
How does PDE order affect the convergence of PINNs?

Changhoon Song^{1*} Yesom Park^{1,2*} Myungjoo Kang³

¹ Research Institute of Mathematics, Seoul National University

² Department of Mathematics, University of California, Los Angeles

³ Department of Mathematical Sciences, Seoul National University

changhoon.song93@gmail.com, yeisom@math.ucla.edu, mkang@snu.ac.kr

Abstract

This paper analyzes the inverse relationship between the order of partial differential equations (PDEs) and the convergence of gradient descent in physics-informed neural networks (PINNs) with the power of ReLU activation. The integration of the PDE into a loss function endows PINNs with a distinctive feature to require computing derivatives of model up to the PDE order. Although it has been empirically observed that PINNs encounter difficulties in convergence when dealing with high-order or high-dimensional PDEs, a comprehensive theoretical understanding of this issue remains elusive. This paper offers theoretical support for this pathological behavior by demonstrating that the gradient flow converges in a lower probability when the PDE order is higher. In addition, we show that PINNs struggle to address high-dimensional problems because the influence of dimensionality on convergence is exacerbated with increasing PDE order. To address the pathology, we use the insights garnered to consider variable splitting that decomposes the high-order PDE into a system of lower-order PDEs. We prove that by reducing the differential order, the gradient flow of variable splitting is more likely to converge to the global optimum. Furthermore, we present numerical experiments in support of our theoretical claims.

1 Introduction

Understanding of partial differential equations (PDEs) is fundamental in describing diverse phenomena in science and engineering, including fluid dynamics [60, 17], weather prediction [51], disease progression [3, 46], and quantum mechanics [20, 8]. This underscores the imperative necessity for the effective acquisition of their solutions. Given that analytically solving PDEs is often infeasible or even impossible for numerous practical scenarios due to their complexity, numerical methodologies play a pivotal role in approximating solutions to PDEs, enabling researchers and engineers to address real-world problems effectively.

The advent of deep learning has led to a surge in attempts to leverage it to solve PDEs [59, 42, 30]. Among these, physics-informed neural networks (PINNs) [39, 21, 38, 58] stand out as a prominent methodology. Coupled with the automatic differentiation technique [7], they integrate the residuals of PDEs and boundary conditions into the loss function, thereby enforcing the approximation of solutions using artificial neural networks. This distinctive incorporation of PDEs into the loss function introduces partial differential operators in calculating the loss, distinguishing PINNs from conventional deep learning models. Renowned for their accessibility and versatility in being capable of easily handling arbitrary PDEs and being mesh-free, PINNs have garnered significant attention and demonstrated promising outcomes across various fields [13, 28, 2, 65].

*Equal contribution authors. Correspondence to: <mkang@snu.ac.kr>.

Despite their potential, PINNs frequently encounter difficulties in accurately approximating solutions, particularly when the governing PDE contains high-order derivatives [48, 33]. They also exhibit sensitivity to increasing dimensions [33]. These challenges impede the practicality of PINNs due to the pervasiveness of high-order or high-dimensional PDEs in numerous physical and engineering descriptions, such as control problems [22, 57], finance [5, 53], phase separation [12, 27], and mechanical engineering [4, 32]. Several studies have indicated that neural network architectures possess sufficient expressive power to approximate solutions [31, 41]. However, it has been purported that the inferior performance may be attributed to the difficulty in optimizing PINNs, which arises from including the PDE in the loss function [37, 61, 62]. Despite the widespread use of PINNs, a rigorous mathematical understanding of these pathological behaviors of PINNs has been lacking.

In this paper, we endeavor to provide a mathematical understanding of the pathological behaviors of PINNs by analyzing the convergence of their gradient flow (GF), which reveals a profound sensitivity of the GF convergence with respect to the PDE order and the power of the activation. Building upon the work of Gao et al. [25], we extend the analysis of the GF of PINNs, composed of two-layer multilayer perceptrons (MLPs), to general k th-order PDEs and the p -th power of Rectified Linear Unit (ReLU) activation function with general p . We achieve tighter bounds than those obtained by Gao et al., shedding light on the underlying causes of the pathological behaviors of PINNs. Our theoretical findings demonstrate that the width size of the network necessary for the convergence of the GF increases exponentially with the power p of ReLU^p activation. Furthermore, our results indicate that the optimal power p is determined by the order k of the governing PDE, specifically to be $k + 1$. We also find that the PDE order impedes the convergence of GF, where this negative impact of the PDE order stems from incorporating the PDE into the PINN loss function, which necessitates network differentiation up to the order of the PDE. Moreover, our theoretical investigation unveils that the GF convergence of PINNs also deteriorates with increasing dimensions, and the differential operators included in the PINN loss further exacerbate the sensitivity of PINNs to dimensionality. This elucidates why PINNs are relatively sensitive to dimensionality compared to conventional deep learning models that do not involve differentiation in the loss function.

To address these challenges, we mathematically demonstrate the efficacy of a variable splitting strategy [54, 55, 6], which represents derivatives of the solution as additional auxiliary variables. The key point of variable splitting is that learning a high-order PDE boils down to learning a system of lower-order PDEs. Reducing the order of derivatives included in the loss function, the strategy alleviates the difficulties associated with the PDE order. It further enables to utilize more general ReLU^p activation with lower power p than PINNs. The lower differential orders that the network computes, the more likely it is that the GF will converge, so the most suitable one among the various constructions of the variable splitting method is the finest splitting, which separates all the derivatives into auxiliary variables and reformulates the PDE into a system of first-order PDEs. This strategy results in a loss function comprising only first-order derivatives, and the efficacy of this finest variable splitting would be magnified as the order of the governing PDE or dimension increases. Therefore, the finest splitting approach would exhibit a pronounced discrepancy from the vanilla PINNs for high-order PDEs. Moreover, a reduction in the differential orders enhances the resilience of the model with respect to dimensionality. Finally, we present numerical experiments to verify our theoretical findings and validate the effectiveness of the variable splitting.

1.1 Related Work

Characterization of Gradient Descent for PINNs As significant issues have been identified within physics-informed machine learning, numerous mathematical studies have been conducted to elucidate the behavior of PINNs. While studies have been mainly dedicated to examining the generalization capacity of PINNs [19, 49, 23], there has also been work on understanding the difficulty of optimization, which is believed to be the primary source of failure for PINNs. Wang et al. [61] found that PINNs exhibit stiff gradient flow dynamics, resulting in imbalanced gradients during training. Ryck et al. [18] characterized the rate of convergence in terms of the conditioning of an operator and suggested that the difficulty of training PINNs is closely related to the conditioning of the differential operators in the governing PDEs. Another work [62] utilized the neural tangent kernel (NTK) theory to indicate that spectral biases and discrepancies between convergence rates of various loss components can lead to training instabilities. Global convergence properties of PINNs for second-order linear PDEs have also been studied within the NTK regime [34] and using the Rademacher complexity [47]. Most closely related to this paper, Gao et al. [25] demonstrated the

convergence of the gradient descent for two-layer PINNs. However, their discussion is limited to second-order linear PDEs. We extend the analysis to general k th-order linear PDEs and p -th power of activation functions and provide tighter bounds than Gao et al.. These advances allow us to observe further the effect of the order and dimensionality of the PDE on the convergence.

Variable Splitting The method of separation of variables, which simplifies differential equations by reformulating them into a more manageable system, is a classical method for solving differential equations [9]. In particular, it has been widely employed when dealing with high-order PDEs as augmenting high-order derivatives as additional variables allows the governing equation to be decoupled into a set of lower-order PDEs that are comparatively easy to solve [16, 26, 63]. Recent endeavors have explored the integration of separable variables within the PINN approach. In this paper, we refer to this approach as variable splitting according to [54, 55, 56]. Augmented variables have been introduced to represent vorticity in the Stokes equation [6], the gradient of the solutions for solving the p -Poisson equation [54], and the eikonal equation [55]. Additionally, second-order derivatives have been separately parameterized to solve bi-harmonic equations effectively [45]. The rationale for introducing auxiliary variables in previous works is to enhance the efficiency and accuracy of PINNs, but they lack a comprehensive theoretical elucidation of its effect. A recent study [56] has theoretically analyzed variable splitting, demonstrating that while PINNs do not guarantee convergence to the PDE solution even when the loss converges to zero, variable splitting does ensure convergence to the solution for second-order linear PDEs. In this study, we analyze the impact of variable splitting for PINNs with ReLU^p activation in terms of the convergence of the GF.

1.2 Main Contributions

The contribution of the paper is summarized as follows.

- We analyze that the GF of PINNs with ReLU^p activation converges to the global minimum for general k th-order linear PDEs. This extends the findings of Gao et al. [25] to encompass a broader range of PDEs and activations and provides an even tighter bound.
- We demonstrate the inverse relation between PDE order and the GF convergence, unveiling the adverse effect of the differentials included in the PINN loss on the GF convergence.
- We provide a theoretical understanding of the reasons why PINNs encounter difficulties in addressing high-dimensional problems.
- We prove that the order reduction of variable splitting, which reformulates the PDEs into a system of lower-order PDEs, results in the convergence enhancement of GF.

2 Mathematical Setup

Arbitrary Order Linear PDEs We consider a general form of k th-order linear partial differential equations (PDEs) defined on a bounded domain $\Omega \subset \mathbb{R}^d$ (in which the temporal dimension could be a subcomponent)

$$\begin{cases} \mathcal{N}[u](\mathbf{x}) = f(\mathbf{x}), & \mathbf{x} \in \Omega, \\ \mathcal{B}[u](\mathbf{x}) = g(\mathbf{x}), & \mathbf{x} \in \partial\Omega, \end{cases} \quad (1)$$

where $\mathcal{N}[u] = \sum_{|\alpha| \leq k} a_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} u$ is a k th-order linear differential operator with coefficient functions $a_\alpha : \Omega \rightarrow \mathbb{R}$ for each multi-index $\alpha \in \mathbb{N}_0^d$, $\mathcal{B}[u] = \sum_{|\alpha| \leq 1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} u$ represents the boundary condition operator with coefficient functions $\tilde{a}_\alpha : \partial\Omega \rightarrow \mathbb{R}$, which could reflect Dirichlet, Neumann, and Robin conditions², $f : \Omega \rightarrow \mathbb{R}$ is a given source function, and $g : \partial\Omega \rightarrow \mathbb{R}$ is a given boundary function, and $u : \bar{\Omega} \rightarrow \mathbb{R}$ is the unknown solution of interest.

²The boundary condition for high-order PDEs is typically given by multiple conditions of a higher order than one. Our approach also encompasses such general boundary conditions by incorporating the residuals of each boundary condition into the loss. The sole distinction is the utilization of induction not only on the derivative matching losses but also on boundary losses to prove Proposition 2. For the sake of brevity, we assume that the boundary condition is the most prevalent (weighted combination of) Dirichlet and Neumann conditions.

Physics-Informed Neural Networks Physics-informed neural networks (PINNs) [58] aim to approximate the solution u of the PDE by neural networks. Following the prior work [25], we approximate the solution u by a two-layer multi-layer perceptron $\phi: \mathbb{R}^d \rightarrow \mathbb{R}$ of width m , defined as

$$\phi(\mathbf{x}; \mathbf{w}, \mathbf{v}) = \frac{1}{\sqrt{m}} \sum_{r=1}^m v_r \cdot \sigma(\mathbf{w}_r^\top \mathbf{y}), \quad (2)$$

where $\mathbf{w}_r \in \mathbb{R}^{d+2}$, $v_r \in \mathbb{R}$, $\mathbf{w} = [\mathbf{w}_0^\top, \dots, \mathbf{w}_m^\top]^\top \in \mathbb{R}^{m(d+2)}$, $\mathbf{v} = [v_0, \dots, v_m]^\top \in \mathbb{R}^m$, $\mathbf{y} = [\mathbf{x}^\top \quad \frac{1}{2}]^\top \in \mathbb{R}^{d+1}$, and $\sigma(\cdot)$ is the activation function. For brief notations, we assume that Ω is bounded so that $\|\mathbf{y}\|_2 \leq 1$ for $\mathbf{x} \in \Omega$. We consider the case where σ is the ReLU ^{p} activation function for an integer p , which is also known as Rectified Power(RePU) activation [1, 10, 15]. As it will be clear in the context, the power p necessitates surpassing the order k of the PDE (1) to ensure that the loss function and gradient descent flow are well-defined. Therefore, our analysis is focused on scenarios where $p \geq k+1$. PINNs learn the parameters of ϕ by minimizing a composite loss function, comprising the residual of the PDE and the boundary condition of (1), which enforces the network's compliance with the governing physics. For given the training data $\{\mathbf{x}_i, f(\mathbf{x}_i)\}_{i=1}^{n_o} \subset \Omega \times \mathbb{R}$ and $\{\tilde{\mathbf{x}}_j, g(\tilde{\mathbf{x}}_j)\}_{j=1}^{n_b} \subset \partial\Omega \times \mathbb{R}$ of respective sizes $n_o \in \mathbb{N}$ and $n_b \in \mathbb{N}$, PINN loss function is given by

$$\mathcal{L}_{PINN}(\mathbf{w}, \mathbf{v}) = \frac{1}{2} \left(\|\mathbf{s}(\mathbf{w}, \mathbf{v})\|^2 + \|\mathbf{h}(\mathbf{w}, \mathbf{v})\|^2 \right), \quad (3)$$

where $\mathbf{s}(\mathbf{w}, \mathbf{v}) = [s_1(\mathbf{w}, \mathbf{v}) \quad \dots \quad s_{n_o}(\mathbf{w}, \mathbf{v})]^\top$ and $\mathbf{h}(\mathbf{w}, \mathbf{v}) = [h_1(\mathbf{w}, \mathbf{v}) \quad \dots \quad h_{n_b}(\mathbf{w}, \mathbf{v})]^\top$ with

$$s_i(\mathbf{w}, \mathbf{v}) = \sqrt{\frac{1}{n_o}} (\mathcal{N}[\phi(\cdot; \mathbf{w}, \mathbf{v})](\mathbf{x}_i) - f(\mathbf{x}_i)), \quad (4)$$

$$h_j(\mathbf{w}, \mathbf{v}) = \sqrt{\frac{\nu}{n_b}} (\mathcal{B}[\phi(\cdot; \mathbf{w}, \mathbf{v})](\tilde{\mathbf{x}}_j) - g(\tilde{\mathbf{x}}_j)), \quad (5)$$

and $\nu > 0$ is a regularization parameter that relatively balances the two components of the loss.

Gradient Flow As the limiting dynamics of the gradient descent (GD) with infinitesimal step-sizes [40], gradient flow (GF) is continuous time dynamics that starts at $\mathbf{w}(0)$ and $\mathbf{v}(0)$ and evolves as

$$\begin{cases} \frac{d\mathbf{w}_r(t)}{dt} &= -\frac{\partial \mathcal{L}_{PINN}(\mathbf{w}, \mathbf{v})}{\partial \mathbf{w}_r} = -\sum_{i=1}^{n_o} s_i(\mathbf{w}, \mathbf{v}) \cdot \frac{\partial s_i(\mathbf{w}, \mathbf{v})}{\partial \mathbf{w}_r} - \sum_{j=1}^{n_b} h_j(\mathbf{w}, \mathbf{v}) \cdot \frac{\partial h_j(\mathbf{w}, \mathbf{v})}{\partial \mathbf{w}_r}, \\ \frac{dv_r(t)}{dt} &= -\frac{\partial \mathcal{L}_{PINN}(\mathbf{w}, \mathbf{v})}{\partial v_r} = -\sum_{i=1}^{n_o} s_i(\mathbf{w}, \mathbf{v}) \cdot \frac{\partial s_i(\mathbf{w}, \mathbf{v})}{\partial v_r} - \sum_{j=1}^{n_b} h_j(\mathbf{w}, \mathbf{v}) \cdot \frac{\partial h_j(\mathbf{w}, \mathbf{v})}{\partial v_r}. \end{cases} \quad (6)$$

Initial weights are supposed to follow the normal and uniform distributions, $\mathbf{w}(0) \sim N(0, I_m)$ and $\mathbf{v} \sim U(\{-1, 1\})$, respectively³. GF can be regarded as a continuous-time analog of GD and is frequently employed to comprehend the behavior of GD optimization algorithm in the limit. By the chain rule in conjunction with (6), the following characterizes how the loss function evolves during training by gradient descent:

$$\frac{d}{dt} \begin{bmatrix} \mathbf{s}(\mathbf{w}(t), \mathbf{v}(t)) \\ \mathbf{h}(\mathbf{w}(t), \mathbf{v}(t)) \end{bmatrix} = -(\mathbf{G}_w(\mathbf{w}(t), \mathbf{v}(t)) + \mathbf{G}_v(\mathbf{w}(t), \mathbf{v}(t))) \begin{bmatrix} \mathbf{s}(\mathbf{w}(t), \mathbf{v}(t)) \\ \mathbf{h}(\mathbf{w}(t), \mathbf{v}(t)) \end{bmatrix}, \quad (7)$$

where $\mathbf{G}$ and $\tilde{\mathbf{G}}$ are Gram matrices for the dynamics, defined by

$$\mathbf{G}_w(\mathbf{w}, \mathbf{v}) = \mathbf{D}_w^\top \mathbf{D}_w, \quad \mathbf{D}_w = \begin{bmatrix} \frac{\partial s_1}{\partial \mathbf{w}}(\mathbf{w}, \mathbf{v}) & \dots & \frac{\partial s_{n_o}}{\partial \mathbf{w}}(\mathbf{w}, \mathbf{v}) & \frac{\partial h_1}{\partial \mathbf{w}}(\mathbf{w}, \mathbf{v}) & \dots & \frac{\partial h_{n_b}}{\partial \mathbf{w}}(\mathbf{w}, \mathbf{v}) \end{bmatrix} \quad (8)$$

$$\mathbf{G}_v(\mathbf{w}, \mathbf{v}) = \mathbf{D}_v^\top \mathbf{D}_v, \quad \mathbf{D}_v = \begin{bmatrix} \frac{\partial s_1}{\partial \mathbf{v}}(\mathbf{w}, \mathbf{v}) & \dots & \frac{\partial s_{n_o}}{\partial \mathbf{v}}(\mathbf{w}, \mathbf{v}) & \frac{\partial h_1}{\partial \mathbf{v}}(\mathbf{w}, \mathbf{v}) & \dots & \frac{\partial h_{n_b}}{\partial \mathbf{v}}(\mathbf{w}, \mathbf{v}) \end{bmatrix}. \quad (9)$$

We are interested in analyzing the effect of the PDE order on the convergence of the PINN loss, which evolved in accordance with the dynamics (7), to the global minimum zero.

³Indeed, our analysis covers more general initialize distributions than [25]. In our analysis, it is enough that a probability density function of $\mathbf{w}(0)$ is in Schwartz space, and that of $\mathbf{v}(0)$ is bounded and has zero expectation.

3 Impact of PDE Order on Convergence of PINNs

Despite the demonstrated promise and versatility of PINNs in addressing a wide range of problems [29, 44, 36], they often encounter difficulties in constructing an accurate approximation to the desired solution of PDEs, particularly with high-order PDEs. Moreover, in contrast to the confirmed efficacy of neural networks in modeling high-dimensional data such as images and text, the exploration of PINNs for high-dimensional PDEs has been apparently limited. While neural network architectures possess sufficient expressive power to approximate solutions [11], inferior performance has been attributed to the difficulty in optimization in practice [61, 62]. Additionally, it has been postulated that the optimization difficulty may stem from the partial differential operators included in the loss function [37, 48, 33]. Nevertheless, despite the significant challenge posed by these pathological phenomena, there remains a paucity of theoretical understanding of them.

In this section, we theoretically elucidate these pathological phenomena by studying the convergence condition of GF (7) of PINN loss (3). Specifically, we provide a width condition for (7) to converge to global optimum in terms of order k , dimension d , and the power p of ReLU activation. Analyzing how those factors are related to the convergence condition, we explain why optimizing PINNs is harder when the order or degree is higher.

Following [24] and [25], we first prove the positive definiteness of the limiting Gram matrix of PINNs for general k th-order linear PDE and p without any further strict assumption other than $p > k$.

Proposition 3.1 (Special Case). *The limiting Gram matrix $\mathbf{G}_v^\infty = \mathbb{E}_{\mathbf{w}, \mathbf{v}} [\mathbf{G}_v(\mathbf{w}, \mathbf{v})]$ is strictly positive definite and independent of m .*

This is a special case of the general state in Proposition C.3 with $L = 0$, and the proof for the general case is provided in Appendix C. We denote the smallest eigenvalue of $\mathbf{G}_v^\infty$ by $\lambda_0 > 0$. The following presents our main theorem in this section, the requisite width size m for the GF of PINN loss to converge to the global minimum with high probability. The result demonstrates that the required width grows exponentially as the PDE order k and the dimension of the domain d increase.

Theorem 3.2 (Special Case). *There exists a constant C , independent of d , k , and p , such that for any $\delta \ll 1$, if*

$$m > C \binom{d+k}{d}^{14} p^{7k+4} 2^{6p} \left(\log \frac{md}{\delta} \right)^{4p} \quad (10)$$

then with probability of at least $1 - \delta$ over the initialization, we have

$$\mathcal{L}_{PINN}(\mathbf{w}(t), \mathbf{v}(t)) \leq \exp(-\lambda_0 t) \mathcal{L}_{PINN}(\mathbf{w}(0), \mathbf{v}(0)), \quad \forall t \geq 0. \quad (11)$$

It is a special case of Theorem 4.3 with $L = 0$, the proof of which can be found in Appendix C.1. It extends, inspired by [25], the convergence of the GF of PINNs of second-order linear PDEs to k th-order linear PDEs and the general p -th power of ReLU. It states that even in these general settings, the GF of PINNs converges to the global minimum with a high probability when the width of the network is sufficiently large. Moreover, we obtain a polylogarithmic bound $(\log \frac{1}{\delta})^{4p}$, which is much tighter than polynomial bound δ^{-3} in [25] for $p = 3$. These improvements permit the derivation of the following valuable explanations for the deficiencies observed when optimizing PINNs.

Optimal Power of ReLU Function in Training Theorem 3.2 sheds light on the suitable choice of activation function for PINNs. In the training process, the activation function plays an important role. However, there are no clues as to which activation function is favorable to the given optimization process. Especially in the case of PINNs, it depends heavily on the PDE at hand. Despite its pervasive use in deep learning due to its numerous advantages and performance benefits, the ReLU activation function is not admissible in the PINN framework, which necessitates the activation function to provide high-order derivatives for optimizing PDE-based constraints. Instead, PINNs harness the p -th power of ReLU as the activation function. It is apparent that p must satisfy $p \geq k + 1$ for the PINN loss and gradient descent to be computed. Theorem 3.2 indicates that the smaller p is, the more likely the gradient descent will converge; that is, it is most optimal⁴ to adjust p to $k + 1$ regarding the training process.

⁴In terms of approximating the solution of the PDE, a larger p makes the network smoother and has better expressive power [11]. However, it means that the set of networks covers broader function spaces as p increases,

Understanding Difficulty in High-order PDEs A significant observation of Theorem 3.2 is that it provides a theoretical understanding of why PINNs struggle with high-order PDEs. From (10), we can see that the bound increases exponentially with the order of the PDE. Moreover, for the GF of PINNs to converge with high probability, that is, $\delta \ll 1$, a small increment of the power p would contribute to non-negligible degradation in the convergence, which could ultimately prevent the network from reaching a minimizer of the loss. Hence, given that k determines the admissible p by $p \geq k + 1$, the order k of PDE primarily influences the convergence of PINNs and increasing the exponential term in (10).

Understanding Difficulty in High-dimensional Problems The above theorem, which shows that the lower bound of m depends on the exponential of d , explains why PINNs cannot completely combat the curse of dimensionality. As PINNs are regarded as a versatile method capable of being mesh-free, they have been expected to be free from the curse of dimensionality [58]. However, the GF of PINNs becomes harder to converge as d increases, requiring the network to be wider. Furthermore, it can be observed that the magnitude of change in d is amplified with respect to the exponent of k . This explains why PINNs are relatively sensitive to increasing dimensionality in comparison to other deep learning models whose loss functions do not contain derivatives. In other words, the presence of derivatives in the loss makes PINNs sensitive to changes in dimensionality, and the larger k is, the more difficult PINNs are for high-dimensionality.

Combining all crucial observations from our main theorem, we believe that the impact of the PDE order is one of the primary underlying reasons why PINNs often fail to minimize their loss. In light of this theoretical evidence, the next section describes a variable splitting strategy that addresses these pathologies by properly reducing the differential order in the PINN loss function.

4 Order Reduction through Variable Splitting

The previous section indicates that the PDE order k significantly affects the width requirement for the GF to converge. Concurrently, for k th-order PDEs, it is necessary to increase the ReLU activation to at least the $k + 1$ power in order to ensure a well-defined GF for the PINN loss. Consequently, lowering k could potentially lead to better convergence of the GF. In this section, we introduce variable splitting strategy to decrease the differential order by reformulating the given PDE into a system of lower-order PDEs. We then extend Theorem 3.2 to a more general form in Theorem 4.3.

4.1 Variable Splitting

The concept of variable splitting [54, 55, 56] is to rewrite a higher-order PDE into a lower-order system, after which the PINN approach is applied to the system. A crucial aspect of the success of such methods is the reduction of the derivative order present in the training loss function.

Augment Variables For $L \geq 0$ and increasing integers $0 = \xi_0 < \xi_1 < \dots < \xi_{L+1} = k$, variable splitting augment the derivatives of the solution $\frac{\partial^{\xi_1}}{\partial \mathbf{x}^{\xi_1}} u, \dots, \frac{\partial^{\xi_L}}{\partial \mathbf{x}^{\xi_L}} u$ as additional auxiliary variables $\phi_1, \dots, \phi_L$, respectively. For notational simplicity, we abbreviate the integer set $\{1, \dots, m\}$ for a positive integer m by $[m]$. For $\ell \in [L]$, each term in $\frac{\partial^{\xi_\ell}}{\partial \mathbf{x}^{\xi_\ell}}$ corresponds to $\frac{\partial^{\xi_\ell}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} u$ for a multi-index $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ with the size $|\alpha| = \sum_{i=1}^d \alpha_i = n$. Therefore, ϕ_ℓ is a vector-valued function of size $|I_{\xi_\ell}|$ for the index set I defined in (18). We denote the component of ϕ_ℓ that corresponds to $\frac{\partial^{\xi_\ell}}{\partial \mathbf{x}^\alpha}$ by $(\phi_\ell)_\alpha$.

Reformulate PDE into Lower-order System By replacing each of the differential terms $\frac{\partial^{\xi_\ell}}{\partial \mathbf{x}^{\xi_\ell}} u$ with the corresponding auxiliary variables ϕ_ℓ , the differential operator $\mathcal{N}$ in (1) can be rewritten as:

$$\mathcal{N}[u] = \sum_{|\alpha| \leq k} a_\alpha \frac{\partial^{|\alpha|}}{\partial \mathbf{x}^\alpha} u = \sum_{\ell=0}^L \sum_{|\alpha| \leq \xi_\ell} \sum_{|\beta| \leq \Delta \xi_{\ell+1}} \hat{a}_{\ell, \alpha, \beta} \frac{\partial^{\Delta \xi_{\ell+1}}}{\partial \mathbf{x}^\beta} (\phi_\ell)_\alpha, \quad (12)$$

not that a network can be easily trained. Indeed, our result leads us to the opposite conclusion that large p could be detrimental to convergence from the optimization perspective. In this paper, we refer to ‘optimal’ as the sense of being likely trained under the mildest condition rather than approximating the solution with the smallest error.

for some coefficient functions $\hat{a}_{\ell,\alpha,\beta} : \Omega \rightarrow \mathbb{R}$ and $\Delta\xi_\ell = \xi_\ell - \xi_{\ell-1}$. Since ϕ_ℓ represents a function that differentiates the PDE solution $\Delta\xi_\ell$ -times more than $\phi_{\ell-1}$, the components of two consecutive variables $\phi_{\ell-1}$ and ϕ_ℓ are governed by

$$\frac{\partial^{\Delta\xi_\ell}}{\partial \mathbf{x}^\beta} (\phi_{\ell-1})_\alpha(\mathbf{x}) = (\phi_\ell)_{\alpha+\beta}(\mathbf{x}), \quad \alpha \in I_{\xi_{\ell-1}}, \quad \beta \in I_{\Delta\xi_\ell}. \quad (13)$$

From these, the PDE (1) can be identically reformulated by the system of lower-order PDEs:

$$\begin{cases} \hat{\mathcal{N}}[\phi_0, \dots, \phi_L](\mathbf{x}) = f(\mathbf{x}), & \mathbf{x} \in \Omega, \\ \frac{\partial^\beta}{\partial \mathbf{x}^\beta} (\phi_{\ell-1})_\alpha(\mathbf{x}) = (\phi_\ell)_{\alpha+\beta}(\mathbf{x}), & \mathbf{x} \in \Omega, \ell \in [L], \alpha \in I_{\xi_{\ell-1}}, \beta \in I_{\Delta\xi_\ell}, \\ \mathcal{B}[\phi_0](\mathbf{x}) = g, & \mathbf{x} \in \partial\Omega. \end{cases} \quad (14)$$

It is of paramount importance to note that the maximum differential order of this system of PDEs is the highest difference of derivative order between consecutive auxiliary variables $|\xi| = \max\{\Delta\xi_\ell : \ell \in [L+1]\}$, which is less than k . This aspect gives rise to notable ramifications in our analysis of VS-PINNs, which will be discussed in the next subsection.

Variable Splitting for PINNs In this paper, we consider the parameterization of all variables ϕ_ℓ with two-layer MLPs with ReLU^p activation function, in a manner analogous to that described in Section 2 for PINNs. The weights in the first and second layers of ϕ_ℓ are denoted by $\mathbf{w}_\ell$ and $\mathbf{v}_\ell$, respectively. We use $\mathbf{w} = [\mathbf{w}_1^\top \dots \mathbf{w}_L^\top]^\top$ and $\mathbf{v} = [\mathbf{v}_1^\top \dots \mathbf{v}_L^\top]^\top$ to refer to the respective collections of all weights. Similar to PINNs, *Variable Splitting for PINNs (VS-PINNs)* employ the linear sum of penalized residuals of each term of the induced system of PDEs (14) as the training loss:

$$\begin{aligned} \mathcal{L}_{PINN}^{VS}(\mathbf{w}, \mathbf{v}) = & \frac{1}{2} \left(\frac{1}{n_o} \sum_{i=1}^{n_o} \left(\hat{\mathcal{N}}[\phi_0, \dots, \phi_L](\mathbf{x}_i) - f(\mathbf{x}_i) \right)^2 \right. \\ & + \frac{\nu_\ell}{n_o} \sum_{\ell=1}^L \sum_{|\alpha| \leq \xi_\ell} \sum_{|\beta| \leq \Delta\xi_{\ell+1}} \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} (\phi_{\ell-1})_\alpha(\mathbf{x}_i) - (\phi_\ell)_{\alpha+\beta}(\mathbf{x}_i) \right)^2 \\ & \left. + \frac{\nu}{n_b} (\mathcal{B}[\phi_0](\tilde{\mathbf{x}}_j) - g(\tilde{\mathbf{x}}_j))^2 \right), \end{aligned} \quad (15)$$

where $\nu, \nu_1, \dots, \nu_L$ are regularization parameters. As the GF of $\mathcal{L}_{PINN}$ is characterized by Gram matrices $\mathbf{G}_w$ and $\mathbf{G}_v$ induced from the gradients of the residuals of each term in (1), the GF of $\mathcal{L}_{PINN}^{VS}$ is characterized by Gram matrices $\hat{\mathbf{G}}_w$ and $\hat{\mathbf{G}}_v$, which is induced from the gradients of the residuals of each term in (14). Appendix A gives more details for $\hat{\mathbf{G}}_w$ and $\hat{\mathbf{G}}_v$.

Remark 4.1. In order for high-order PDEs with $k > 2$ to be well-posed, it is necessary to have more boundary conditions than those defined by the boundary operator $\mathcal{B}$ in (1). Although our analysis concentrated on $\mathcal{B}$ that reflect only up to first-order derivatives for the sake of simplicity, our theory can also be applied to more general boundary conditions. Furthermore, the high-order boundary conditions $\mathcal{B}$ can also be reformulated using the auxiliary variables used for $\hat{\mathcal{N}}$. In that case, relations (13) should hold on the boundary $\mathbf{x} \in \partial\Omega$. As the reduced system (14) with reformulated boundary condition is equivalent to (1), instability issues were not observed in our numerical experiments even in the absence of artificial boundary conditions on the auxiliary variables unlike to grid-based conventional numerical schemes.

4.2 Analysis

A key advantage of VS-PINNs is that the derivative order of the induced system of PDEs (14) is $|\xi|$, which is lower than that of the original PDE (1). We prove its effectiveness in this section. As analogous to PINNs, we begin by proving the positive definiteness of the limiting Gram matrix, providing its proof in Appendix C.

Proposition 4.2 (General Case). *The limiting Gram matrix $\hat{\mathbf{G}}_v^\infty = \mathbb{E}_{w,v} [\hat{\mathbf{G}}_v(w, v)]$ is strictly positive definite and independent of m .*

We denote the smallest eigenvalue of $\hat{\mathbf{G}}_v^\infty$ by $\lambda_0 > 0$. We now present our main theorem, which demonstrates the profound impact of order reduction in variable splitting. The proof of the following theorem can be found in C.1.

Theorem 4.3 (General Case). *There exists a constant C , independent of d , k , $|\xi|$, and p , such that for any $\delta \ll 1$, if*

$$m > C \binom{d+k}{d}^6 \binom{d+|\xi|}{d}^8 p^{7|\xi|+4} 2^{6p} \left(\log \frac{md}{\delta} \right)^{4p}, \quad (16)$$

then with probability of at least $1 - \delta$ over the initialization, we have

$$\mathcal{L}_{PINN}^{VS}(\mathbf{w}(t), \mathbf{v}(t)) \leq \exp(-\lambda_0 t) \mathcal{L}_{PINN}^{VS}(\mathbf{w}(0), \mathbf{v}(0)), \quad \forall t \geq 0. \quad (17)$$

The right-hand-side of (16) grows exponentially with respect to d , k , and p , thereby indicating the substantial influence of these factors on the convergence of VS-PINNs, including PINNs as a specific case ($L = 0$). This analysis reveals several significant advantages of VS-PINNs:

Improved Convergence: VS-PINNs are more likely to converge to the global optimum than PINNs due to the reduction in the derivative order $|\xi| < k$. This also relaxes the condition on p from $p \geq k+1$ to $p \geq |\xi| + 1$. As previously discussed in Section 3, the optimal value of p is $|\xi| + 1$. Given that $6(|\xi| + 1)$ is an exponent of $\log(1/\delta)$, the most dominant term, reducing the order from k to $|\xi|$ leads to an immense improvement. There is another noteworthy observation we can see here. Given a k th-order PDE, there are numerous possible partitions ξ that could be employed to decompose it to a system of lower-order PDEs. Consequently, there are a many of potential VS-PINNs that could be constructed. The aforementioned result indicates which of these is the most effective. As the convergence improves dramatically with a reduction in the derivative order, the optimal approach for splitting variables among various ways is to separate the given PDE into a system of first-order PDEs by parameterizing all derivatives of the solution as auxiliary variables. In other words, the finest splitting with $\xi_0 = 0, \xi_1 = 1, \dots, \xi_{k-1} = k-1$ would be the most effective in terms of the convergence of GF, as the differential order $|\xi|$ is reduced the most to 1. Taken all together, the most optimal VS-PINNs that reduce the PDE order k to 1 will markedly enhance the convergence of GF.

Reduced Dimensional Impact: The reduction of orders in VS-PINNs enhances the resilience of the model to high-dimensionality. From Theorem 3.2, we observed the effect of d being exponentially enlarged for the PDE order k due to the k th-order partial differential operators in the loss function. It can be alleviated by VS-PINNs reducing the order, thereby easing the amplified scale to exponential of $|\xi|$. This indicates that VS-PINNs are more effective in combating the curse of dimensionality. Since the curse of dimensionality is a serious issue that is prevalent in various fields, including Hamilton-Jacobi-Bellman equation in control problems, Schrodinger equation in quantum physics, and Black-Scholes equation in finance, it is evident that enhancements to the robustness of VS-PINNs with respect to their dimensionality would facilitate considerable advancements in various fields.

Memory Efficiency: VS-PINNs are memory-efficient despite the presence of multiple auxiliary networks. As the order of the derivative increases, the complexity in automatic differentiation in modern deep-learning frameworks like PyTorch increases and it becomes computationally expensive [7]. Adopting the order-reduced representation in the proposed variable splitting can overcome the difficulty in calculating the high-order derivative via automatic differentiation. The loss function for the finest VS-PINNs involves only first-order derivatives, which reduces the memory usage and computational requirements. Despite the increase in the number of networks, VS-PINNs exhibit greater efficiency because memory usage and computation scale linearly with the number of networks in contrast to the exponential scaling with the order of derivatives. Table 3 in appendix demonstrates the memory reduction of VS-PINNs.

Remark 4.4. The current approach to parameterizing the ξ_ℓ -th order differential operator on all axes ∇^{ξ_ℓ} as an auxiliary variable may be suboptimal in certain cases. In a given PDE, if the order of the derivative varies significantly along the axes, that is, $a_\alpha \neq 0$ for only a few α in (1), it may be more efficient to approximate the partial derivatives using auxiliary variables separately for each axis. To illustrate, for the PDE $u_{tt} = u_{xxx}$, it is more suitable to parameterize variables $\phi_0 \approx u$, $\phi_1 \approx (u_t, u_x)$, $\phi_2 \approx u_{xx}$, and $\phi_3 \approx u_{xxx}$, rather than approximating all tensors $\nabla_{(t,x)}^1$, $\nabla_{(t,x)}^2$, and $\nabla_{(t,x)}^3$. The theoretical framework presented in this paper is capable of addressing this scenario by constructing each ϕ_ℓ to replace $\frac{\partial^{\xi_\ell}}{\partial x^\alpha}$ for only part of α with $|\alpha| = \xi_\ell$. However, we exclude it due to the intricate nature of the states and the lack of a meaningful impact on the PDE order.

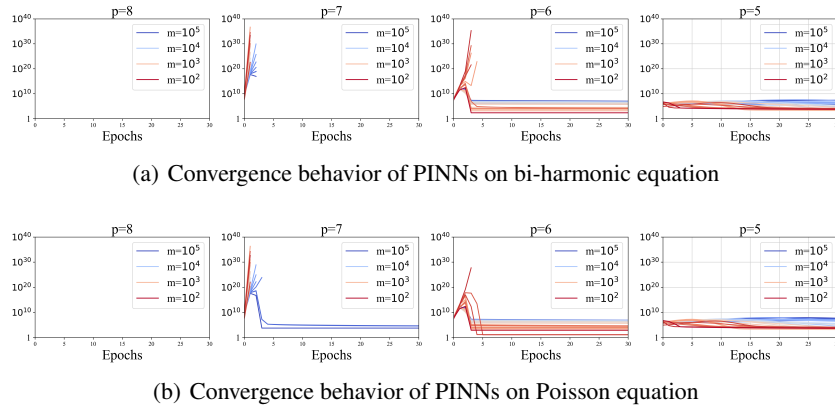


Figure 1: Training losses of PINNs solving (a) bi-harmonic equation and (b) Poisson equation.

Remark 4.5. Although the sharpness of the bound in Theorem 4.3 is open, it is important to note that the leading term of the bound is based on conditions necessary for the Gram matrix to be positive definite, which is a crucial property of the Gram matrix for ensuring the convergence of the GF to a global optimizer. Since the Gram matrix is defined by the PDE loss and the network structure, we believe it can still provide valuable insight into how order and power affect convergence.

5 Experiments

This section presents experimental results that validate the theory. Throughout numerical experiments, two-layer MLPs with ReLU^p activation function were utilized in order to align with our theoretical framework. Throughout all experiments, the training collocation points consists of uniform grid and regularization parameters are set to $\nu_1, \dots, \nu_L = 1$ and $\nu = 10$. We implement all numerical experiments on a single NVIDIA RTX 3090 GPU. Experimental details are provided in Appendix D.

Convergence behavior of PINNs To investigate the influence of the activation order p and the PDE order k on the width m required for convergence, we examined both the second-order Poisson equation and the fourth-order bi-harmonic equation, both of which yield the same solution. We trained networks with varying widths m , ranging from 10^2 to 10^6 , for each combination of p and k using GD optimization with a learning rate of 10^{-8} . Figure 1 illustrates the training losses at the initial stage on a logarithmic scale, supporting our theoretical findings that a larger width is needed for higher values of p to ensure convergence. Moreover, we can observe that narrower networks tend to converge more readily when solving lower-order PDEs (Poisson) compared to higher-order PDEs (bi-harmonic). This observation aligns with Theorem 4.3 that higher-order PDEs necessitate larger network widths for guaranteed convergence.

Validation on the effect of p To verify the influence of the power p of the ReLU activation function, we test PINNs with varying p values between 3 and 10. Since the training process became highly unstable as p increases, we consider second-order heat equation (509) [14] to gain a more precise investigation of the effect of p . The results are summarized in Figure 2 (a). We can see that the convergence of loss is enhanced as p decreases, which supports our theoretical finding.

Comparison between PINNs and VS-PINNs To validate the order reduction effect of VS-PINNs, we conducted an experiment comparing PINNs with VS-PINNs on the second-order heat equation. Each model was run five times with different random seeds, and Figure 2 depicts the training loss for both PINNs and VS-PINNs along with their variance. The results show that the training loss for VS-PINNs converges more effectively than that of PINNs. This indicates that VS-PINNs, which optimize a loss function incorporating lower-order derivatives using networks with smaller p , facilitate convergence of GD, consistent with the theoretical findings in Section 4. Furthermore, we performed a similar experiment on the convection-diffusion equation (511) in the Appendix E and obtained results that were consistent with those observed for the heat equation.

Effect of splitting level For higher-order PDEs, there are several ways to transform a given PDE into a lower-order system through variable splitting. To investigate this effect, we conducted experiments on the fourth-order elastic beam equation (510) [52] with two cases: (i) $\phi_0 \approx u$, $\phi_1 \approx u_t$, $\phi_2 \approx u_{xx}$

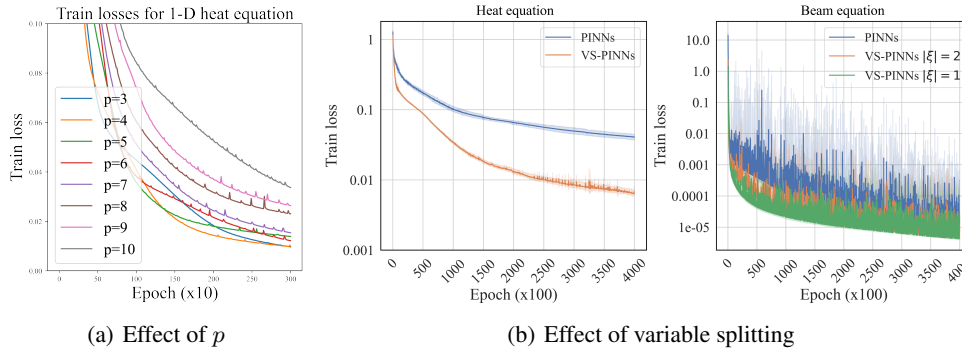


Figure 2: Loss curves of (a) effect of the power p of ReLU^p and (b) comparison between PINNs with VS-PINNs.

with $|\xi| = 2$ and $p = 3$ and (ii) the finest splitting of $\phi_0 \approx u$, $\phi_1 \approx \nabla u$, $\phi_2 \approx u_{xx}$, $\phi_3 \approx u_{xxx}$ with $|\xi| = 1$ and $p = 2$. In order to train PINNs for a fourth-order PDE, p should be at least five, but training such PINNs with GD does not proceed properly, as illustrated in Figure 1. Consequently, the experiments were conducted using the Adam optimizer. In contrast to the underperforming PINNs, VS-PINNs are effectively trained even with GD, as illustrated in Figure 6 of the Appendix E. We run each model five times with different random seeds, and Figure 2 (b) depicts the training loss of PINN and two VS-PINNs with variance. The results show that the model with a lower PDE order k and a smaller power p of the activation exhibits a more pronounced reduction in the loss function, in accordance with our theoretical findings. Furthermore, it can be observed that the variance of the training loss is significantly smaller for the models with smaller values of k and p . This indicates that the learning process is much more stable for a smaller k and p . We also conduct numerical studies on the fourth-order bi-harmonic equation (508). However, the results exhibit a similar trend to that observed in the beam equation and are therefore presented in Appendix E.

6 Conclusion

In this paper, we proved that the gradient flow of PINNs converges to a global minimum and provides sufficient width for this convergence. It extends the results in [25] to general PDEs and activation functions and provides even tighter conditions on the width size. The main theorem demonstrates that the PDE order or dimension exponentially increases the width requirement, theoretically indicating that PINNs are challenging to optimize for high-order or high-dimensional PDEs. We also substantiate that the PDE order amplifies the adverse effects of dimensionality, which explains why PINNs are more susceptible to dimensionality than other deep learning losses without differentiation. Furthermore, We showed that the variable splitting strategy improves convergence by reducing the differential order included in the training loss function.

It is acknowledged that we only provided sufficient conditions for convergence. To fully comprehend the role of these factors in optimizing PINNs, it is also necessary to establish the necessary conditions linking PDE order, dimension, and convergence. Given that the primary goal of PINNs is to approximate the solution of PDEs, it could also be a limitation that all discussions were limited to empirical losses with fixed collocation points. It would therefore be a worthwhile future direction to analyze the conditions under which the expected loss converges when training collocation points are randomly sampled per epoch. Extending our theoretical framework to analyze the impact of the variable splitting strategy on the generalization error of PINNs, as suggested in [64], would also be an interesting and important research direction. Moreover, as our analysis was confined to continuous time flows, a comprehensive understanding of gradient descent would necessitate the analysis of discretized flows, since GF and GD have different dynamics [50]. We expect that our theory could be adapted to GD dynamics by using Theorem 3.3 of [50], which treats GD as GF with a counter term, but we leave it for future work. In a practical context, the convergence of PINNs for adaptive optimizers, such as Adam [35] or L-BFGS [43], and other activation functions, including hyperbolic tangent, remains an open question.

7 Acknowledgements

This work was supported by the NRF grant [2021R1A2C3010887, RS-2024-00406127, RS-2024-00343226, RS-2024-00421203] and MSIT/IITP[NO.2021-0-01343Artificial Intelligence Graduate School Program(SNU)].

References

- [1] Ahmed Abdeljawad and Philipp Grohs. Integral representations of shallow neural network with rectified power unit activation function. *Neural Networks*, 155:536–550, 2022.
- [2] Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, Jose Santamaría, Ahmed Shihab Albahri, Bashar Sami Nayyef Al-dabbagh, Mohammed A Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali H Al-Timemy, et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, 10(1):46, 2023.
- [3] Abdon Atangana and Ernestine Alabaraoye. Solving a system of fractional partial differential equations arising in the model of hiv infection of cd4+ cells and attractor one-dimensional Keller-Segel equations. *Advances in Difference Equations*, 2013:1–14, 2013.
- [4] Roberto Ballarini. The Da Vinci-Euler-Bernoulli Beam theory? *Mechanical Engineering Magazine Online*, 2003.
- [5] Guy Barles and Halil Mete Soner. Option pricing with transaction costs and a nonlinear black-scholes equation. *Finance and Stochastics*, 2:369–397, 1998.
- [6] Shamsulhaq Basir. Investigating and mitigating failure modes in physics-informed neural networks (PINNs). *Communications in Computational Physics*, 33(5):1240–1269, 2023.
- [7] Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18(153):1–43, 2018.
- [8] Feliks Aleksandrovich Berezin and Mikhail Shubin. *The Schrödinger Equation*, volume 66. Springer Science & Business Media, 2012.
- [9] Jiang Bo-Nan and CL Chang. Least-squares finite elements for the stokes problem. *Computer Methods in Applied Mechanics and Engineering*, 78(3):297–311, 1990.
- [10] Laxman Bokati, Vladik Kreinovich, Joseph Baca, and Natasha Rovelli. Why rectified power (repu) activation functions are efficient in deep learning: A theoretical explanation. In *Uncertainty, Constraints, and Decision Making*, pages 7–13. Springer, 2023.
- [11] Kurt Izak M. Cabanilla, Rhudaina Z. Mohammad, and Jose Ernie C. Lope. Neural networks with relu powers need less depth. *Neural Networks*, 172:106073, 2024.
- [12] John W Cahn and John E Hilliard. Free energy of a nonuniform system. i. interfacial free energy. *The Journal of chemical physics*, 28(2):258–267, 1958.
- [13] 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(12):1727–1738, 2021.
- [14] John Rozier Cannon. *The one-dimensional heat equation*. Number 23. Cambridge University Press, 1984.
- [15] Charles K Chui, Xin Li, and Hrushikesh Narhar Mhaskar. Neural networks for localized approximation. *mathematics of computation*, 63(208):607–623, 1994.
- [16] Bernardo Cockburn and Chi-Wang Shu. The local discontinuous galerkin method for time-dependent convection-diffusion systems. *SIAM journal on numerical analysis*, 35(6):2440–2463, 1998.

- [17] Peter Constantin. Euler equations, Navier-Stokes equations and turbulence. In *Mathematical Foundation of Turbulent Viscous Flows: Lectures given at the CIME Summer School held in Martina Franca, Italy, SEptember 1-5, 2003*, pages 1–43. Springer, 2005.
- [18] Tim De Ryck, Florent Bonnet, Siddhartha Mishra, and Emmanuel de Bézenac. An operator preconditioning perspective on training in physics-informed machine learning. *arXiv preprint arXiv:2310.05801*, 2023.
- [19] Tim De Ryck and Siddhartha Mishra. Error analysis for physics-informed neural networks (pinns) approximating kolmogorov pdes. *Advances in Computational Mathematics*, 48(6):79, 2022.
- [20] Lokenath Debnath and Lokenath Debnath. *Nonlinear partial differential equations for scientists and engineers*. Springer, 2005.
- [21] MWMG Dissanayake and Nhan Phan-Thien. Neural-network-based approximations for solving partial differential equations. *communications in Numerical Methods in Engineering*, 10(3):195–201, 1994.
- [22] E Kirk Donald et al. Optimal control theory: an introduction. *Mineola, NY: Dover Publications, Inc*, 1970.
- [23] Nathan Doumèche, Gérard Biau, and Claire Boyer. Convergence and error analysis of pinns. *arXiv preprint arXiv:2305.01240*, 2023.
- [24] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.
- [25] Yihang Gao, Yiqi Gu, and Michael Ng. Gradient descent finds the global optima of two-layer physics-informed neural networks. In *International Conference on Machine Learning*, pages 10676–10707. PMLR, 2023.
- [26] Roland Glowinski, Hao Liu, Shingyu Leung, and Jianliang Qian. A finite element/operator-splitting method for the numerical solution of the two dimensional elliptic monge–ampère equation. *Journal of Scientific Computing*, 79:1–47, 2019.
- [27] John B Greer, Andrea L Bertozzi, and Guillermo Sapiro. Fourth order partial differential equations on general geometries. *Journal of Computational Physics*, 216(1):216–246, 2006.
- [28] Ehsan Haghighat, Maziar Raissi, Adrian Moure, Hector Gomez, and Ruben Juanes. A physics-informed deep learning framework for inversion and surrogate modeling in solid mechanics. *Computer Methods in Applied Mechanics and Engineering*, 379:113741, 2021.
- [29] QiZhi He, David Barajas-Solano, Guzel Tartakovsky, and Alexandre M Tartakovsky. Physics-informed neural networks for multiphysics data assimilation with application to subsurface transport. *Advances in Water Resources*, 141:103610, 2020.
- [30] Masanobu Horie and Naoto Mitsume. Physics-embedded neural networks: Graph neural pde solvers with mixed boundary conditions. *Advances in Neural Information Processing Systems*, 35:23218–23229, 2022.
- [31] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural networks*, 3(5):551–560, 1990.
- [32] K Hosseini, K Sadri, M Mirzazadeh, YM Chu, A Ahmadian, BA Pansera, and S Salahshour. A high-order nonlinear schrödinger equation with the weak non-local nonlinearity and its optical solitons. *Results in Physics*, 23:104035, 2021.
- [33] Zheyuan Hu, Zekun Shi, George Em Karniadakis, and Kenji Kawaguchi. Hutchinson trace estimation for high-dimensional and high-order physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 424:116883, 2024.

- [34] Deqing Jiang, Justin Sirignano, and Samuel N Cohen. Global convergence of deep galerkin and pinns methods for solving partial differential equations. *arXiv preprint arXiv:2305.06000*, 2023.
- [35] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [36] Georgios Kissas, Yibo Yang, Eileen Hwuang, Walter R Witschey, John A Detre, and Paris Perdikaris. Machine learning in cardiovascular flows modeling: Predicting arterial blood pressure from non-invasive 4d flow mri data using physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 358:112623, 2020.
- [37] Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems*, 34:26548–26560, 2021.
- [38] Isaac E Lagaris, Aristidis Likas, and Dimitrios I Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE transactions on neural networks*, 9(5):987–1000, 1998.
- [39] Hyuk Lee and In Seok Kang. Neural algorithm for solving differential equations. *Journal of Computational Physics*, 91(1):110–131, 1990.
- [40] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.
- [41] Xin Li. Simultaneous approximations of multivariate functions and their derivatives by neural networks with one hidden layer. *Neurocomputing*, 12(4):327–343, 1996.
- [42] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.
- [43] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.
- [44] Minliang Liu, Liang Liang, and Wei Sun. A generic physics-informed neural network-based constitutive model for soft biological tissues. *Computer methods in applied mechanics and engineering*, 372:113402, 2020.
- [45] Yu Liu and Wentao Ma. Gradient auxiliary physics-informed neural network for nonlinear biharmonic equation. *Engineering Analysis with Boundary Elements*, 157:272–282, 2023.
- [46] El Mehdi Lotfi, Mehdi Maziane, Khalid Hattaf, and Noura Yousfi. Partial differential equations of an epidemic model with spatial diffusion. *International Journal of Partial Differential Equations*, 2014:1–6, 2014.
- [47] Tao Luo and Haizhao Yang. Two-layer neural networks for partial differential equations: Optimization and generalization theory. *arXiv preprint arXiv:2006.15733*, 2020.
- [48] Revanth Mathey and Susanta Ghosh. A physics informed neural network for time-dependent nonlinear and higher order partial differential equations. *arXiv preprint arXiv:2106.07606*, 2021.
- [49] Siddhartha Mishra and Roberto Molinaro. Estimates on the generalization error of physics-informed neural networks for approximating a class of inverse problems for pdes. *IMA Journal of Numerical Analysis*, 42(2):981–1022, 2022.
- [50] Taiki Miyagawa. Toward equation of motion for deep neural networks: Continuous-time gradient descent and discretization error analysis. *Advances in Neural Information Processing Systems*, 35:37778–37791, 2022.

- [51] Eike H Müller and Robert Scheichl. Massively parallel solvers for elliptic partial differential equations in numerical weather and climate prediction. *Quarterly Journal of the Royal Meteorological Society*, 140(685):2608–2624, 2014.
- [52] Frithiof I Niordson. On the optimal design of a vibrating beam. *Quarterly of Applied Mathematics*, 23(1):47–53, 1965.
- [53] Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.
- [54] Yesom Park, Taekyung Lee, Jooyoung Hahn, and Myungjoo Kang. p -Poisson surface reconstruction in curl-free flow from point clouds. *Advances in Neural Information Processing Systems*, 2023.
- [55] Yesom Park, Chang Hoon Song, Jooyoung Hahn, and Myungjoo Kang. ReSDF: Redistancing implicit surfaces using neural networks. *Journal of Computational Physics*, 502, 2024.
- [56] Yesom Park, Changhoon Song, and Myungjoo Kang. Beyond derivative pathology of pinns: Variable splitting strategy with convergence analysis. *arXiv preprint arXiv:2409.20383*, 2024.
- [57] Shige Peng. A generalized dynamic programming principle and hamilton-jacobi-bellman equation. *Stochastics: An International Journal of Probability and Stochastic Processes*, 38(2):119–134, 1992.
- [58] Maziar Raissi, Paris Perdikaris, and George 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.
- [59] Justin Sirignano and Konstantinos Spiliopoulos. DGM: A deep learning algorithm for solving partial differential equations. *Journal of computational physics*, 375:1339–1364, 2018.
- [60] Roger Temam. *Navier-Stokes equations: theory and numerical analysis*, volume 343. American Mathematical Soc., 2001.
- [61] Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021.
- [62] Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why pinns fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022.
- [63] Jue Yan and Chi-Wang Shu. Local discontinuous galerkin methods for partial differential equations with higher order derivatives. *Journal of Scientific Computing*, 17:27–47, 2002.
- [64] Marius Zeinhofer, Rami Masri, and Kent-André Mardal. A unified framework for the error analysis of physics-informed neural networks. *arXiv preprint arXiv:2311.00529*, 2023.
- [65] Enrui Zhang, Ming Dao, George Em Karniadakis, and Subra Suresh. Analyses of internal structures and defects in materials using physics-informed neural networks. *Science advances*, 8(7):eabk0644, 2022.

A Mathematical Notations

We introduce the symbols and mathematical notations that are frequently used in this paper.

Symbol	Description
$\mathbb{N}_0$	The set of non-negative integers
$d \in \mathbb{N}$	dimension of domain
$\Omega \subset \mathbb{R}^d$	d -dimensional domain
$\alpha, \beta \in \mathbb{N}_0^d$	multi-index of dimension d
$k \in \mathbb{N}$	order of the governing PDE
$\mathcal{N}$	differential operator
$\hat{\mathcal{N}}$	Splitted differential operator
$\mathcal{B}$	boundary condition operator
ξ	integral partition of $[0, k]$
$f : \Omega \rightarrow \mathbb{R}$	source function
$g : \partial\Omega \rightarrow \mathbb{R}$	boundary function
$\mathbf{x}(\mathbf{y})$	(augmented)point in Ω
$\tilde{\mathbf{x}}(\tilde{\mathbf{y}})$	(augmented)point on $\partial\Omega$
ϕ	neural network
$\mathbf{G}_w, \mathbf{G}_v$	Gram matrices
λ_0	The minimal eigenvalue of $\mathbf{G}_v$
$m \in \mathbb{N}$	the number of the width of the network
σ	activation function
$p \in \mathbb{N}$	power of ReLU activation function
$n_o, n_b \in \mathbb{N}$	the number of collocation points sampled from the domain $\Omega, \partial\Omega$
$\mathbf{s}$	residual of PDE loss
$\boldsymbol{\mu}$	residual of gradient matching loss
$\mathbf{h}$	residual of boundary loss

For a positive integer m , the set $\{1, \dots, m\}$ is abbreviated as $[m]$. The set of multi-indexes whose size is m or is at most m is referred by

$$I_m = \left\{ \alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d : |\alpha| = \sum_{i=1}^d \alpha_i = m \right\}, \quad (18)$$

$$J_m = \left\{ \alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d : |\alpha| = \sum_{i=1}^d \alpha_i \leq m \right\}.$$

In this paper, we use two-layer MLPs. For given $k \in \mathbb{N}$ and a partition $0 = \xi_0 < \xi_1 < \dots < \xi_L < \xi_{L+1} = k$, we define $\phi_\ell : \mathbb{R}^d \rightarrow \mathbb{R}^{|I_{\xi_\ell}|}$ of width m as

$$\begin{aligned} \phi_\ell(\mathbf{x}; \mathbf{w}_\ell, \mathbf{v}_\ell) &= \frac{1}{\sqrt{m}} \sum_{r=1}^m (\mathbf{v}_\ell)_r \sigma \left([(\mathbf{w}_\ell)_{r,1} \quad (\mathbf{w}_\ell)_{r,2} \quad \dots \quad (\mathbf{w}_\ell)_{r,d}] \mathbf{x} + \frac{1}{2} (\mathbf{w}_\ell)_{r,d+1} \right) \\ &= \frac{1}{\sqrt{m}} \sum_{r=1}^m (\mathbf{v}_\ell)_r \sigma \left((\mathbf{w}_\ell)_r^\top \mathbf{y} \right), \end{aligned} \quad (19)$$

where weights in the first and second layer are

$$(\mathbf{w}_\ell)_r = [(\mathbf{w}_\ell)_{r,1} \quad (\mathbf{w}_\ell)_{r,2} \quad \dots \quad (\mathbf{w}_\ell)_{r,d} \quad \frac{1}{2} (\mathbf{w}_\ell)_{r,d+1}]^\top \in \mathbb{R}^{(d+1)}, \quad (20)$$

$$(\mathbf{v}_\ell)_r = [(\mathbf{v}_\ell)_{r,1} \quad (\mathbf{v}_\ell)_{r,2} \quad \dots \quad (\mathbf{v}_\ell)_{r,|I_{\xi_\ell}|}]^\top \in \mathbb{R}^{|I_{\xi_\ell}|}, \quad (21)$$

and $\mathbf{y} = [\mathbf{x}^\top \quad \frac{1}{2}]^\top \in \mathbb{R}^{d+1}$. Note that the output dimension of ϕ_ℓ is $\mathbb{R}^{|I_{\xi_\ell}|}$ because each component of $\phi_\ell(\mathbf{x}; \mathbf{w}_\ell, \mathbf{v}_\ell)$ represents partial derivative $\frac{\partial^{\xi_\ell}}{\partial \mathbf{x}^\alpha} u$, for each $\alpha \in I_{\xi_\ell}$. Hence, even though $(\mathbf{v}_\ell)_r$ is a flattened vector, it is more convenient to use $\alpha \in I_{\xi_\ell}$ as an index of component of $(\mathbf{v}_\ell)_r$. For example, if $d = 2$, $\xi_1 = 1$, and $\alpha = (0, 1) \in I_1$, $(\mathbf{v}_1)_{r,\alpha} = (\mathbf{v}_1)_{r,(0,1)}$ refers a weight between the

r -th hidden node and output that represents $\frac{\partial}{\partial y} u(x, y)$. Similarly, we use multi-index itself as an index that is related to I_{ξ_ℓ} , $I_{\Delta\xi_\ell}$, or $J_{\Delta\xi_\ell}$.

Moreover, as each $(\phi_\ell)_\alpha$ is differentiated by some $\frac{\partial^{\Delta\xi_\ell}}{\partial \mathbf{x}^\beta}$ for $\alpha \in I_{\xi_{\ell-1}}$ and $\beta \in I_{\Delta\xi_\ell}$, index we use pair of multi-index (α, β) to refer a component of $\frac{\partial^{\Delta\xi_\ell}}{\partial \mathbf{x}^\beta} (\phi_\ell)_{r,\alpha}$. Sometimes, this indexing by multi-index is used in conjunction with normal indexes, like $(\boldsymbol{\mu}_\ell)_{(\alpha,\beta),i}$ for some $i \in [n_o]$.

We denote the collection of weights of all ϕ_ℓ 's by $\mathbf{w} = [\mathbf{w}_0^\top, \dots, \mathbf{w}_L^\top]^\top$ and $\mathbf{v} = [\mathbf{v}_0^\top, \dots, \mathbf{v}_L^\top]^\top$. Similarly, $\boldsymbol{\mu}$ is the collection of all $\boldsymbol{\mu}_\ell$'s, $\boldsymbol{\mu} = [\boldsymbol{\mu}_0^\top, \dots, \boldsymbol{\mu}_L^\top]$. With regularization parameters $\nu, \nu_\ell > 0$, residuals s_i , $(\boldsymbol{\mu}_\ell)_i$, and h_j of variable splitting for each training sample

$$s_i(\mathbf{w}, \mathbf{v}) = \sqrt{\frac{1}{n_o}} \left(\hat{\mathcal{N}}[\phi_0(\cdot; \mathbf{w}_0, \mathbf{v}_0), \dots, \phi_L(\cdot; \mathbf{w}_L, \mathbf{v}_L)](\mathbf{x}_i) - f(\mathbf{x}_i) \right) \quad (22)$$

$$(\boldsymbol{\mu}_\ell)_i(\mathbf{w}, \mathbf{v}) = [(\boldsymbol{\mu}_\ell)_{(\alpha,\beta),i}(\mathbf{w}, \mathbf{v})]_{\alpha \in I_{\xi_{\ell-1}}, \beta \in I_{\Delta\xi_\ell}} \quad (23)$$

$$= \left[\frac{\partial^\beta}{\partial \mathbf{x}^\beta} (\phi_{\ell-1})_\alpha(\mathbf{x}_i; \mathbf{w}_{\ell-1}, \mathbf{v}_{\ell-1}) - \phi_\ell(\mathbf{x}_i; \mathbf{w}_\ell, \mathbf{v}_\ell)_{\alpha+\beta} \right]_{\alpha \in I_{\xi_{\ell-1}}, \beta \in I_{\Delta\xi_\ell}}, \quad (24)$$

and

$$h_j(\mathbf{w}, \mathbf{v}) = \sqrt{\frac{\nu}{n_b}} (\mathcal{B}[\phi(\cdot; \mathbf{w}_0, \mathbf{v}_0)](\tilde{\mathbf{x}}_j) - g(\tilde{\mathbf{x}}_j)) \quad (25)$$

define the loss $\mathcal{L}_{PINN}^{VS}$ for the system,

$$\mathcal{L}_{PINN}^{VS}(\mathbf{w}, \mathbf{v}) \quad (26)$$

$$= \frac{1}{2} \left(\sum_{i=1}^{n_o} s_i(\mathbf{w}, \mathbf{v})^2 + \sum_{i=1}^{n_o} \sum_{\ell=1}^L \sum_{\alpha \in I_{\xi_{\ell-1}}} \sum_{\beta \in J_{\Delta\xi_\ell}} (\boldsymbol{\mu}_\ell)_{(\alpha,\beta),i}(\mathbf{w}, \mathbf{v})^2 + \sum_{j=1}^{n_b} h_j(\mathbf{w}, \mathbf{v})^2 \right). \quad (27)$$

we define

$$a_{\max} = \max \left(\left\{ |\hat{a}_{\ell,\alpha,\beta}(\mathbf{x}_i)| : \begin{array}{l} 0 \leq \ell \leq L, i \in [n_o], \\ \alpha \in I_{\xi_\ell}, \beta \in J_{\Delta\xi_\ell} \end{array} \right\} \cup \left\{ |\tilde{a}_\alpha(\tilde{\mathbf{x}}_j)| : \begin{array}{l} \alpha \in J_1, \\ j \in [n_b] \end{array} \right\} \right), \quad (28)$$

$$\nu_{\max} = \max(\{1, \nu\} \cup \{\nu_\ell : \ell \in [L]\}). \quad (29)$$

B Calculations

The proof of the main theorem includes intricate calculations. To keep the proof clear, we separate some tedious computations that are used frequently. As this section is a reference for the proofs, we use some notation or symbols without any mention if they are defined in the other part of the paper.

Sizes of index set I_m and J_m . For any $\ell \in [L]$,

$$|I_{\xi_\ell}| = \binom{d + \xi_\ell - 1}{\xi_\ell}, |I_{\Delta\xi_\ell}| = \binom{d + \Delta\xi_\ell - 1}{\Delta\xi_\ell}, |J_{\Delta\xi_\ell}| = \binom{d + \Delta\xi_\ell}{\Delta\xi_\ell}. \quad (30)$$

We bound $|I_{\xi_\ell}|$, $|\Delta\xi_\ell|$ and $J_{\Delta\xi_\ell}$ by

$$|I_{\Delta\xi_\ell}| \leq |J_{\Delta\xi_\ell}| \leq \binom{d + |\xi|}{d}, \quad (31)$$

and

$$\sum_{\ell=0}^L |I_{\xi_\ell}| \leq |J_{\xi_L}| \leq |J_{k-1}| = \binom{d + k}{d}. \quad (32)$$

Partial derivative of $\hat{s}$, μ , and h . Note that $\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_i \right)_+^p$ in (34), $\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_i \right)_+^p$ in (35), $\left((\mathbf{w}_\ell)_r^\top \mathbf{y}_i \right)_+^p$ in (36), and $\frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p$ in (37) are polynomial of $(\mathbf{w}_\ell)_r$, $(\mathbf{w}_{\ell-1})_r$, $(\mathbf{w}_\ell)_r$, and $(\mathbf{w}_0)_r$ of degree p , respectively.

For $i \in [n_o]$, $\ell = 0, 1, \dots, L$, $r \in [m]$, and $\alpha \in I_{\xi_\ell}$,

$$\frac{\partial}{\partial (\mathbf{v}_\ell)_{r,\alpha}} s_i = \sqrt{\frac{1}{n_o}} \sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell,\alpha,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \frac{\partial}{\partial (\mathbf{v}_\ell)_{r,\alpha}} \phi_\ell(\mathbf{x}_i) \quad (33)$$

$$= \sqrt{\frac{1}{mn_o}} \sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell,\alpha,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_i \right)_+^p. \quad (34)$$

Similarly, for $\ell \in [L]$, $r \in [m]$, $\alpha_1 \in I_{\xi_{\ell-1}}$, $\alpha_2 \in I_{\xi_{\ell-1}}$, $\beta \in I_{\Delta \xi_\ell}$, and $i \in [n_o]$, we have

$$\frac{\partial}{\partial (\mathbf{v}_{\ell-1})_{r,\alpha_1}} (\mu_\ell)_{(\alpha_2,\beta),i} = \begin{cases} \sqrt{\frac{\nu_\ell}{mn_o}} \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_i \right)_+^p \right) & \text{if } \alpha_1 = \alpha_2 \\ 0 & \text{otherwise.} \end{cases} \quad (35)$$

For $\ell \in [L]$, $r \in [m]$, $\alpha_1 \in I_{\xi_\ell}$, $\alpha_2 \in I_{\xi_{\ell-1}}$, $\beta \in I_{\Delta \xi_\ell}$, and $i \in [n_o]$,

$$\frac{\partial}{\partial (\mathbf{v}_\ell)_{r,\alpha_1}} (\mu_\ell)_{(\alpha_2,\beta),i} = \begin{cases} -\sqrt{\frac{\nu_\ell}{mn_o}} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_i \right)_+^p & \text{if } \alpha_1 = \alpha_2 + \beta, \\ 0 & \text{otherwise.} \end{cases} \quad (36)$$

For $r \in [m]$ and $j \in [n_b]$,

$$\frac{\partial}{\partial (\mathbf{v}_0)_{r,1}} h_j = \sqrt{\frac{\nu}{mn_b}} \sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p. \quad (37)$$

Norms of partial derivative of $\hat{s}$, μ , and h . Note that for $i \in [n_o]$, $\ell = 0, 1, \dots, L$, $r \in [m]$,

$$\left\| \frac{\partial s_i(\mathbf{w}, \mathbf{v})}{\partial (\mathbf{w}_\ell)_r} \right\|_2 = \sqrt{\frac{1}{mn_o}} \left\| \frac{\partial}{\partial (\mathbf{w}_\ell)_r} \left(\sum_{\alpha \in I_{\xi_\ell}} \sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell,\alpha,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{v}_\ell)_r \left((\mathbf{w}_\ell)_r^\top \mathbf{y} \right)_+^p \right) \right) \right\|_2 \quad (38)$$

$$= \sqrt{\frac{1}{mn_o}} \left\| \frac{\partial}{\partial (\mathbf{w}_\ell)_r} \left(\sum_{\alpha \in I_{\xi_\ell}} \sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell,\alpha,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{v}_\ell)_{r,\alpha} \left((\mathbf{w}_\ell)_r^\top \mathbf{y} \right)_+^p \right) \right) \right\|_2 \quad (39)$$

$$= \sqrt{\frac{1}{mn_o}} \left\| \left(\sum_{\alpha \in I_{\xi_\ell}} (\mathbf{v}_\ell)_{r,\alpha} \sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell,\alpha,\beta} \frac{\partial}{\partial (\mathbf{w}_\ell)_r} \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y} \right)_+^p \right) \right) \right\|_2 \quad (40)$$

$$\leq \sqrt{\frac{1}{mn_o}} a_{\max} \sum_{\alpha \in I_{\xi_\ell}} |(\mathbf{v}_\ell)_{r,\alpha}| \sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \left\| \frac{\partial}{\partial (\mathbf{w}_\ell)_r} \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y} \right)_+^p \right) \right\|_2 \quad (41)$$

$$\leq \sqrt{\frac{1}{mn_o}} a_{\max} \sum_{\alpha \in I_{\xi_\ell}} |(\mathbf{v}_\ell)_{r,\alpha}| \sum_{\beta \in J_{\Delta \xi_{\ell+1}}} p^{\Delta \xi_{\ell+1}+1} \|\mathbf{w}_\ell\|_2^{p-1} \quad (42)$$

$$= \sqrt{\frac{1}{mn_o}} a_{\max} \sum_{\alpha \in I_{\xi_\ell}} |(\mathbf{v}_\ell)_{r,\alpha}| |J_{\Delta \xi_{\ell+1}}| p^{\Delta \xi_{\ell+1}+1} \|\mathbf{w}_\ell\|_2^{p-1} \quad (43)$$

$$\leq \sqrt{\frac{1}{mn_o}} a_{\max} \sqrt{|I_{\xi_\ell}|} \|\mathbf{v}_\ell\|_2 |J_{\Delta \xi_{\ell+1}}| p^{\Delta \xi_{\ell+1}+1} \|\mathbf{w}_\ell\|_2^{p-1}, \quad (44)$$

for $\ell \in [L]$, $r \in [m]$, $\alpha \in I_{\xi_{\ell-1}}$, $\beta \in I_{\Delta\xi_\ell}$, and $i \in [n_o]$,

$$\left\| \frac{\partial(\boldsymbol{\mu}_\ell)_{(\alpha,\beta),i}(\mathbf{w}, \mathbf{v})}{\partial(\mathbf{w}_{\ell-1})_r} \right\|_2 = \left\| \frac{\partial}{\partial(\mathbf{w}_{\ell-1})_r} \left(\sqrt{\frac{\nu_\ell}{mn_o}} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{v}_{\ell-1})_{r,\alpha} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_i \right)_+^p \right) \right) \right\|_2 \quad (45)$$

$$= \sqrt{\frac{\nu_\ell}{mn_o}} |(\mathbf{v}_{\ell-1})_{r,\alpha}| \left\| \frac{\partial}{\partial(\mathbf{w}_{\ell-1})_r} \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left(((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_i)_+^p \right) \right) \right\|_2 \quad (46)$$

$$\leq \sqrt{\frac{\nu_\ell}{mn_o}} |(\mathbf{v}_{\ell-1})_{r,\alpha}| p^{\Delta\xi_\ell+1} \|(\mathbf{w}_{\ell-1})_r\|_2^{p-1} \quad (47)$$

$$\leq \sqrt{\frac{\nu_{\max}}{mn_o}} p^{|\xi|+1} \|(\mathbf{v}_{\ell-1})_r\|_2 \|(\mathbf{w}_{\ell-1})_r\|_2^{p-1}, \quad (48)$$

$$\left\| \frac{\partial(\boldsymbol{\mu}_\ell)_{(\alpha,\beta),i}(\mathbf{w}, \mathbf{v})}{\partial(\mathbf{w}_\ell)_r} \right\|_2 = \left\| \frac{\partial}{\partial(\mathbf{w}_\ell)_r} \left(\sqrt{\frac{\nu_\ell}{mn_o}} \left((\mathbf{v}_\ell)_{r,\alpha} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_i \right)_+^p \right) \right) \right\|_2 \quad (49)$$

$$= \sqrt{\frac{\nu_\ell}{mn_o}} \left\| (\mathbf{v}_\ell)_{r,\alpha} \frac{\partial}{\partial(\mathbf{w}_\ell)_r} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_i \right)_+^p \right\|_2 \quad (50)$$

$$\leq \sqrt{\frac{\nu_\ell}{mn_o}} p |(\mathbf{v}_\ell)_{r,\alpha}| \|(\mathbf{w}_\ell)_r\|_2^{p-1} \quad (51)$$

$$\leq \sqrt{\frac{\nu_{\max}}{mn_o}} p \|(\mathbf{v}_\ell)_r\|_2 \|(\mathbf{w}_\ell)_r\|_2^{p-1}, \quad (52)$$

$$\left\| \frac{\partial(\boldsymbol{\mu}_\ell)_i(\mathbf{w}, \mathbf{v})}{\partial(\mathbf{w}_{\ell'})_r} \right\|_2 = 0, \text{ if } \ell' \neq \ell - 1, \ell, \quad (53)$$

and for $r \in [m]$ and $j \in [n_b]$,

$$\left\| \frac{\partial h_j(\mathbf{w}, \mathbf{v})}{\partial(\mathbf{w}_0)_r} \right\|_2 = \left\| \frac{\partial}{\partial(\mathbf{w}_0)_r} \left(\sqrt{\frac{\nu}{mn_b}} \sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((v_0)_{r,0} \cdot \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \right) \right) \right\|_2 \quad (54)$$

$$\leq \frac{\sqrt{\nu}}{\sqrt{mn_b}} a_{\max} \sum_{\alpha \in J_1} \left\| \frac{\partial}{\partial(\mathbf{w}_0)_r} \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((v_0)_{r,0} \cdot \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \right) \right\|_2 \quad (55)$$

$$\leq \frac{\sqrt{\nu}}{\sqrt{mn_b}} a_{\max} \sum_{\alpha \in J_1} p^2 |(v_0)_{r,0}| \|(\mathbf{w}_0)_r\|_2^{p-1} \quad (56)$$

$$\leq \sqrt{\frac{\nu_{\max}}{mn_b}} a_{\max} (d+1) p^2 \|(\mathbf{v}_0)_r\|_2 \|(\mathbf{w}_0)_r\|_2^{p-1}, \quad (57)$$

$$\left\| \frac{\partial h_j(\mathbf{w}, \mathbf{v})}{\partial(\mathbf{w}_{\ell'})_{r,1}} \right\|_2 = 0 \text{ if } \ell' \in [L]. \quad (58)$$

Similarly, for $\ell \in [L]$, $r \in [m]$, and $i \in [n_o]$, we have the following inequalities:
for $\alpha \in I_{\xi_\ell}$,

$$\left| \frac{\partial s_i(\mathbf{w}, \mathbf{v})}{\partial(\mathbf{v}_\ell)_{r,\alpha}} \right| \leq \sqrt{\frac{1}{mn_o}} \sum_{\beta \in J_{\Delta\xi_{\ell+1}}} \left| \hat{a}_{\ell,\alpha,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_i \right)_+^p \right| \quad (59)$$

$$\leq \sqrt{\frac{1}{mn_o}} a_{\max} |J_{\Delta\xi_{\ell+1}}| p^{\Delta\xi_{\ell+1}} \|(\mathbf{w}_\ell)_r\|_2^p, \quad (60)$$

$$\left\| \frac{\partial s_i(\mathbf{w}, \mathbf{v})}{\partial(\mathbf{v}_\ell)_r} \right\|_2 \leq \sqrt{\frac{\nu_{\max}}{mn_o}} a_{\max} \sqrt{|I_{\xi_\ell}|} |J_{\Delta\xi_{\ell+1}}| p^{|\xi|} \|(\mathbf{w}_\ell)_r\|_2^p \quad (61)$$

for $\alpha_1 \in I_{\xi_{\ell-1}}$, $\alpha_2 \in I_{\xi_{\ell-1}}$, and $\beta \in I_{\Delta\xi_\ell}$,

$$\left| \frac{\partial(\boldsymbol{\mu}_\ell)_{(\alpha_2,\beta),i}(\mathbf{w}(t), \mathbf{v}(t))}{\partial(\mathbf{v}_{\ell-1})_{r,\alpha_1}} \right| = \mathbf{1}_{\{\alpha_1=\alpha_2\}} \cdot \sqrt{\frac{\nu_\ell}{mn_o}} \left| \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_i \right)_+^p \right| \quad (62)$$

$$\leq \mathbf{1}_{\{\alpha_1=\alpha_2\}} \cdot \sqrt{\frac{\nu_\ell}{mn_o}} p^{\Delta\xi_\ell} \|(\mathbf{w}_{\ell-1})_r\|_2^p \quad (63)$$

$$\left\| \frac{\partial(\boldsymbol{\mu}_\ell)_{(\alpha_2,\beta),i}(\mathbf{w},\mathbf{v})}{\partial(\mathbf{v}_{\ell-1})_r} \right\|_2 \leq \sqrt{\frac{\nu_\ell}{mn_o}} p^{\Delta\xi_\ell} \sqrt{|I_{\xi_{\ell-1}}| |I_{\Delta\xi_\ell}|} \|(\mathbf{w}_{\ell-1})_r\|_2^p, \quad (64)$$

for $\alpha_1 \in I_{\xi_\ell}$, $\alpha_2 \in I_{\xi_{\ell-1}}$, and $\beta \in I_{\Delta\xi_\ell}$,

$$\left| \frac{\partial(\boldsymbol{\mu}_\ell)_{(\alpha_2,\beta),i}(\mathbf{w},\mathbf{v})}{\partial(\mathbf{v}_\ell)_{r,\alpha_1}} \right| \leq \mathbf{1}_{\{\alpha_1=\alpha_2+\beta\}} \cdot \sqrt{\frac{\nu_\ell}{mn_o}} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_i \right)_+^p \quad (65)$$

$$\leq \mathbf{1}_{\{\alpha_1=\alpha_2+\beta\}} \cdot \sqrt{\frac{\nu_\ell}{mn_o}} \|(\mathbf{w}_\ell)_r\|_2^p \quad (66)$$

$$\left\| \frac{\partial(\boldsymbol{\mu}_\ell)_{(\alpha_2,\beta),i}(\mathbf{w},\mathbf{v})}{\partial(\mathbf{v}_\ell)_r} \right\|_2 \leq \sqrt{\frac{\nu_\ell}{mn_o}} \sqrt{|I_{\xi_{\ell-1}}| |I_{\Delta\xi_\ell}|} \|(\mathbf{w}_\ell)_r\|_2^p, \quad (67)$$

for $\ell' \neq \ell - 1, \ell$,

$$\left\| \frac{\partial(\boldsymbol{\mu}_\ell)_i(\mathbf{w},\mathbf{v})}{\partial(\mathbf{v}_{\ell'})_r} \right\|_2 = 0, \quad (68)$$

and

$$\left| \frac{\partial h_j(\mathbf{w},\mathbf{v})}{\partial(\mathbf{v}_0)_{r,0}} \right| \leq \sqrt{\frac{\nu}{mn_b}} \sum_{\alpha \in J_1} \left| \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left(\left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \right) \right| \quad (69)$$

$$\leq \sqrt{\frac{\nu_{\max}}{mn_b}} (d+1) a_{\max} p \|(\mathbf{w}_0)_r\|_2^p, \quad (70)$$

$$\left\| \frac{\partial h_j(\mathbf{w},\mathbf{v})}{\partial(\mathbf{v}_{\ell'})_r} \right\|_2 = 0 \text{ if } \ell' \in [L]. \quad (71)$$

Hoeffding's inequalities for $s_{i,r}$, $(\boldsymbol{\mu}_\ell)_{r,(\alpha,\beta),i}$, and $h_{j,r}$. For each $i \in [n_o]$ and $r \in [m]$, let

$$s_i(\mathbf{w},\mathbf{v}) = \sqrt{\frac{1}{n_o}} \left(\sum_{\ell=0}^L \sum_{\alpha \in I_{\xi_\ell}} \sum_{\beta \in J_{\Delta\xi_{\ell+1}}} \hat{a}_{\ell,\alpha,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} (\phi_\ell)_\alpha - f(\mathbf{x}_i) \right) \quad (72)$$

$$= \sum_{r=1}^m s_{i,r}(\mathbf{w},\mathbf{v}) - \frac{1}{\sqrt{n_o}} f(\mathbf{x}_i), \quad (73)$$

where

$$s_{i,r}(\mathbf{w},\mathbf{v}) = \frac{1}{\sqrt{mn_o}} \sum_{\ell=0}^L \sum_{\alpha \in I_{\xi_\ell}} \sum_{\beta \in J_{\Delta\xi_{\ell+1}}} \hat{a}_{\ell,\alpha,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{v}_\ell)_{r,\alpha} \left((\mathbf{w}_\ell)_r^\top \mathbf{y} \right)_+^p \right). \quad (74)$$

As $s_{i,r}$ is a p -th degree polynomial of $(\mathbf{w}_\ell)_r$, we have

$$|s_{i,r}(\mathbf{w},\mathbf{v})| \leq \frac{a_{\max}}{\sqrt{mn_o}} \sum_{\ell=0}^L \sum_{\alpha \in I_{\xi_\ell}} \sum_{\beta \in J_{\Delta\xi_{\ell+1}}} \left| \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{v}_\ell)_{r,\alpha} \left((\mathbf{w}_\ell)_r^\top \mathbf{y} \right)_+^p \right) \right| \quad (75)$$

$$\leq \frac{a_{\max}}{\sqrt{mn_o}} \sum_{\ell=0}^L \sum_{\alpha \in I_{\xi_\ell}} \sum_{\beta \in J_{\Delta\xi_{\ell+1}}} \left| (\mathbf{v}_\ell)_{r,\alpha} \right| p^{\Delta\xi_{\ell+1}} \|(\mathbf{w}_\ell)_r\|_2^p, \quad (76)$$

for all $i \in [n_o]$ and $r \in [m]$.

Since $\|(\mathbf{w}_\ell)_r(0)\|_2 < R$ and $|(v_\ell)_{r,\alpha}(0)| \leq 1$ for all $\ell = 0, \dots, L$, $\alpha \in I_{\xi_\ell}$, and $r \in [m]$, we have

$$|s_{i,r}(\mathbf{w}(0),\mathbf{v}(0))| \leq \sqrt{\frac{1}{mn_o}} a_{\max} \sum_{\ell=0}^L \sum_{\alpha \in I_{\xi_\ell}} \sum_{\beta \in J_{\Delta\xi_{\ell+1}}} p^{\Delta\xi_{\ell+1}} R^p \quad (77)$$

$$\leq \sqrt{\frac{\nu_{\max}}{mn_o}} a_{\max} \sum_{\ell=0}^L |I_{\xi_\ell}| \cdot |J_{\Delta \xi_{\ell+1}}| \cdot p^{|\xi|} R^p \quad (78)$$

$$\leq \sqrt{\frac{\nu_{\max}}{mn_o}} a_{\max} \binom{d+k}{d} \binom{d+|\xi|}{d} p^{|\xi|} R^p, \quad (79)$$

for all $i \in [n_o]$ and $r \in [m]$.

Using $\mathbb{E}[(\mathbf{v}_\ell)_{r,\alpha}] = 0$, we attain $\mathbb{E}[s_{i,r}(\mathbf{w}(0), \mathbf{v}(0))] = 0$, and Hoeffding's inequality gives

$$\mathbb{P}\left[\left|\sum_{r=1}^m s_{i,r}(\mathbf{w}(0), \mathbf{v}(0))\right| > \frac{\varepsilon}{\sqrt{n_o}}\right] \leq 2 \exp\left(-\frac{2\left(\frac{\varepsilon}{\sqrt{n_o}}\right)^2}{m\left(2\sqrt{\frac{\nu_{\max}}{mn_o}} a_{\max} \binom{d+k}{d} \binom{d+|\xi|}{d} p^{|\xi|} R^p\right)^2}\right) \quad (80)$$

$$\leq 2 \exp\left(-\frac{\varepsilon^2}{2\nu_{\max} a_{\max}^2 \binom{d+k}{d}^2 \binom{d+|\xi|}{d}^2 p^{2|\xi|} R^{2p}}\right), \quad (81)$$

for all $\varepsilon > 0$.

Similarly, for each $\ell \in [L]$, $\alpha \in I_{\xi_{\ell-1}}$, $\beta \in I_{\Delta \xi_\ell}$, and $i \in [n_o]$,

$$(\boldsymbol{\mu}_\ell)_{(\alpha,\beta),i}(\mathbf{w}, \mathbf{v}) = \sqrt{\frac{\nu_\ell}{n_o}} \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} (\phi_{\ell-1})_\alpha(\mathbf{x}_i; \mathbf{w}_{\ell-1}, \mathbf{v}_{\ell-1}) - \phi_\ell(\mathbf{x}_i; \mathbf{w}_\ell, \mathbf{v}_\ell)_{\alpha+\beta} \right) \quad (82)$$

$$= \sum_{r=1}^m (\boldsymbol{\mu}_\ell)_{r,(\alpha,\beta),i}(\mathbf{w}, \mathbf{v}), \quad (83)$$

where

$$(\boldsymbol{\mu}_\ell)_{r,(\alpha,\beta),i}(\mathbf{w}, \mathbf{v}) = \sqrt{\frac{\nu_\ell}{mn_o}} \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{v}_{\ell-1})_{r,\alpha} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_i \right)_+^p \right) - (\mathbf{v}_\ell)_{r,\alpha+\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_i \right)_+^p \right). \quad (84)$$

Then, we attain that

$$\left| (\boldsymbol{\mu}_\ell)_{r,(\alpha,\beta),i}(\mathbf{w}, \mathbf{v}) \right| \leq \sqrt{\frac{\nu_\ell}{mn_o}} \left(\left| \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{v}_{\ell-1})_{r,\alpha} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_i \right)_+^p \right) \right| + \left| (\mathbf{v}_\ell)_{r,\alpha+\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_i \right)_+^p \right| \right) \quad (85)$$

$$\leq \sqrt{\frac{\nu_\ell}{mn_o}} \left(p^{\Delta \xi_\ell} \left| (\mathbf{v}_{\ell-1})_{r,\alpha} \right| \left\| (\mathbf{w}_{\ell-1})_r \right\|_2^p + \left| (\mathbf{v}_\ell)_{r,\alpha+\beta} \right| \left\| (\mathbf{w}_\ell)_r \right\|_2^p \right), \quad (86)$$

and

$$\left| (\boldsymbol{\mu}_\ell)_{r,(\alpha,\beta),i}(\mathbf{w}(0), \mathbf{v}(0)) \right| \leq \sqrt{\frac{\nu_{\max}}{mn_o}} \left(p^{|\xi|} + 1 \right) R^p. \quad (87)$$

$\mathbb{E}[(\mathbf{v}_\ell)_{r,\alpha}] = 0$ implies $\mathbb{E}[(\boldsymbol{\mu}_\ell)_{r,(\alpha,\beta),i}] = 0$ and for each $i \in [n_o]$, Hoeffding's inequality with the sum over $r \in [m]$ of variables

$$X_{r,i}(\mathbf{w}(0), \mathbf{v}(0)) = \sum_{\alpha \in I_{\xi_{\ell-1}}} \sum_{\beta \in I_{\Delta \xi_\ell}} (\boldsymbol{\mu}_\ell)_{r,(\alpha,\beta),i}(\mathbf{w}(0), \mathbf{v}(0)), \quad (88)$$

with

$$|X_{r,i}(\mathbf{w}(0), \mathbf{v}(0))| \leq |I_{\xi_{\ell-1}}| \cdot |I_{\Delta \xi_\ell}| \sqrt{\frac{\nu_{\max}}{mn_o}} \left(p^{|\xi|} + 1 \right) R^p \quad (89)$$

$$< \sqrt{\frac{\nu_{\max}}{mn_o}} \binom{d+k}{d} \binom{d+|\xi|}{d} \left(p^{|\xi|} + 1 \right) R^p \quad (90)$$

gives

$$\mathbb{P} \left[\left| \sum_{r=1}^m (\boldsymbol{\mu}_\ell)_{r,(\alpha,\beta),i} (\mathbf{w}(0), \mathbf{v}(0)) \right| > \frac{\varepsilon}{\sqrt{n_o}} \right] \quad (91)$$

$$\leq 2 \exp \left(- \frac{2 \left(\frac{\varepsilon}{\sqrt{n_o}} \right)^2}{m \left(\sqrt{\frac{\nu_{\max}}{mn_o}} \binom{d+k}{d} \binom{d+|\xi|}{d} (p^{|\xi|} + 1) R^p \right)^2} \right) \quad (92)$$

$$\leq 2 \exp \left(- \frac{\varepsilon^2}{\nu_{\max} \binom{d+k}{d}^2 \binom{d+|\xi|}{d}^2 (p^{|\xi|} + 1)^2 R^{2p}} \right). \quad (93)$$

For each $j \in [n_b]$, let us denote

$$h_j(\mathbf{w}, \mathbf{v}) = \sqrt{\frac{\nu}{n_b}} \sum_{\alpha \in J_1} \tilde{a}_\alpha \phi_0(\tilde{\mathbf{x}}_j; \mathbf{w}_0, \mathbf{v}_0) - \sqrt{\frac{\nu}{n_b}} g(\tilde{\mathbf{x}}_j) \quad (94)$$

$$= \sum_{r=1}^m h_{j,r}((\mathbf{w}_0)_r, (\mathbf{v}_0)_r) - \sqrt{\frac{\nu}{n_b}} g(\tilde{\mathbf{x}}_j), \quad (95)$$

where

$$h_{j,r}(\mathbf{w}, \mathbf{v}) = \sqrt{\frac{\nu}{mn_b}} \sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{v}_0)_{r,\mathbf{0}} \cdot \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \right). \quad (96)$$

It is clear that $h_{j,r}$ is expressed as a p -th order polynomial of $(\mathbf{w}_0)_r$ and we can deduce that

$$|h_{j,r}(\mathbf{w}(0), \mathbf{v}(0))| \leq \sqrt{\frac{\nu_{\max}}{mn_b}} a_{\max} (d+1) p R^p, \quad (97)$$

for all $j \in [n_b]$ and $r \in [m]$.

Similar to the case of $s_{i,r}$ and $(\boldsymbol{\mu}_\ell)_{r,(\alpha,\beta),i}$, Hoeffding's inequality gives

$$\mathbb{P} \left[\left| \sum_{r=1}^m h_{j,r}(\mathbf{w}(0), \mathbf{v}(0)) \right| > \frac{\varepsilon}{\sqrt{n_b}} \right] \leq 2 \exp \left(- \frac{\varepsilon^2}{2 \nu_{\max} a_{\max}^2 (d+1)^2 p^2 R^{2p}} \right). \quad (98)$$

Components of $\mathbf{Q}_{\ell_1, \ell_2}$ Each component of $\mathbf{Q}_{\ell_1, \ell_2}$ is polynomial of order p as follows:

$$(\mathbf{Q}_{0,0})_{i_1, i_2} \quad (99)$$

$$= \sum_{\ell=0}^L \sum_{r=1}^m \sum_{\alpha \in I_{\xi_\ell}} \left(\frac{\partial}{\partial (\mathbf{v}_\ell)_{r,\alpha}} s_{i_1} \right) \left(\frac{\partial}{\partial (\mathbf{v}_\ell)_{r,\alpha}} s_{i_2} \right) \quad (100)$$

$$= \frac{1}{mn_o} \sum_{\ell=0}^L \sum_{r=1}^m \sum_{\alpha \in I_{\xi_\ell}} \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right). \quad (101)$$

$$(\mathbf{Q}_{0,\ell})_{i_1, ((\alpha, \beta), i_2)} \quad (102)$$

$$= \sum_{\ell'=0}^L \sum_{r=1}^m \sum_{\alpha' \in I_{\xi_{\ell'}}} \left(\frac{\partial}{\partial (\mathbf{v}_{\ell'})_{r,\alpha'}} s_{i_1} \right) \left(\frac{\partial}{\partial (\mathbf{v}_{\ell'})_{r,\alpha'}} (\boldsymbol{\mu}_\ell)_{(\alpha, \beta), i_2} \right) \quad (103)$$

$$= \sum_{r=1}^m \sum_{\alpha' \in I_{\xi_{\ell-1}}} \left(\frac{\partial}{\partial (\mathbf{v}_{\ell-1})_{r,\alpha'}} s_{i_1} \right) \left(\frac{\partial}{\partial (\mathbf{v}_{\ell-1})_{r,\alpha'}} (\boldsymbol{\mu}_\ell)_{(\alpha, \beta), i_2} \right) \quad (104)$$

$$+ \sum_{r=1}^m \sum_{\alpha' \in I_{\xi_\ell}} \left(\frac{\partial}{\partial (\mathbf{v}_\ell)_{r,\alpha'}} s_{i_1} \right) \left(\frac{\partial}{\partial (\mathbf{v}_\ell)_{r,\alpha'}} (\boldsymbol{\mu}_\ell)_{(\alpha,\beta),i_2} \right) \quad (105)$$

$$= \sum_{r=1}^m \sum_{\alpha' \in I_{\xi_{\ell-1}}} \left(\frac{\partial}{\partial (\mathbf{v}_{\ell-1})_{r,\alpha'}} s_{i_1} \right) \left(\frac{\partial}{\partial (\mathbf{v}_{\ell-1})_{r,\alpha'}} (\boldsymbol{\mu}_\ell)_{(\alpha,\beta),i_2} \right) \quad (106)$$

$$+ \sum_{r=1}^m \left(\frac{\partial}{\partial (\mathbf{v}_\ell)_{r,\alpha+\beta}} s_{i_1} \right) \left(\frac{\partial}{\partial (\mathbf{v}_\ell)_{r,\alpha+\beta}} (\boldsymbol{\mu}_\ell)_{(\alpha,\beta),i_2} \right) \quad (107)$$

$$= \frac{\sqrt{\nu_\ell}}{mn_o} \sum_{r=1}^m \left(\left(\sum_{\beta' \in J_{\Delta \xi_\ell}} \hat{a}_{\ell-1,\alpha,\beta'} \frac{\partial^{\beta'}}{\partial \mathbf{x}^{\beta'}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \right. \quad (108)$$

$$\left. - \left(\sum_{\beta' \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell,\alpha,\beta'} \frac{\partial^{\beta'}}{\partial \mathbf{x}^{\beta'}} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right). \quad (109)$$

$$(\mathbf{Q}_{0,L+1})_{i,j} = \sum_{\ell=0}^L \sum_{r=1}^m \sum_{\alpha \in I_{\xi_\ell}} \left(\frac{\partial}{\partial (\mathbf{v}_\ell)_{r,\alpha}} s_i \right) \left(\frac{\partial}{\partial (\mathbf{v}_\ell)_{r,\alpha}} h_j \right) \quad (110)$$

$$= \sum_{r=1}^m \left(\frac{\partial}{\partial (\mathbf{v}_0)_{r,1}} s_i \right) \left(\frac{\partial}{\partial (\mathbf{v}_0)_{r,1}} h_j \right) \quad (111)$$

$$= \frac{1}{m} \sqrt{\frac{\nu}{n_o n_b}} \sum_{r=1}^m \left(\sum_{\beta \in J_{\Delta \xi_1}} \hat{a}_{0,1,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_0)_r^\top \mathbf{y}_i \right)_+^p \right) \left(\sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \right). \quad (112)$$

$$(\mathbf{Q}_{\ell,\ell})_{((\alpha_1,\beta_1),i_1),((\alpha_2,\beta_2),i_2)} \quad (113)$$

$$= \sum_{\ell'=0}^L \sum_{r=1}^m \sum_{\alpha \in I_{\xi_{\ell'}}} \left(\frac{\partial}{\partial (\mathbf{v}_{\ell'})_{r,\alpha}} (\boldsymbol{\mu}_\ell)_{(\alpha_1,\beta_1),i_1} \right) \left(\frac{\partial}{\partial (\mathbf{v}_{\ell'})_{r,\alpha}} (\boldsymbol{\mu}_\ell)_{(\alpha_2,\beta_2),i_2} \right) \quad (114)$$

$$= \sum_{r=1}^m \sum_{\alpha \in I_{\xi_{\ell-1}}} \left(\frac{\partial}{\partial (\mathbf{v}_{\ell-1})_{r,\alpha}} (\boldsymbol{\mu}_\ell)_{(\alpha_1,\beta_1),i_1} \right) \left(\frac{\partial}{\partial (\mathbf{v}_{\ell-1})_{r,\alpha}} (\boldsymbol{\mu}_\ell)_{(\alpha_2,\beta_2),i_2} \right) \quad (115)$$

$$+ \sum_{r=1}^m \sum_{\alpha \in I_{\xi_\ell}} \left(\frac{\partial}{\partial (\mathbf{v}_\ell)_{r,\alpha}} (\boldsymbol{\mu}_\ell)_{(\alpha_1,\beta_1),i_1} \right) \left(\frac{\partial}{\partial (\mathbf{v}_\ell)_{r,\alpha}} (\boldsymbol{\mu}_\ell)_{(\alpha_2,\beta_2),i_2} \right) \quad (116)$$

$$= \sum_{r=1}^m \left(\sqrt{\frac{\nu_\ell}{mn_o}} \frac{\partial^{\beta_1}}{\partial \mathbf{x}^{\beta_1}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\sqrt{\frac{\nu_\ell}{mn_o}} \frac{\partial^{\beta_2}}{\partial \mathbf{x}^{\beta_2}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \mathbf{1}_{\{\alpha_1=\alpha_2\}} \quad (117)$$

$$+ \sum_{r=1}^m \left(-\sqrt{\frac{\nu_\ell}{mn_o}} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(-\sqrt{\frac{\nu_\ell}{mn_o}} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \mathbf{1}_{\{\alpha_1+\beta_1=\alpha_2+\beta_2\}} \quad (118)$$

$$= \frac{\nu_\ell}{mn_o} \sum_{r=1}^m \left(\left(\frac{\partial^{\beta_1}}{\partial \mathbf{x}^{\beta_1}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\frac{\partial^{\beta_2}}{\partial \mathbf{x}^{\beta_2}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \mathbf{1}_{\{\alpha_1=\alpha_2\}} \right. \quad (119)$$

$$\left. + \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \mathbf{1}_{\{\alpha_1+\beta_1=\alpha_2+\beta_2\}} \right). \quad (120)$$

$$(\mathbf{Q}_{\ell,\ell+1})_{((\alpha_1,\beta_1),i_1),((\alpha_2,\beta_2),i_2)} \quad (121)$$

$$= \sum_{\ell'=0}^L \sum_{r=1}^m \sum_{\alpha \in I_{\xi_{\ell'}}} \left(\frac{\partial}{\partial (\mathbf{v}_{\ell'})_{r,\alpha}} (\boldsymbol{\mu}_{\ell})_{(\alpha_1, \beta_1), i_1} \right) \left(\frac{\partial}{\partial (\mathbf{v}_{\ell'})_{r,\alpha}} (\boldsymbol{\mu}_{\ell+1})_{(\alpha_2, \beta_2), i_2} \right) \quad (122)$$

$$= \sum_{r=1}^m \sum_{\alpha \in I_{\xi_{\ell}}} \left(\frac{\partial}{\partial (\mathbf{v}_{\ell})_{r,\alpha}} (\boldsymbol{\mu}_{\ell})_{(\alpha_1, \beta_1), i_1} \right) \left(\frac{\partial}{\partial (\mathbf{v}_{\ell})_{r,\alpha}} (\boldsymbol{\mu}_{\ell+1})_{(\alpha_2, \beta_2), i_2} \right) \quad (123)$$

$$= \sum_{r=1}^m \sum_{\alpha \in I_{\xi_{\ell}}} \left(-\sqrt{\frac{\nu_{\ell}}{mn_o}} \left((\mathbf{w}_{\ell})_r^{\top} \mathbf{y}_{i_1} \right)_+^p \right) \mathbf{1}_{\{\alpha = \alpha_1 + \beta_1\}} \left(\sqrt{\frac{\nu_{\ell+1}}{mn_o}} \frac{\partial^{\beta_2}}{\partial \mathbf{x}^{\beta_2}} \left((\mathbf{w}_{\ell})_r^{\top} \mathbf{y}_{i_2} \right)_+^p \right) \mathbf{1}_{\{\alpha = \alpha_2\}} \quad (124)$$

$$= \frac{\sqrt{\nu_{\ell} \nu_{\ell+1}}}{mn_o} \sum_{r=1}^m \left(- \left((\mathbf{w}_{\ell})_r^{\top} \mathbf{y}_{i_1} \right)_+^p \right) \left(\frac{\partial^{\beta_2}}{\partial \mathbf{x}^{\beta_2}} \left((\mathbf{w}_{\ell})_r^{\top} \mathbf{y}_{i_2} \right)_+^p \right) \mathbf{1}_{\{\alpha_2 = \alpha_1 + \beta_1\}}. \quad (125)$$

$$(\mathbf{Q}_{\ell, L+1})_{((\alpha, \beta), i), j} \quad (126)$$

$$= \sum_{\ell'=0}^L \sum_{r=1}^m \sum_{\alpha' \in I_{\xi_{\ell'}}} \left(\frac{\partial}{\partial (\mathbf{v}_{\ell'})_{r,\alpha'}} (\boldsymbol{\mu}_{\ell})_{(\alpha, \beta), i} \right) \left(\frac{\partial}{\partial (\mathbf{v}_{\ell'})_{r,\alpha'}} h_j \right) \quad (127)$$

$$= \sum_{r=1}^m \left(\frac{\partial}{\partial (\mathbf{v}_0)_{r,1}} (\boldsymbol{\mu}_{\ell})_{(\alpha, \beta), i} \right) \left(\frac{\partial}{\partial (\mathbf{v}_0)_{r,1}} h_j \right) \quad (128)$$

$$= \sum_{r=1}^m \left(\sqrt{\frac{\nu_1}{mn_o}} \frac{\partial^{\beta}}{\partial \mathbf{x}^{\beta}} \left((\mathbf{w}_0)_r^{\top} \mathbf{y}_i \right)_+^p \right) \left(\sqrt{\frac{\nu}{mn_b}} \sum_{\alpha \in J_1} \tilde{a}_{\alpha} \left((\mathbf{w}_0)_r^{\top} \tilde{\mathbf{y}}_j \right)_+^p \right) \mathbf{1}_{\{\ell=1\}} \quad (129)$$

$$= \frac{1}{m} \sqrt{\frac{\nu_1 \nu}{n_o n_b}} \sum_{r=1}^m \left(\frac{\partial^{\beta}}{\partial \mathbf{x}^{\beta}} \left((\mathbf{w}_0)_r^{\top} \mathbf{y}_i \right)_+^p \right) \left(\sum_{\alpha \in J_1} \tilde{a}_{\alpha} \frac{\partial^{\alpha}}{\partial \mathbf{x}^{\alpha}} \left((\mathbf{w}_0)_r^{\top} \tilde{\mathbf{y}}_j \right)_+^p \right) \mathbf{1}_{\{\ell=1\}}. \quad (130)$$

$$(\mathbf{Q}_{L+1, L+1})_{j_1, j_2} \quad (131)$$

$$= \sum_{\ell=0}^L \sum_{r=1}^m \sum_{\alpha \in I_{\xi_{\ell}}} \left(\frac{\partial}{\partial (\mathbf{v}_{\ell})_{r,\alpha}} h_{j_1} \right) \left(\frac{\partial}{\partial (\mathbf{v}_{\ell})_{r,\alpha}} h_{j_2} \right) \quad (132)$$

$$= \sum_{r=1}^m \left(\frac{\partial}{\partial (\mathbf{v}_0)_{r,1}} h_{j_1} \right) \left(\frac{\partial}{\partial (\mathbf{v}_0)_{r,1}} h_{j_2} \right) \quad (133)$$

$$= \sum_{r=1}^m \left(\sqrt{\frac{\nu}{mn_b}} \sum_{\alpha \in J_1} \tilde{a}_{\alpha} \frac{\partial^{\alpha}}{\partial \mathbf{x}^{\alpha}} \left((\mathbf{w}_0)_r^{\top} \tilde{\mathbf{y}}_{j_1} \right)_+^p \right) \left(\sqrt{\frac{\nu}{mn_b}} \sum_{\alpha \in J_1} \tilde{a}_{\alpha} \frac{\partial^{\alpha}}{\partial \mathbf{x}^{\alpha}} \left((\mathbf{w}_0)_r^{\top} \tilde{\mathbf{y}}_{j_2} \right)_+^p \right) \quad (134)$$

$$= \frac{\nu}{mn_b} \sum_{r=1}^m \left(\sum_{\alpha \in J_1} \tilde{a}_{\alpha} \frac{\partial^{\alpha}}{\partial \mathbf{x}^{\alpha}} \left((\mathbf{w}_0)_r^{\top} \tilde{\mathbf{y}}_{j_1} \right)_+^p \right) \left(\sum_{\alpha \in J_1} \tilde{a}_{\alpha} \frac{\partial^{\alpha}}{\partial \mathbf{x}^{\alpha}} \left((\mathbf{w}_0)_r^{\top} \tilde{\mathbf{y}}_{j_2} \right)_+^p \right). \quad (135)$$

The number of possible nonzero elements in $\mathbf{Q}_{\ell_1, \ell_2}$ and $\hat{\mathbf{G}}_{\mathbf{v}}$. There are possibly n_o^2 , $|I_{\xi_{\ell-1}}| |I_{\Delta \xi_{\ell}}| n_o^2$, and $n_o n_b$ nonzero elements in $\mathbf{Q}_{0,0}$, $\mathbf{Q}_{0,\ell}$, and $\mathbf{Q}_{0,L+1}$, respectively. For each $(\alpha_1, \beta_1) \in I_{\xi_{\ell-1}} \times I_{\Delta \xi_{\ell}}$, there are at most $|I_{\Delta \xi_{\ell}}|$ pairs of $(\alpha_2, \beta_2) \in I_{\xi_{\ell-1}} \times I_{\Delta \xi_{\ell}}$ such that $\alpha_1 + \beta_1 = \alpha_2 + \beta_2$ and at most $|I_{\Delta \xi_{\ell}}|$ pairs of $(\alpha_2, \beta_2) \in I_{\xi_{\ell-1}} \times I_{\Delta \xi_{\ell}}$ such that $\alpha_1 = \alpha_2$. Thus, there are at most $2 |I_{\Delta \xi_{\ell}}|$ pairs of $(\alpha_2, \beta_2) \in I_{\xi_{\ell-1}} \times I_{\Delta \xi_{\ell}}$ with $\alpha_1 + \beta_1 = \alpha_2 + \beta_2$ or $\alpha_1 = \alpha_2$, and $\mathbf{Q}_{\ell, \ell}$ has possibly $(|I_{\xi_{\ell-1}}| |I_{\Delta \xi_{\ell}}| n_o) (2 |I_{\Delta \xi_{\ell}}| n_o) = 2 |I_{\xi_{\ell-1}}| \cdot |I_{\Delta \xi_{\ell}}|^2 n_o^2$ nonzero elements. Similarly, each $(\alpha_1, \beta_1) \in I_{\xi_{\ell-1}} \times I_{\Delta \xi_{\ell}}$ uniquely determines $\alpha_2 = \alpha_1 + \beta_1 \in I_{\xi_{\ell}}$, there are $|I_{\Delta \xi_{\ell+1}}|$ pairs of $(\alpha_2, \beta_2) \in I_{\xi_{\ell}} \times I_{\Delta \xi_{\ell+1}}$ with $\alpha_2 = \alpha_1 + \beta_1$. Hence, $\mathbf{Q}_{\ell, \ell+1}$ has $|I_{\xi_{\ell-1}}| \cdot |I_{\Delta \xi_{\ell}}| \cdot |I_{\Delta \xi_{\ell+1}}| n_o^2$ nonzero elements. $\mathbf{Q}_{1, L+1}$ has $|I_{\Delta \xi_1}| n_o n_b$ nonzero elements. The number of nonzero elements in $\mathbf{Q}_{L+1, L+1}$ is n_o^2 . After all, the number of nonzero elements in $\hat{\mathbf{G}}_{\mathbf{v}}$ is

Matrix	The number of nonzero elements
$\mathbf{Q}_{0,0}$	n_o^2
$\mathbf{Q}_{0,\ell}$	$ I_{\xi_{\ell-1}} I_{\Delta\xi_\ell} n_o^2$
$\mathbf{Q}_{0,L+1}$	$n_o n_b$
$\mathbf{Q}_{\ell,\ell}$	$2 I_{\xi_{\ell-1}} \cdot I_{\Delta\xi_\ell} ^2 n_o^2$
$\mathbf{Q}_{\ell,\ell+1}$	$ I_{\xi_{\ell-1}} \cdot I_{\Delta\xi_\ell} \cdot I_{\Delta\xi_{\ell+1}} n_o^2$
$\mathbf{Q}_{1,L+1}$	$ I_{\Delta\xi_1} n_o n_b$
$\mathbf{Q}_{L+1,L+1}$	n_b^2

Table 1: The number of nonzero elements in blocks in $\mathbf{G}$.

at most $N_1 = 7 \binom{d+k}{d} \binom{d+|\xi|}{d}^2 n_o^2 + 4 \binom{d+|\xi|}{d} n_o n_b + n_b^2$, because

$$\left(\text{the number of nonzero elements of } \widehat{\mathbf{G}}_v \right) \quad (136)$$

$$= \sum_{\ell=0}^{L+1} \left(\text{the number of nonzero elements in } \mathbf{Q}_{\ell,\ell} \right) \quad (137)$$

$$+ 2 \sum_{\ell=0}^L \sum_{\ell'=\ell+1}^{L+1} \left(\text{the number of nonzero elements in } \mathbf{Q}_{\ell,\ell'} \right) \quad (138)$$

$$\leq n_o^2 + \sum_{\ell=1}^L \left(2 |I_{\xi_{\ell-1}}| |I_{\Delta\xi_\ell}|^2 n_o^2 \right) + n_b^2 \quad (139)$$

$$+ 2 \left\{ \sum_{\ell=1}^L \left(|I_{\xi_{\ell-1}}| |I_{\Delta\xi_\ell}| n_o^2 \right) + n_o n_b + \sum_{\ell=1}^L \left(|I_{\xi_{\ell-1}}| |I_{\Delta\xi_\ell}| |I_{\Delta\xi_{\ell+1}}| n_o^2 \right) + |I_{\Delta\xi_1}| n_o n_b \right\} \quad (140)$$

$$\leq n_o^2 + 2 \binom{d+|\xi|}{d}^2 \sum_{\ell=1}^L |I_{\xi_{\ell-1}}| n_o^2 + n_b^2 \quad (141)$$

$$+ 2 \left\{ \binom{d+|\xi|}{d} \sum_{\ell=1}^L |I_{\xi_{\ell-1}}| n_o^2 + n_o n_b + \binom{d+|\xi|}{d}^2 \sum_{\ell=1}^L |I_{\xi_{\ell-1}}| n_o^2 + \binom{d+|\xi|}{d} n_o n_b \right\} \quad (142)$$

$$\leq \left(1 + 2 \binom{d+|\xi|}{d} \binom{d+k}{d} + 4 \binom{d+|\xi|}{d}^2 \binom{d+k}{d} \right) n_o^2 + 2 \left(1 + \binom{d+|\xi|}{d} \right) n_o n_b + n_b^2 \quad (143)$$

$$\leq 7 \binom{d+k}{d} \binom{d+|\xi|}{d}^2 n_o^2 + 4 \binom{d+|\xi|}{d} n_o n_b + n_b^2 \quad (144)$$

$$= N_1. \quad (145)$$

Norms of partial derivative of $\mathbf{Q}_{\ell_1,\ell_2}$. For each $i_1, i_2 \in [n_o]$, $r \in [m]$, and $\ell \in \{0\} \cup [L]$, we have

$$(\mathbf{Q}_{0,0})_{i_1,i_2} \quad (146)$$

$$= \frac{1}{mn_o} \sum_{\ell=0}^L \sum_{r=1}^m \sum_{\alpha \in I_{\xi_\ell}} \left(\sum_{\beta \in J_{\Delta\xi_{\ell+1}}} \hat{a}_{\ell,\alpha,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\sum_{\beta \in J_{\Delta\xi_{\ell+1}}} \hat{a}_{\ell,\alpha,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right), \quad (147)$$

$$\left\| \frac{\partial}{\partial (\mathbf{w}_\ell)_r} (\mathbf{Q}_{0,0})_{i_1,i_2} \right\|_2 \quad (148)$$

$$= \left\| \frac{1}{mn_o} \frac{\partial}{\partial (\mathbf{w}_\ell)_r} \left\{ \sum_{\alpha \in I_{\xi_\ell}} \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \right\} \right\|_2 \quad (149)$$

$$\leq \frac{1}{mn_o} |I_{\xi_\ell}| a_{\max}^2 \left\| \frac{\partial}{\partial (\mathbf{w}_\ell)_r} \left\{ \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \right\} \right\|_2 \quad (150)$$

$$\leq \frac{1}{mn_o} a_{\max}^2 |I_{\xi_\ell}| |J_{\Delta \xi_{\ell+1}}|^2 p^{2|\xi|+1} \|(\mathbf{w}_\ell)_r\|_2^{2p-1}, \quad (151)$$

$$\frac{\partial}{\partial (\mathbf{v}_\ell)_r} (\mathbf{Q}_{0,0})_{i_1, i_2} = \mathbf{0}. \quad (152)$$

For each $\ell \in [L]$, $\alpha \in I_{\xi_{\ell-1}}$, $\beta \in I_{\Delta \xi_\ell}$, and $i_1, i_2 \in [n_o]$,

$$(\mathbf{Q}_{0,\ell})_{i_1, ((\alpha, \beta), i_2)} \quad (153)$$

$$= \frac{\sqrt{\nu_\ell}}{mn_o} \sum_{r=1}^m \left(\left(\sum_{\beta' \in J_{\Delta \xi_\ell}} \hat{a}_{\ell-1, \alpha, \beta'} \frac{\partial^{\beta'}}{\partial \mathbf{x}^{\beta'}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \right. \quad (154)$$

$$\left. - \left(\sum_{\beta' \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta'} \frac{\partial^{\beta'}}{\partial \mathbf{x}^{\beta'}} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right), \quad (155)$$

$$\left\| \frac{\partial}{\partial (\mathbf{w}_{\ell-1})_r} (\mathbf{Q}_{0,\ell})_{i_1, ((\alpha, \beta), i_2)} \right\|_2 \quad (156)$$

$$= \left\| \frac{\sqrt{\nu_\ell}}{mn_o} \frac{\partial}{\partial (\mathbf{w}_{\ell-1})_r} \left\{ \left(\sum_{\beta' \in J_{\Delta \xi_\ell}} \hat{a}_{\ell-1, \alpha, \beta'} \frac{\partial^{\beta'}}{\partial \mathbf{x}^{\beta'}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \right\} \right\|_2 \quad (157)$$

$$\leq \frac{\sqrt{\nu_\ell}}{mn_o} a_{\max} |J_{\Delta \xi_\ell}| p^{2\Delta \xi_\ell+1} \|(\mathbf{w}_{\ell-1})_r\|_2^{2p-1} \quad (158)$$

$$\leq \frac{\sqrt{\nu_{\max}}}{mn_o} a_{\max} \binom{d+|\xi|}{d} p^{2|\xi|+1} \|(\mathbf{w}_{\ell-1})_r\|_2^{2p-1}, \quad (159)$$

and

$$\left\| \frac{\partial}{\partial (\mathbf{w}_\ell)_r} (\mathbf{Q}_{0,\ell})_{i_1, ((\alpha, \beta), i_2)} \right\|_2 \quad (160)$$

$$= \left\| -\frac{\sqrt{\nu_\ell}}{mn_o} \frac{\partial}{\partial (\mathbf{w}_\ell)_r} \left(\left(\sum_{\beta' \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta'} \frac{\partial^{\beta'}}{\partial \mathbf{x}^{\beta'}} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \right\|_2 \quad (161)$$

$$\leq \frac{\sqrt{\nu_\ell}}{mn_o} a_{\max} |J_{\Delta \xi_{\ell+1}}| p^{\Delta \xi_{\ell+1}+1} \|(\mathbf{w}_\ell)_r\|_2^{2p-1} \quad (162)$$

$$\leq \frac{\sqrt{\nu_{\max}}}{mn_o} a_{\max} \binom{d+|\xi|}{d} p^{2|\xi|+1} \|(\mathbf{w}_\ell)_r\|_2^{2p-1}, \quad (163)$$

$$\frac{\partial}{\partial (\mathbf{w}_{\ell'})_r} (\mathbf{Q}_{0,\ell})_{i_1, ((\alpha, \beta), i_2)} = \mathbf{0}, \text{ if } \ell' \neq \ell-1, \ell, \quad (164)$$

$$\frac{\partial}{\partial (\mathbf{v}_{\ell'})_r} (\mathbf{Q}_{0,\ell})_{i_1, ((\alpha, \beta), i_2)} = \mathbf{0}, \forall \ell' \in [L]. \quad (165)$$

For each $i \in [n_o]$ and $j \in [n_b]$,

$$(\mathbf{Q}_{0,L+1})_{i,j} \quad (166)$$

$$= \frac{1}{m} \sqrt{\frac{\nu}{n_o n_b}} \sum_{r=1}^m \left(\sum_{\beta \in J_{\Delta \xi_1}} \hat{a}_{0,1,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_0)_r^\top \mathbf{y}_i \right)_+^p \right) \left(\sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \right), \quad (167)$$

$$\left\| \frac{\partial}{\partial (\mathbf{w}_0)_r} (\mathbf{Q}_{0,L+1})_{i,j} \right\|_2 \quad (168)$$

$$= \left\| \frac{1}{m} \sqrt{\frac{\nu}{n_o n_b}} \frac{\partial}{\partial (\mathbf{w}_0)_r} \left\{ \left(\sum_{\beta \in J_{\Delta \xi_1}} \hat{a}_{0,1,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_0)_r^\top \mathbf{y}_i \right)_+^p \right) \left(\sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \right) \right\} \right\|_2 \quad (169)$$

$$\leq \frac{1}{m} \sqrt{\frac{\nu}{n_o n_b}} \binom{d + \Delta \xi_1}{d} (d+1) a_{\max}^2 p^{\Delta \xi_1 + 2} \|(\mathbf{w}_0)_r\|_2^{2p-1} \quad (170)$$

$$\leq \frac{1}{m} \sqrt{\frac{\nu_{\max}}{n_o n_b}} \binom{d + |\xi|}{d} (d+1) a_{\max}^2 p^{|\xi| + 2} \|(\mathbf{w}_0)_r\|_2^{2p-1}, \quad (171)$$

$$\frac{\partial}{\partial (\mathbf{w}_\ell)_r} (\mathbf{Q}_{0,L+1})_{i,j} = \mathbf{0}, \quad \forall \ell \neq 0, \quad (172)$$

$$\frac{\partial}{\partial (\mathbf{v}_\ell)_r} \mathbf{Q}_{i,j} = \mathbf{0}, \quad \forall \ell. \quad (173)$$

For each $\ell \in [L]$, $\alpha_1, \alpha_2 \in I_{\xi_{\ell-1}}$, $\beta_1, \beta_2 \in I_{\Delta \xi_\ell}$, and $i_1, i_2 \in [n_o]$,

$$(\mathbf{Q}_{\ell,\ell})_{((\alpha_1, \beta_1), i_1), ((\alpha_2, \beta_2), i_2)} \quad (174)$$

$$= \frac{\nu_\ell}{mn_o} \sum_{r=1}^m \left(\left(\frac{\partial^{\beta_1}}{\partial \mathbf{x}^{\beta_1}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\frac{\partial^{\beta_2}}{\partial \mathbf{x}^{\beta_2}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \mathbf{1}_{\{\alpha_1 = \alpha_2\}} \right. \quad (175)$$

$$\left. + \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \mathbf{1}_{\{\alpha_1 + \beta_1 = \alpha_2 + \beta_2\}} \right), \quad (176)$$

$$\left\| \frac{\partial}{\partial (\mathbf{w}_{\ell-1})_r} (\mathbf{Q}_{\ell,\ell})_{((\alpha_1, \beta_1), i_1), ((\alpha_2, \beta_2), i_2)} \right\|_2 \quad (177)$$

$$= \mathbf{1}_{\{\alpha_1 = \alpha_2\}} \left\| \frac{\nu_\ell}{mn_o} \frac{\partial}{\partial (\mathbf{w}_{\ell-1})_r} \left\{ \left(\frac{\partial^{\beta_1}}{\partial \mathbf{x}^{\beta_1}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\frac{\partial^{\beta_2}}{\partial \mathbf{x}^{\beta_2}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \right\} \right\|_2 \quad (178)$$

$$\leq \mathbf{1}_{\{\alpha_1 = \alpha_2\}} \frac{\nu_\ell}{mn_o} p^{2\Delta \xi_\ell + 1} \|(\mathbf{w}_{\ell-1})_r\|_2^{2p-1} \quad (179)$$

$$\leq \mathbf{1}_{\{\alpha_1 = \alpha_2\}} \frac{\nu_{\max}}{mn_o} p^{2|\xi| + 1} \|(\mathbf{w}_{\ell-1})_r\|_2^{2p-1}, \quad (180)$$

and

$$\left\| \frac{\partial}{\partial (\mathbf{w}_\ell)_r} (\mathbf{Q}_{\ell,\ell})_{((\alpha_1, \beta_1), i_1), ((\alpha_2, \beta_2), i_2)} \right\|_2 \quad (181)$$

$$= \mathbf{1}_{\{\alpha_1 + \beta_1 = \alpha_2 + \beta_2\}} \left\| \frac{\nu_\ell}{mn_o} \frac{\partial}{\partial (\mathbf{w}_\ell)_r} \left\{ \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right\} \right\|_2 \quad (182)$$

$$\leq \mathbf{1}_{\{\alpha_1 + \beta_1 = \alpha_2 + \beta_2\}} \frac{\nu_\ell}{mn_o} p \|(\mathbf{w}_\ell)_r\|_2^{2p-1} \quad (183)$$

$$\leq \mathbf{1}_{\{\alpha_1 + \beta_1 = \alpha_2 + \beta_2\}} \frac{\nu_{\max}}{mn_o} p \|(\mathbf{w}_\ell)_r\|_2^{2p-1}, \quad (184)$$

$$\frac{\partial}{\partial (\mathbf{w}_{\ell'})_r} (\mathbf{Q}_{\ell,\ell})_{i_1, i_2} = \mathbf{0}, \quad \forall \ell' \neq \ell - 1, \ell, \quad (185)$$

$$\frac{\partial}{\partial (\mathbf{v}_{\ell'})_r} (\mathbf{Q}_{\ell,\ell})_{i_1, i_2} = \mathbf{0}, \quad \forall \ell'. \quad (186)$$

Similarly, for each $\ell \in \{1, \dots, L-1\}$, $\alpha_1 \in I_{\xi_{\ell-1}}$, $\alpha_2 \in I_{\xi_\ell}$, $\beta_1 \in I_{\Delta\xi_\ell}$, $\beta_2 \in I_{\Delta\xi_{\ell+1}}$ and $i_1, i_2 \in [n_o]$,

$$(\mathbf{Q}_{\ell, \ell+1})_{((\alpha_1, \beta_1), i_1), ((\alpha_2, \beta_2), i_2)} \quad (187)$$

$$= \frac{\sqrt{\nu_\ell \nu_{\ell+1}}}{mn_o} \sum_{r=1}^m \left(- \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\frac{\partial^{\beta_2}}{\partial \mathbf{x}^{\beta_2}} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \mathbf{1}_{\{\alpha_2 = \alpha_1 + \beta_1\}}, \quad (188)$$

$$(189)$$

$$\left\| \frac{\partial}{\partial (\mathbf{w}_\ell)_r} (\mathbf{Q}_{\ell, \ell+1})_{((\alpha_1, \beta_1), i_1), ((\alpha_2, \beta_2), i_2)} \right\|_2 \quad (190)$$

$$\leq \mathbf{1}_{\{\alpha_2 = \alpha_1 + \beta_1\}} \frac{\sqrt{\nu_\ell \nu_{\ell+1}}}{mn_o} p^{\Delta\xi_{\ell+1}+1} \|(\mathbf{w}_\ell)_r\|_2^{2p-1} \quad (191)$$

$$\leq \mathbf{1}_{\{\alpha_2 = \alpha_1 + \beta_1\}} \frac{\nu_{\max}}{mn_o} p^{|\xi|+1} \|(\mathbf{w}_\ell)_r\|_2^{2p-1}, \quad (192)$$

$$\frac{\partial}{\partial (\mathbf{w}_{\ell'})_r} (\mathbf{Q}_{\ell, \ell+1})_{((\alpha_1, \beta_1), i_1), ((\alpha_2, \beta_2), i_2)} = \mathbf{0}, \quad \forall \ell' \neq \ell, \quad (193)$$

$$(194)$$

$$\frac{\partial}{\partial (\mathbf{v}_{\ell'})_r} (\mathbf{Q}_{\ell, \ell+1})_{((\alpha_1, \beta_1), i_1), ((\alpha_2, \beta_2), i_2)} = \mathbf{0}, \quad \forall \ell'. \quad (195)$$

For each $\ell \in [L]$, $\alpha = 1$, $\beta \in I_{\Delta\xi_1}$, $i \in [n_o]$ and $j \in [n_b]$,

$$(\mathbf{Q}_{\ell, L+1})_{((\alpha, \beta), i), j} \quad (196)$$

$$= \frac{1}{m} \sqrt{\frac{\nu_1 \nu}{n_o n_b}} \sum_{r=1}^m \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_0)_r^\top \mathbf{y}_i \right)_+^p \right) \left(\sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \right) \mathbf{1}_{\{\ell=1\}}, \quad (197)$$

$$\left\| \frac{\partial}{\partial (\mathbf{w}_0)_r} (\mathbf{Q}_{1, L+1})_{((\alpha, \beta), i), j} \right\|_2 \quad (198)$$

$$= \left\| \frac{\sqrt{\nu_1 \nu}}{m \sqrt{n_o n_b}} \frac{\partial}{\partial (\mathbf{w}_0)_r} \left\{ \sum_{r=1}^m \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_0)_r^\top \mathbf{y}_i \right)_+^p \right) \left(\sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \right) \right\} \right\|_2 \quad (199)$$

$$\leq \frac{\sqrt{\nu_1 \nu}}{m \sqrt{n_o n_b}} a_{\max} (d+1) p^{\Delta\xi_1+2} \|(\mathbf{w}_0)_r\|_2^{2p-1} \quad (200)$$

$$\leq \frac{\nu_{\max}}{m \sqrt{n_o n_b}} a_{\max} (d+1) p^{|\xi|+2} \|(\mathbf{w}_0)_r\|_2^{2p-1}, \quad (201)$$

$$\frac{\partial}{\partial (\mathbf{w}_\ell)_r} (\mathbf{Q}_{1, L+1})_{((\alpha, \beta), i), j} = \mathbf{0}, \quad \forall \ell \neq 0, \quad (202)$$

$$\frac{\partial}{\partial (\mathbf{v}_\ell)_r} (\mathbf{Q}_{1, L+1})_{((\alpha, \beta), i), j} = \mathbf{0}, \quad \forall \ell. \quad (203)$$

For each $j_1, j_2 \in [n_b]$, the direct calculations lead to

$$(\mathbf{Q}_{L+1, L+1})_{j_1, j_2} = \frac{\nu}{mn_b} \sum_{r=1}^m \left(\sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_{j_1} \right)_+^p \right) \left(\sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_{j_2} \right)_+^p \right), \quad (204)$$

$$\left\| \frac{\partial}{\partial (\mathbf{w}_0)_r} (\mathbf{Q}_{L+1, L+1})_{j_1, j_2} \right\|_2 \leq \frac{\nu}{mn_b} a_{\max}^2 (d+1)^2 p^3 \|(\mathbf{w}_0)_r\|_2^{2p-1}, \quad (205)$$

$$\frac{\partial}{\partial (\mathbf{w}_\ell)_r} (\mathbf{Q}_{L+1, L+1})_{j_1, j_2} = \mathbf{0}, \quad \forall \ell \neq 0, \quad (206)$$

$$\frac{\partial}{\partial (\mathbf{v}_\ell)_r} (\mathbf{Q}_{L+1, L+1})_{j_1, j_2} = \mathbf{0}, \quad \forall \ell. \quad (207)$$

Hoeffding's inequalities for $\mathbf{Q}_{\ell_1, \ell_2}$. For each $i \in [n_o]$ and $j \in [n_b]$, we have

$$\left| \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_i \right)_+^p \right| \leq p^{|\beta|} \|\mathbf{w}_\ell\|_2^p, \quad (208)$$

$$\left| \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_i \right)_+^p \right| \leq p^{|\beta|} \|\mathbf{w}_{\ell-1}\|_2^p, \quad (209)$$

$$\left| \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_i \right)_+^p \right| \leq \|\mathbf{w}_\ell\|_2^p, \quad (210)$$

$$\left| \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \right| \leq p^{|\alpha|} \|\mathbf{w}_0\|_2^p. \quad (211)$$

Then, for each $i_1, i_2 \in [n_o]$,

$$\left| \sum_{\ell=0}^L \sum_{\alpha \in I_{\xi_\ell}} \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \right| \quad (212)$$

$$\leq \sum_{\ell=0}^L \sum_{\alpha \in I_{\xi_\ell}} \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} a_{\max} p^{|\beta|} \|\mathbf{w}_\ell\|_2^p \right) \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} a_{\max} p^{|\beta|} \|\mathbf{w}_\ell\|_2^p \right) \quad (213)$$

$$\leq \sum_{\ell=0}^L \sum_{\alpha \in I_{\xi_\ell}} |J_{\Delta \xi_{\ell+1}}|^2 a_{\max}^2 p^{2|\Delta \xi_{\ell+1}|} R^{2p} \quad (214)$$

$$\leq \sum_{\ell=0}^L |I_{\xi_\ell}| \binom{d+|\xi|}{d}^2 a_{\max}^2 p^{2|\xi|} R^{2p} \quad (215)$$

$$= \binom{d+k}{d} \binom{d+|\xi|}{d}^2 a_{\max}^2 p^{2|\xi|} R^{2p}. \quad (216)$$

By Hoeffding's inequality with

$$(\mathbf{Q}_{0,0})_{i_1, i_2} \quad (217)$$

$$= \sum_{r=1}^m \sum_{\ell=0}^L \sum_{\alpha \in I_{\xi_\ell}} \left(\frac{\partial}{\partial (\mathbf{v}_\ell)_r, \alpha} \mathbf{s}_{i_1} \right) \left(\frac{\partial}{\partial (\mathbf{v}_\ell)_r, \alpha} \mathbf{s}_{i_2} \right) \quad (218)$$

$$= \frac{1}{mn_o} \sum_{\ell=0}^L \sum_{r=1}^m \sum_{\alpha \in I_{\xi_\ell}} \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right), \quad (219)$$

we have

$$\mathbb{P} \left[\left| (\mathbf{Q}_{0,0})_{i_1, i_2} - \mathbb{E} \left[(\mathbf{Q}_{0,0})_{i_1, i_2} \right] \right| > \varepsilon \right] < 2 \exp \left(- \frac{mn_o^2 \varepsilon^2}{2 \binom{d+k}{d}^2 \binom{d+|\xi|}{d}^4 a_{\max}^4 p^{4|\xi|} R^{4p}} \right), \quad (220)$$

for any $\varepsilon > 0$.

Similarly, for each $\ell \in [L]$, $\alpha \in I_{\xi_{\ell-1}}$, $\beta \in \Delta \xi_\ell$, and $i_1, i_2 \in [n_o]$,

$$\left| \left(\sum_{\beta' \in J_{\Delta \xi_\ell}} \hat{a}_{\ell-1, \alpha, \beta'} \frac{\partial^{\beta'}}{\partial \mathbf{x}^{\beta'}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \right| \quad (221)$$

$$- \left(\sum_{\beta' \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta'} \frac{\partial^{\beta'}}{\partial \mathbf{x}^{\beta'}} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \quad (222)$$

$$\leq \binom{d + \Delta_{\xi_\ell}}{\Delta_{\xi_\ell}} a_{\max} p^{\Delta_{\xi_\ell}} \|(\mathbf{w}_{\ell-1})_r\|_2^p \cdot p^{|\beta|} \|(\mathbf{w}_{\ell-1})_r\|_2^p + \binom{d + \Delta_{\xi_{\ell+1}}}{\Delta_{\xi_{\ell+1}}} a_{\max} p^{\Delta_{\xi_{\ell+1}}} \|(\mathbf{w}_\ell)_r\|_2^p \|(\mathbf{w}_\ell)_r\|_2^p \quad (223)$$

$$\leq \binom{d + |\xi|}{d} (p^{|\xi|} + 1) a_{\max} p^{|\xi|} R^{2p}. \quad (224)$$

By Hoeffding's inequality with

$$(\mathbf{Q}_{0,\ell})_{i_1,((\alpha,\beta),i_2)} \quad (225)$$

$$= \frac{\sqrt{\nu_\ell}}{mn_o} \sum_{r=1}^m \left(\left(\sum_{\beta' \in J_{\Delta_{\xi_\ell}}} \hat{a}_{\ell-1,\alpha,\beta'} \frac{\partial^{\beta'}}{\partial \mathbf{x}^{\beta'}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \right. \quad (226)$$

$$\left. - \left(\sum_{\beta' \in J_{\Delta_{\xi_{\ell+1}}}} \hat{a}_{\ell,\alpha,\beta'} \frac{\partial^{\beta'}}{\partial \mathbf{x}^{\beta'}} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right), \quad (227)$$

we have

$$\mathbb{P} \left[\left| (\mathbf{Q}_{0,\ell})_{i_1,((\alpha,\beta),i_2)} - \mathbb{E} \left[(\mathbf{Q}_{0,\ell})_{i_1,((\alpha,\beta),i_2)} \right] \right| > \varepsilon \right] < 2 \exp \left(- \frac{mn_o^2 \varepsilon^2}{2\nu_\ell \binom{d+|\xi|}{d}^2 (p^{|\xi|} + 1)^2 a_{\max}^2 p^{2|\xi|} R^{4p}} \right), \quad (228)$$

for any $\varepsilon > 0$.

For each $i \in [n_o]$ and $j \in [n_b]$,

$$\left| \left(\sum_{\beta \in J_{\Delta_{\xi_1}}} \hat{a}_{0,1,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_0)_r^\top \mathbf{y}_i \right)_+^p \right) \left(\sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \right) \right| \quad (229)$$

$$\leq \binom{d + \Delta_{\xi_1}}{\Delta_{\xi_1}} a_{\max} p^{\Delta_{\xi_1}} \|(\mathbf{w}_0)_r\|_2^p \binom{d+1}{1} a_{\max} p^1 \|(\mathbf{w}_0)_r\|_2^p \quad (230)$$

$$\leq (d+1) \binom{d + |\xi|}{d} a_{\max}^2 p^{|\xi|+1} R^{2p}. \quad (231)$$

By Hoeffding's inequality with

$$(\mathbf{Q}_{0,L+1})_{i,j} = \frac{1}{m} \sqrt{\frac{\nu}{n_o n_b}} \sum_{r=1}^m \left(\sum_{\beta \in J_{\Delta_{\xi_1}}} \hat{a}_{0,1,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_0)_r^\top \mathbf{y}_i \right)_+^p \right) \left(\sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \right), \quad (232)$$

we have

$$\mathbb{P} \left[\left| (\mathbf{Q}_{0,L+1})_{i,j} - \mathbb{E} \left[(\mathbf{Q}_{0,L+1})_{i,j} \right] \right| > \varepsilon \right] < 2 \exp \left(- \frac{mn_o n_b \varepsilon^2}{2\nu (d+1)^2 \binom{d+|\xi|}{d}^2 a_{\max}^4 p^{2|\xi|+2} R^{4p}} \right), \quad (233)$$

for any $\varepsilon > 0$.

For each $\ell \in [L]$, $\alpha_1, \alpha_2 \in I_{\xi_{\ell-1}}$, $\beta_1, \beta_2 \in J_{\Delta_{\xi_\ell}}$, and $i_1, i_2 \in [n_o]$,

$$\left| \left(\frac{\partial^{\beta_1}}{\partial \mathbf{x}^{\beta_1}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\frac{\partial^{\beta_2}}{\partial \mathbf{x}^{\beta_2}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \mathbf{1}_{\{\alpha_1=\alpha_2\}} \right. \quad (234)$$

$$\left. + \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \mathbf{1}_{\{\alpha_1+\beta_1=\alpha_2+\beta_2\}} \right| \quad (235)$$

$$\leq p^{|\beta_1|+|\beta_2|} R^{2p} \mathbf{1}_{\{\alpha_1=\alpha_2\}} + R^{2p} \mathbf{1}_{\{\alpha_1+\beta_1=\alpha_2+\beta_2\}} \quad (236)$$

$$\leq (p^{2|\xi|} + 1) R^{2p}. \quad (237)$$

By Hoeffding's inequality with

$$(\mathbf{Q}_{\ell,\ell})_{((\alpha_1,\beta_1),i_1),((\alpha_2,\beta_2),i_2)} \quad (238)$$

$$= \frac{\nu_\ell}{mn_o} \sum_{r=1}^m \left(\left(\frac{\partial^{\beta_1}}{\partial \mathbf{x}^{\beta_1}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\frac{\partial^{\beta_2}}{\partial \mathbf{x}^{\beta_2}} \left((\mathbf{w}_{\ell-1})_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \mathbf{1}_{\{\alpha_1=\alpha_2\}} \right. \quad (239)$$

$$\left. + \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \mathbf{1}_{\{\alpha_1+\beta_1=\alpha_2+\beta_2\}} \right), \quad (240)$$

we have

$$\mathbb{P} \left[\left| (\mathbf{Q}_{\ell,\ell})_{((\alpha_1,\beta_1),i_1),((\alpha_2,\beta_2),i_2)} - \mathbb{E} \left[(\mathbf{Q}_{\ell,\ell})_{((\alpha_1,\beta_1),i_1),((\alpha_2,\beta_2),i_2)} \right] \right| > \varepsilon \right] < 2 \exp \left(- \frac{mn_o^2 \varepsilon^2}{2\nu_\ell^2 (p^{2|\xi|} + 1)^2 R^{4p}} \right), \quad (241)$$

for any $\varepsilon > 0$.

For each $\ell \in [L]$, $\alpha_1 \in I_{\xi_{\ell-1}}$, $\alpha_2 \in I_{\xi_\ell}$, $\beta_1 \in I_{\Delta\xi_\ell}$, $\beta_2 \in I_{\Delta\xi_{\ell+1}}$, and $i_1, i_2 \in [n_o]$,

$$\left| \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \left(\frac{\partial^{\beta_2}}{\partial \mathbf{x}^{\beta_2}} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \mathbf{1}_{\{\alpha_2=\alpha_1+\beta_2\}} \right| \leq p^{|\xi|} R^{2p}. \quad (242)$$

By Hoeffding's inequality with

$$(\mathbf{Q}_{\ell,\ell+1})_{((\alpha_1,\beta_1),i_1),((\alpha_2,\beta_2),i_2)} \quad (243)$$

$$= \frac{\sqrt{\nu_\ell \nu_{\ell+1}}}{mn_o} \sum_{r=1}^m \left(- \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\frac{\partial^{\beta_2}}{\partial \mathbf{x}^{\beta_2}} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right) \mathbf{1}_{\{\alpha_2=\alpha_1+\beta_2\}}, \quad (244)$$

we have

$$\mathbb{P} \left[\left| (\mathbf{Q}_{\ell,\ell+1})_{((\alpha_1,\beta_1),i_1),((\alpha_2,\beta_2),i_2)} - \mathbb{E} \left[(\mathbf{Q}_{\ell,\ell+1})_{((\alpha_1,\beta_1),i_1),((\alpha_2,\beta_2),i_2)} \right] \right| > \varepsilon \right] < 2 \exp \left(- \frac{mn_o^2 \varepsilon^2}{2\nu_\ell \nu_{\ell+1} p^{2|\xi|} R^{4p}} \right), \quad (245)$$

for any $\varepsilon > 0$.

For each $\alpha \in I_{\xi_{\ell-1}}$, $\beta \in J_{\Delta\xi_\ell}$, $i \in [n_o]$, and $j \in [n_b]$,

$$\left| \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_0)_r^\top \mathbf{y}_i \right)_+^p \left(\sum_{\alpha' \in J_1} \tilde{a}_{\alpha'} \frac{\partial^{\alpha'}}{\partial \mathbf{x}^{\alpha'}} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \right) \right| \quad (246)$$

$$\leq p^{|\beta|} R^p \sum_{\alpha' \in J_1} a_{\max} p^{|\alpha'|} R^p \quad (247)$$

$$\leq (d+1) a_{\max} p^{|\xi|+1} R^{2p}. \quad (248)$$

By Hoeffding's inequality with

$$(\mathbf{Q}_{1,L+1})_{((\alpha,\beta),i),j} \quad (249)$$

$$= \frac{1}{m} \sqrt{\frac{\nu_1 \nu}{n_o n_b}} \sum_{r=1}^m \left(\frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_0)_r^\top \mathbf{y}_i \right)_+^p \right) \left(\sum_{\alpha' \in J_1} \tilde{a}_{\alpha'} \frac{\partial^{\alpha'}}{\partial \mathbf{x}^{\alpha'}} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \right), \quad (250)$$

we have

$$\mathbb{P} \left[\left| (\mathbf{Q}_{1,L+1})_{((\alpha,\beta),i),j} - \mathbb{E} \left[(\mathbf{Q}_{1,L+1})_{((\alpha,\beta),i),j} \right] \right| > \varepsilon \right] < 2 \exp \left(- \frac{mn_o n_b \varepsilon^2}{2\nu_1 \nu (d+1)^2 a_{\max}^2 p^{2|\xi|+2} R^{4p}} \right), \quad (251)$$

for any $\varepsilon > 0$.

For $j_1, j_2 \in [n_b]$,

$$\left(\sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_{j_1} \right)_+^p \right) \left(\sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_{j_2} \right)_+^p \right) \leq (d+1)^2 a_{\max}^2 p^2 R^{2p}. \quad (252)$$

By Hoeffding's inequality with

$$(\mathbf{Q}_{L+1,L+1})_{j_1,j_2} = \frac{\nu}{mn_b} \sum_{r=1}^m \left(\sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_{j_1} \right)_+^p \right) \left(\sum_{\alpha \in J_1} \tilde{a}_\alpha \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_{j_2} \right)_+^p \right), \quad (253)$$

we have

$$\mathbb{P} \left[\left| (\mathbf{Q}_{L+1, L+1})_{j_1, j_2} - \mathbb{E} \left[(\mathbf{Q}_{L+1, L+1})_{j_1, j_2} \right] \right| > \varepsilon \right] < 2 \exp \left(- \frac{mn_b^2 \varepsilon^2}{2\nu^2 (d+1)^4 a_{\max}^4 p^4 R^{4p}} \right), \quad (254)$$

for any $\varepsilon > 0$.

C Proofs

This section proves the main theorem with some lemmas and propositions. Similar to the [25], we first prove that the Gram matrix $\widehat{\mathbf{G}}_v^\infty$ is independent of m and strictly positive definite, in Proposition C.3. Second, as $\widehat{\mathbf{G}}_v(\mathbf{w}(0), \mathbf{v}(0))$ is the sample mean, Proposition C.4 shows that $\widehat{\mathbf{G}}_v(\mathbf{w}(0), \mathbf{v}(0))$ is close to its expected value $\widehat{\mathbf{G}}_v^\infty$ with a high probability, if m is large enough. This implies that $\widehat{\mathbf{G}}_v(\mathbf{w}(0), \mathbf{v}(0))$ is strictly positive definite with high probability. Then, we show that the initial loss is bounded with a high probability if m is large. This is because $\mathbf{v}(0)$ has zero mean, which hinders the output of m variables to diverge as m increases. Finally, we prove that the Gram matrix $\widehat{\mathbf{G}}_v(\mathbf{w}(t), \mathbf{v}(t))$ remains within a small neighborhood of the initial point, where it maintains its strict positive definiteness. The gradient flow converges within the neighborhood, as the smallest eigenvalue of $\widehat{\mathbf{G}}_v$ is distinct from zero, resulting in a rapid reduction of the loss.

Lemma C.1. *Let $d, n \in \mathbb{N}$ and $\mathbf{y}_1, \dots, \mathbf{y}_n \in \mathbb{R}^d$ be vectors such that $\mathbf{y}_i \parallel \mathbf{y}_j$ if and only if $i = j$. Then, for each $i \in [n]$, there exists $\mathbf{w}_i$ such that $\mathbf{w}_i^\top \mathbf{y}_j = 0$ if and only if $i = j$.*

Proof. Let $\mathbf{y}_i^\perp = \{\mathbf{w} \in \mathbb{R}^d : \mathbf{w}^\top \mathbf{y}_i = 0\}$. Since $\mathbf{y}_i \parallel \mathbf{y}_j$ for $i \neq j$, $\mathbf{y}_i \cap \mathbf{y}_j$ is nowhere dense in $\mathbf{y}_i^\perp$. Hence, finite union of $\cup_{j \neq i} \mathbf{y}_j^\perp$ is nowhere dense in $\mathbf{y}_i^\perp$, and in particular, there exists $\mathbf{w}_i \in \mathbf{y}_i^\perp$ such that $\mathbf{w}_i \notin \cup_{j \neq i} \mathbf{y}_j^\perp$. $\square$

Lemma C.2. *Let $n, d \in \mathbb{N}$, and $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a homogeneous polynomial of degree n . For any $i \in \mathbb{N}$ and nonzero $\mathbf{y} = (y_1, \dots, y_d) \in \mathbb{R}^d$, if a function $\psi(\mathbf{w}) (\mathbf{w}^\top \mathbf{y})_+^i$ of $\mathbf{w} = (w_1, \dots, w_d) \in \mathbb{R}^d$ is $(n+i)$ -times continuously differentiable at some $\mathbf{w}^* \in \mathbb{R}^d$ with $(\mathbf{w}^*)^\top \mathbf{y} = 0$, then $\psi \equiv 0$.*

Proof. Without loss of the generality, we may assume the $y_1 \neq 0$. Write $\psi(\mathbf{w}) = \sum_{\alpha \in I_n} a_\alpha \mathbf{w}^\alpha$ for coefficients $a_\alpha \in \mathbb{R}$ and define $\varphi(\mathbf{w}) = \psi(\mathbf{w}) (\mathbf{w}^\top \mathbf{y})_+^i$. For any $\alpha \in I_n$, the assumption on differentiability implies that $\frac{\partial^i}{\partial w_1^i} \frac{\partial^n}{\partial \mathbf{w}^\alpha} \varphi$ is continuous. Indeed, we have $\varphi(\mathbf{w}) = 0$ for $\mathbf{w}^\top \mathbf{y} \leq 0$, and

$$\frac{\partial^i}{\partial w_1^i} \frac{\partial^n}{\partial \mathbf{w}^\alpha} \varphi(\mathbf{w}) = \alpha! a_\alpha y_1^i, \quad (255)$$

for $\mathbf{w}^\top \mathbf{y} > 0$. The continuity induces $a_\alpha = 0$ and thus $\psi \equiv 0$ for all $\mathbf{w} \in \mathbb{R}^d$. $\square$

Proposition C.3. $\widehat{\mathbf{G}}_v^\infty = \mathbb{E}_{\mathbf{w}, \mathbf{v}} [\widehat{\mathbf{G}}_v(\mathbf{w}, \mathbf{v})]$ is strictly positive definite and independent of m .

Proof. For $i_1, i_2 \in [n_o]$,

$$(\mathbf{Q}_{0,0})_{i_1, i_2} \quad (256)$$

$$= \frac{1}{m} \sum_{r=1}^m \frac{1}{n_o} \sum_{\ell=0}^L \sum_{\alpha \in I_{\xi_\ell}} \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right), \quad (257)$$

and m random variables

$$\frac{1}{n_o} \sum_{\ell=0}^L \sum_{\alpha \in I_{\xi_\ell}} \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_1} \right)_+^p \right) \left(\sum_{\beta \in J_{\Delta \xi_{\ell+1}}} \hat{a}_{\ell, \alpha, \beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_{i_2} \right)_+^p \right), \quad (258)$$

for $r \in [m]$, are independent and identically distributed. Therefore, the expected value of each component of $\mathbf{Q}_{0,0}$ is independent of m , as the expected value of the sample mean of independent and identically distributed random variables is independent of the number of samples. Similarly, each component of $\hat{\mathbf{G}}_v(\mathbf{w}, \mathbf{v})$ is the sample mean of m independent and identically distributed random variables, and its expected value $\hat{\mathbf{G}}_v^\infty$ is independent of m . To show that $\hat{\mathbf{G}}_v^\infty$ is strictly positive definite, assume that there exists $\mathbf{z}_0 = [(\mathbf{z}_0)_1 \cdots (\mathbf{z}_0)_{n_o}] \in \mathbb{R}^{n_o}$, $\tilde{\mathbf{z}} = [\tilde{z}_1 \cdots \tilde{z}_{n_b}]^\top \in \mathbb{R}^{n_b}$, and

$$\mathbf{z}_\ell = [(\mathbf{z}_\ell)_{(\alpha,\beta),i}]_{(\alpha \in I_{\xi_\ell}, \beta \in I_{\Delta\xi_\ell}, i \in [n_o])} \in \mathbb{R}^{|I_{\xi_{\ell-1}}| \cdot |I_{\Delta\xi_\ell}| n_o} \quad (259)$$

for $\ell \in [L]$, such that $\mathbf{z} = [\mathbf{z}_0^\top \mathbf{z}_1^\top \cdots \mathbf{z}_L^\top \tilde{\mathbf{z}}^\top]^\top \in \mathbb{R}^{n_o + \sum_{\ell=1}^L (|I_{\xi_{\ell-1}}| \cdot |I_{\Delta\xi_\ell}| n_o) + n_b}$ satisfies $\hat{\mathbf{G}}_v^\infty \mathbf{z} = \mathbf{0}$. Since $\hat{\mathbf{G}}_v^\infty \mathbf{z} = \mathbb{E}_{\mathbf{w}, \mathbf{v}} [\hat{\mathbf{D}}_v^\top \hat{\mathbf{D}}_v] \mathbf{z} = \mathbf{0}$ if and only if $\mathbb{E}_{\mathbf{w}, \mathbf{v}} [\|\hat{\mathbf{D}}_v \mathbf{z}\|_2^2] = \mathbf{0}$, we show that $\mathbb{E}_{\mathbf{w}, \mathbf{v}} [\|\hat{\mathbf{D}}_v \mathbf{z}\|_2^2] = \mathbf{0}$ implies $\mathbf{z} = \mathbf{0}$ and thereby 0 is not an eigenvalue of $\hat{\mathbf{G}}_v^\infty$.

Now assume that there exists $\mathbf{z} \in \mathbb{R}^d$ such that $\hat{\mathbf{D}}_v \mathbf{z} = \mathbf{0}$ for almost all $\mathbf{w}$ and $\mathbf{v}$. Since $\hat{\mathbf{D}}_v$ is continuous with respect to $\mathbf{w}$ and $\mathbf{v}$, we have $\hat{\mathbf{D}}_v \mathbf{z} = \mathbf{0}$ for all $\mathbf{w}$ and $\mathbf{v}$. Hence, the function $\hat{\mathbf{D}}_v \mathbf{z}$ of $\mathbf{w}$ and $\mathbf{v}$, which is identically zero, is smooth with respect to $\mathbf{w}$ and $\mathbf{v}$.

In this proof, we denote an index for a component of $\hat{\mathbf{D}}_v \mathbf{z}$ by $\ell \in [L]$, $r \in [m]$ and $\alpha \in I_{\xi_\ell}$, because each row of $\hat{\mathbf{D}}_v$ corresponds to the partial derivatives with respect to $(\mathbf{v}_\ell)_{r,\alpha}$ for $\ell \in [L]$, $r \in [m]$ and $\alpha \in I_{\xi_\ell}$. For instance, $(\hat{\mathbf{D}}_v \mathbf{z})_{0,1,0}$ is the first component of $(\hat{\mathbf{D}}_v \mathbf{z})$. First, for $\ell = L$ and any $r \in [m]$ and $\alpha \in I_{\xi_L}$, we have

$$(\hat{\mathbf{D}}_v \mathbf{z})_{L,r,\alpha} \quad (260)$$

$$= \sum_{i=1}^{n_o} \left(\frac{\partial}{\partial (\mathbf{v}_L)_{r,\alpha}} s_i \right) (\mathbf{z}_0)_i + \sum_{i=1}^{n_o} \sum_{\alpha' \in I_{\xi_{L-1}}} \sum_{\beta \in I_{\Delta\xi_L}} \left(\frac{\partial}{\partial (\mathbf{v}_L)_{r,\alpha}} (\boldsymbol{\mu}_L)_{(\alpha',\beta),i} \right) (\mathbf{z}_L)_{(\alpha',\beta),i} \quad (261)$$

$$= \sqrt{\frac{1}{mn_o}} \sum_{i=1}^{n_o} \left(\sum_{\beta \in J_{\Delta\xi_{L+1}}} \hat{a}_{L,\alpha,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_L)_r^\top \mathbf{y}_i \right)_+^p \right) (\mathbf{z}_0)_i \quad (262)$$

$$- \sqrt{\frac{\nu_L}{mn_o}} \sum_{i=1}^{n_o} \sum_{\alpha' \in I_{\xi_{L-1}}} \sum_{\beta \in I_{\Delta\xi_L}} \left((\mathbf{w}_L)_r^\top \mathbf{y}_i \right)_+^p (\mathbf{z}_L)_{(\alpha',\beta),i} \mathbf{1}_{\{\alpha=\alpha'+\beta\}}. \quad (263)$$

Fix $i \in [n_o]$, and by Lemma C.1, there exists $(\mathbf{w}_L^*)_r \in \mathbb{R}^d$ such that $(\mathbf{w}_L^*)_r^\top \mathbf{y}_i = 0$, $(\mathbf{w}_L^*)_r^\top \mathbf{y}_{i'} \neq 0$ for $i \neq i' \in [n_o]$ and $(\mathbf{w}_L^*)_r^\top \tilde{\mathbf{y}}_j \neq 0$ for $j \in [n_b]$. As a function of $(\mathbf{w}_L)_r$, $\hat{\mathbf{D}}_v \mathbf{z} = \mathbf{0}$ implies that

$$\left(\sum_{\beta \in J_{\Delta\xi_{L+1}}} \hat{a}_{L,\alpha,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_L)_r^\top \mathbf{y}_i \right)_+^p \right) (\mathbf{z}_0)_i \quad (264)$$

$$- \sqrt{\nu_L} \sum_{\alpha' \in I_{\xi_{L-1}}} \sum_{\beta \in I_{\Delta\xi_L}} \left((\mathbf{w}_L)_r^\top \mathbf{y}_i \right)_+^p (\mathbf{z}_L)_{(\alpha',\beta),i} \mathbf{1}_{\{\alpha=\alpha'+\beta\}} \quad (265)$$

$$= - \sum_{i' \neq i} \left(\sum_{\beta \in J_{\Delta\xi_{L+1}}} \hat{a}_{L,\alpha,\beta} \frac{\partial^\beta}{\partial \mathbf{x}^\beta} \left((\mathbf{w}_L)_r^\top \mathbf{y}_i \right)_+^p \right) (\mathbf{z}_0)_i \quad (266)$$

$$+ \sqrt{\nu_L} \sum_{i' \neq i} \sum_{\alpha' \in I_{\xi_{L-1}}} \sum_{\beta \in I_{\Delta\xi_L}} \left((\mathbf{w}_L)_r^\top \mathbf{y}_i \right)_+^p (\mathbf{z}_L)_{(\alpha',\beta),i} \mathbf{1}_{\{\alpha=\alpha'+\beta\}}. \quad (267)$$

is smooth at $(\mathbf{w}_L)_r = (\mathbf{w}_L^*)_r$. Note that the coefficient of $\left((\mathbf{w}_L)_r^\top \mathbf{y}_i\right)_+^{p-\Delta\xi_{L+1}}$ in (266) is given by

$$\sum_{\beta \in I_{\Delta\xi_{L+1}}} \hat{a}_{L,\alpha,\beta}(\mathbf{x}_i) (\mathbf{w}_L)_{r,1}^{\beta_1} \cdots (\mathbf{w}_L)_{r,d}^{\beta_d} (z_0)_i, \quad (268)$$

which is a homogeneous polynomial of $\mathbf{w}_L$. Together with the fact that at least one of $\hat{a}_{L,\alpha,\beta}(\mathbf{x}_i)$ is nonzero, Lemma C.2 implies $(z_0)_i = 0$. Similarly, for any $r \in [m]$,

$$(\hat{D}_v z)_{0,r,0} = \sum_{j=1}^{n_b} \left(\frac{\partial}{\partial (v_0)_{r,0}} h_j \right) \tilde{z}_j \quad (269)$$

$$= \sqrt{\frac{\nu}{mn_b}} \sum_{j=1}^{n_b} \sum_{\alpha \in J_1} \tilde{a}_\alpha(\tilde{\mathbf{x}}_j) \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^p \tilde{z}_j \quad (270)$$

$$= \sqrt{\frac{\nu}{mn_b}} \sum_{j=1}^{n_b} \sum_{\alpha \in J_1} \tilde{a}_\alpha(\tilde{\mathbf{x}}_j) \tilde{z}_j (\mathbf{w}_0)_{r,1}^{\alpha_1} \cdots (\mathbf{w}_0)_{r,d}^{\alpha_d} \left((\mathbf{w}_0)_r^\top \tilde{\mathbf{y}}_j \right)_+^{p-|\alpha|} \quad (271)$$

deduces $\tilde{z}_j = 0$ with $\mathbf{w}_0^*$, instead of $\mathbf{w}_L^*$, such that $(\mathbf{w}_0^*)^\top \tilde{\mathbf{y}}_j = 0$, $(\mathbf{w}_0^*)^\top \tilde{\mathbf{y}}_{j'} \neq 0$ for $j' \neq j$, and $(\mathbf{w}_0^*)^\top \mathbf{y}_i \neq 0$ for $i \in [n_o]$. Now suppose that we have $\tilde{z} = 0$ and $z_{\ell'} = 0$ for $\ell' = 0, 1, \dots, \ell - 1$ for some $\ell \geq 1$. Then, for any $r \in [m]$ and $\alpha \in I_{\xi_{\ell-1}}$,

$$(\hat{D}_v z)_{\ell-1,r,\alpha} \quad (272)$$

$$= \sum_{i=1}^{n_o} \left(\frac{\partial}{\partial (v_{\ell-1})_{r,\alpha}} s_i \right) (z_0)_i + \sum_{i=1}^{n_o} \sum_{\alpha' \in I_{\xi_{\ell-1}}} \sum_{\beta \in I_{\Delta\xi_\ell}} \left(\frac{\partial}{\partial (v_{\ell-1})_{r,\alpha}} (\boldsymbol{\mu}_\ell)_{(\alpha',\beta),i} \right) (z_\ell)_{(\alpha',\beta),i} \quad (273)$$

$$+ \sum_{i=1}^{n_o} \sum_{\alpha' \in I_{\xi_{\ell-2}}} \sum_{\beta \in I_{\Delta\xi_{\ell-1}}} \left(\frac{\partial}{\partial (v_{\ell-1})_{r,\alpha}} (\boldsymbol{\mu}_{\ell-1})_{(\alpha',\beta),i} \right) (z_{\ell-1})_{(\alpha',\beta),i} \quad (274)$$

$$= \sum_{i=1}^{n_o} \sum_{\alpha' \in I_{\xi_{\ell-1}}} \sum_{\beta \in I_{\Delta\xi_\ell}} \left(\frac{\partial}{\partial (v_{\ell-1})_{r,\alpha}} (\boldsymbol{\mu}_\ell)_{(\alpha',\beta),i} \right) (z_\ell)_{(\alpha',\beta),i} \quad (275)$$

$$= \sqrt{\frac{\nu_\ell}{n_o}} \sum_{i=1}^{n_o} \sum_{\beta \in I_{\Delta\xi_\ell}} \left((\mathbf{w}_\ell)_{r,1}^{\beta_1} \cdots (\mathbf{w}_\ell)_{r,d}^{\beta_d} \left((\mathbf{w}_\ell)_r^\top \mathbf{y}_i \right)_+^{p-\Delta\xi_\ell} \right) (z_\ell)_{(\alpha,\beta),i}. \quad (276)$$

Then the coefficient of $\left((\mathbf{w}_\ell)_r^\top \mathbf{y}_i\right)_+^{p-\Delta\xi_\ell}$ is a homogeneous polynomial

$$\sqrt{\frac{\nu_\ell}{n_o}} \sum_{\beta \in I_{\Delta\xi_\ell}} (z_\ell)_{(\alpha,\beta),i} (\mathbf{w}_\ell)_{r,d}^{\beta_d}. \quad (277)$$

Lemma C.2 shows $(z_\ell)_{(\alpha,\beta),i} = 0$ for all α, β . The induction concludes $z = 0$. $\square$

In the sequel, $\lambda_0 = \lambda_{\min}(\hat{\mathbf{G}}_v) > 0$ is the minimal eigenvalue of $\hat{\mathbf{G}}_v$. Furthermore, we assume the occurrence of the following event: there exists $R > 1$ such that $\|(\mathbf{w}_\ell)_r(0)\|_2 < R$ and $\|(\mathbf{v}_\ell)_r(0)\|_2 < R$ for all $\ell = 0, 1, \dots, L$ and $r \in [m]$. This assumption is employed solely for the purpose of simplifying propositions and their respective proofs. In the main theorem and its proof, we use the following arguments with regard to the conditional probability of the aforementioned event.

Proposition C.4. Let $\delta > 0$ and

$$C_1 = \binom{d+k}{k} \binom{d+|\xi|}{|\xi|}^2 \nu_{\max}^2 \left(\frac{1}{n_o} + \frac{1}{n_b} \right) \sqrt{1+a_{\max}^4} \left(p^{2|\xi|} + 1 \right), \quad (278)$$

$$N_1 = 7 \binom{d+k}{d} \binom{d+|\xi|}{d}^2 n_o^2 + 4 \binom{d+|\xi|}{d} n_o n_b + n_b^2, \quad (279)$$

be constants. If m is large enough so that

$$m \geq \frac{32N_1 C_1^2 R^{4p}}{\lambda_0^2} \log \left(\frac{2N_1}{\delta} \right), \quad (280)$$

then with the probability of at least $1 - \delta$ over the initialization, we have

$$\left\| \widehat{\mathbf{G}}_{\mathbf{v}}(\mathbf{w}(0), \mathbf{v}(0)) - \widehat{\mathbf{G}}_{\mathbf{v}}^{\infty} \right\|_2 < \frac{\lambda_0}{4}. \quad (281)$$

Proof. We use Hoeffding's inequality to measure the error between $\widehat{\mathbf{G}}_{\mathbf{v}}(\mathbf{w}, \mathbf{v})$ and $\widehat{\mathbf{G}}_{\mathbf{v}}^{\infty}$, as Gram matrix $\widehat{\mathbf{G}}_{\mathbf{v}}^{\infty}$ is the expectation of the sample mean. In the Hoeffding's inequalities (220), (228), (233), (241), (245), (251), and (254), we have

$$\frac{1}{n_o^2} \binom{d+k}{d}^2 \binom{d+|\xi|}{d}^4 a_{\max}^4 p^{4|\xi|} \leq C_1^2, \quad (282)$$

$$\frac{\nu_{\ell}}{n_o^2} \binom{d+|\xi|}{d}^2 (p^{|\xi|} + 1)^2 a_{\max}^2 p^{2|\xi|} \leq C_1^2, \quad (283)$$

$$\frac{\nu}{n_o n_b} (d+1)^2 \binom{d+|\xi|}{d}^2 a_{\max}^4 p^{2|\xi|+2} \leq C_1^2, \quad (284)$$

$$\frac{\nu_{\ell}^2}{n_o^2} (p^{2|\xi|} + 1)^2 \leq C_1^2 \quad (285)$$

$$\frac{\nu_{\ell} \nu_{\ell+1}}{n_o^2} p^{2|\xi|} \leq C_1^2, \quad (286)$$

$$\frac{\nu_1 \nu}{n_o n_b} (d+1)^2 a_{\max}^2 p^{2|\xi|+2} \leq C_1^2, \quad (287)$$

$$\frac{\nu^2}{n_b^2} (d+1)^4 a_{\max}^4 p^4 \leq C_1^2. \quad (288)$$

Consequently, all inequalities induce that each component of $\left| \widehat{\mathbf{G}}_{\mathbf{v}} - \widehat{\mathbf{G}}_{\mathbf{v}}^{\infty} \right|$ is greater than ε , with a probability of at most $2 \exp \left(-\frac{m\varepsilon^2}{2C_1^2 R^{4p}} \right)$. Since there exists at most N_1 nonzero elements in $\left| \widehat{\mathbf{G}}_{\mathbf{v}} - \widehat{\mathbf{G}}_{\mathbf{v}}^{\infty} \right|$ by (145),

$$\left| \widehat{\mathbf{G}}_{\mathbf{v}}(\mathbf{w}, \mathbf{v}) - \widehat{\mathbf{G}}_{\mathbf{v}}^{\infty} \right| \leq \varepsilon \quad (289)$$

holds componentwise, with the probability of at least

$$\left(1 - 2 \exp \left(-\frac{m\varepsilon^2}{2C_1^2 R^{4p}} \right) \right)^{N_1} \geq 1 - 2N_1 \exp \left(-\frac{m\varepsilon^2}{2C_1^2 R^{4p}} \right). \quad (290)$$

Then, we have

$$\left\| \widehat{\mathbf{G}}_{\mathbf{v}}(\mathbf{w}, \mathbf{v}) - \widehat{\mathbf{G}}_{\mathbf{v}}^{\infty} \right\|_2 \leq \left\| \widehat{\mathbf{G}}_{\mathbf{v}}(\mathbf{w}, \mathbf{v}) - \widehat{\mathbf{G}}_{\mathbf{v}}^{\infty} \right\|_F \leq \varepsilon \sqrt{N_1}. \quad (291)$$

Set ε and m to satisfy $\varepsilon = \frac{1}{4\sqrt{N_1}} \lambda_0$ and $2N \exp \left(-\frac{m\varepsilon^2}{2C_1^2 R^{4p}} \right) < \delta$. In other words, if

$$m \geq \frac{2C_1^2 R^{4p}}{\varepsilon^2} \log \left(\frac{2N_1}{\delta} \right) \quad (292)$$

$$= \frac{32N_1 C_1^2 R^{4p}}{\lambda_0^2} \log \left(\frac{2N_1}{\delta} \right), \quad (293)$$

then $\left\| \widehat{\mathbf{G}}_{\mathbf{v}}(\mathbf{w}, \mathbf{v}) - \widehat{\mathbf{G}}_{\mathbf{v}}^{\infty} \right\|_2 \leq \frac{1}{4} \lambda_0$ with probability of at least $1 - \delta$. $\square$

By the above proposition, the initial Gram matrix $\widehat{\mathbf{G}}_{\mathbf{v}}(\mathbf{w}(0), \mathbf{v}(0))$ is likely to be strictly positive definite, and its smallest eigenvalue remains greater than $\frac{\lambda_0}{2}$, at the beginning of the flow. The following lemma implies that at the early stage of the flow, for which $\mathbf{w}$ and $\mathbf{v}$ are not far from the initial values, the positive definiteness is preserved, and its smallest eigenvalue remains distinct from zero.

Lemma C.5. Suppose $\|\tilde{\mathbf{w}}_\ell - \mathbf{w}_\ell(0)\|_2 \leq R_w < R$, for

$$R_w = \frac{\lambda_0}{8\sqrt{2}} \left(\nu_{\max} \sqrt{1 + a_{\max}^4} \binom{d+k}{d} \binom{d+|\xi|}{d}^2 p^{2|\xi|+1} (2R)^{2p-1} \right)^{-1}. \quad (294)$$

Then

$$\left\| \widehat{\mathbf{G}}_v(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - \widehat{\mathbf{G}}_v(\mathbf{w}(0), \mathbf{v}(0)) \right\|_2 \leq \frac{1}{4} \lambda_0. \quad (295)$$

Proof. **1. Case of $\mathbf{Q}_{0,0}$:** By the mean value theorem with (151), we can induce that

$$(\mathbf{Q}_{0,0}(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - \mathbf{Q}_{0,0}(\mathbf{w}(0), \mathbf{v}(0)))_{i_1, i_2} \quad (296)$$

$$\leq \left| \mathbf{Q}_{0,0}(\tilde{\mathbf{w}}, \tilde{\mathbf{v}})_{i_1, i_2} - \mathbf{Q}_{0,0}(\mathbf{w}(0), \tilde{\mathbf{v}})_{i_1, i_2} \right| + \left| \mathbf{Q}_{0,0}(\tilde{\mathbf{w}}, \tilde{\mathbf{v}})_{i_1, i_2} - \mathbf{Q}_{0,0}(\tilde{\mathbf{w}}, \mathbf{v}(0))_{i_1, i_2} \right| \quad (297)$$

$$= \left| \frac{\partial}{\partial \mathbf{w}} \mathbf{Q}_{0,0}(\mathbf{w}^*, \mathbf{v}(0))_{i_1, i_2} \cdot (\tilde{\mathbf{w}} - \mathbf{w}(0)) \right| + \left| \frac{\partial}{\partial \mathbf{v}} \mathbf{Q}_{0,0}(\mathbf{w}(0), \mathbf{v}^*)_{i_1, i_2} (\tilde{\mathbf{v}} - \mathbf{v}(0)) \right| \quad (298)$$

$$\leq \sum_{\ell=0}^L \sum_{r=1}^m \left(\left\| \frac{\partial}{\partial (\mathbf{w}_\ell)_r} (\mathbf{Q}_{0,0})(\mathbf{w}^*, \mathbf{v}(0))_{i_1, i_2} \right\|_2 \cdot \|(\tilde{\mathbf{w}}_\ell)_r - (\mathbf{w}_\ell)_r(0)\|_2 \right) \quad (299)$$

$$\leq \frac{1}{mn_o} a_{\max}^2 p^{2|\xi|+1} \sum_{\ell=0}^L \sum_{r=1}^m |I_{\xi_\ell}| |J_{\Delta \xi_{\ell+1}}|^2 \left(\|(\mathbf{w}_\ell^*)_r\|_2^{2p-1} \cdot \|(\tilde{\mathbf{w}}_\ell)_r - (\mathbf{w}_\ell)_r(0)\|_2 \right), \quad (300)$$

for some $\mathbf{w}^* = [(\mathbf{w}_0^*)_1^\top \cdots (\mathbf{w}_0^*)_m^\top \cdots (\mathbf{w}_L^*)_0^\top \cdots (\mathbf{w}_L^*)_m^\top]^\top$ lies on a line between $\tilde{\mathbf{w}}$ and $\mathbf{w}(0)$ and for some $\mathbf{v}^*$ lies on a line connecting $\tilde{\mathbf{v}}$ and $\mathbf{v}(0)$. Since $\|(\mathbf{w}_\ell^*)_r\|_2 \leq \|(\tilde{\mathbf{w}}_\ell)_r - (\mathbf{w}_\ell)_r(0)\|_2 + \|(\mathbf{w}_\ell)_r(0)\|_2 \leq R_w + R \leq 2R$ for each $\ell \in \{0\} \cup [L]$ and $r \in [m]$,

$$\begin{aligned} & (\mathbf{Q}_{0,0}(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - \mathbf{Q}_{0,0}(\mathbf{w}(0), \mathbf{v}(0)))_{i_1, i_2} \\ & \leq \frac{a_{\max}^2}{mn_o} p^{2|\xi|+1} \sum_{\ell=0}^L \sum_{r=1}^m |I_{\xi_\ell}| |J_{\Delta \xi_{\ell+1}}|^2 \left((2R)^{2p-1} \|(\tilde{\mathbf{w}}_\ell)_r - (\mathbf{w}_\ell)_r(0)\|_2 \right) \\ & \leq \frac{a_{\max}^2}{n_o} p^{2|\xi|+1} \binom{d+|\xi|}{d}^2 \sum_{\ell=0}^L |I_{\xi_\ell}| (2R)^{2p-1} R_w \\ & \leq \frac{a_{\max}^2}{n_o} p^{2|\xi|+1} \binom{d+k}{d} \binom{d+|\xi|}{d}^2 (2R)^{2p-1} R_w \end{aligned} \quad (301)$$

Therefore, we can attain that

$$\begin{aligned} & \sum_{i_1=1}^{n_o} \sum_{i_2=1}^{n_o} (\mathbf{Q}_{0,0}(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - \mathbf{Q}_{0,0}(\mathbf{w}(0), \mathbf{v}(0)))_{i_1, i_2}^2 \\ & \leq a_{\max}^4 \binom{d+k}{d}^2 \binom{d+|\xi|}{d}^4 2p^{2(2|\xi|+1)} (2R)^{2(2p-1)} R_w^2. \end{aligned} \quad (302)$$

2. Case of $\mathbf{Q}_{0,\ell}$: Similar to $\mathbf{Q}_{0,0}$, the mean value theorem with (159) and (163) gives

$$(\mathbf{Q}_{0,\ell}(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - \mathbf{Q}_{0,\ell}(\mathbf{w}(0), \mathbf{v}(0)))_{i_1, ((\alpha, \beta), i_2)} \quad (303)$$

$$\leq \frac{\sqrt{\nu_{\max}}}{mn_o} a_{\max} \binom{d+|\xi|}{d} p^{2|\xi|+1} \sum_{r=1}^m \left\{ \|(\tilde{\mathbf{w}}_{\ell-1})_r\|_2^{2p-1} \|(\tilde{\mathbf{w}}_{\ell-1})_r - (\mathbf{w}_{\ell-1})_r(0)\|_2 \right. \quad (304)$$

$$\left. + \|(\tilde{\mathbf{w}}_\ell)_r\|_2^{2p-1} \|(\tilde{\mathbf{w}}_\ell)_r - (\mathbf{w}_\ell)_r(0)\|_2 \right\} \quad (305)$$

$$\leq \frac{\sqrt{\nu_{\max}}}{mn_o} a_{\max} \binom{d+|\xi|}{d} p^{2|\xi|+1} \sum_{r=1}^m 2 \left((2R)^{2p-1} R_w \right) \quad (306)$$

$$= \frac{\sqrt{\nu_{\max}}}{n_o} a_{\max} \binom{d+|\xi|}{d} p^{2|\xi|+1} 2(2R)^{2p-1} R_w. \quad (307)$$

Consequently, we have

$$\sum_{i_1, i_2=1}^{n_o} \sum_{\ell=1}^L \sum_{\alpha \in I_{\xi_{\ell-1}}} \sum_{\beta \in I_{\Delta \xi_{\ell}}} (\mathbf{Q}_{0,\ell}(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - \mathbf{Q}_{0,\ell}(\mathbf{w}(0), \mathbf{v}(0)))_{i_1, ((\alpha, \beta), i_2)}^2 \quad (308)$$

$$\leq \frac{\nu_{\max}^2}{n_o^2} a_{\max}^2 \binom{d+|\xi|}{d}^2 p^{2(2|\xi|+1)} \sum_{i_1, i_2=1}^{n_o} \sum_{\ell=1}^L \sum_{\alpha \in I_{\xi_{\ell-1}}} \sum_{\beta \in I_{\Delta \xi_{\ell}}} 4(2R)^{2(2p-1)} R_w^2 \quad (309)$$

$$\leq \nu_{\max}^2 a_{\max}^2 \binom{d+|\xi|}{d}^2 p^{2(2|\xi|+1)} \sum_{\ell=1}^L |I_{\xi_{\ell-1}}| |I_{\Delta \xi_{\ell}}| 4(2R)^{2(2p-1)} R_w^2 \quad (310)$$

$$\leq \nu_{\max}^2 a_{\max}^2 \binom{d+k}{d} \binom{d+|\xi|}{d}^4 4(2R)^{2(2p-1)} R_w^2. \quad (311)$$

3. Case of $\mathbf{Q}_{0,L+1}$: The mean value theorem with (171) induces

$$\begin{aligned} & ((\mathbf{Q}_{0,L+1})(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - (\mathbf{Q}_{0,L+1})(\mathbf{w}(0), \mathbf{v}(0)))_{i,j} \\ & \leq \frac{\sqrt{\nu_{\max}}}{m\sqrt{n_o n_b}} a_{\max}^2 (d+1) \binom{d+|\xi|}{d} p^{|\xi|+2} \sum_{r=1}^m \left(\|(\tilde{\mathbf{w}}_0)_r\|_2^{2p-1} \|(\tilde{\mathbf{w}}_0)_r - (\mathbf{w}_0)_r(0)\|_2 \right) \\ & \leq \frac{\sqrt{\nu_{\max}}}{\sqrt{n_o n_b}} a_{\max}^2 (d+1) \binom{d+|\xi|}{d} p^{|\xi|+2} (2R)^{2p-1} R_w, \end{aligned} \quad (312)$$

and hence

$$\begin{aligned} & \sum_{i=1}^{n_o} \sum_{j=1}^{n_b} ((\mathbf{Q}_{0,L+1})(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - (\mathbf{Q}_{0,L+1})(\mathbf{w}(0), \mathbf{v}(0)))_{i,j}^2 \\ & \leq \nu_{\max} a_{\max}^4 (d+1)^2 \binom{d+|\xi|}{d}^2 p^{2(|\xi|+2)} (2R)^{2(2p-1)} R_w^2 \end{aligned} \quad (313)$$

4. Case of $\mathbf{Q}_{\ell_1, \ell_2}$: From the mean value theorem with (180) and (184), we attain

$$((\mathbf{Q}_{\ell,\ell})(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - (\mathbf{Q}_{\ell,\ell})(\mathbf{w}(0), \mathbf{v}(0)))_{((\alpha_1, \beta_1), i_1), ((\alpha_2, \beta_2), i_2)} \quad (314)$$

$$\leq \frac{\nu_{\max}}{mn_o} p^{2|\xi|+1} \sum_{r=1}^m \left\{ \mathbf{1}_{\{\alpha_1=\alpha_2\}} \|(\tilde{\mathbf{w}}_{\ell-1})_r\|_2^{2p-1} \|(\tilde{\mathbf{w}}_{\ell-1})_r - (\mathbf{w}_{\ell-1})_r(0)\|_2 \right. \quad (315)$$

$$\left. + \mathbf{1}_{\{\alpha_1+\beta_1=\alpha_2+\beta_2\}} \|(\tilde{\mathbf{w}}_{\ell})_r\|_2^{2p-1} \|(\tilde{\mathbf{w}}_{\ell})_r - (\mathbf{w}_{\ell})_r(0)\|_2 \right\} \quad (316)$$

$$\leq \frac{\nu_{\max}}{n_o} p^{2|\xi|+1} \left\{ \mathbf{1}_{\{\alpha_1=\alpha_2\}} (2R)^{2p-1} + \mathbf{1}_{\{\alpha_1+\beta_1=\alpha_2+\beta_2\}} (2R)^{2p-1} \right\} R_w, \quad (317)$$

and further induces

$$\sum_{i_1, i_2=1}^{n_o} \sum_{\ell=1}^L \sum_{\alpha_1, \alpha_2 \in I_{\xi_{\ell-1}}} \sum_{\beta_1, \beta_2 \in I_{\Delta \xi_{\ell}}} ((\mathbf{Q}_{\ell,\ell})(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - (\mathbf{Q}_{\ell,\ell})(\mathbf{w}(0), \mathbf{v}(0)))_{((\alpha_1, \beta_1), i_1), ((\alpha_2, \beta_2), i_2)}^2 \quad (318)$$

$$\leq \nu_{\max}^2 p^{2(2|\xi|+1)} \sum_{\ell=1}^L \sum_{\alpha_1 \in I_{\xi_{\ell-1}}} \sum_{\beta_1 \in I_{\Delta \xi_{\ell}}} \sum_{\alpha_2 \in I_{\xi_{\ell-1}}} \sum_{\beta_2 \in I_{\Delta \xi_{\ell}}} \left\{ \mathbf{1}_{\{\alpha_1=\alpha_2\}} (2R)^{2p-1} \right. \quad (319)$$

$$\left. + \mathbf{1}_{\{\alpha_1+\beta_1=\alpha_2+\beta_2\}} (2R)^{2(2p-1)} \right\}^2 R_w^2 \quad (320)$$

$$\leq \nu_{\max}^2 p^{2(2|\xi|+1)} \sum_{\ell=1}^L \sum_{\alpha_1 \in I_{\xi_{\ell-1}}} \sum_{\beta_1 \in I_{\Delta \xi_{\ell}}} \left((|I_{\Delta \xi_{\ell}}| - 1) (2R)^{2(2p-1)} + (|I_{\Delta \xi_{\ell}}| - 1) (2R)^{2p-1} \right) \quad (321)$$

$$+4(2R)^{2(2p-1)} R_w^2 \quad (322)$$

$$\leq \nu_{\max}^2 p^{2(2|\xi|+1)} \sum_{\ell=1}^L \sum_{\alpha_1 \in I_{\xi_{\ell-1}}} \sum_{\beta_1 \in I_{\Delta \xi_{\ell}}} 4 |I_{\Delta \xi_{\ell}}| (2R)^{2(2p-1)} R_w^2 \quad (323)$$

$$\leq 4\nu_{\max}^2 p^{2(2|\xi|+1)} \sum_{\ell=1}^L |I_{\xi_{\ell-1}}| |I_{\Delta \xi_{\ell}}|^2 (2R)^{2p-1} R_w^2 \quad (324)$$

$$\leq 4\nu_{\max}^2 p^{2(2|\xi|+1)} \binom{d+k}{d} \binom{d+|\xi|}{d}^2 (2R)^{2p-1} R_w^2. \quad (325)$$

(192) follows that

$$((Q_{\ell,\ell+1})(\tilde{w}, \tilde{v}) - (Q_{\ell,\ell+1})(w(0), v(0)))_{((\alpha_1, \beta_1), i_1), ((\alpha_2, \beta_2), i_2)} \quad (326)$$

$$\leq \mathbf{1}_{\{\alpha_2 = \alpha_1 + \beta_1\}} \frac{\nu_{\max}}{mn_o} p^{|\xi|+1} \sum_{r=1}^m \left(\|(\tilde{w}_{\ell})_r\|_2^{2p-1} \|(\tilde{w}_{\ell})_r - (w_{\ell})_r(0)\|_2 \right) \quad (327)$$

$$\leq \mathbf{1}_{\{\alpha_2 = \alpha_1 + \beta_1\}} \frac{\nu_{\max}}{n_o} p^{|\xi|+1} (2R)^{2p-1} R_w. \quad (328)$$

It deduces that

$$\sum_{i_1, i_2=1}^{n_o} \sum_{\ell=1}^{L-1} \sum_{\substack{\alpha_1 \in I_{\xi_{\ell-1}} \\ \alpha_2 \in I_{\xi_{\ell}}}} \sum_{\substack{\beta_1 \in I_{\Delta \xi_{\ell}} \\ \beta_2 \in I_{\Delta \xi_{\ell+1}}}} ((Q_{\ell,\ell+1})(\tilde{w}, \tilde{v}) - (Q_{\ell,\ell+1})(w(0), v(0)))_{((\alpha_1, \beta_1), i_1), ((\alpha_2, \beta_2), i_2)}^2 \quad (329)$$

$$\leq \nu_{\max}^2 p^{2|\xi|+2} \sum_{\ell=1}^{L-1} \sum_{\alpha_1 \in I_{\xi_{\ell-1}}} \sum_{\beta_1 \in I_{\Delta \xi_{\ell}}} \sum_{\alpha_2 \in I_{\xi_{\ell}}} \sum_{\beta_2 \in I_{\Delta \xi_{\ell+1}}} \mathbf{1}_{\{\alpha_2 = \alpha_1 + \beta_1\}} (2R)^{2(2p-1)} R_w^2 \quad (330)$$

$$\leq \nu_{\max}^2 p^{2|\xi|+2} \sum_{\ell=1}^{L-1} \sum_{\alpha_1 \in I_{\xi_{\ell-1}}} \sum_{\beta_1 \in I_{\Delta \xi_{\ell}}} |I_{\Delta \xi_{\ell+1}}| (2R)^{2(2p-1)} R_w^2 \quad (331)$$

$$= \nu_{\max}^2 p^{2|\xi|+2} \sum_{\ell=1}^{L-1} |I_{\xi_{\ell-1}}| |I_{\Delta \xi_{\ell}}| |I_{\Delta \xi_{\ell+1}}| (2R)^{2(2p-1)} R_w^2 \quad (332)$$

$$\leq \nu_{\max}^2 p^{2|\xi|+2} \binom{d+k}{d} \binom{d+|\xi|}{d}^2 (2R)^{2(2p-1)} R_w^2. \quad (333)$$

In accordance with the definition of the loss μ_{ℓ} , it is clear that Q_{ℓ_1, ℓ_2} is identical to the zero matrix if the difference between ℓ_1 and ℓ_2 is greater than 1. Consequently, together with (325) and (333) we have

$$\sum_{\ell_1, \ell_2=1}^L \sum_{\alpha_1, \alpha_2 \in I_{\xi_{\ell-1}}} \sum_{\beta_1, \beta_2 \in I_{\Delta \xi_{\ell}}} \sum_{i_1, i_2=1}^{n_o} ((Q_{\ell_1, \ell_2})(w(0), v(0)))_{((\alpha_1, \beta_1), i_1), ((\alpha_2, \beta_2), i_2)}^2 \quad (334)$$

$$= \sum_{i_1, i_2=1}^{n_o} \sum_{\ell=1}^L \sum_{\alpha_1, \alpha_2 \in I_{\xi_{\ell-1}}} \sum_{\beta_1, \beta_2 \in I_{\Delta \xi_{\ell}}} ((Q_{\ell, \ell})(\tilde{w}, \tilde{v}) - (Q_{\ell, \ell})(w(0), v(0)))_{((\alpha_1, \beta_1), i_1), ((\alpha_2, \beta_2), i_2)}^2 \quad (335)$$

$$+ 2 \sum_{i_1, i_2=1}^{n_o} \sum_{\ell=1}^{L-1} \sum_{\substack{\alpha_1 \in I_{\xi_{\ell-1}} \\ \alpha_2 \in I_{\xi_{\ell}}}} \sum_{\substack{\beta_1 \in I_{\Delta \xi_{\ell}} \\ \beta_2 \in I_{\Delta \xi_{\ell+1}}}} ((Q_{\ell, \ell+1})(\tilde{w}, \tilde{v}) - (Q_{\ell, \ell+1})(w(0), v(0)))_{((\alpha_1, \beta_1), i_1), ((\alpha_2, \beta_2), i_2)}^2 \quad (336)$$

$$\leq 4\nu_{\max}^2 p^{2(2|\xi|+1)} \binom{d+k}{d} \binom{d+|\xi|}{d}^2 (2R)^{2p-1} R_w^2 \quad (337)$$

$$+ 2\nu_{\max}^2 p^{2|\xi|+2} \binom{d+k}{d} \binom{d+|\xi|}{d}^2 (2R)^{2(2p-1)} R_w^2 \quad (338)$$

$$\leq 6\nu_{\max}^2 \binom{d+k}{d} \binom{d+|\xi|}{d}^2 p^{2(2|\xi|+1)} (2R)^{2p-1} R_w^2. \quad (339)$$

5. Case of $\mathbf{Q}_{\ell,L+1}$: From (201), we have

$$((\mathbf{Q}_{1,L+1})(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - (\mathbf{Q}_{1,L+1})(\mathbf{w}(0), \mathbf{v}(0)))_{((\alpha,\beta),i),j} \quad (340)$$

$$\leq \frac{\nu_{\max}}{m\sqrt{n_o n_b}} a_{\max} (d+1) p^{|\xi|+2} \sum_{r=1}^m \left(\|(\tilde{\mathbf{w}}_0)_r\|_2^{2p-1} \|(\tilde{\mathbf{w}}_0)_r - (\mathbf{w}_0)_r(0)\|_2 \right) \quad (341)$$

$$\leq \frac{\nu_{\max}}{\sqrt{n_o n_b}} a_{\max} (d+1) p^{|\xi|+2} (2R)^{2p-1} R_w. \quad (342)$$

By the definition of the loss μ_ℓ , we have $\mathbf{Q}_{\ell,L+1}$ is identical to the zero matrix for every $\ell > 1$. Therefore, we have

$$\sum_{i=1}^{n_o} \sum_{j=1}^{n_b} \sum_{\ell=1}^L \sum_{\beta \in I_{\Delta\xi_1}} ((\mathbf{Q}_{\ell,L+1})(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - (\mathbf{Q}_{\ell,L+1})(\mathbf{w}(0), \mathbf{v}(0)))_{((1,\beta),i),j}^2 \quad (343)$$

$$= \sum_{i=1}^{n_o} \sum_{j=1}^{n_b} \sum_{\beta \in I_{\Delta\xi_1}} ((\mathbf{Q}_{1,L+1})(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - (\mathbf{Q}_{1,L+1})(\mathbf{w}(0), \mathbf{v}(0)))_{((1,\beta),i),j}^2 \quad (344)$$

$$\leq |I_{\Delta\xi_1}| \nu_{\max}^2 a_{\max}^2 (d+1)^2 p^{2(|\xi|+2)} (2R)^{2(2p-1)} R_w^2 \quad (345)$$

$$\leq \nu_{\max}^2 a_{\max}^2 \binom{d+|\xi|-1}{|\xi|} (d+1)^2 p^{2(|\xi|+2)} (2R)^{2(2p-1)} R_w^2 \quad (346)$$

$$\leq \nu_{\max}^2 a_{\max}^2 (d+1)^2 \binom{d+|\xi|}{d} p^{2(|\xi|+2)} (2R)^{2(2p-1)} R_w^2. \quad (347)$$

6. Case of $\mathbf{Q}_{L+1,L+1}$:

From (205), we have

$$((\mathbf{Q}_{L+1,L+1})(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - (\mathbf{Q}_{L+1,L+1})(\mathbf{w}(0), \mathbf{v}(0)))_{j_1,j_2} \quad (348)$$

$$\leq \frac{\nu}{mn_b} a_{\max}^2 (d+1)^2 p^3 \sum_{r=1}^m \left(\|(\tilde{\mathbf{w}}_0)_r\|_2^{2p-1} \|(\tilde{\mathbf{w}}_0)_r - (\mathbf{w}_0)_r(0)\|_2 \right) \quad (349)$$

$$\leq \frac{\nu_{\max}}{n_b} a_{\max}^2 (d+1)^2 p^3 (2R)^{2p-1} R_w. \quad (350)$$

It further follows that

$$\begin{aligned} & \sum_{j_1=1}^{n_b} \sum_{j_2=1}^{n_b} ((\mathbf{Q}_{L+1,L+1})(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - (\mathbf{Q}_{L+1,L+1})(\mathbf{w}(0), \mathbf{v}(0)))_{j_1,j_2}^2 \\ & \leq \nu_{\max}^2 a_{\max}^4 (d+1)^4 p^6 (2R)^{2(2p-1)} R_w^2. \end{aligned} \quad (351)$$

Therefore, combining (302), (311), (313), (339), (347), and (351) concludes that

$$\left\| \hat{\mathbf{G}}_{\mathbf{v}}(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - \hat{\mathbf{G}}_{\mathbf{v}}(\mathbf{w}(0), \mathbf{v}(0)) \right\|_2^2 \quad (352)$$

$$\leq \left\| \hat{\mathbf{G}}_{\mathbf{v}}(\tilde{\mathbf{w}}, \tilde{\mathbf{v}}) - \hat{\mathbf{G}}_{\mathbf{v}}(\mathbf{w}(0), \mathbf{v}(0)) \right\|_F^2 \quad (353)$$

$$\leq a_{\max}^4 \binom{d+k}{d}^2 \binom{d+|\xi|}{d}^4 p^{2(2|\xi|+1)} (2R)^{2(2p-1)} R_w^2 \quad (354)$$

$$+ 4\nu_{\max}^2 a_{\max}^2 \binom{d+k}{d} \binom{d+|\xi|}{d}^4 p^{2(2|\xi|+1)} (2R)^{2(2p-1)} R_w^2 \quad (355)$$

$$+ \nu_{\max} a_{\max}^4 (d+1)^2 \binom{d+|\xi|}{d}^2 p^{2(|\xi|+2)} (2R)^{2(2p-1)} R_w^2 \quad (356)$$

$$+ 6\nu_{\max}^2 p^{2(2|\xi|+1)} \binom{d+k}{d} \binom{d+|\xi|}{d}^2 (2R)^{2p-1} R_w^2 \quad (357)$$

$$+ \nu_{\max}^2 a_{\max}^2 (d+1)^2 \binom{d+|\xi|}{d} p^{2(|\xi|+2)} (2R)^{2(2p-1)} R_w^2 \quad (358)$$

$$+ \nu_{\max}^2 a_{\max}^4 (d+1)^4 p^6 (2R)^{2(2p-1)} R_w^2 \quad (359)$$

$$\leq 8\nu_{\max}^2 (1 + a_{\max}^4) \binom{d+k}{d}^2 \binom{d+|\xi|}{d}^4 p^{2(2|\xi|+1)} (2R)^{2(2p-1)} R_w^2 \quad (360)$$

$$= \frac{\lambda_0^2}{16}. \quad (361)$$

□

The preceding lemma indicates that the loss will decrease rapidly in cases where $\mathbf{w}$ is not significantly distant from $\mathbf{w}(0)$. Indeed, the subsequent lemma and proposition demonstrate that $\mathbf{w}(t)$ remains within the designated region for any given value of $t > 0$, provided sufficiently large m .

Proposition C.6. Set constants C_2 and C_3 as

$$C_2 = \sqrt{\nu_{\max}} \sqrt{(1 + a_{\max}^2)} \binom{d+k}{d} \binom{d+|\xi|}{d} (p^{|\xi|} + 1), \quad (362)$$

$$C_3 = \frac{1}{n_o} \sum_{i=1}^{n_o} f(\mathbf{x}_i)^2 + \frac{1}{n_b} \sum_{j=1}^{n_b} g(\tilde{\mathbf{x}}_j)^2, \quad (363)$$

$$N_2 = \binom{d+k}{d} \binom{d+|\xi|}{d} n_o + n_b. \quad (364)$$

For any $C > 0$ and $\delta < 2N_2 \exp\left(-\frac{C_3^2}{2C_2^2 R^{2p}}\right)$, if m is large enough so that

$$m \geq \frac{2C_2^2}{\lambda_0^2} \left(C_2 \left(4 + \binom{d+k}{d} \binom{d+|\xi|}{d} \right) R^{4p} \log\left(\frac{2N_2}{\delta}\right) + C_3 R^{2p} \right), \quad (365)$$

then, with the probability of at least $1 - \delta$ over the initialization, we have

$$\frac{1}{\sqrt{m}} \left\| \begin{bmatrix} \mathbf{s}(\mathbf{w}(0), \mathbf{v}(0)) \\ \boldsymbol{\mu}_1(\mathbf{w}(0), \mathbf{v}(0)) \\ \vdots \\ \boldsymbol{\mu}_L(\mathbf{w}(0), \mathbf{v}(0)) \\ \mathbf{h}(\mathbf{w}(0), \mathbf{v}(0)) \end{bmatrix} \right\|_2 < \frac{\lambda_0}{C} R^{-p}. \quad (366)$$

Proof. Since

$$\nu_{\max} a_{\max}^2 \binom{d+k}{d}^2 \binom{d+|\xi|}{d}^2 p^{2|\xi|} \leq C_2^2, \quad (367)$$

$$\nu_{\max} \binom{d+k}{d}^2 \binom{d+|\xi|}{d}^2 (p^{|\xi|} + 1)^2 \leq C_2^2, \quad (368)$$

$$\nu_{\max} a_{\max}^2 (d+1)^2 p^2 \leq C_2^2, \quad (369)$$

we have

$$\mathbb{P} \left[\left| \sum_{r=1}^m s_{i,r}(\mathbf{w}(0), \mathbf{v}(0)) \right| > \frac{\varepsilon}{\sqrt{n_o}} \right] \leq 2 \exp \left(-\frac{\varepsilon^2}{2C_2^2 R^{2p}} \right), \quad (370)$$

$$\mathbb{P} \left[\left| \sum_{r=1}^m (\boldsymbol{\mu}_\ell)_{r,(\alpha,\beta),i}(\mathbf{w}(0), \mathbf{v}(0)) \right| > \frac{\varepsilon}{\sqrt{n_o}} \right] \leq 2 \exp \left(-\frac{\varepsilon^2}{2C_2^2 R^{2p}} \right), \quad (371)$$

$$\mathbb{P} \left[\left| \sum_{r=1}^m h_{j,r}(\mathbf{w}(0), \mathbf{v}(0)) \right| > \frac{\varepsilon}{\sqrt{n_b}} \right] \leq 2 \exp \left(-\frac{\varepsilon^2}{2C_2^2 R^{2p}} \right). \quad (372)$$

Hence, from

$$n_o + \sum_{\ell=1}^L |I_{\xi_{\ell-1}}| \cdot |I_{\Delta \xi_{\ell}}| \cdot n_o + n_b \leq \binom{d+k}{d} \binom{d+|\xi|}{d} n_o + n_b \leq N_2, \quad (373)$$

with the probability of at least

$$\left(1 - 2 \exp \left(-\frac{\varepsilon^2}{2C_2^2 R^{2p}} \right) \right)^{N_2} \geq 1 - 2N \exp \left(-\frac{\varepsilon^2}{2C_2^2 R^{2p}} \right), \quad (374)$$

all the conditions

$$\left| \sum_{r=1}^m s_{i,r}(\mathbf{w}(0), \mathbf{v}(0)) \right| \leq \frac{\varepsilon}{\sqrt{n_o}}, \quad (375)$$

$$\left| \sum_{r=1}^m (\boldsymbol{\mu}_{\ell})_{r,(\alpha,\beta),i}(\mathbf{w}(0), \mathbf{v}(0)) \right| \leq \frac{\varepsilon}{\sqrt{n_o}}, \quad (376)$$

$$\left| \sum_{r=1}^m h_{j,r}(\mathbf{w}(0), \mathbf{v}(0)) \right| \leq \frac{\varepsilon}{\sqrt{n_b}} \quad (377)$$

are satisfied. Then, the square of the initial loss is bounded as

$$\|\mathbf{s}(\mathbf{w}(0), \mathbf{v}(0))\|_2^2 + \|\mathbf{h}(\mathbf{w}(0), \mathbf{v}(0))\|_2^2 + \sum_{\ell=1}^L \|\boldsymbol{\mu}_{\ell}(\mathbf{w}(0), \mathbf{v}(0))\|_2^2 \quad (378)$$

$$= \sum_{i=1}^{n_o} s_i(\mathbf{w}(0), \mathbf{v}(0))^2 + \sum_{j=1}^{n_b} h_j(\mathbf{w}(0), \mathbf{v}(0))^2 \quad (379)$$

$$+ \sum_{i=1}^{n_o} \sum_{\ell=1}^L \sum_{\alpha \in I_{\xi_{\ell-1}}} \sum_{\beta \in I_{\Delta \xi_{\ell}}} (\boldsymbol{\mu}_{\ell})_{(\alpha,\beta),i}(\mathbf{w}(0), \mathbf{v}(0))^2 \quad (380)$$

$$\leq \frac{1}{n_o} \sum_{i=1}^{n_o} (\varepsilon + |f(\mathbf{x}_i)|)^2 + \frac{1}{n_b} \sum_{j=1}^{n_b} (\varepsilon + |g(\tilde{\mathbf{x}}_j)|)^2 + \sum_{\ell=1}^L |I_{\xi_{\ell-1}}| \cdot |I_{\Delta \xi_{\ell}}| \varepsilon^2 \quad (381)$$

$$\leq \frac{1}{n_o} \sum_{i=1}^{n_o} (\varepsilon + |f(\mathbf{x}_i)|)^2 + \frac{1}{n_b} \sum_{j=1}^{n_b} (\varepsilon + |g(\tilde{\mathbf{x}}_j)|)^2 + \binom{d+k}{d} \binom{d+|\xi|}{d} \varepsilon^2 \quad (382)$$

$$\leq \left(4 + \binom{d+k}{d} \binom{d+|\xi|}{d} \right) \varepsilon^2 + 2 \left(\frac{1}{n_o} \sum_{i=1}^{n_o} f(\mathbf{x}_i)^2 + \frac{1}{n_b} \sum_{j=1}^{n_b} g(\tilde{\mathbf{x}}_j)^2 \right) \quad (383)$$

$$= \left(4 + \binom{d+k}{d} \binom{d+|\xi|}{d} \right) \varepsilon^2 + 2C_3. \quad (384)$$

For $\delta > 0$ and $C > 0$, if ε and m satisfy the following inequalities,

$$\delta \geq 2N_2 \exp \left(-\frac{\varepsilon^2}{2C_2^2 R^{2p}} \right), \quad (385)$$

$$\varepsilon > C_3, \quad (386)$$

$$m \geq \frac{C^2}{\lambda_0^2} R^{2p} \left(\left(4 + \binom{d+k}{d} \binom{d+|\xi|}{d} \right) \varepsilon^2 + 2C_3 \right), \quad (387)$$

then for enough small δ such that

$$\delta < 2N_2 \exp \left(-\frac{C_3^2}{2C_2^2 R^{2p}} \right), \quad (388)$$

we have that with the probability of at least $1 - \delta$,

$$\frac{1}{m} \left(\|s(w(0), v(0))\|_2^2 + \|h(w(0), v(0))\|_2^2 + \sum_{\ell=1}^L \|\mu_\ell(w(0), v(0))\|_2^2 \right) \quad (389)$$

$$\leq \frac{1}{m} \left(\left(4 + \binom{d+k}{d} \binom{d+|\xi|}{d} \right) \varepsilon^2 + 2C_3 \right) \quad (390)$$

$$\leq \frac{\lambda_0^2}{C^2} R^{2p}. \quad (391)$$

In other words, if

$$m \geq \frac{2C^2}{\lambda_0^2} \left(C_2 \left(4 + \binom{d+k}{d} \binom{d+|\xi|}{d} \right) R^{4p} \log \left(\frac{2N_2}{\delta} \right) + C_3 R^{2p} \right), \quad (392)$$

with the probability of at least $1 - \delta$, we have

$$\frac{1}{\sqrt{m}} \left\| \begin{bmatrix} s(w(0), v(0)) \\ \mu_1(w(0), v(0)) \\ \vdots \\ \mu_L(w(0), v(0)) \\ h(w(0), v(0)) \end{bmatrix} \right\|_2 < \frac{\lambda_0}{2C} R^{-p}. \quad (393)$$

□

Proposition C.7. Suppose the conditions (281) holds and R_w are given as in Lemma C.5. For

$$C_4 = 6\sqrt{1 + a_{\max}^2} \sqrt{\nu_{\max}} \binom{d+k}{d} \binom{d+|\xi|}{d} p^{|\xi|+1} 2^p \frac{1}{R_w}, \quad (394)$$

if m is large enough so that

$$\frac{1}{\sqrt{m}} \left\| \begin{bmatrix} s(w(0), v(0)) \\ \mu_1(w(0), v(0)) \\ \vdots \\ \mu_L(w(0), v(0)) \\ h(w(0), v(0)) \end{bmatrix} \right\|_2 < \frac{\lambda_0}{C_4} R^{-p}, \quad (395)$$

then

$$\|(w_\ell)_r(t) - (w_\ell)_r(0)\|_2 < R_w, \quad (396)$$

$$\|(v_\ell)_r(t) - (v_\ell)_r(0)\|_2 < R_w, \quad (397)$$

for all $t > 0$, $r \in [m]$ and $\ell = 0, \dots, L$.

Proof. We begin the proof by defining

$$T = \sup \{ t \geq 0 : \|(w_\ell)_r(t) - (w_\ell)_r(0)\|_2 < R_w \text{ and } \|(v_\ell)_r(t) - (v_\ell)_r(0)\|_2 < R_w, \forall \ell = 0, \dots, L \}. \quad (398)$$

Since $(w_\ell)_r$ and (v_ℓ) are continuous, the above set in supremum is nonempty. For $t \in (0, T)$,

$$\frac{d}{dt} \left\| \begin{bmatrix} s(w(t), v(t)) \\ \mu_1(w(t), v(t)) \\ \vdots \\ \mu_L(w(t), v(t)) \\ h(w(t), v(t)) \end{bmatrix} \right\|_2^2 \quad (399)$$

$$= -2 \begin{bmatrix} s(w(t), v(t)) \\ \mu_1(w(t), v(t)) \\ \vdots \\ \mu_L(w(t), v(t)) \\ h(w(t), v(t)) \end{bmatrix}^\top (\hat{G}_w + \hat{G}_v) \begin{bmatrix} s(w(t), v(t)) \\ \mu_1(w(t), v(t)) \\ \vdots \\ \mu_L(w(t), v(t)) \\ h(w(t), v(t)) \end{bmatrix} \quad (400)$$

$$\leq -2 \cdot \frac{1}{2} \lambda_0 \left\| \begin{bmatrix} s(\mathbf{w}(t), \mathbf{v}(t)) \\ \boldsymbol{\mu}_1(\mathbf{w}(t), \mathbf{v}(t)) \\ \vdots \\ \boldsymbol{\mu}_L(\mathbf{w}(t), \mathbf{v}(t)) \\ \mathbf{h}(\mathbf{w}(t), \mathbf{v}(t)) \end{bmatrix} \right\|_2^2, \quad (401)$$

where the last inequality comes from Lemma C.5 and the positive semi-definiteness of $\widehat{\mathbf{G}}_{\mathbf{w}}$. This implies that

$$\left\| \begin{bmatrix} s(\mathbf{w}(t), \mathbf{v}(t)) \\ \boldsymbol{\mu}_1(\mathbf{w}(t), \mathbf{v}(t)) \\ \vdots \\ \boldsymbol{\mu}_L(\mathbf{w}(t), \mathbf{v}(t)) \\ \mathbf{h}(\mathbf{w}(t), \mathbf{v}(t)) \end{bmatrix} \right\|_2 \leq \exp\left(-\frac{\lambda_0}{2}t\right) \left\| \begin{bmatrix} s(\mathbf{w}(0), \mathbf{v}(0)) \\ \boldsymbol{\mu}_1(\mathbf{w}(0), \mathbf{v}(0)) \\ \vdots \\ \boldsymbol{\mu}_L(\mathbf{w}(0), \mathbf{v}(0)) \\ \mathbf{h}(\mathbf{w}(0), \mathbf{v}(0)) \end{bmatrix} \right\|_2. \quad (402)$$

For $t \in (0, T)$ and $\ell \in \{0, \dots, L\}$, from $\|(\mathbf{w}_\ell)_r(0)\|_2 \leq R$ and $\|(\mathbf{v}_\ell)_r(0)\|_2 \leq |I_{\xi_\ell}|$ and (398), we have

$$\|(\mathbf{w}_\ell)_r(t)\|_2 \leq \|(\mathbf{w}_\ell)_r(t) - (\mathbf{w}_\ell)_r(0)\|_2 + \|(\mathbf{w}_\ell)_r(0)\|_2 \quad (403)$$

$$< R_w + R \quad (404)$$

$$\leq 2R, \quad (405)$$

$$\|(\mathbf{v}_\ell)_r(t)\|_2 \leq \|(\mathbf{v}_\ell)_r(t) - (\mathbf{v}_\ell)_r(0)\|_2 + \|(\mathbf{v}_\ell)_r(0)\|_2 \quad (406)$$

$$< R_w + |I_{\xi_\ell}| \quad (407)$$

$$\leq R. \quad (408)$$

From (44), (57), (52), and (48), we can attain

$$\left\| \frac{\partial s_i(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{w}_\ell)_r} \right\|_2 \leq \sqrt{\frac{1}{mn_o}} a_{\max} \sqrt{|I_{\xi_\ell}|} |J_{\Delta \xi_{\ell+1}}| p^{\Delta \xi_{\ell+1}+1} 2^{p-1} R^p, \quad (409)$$

$$\left\| \frac{\partial h_j(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{w}_0)_r} \right\|_2 \leq \sqrt{\frac{\nu_{\max}}{mn_b}} a_{\max} (d+1) p^2 2^{p-1} R^p, \quad (410)$$

$$\left\| \frac{\partial (\boldsymbol{\mu}_\ell)_{(\alpha, \beta), i}(\mathbf{w}, \mathbf{v})}{\partial (\mathbf{w}_\ell)_r} \right\|_2 \leq \sqrt{\frac{\nu_{\max}}{mn_o}} p 2^{p-1} R^p, \quad (411)$$

$$\left\| \frac{\partial (\boldsymbol{\mu}_\ell)_{(\alpha, \beta), i}(\mathbf{w}, \mathbf{v})}{\partial (\mathbf{w}_{\ell-1})_r} \right\|_2 \leq \sqrt{\frac{\nu_{\max}}{mn_o}} p^{|\xi|+1} 2^{p-1} R^p, \quad (412)$$

for all $t \in (0, T)$.

Therefore, we can induce that

$$\left\| \frac{d}{dt} (\mathbf{w}_\ell)_r(t) \right\|_2 \quad (413)$$

$$= \left\| \sum_{i=1}^{n_o} s_i(\mathbf{w}(t), \mathbf{v}(t)) \cdot \frac{\partial s_i(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{w}_\ell)_r} + \sum_{j=1}^{n_b} h_j(\mathbf{w}(t), \mathbf{v}(t)) \cdot \frac{\partial h_j(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{w}_\ell)_r} \right\|_2 \quad (414)$$

$$+ \sum_{i=1}^{n_o} \sum_{\ell'=1}^L \sum_{\alpha \in I_{\xi_{\ell'-1}}} \sum_{\beta \in I_{\Delta \xi_{\ell'}}} \left\| \frac{\partial (\boldsymbol{\mu}_{\ell'})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{w}_\ell)_r} \cdot (\boldsymbol{\mu}_{\ell'})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t)) \right\|_2 \quad (415)$$

$$\leq \sum_{i=1}^{n_o} \left\| \frac{\partial s_i(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{w}_\ell)_r} \right\|_2 \cdot |s_i(\mathbf{w}(t), \mathbf{v}(t))| \quad (416)$$

$$+ \mathbf{1}_{\{\ell=0\}} \cdot \sum_{j=1}^{n_b} \left\| \frac{\partial h_j(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{w}_\ell)_r} \right\|_2 \cdot |h_j(\mathbf{w}(t), \mathbf{v}(t))| \quad (417)$$

$$+ \left(\mathbf{1}_{\{\ell \neq 0\}} \cdot \sum_{i=1}^{n_o} \sum_{\alpha \in I_{\xi_{\ell-1}}} \sum_{\beta \in I_{\Delta \xi_{\ell}}} \left\| \frac{\partial (\boldsymbol{\mu}_{\ell})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{w}_{\ell})_r} \right\|_2 \cdot |(\boldsymbol{\mu}_{\ell})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))| \right. \quad (418)$$

$$\left. + \mathbf{1}_{\{\ell \neq L\}} \cdot \sum_{i=1}^{n_o} \sum_{\alpha \in I_{\xi_{\ell}}} \sum_{\beta \in I_{\Delta \xi_{\ell+1}}} \left\| \frac{\partial (\boldsymbol{\mu}_{\ell+1})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{w}_{\ell})_r} \right\|_2 \cdot |(\boldsymbol{\mu}_{\ell+1})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))| \right). \quad (419)$$

The first term and second term are bounded by

$$\sum_{i=1}^{n_o} \left\| \frac{\partial s_i(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{w}_{\ell})_r} \right\|_2 \cdot |s_i(\mathbf{w}(t), \mathbf{v}(t))| \quad (420)$$

$$\leq \sum_{i=1}^{n_o} \sqrt{\frac{\nu_{\max}}{mn_o}} a_{\max} \sqrt{|I_{\xi_{\ell}}|} |J_{\Delta \xi_{\ell+1}}| p^{|\xi|+1} 2^{p-1} R^p \cdot |s_i(\mathbf{w}(t), \mathbf{v}(t))| \quad (421)$$

$$= \sqrt{\frac{\nu_{\max}}{m}} a_{\max} \sqrt{|I_{\xi_{\ell}}|} |J_{\Delta \xi_{\ell+1}}| p^{|\xi|+1} 2^{p-1} R^p \cdot \frac{1}{\sqrt{n_o}} \sum_{i=1}^{n_o} |s_i(\mathbf{w}(t), \mathbf{v}(t))| \quad (422)$$

$$\leq \sqrt{\frac{\nu_{\max}}{m}} a_{\max} \sqrt{|I_{\xi_{\ell}}|} |J_{\Delta \xi_{\ell+1}}| p^{|\xi|+1} 2^{p-1} R^p \cdot \|\mathbf{s}(\mathbf{w}(t), \mathbf{v}(t))\|_2, \quad (423)$$

and

$$\sum_{j=1}^{n_b} \left\| \frac{\partial h_j(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{w}_{\ell})_r} \right\|_2 \cdot |h_j(\mathbf{w}(t), \mathbf{v}(t))| \quad (424)$$

$$\leq \sqrt{\frac{\nu_{\max}}{m}} a_{\max} (d+1) p^2 2^{p-1} R^p \cdot \frac{1}{\sqrt{n_b}} \sum_{j=1}^{n_b} |h_j(\mathbf{w}(t), \mathbf{v}(t))| \quad (425)$$

$$\leq \sqrt{\frac{\nu_{\max}}{m}} a_{\max} (d+1) p^2 2^{p-1} R^p \cdot \|\mathbf{h}(\mathbf{w}(t), \mathbf{v}(t))\|_2. \quad (426)$$

The third term is bounded by

$$\sum_{i=1}^{n_o} \sum_{\alpha \in I_{\xi_{\ell-1}}} \sum_{\beta \in I_{\Delta \xi_{\ell}}} \left\| \frac{\partial (\boldsymbol{\mu}_{\ell})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{w}_{\ell})_r} \right\|_2 \cdot |(\boldsymbol{\mu}_{\ell})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))| \quad (427)$$

$$\leq \sqrt{\frac{\nu_{\max}}{mn_o}} p 2^{p-1} R^p \cdot \sum_{i=1}^{n_o} \sum_{\alpha \in I_{\xi_{\ell-1}}} \sum_{\beta \in I_{\Delta \xi_{\ell}}} |(\boldsymbol{\mu}_{\ell})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))| \quad (428)$$

$$\leq \sqrt{\frac{\nu_{\max}}{m}} p 2^{p-1} R^p \sqrt{|I_{\xi_{\ell-1}}| \cdot |I_{\Delta \xi_{\ell}}|} \cdot \|\boldsymbol{\mu}_{\ell}\|_2, \quad (429)$$

and the fourth term is

$$\sum_{i=1}^{n_o} \sum_{\alpha \in I_{\xi_{\ell}}} \sum_{\beta \in I_{\Delta \xi_{\ell+1}}} \left\| \frac{\partial (\boldsymbol{\mu}_{\ell+1})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{w}_{\ell})_r} \right\|_2 \cdot |(\boldsymbol{\mu}_{\ell+1})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))| \quad (430)$$

$$\leq \sqrt{\frac{\nu_{\max}}{mn_o}} p^{|\xi|+1} 2^{p-1} R^p \cdot \sum_{i=1}^{n_o} \sum_{\alpha \in I_{\xi_{\ell}}} \sum_{\beta \in I_{\Delta \xi_{\ell+1}}} |(\boldsymbol{\mu}_{\ell+1})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))| \quad (431)$$

$$\leq \sqrt{\frac{\nu_{\max}}{m}} p^{|\xi|+1} 2^{p-1} R^p \sqrt{|I_{\xi_{\ell}}| \cdot |I_{\Delta \xi_{\ell+1}}|} \cdot \|\boldsymbol{\mu}_{\ell+1}\|_2. \quad (432)$$

Therefore, for $t \in (0, T)$ and $\ell = \{0, \dots, L\}$,

$$\left\| \frac{d}{dt} (\mathbf{w}_{\ell})_r(t) \right\|_2 \quad (433)$$

$$\leq \sqrt{\frac{\nu_{\max}}{m}} a_{\max} \sqrt{\binom{d+k}{d} \binom{d+|\xi|}{d}} p^{|\xi|+1} 2^{p-1} R^p \|s(w(t), v(t))\|_2 \quad (434)$$

$$+ \mathbf{1}_{\{\ell=0\}} \cdot \frac{\sqrt{\nu_{\max}}}{\sqrt{m}} a_{\max} (d+1) p^2 2^{p-1} R^p \|h(w(t), v(t))\|_2 \quad (435)$$

$$+ \mathbf{1}_{\{\ell \neq 0\}} \cdot \sqrt{\frac{\nu_{\max}}{m}} \sqrt{\binom{d+k}{d} \binom{d+|\xi|}{d}} p 2^{p-1} R^p \|\mu_\ell(w(t), v(t))\|_2 \quad (436)$$

$$+ \mathbf{1}_{\{\ell \neq L\}} \cdot \sqrt{\frac{\nu_{\max}}{m}} \sqrt{\binom{d+k}{d} \binom{d+|\xi|}{d}} p^{|\xi|+1} 2^{p-1} R^p \|\mu_{\ell+1}(w(t), v(t))\|_2 \quad (437)$$

$$\leq 3 \sqrt{\frac{\nu_{\max}}{m}} \sqrt{1 + a_{\max}^2} \sqrt{\binom{d+k}{d} \binom{d+|\xi|}{d}} p^{|\xi|+1} 2^{p-1} R^p \left\| \begin{bmatrix} s(w(t), v(t)) \\ \mu_1(w(t), v(t)) \\ \vdots \\ \mu_L(w(t), v(t)) \\ h(w(t), v(t)) \end{bmatrix} \right\|_2 \quad (438)$$

$$\leq 3 \sqrt{1 + a_{\max}^2} \sqrt{\frac{\nu_{\max}}{m}} \sqrt{\binom{d+k}{d} \binom{d+|\xi|}{d}} p^{|\xi|+1} 2^{p-1} R^p \exp\left(-\frac{\lambda_0}{2} t\right) \left\| \begin{bmatrix} s(w(0), v(0)) \\ \mu_1(w(0), v(0)) \\ \vdots \\ \mu_L(w(0), v(0)) \\ h(w(0), v(0)) \end{bmatrix} \right\|_2. \quad (439)$$

On the other hand,

$$\left\| \int_0^T \frac{d}{ds} (w_\ell)_r(s) ds \right\|_2^2 = \sum_{q=1}^d \left(\int_0^T \frac{d}{ds} (w_\ell)_{r,q}(s) ds \right)^2 \quad (440)$$

$$= \int_0^T \sum_{q=1}^d \left(\int_0^T \frac{d}{ds} (w_\ell)_{r,q}(s) ds \right) \cdot \frac{d}{ds} (w_\ell)_{r,q}(s) ds \quad (441)$$

$$\leq \int_0^T \left\| \int_0^T \frac{d}{ds} (w_\ell)_r(s) ds \right\|_2 \cdot \left\| \frac{d}{ds} (w_\ell)(s) \right\|_2 ds \quad (442)$$

$$= \left\| \int_0^T \frac{d}{ds} (w_\ell)_r(s) ds \right\|_2 \cdot \int_0^T \left\| \frac{d}{ds} (w_\ell)(s) \right\|_2 ds, \quad (443)$$

and hence

$$\|(w_\ell)_r(T) - (w_\ell)_r(0)\|_2 \quad (444)$$

$$= \left\| \int_0^T \frac{d}{ds} (w_\ell)_r(s) ds \right\|_2 \quad (445)$$

$$\leq \int_0^T \left\| \frac{d}{ds} (w_\ell)_r(s) \right\|_2 ds \quad (446)$$

$$\leq 3 \sqrt{1 + a_{\max}^2} \sqrt{\frac{\nu_{\max}}{m}} \sqrt{\binom{d+k}{d} \binom{d+|\xi|}{d}} p^{|\xi|+1} 2^{p-1} R^p \int_0^T \exp\left(-\frac{\lambda_0}{2} s\right) ds \left\| \begin{bmatrix} s(w(0), v(0)) \\ \mu_1(w(0), v(0)) \\ \vdots \\ \mu_L(w(0), v(0)) \\ h(w(0), v(0)) \end{bmatrix} \right\|_2 \quad (447)$$

$$\leq \frac{6}{\lambda_0} \sqrt{1 + a_{\max}^2} \sqrt{\frac{\nu_{\max}}{m}} \sqrt{\binom{d+k}{d}} \binom{d+|\xi|}{d} p^{|\xi|+1} 2^{p-1} R^p \left\| \begin{bmatrix} s(\mathbf{w}(0), \mathbf{v}(0)) \\ \boldsymbol{\mu}_1(\mathbf{w}(0), \mathbf{v}(0)) \\ \vdots \\ \boldsymbol{\mu}_L(\mathbf{w}(0), \mathbf{v}(0)) \\ \mathbf{h}(\mathbf{w}(0), \mathbf{v}(0)) \end{bmatrix} \right\|_2 \quad (448)$$

If m is large enough to satisfy (395), we have

$$\|(\mathbf{w}_\ell)_r(T) - (\mathbf{w}_\ell)_r(0)\|_2 \leq \frac{6\sqrt{1 + a_{\max}^2} \sqrt{\nu_{\max}} \sqrt{\binom{d+k}{d}} \binom{d+|\xi|}{d} p^{|\xi|+1} 2^{p-1}}{\lambda_0} \cdot \frac{\lambda_0}{C_4} < \frac{1}{2} R_w < R_w. \quad (449)$$

Similarly, we obtain

$$\left\| \frac{d}{dt} (\mathbf{v}_\ell)_r(t) \right\|_2 \quad (450)$$

$$= \left\| \sum_{i=1}^{n_o} s_i(\mathbf{w}(t), \mathbf{v}(t)) \cdot \frac{\partial s_i(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{v}_\ell)_r} + \sum_{j=1}^{n_b} h_j(\mathbf{w}(t), \mathbf{v}(t)) \cdot \frac{\partial h_j(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{v}_\ell)_r} \right. \quad (451)$$

$$\left. + \sum_{i=1}^{n_o} \sum_{\ell'=1}^L \sum_{\alpha \in I_{\xi_{\ell'-1}}} \sum_{\beta \in I_{\Delta \xi'_\ell}} \frac{\partial (\boldsymbol{\mu}_{\ell'})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{v}_\ell)_r} \cdot (\boldsymbol{\mu}_{\ell'})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t)) \right\|_2 \quad (452)$$

$$\leq \sum_{i=1}^{n_o} |s_i(\mathbf{w}(t), \mathbf{v}(t))| \cdot \left\| \frac{\partial s_i(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{v}_\ell)_r} \right\|_2 \quad (453)$$

$$+ \mathbf{1}_{\{\ell=0\}} \cdot \sum_{j=1}^{n_b} |h_j(\mathbf{w}(t), \mathbf{v}(t))| \cdot \left\| \frac{\partial h_j(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{v}_\ell)_r} \right\|_2 \quad (454)$$

$$+ \mathbf{1}_{\{\ell \neq 0\}} \cdot \sum_{i=1}^{n_o} \sum_{\alpha \in I_{\xi_{\ell-1}}} \sum_{\beta \in I_{\Delta \xi_\ell}} \left\| \frac{\partial (\boldsymbol{\mu}_\ell)_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{v}_\ell)_r} \right\|_2 \cdot |(\boldsymbol{\mu}_\ell)_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))| \quad (455)$$

$$+ \mathbf{1}_{\{\ell \neq L\}} \cdot \sum_{i=1}^{n_o} \sum_{\alpha \in I_{\xi_\ell}} \sum_{\beta \in I_{\Delta \xi_{\ell+1}}} \left\| \frac{\partial (\boldsymbol{\mu}_{\ell+1})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{v}_\ell)_r} \right\|_2 \cdot |(\boldsymbol{\mu}_{\ell+1})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))|. \quad (456)$$

The first term and second term are bounded by

$$\sum_{i=1}^{n_o} |s_i(\mathbf{w}(t), \mathbf{v}(t))| \cdot \left\| \frac{\partial s_i(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{v}_\ell)_r} \right\|_2 \quad (457)$$

$$\leq \sqrt{\frac{\nu_{\max}}{m}} a_{\max} \sqrt{\binom{d+k}{d}} \binom{d+|\xi|}{d} p^{|\xi|} (2R)^p \frac{1}{\sqrt{n_o}} \sum_{i=1}^{n_o} |s_i(\mathbf{w}(t), \mathbf{v}(t))| \quad (458)$$

$$\leq \sqrt{\frac{\nu_{\max}}{m}} a_{\max} \sqrt{\binom{d+k}{d}} \binom{d+|\xi|}{d} p^{|\xi|} (2R)^p \|\hat{\mathbf{s}}\|_2, \quad (459)$$

and

$$\sum_{j=1}^{n_b} |h_j(\mathbf{w}(t), \mathbf{v}(t))| \cdot \left\| \frac{\partial h_j(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{v}_\ell)_r} \right\|_2 \quad (460)$$

$$\leq \sqrt{\frac{\nu_{\max}}{m}} (d+1) a_{\max} p (2R)^p \frac{1}{\sqrt{n_b}} \sum_{j=1}^{n_b} |h_j(\mathbf{w}(t), \mathbf{v}(t))| \quad (461)$$

$$\leq \sqrt{\frac{\nu_{\max}}{m}} (d+1) a_{\max} p (2R)^p \|\mathbf{h}\|_2. \quad (462)$$

The third term is bounded by

$$\sum_{i=1}^{n_o} \sum_{\alpha \in I_{\xi_{\ell-1}}} \sum_{\beta \in I_{\Delta \xi_{\ell}}} \left\| \frac{\partial (\boldsymbol{\mu}_{\ell})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{v}_{\ell})_r} \right\|_2 \cdot |(\boldsymbol{\mu}_{\ell})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))| \quad (463)$$

$$\leq \sqrt{\frac{\nu_{\max}}{mn_o}} \sqrt{\binom{d+k}{d} \binom{d+|\xi|}{d}} (2R)^p \cdot \sum_{i=1}^{n_o} \sum_{\alpha \in I_{\xi_{\ell-1}}} \sum_{\beta \in I_{\Delta \xi_{\ell}}} |(\boldsymbol{\mu}_{\ell})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))| \quad (464)$$

$$\leq \sqrt{\frac{\nu_{\max}}{m}} \sqrt{\binom{d+k}{d} \binom{d+|\xi|}{d}} (2R)^p \sqrt{|I_{\xi_{\ell-1}}| \cdot |I_{\Delta \xi_{\ell}}|} \cdot \|\boldsymbol{\mu}_{\ell}\|_2 \quad (465)$$

$$\leq \sqrt{\frac{\nu_{\max}}{m}} \binom{d+k}{d} \binom{d+|\xi|}{d} (2R)^p \|\boldsymbol{\mu}_{\ell}\|_2, \quad (466)$$

and the fourth term is

$$\sum_{i=1}^{n_o} \sum_{\alpha \in I_{\xi_{\ell}}} \sum_{\beta \in I_{\Delta \xi_{\ell+1}}} \left\| \frac{\partial (\boldsymbol{\mu}_{\ell+1})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))}{\partial (\mathbf{v}_{\ell})_r} \right\|_2 \cdot |(\boldsymbol{\mu}_{\ell+1})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))| \quad (467)$$

$$\leq \sqrt{\frac{\nu_{\max}}{mn_o}} p^{|\xi|} \sqrt{\binom{d+k}{d} \binom{d+|\xi|}{d}} (2R)^p \cdot \sum_{\alpha \in I_{\xi_{\ell}}} \sum_{\beta \in I_{\Delta \xi_{\ell+1}}} |(\boldsymbol{\mu}_{\ell+1})_{(\alpha, \beta), i}(\mathbf{w}(t), \mathbf{v}(t))| \quad (468)$$

$$\leq \sqrt{\frac{\nu_{\max}}{m}} \binom{d+k}{d} \binom{d+|\xi|}{d} p^{|\xi|} (2R)^p \|\boldsymbol{\mu}_{\ell+1}\|_2. \quad (469)$$

Therefore, for $t \in (0, T)$ and $\ell = \{0, \dots, L\}$,

$$\left\| \frac{d}{dt} (\mathbf{v}_{\ell})_r(t) \right\|_2 \quad (470)$$

$$\leq \sqrt{\frac{\nu_{\max}}{m}} a_{\max} \sqrt{\binom{d+k}{d} \binom{d+|\xi|}{d}} p^{|\xi|} (2R)^p \|\mathbf{s}\|_2 \quad (471)$$

$$+ \mathbf{1}_{\{\ell=0\}} \cdot \sqrt{\frac{\nu_{\max}}{m}} (d+1) a_{\max} p (2R)^p \|\mathbf{h}\|_2 \quad (472)$$

$$+ \mathbf{1}_{\{\ell \neq 0\}} \cdot \sqrt{\frac{\nu_{\max}}{m}} \binom{d+k}{d} \binom{d+|\xi|}{d} (2R)^p \|\boldsymbol{\mu}_{\ell}\|_2 \quad (473)$$

$$+ \mathbf{1}_{\{\ell \neq L\}} \cdot \sqrt{\frac{\nu_{\max}}{m}} \binom{d+k}{d} \binom{d+|\xi|}{d} p^{|\xi|} (2R)^p \|\boldsymbol{\mu}_{\ell+1}\|_2 \quad (474)$$

$$\leq 3 \sqrt{\frac{\nu_{\max}}{m}} \sqrt{1 + a_{\max}^2} \binom{d+k}{d} \binom{d+|\xi|}{d} p^{|\xi|} (2R)^p \left\| \begin{bmatrix} \mathbf{s}(\mathbf{w}(t), \mathbf{v}(t)) \\ \boldsymbol{\mu}_1(\mathbf{w}(t), \mathbf{v}(t)) \\ \vdots \\ \boldsymbol{\mu}_L(\mathbf{w}(t), \mathbf{v}(t)) \\ \mathbf{h}(\mathbf{w}(t), \mathbf{v}(t)) \end{bmatrix} \right\|_2 \quad (475)$$

$$\leq 3 \sqrt{\frac{\nu_{\max}}{m}} \sqrt{1 + a_{\max}^2} \binom{d+k}{d} \binom{d+|\xi|}{d} p^{|\xi|} (2R)^p \exp\left(-\frac{\lambda_0}{2} t\right) \left\| \begin{bmatrix} \mathbf{s}(\mathbf{w}(0), \mathbf{v}(0)) \\ \boldsymbol{\mu}_1(\mathbf{w}(0), \mathbf{v}(0)) \\ \vdots \\ \boldsymbol{\mu}_L(\mathbf{w}(0), \mathbf{v}(0)) \\ \mathbf{h}(\mathbf{w}(0), \mathbf{v}(0)) \end{bmatrix} \right\|_2 \quad (476)$$

As a consequence, we attain

$$\|(\mathbf{v}_{\ell})_r(T) - (\mathbf{v}_{\ell})_r(0)\|_2 \quad (477)$$

$$\leq \int_0^T \left\| \frac{d}{ds} (\mathbf{v}_\ell)_r(s) \right\|_2 ds \quad (478)$$

$$\leq 3 \sqrt{\frac{\nu_{\max}}{m}} \sqrt{1 + a_{\max}^2} \binom{d+k}{d} \binom{d+|\xi|}{d} p^{|\xi|} (2R)^p \int_0^T \exp\left(-\frac{\lambda_0}{2}s\right) ds \left\| \begin{bmatrix} \mathbf{s}(\mathbf{w}(0), \mathbf{v}(0)) \\ \boldsymbol{\mu}_1(\mathbf{w}(0), \mathbf{v}(0)) \\ \vdots \\ \boldsymbol{\mu}_L(\mathbf{w}(0), \mathbf{v}(0)) \\ \mathbf{h}(\mathbf{w}(0), \mathbf{v}(0)) \end{bmatrix} \right\|_2 \quad (479)$$

$$\leq \frac{6}{\lambda_0} \sqrt{1 + a_{\max}^2} \sqrt{\frac{\nu_{\max}}{m}} \binom{d+k}{d} \binom{d+|\xi|}{d} p^{|\xi|} (2R)^p \left\| \begin{bmatrix} \mathbf{s}(\mathbf{w}(0), \mathbf{v}(0)) \\ \boldsymbol{\mu}_1(\mathbf{w}(0), \mathbf{v}(0)) \\ \vdots \\ \boldsymbol{\mu}_L(\mathbf{w}(0), \mathbf{v}(0)) \\ \mathbf{h}(\mathbf{w}(0), \mathbf{v}(0)) \end{bmatrix} \right\|_2. \quad (480)$$

Since m is large enough to satisfy (395), we have

$$\|(\mathbf{v}_\ell)_r(T) - (\mathbf{v}_\ell)_r(0)\|_2 \leq \frac{6\sqrt{1 + a_{\max}^2} \sqrt{\nu_{\max}} \binom{d+k}{d} \binom{d+|\xi|}{d} p^{|\xi|} 2^p}{\lambda_0} \cdot \frac{\lambda_0}{C_4} = \frac{1}{p} R_w < R_w. \quad (481)$$

□

C.1 Proof of Theorem 4.3

Proof. For $R \in \mathbb{R}$, $r \in [m]$, and $q \in [d]$, each $(\mathbf{w}_\ell)_{r,q}$ satisfies

$$\mathbb{P} \left[|(\mathbf{w}_\ell)_{r,q}| > R \right] < \exp\left(-\frac{1}{2}R^2\right). \quad (482)$$

If we set $\delta_1 = (L+1)md \exp(-\frac{1}{2}R^2)$, then δ_1 satisfies

$$\left(1 - \exp\left(-\frac{1}{2}R^2\right)\right)^{(L+1)md} \geq 1 - (L+1)md \exp\left(-\frac{1}{2}R^2\right) = 1 - \delta_1. \quad (483)$$

Hence, with probability of at least $1 - \delta_1$, we have $|(\mathbf{w}_\ell)_{r,q}| \leq R$ for all $\ell = \{0, 1, \dots, L+1\}$, $r \in [m]$, and $q \in [d]$.

Provided $|(\mathbf{w}_\ell)_{r,q}| \leq R$ for all $\ell = \{0, 1, \dots, L+1\}$, if

$$m \geq \frac{32N_1C_1^2R^{4p}}{\lambda_0^2} \log\left(\frac{2N_1}{\delta_2}\right), \quad (484)$$

where

$$C_1 = \binom{d+k}{k} \binom{d+|\xi|}{|\xi|}^2 \nu_{\max}^2 \left(\frac{1}{n_o} + \frac{1}{n_b}\right) \sqrt{1 + a_{\max}^4} (p^{2|\xi|} + 1), \quad (485)$$

$$N_1 = 7 \binom{d+k}{d} \binom{d+|\xi|}{d}^2 n_o^2 + 4 \binom{d+|\xi|}{d} n_o n_b + n_b^2, \quad (486)$$

then by Proposition C.4, with probability of at least $1 - \delta_2$, we have

$$\left\| \widehat{\mathbf{G}}_{\mathbf{v}}(\mathbf{w}(0), \mathbf{v}(0)) - \widehat{\mathbf{G}}_{\mathbf{v}}^\infty \right\|_2 < \frac{\lambda_0}{4}. \quad (487)$$

On the other hand, with the condition $|(\mathbf{w}_\ell)_{r,q}| \leq R$ for all $\ell = \{0, 1, \dots, L+1\}$, $r \in [m]$, and $q \in [d]$, Proposition C.6 shows that if

$$m \geq \frac{2C_4^2}{\lambda_0^2} \left(C_2 \left(4 + \binom{d+k}{d} \binom{d+|\xi|}{d} \right) R^{4p} \log\left(\frac{2N_2}{\delta}\right) + C_3 R^{2p} \right), \quad (488)$$

$$\delta_3 \leq 2N_2 \exp\left(-\frac{C_3^2}{2C_2^2 R^{2p}}\right), \quad (489)$$

$$R_w \leq R, \quad (490)$$

where

$$R_w = \frac{\lambda_0}{8\sqrt{2}} \left(\nu_{\max} \sqrt{1 + a_{\max}^4} \binom{d+k}{d} \binom{d+|\xi|}{d}^2 p^{2|\xi|+1} (2R)^{2p-1} \right)^{-1}, \quad (491)$$

$$C_2 = \sqrt{\nu_{\max}} \sqrt{1 + a_{\max}^2} \binom{d+k}{d} \binom{d+|\xi|}{d} (p^{|\xi|} + 1), \quad (492)$$

$$C_3 = \frac{1}{n_o} \sum_{i=1}^{n_o} f(\mathbf{x}_i)^2 + \frac{1}{n_b} \sum_{j=1}^{n_b} g(\tilde{\mathbf{x}}_j)^2, \quad (493)$$

$$C_4 = \sqrt{\nu_{\max}} \sqrt{1 + a_{\max}^2} \binom{d+k}{d} \binom{d+|\xi|}{d} p^{|\xi|+1} 2^p \frac{1}{R_w}, \quad (494)$$

$$N_2 = \binom{d+k}{d} \binom{d+|\xi|}{d} n_o + n_b, \quad (495)$$

then, with the probability of at least $1 - \delta_3$, we have

$$\frac{1}{\sqrt{m}} \left\| \begin{bmatrix} \mathbf{s}(\mathbf{w}(0), \mathbf{v}(0)) \\ \boldsymbol{\mu}_1(\mathbf{w}(0), \mathbf{v}(0)) \\ \vdots \\ \boldsymbol{\mu}_L(\mathbf{w}(0), \mathbf{v}(0)) \\ \mathbf{h}(\mathbf{w}(0), \mathbf{v}(0)) \end{bmatrix} \right\|_2 < \frac{\lambda_0}{C_4} R^{-p}. \quad (496)$$

Consequently, Proposition C.7 implies

$$\|(\mathbf{w}_\ell)_r(t) - (\mathbf{w}_\ell)_r(0)\|_2 < R_w, \quad (497)$$

$$\|(\mathbf{v}_\ell)_r(t) - (\mathbf{v}_\ell)_r(0)\|_2 < R_w, \quad (498)$$

for all $t \geq 0$. Then, we have

$$\left\| \begin{bmatrix} \mathbf{s}(\mathbf{w}(t), \mathbf{v}(t)) \\ \boldsymbol{\mu}_1(\mathbf{w}(t), \mathbf{v}(t)) \\ \vdots \\ \boldsymbol{\mu}_L(\mathbf{w}(t), \mathbf{v}(t)) \\ \mathbf{h}(\mathbf{w}(t), \mathbf{v}(t)) \end{bmatrix} \right\|_2 \leq \exp\left(-\frac{\lambda_0}{2}t\right) \left\| \begin{bmatrix} \mathbf{s}(\mathbf{w}(0), \mathbf{v}(0)) \\ \boldsymbol{\mu}_1(\mathbf{w}(0), \mathbf{v}(0)) \\ \vdots \\ \boldsymbol{\mu}_L(\mathbf{w}(0), \mathbf{v}(0)) \\ \mathbf{h}(\mathbf{w}(0), \mathbf{v}(0)) \end{bmatrix} \right\|_2 \quad (499)$$

from Lemam C.5.

Set $\delta_1 = \delta_2 = \delta_3 = \frac{1}{3}\delta$ and consider all inequality conditions. First, for $\delta_1 = \frac{1}{3}\delta$,

$$R = \sqrt{\log \frac{6(L+1)md}{\delta}} \quad (500)$$

satisfies the condition (483). Then, (484) is satisfied if

$$m \geq \frac{32N_1 C_1^2}{\lambda_0^2} \left(\log \frac{6(L+1)md}{\delta} \right)^{2p} \log \left(\frac{6N_1}{\delta} \right). \quad (501)$$

Similarly, if m is large enough so that

$$m \geq \frac{2C_2 C_4^2}{\lambda_0^2} \left(4 + \binom{d+k}{d} \binom{d+|\xi|}{d} \right) \left(\log \frac{6(L+1)md}{\delta} \right)^{2p} \log \left(\frac{2N_2}{\delta} \right) \quad (502)$$

$$+ \frac{2C_3 C_4^2}{\lambda_0^2} \left(\log \frac{6(L+1)md}{\delta} \right)^p, \quad (503)$$

then (488) holds.

Therefore, there exists $C_5 = C_5(a_{\max}, \nu_{\max}, n_o, n_b, \lambda_0)$ such that if

$$m \geq C \binom{d+k}{d}^6 \binom{d+|\xi|}{d}^8 p^{7|\xi|+4} 2^{6p} \left(\log \frac{md}{\delta} \right)^{4p} \quad (504)$$

then with the probability of at least $(1 - \delta_1)(1 - \delta_2)(1 - \delta_3) \geq 1 - (\delta_1 + \delta_2 + \delta_3) = 1 - \delta$, we have

$$\left\| \begin{bmatrix} s(\mathbf{w}(t), \mathbf{v}(t)) \\ \mu_1(\mathbf{w}(t), \mathbf{v}(t)) \\ \vdots \\ \mu_L(\mathbf{w}(t), \mathbf{v}(t)) \\ \mathbf{h}(\mathbf{w}(t), \mathbf{v}(t)) \end{bmatrix} \right\|_2 \leq \exp\left(-\frac{\lambda_0}{2}t\right) \left\| \begin{bmatrix} s(\mathbf{w}(0), \mathbf{v}(0)) \\ \mu_1(\mathbf{w}(0), \mathbf{v}(0)) \\ \vdots \\ \mu_L(\mathbf{w}(0), \mathbf{v}(0)) \\ \mathbf{h}(\mathbf{w}(0), \mathbf{v}(0)) \end{bmatrix} \right\|_2. \quad (505)$$

□

D Experimental Details

D.1 Problem Formulations

Poisson equation

$$\begin{cases} u_{xx} + u_{yy} = f_1 & x, y \in (0, \pi) \\ u(x, 0) = u(x, \pi) = u(0, y) = u(\pi, y) = 0 & x, y \in [0, \pi], \end{cases} \quad (506)$$

where $f_1 : (0, \pi)^2 \rightarrow \mathbb{R}$ is defined so that the exact solution is given by

$$u(x, y) = \frac{1}{5} \sin^2(x) y^2 (\pi - y)^2. \quad (507)$$

Bi-harmonic equation

$$\begin{cases} u_{xxxx} + 2u_{xxyy} + u_{yyyy} = f_2 & x, y \in (0, \pi) \\ u(x, 0) = u(x, \pi) = u(0, y) = u(\pi, y) = 0 & x, y \in [0, \pi] \\ \frac{\partial}{\partial \mathbf{n}} u(x, 0) = \frac{\partial}{\partial \mathbf{n}} u(x, \pi) = \frac{\partial}{\partial \mathbf{n}} u(0, y) = \frac{\partial}{\partial \mathbf{n}} u(\pi, y) = 0 & x, y \in (0, \pi), \end{cases} \quad (508)$$

where the flux function $f_2 : (0, \pi)^2 \rightarrow \mathbb{R}$ is set to ensure that it has a solution that is identical to that of Poisson equation (507).

Heat equation

$$\begin{cases} u_t = u_{xx} & x \in (-1, 1), t \in (0, 1) \\ u(t, -1) = u(t, 1) = 0 & t \in [0, 1] \\ u(0, x) = \sin(\pi x) & x \in (-1, 1), \end{cases} \quad (509)$$

Elastic beam equation

$$\begin{cases} u_t + u_{xxxx} = 0 & x \in (0, \pi), t \in (0, 1) \\ u(t, 0) = u(t, \pi) = u_{xx}(t, 0) = u_{xx}(t, \pi) = 0 & t \in [0, 1] \\ u(0, x) = 2 \sin(x) & x \in (0, \pi). \end{cases} \quad (510)$$

The exact solution is given by $u(t, x) = 2e^{-t} \sin(x)$.

Convection-diffusion equation

$$\begin{cases} u_t + u_x - \frac{1}{4}u_{xx} = 0 & (t, x) \in \Omega = (0, 1) \times (0, \pi), \\ u(0, x) = \sin(x) & x \in [0, \pi], \\ u(t, 0) = -e^{-\frac{1}{4}t} \sin(t) & t \in [0, 1], \\ u(t, \pi) = e^{-\frac{1}{4}t} \sin(\pi - t) & t \in [0, 1], \end{cases} \quad (511)$$

whose exact solution is $u(t, x) = e^{-\frac{1}{4}t} \sin(x - t)$.

Table 2: Experimental settings for each PDE

PDE	width m	power p	optimizer(lr)	n_o	n_b
Harmonic	$10^2 \sim 10^6$	$5 \sim 8$	GD(10^{-8})	400	400
Bi-harmonic	$10^2 \sim 10^6$	$5 \sim 8$	GD(10^{-8})	400	400
Heat	10^3	$3 \sim 10$	Adam(10^{-3})	300	300
Heat(PINNs)	10^3	3	GD(10^{-1})	300	300
Heat(VS-PINNs)	10^3	2	GD(10^{-1})	300	300
Beam (PINNs)	10^3	5	Adam(10^{-3})	10,000	300
Beam (VS-PINNs $ \xi = 2$)	10^3	3	Adam(10^{-3})	10,000	300
Beam (VS-PINNs $ \xi = 1$)	10^3	2	Adam(10^{-3})	10,000	300
Convection-diffusion (PINNs)	10^3	3	Adam(10^{-3}) / GD(10^{-2})	300	300
Convection-diffusion (VS-PINNs)	10^3	2	Adam(10^{-3}) / GD(10^{-1})	300	300
Bi-harmonic (PINNs)	10^3	5	Adam(10^{-3})	10,000	400
Bi-harmonic (VS-PINNs $ \xi = 2$)	10^3	3	Adam(10^{-3})	10,000	400
Bi-harmonic (VS-PINNs $ \xi = 1$)	10^3	2	Adam(10^{-3})	10,000	400

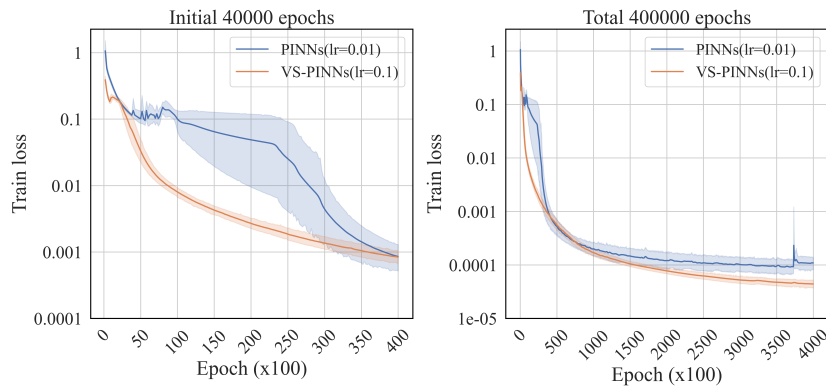


Figure 3: Loss of convection-diffusion equation trained by GD.

D.2 Parameter settings for experiments

Convection-diffusion equation We conducted experiments on a convection-diffusion equation: We train 100,000 epochs of PINNs with $p = 3$ and VS-PINNs with $p = 2$, using the same settings as represented for the heat equation (509). Figures 3 and 4 show that VS-PINNs reach lower loss and achieve more stable convergence for both GD and Adam.

Bi-harmonic equation Consider the bi-harmonic equation

$$\begin{cases} \Delta^2 u(x, y) = f(x, y) & x, y \in [0, \pi] \\ u(x, 0) = u(x, \pi) = u(0, y) = u(\pi, y) = 0 & x, y \in [0, \pi] \\ \frac{\partial}{\partial \mathbf{n}} u(x, 0) = \frac{\partial}{\partial \mathbf{n}} u(x, \pi) = \frac{\partial}{\partial \mathbf{n}} u(0, y) = \frac{\partial}{\partial \mathbf{n}} u(\pi, y) = 0 & x, y \in [0, \pi], \end{cases} \quad (512)$$

where $f(x, y)$ are defined so that the exact solution is given by $u(x, y) = \frac{1}{5} \sin^2(x) y^2 (\pi - y)^2$. We set $m = 1,000$, $n_o = 10,000$, $n_b = 400$, and the training collocation points are fixed once they are randomly selected. We experiment VS-PINNs with two cases: (i) $\phi_0 \approx u$, $\phi_1 \approx (u_{xx}, u_{yy})$ with $|\xi| = 2$ and $p = 3$ and (ii) the finest splitting of $\phi_0 \approx u$, $\phi_1 \approx \nabla u$, $\phi_2 \approx (u_{xx}, u_{yy})$, $\phi_3 \approx (u_{xxx}, u_{xxy}, u_{yyy})$ with $|\xi| = 1$ and $p = 2$. Regularization parameters are $\nu_1 = \nu_2 = \nu_3 = 1$ for derivative matching loss μ and $\nu = 10$ for boundary loss h . Figure 5 depicts the training loss of PINN and two VS-PINNs.

E Additional Experimental Results

Computational efficiency of VS-PINNs Table 3 measures the GPU memory, running time, and the number of model parameters corresponding to experiments on elastic beam and bi-harmonic equations. Because VS-PINNs need as many networks as auxiliary variables, finer VS-PINN requires

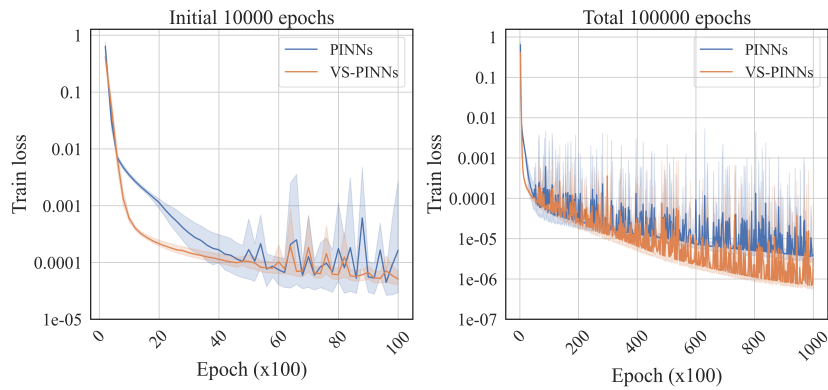


Figure 4: Loss of solving convection-diffusion equation trained by Adam.

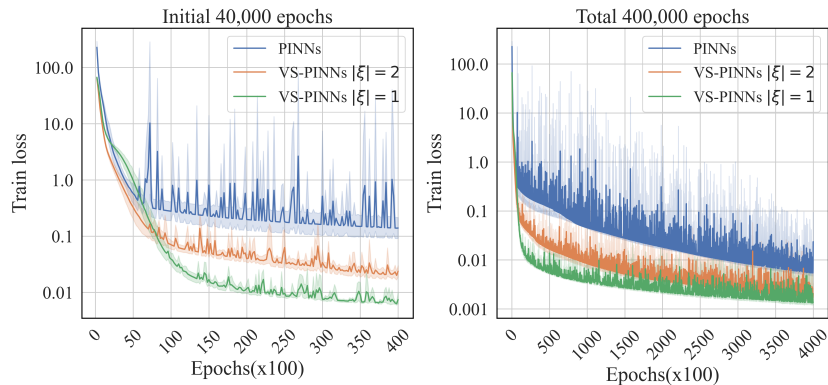


Figure 5: Loss comparison between PINNs with two VS-PINNs with $|\xi| = 1$ and 2.

more parameters to be trained. However, reducing the differentiation order in the Loss significantly reduces memory consumption, despite the additional loss terms.

Error with true solution As the ultimate objective of PINNs is to approximate the solution to the PDE, we also quantified the discrepancy between the exact solution of the predicted solutions of the three models we tested. We tested five times for each PDE and splitting setup from different random seeds, and the averages of the Mean Squared Error (MSE) with the exact solution are reported in Table 4. We can observe that the results do not exhibit the same tendency as depicted in Figure 2. Since there is a discrepancy between optimization ability and the generalization error, it is not within the scope of this paper, and we believe that this is an important area for future research.

GD results for fourth-order PDEs In order to address fourth-order PDEs (elastic beam and bi-harmonic equations), PINNs should employ $p = 5$ of the power of the ReLU^p activation. When optimized with GD, this configuration results in a loss value of NaN, indicating that PINNs fail to train. Consequently, we present the results of our experiments with the Adam optimizer in Section 5. However, in contrast to PINNs, which exhibit difficulty in training, variable splitting demonstrates superior performance in terms of loss reduction for GD, as illustrated in Figure 6. This is attributed to the fact that variable splitting employs a lower derivative order and lower power p .

F Broader Impacts

We analyzed the behavior of PINNs when applied to general k th-order PDEs and provided theoretical understanding of the reasons why PINNs encounter difficulties in optimizing when dealing with high-order or high-dimensional problems. This paper has societal impact as it bridges artificial intelligence and physics, and hence has a wide range of potential applications. It could contribute to the advancement in scientific understanding, and foster scientific progress and innovation, benefiting

Table 3: Computation costs with width=1000 for experiments on the paper.

PDE	Method	GPU memory	running time	parameters
Bi-harmonic	PINN	801.052 Mb	0.053 s/epoch	4000
	VS-PINN $ \xi = 2$	481.094 Mb	0.049 s/epoch	9000
	VS-PINN $ \xi = 1$	81,466 Mb	0.053 s/epoch	20000
Beam	PINN	323.772 Mb	0.037 s/epoch	4000
	VS-PINN $ \xi = 2$	240.689 Mb	0.038 s/epoch	9000
	VS-PINN $ \xi = 1$	80.836 Mb	0.040 s/epoch	17000

Table 4: The average of Mean Square Error (MSE) with exact solution.

PDE	PINNs	VS-PINNs $ \xi = 2$	VS-PINNs $ \xi = 1$
Elastic Beam	3.0598E-08	2.2868E-08	4.3826E-07
Bi-harmonic	3.9319E-05	5.3055E-05	5.4418E-03

researchers in physics, mathematics, and artificial intelligence fields. By providing variable splitting which can address the convergence pathologies of PINNs, the paper suggests potential applications in industrial and technological fields, such as engineering, medicine, materials science, and environmental modeling, for complex system modeling, prediction, and optimization. Furthermore, as many problems have arisen in sciences and engineering tied with complex PDE systems, we expect that our work has the potential to be applicable in the enormous area such as climate forecasting, epidemics, molecular simulations, micro-mechanics, and modeling turbulent flows. Additionally, PDEs can be utilized in the development of military equipment. As with all numerical algorithms, however, it is not a work of developing a technique to go to warfare; rather, it is a tool for scientific inquiry. We encourage users of our model to focus on the positive impact of this work.

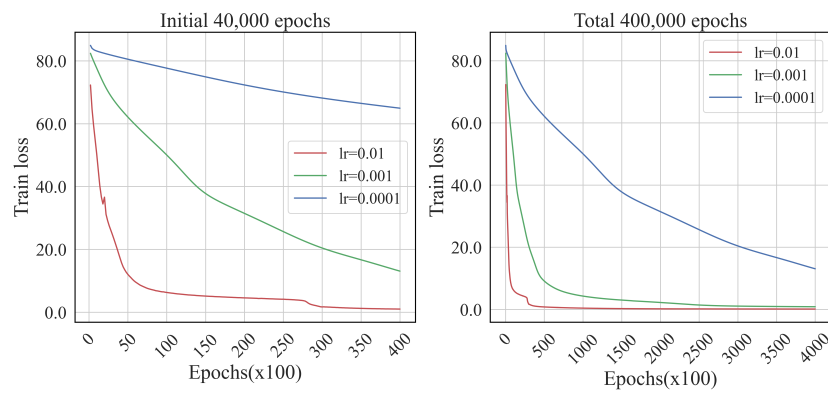


Figure 6: Loss of VS-PINNs trained by gradient descent with variant learning rates.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: [Yes]

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.