confirm the applicability of our results to the standard PINNs in Section 5.*

**Remark 4.11.** *Since the variational model $V_\theta$ goes to the standard PINN as $c \to \infty$, the Rademacher complexity bound diverges as the model goes to the standard PINNs. This implies the advantage of VPINNs over the standard PINNs in the sense of the generalization. We confirm this in Subsection 5.1.*

**Remark 4.12.** *We can derive a faster rate than $1/\sqrt{N}$ using local Rademacher complexity. See Appendix B for more details.*

## 5 Numerical results

We numerically confirm the validity of our bound. In Subsection 5.1, we consider a regularization based on our bound and show its effectiveness. In Subsection 5.2, we investigate the impact of the nonlinearity $r$ on our bound. All the experiments in this paper were executed with Python 3.10 on an Intel(R) Core(TM) i9-10885H 2.4GHz processor with the Windows 11 operating system.

### 5.1 Navier–Stokes equation

We consider the 2D steady incompressible Navier–Stokes equations on the unit square domain $\mathcal{X} = [0,1]^2$, with the Reynolds number number $\mathrm{Re} = 100$.

$$(u \cdot \nabla)u + \nabla p - \frac{1}{\mathrm{Re}}\Delta u = 0, \qquad \nabla u = 0,$$

where $u \in H^2(\mathcal{X})$ denotes the velocity field and $p \in H^1(\mathcal{X})$ denotes the pressure. The boundary condition is $u(x) = [1,0]$ for $x \in \{[x_1, 1] \mid x_1 \in [0,1]\}$ and $u(x) = [0,0]$ otherwise. The velocity–pressure pair $[u,p]$ is approximated by a single fully connected network $u_\theta : \mathbb{R}^2 \to \mathbb{R}^3$ defined as $u_\theta(x) = W_4\sigma_3(W_3\sigma_2(W_2\sigma_1(W_1x + b_1) + b_2) + b_3)$, where $W_1 \in \mathbb{R}^{64\times 2}$, $W_2, W_3 \in \mathbb{R}^{64\times 64}$, and $W_4 \in \mathbb{R}^{3\times 64}$. We set $\sigma_1, \sigma_2, \sigma_3$ as the pointwise hyperbolic tangent activation function. The two-dimensional input is normalized to $[-1,1]^2$ via $\tilde{x} = 2x - 1$. Weights are initialized with the Glorot uniform scheme (Glorot & Bengio, 2010). The weak residual is approximated by midpoint-rule quadrature on a $25 \times 25$ grid over $\mathcal{X}$. Quadrature nodes are placed at cell centers to avoid boundary points. We set $p_{x_n}$ as in Example 3.1 with $c = 0.1$. The total training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{vpinn}} + \lambda_{\mathrm{bc}}\,\mathcal{L}_{\mathrm{BC}} + \lambda_p\,\mathcal{L}_p + \lambda_K\,\mathcal{L}_K,$$

where each component is defined as follows. The VPINN residual $\mathcal{L}_{\mathrm{vpinn}}$ is set as Eq. (3) with $\mathcal{L}_n(z) = z^2$. Regarding the boundary-condition loss $\mathcal{L}_{\mathrm{BC}}$, boundary conditions are enforced via a penalty term:

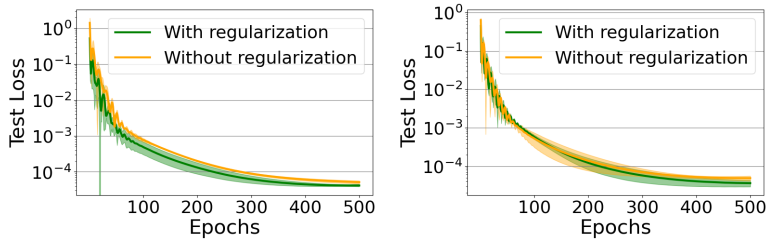
$$\mathcal{L}_{\mathrm{BC}} = \frac{1}{N_{\mathrm{BC}}}\sum_{i=1}^{N_{\mathrm{BC}}} \|u_\theta(x_i) - u_{b,i}\|^2, \qquad \lambda_{\mathrm{bc}} = 0.1,$$

where $u_{b,i}$ represents the boundary condition at $x_i \in \partial\mathcal{X}$. We used $N_{\mathrm{BC}} = 240$ points on $\partial\mathcal{X}$. About the pressure reference $\mathcal{L}_p$, since the pressure is determined only up to an additive constant in incompressible flow, we impose the soft constraint $p(0.5,\, 0.5) = 0$ by setting

$$\mathcal{L}_p = p(0.5,\, 0.5)^2, \qquad \lambda_p = 0.1.$$

We set the number $N$ of training points as $N = 100$. We consider a regularization based on the bound in Theorem 4.8. We computed $\tilde{A}_l$ and $D_l$ based on Appendix C. We added a regularization term $0.01\tilde{A}_l/D_l^{1/2}$ to make the factor $\|K_{\sigma_l}\|/|\det W_l^* W_l|^{1/4}$ in the bound in Theorem 4.8 small. We compared the test loss $\mathcal{L}_{\mathrm{vPINN}}$ with and without the regularization. We used the Adam optimizer with learning rate $10^{-3}$. The results are illustrated in Figure 1 (a). We can see that with the regularization, the test loss becomes smaller than without the regularization, which implies the validity of our bound.

To confirm the applicability of our bound to the standard PINN (the case of $c \to \infty$), we also applied the above regularization to the PINN residual $\mathcal{L}_{\mathrm{PINN}} := 1/N \sum_{n=1}^{N} \mathcal{L}_n(\mathcal{D}(u_\theta)(x_n))$. The results are illustrated in Figure 1 (b). We can see that even for the standard PINN, with the regularization, the test loss becomes smaller than without the regularization, which shows that our bound is also valid for the standard PINNs as we discussed in Remark 4.11.




(a) (b)

Figure 1: (a) Test loss $\mathcal{L}_{\mathrm{VPINN}}$ with and without the regularization based on Therem 4.8 (Average $\pm$ standard deviation of three independent runs). (b) Test loss $\mathcal{L}_{\mathrm{PINN}}$ with and without the regularization based on Therem 4.8 (Average $\pm$ standard deviation of three independent runs).

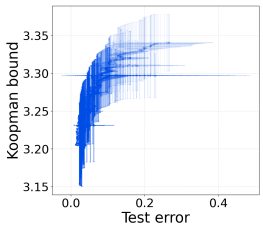



Figure 2: Scatter plot of the test error versus $\tilde{A}_l/D_l^{1/2}$ (Average of 3 independent runs).

Table 1: : Correlation coefficient between the test error and $(\tilde{A}/D_l^{1/2})^r$.

| $r$ | correlation coefficient |
|---|---|
| 1 | 0.8794 |
| 2 | 0.8842 |
| 3 | 0.8882 |

### 5.2 Monge-Ampère equation

We consider the 2D parabolic Monge–Ampère (PMA) equation on $\mathcal{X} = [0,1]^2 \times [0,1]$ studied by Gong et al. (2020), which is defined as

$$-u_t + \log\det D^2 u = f,$$

where $D^2u$ denotes the spatial Hessian of $u$. The boundary condition is $u(x,t) = [1,0]$ for $x \in \{[x_1,1] \mid x_1 \in [0,1]\}$, $t \in [0,1]$ and $u(x,t) = [0,0]$ otherwise. We set the same architecture for the neural network as Subsection 5.1, but the input dimension is 3 since the equation in this case depends also on time. We used the same loss as in Subsection 5.1 but we do not need the loss $\mathcal{L}_p$ in this case. We set $f(x,y) = 1 + \log 2 + 2x$ and consider the same regularization based on our bound. We set the number $N$ of training points as $N = 100$. We used $N_{\mathrm{BC}} = 720$ points on $\partial\mathcal{X}$ to compute $\mathcal{L}_{\mathrm{BC}}$ for the boundary condition. We used the Adam optimizer with learning rate $10^{-3}$. Figure 2 shows the scatter plot of the test error versus $\tilde{A}_l/D_l^{1/2}$. Since $\mathcal{D}$ is not polynomial in this case, $A_l$ in Theorem 4.9 can become larger. We can see that the relationship between the test error and $\tilde{A}_l/D_l^{1/2}$ is not linear, and described by a high-order polynomial. Indeed, Table 1 shows the correlation coefficient between the test error and $(\tilde{A}_l/D_l^{1/2})^r$. We can see that the correlation coefficient becomes larger as $r$ becomes larger, which corresponds to the fact that the differential operator is described by the log function.

## 6 Conclusion and Limitation

In this paper, we derived a new generalization bound for neural networks that involve differential operators with respect to input variables. The proposed bound allows us to apply the Koopman-based approach, which shows that high-rank networks can generalize well even in settings involving differential operators. It also shows that the nonlinearity of the differential operator exponentially enlarges the bound.

Our bound is a uniform bound that does not take the learning algorithm into account. Deriving algorithm-dependent bounds using Koopman-based approach is left for future work.

### AI USE STATEMENT

In this work, we used generative AI tools for formulate mathematical claims, assist in the writing of proofs, implement methods, assist with translation. The mathematical claims in Appendix B is formulated with the assist of generative AI tools. However, we have not used generative AI tools to formulate mathematical claims in the main text. We have not used generative AI tools for generate synthetic data sets, help develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, clean and reformat dataset support qualitative and thematic data analysis, interpret results. Additionally, we used generative AI tools for summarize or analyse existing literature, discover research topics or identify gaps, sourcing/searching for information, identify relevant literature, format references. We have reviewed all AI-assisted work. We checked LLM-generated mathematical claims and codes manually. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

All assumptions are stated in the main text. Complete proofs and additional theoretical results are included in the appendix.

## REFERENCES

Martin Anthony and Peter L. Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

Stefano Berrone, Claudio Canuto, and Moreno Pintore. Solving pdes by variational physics-informed neural networks: an a posteriori error analysis. *Annali dell'Universita di Ferrara*, 68:575–595, 2022.

Tommaso Botarelli, Marco Fanfani, Paolo Nesi, and Lorenzo Pinelli. Using physics-informed neural networks for solving Navier-Stokes equations in complex scenarios. *Engineering Applications of Artificial Intelligence*, 148:110347, 2025.

Shengze Cai, Zhiping Mao, Zhicheng Wang, Minglang Yin, and George Em Karniadakis. Physics-informed neural networks (pinns) for fluid mechanics: a review. *Acta Mechanica Sinica*, 37: 1727–1738, 2021.

Michael G. Crandall and Pierre-Louis Lions. Viscosity solutions of Hamilton–Jacobi equations. *Transactions of the American Mathematical Society*, 277(1):1–42, 1983.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. In *ICLR Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Proceedings of the 2018 Conference On Learning Theory (COLT)*, 2018.

Shuyu Gong, Ziwei Zhou, and Jiguang Bao. Existence and uniqueness of viscosity solutions to the exterior problem of a parabolic Monge–Ampère equation. *Communications on Pure and Applied Analysis*, 19(10):4921–4936, 2020.

Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In *Proceedings of the 2017 Conference on Learning Theory (COLT)*, pp. 1064–1068, 2017.

Yuka Hashimoto, Sho Sonoda, Isao Ishikawa, Atsushi Nitanda, and Taiji Suzuki. Koopman-based generalization bound: New aspect for full-rank weights. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Yuka Hashimoto, Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. Why high-rank neural networks generalize?: An algebraic framework with RKHSs. In *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026.

Haotian Ju, Dongyue Li, and Hongyang R Zhang. Robust fine-tuning of deep neural networks with Hessian-based generalization guarantees. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Vijay Kumar and Gautam Singh. A variational physics-informed neural network framework using Petrov–Galerkin method for solving singularly perturbed boundary value problems. *International Journal of Computational Methods*, 23:2650043, 2026.

Shuai Li, Kui Jia, Yuxin Wen, Tongliang Liu, and Dacheng Tao. Orthogonal deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(04):1352–1368, 2021.

Siddhartha Mishra and Roberto Molinaro. Estimates for generalization error of physics-informed neural networks for approximating pdes. *IMA Journal of Numerical Analysis*, 43(1):1–43, 2023.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2nd edition, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Proceedings of the 2015 Conference on Learning Theory (COLT)*, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

Maziar Raissi, Paris Perdikaris, and George 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.

Zhiyuan Ren, Shijie Zhou, Dong Liu, and Qihe Liu. Physics-informed neural networks: A review of methodological evolution, theoretical foundations, and interdisciplinary frontiers toward next-generation scientific computing. *Applied Sciences*, 15(14):8092, 2025.

Taiji Suzuki, Hiroshi Abe, and Tomoaki Nishimura. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via Lipschitz augmentation. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Colin Wei and Tengyu Ma. Improved sample complexities for deep neural networks and robust classification via an all-layer margin. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

Weinan E, Chao Ma, and Lei Wu. The Barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55:369–406, 2022.

Xianliang Xu, Ye Li, and Zhongyi Huang. Refined generalization analysis of the deep ritz method and physics-informed neural networks. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

# A PROOFS

*Proof of Theorem 4.1.* By the Cauchy–Schwartz inequality and the Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{u_\theta\in\mathcal{U}_\Theta}\frac{1}{N}\sum_{n=1}^{N}\langle p_{x_n},\mathcal{D}(u_\theta)\rangle\,\epsilon_n\Big] \le \frac{1}{N}\mathbb{E}\Big[\sup_{u_\theta\in\mathcal{U}_\Theta}\|u_\theta\|\,\Big\|\sum_{n=1}^{N}\mathcal{D}^*p_{x_n}\epsilon_n\Big\|\Big]\\
&=\frac{1}{N}\sup_{u_\theta\in\mathcal{U}_\Theta}\|u_\theta\|\mathbb{E}\Big[\Big(\sum_{n,m=1}^{N}\langle\mathbb{D}_{x_n}\epsilon_n,\,\mathbb{D}_{x_m}\epsilon_m\rangle\Big)^{1/2}\Big]\\
&=\frac{1}{N}\sup_{u_\theta\in\mathcal{U}_\Theta}\|u_\theta\|\mathbb{E}\Big(\Big[\sum_{n,m=1}^{N}\langle\mathbb{D}_{x_n}\epsilon_n,\,\mathbb{D}_{x_m}\epsilon_m\rangle\Big]\Big)^{1/2}\\
&=\frac{1}{N}\sup_{u_\theta\in\mathcal{U}_\Theta}\|u_\theta\|\Big(\sum_{n,m=1}^{N}\langle\mathbb{D}_{x_n},\,\mathbb{D}_{x_m}\rangle\,\mathbb{E}[\epsilon_n\epsilon_m]\Big)^{1/2}\\
&=\frac{1}{N}\sup_{u_\theta\in\mathcal{U}_\Theta}\|u_\theta\|\Big(\sum_{n=1}^{N}\langle\mathbb{D}_{x_n},\,\mathbb{D}_{x_n}\rangle\Big)^{1/2}\le\frac{F}{\sqrt{N}}\sup_{u_\theta\in\mathcal{U}_\Theta}\|u_\theta\|.
\end{aligned}$$

□

# B FAST RATE

We can apply local Rademacher complexity to derive fast rates.

## B.1 LINEAR CASE

Let $\Phi(x) := \mathcal{D}^*p_x$ and

$$k(x,x') = \langle \mathcal{D}^*p_x, \mathcal{D}^*p_{x'}\rangle\,.$$

Consider the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ associated with $k$.

The Rademacher complexity bound is finite, if there exists $u^* \in L^2(\mathcal{X},\mathbb{R}^{\tilde{d}})$ and if there exists $R>0$ such that $\|u_\theta - u^*\|_{L^2} \le R$. Let $P : L^2(\mathcal{X},\mathbb{R}^{\tilde{d}}) \to \overline{\mathrm{Span}\{\Phi(x)\mid x\in\mathcal{X}\}}$ be the projection onto $\overline{\mathrm{Span}\{\Phi(x)\mid x\in\mathcal{X}\}}$, and let

$$v_\theta\colon x\mapsto \langle \mathcal{D}^*p_x, u_\theta\rangle = \langle \mathcal{D}^*p_x, Pu_\theta\rangle\,.$$

Then, $v_\theta\in\mathcal{H}_k$ and we have $\|v_\theta\|_{\mathcal{H}_k} = \|Pu_\theta\|_{L^2} \le \|u_\theta\|_{L^2}$ (see Hashimoto et al. (2026)). Thus, we have

$$\{u_\theta\in\mathcal{U}_\Theta \mid \|u_\theta-u^*\|_{L^2}\le R\} \subseteq \{v_\theta\in\mathcal{H}_k \mid \|v_\theta - v^*\|_{\mathcal{H}_k}\le R\} =: B_R.$$

Let $r>0$ and the localization $B_R(r)$ of $B_R$ be $B_R(r) = \{v\in B_R \mid \int_{\mathcal{X}}|v_\theta(x)-v^*(x)|^2\mathrm{d}P_X(x)\le r\}$, where $P_X$ is the distribution of samples $x_1,\dots,x_N$. We have the following lemma by Bartlett et al. (2005).

**Lemma B.1.** *Let $T_k$ be the linear operator on $L^2(\mathcal{X},\mathbb{R}^{\tilde{d}})$ defined as $T_ku(x) = \int_{\mathcal{X}}k(x,y)u(y)\mathrm{d}P_X(y)$. Let $\mu_1\ge\mu_2\ge\dots$ be the eignvalues of $T_k$. Then, we have*

$$\hat{R}\big(B_R(r),x_1,\dots,x_N\big) \le \sqrt{\frac{2}{N}\sum_j \min\big(r,R^2\mu_j\big)} =: \psi(r). \tag{7}$$

**Theorem B.2.** *Let $\mu_j \asymp |j|^{-\alpha}$ ($j\in\mathbb{Z}^d$) for $\alpha>d$. Then, we have*

$$\psi(r^*)\asymp r^* \asymp R^{\frac{2d}{\alpha+d}}N^{-\frac{\alpha}{\alpha+d}},$$

*which is faster than $1/2$ since $\alpha>d$.*

*Proof.* Let $T = (R^2/r)^{1/\alpha}$. Note that $T$ satisfies "$R^2\mu_j > r \iff |j| > T$". Let $C(r) = \#\{r < R^2\mu_j\}$ and $D(r) = \sum_{r \geq R^2\mu_j} R^2\mu_j$. Then, we have

$$C(r) \cdot r = \#\{|j| < T\} \cdot r \asymp T^d r = R^{\frac{2d}{\alpha}} r^{1-\frac{d}{\alpha}},$$
$$D(r) \asymp R^2 \int_T^\infty t^{-\alpha} t^{d-1} \mathrm{d}t = \frac{R^2 T^{d-\alpha}}{\alpha - d} = \frac{1}{\alpha - d} R^{\frac{2d}{\alpha}} r^{1-\frac{d}{\alpha}}.$$

Let $c_\alpha = 1 + 1/(\alpha - d)$. The fixed point $r^* = \psi(r^*)$ satisfies

$$(r^*)^2 \asymp \frac{2}{N}\big(C(r^*)r^* + D(r^*)\big) \asymp \frac{2c_\alpha}{N} R^{\frac{2d}{\alpha}} (r^*)^{1-\frac{d}{\alpha}}.$$

Thus, we obtain a rate

$$\psi(r^*) \asymp r^* \asymp R^{\frac{2d}{\alpha+d}} N^{-\frac{\alpha}{\alpha+d}},$$

□

**Corollary B.3.** *Let $p_x(y) = c^{-d}\varphi((y-x)/c)$ for a radial function $\varphi \in L^2(\mathcal{X})$ with a compact support. Let $s > 0$ and $\mathcal{D} = \sum_{|\beta| \leq s} c_\beta \partial^\beta$. Assume $\varphi$ satisfies $|\hat{\varphi}(\xi)| \asymp |\xi|^{-\gamma}$ for $\gamma > s + d/2$. Then, we have*

$$r^* \asymp R^{\frac{2d}{2(\gamma-s)+d}} N^{-\frac{2(\gamma-s)}{2(\gamma-s)+d}}.$$

*Proof.* For simplicity, we assume the support of $\varphi$ is $[-\pi, \pi]$. Let $g = \mathcal{D}^* p_0$ and $\tau_x g(y) = g(x-y)$. Then, we have $k(x, x') = \langle \tau_x g, \tau_{x'} g \rangle_{L^2}$ and $\widehat{\tau_x g}(\xi) = e^{-\mathrm{i}\xi \cdot x} \hat{g}(\xi)$, where $\hat{}$ denotes the Fourier transformation. By the Plancherel theorem, we have

$$\begin{aligned} k(x, x') &= \frac{1}{(2\pi)^d} \langle \widehat{\tau_x g}, \widehat{\tau_{x'} g} \rangle = \frac{1}{(2\pi)^d} \int e^{\mathrm{i}\xi \cdot (x - x')} \left|\hat{g}(\xi)\right|^2 \mathrm{d}\xi \\ &= \mathcal{F}^{-1}\left[\left|\hat{g}\right|^2\right](x - x') =: \kappa(x - x'), \end{aligned}$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transformation. Thus, we have $\hat{\kappa}(\xi) = \left|\hat{g}(\xi)\right|^2$. We have

$$\hat{g}(\xi) = \widehat{\mathcal{D}^* p_0}(\xi) = \sum_\beta c_\beta (-\mathrm{i}\xi)^\beta \widehat{p_0}(\xi)$$

and

$$\widehat{p_0}(\xi) = \int c^{-d} \varphi\Big(\frac{y}{c}\Big) e^{-\mathrm{i}\xi \cdot y} \mathrm{d}y = \int c^{-d} \varphi(y) e^{-\mathrm{i}\xi \cdot cy} \cdot c^d \mathrm{d}y = \hat{\varphi}(c\xi).$$

Therefore, we obtain

$$\hat{\kappa}(\xi) = \left|\hat{g}(\xi)\right|^2 \asymp |\xi|^{2s} \left|\hat{\varphi}(c\xi)\right|^2.$$

Thus, we have $\mu_j = \hat{\kappa}(j) \asymp |j|^{2s} \left|\hat{\varphi}(cj)\right|^2$. When $\left|\hat{\varphi}(\xi)\right| \asymp |\xi|^{-\gamma}$, $\mu_j$ satisfies $\mu_j \asymp |j|^{2s-2\gamma}$. Thus, we have $\alpha = 2s - 2\gamma$ and obtain the result. □

**Remark B.4.** *Corollary B.3 shows that if we choose a sufficiently smooth test function $p_x$ so that it satisfies $\gamma > s + d/2$, then we obtain the fast rate. The test function $p_x$ in Example 3.1 satisfies the assumption of Corollary B.3. In this case, since $\varphi$ is infinitely differentiable, $\gamma$ can be set as an arbitrary large value, and the rate becomes $N^{-1}$.*

### B.2 Polynomial case

We extend Subsection B.1 to the nonlinear setting of Subsection 4.2. In addition to replacing $\mathcal{D}^* p_x$ by $\mathbb{D}_x = \bigoplus_{i=0}^r \frac{1}{i!} \tilde{\mathcal{D}}^{(i)}(0)^* p_x \oplus p_x$, we replace $u_\theta$ by

$$\Psi(u_\theta) = \bigoplus_{i=0}^r u_\theta^{\otimes i} \oplus (g \circ u_\theta) \in \mathcal{H},$$

where $\mathcal{H} = \bigoplus_{i=0}^r L^2(\mathcal{X}, \mathbb{R}^{\tilde{d}})^{\otimes i} \oplus L^2(\mathcal{X}, \mathbb{R}^{\tilde{d}})$. Then, we have

$$V_\theta(x) = \big\langle \mathbb{D}_x, \Psi(u_\theta) \big\rangle_{\mathcal{H}} + \big\langle p_x, w_\theta \big\rangle,$$

where $w_\theta := S(0, u_\theta) u_\theta^{\otimes r}$. The first term is a linear functional of a $\theta$-dependent vector, so the same argument in Subsection B.1 is applicable to it with the kernel

$$k(x, x') = \langle \mathbb{D}_x, \mathbb{D}_{x'} \rangle_{\mathcal{H}} = \sum_{i=0}^{r} \frac{1}{(i!)^2} \Big\langle \tilde{\mathcal{D}}^{(i)}(0)^* p_x,\, \tilde{\mathcal{D}}^{(i)}(0)^* p_{x'} \Big\rangle + \langle p_x, p_{x'} \rangle. \tag{8}$$

Since $\Psi$ is nonlinear, the radius $R$ around $u^*$ is replaced by a larger radius $L_\Psi R$ in $\mathcal{H}_k$ with a constant $L_\Psi$ that grows exponentially in $r$. In addition, the remainder $\rho_\theta : x \mapsto \langle p_x, w_\theta \rangle$ does not belong to $\mathcal{H}_k$ and limits how fast the rate can be.

**Theorem B.5.** *Assume $\mathcal{D}$ is in the form (5) and let $\mathbb{D}_x$ and $k$ be as in (8). Assume $\|\mathcal{D}_{x_n}\| \le F$, $\|p_{x_n}\| \le F$ and $\|S(0, u_\theta)\| \le \epsilon$. Assume $g$ is $L_g$-Lipschitz continuous and we set $L_\Psi = (\sum_{i=1}^r i^2 C^{2(i-1)} + L_g^2)^{1/2}$ and $\delta = F\epsilon C^r$. Let $u^* \in \mathcal{F}$, $R > 0$, and $\mathcal{U}_\Theta(R) = \{u_\theta \in \mathcal{U}_\Theta \mid \|u_\theta - u^*\|_{L^2} \le R\}$. Let $\mu_1 \ge \mu_2 \ge \dots$ be the eigenvalues of $T_k$ and assume $\mu_j \asymp |j|^{-\alpha}$ $(j \in \mathbb{Z}^d)$ for $\alpha > d$. Then the fixed point $r^*$ of the local Rademacher complexity of $\{V_\theta \mid u_\theta \in \mathcal{U}_\Theta(R)\}$ satisfies*

$$r^* \;\lesssim\; \max\Big\{ \big(L_\Psi R\big)^{\frac{2d}{\alpha+d}} N^{-\frac{\alpha}{\alpha+d}},\;\; \delta^2,\;\; \delta N^{-1/2} \Big\}. \tag{9}$$

*In particular, if $\mathcal{D}$ is a polynomial of degree $r$, so that $\epsilon = 0$ and $\delta = 0$, then*

$$r^* \;\asymp\; \big(L_\Psi R\big)^{\frac{2d}{\alpha+d}} N^{-\frac{\alpha}{\alpha+d}}.$$

*Proof.* For $u, u' \in \mathcal{F}$, we have $\|u\|_{L^2}, \|u'\|_{L^2} \le C$. In addition, by the identity $u^{\otimes i} - u'^{\otimes i} = \sum_{k=0}^{i-1} u^{\otimes k} \otimes (u - u') \otimes u'^{\otimes(i-1-k)}$, we obtain $\|u^{\otimes i} - u'^{\otimes i}\| \le i C^{i-1} \|u - u'\|$. Together with the Lipschitz continuity of $g$, we have

$$\|\Psi(u) - \Psi(u')\|_{\mathcal{H}}^2 \le \Big( \sum_{i=1}^{r} i^2 C^{2(i-1)} + L_g^2 \Big) \|u - u'\|^2 = L_\Psi^2 \|u - u'\|^2.$$

Let $P$ be the projection onto $\overline{\mathrm{Span}\{\mathbb{D}_x \mid x \in \mathcal{X}\}}$ and $v^*(x) = \langle \mathcal{D}_x, \Psi(u^*) \rangle$. As in Subsection B.1, $v_\theta \in \mathcal{H}_k$ and $\|v_\theta - v^*\|_{\mathcal{H}_k} = \|P(\Psi(u_\theta) - \Psi(u^*))\| \le L_\Psi R$. Hence $\{v_\theta \mid u_\theta \in \mathcal{U}_\Theta(R)\} \subseteq B_{\tilde{R}}$ with $\tilde{R} := L_\Psi R$, and we can apply Lemma B.1 to $B_{\tilde{R}}$. we write $\psi_{\tilde{R}}$ for the function $\psi$ in (7).

Since $\|w_\theta\| \le \|S(0, u_\theta)\| \, \|u_\theta\|^r \le \epsilon C^r$ and $\|p_x\| \le F$, we have $\sup_x |\rho_\theta(x)| \le \delta$, and the same calculation as in the proof of Theorem 4.1 gives $\hat{R}(\{\rho_\theta \mid \theta \in \Theta\}, x_1, \dots, x_N) \le \delta / \sqrt{N}$.

If $\|V_\theta - v^*\|_{L^2(P_{\mathcal{X}})} \le \sqrt{r}$, then $\|v_\theta - v^*\|_{L^2(P_{\mathcal{X}})} \le \sqrt{r} + \delta$. Thus, by the subadditivity of the empirical Rademacher complexity,

$$\hat{R}\big(\{V_\theta \mid \|V_\theta - v^*\| \le \sqrt{r}\}, x_1, \dots, x_N\big) \le \psi_{\tilde{R}}\big((\sqrt{r} + \delta)^2\big) + \frac{\delta}{\sqrt{N}} =: \psi_{\mathrm{nl}}(r).$$

If $r \ge \delta^2$, then $(\sqrt{r} + \delta)^2 \le 4r$ and, since $\psi_{\tilde{R}}(4r) \le 2\psi_{\tilde{R}}(r)$ by the concavity of $t \mapsto \min(t, a)$, we get $\psi_{\mathrm{nl}}(r) \le 2\psi_{\tilde{R}}(r) + \delta/\sqrt{N}$. The fixed point of the right hand side is bounded by a constant multiple of $\max\{r^*_{\tilde{R}}, \delta/\sqrt{N}\}$, where $r^*_{\tilde{R}}$ is the fixed point of $\psi_{\tilde{R}}$, which is evaluated by Theorem B.2. Combining this with the bound $r^* \le \delta^2$ yields (9). □

We can apply the same argument to the form (6) of Subsection 4.3. We set $\Psi(u_\theta) = \bigoplus_{i=0}^r u_\theta^{\otimes i}$, $\mathcal{D}_x = \bigoplus_{i=0}^r \frac{1}{i!} (\tilde{\mathcal{D}}^{\otimes i})^* (P^i)^* g^{(i)}(0)^* p_x$, $L_\Psi = (\sum_{i=1}^r i^2 C^{2(i-1)})^{1/2}$ and $\delta = F\epsilon \|\tilde{\mathcal{D}}\|^r C^r$.

## C Regularization based on the proposed bound

We consider a regularization based on the bounds in theorems in Subsection 4.4. We set the space of pre-activation vectors as $\tilde{\mathcal{X}}_l = \|W_l\|[-1, 1]^{d_{l-1}} + \|b_l\|_\infty$ for $l = 1, \dots, 4$, where $d_0 = d$ and $d_l$ is the output dimension of the $l$th layer for $l = 1, \dots, 3$. According to Lemma 2.5, we computed the upper bound $A_l := \sup_{x \in \sigma_l(\tilde{\mathcal{X}}_j)} |J\sigma^{-1}(x)|$ of $\|K_{\sigma_l}\|^2$. To avoid that the upper bound becomes too large, we computed $\tilde{A}_l := 1/(1 + 1/A_l)$ instead of $A_l$. The factor $\tilde{A}_l$ increases monotonically

as $A_l$ increases but is bounded. Regarding the factor $(\det W_l^* W_l)^{1/4}$, since $(\det W_l^* W_l)^{1/2}$ is described by the product of the singular values of $W_l$, we consider the singular values of $W_l$. Let $\sigma_1 \geq \cdots \geq \sigma_R > 0$ be the singular values of $W_l$, where $R$ is the rank of $W_l$. The geometric mean

$$D_l = \Big(\prod_{k=1}^{R} \sigma_k\Big)^{1/R} = |\det W_l|^{1/R} = \exp\left(\frac{1}{R}\sum_{k=1}^{R} \ln \sigma_k\right)$$

is computed in log-space for numerical stability with an offset $10^{-8}$ before taking logarithms. By adding a regularization term $0.01\tilde{A}_l/D_l^{1/2}$, we can make the factor $\|K_{\sigma_l}\|/|\det W_l^* W_l|^{1/4}$ in the bound in Theorem 4.8 small.