Gradient Descent Finds the Global Optima of Two-Layer Physics-Informed Neural Networks

Yihang Gao[1] **Yiqi Gu**[2] **Michael K. Ng**[1]

Abstract

The main aim of this paper is to conduct the convergence analysis of the gradient descent for two-layer physics-informed neural networks (PINNs). Here, the loss function involves derivatives of neural network outputs with respect to its inputs, so the interaction between the trainable parameters is more complicated compared with simple regression and classification tasks. We first develop the positive definiteness of Gram matrices and prove that the gradient flow finds the global optima of the empirical loss under over-parameterization. Then, we demonstrate that the standard gradient descent converges to the global optima of the loss with proper choices of learning rates. The framework of our analysis works for various categories of PDEs (e.g., linear second-order PDEs) and common types of network initialization (LecunUniform etc.). Our theoretical results do not need a very strict hypothesis for training samples and have a looser requirement on the network width compared with some previous works.

1. Introduction

Physics-informed neural networks (PINNs) have attracted significant attention in solving high-dimensional and nonlinear partial differential equations (PDEs) due to their evasion of the curse of dimensionality and friendly implementation (Raissi et al., 2019; Mao et al., 2020; Cai et al., 2022). For a given PDE

$$
\begin{aligned}
\mathcal{D}[\boldsymbol{u}, \boldsymbol{x}] &= \boldsymbol{f}(\boldsymbol{x}), \quad \boldsymbol{x} \in \Gamma \subset \mathbb{R}^d, \\
\mathcal{B}[\boldsymbol{u}, \boldsymbol{x}] &= \boldsymbol{g}(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Gamma,
\end{aligned} \tag{1}
$$

where $\boldsymbol{u}$ is the unknown solution; $\mathcal{D}$ and $\mathcal{B}$ are differential operators in the interior and on the boundary respectively; $\boldsymbol{f}$ and $\boldsymbol{g}$ are given smooth functions; Γ is an open bounded domain of our interest and $\partial\Gamma$ is its boundary. In PINNs, we adopt a neural network $\phi(\boldsymbol{x}; \boldsymbol{w})$ parameterized by $\boldsymbol{w}$, as a surrogate to the solution $\boldsymbol{u}(\boldsymbol{x})$, and then solve the following optimization problem

$$
\begin{aligned}
\min_{\boldsymbol{w}} \frac{1}{n_1} \sum_{p=1}^{n_1} \frac{1}{2} \left| \mathcal{D}[\phi(\boldsymbol{x}_p; \boldsymbol{w}), \boldsymbol{x}_p] - \boldsymbol{f}(\boldsymbol{x}_p) \right|^2 \\
+ \nu \cdot \frac{1}{n_2} \sum_{k=1}^{n_2} \frac{1}{2} \left| \mathcal{B}[\phi(\tilde{\boldsymbol{x}}_k; \boldsymbol{w}), \tilde{\boldsymbol{x}}_k] - \boldsymbol{g}(\tilde{\boldsymbol{x}}_k) \right|^2,
\end{aligned} \tag{2}
$$

where ν is the hyperparameter to balance the interior and boundary conditions; n_1 and n_2 are numbers of samples in the interior and on the boundary, respective; $\{\boldsymbol{x}_p\}_{p=1}^{n_1}$ and $\{\tilde{\boldsymbol{x}}_k\}_{k=1}^{n_2}$ are the training datasets sampled from Γ and $\partial\Gamma$, respectively. We aim to find the optimal neural network $\phi(\boldsymbol{x}; \boldsymbol{w})$ as an approximate solution by solving (2). Raissi et al. (Raissi et al., 2019) recommended using L-BFGS (Liu & Nocedal, 1989), which is a quasi-Newton method. However, first-order methods (i.e., gradient descent and its variants) are more popular and perform really well in implementations, see for instance (Gu et al., 2021; Cai et al., 2021; Meng et al., 2021; Mao et al., 2020).

Numerical examples of PINNs can be widely found in recent literature, including linear elliptic/parabolic/hyperbolic equations (Gu et al., 2021), Schrodinger equation (Raissi et al., 2019), Allen–Cahn equation (Raissi et al., 2019), compressible/incompressible flows (Cai et al., 2022; Mao et al., 2020), Hamilton–Jacobi–Bellman equation (Sirignano & Spiliopoulos, 2018), Burgers' equation (Sirignano & Spiliopoulos, 2018), etc. Although losses as small as $O(10^{-4})$ can be achieved in these experiments, optimization errors caused by the algorithms prevent the losses from being reduced to the machine precision. In theory, however, it is conjectured that the training loss of PINNs can be reduced to zero by gradient descent under the over-parameterization setting.

Existing analysis of the gradient descent usually depends on the smoothness (Carmon et al., 2018; Li & Orabona, 2019), the Lipschitzness of the Hessian (Carmon et al., 2018;

[1] Department of Mathematics, The University of Hong Kong, Pokfulam, Hong Kong [2] School of Mathematical Sciences, University of Electronic Science and Technology of China, Sichuan 611731, China. Correspondence to: Michael K. Ng <mng@maths.hku.hk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Nesterov & Polyak, 2006) and even the convexity (Duchi et al., 2011; Kingma & Ba, 2015; Reddi et al., 2019), so these works are not applicable to deep learning whose loss function is highly non-convex and not necessarily smooth. Also, much literature merely considers the convergence to local optima, but numerical experiments show that the gradient descent can nearly find the global optima, where the mean square loss decreases almost to zero (Zhang et al., 2021). This phenomenon in deep learning cannot be explained by classical convergence analysis but by the over-parameterization of neural networks. Soudry and Carmon (Soudry & Carmon, 2016) showed that all local minima are actually the global ones for over-parameterized neural networks. Du et al. (Du et al., 2019) proved that the gradient descent finds the global optima of over-parameterized ReLU neural networks for least squares problems. Wang et al. (Wang et al., 2022) analyzed the gradient flow of PINNs with positive-definiteness assumptions on Gram matrices and accelerated the convergence by involving eigenvalues of Gram matrices in the loss function. We refer to readers for more related works (Soltanolkotabi, 2017; Xie et al., 2017; Chizat & Bach, 2018; Jacot et al., 2018; Soltanolkotabi et al., 2018; Chatterjee, 2022). However, the (nearly) global convergence of the gradient descent for training PINNs observed in numerical experiments (Raissi et al., 2019; Pang et al., 2020; Mao et al., 2020) cannot be explained by the aforementioned results.

1.1. The Contributions

In this paper, we develop the convergence analysis of the gradient descent in optimizing two-layer PINNs. Here are our contributions.

- We provide a scheme for proving the positive definiteness of Gram matrices of PINNs without strict assumptions (Proposition 3.1 and Lemma 3.2). It can be applied to various types of PDEs with some minor modifications.
- We first theoretically prove that the gradient flow finds the global optima of over-parameterized physics-informed neural networks in solving a heat equation as a pedagogical example (Theorem 3.8):

$$\begin{aligned}&u_t(t,\boldsymbol{x})-\Delta_{\boldsymbol{x}}u(t,\boldsymbol{x})=f(t,\boldsymbol{x}),\quad (t,\boldsymbol{x})\in(0,T)\times\Gamma,\\&u(0,\boldsymbol{x})=g_1(\boldsymbol{x}),\quad \boldsymbol{x}\in\Gamma,\\&u(t,\boldsymbol{x})=g_2(t,\boldsymbol{x}),\quad (t,\boldsymbol{x})\in[0,T]\times\partial\Gamma.\end{aligned}\tag{3}$$

  Similar analysis and results can be achieved for a class of second-order linear PDEs with some minor modifications; see Section 3.1.
- We next prove that the gradient descent finds the global optima of the empirical loss of PINNs (Theorem 4.5). Here, the learning rate does not depend on the size of neural networks but relies on the PDE itself. We then extend our results from the pedagogical example to more general second-order linear PDEs; see Section 4.1. Our results also apply to some popular initialization methods, e.g., HeNormal (He et al., 2015), HeUniform (He et al., 2015), LecunNormal (Klambauer et al., 2017) and LecunUniform (Klambauer et al., 2017); see Corollary 4.7.

We should mention that there exist significant differences between our work for PINNs and the similar work for least squares regressions given in (Du et al., 2019). Firstly, to prove the positive definiteness of Gram matrices, they have a hypothesis for training samples that they cannot be (nearly) parallel, and this is difficult to be verified and satisfied in real applications. But our work does not need such a requirement for training samples (see Proposition 3.1 and Remark 3.4). Next, the loss function of PINNs is the residual of PDEs and involves partial derivatives of the network. Accordingly, our analysis is novel and more technical. Moreover, our results and frameworks can be simply applied to various PDE types and initialization types. In addition, we discuss the networks having bias terms. So our results are more applicable in practice compared with the works which do not study bias terms.

Another similar work is the optimization analysis of gradient descent in training PINNs for second-order linear PDEs (Luo & Yang, 2020), where the gradient flow representation and Rademacher complexity are utilized. In comparison, (Luo & Yang, 2020) only studies the gradient flow of the training, but our work considers both the continuous gradient flow and, more practically, the discrete gradient descent. Moreover, our theory (Corollary 4.7) gives a looser requirement of the network width ($\widetilde{\Omega}(n^{2/3})$) than that given in (Luo & Yang, 2020) ($\widetilde{\Omega}(n^4)$), where n is the number of training samples.

This paper is organized as follows. Some preliminaries, including a brief introduction to PINNs and some preparation works, are presented in Section 2. We provide detailed results for the convergence of the continuous gradient flow and the discrete gradient descent in Sections 3 and 4, respectively. In Section 5, results of numerical experiments on 1-d heat equation are displayed. Some concluding and potential works are discussed in Section 6.

2. Preliminaries

We write two functions with relations $f_1(n)=\mathcal{O}\left(f_2(n)\right)$, or equivalently $f_2=\Omega\left(f_1(n)\right)$, if there exists a constant C such that $f_1(n)\leq C\cdot f_2(n)$. If we further omit some logarithmic terms with the existence of polynomial terms, we adopt $f_1(n)=\widetilde{\mathcal{O}}\left(f_2(n)\right)$ and $f_2=\widetilde{\Omega}\left(f_1(n)\right)$. We use boldface capital and lowercase letters to denote matrices and vectors respectively. Non-bold letters represent the elements of matrices or vectors. For example, $A_{i,j}$ denotes the (i,j)-th element of the matrix $\boldsymbol{A}$. For a positive integer m, the set $\{1,\cdots,m\}$ is abbreviated as $[m]$. Especially, we use $\boldsymbol{e}_i \in \mathbb{R}^{d+2}$ to denote the elementary vector whose i-th $(0 \le i \le d+1)$ element is 1 and others are 0.

Denote the variable $\boldsymbol{x} = [x_0 \; x_1 \; \cdots \; x_d]^\top \in \mathbb{R}^{d+1}$, where $x_0 \in [0,T]$ and $[x_1 \; \cdots \; x_d]^\top \in \overline{\Gamma}$. Without loss of generality, we assume the domain of our interest $[0,T]\times\overline{\Gamma}$ is bounded such that $\|\boldsymbol{x}\|_2 \le \frac{\sqrt{3}}{2}$ for all $\boldsymbol{x} \in [0,T]\times\overline{\Gamma}$. Note that the upper bound $\frac{\sqrt{3}}{2}$ is artificially chosen for convenience. For any PDEs with a larger but bounded domain, it can be rescaled so that the domain is small enough below the upper bound. Next, the PDE (3) is rewritten as

$$\begin{aligned} &\frac{\partial u}{\partial x_0}(\boldsymbol{x}) - \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in (0,T)\times\Gamma, \\ &u(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \{0\}\times\Gamma \cup [0,T]\times\partial\Gamma, \end{aligned} \tag{4}$$

Moreover, we consider $\phi(\boldsymbol{x};\boldsymbol{w},\boldsymbol{a})$ as a shallow (with 1 hidden layer) but wide neural network with bias terms, defined as

$$\begin{aligned} &\phi(\boldsymbol{x};\boldsymbol{w},\boldsymbol{a}) \\ =&\frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \cdot \sigma\left([w_{r0} \; w_{r1} \cdots w_{rd}]\boldsymbol{x} + \frac{1}{2} w_{r,d+1} \right) \\ =&\frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \cdot \sigma\left(\boldsymbol{w}_r^\top \boldsymbol{y} \right), \end{aligned} \tag{5}$$

where $\boldsymbol{w}_r = [w_{r0} \; w_{r1} \; \cdots \; w_{rd} \; w_{r,d+1}]^\top \in \mathbb{R}^{d+2}$, $\boldsymbol{w} = [\boldsymbol{w}_1^\top \; \cdots \; \boldsymbol{w}_m^\top]^\top \in \mathbb{R}^{m(d+2)}$, $\boldsymbol{a} = [a_1 \; \cdots \; a_m]^\top \in \mathbb{R}^m$, $\boldsymbol{y} = [\boldsymbol{x}^\top \; 1/2]^\top$ and $\sigma(\cdot)$ is the activation function. In this paper, we consider the case that $\sigma(\cdot)$ is the ReLU^3 activation function (i.e., $\sigma(x) = \max(0, x^3)$), which is widely used in solving second-order PDEs. Throughout the paper, we use $\boldsymbol{y} \in \mathbb{R}^{d+2}$ to denote the augmented vector whose first d+1 elements are copied from $\boldsymbol{x}$ and the last element is assigned to be $1/2$. Therefore, we have $\|\boldsymbol{y}\|_2 \le 1$ for all $\boldsymbol{x} \in [0,T]\times\overline{\Gamma}$. Note that $\frac{1}{2}w_{r,d+1}$ is the bias term of the neural network. Here, we use $\frac{1}{2}w_{r,d+1}$ rather than $w_{r,d+1}$ because we hope $\|\boldsymbol{y}\|_2 \le 1$ to simplify the analysis. In the setting, $\phi(\boldsymbol{x};\boldsymbol{w},\boldsymbol{a})$ is second-order continuously differentiable and is a good approximation structure of the true solution of the PDE (4) (Siegel & Xu, 2022).

Therefore, the corresponding optimization is given by

$$\begin{aligned} &\min_{\boldsymbol{w},\boldsymbol{a}} \mathcal{L}(\boldsymbol{w},\boldsymbol{a}) := \\ &\sum_{p=1}^{n_1} \frac{1}{2n_1} \left(\frac{\partial \phi}{\partial x_0}(\boldsymbol{x}_p;\boldsymbol{w},\boldsymbol{a}) - \sum_{i=1}^{d} \frac{\partial^2 \phi}{\partial x_i^2}(\boldsymbol{x}_p;\boldsymbol{w},\boldsymbol{a}) - f(\boldsymbol{x}_p) \right)^2 \\ &+ \nu \cdot \frac{1}{n_2} \sum_{k=1}^{n_2} \frac{1}{2} \left(\phi(\tilde{\boldsymbol{x}}_k;\boldsymbol{w},\boldsymbol{a}) - g(\tilde{\boldsymbol{x}}_k) \right)^2, \end{aligned} \tag{6}$$

where $\{\boldsymbol{x}_p\}_{p=1}^{n_1} \bigcup \{\tilde{\boldsymbol{x}}_k\}_{k=1}^{n_2}$ is the set of training samples in the interior or on the boundary. Correspondingly, we use $\{\boldsymbol{y}_p\}_{p=1}^{n_1} \bigcup \{\tilde{\boldsymbol{y}}_k\}_{k=1}^{n_2}$ to denote the augmented training samples.

The gradient descent method solves (6) by the following formulation:

$$\boldsymbol{w}_r(t+1) = \boldsymbol{w}_r(t) - \eta \cdot \frac{\partial \mathcal{L}(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}_r}, \tag{7}$$

$$a_r(t+1) = a_r(t) - \eta \cdot \frac{\partial \mathcal{L}(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial a_r} \tag{8}$$

for all $r \in [m]$, where $t \in \mathbb{N}$ and $\eta > 0$ is the learning rate. Note that the activation function $\sigma(x) := \max\{0, x^3\}$ is third-order differentiable except at $x = 0$, we may define its third-order derivative as $\sigma'''(x) = 6\mathbb{I}(x > 0)$, where $\mathbb{I}(\cdot)$ is the indicator function. Throughout this paper, we consider the initialization

$$\boldsymbol{w}_r(0) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d+2}), \quad a_r(0) \sim \text{Unif}(\{-1, 1\}). \tag{9}$$

Our scheme is valid for other types of PDEs and initialization methods, and we will discuss them later. Here, we adopt the $\{-1,1\}$ initialization for $a_r(0)$ to simplify the proof, as in (Du et al., 2019). Readers can directly apply our results for Normal/Uniform initialization (e.g., $a_r(0) \sim \mathcal{N}(0,1)$) without much modification (up to some constants).

3. Continuous Time Analysis

In this section, we formulate the training task (7)-(8) as a gradient flow, which can be viewed as a continuous form of gradient descent with an infinitesimal time step size. This continuous time analysis of gradient flow is a stepping stone toward understanding the discrete gradient descent algorithms. We prove that the gradient flow converges to the global optima of the loss under over-parameterization and some mild conditions on training samples. Without ambiguity, we regard $t \ge 0$ as a real number in this section.

The time continuous form of (7)-(8) is characterized as the following dynamics

$$\frac{d\boldsymbol{w}(t)}{dt} = -\frac{\partial \mathcal{L}(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}}, \quad \frac{d\boldsymbol{a}(t)}{dt} = -\frac{\partial \mathcal{L}(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{a}}. \tag{10}$$

Since the third derivative of the activation function is a Heaviside function, the right-hand sides are discontinuous at points of zero measure, so Equation (10) may not have a solution in the classical sense. But it has a weak solution (i.e., $\boldsymbol{w}$ and $\boldsymbol{a}$ have a weak derivative with respect to t) depending on the initial condition. Let

$$s_p(\boldsymbol{w},\boldsymbol{a}) = \sqrt{\frac{1}{n_1}}\left(\frac{\partial \phi}{\partial x_0}(\boldsymbol{x}_p;\boldsymbol{w},\boldsymbol{a}) - \sum_{i=1}^{d}\frac{\partial^2 \phi}{\partial x_i^2}(\boldsymbol{x}_p;\boldsymbol{w},\boldsymbol{a}) - f(\boldsymbol{x}_p)\right), \tag{11}$$

and

$$h_k(\boldsymbol{w},\boldsymbol{a}) = \sqrt{\frac{\nu}{n_2}}\left(\phi(\tilde{\boldsymbol{x}}_k;\boldsymbol{w},\boldsymbol{a}) - g(\tilde{\boldsymbol{x}}_k)\right). \tag{12}$$

We have

$$\mathcal{L}(\boldsymbol{w},\boldsymbol{a}) = \frac{1}{2}\left(\|\boldsymbol{s}(\boldsymbol{w},\boldsymbol{a})\|_2^2 + \|\boldsymbol{h}(\boldsymbol{w},\boldsymbol{a})\|_2^2\right), \tag{13}$$

where vectors $\boldsymbol{s}(\boldsymbol{w},\boldsymbol{a}) = [s_1(\boldsymbol{w},\boldsymbol{a}) \ \cdots \ s_{n_1}(\boldsymbol{w},\boldsymbol{a})]^\top$ and $\boldsymbol{h}(\boldsymbol{w},\boldsymbol{a}) = [h_1(\boldsymbol{w},\boldsymbol{a}) \ \cdots \ h_{n_2}(\boldsymbol{w},\boldsymbol{a})]^\top$. Therefore,

$$\begin{aligned}\frac{d\boldsymbol{w}_r}{dt} &= -\frac{\partial \mathcal{L}(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r}\\ &= -\sum_{p=1}^{n_1} s_p(\boldsymbol{w},\boldsymbol{a})\cdot\frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r} - \sum_{k=1}^{n_2} h_k(\boldsymbol{w},\boldsymbol{a})\cdot\frac{\partial h_k(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r},\end{aligned} \tag{14}$$

and

$$\begin{aligned}\frac{da_r}{dt} &= -\frac{\partial \mathcal{L}(\boldsymbol{w},\boldsymbol{a})}{\partial a_r}\\ &= -\sum_{p=1}^{n_1} s_p(\boldsymbol{w},\boldsymbol{a})\cdot\frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial a_r} - \sum_{k=1}^{n_2} h_k(\boldsymbol{w},\boldsymbol{a})\cdot\frac{\partial h_k(\boldsymbol{w},\boldsymbol{a})}{\partial a_r}.\end{aligned} \tag{15}$$

Using the chain rule and (14)-(15), we can derive the following gradient flow (see more details in Appendix A.2):

$$\frac{d}{dt}\begin{bmatrix}\boldsymbol{s}(\boldsymbol{w},\boldsymbol{a})\\ \boldsymbol{h}(\boldsymbol{w},\boldsymbol{a})\end{bmatrix} = -\left(\boldsymbol{G}(\boldsymbol{w},\boldsymbol{a}) + \widetilde{\boldsymbol{G}}(\boldsymbol{w},\boldsymbol{a})\right)\begin{bmatrix}\boldsymbol{s}(\boldsymbol{w},\boldsymbol{a})\\ \boldsymbol{h}(\boldsymbol{w},\boldsymbol{a})\end{bmatrix}, \tag{16}$$

where $\boldsymbol{G}(\boldsymbol{w},\boldsymbol{a})$ and $\widetilde{\boldsymbol{G}}(\boldsymbol{w},\boldsymbol{a})$ are the Gram matrices for the dynamics, defined as

$$\boldsymbol{G}(\boldsymbol{w},\boldsymbol{a}) = \boldsymbol{D}^\top\boldsymbol{D}, \quad \boldsymbol{D} = \begin{bmatrix}\frac{\partial s_1}{\partial \boldsymbol{w}} & \cdots & \frac{\partial s_{n_1}}{\partial \boldsymbol{w}} & \frac{\partial h_1}{\partial \boldsymbol{w}} & \cdots & \frac{\partial h_{n_2}}{\partial \boldsymbol{w}}\end{bmatrix} \tag{17}$$

and

$$\widetilde{\boldsymbol{G}}(\boldsymbol{w},\boldsymbol{a}) = \widetilde{\boldsymbol{D}}^\top\widetilde{\boldsymbol{D}}, \quad \widetilde{\boldsymbol{D}} = \begin{bmatrix}\frac{\partial s_1}{\partial \boldsymbol{a}} & \cdots & \frac{\partial s_{n_1}}{\partial \boldsymbol{a}} & \frac{\partial h_1}{\partial \boldsymbol{a}} & \cdots & \frac{\partial h_{n_2}}{\partial \boldsymbol{a}}\end{bmatrix} \tag{18}$$

Note that $\widetilde{\boldsymbol{G}}(\boldsymbol{w},\boldsymbol{a})$ is independent of $\boldsymbol{a}$, but we keep the variable $\boldsymbol{a}$ here for the consistent symbol format with $\boldsymbol{G}(\boldsymbol{w},\boldsymbol{a})$. For readability, we provide some preparatory computation in Appendix A.1. In Lemma 3.2, we will first prove the positive definiteness of the expectation of the Gram matrix. And then the initialized Gram matrices are positive definite with high probabilities (see Lemma 3.5). The following Proposition 3.1 provides sufficient conditions for the positive definiteness of the expectation of the Gram matrix.

Proposition 3.1. *If two samples in $\{\boldsymbol{y}_p\}_{p=1}^{n_1}\bigcup\{\tilde{\boldsymbol{y}}_k\}_{k=1}^{n_2}$ are parallel, say, $\boldsymbol{y} = \alpha\cdot\bar{\boldsymbol{y}}$ for some $\boldsymbol{y},\bar{\boldsymbol{y}} \in \{\boldsymbol{y}_p\}_{p=1}^{n_1}\bigcup\{\tilde{\boldsymbol{y}}_k\}_{k=1}^{n_2}$ and $\alpha\in\mathbb{R}$, then $\boldsymbol{y} = \bar{\boldsymbol{y}}$.*

Proof. Note that our model (5) involves a bias term that the last element of all training samples $\boldsymbol{y}$ are $1/2$. Then two data points $\boldsymbol{y}$ and $\bar{\boldsymbol{y}}$ are parallel if and only if $\boldsymbol{y} = \bar{\boldsymbol{y}}$.

□

We assume that all points in the training set $\{\boldsymbol{y}_p\}_{p=1}^{n_1}\bigcup\{\tilde{\boldsymbol{y}}_k\}_{k=1}^{n_2}$ are distinct, then by Proposition 3.1, there do not exist parallel points.

The following useful lemmas are generalizations of the results in the least squares regression model (Du et al., 2019). One of our main contributions is that we provide a scheme to prove the positive definiteness of the Gram matrices for PINNs. Besides the techniques, our results do not need strict assumptions compared with (Du et al., 2019). In the most related papers (Wang et al., 2022; Luo & Yang, 2020), they skipped the theoretical proof for the positive-definiteness of Gram matrices. For the readability and brevity of the paper, we put all detailed proofs in Appendix B.

We would like to mention that Theorem 2.1 in (He et al., 2020) proved the linear independence of ReLU$(\boldsymbol{w}^\top\boldsymbol{y}_p)$, for $p = 1,\cdots,n_1$. Our results in Lemma 3.2 extend their results and show the linear independence of ReLU$(\boldsymbol{w}^\top\boldsymbol{y}_p)$, ReLU$^2(\boldsymbol{w}^\top\boldsymbol{y}_p)$ and ReLU$^3(\boldsymbol{w}^\top\boldsymbol{y}_p)$, for $p = 1,\cdots,n_1$ (i.e., linear independence of columns of $\widetilde{\boldsymbol{D}}$). Moreover, the linear independence of columns of $\widetilde{\boldsymbol{D}}$ also appears in the physics-informed extreme learning machine (PIELM) (Dwivedi & Srinivasan, 2020). Our work tries to explain the phenomenon that the gradient descent finds the global optima of PINNs, while (Dwivedi & Srinivasan, 2020) shows the existence of zero loss under overparameterization.

Lemma 3.2. *Let $\widetilde{\boldsymbol{G}}^\infty = \boldsymbol{E}_{\boldsymbol{w}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}),\boldsymbol{a}\sim\{-1,1\}}\widetilde{\boldsymbol{G}}(\boldsymbol{w},\boldsymbol{a})$, then $\widetilde{\boldsymbol{G}}^\infty$ is strictly positive definite and its minimal eigenvalue $\tilde{\lambda}_0 := \lambda_{\min}(\widetilde{\boldsymbol{G}}^\infty) > 0$ is independent of m (the size of neural network in (5)).*

Lemma 3.3. *Let $\boldsymbol{G}^\infty = \boldsymbol{E}_{\boldsymbol{w}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}),\boldsymbol{a}\sim\{-1,1\}}\boldsymbol{G}(\boldsymbol{w},\boldsymbol{a})$, then $\boldsymbol{G}^\infty$ is strictly positive definite and its minimal eigenvalue $\lambda_0 := \lambda_{\min}(\boldsymbol{G}^\infty) > 0$ is independent of m.*

The proof for Lemma 3.3 can be similarly developed by using the argument of Lemma 3.2 (see Appendix B.1).

Remark 3.4. Note that if samples $\{\boldsymbol{x}_p\}_{p=1}^{n_1}\bigcup\{\tilde{\boldsymbol{x}}_k\}_{k=1}^{n_2}$ are fixed and the neural network is initialized, then eigenvalues $\widetilde{\lambda}_0$ and λ_0 are fixed. According to the definition of $\widetilde{\boldsymbol{D}}$ and $\boldsymbol{D}$ and as well as the proof, if two data points are very close, then their associated columns of $\widetilde{\boldsymbol{D}}$ and $\boldsymbol{D}$ are nearly linear dependent and hence $\widetilde{\lambda}_0, \lambda_0 \approx 0$. But the close singularity will not ruin the dynamics of the gradient flow Equation (16). Suppose that $\boldsymbol{x}_1 \approx \boldsymbol{x}_2$, by the smoothness of the PDE and mean value theorem, there exists some $\boldsymbol{x}_3 = \alpha\boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2$ for some $0 \le \alpha \le 1$ such that $2s_3 = s_1 + s_2$ and $2\frac{\partial s_3}{\partial \boldsymbol{a}} \approx \frac{\partial s_1}{\partial \boldsymbol{a}} + \frac{\partial s_2}{\partial \boldsymbol{a}}$. Then the gradient flow Equation (16) is close to the new gradient flow where the squared terms of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are replaced with the squared terms of $\boldsymbol{x}_3$, which does not have the close singularity. This is not true for problems where the loss (i.e., $s_p(\boldsymbol{w},\boldsymbol{a})$ and $h_k(\boldsymbol{w},\boldsymbol{a})$) is not sufficiently smooth.

Lemma 3.5. *If* $m = \widetilde{\Omega}\left(\frac{(n_1+n_2)^4}{(n_1n_2)^2\cdot\left(\min\{\lambda_0,\widetilde{\lambda}_0\}\right)^2}\cdot\left(\log\frac{1}{\delta}\right)^7\right)$ *over the initialization (9), then with probability of at least* $1-\delta$*, we have* $\|\boldsymbol{G}(\boldsymbol{w}(0),\boldsymbol{a}(0)) - \boldsymbol{G}^\infty\|_2 \le \frac{\lambda_0}{4}$ *and* $\left\|\widetilde{\boldsymbol{G}}(\boldsymbol{w}(0),\boldsymbol{a}(0)) - \widetilde{\boldsymbol{G}}^\infty\right\|_2 \le \frac{\widetilde{\lambda}_0}{4}$. *Moreover, we have* $\lambda_{\min}(\boldsymbol{G}(\boldsymbol{w}(0),\boldsymbol{a}(0))) \ge \frac{3}{4}\lambda_0$ *and* $\lambda_{\min}\left(\widetilde{\boldsymbol{G}}\left(\boldsymbol{w}(0),\boldsymbol{a}(0)\right)\right) \ge \frac{3}{4}\widetilde{\lambda}_0$ *hold.*

Lemma 3.6. *Suppose that* $\boldsymbol{w}_r(0)$ *and* $a_r(0)$*,* $r \in [m]$ *are initialized independently by (9), then with probability of at least* $1-\delta$*, we have*

$$\|\boldsymbol{G}(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}}) - \boldsymbol{G}(\boldsymbol{w}(0),\boldsymbol{a}(0))\|_2 \le \frac{\lambda_0}{4}$$

and

$$\left\|\widetilde{\boldsymbol{G}}(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}}) - \widetilde{\boldsymbol{G}}(\boldsymbol{w}(0),\boldsymbol{a}(0))\right\|_2 \le \frac{\widetilde{\lambda}_0}{4},$$

for all $\|\tilde{\boldsymbol{w}}_r - \boldsymbol{w}_r(0)\|_2 \le R_w$, $|\tilde{a}_r - a_r(0)| \le R_a \le 1$ *and* $r \in [m]$*, where* $R_w = \widetilde{\mathcal{O}}\left(\frac{\min\{\lambda_0,\widetilde{\lambda}_0\}\cdot\delta}{(n_1+n_2)\cdot(\log m)^3}\right)$ *and* $R_a = \widetilde{\mathcal{O}}\left(\frac{\min\{\lambda_0,\widetilde{\lambda}_0\}\cdot\delta}{(n_1+n_2)\cdot(\log m)^2}\right)$.

Lemma 3.7. *With probability of at least* $1-\delta$ *over the initialization (9) for all* $r \in [m]$*, we have* $\left\|\begin{bmatrix}\boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0))\\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{bmatrix}\right\|_2^2 = \mathcal{O}(\frac{1}{\delta})$.

Theorem 3.8. *For given training samples* $\{\boldsymbol{x}_p\}_{p=1}^{n_1}\bigcup\{\tilde{\boldsymbol{x}}_k\}_{k=1}^{n_2}$*, if all weights of PINNs are initialized by (9) for all* $r \in [m]$*, then with probability of at least* $1-\delta$*,*

$$\mathcal{L}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \le \exp\left(-\left(\lambda_0 + \widetilde{\lambda}_0\right)\cdot t\right)\cdot\mathcal{L}(\boldsymbol{w}(0),\boldsymbol{a}(0)), \tag{19}$$

for all $t \ge 0$*, if* $m = \widetilde{\Omega}\left(\frac{(n_1+n_2)^2}{\left(\lambda_0+\widetilde{\lambda}_0\right)^2\cdot\left(\min\{\lambda_0,\widetilde{\lambda}_0\}\right)^2\cdot\delta^3}\right)$.

Proof. Firstly, we prove the positive definiteness of the initialized Gram matrix $\boldsymbol{G}(\boldsymbol{w}(0),\boldsymbol{a}(0))$ and $\widetilde{\boldsymbol{G}}(\boldsymbol{w}(0),\boldsymbol{a}(0))$ (Lemma 3.5). Moreover, Gram matrices $\boldsymbol{G}(\boldsymbol{w},\boldsymbol{a})$ and $\widetilde{\boldsymbol{G}}(\boldsymbol{w},\boldsymbol{a})$ are continuous with respect to $(\boldsymbol{w},\boldsymbol{a})$, with a high probability (Lemma 3.6). With sufficiently large m, $\boldsymbol{w}_r(\tau)$ and $a_r(\tau)$ stay close to the initialization $\boldsymbol{w}_r(0)$ and $a_r(0)$ for all $r \in [m]$ and thus the Gram matrices keep positive definite. Finally, Equation (16) implies the monotonically decreasing of the loss function with positive definite Gram matrices. The detailed proof can be found in Appendix B.5. □

Remark 3.9. If operators $\mathcal{D}$ and $\mathcal{B}$ are polynomials of u and its derivatives, then $\widetilde{\boldsymbol{G}}(\boldsymbol{w},\boldsymbol{a})$, $\boldsymbol{G}(\boldsymbol{w},\boldsymbol{a})$ and $\mathcal{L}(\boldsymbol{w},\boldsymbol{a})$ are polynomials of $(\boldsymbol{w},\boldsymbol{a})$. Under initialization methods whose tails decay faster than polynomials (e.g., Gaussian and uniform distributions), we can adopt concentration inequalities to similarly prove Lemma 3.5, Lemma 3.6 and Theorem 3.8 for PINNs in solving various kinds of PDEs.

3.1. Generalization to Linear Second-Order PDEs

In this section, we extend the main results of gradient flow from the heat equation Equation (4) to more general second-order linear PDEs. Considering the following second-order parabolic PDE

$$\begin{aligned}\frac{\partial u}{\partial x_0}(\boldsymbol{x}) - \sum_{i,j=1}^{d} b_{ij}(\boldsymbol{x})\cdot\frac{\partial^2 u}{\partial x_i \partial x_j}(\boldsymbol{x}) - \sum_{i=1}^{d} c_i(\boldsymbol{x})\cdot\frac{\partial u}{\partial x_i}(\boldsymbol{x})&\\ -\ell(\boldsymbol{x})\cdot u(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in (0,T)\times\Gamma,&\\ u(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \{0\}\times\Gamma \cup [0,T]\times\partial\Gamma.&\end{aligned} \tag{20}$$

Here, we assume that $b_{ij}(\boldsymbol{x}) = b_{ji}(\boldsymbol{x})$ and there exists $M > 0$ such that $|b_{ij}(\boldsymbol{x})| \le M$, $|c_i(\boldsymbol{x})| \le M$ and $|\ell(\boldsymbol{x})| \le M$ for all $1 \le i,j \le d$ and $\boldsymbol{x} \in [0,T]\times\overline{\Gamma}$. Without ambiguity, We use the same notations such as $\phi(\boldsymbol{x};\boldsymbol{w},\boldsymbol{a})$, $\mathcal{L}(\boldsymbol{w},\boldsymbol{a})$, $\boldsymbol{y}$ and λ_0, etc., as the heat equation case. Without much effort, one can reestablish the preceding results for the general PDE (20), where additional terms related to M appear in concentration inequalities, so the proof is similar and we omit the details. Specifically, one can prove that Lemma 3.2 and Lemma 3.3 still hold for (20) without giving any extra hypotheses. Next, it can be proved that Lemma 3.5 is true for (20) by replacing the hypothesis of m with $m = \widetilde{\Omega}\left(\frac{M^4\cdot(n_1+n_2)^4}{(n_1n_2)^2\cdot\left(\min\{\lambda_0,\widetilde{\lambda}_0\}\right)^2}\cdot\left(\log\frac{1}{\delta}\right)^7\right)$. One can also prove that Lemma 3.6 holds for (20) if changing the hypothesis of R_w and R_a as

$$R_w = \widetilde{\mathcal{O}}\left(\frac{\min\{\lambda_0,\widetilde{\lambda}_0\}\cdot\delta}{M^2\cdot(n_1+n_2)\cdot(\log m)^3}\right)$$

and

$$R_a = \widetilde{\mathcal{O}}\left(\frac{\min\{\lambda_0, \widetilde{\lambda}_0\}\cdot\delta}{M^2\cdot(n_1+n_2)\cdot(\log m)^2}\right).$$

And Lemma 3.7 is also true with the conclusion replaced with $\left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{bmatrix}\right\|_2^2 = \mathcal{O}(\frac{M^2}{\delta})$.

Based on the aforementioned generalized lemmas, we can easily develop the following Corollary 3.10, which is a generalization of Theorem 3.8.

Corollary 3.10. *For given training samples $\{\boldsymbol{x}_p\}_{p=1}^{n_1}\bigcup\{\tilde{\boldsymbol{x}}_k\}_{k=1}^{n_2}$ from PDE (20), if all weights of PINNs are initialized by (9) for all $r\in[m]$, then with a probability of at least $1-\delta$, we have*

$$\mathcal{L}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \le \exp\left(-\left(\lambda_0+\widetilde{\lambda}_0\right)\cdot t\right)\cdot\mathcal{L}(\boldsymbol{w}(0),\boldsymbol{a}(0)), \tag{21}$$

for all $t\ge 0$, if $m = \widetilde{\Omega}\left(\frac{M^8\cdot(n_1+n_2)^2}{\left(\lambda_0+\widetilde{\lambda}_0\right)^2\cdot\left(\min\{\lambda_0,\widetilde{\lambda}_0\}\right)^2\cdot\delta^3}\right)$.

One can also obtain the same results for second-order linear elliptic PDEs ((20) without $\frac{\partial u}{\partial x_0}(\boldsymbol{x})$) and second-order linear hyperbolic PDEs ((20) with $\frac{\partial u}{\partial x_0}(\boldsymbol{x})$ replaced by $\frac{\partial^2 u}{\partial x_0^2}(\boldsymbol{x})$) up to constants.

4. Discrete Time Analysis

As the Euler's form of the gradient flow, the gradient descent can also find the global optima of the loss function. In this section, we turn to regard $t\in\mathbb{N}$. The convergence of the gradient descent consists of the following several lemmas. We first prove that the parameters $\boldsymbol{w}(t)$ and $\boldsymbol{a}(t)$ do not go far away from the initialization $\boldsymbol{w}(0)$ and $\boldsymbol{a}(0)$ (Lemma 4.1). Moreover, in each step, the error between the finite difference (i.e., the gradient descent) and the exact continuous dynamic (i.e., the gradient flow) is small if m is large and the learning rate is small enough (Lemma 4.2). Finally, the loss is strictly decreasing by the gradient descent, since the error is minor (Theorem 4.5). Note that the theoretical learning rate should be $\mathcal{O}(\lambda_0+\widetilde{\lambda}_0)$, which relies on the PDE itself but is independent of m. As is shown in Corollary 4.7, our framework can be extended to neural networks initialized by HeNormal, HeUniform (He et al., 2015), LecunNormal or LecunUniform (Klambauer et al., 2017) with some minor modifications to the following lemmas and theorems. For readability and brevity, we put the detailed proofs in Appendix C.

Lemma 4.1. *If $\|\boldsymbol{w}_r(t)\|_2\le R$, $|a_r|\le 2$, for $r\in[m]$ and*

$$\begin{aligned}&\left\|\begin{pmatrix} \boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t))\end{pmatrix}\right\|_2^2 \\ &\le\left(1-\eta\cdot\frac{\lambda_0+\widetilde{\lambda}_0}{2}\right)^t\cdot\left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{bmatrix}\right\|_2^2,\end{aligned} \tag{22}$$

for $t=0,\cdots,T$ and $\eta<\frac{2}{\lambda_0+\widetilde{\lambda}_0}$, then we have

$$\begin{aligned}&\|\boldsymbol{w}_r(t+1)-\boldsymbol{w}_r(0)\|_2 \\ &\le c_0\cdot\frac{R^2}{\sqrt{m}}\cdot\frac{1}{\lambda_0+\widetilde{\lambda}_0}\cdot\left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{bmatrix}\right\|_2 := R_w\end{aligned} \tag{23}$$

and

$$\begin{aligned}&|a_r(t+1)-a_r(0)| \\ &\le c_1\cdot\frac{R^3}{\sqrt{m}}\cdot\frac{1}{\lambda_0+\widetilde{\lambda}_0}\cdot\left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{bmatrix}\right\|_2 := R_a,\end{aligned} \tag{24}$$

for universal constants $c_0>0$ and $c_1>0$. Moreover,

$$\begin{aligned}&\|\boldsymbol{w}_r(t+1)-\boldsymbol{w}_r(t)\|_2 \\ &\le\eta\cdot\frac{c_0\cdot R^2}{4\sqrt{m}}\cdot\left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t))\end{bmatrix}\right\|_2 := \widetilde{R}_w\end{aligned}$$

and

$$\begin{aligned}&|a_r(t+1)-a_r(t)| \\ &\le\eta\cdot\frac{c_1\cdot R^3}{4\sqrt{m}}\cdot\left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t))\end{bmatrix}\right\|_2 := \widetilde{R}_a.\end{aligned}$$

Lemma 4.2. *With probability of at least $1-\delta$ over the initialization (9) of $\boldsymbol{w}_r(0)$ and $a_r(0)$, we have*

$$\|\boldsymbol{w}_r(0)\|_2\le R' := \sqrt{2(d+2)\cdot\log\left(\frac{2m(d+2)}{\delta}\right)},$$

for all $r\in[m]$. Moreover, if conditions in Lemma 4.1 hold for all $t=0,\cdots,T$ with $R=R'$, then

$$\begin{aligned}\left\|\begin{pmatrix} \boldsymbol{\chi}(t) \\ \tilde{\boldsymbol{\chi}}(t)\end{pmatrix}\right\|_2 &\le\tilde{c}_0\cdot\eta\cdot\left(\frac{\sqrt{n_1+n_2}}{\delta\cdot(\lambda_0+\widetilde{\lambda}_0)\cdot\sqrt{m}}\right)R'^8 \\ &\cdot\left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t))\end{bmatrix}\right\|_2\cdot\left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{bmatrix}\right\|_2 \\ &+\tilde{c}_1\cdot\eta^2\cdot\frac{R'^7}{\sqrt{m}}\cdot\left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t))\end{bmatrix}\right\|_2^2,\end{aligned}$$

for some universal constants $\tilde{c}_0>0$ and $\tilde{c}_1>0$, where $\boldsymbol{\chi}(t)=[\chi_1(t)\ \cdots\ \chi_{n_1}(t)]^\top$ and $\tilde{\boldsymbol{\chi}}(t)=[\tilde{\chi}_1(t)\ \cdots\ \tilde{\chi}_{n_2}(t)]^\top$ with

$$\begin{aligned}\chi_p(t) :=\ & s_p(\boldsymbol{w}(t+1),\boldsymbol{a}(t+1))-s_p(\boldsymbol{w}(t),\boldsymbol{a}(t)) \\ &-\left(\left\langle\frac{\partial s_p(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial\boldsymbol{w}},\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\right\rangle\right. \\ &\left.+\left\langle\frac{\partial s_p(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial\boldsymbol{a}},\boldsymbol{a}(t+1)-\boldsymbol{a}(t)\right\rangle\right),\end{aligned} \tag{25}$$

and

$$\begin{aligned}\tilde{\chi}_k(t) &:= h_k(\boldsymbol{w}(t+1),\boldsymbol{a}(t+1)) - h_k(\boldsymbol{w}(t),\boldsymbol{a}(t)) \\ &- \left(\left\langle \frac{\partial h_k(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}}, \boldsymbol{w}(t+1)-\boldsymbol{w}(t)\right\rangle\right. \\ &+ \left.\left\langle \frac{\partial h_k(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{a}}, \boldsymbol{a}(t+1)-\boldsymbol{a}(t)\right\rangle\right). \end{aligned} \tag{26}$$

Here, $\chi_p(t)$ and $\tilde{\chi}_k(t)$ are the residuals of first-order Taylor expansions.

Lemma 4.3. *Let $C_0 = \mathbb{E}\|\boldsymbol{w}_r(0)\|_2^4 + 1$ and $C_1 = \mathbb{E}\|\boldsymbol{w}_r(0)\|_2^6 + 1$. Assume that Lemma 4.1 holds, then with probability of at least $1-\delta$, we have*

(i) If $m = \widetilde{\Omega}\left(\frac{(\log(\frac{1}{\delta}))^4}{(\lambda_0+\widetilde{\lambda}_0)^2} \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix} \right\|_2^2\right)$, then $\frac{1}{m}\sum_{r=1}^{m}\|\boldsymbol{w}_r(t+1)\|_2^4 \leq 2C_0$;

(ii) If $m = \widetilde{\Omega}\left(\frac{(\log(\frac{1}{\delta}))^6}{(\lambda_0+\widetilde{\lambda}_0)^2} \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix} \right\|_2^2\right)$, then $\frac{1}{m}\sum_{r=1}^{m}\|\boldsymbol{w}_r(t+1)\|_2^6 \leq 2C_1$.

Lemma 4.4. *Assume that Lemma 4.1 holds, then with probability of at least $1-\delta$, we have*

(i) if $m = \widetilde{\Omega}\left(\left(\log\frac{1}{\delta}\right)^5\right)$ and $m = \widetilde{\Omega}\left(\frac{(\log(\frac{1}{\delta}))^4}{(\lambda_0+\widetilde{\lambda}_0)^2} \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix} \right\|_2^2\right)$, then

$$\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|_2 \leq \eta \cdot \frac{c_0\cdot\sqrt{2C_0}}{4} \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \end{bmatrix} \right\|_2;$$

(ii) if $m = \widetilde{\Omega}\left(\left(\log\frac{1}{\delta}\right)^7\right)$ and $m = \widetilde{\Omega}\left(\frac{(\log(\frac{1}{\delta}))^6}{(\lambda_0+\widetilde{\lambda}_0)^2} \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix} \right\|_2^2\right)$, then

$$\|\boldsymbol{a}(t+1)-\boldsymbol{a}(t)\|_2 \leq \eta \cdot \frac{c_1\cdot\sqrt{2C_1}}{4} \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \end{bmatrix} \right\|_2.$$

Here the universal constants are defined in Lemma 4.1 and Lemma 4.3.

Proof. Directly combining proofs for Lemma 4.1 and Lemma 4.3, the above results can be achieved. □

Theorem 4.5. *For given training samples $\{\boldsymbol{x}_p\}_{p=1}^{n_1}\bigcup\{\tilde{\boldsymbol{x}}_k\}_{k=1}^{n_2}$, if all weights of PINNs are initialized by (9) for all $r\in[m]$, then with a probability of at least $1-\delta$, the gradient descent algorithm satisfies*

$$\mathcal{L}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \leq \left(1-\eta\cdot\frac{\lambda_0+\widetilde{\lambda}_0}{2}\right)^t \cdot \mathcal{L}(\boldsymbol{w}(0),\boldsymbol{a}(0)), \tag{27}$$

for all $t\in\mathbb{N}$, if $m = \widetilde{\Omega}\left(\frac{(n_1+n_2)^2}{\left(\lambda_0+\widetilde{\lambda}_0\right)^2\cdot\left(\min\{\lambda_0,\widetilde{\lambda}_0\}\right)^2\cdot\delta^3}\right)$ and $\eta = \mathcal{O}\left(\lambda_0+\widetilde{\lambda}_0\right) < \frac{2}{\lambda_0+\widetilde{\lambda}_0}$.

Remark 4.6. If operators $\mathcal{D}$ and $\mathcal{B}$ are polynomials of u and its derivatives, then $\widetilde{\boldsymbol{G}}(\boldsymbol{w},\boldsymbol{a})$, $\boldsymbol{G}(\boldsymbol{w},\boldsymbol{a})$ and $\mathcal{L}(\boldsymbol{w},\boldsymbol{a})$ are polynomials of $(\boldsymbol{w},\boldsymbol{a})$. Under initialization methods whose tails decay faster than polynomials (e.g., Gaussian and uniform distributions), we can similarly prove Lemma 4.1-4.4 and Theorem 4.5 for PINNs in solving various kinds of PDEs by some concentration inequalities.

In many applications, people adopt the following two-layer neural networks without the multiplier $\frac{1}{\sqrt{m}}$,

$$\tilde{\phi}(\boldsymbol{x};\boldsymbol{w},\boldsymbol{a}) = \sum_{r=1}^{m} a_r\cdot\sigma(\boldsymbol{w}_r^\top\boldsymbol{y}), \tag{28}$$

initialized by common methods such as HeNormal (He et al., 2015), HeUniform (He et al., 2015), LecunNormal(Klambauer et al., 2017) or LecunUniform(Klambauer et al., 2017) (see Table 1). By Theorem 4.5, the following Corollary 4.7 holds.

Corollary 4.7. *If the weights of PINNs are initialized by $\boldsymbol{w}_r \sim p_1$ and $a_r \sim p_2$ for all $r\in[m]$. For given training samples $\{\boldsymbol{x}_p\}_{p=1}^{n_1}\bigcup\{\tilde{\boldsymbol{x}}_k\}_{k=1}^{n_2}$, let $\widetilde{\boldsymbol{G}}^\infty = \frac{1}{m}\boldsymbol{E}_{\boldsymbol{w}_r\sim p_1, a_r\sim p_2}\widetilde{\boldsymbol{G}}(\boldsymbol{w})$ and $\lambda_0 := \lambda_{\min}(\widetilde{\boldsymbol{G}}^\infty) > 0$. Here p_1 and p_2 are HeNormal , HeUniform, LecunNormal or LecunUniform; see Table 1. If $m = \widetilde{\Omega}\left(\frac{(n_1+n_2)^{2/3}}{\widetilde{\lambda}_0^{4/3}\cdot\delta}\right)$ and $\eta = \mathcal{O}\left(\frac{1}{m}\right) < \frac{2}{m\widetilde{\lambda}_0}$, then with probability of at least $1-\delta$, the gradient descent algorithm satisfies*

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}(t),\boldsymbol{a}(t)) &\leq \left(1-\eta\cdot\frac{m\widetilde{\lambda}_0}{2}\right)^t \cdot \mathcal{L}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ &\leq \left(1-\mathcal{O}(1)\cdot\widetilde{\lambda}_0\right)^t \cdot \mathcal{L}(\boldsymbol{w}(0),\boldsymbol{a}(0)),\end{aligned} \tag{29}$$

for all $t\geq 0$. Note that $\widetilde{\boldsymbol{G}}^\infty$ is independent of m if with the aforementioned initialization.

4.1. Generalization to Linear Second-Order PDEs

We can also extend the main result Theorem 4.5 from the heat equation Equation (4) to the second-order linear PDE (20) as in Section 3.1. Without much effort, we rewrite Lemma 4.1-Lemma 4.4 and Theorem 4.5 for (20). In the proof, additional terms related to M appear in concentration inequalities, so the proof is similar and we omit the details. Specifically, we have that

Table 1. Several popular initialization methods.

Initialization	w_{rj}	a_r
HeUniform	$\text{Unif}\left(\left[-\sqrt{\frac{6}{d+2}}, \sqrt{\frac{6}{d+2}}\right]\right)$	$\text{Unif}\left(\left[-\sqrt{\frac{6}{m}}, \sqrt{\frac{6}{m}}\right]\right)$
HeNormal	$\mathcal{N}\left(0, \frac{2}{d+2}\right)$	$\mathcal{N}\left(0, \frac{2}{m}\right)$
LecunUniform	$\text{Unif}\left(\left[-\sqrt{\frac{3}{d+2}}, \sqrt{\frac{6}{d+2}}\right]\right)$	$\text{Unif}\left(\left[-\sqrt{\frac{3}{m}}, \sqrt{\frac{6}{m}}\right]\right)$
LecunNormal	$\mathcal{N}\left(0, \frac{1}{d+2}\right)$	$\mathcal{N}\left(0, \frac{1}{m}\right)$

Lemma 4.1 holds for (20) with R_w, R_a, $\widetilde{R}_w$ and $\widetilde{R}_a$ magnified M times; Lemma 4.2 holds for (20) with (25) magnified M times; Lemma 4.3 holds for (20) with $m = \widetilde{\Omega}\left(\frac{M^2 \cdot \left(\log(\frac{1}{\delta})\right)^6}{(\lambda_0+\widetilde{\lambda}_0)^2} \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0), \boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0), \boldsymbol{a}(0)) \end{bmatrix} \right\|_2^2\right)$ for both (i) and (ii); for Lemma 4.4, if $m = \widetilde{\Omega}\left(\left(\log \frac{1}{\delta}\right)^7\right)$ and $m = \widetilde{\Omega}\left(\frac{M^2 \cdot \left(\log(\frac{1}{\delta})\right)^6}{(\lambda_0+\widetilde{\lambda}_0)^2} \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0), \boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0), \boldsymbol{a}(0)) \end{bmatrix} \right\|_2^2\right)$, then

$$\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|_2 \leq \eta \cdot \frac{c_0 \cdot \sqrt{2C_0}}{4} \cdot M \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t), \boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t), \boldsymbol{a}(t)) \end{bmatrix} \right\|_2$$

and

$$\|\boldsymbol{a}(t+1)-\boldsymbol{a}(t)\|_2 \leq \eta \cdot \frac{c_1 \cdot \sqrt{2C_1}}{4} \cdot M \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t), \boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t), \boldsymbol{a}(t)) \end{bmatrix} \right\|_2 .$$

Based on the aforementioned generalized lemmas, we can easily develop the following Corollary 4.8 as a generalization of Theorem 4.5.

Corollary 4.8. *For given training samples $\{\boldsymbol{x}_p\}_{p=1}^{n_1} \bigcup \{\tilde{\boldsymbol{x}}_k\}_{k=1}^{n_2}$ from PDE (20), if all weights of PINNs are initialized by (9) for all $r \in [m]$, then with a probability of at least $1-\delta$, we have that the gradient descent algorithm satisfies*

$$\mathcal{L}(\boldsymbol{w}(t), \boldsymbol{a}(t)) \leq \left(1 - \eta \cdot \frac{\lambda_0 + \widetilde{\lambda}_0}{2}\right)^t \cdot \mathcal{L}(\boldsymbol{w}(0), \boldsymbol{a}(0)), \tag{30}$$

for all $t \in \mathbb{N}$, if $m = \widetilde{\Omega}\left(\frac{M^8 \cdot (n_1+n_2)^2}{\left(\lambda_0+\widetilde{\lambda}_0\right)^2 \cdot \left(\min\{\lambda_0, \widetilde{\lambda}_0\}\right)^2 \cdot \delta^3}\right)$ and $\eta = \mathcal{O}\left(\frac{\lambda_0+\widetilde{\lambda}_0}{M}\right) < \frac{2}{\lambda_0+\widetilde{\lambda}_0}$.

Similar to the continuous time analysis, Corollary 4.8 can be easily generalized to the second-order linear elliptic equation and hyperbolic equation.

5. Numerical Experiments

We validate our theoretical results on the 1-D heat equation, and numerical results show the effectiveness of the over-parameterization in training PINNs.

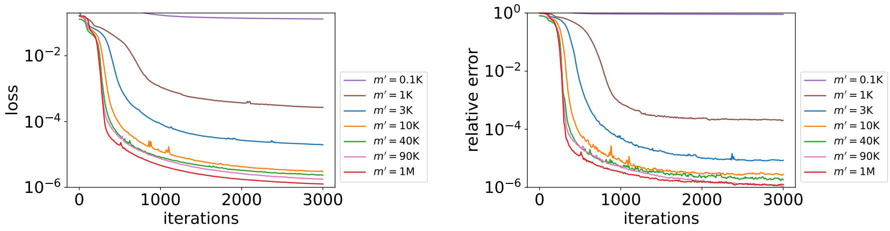
We implement PINNs on the 1-D heat equation, which is given as follows:

$$\begin{aligned} &\frac{\partial u}{\partial t}(t, x) = \frac{\partial^2 u}{\partial x^2}(t, x), \quad (t, x) \in [0, 1] \times [-1, 1], \\ &u(t, -1) = u(t, 1) = 0, \quad t \in [0, 1], \\ &u(0, x) = \sin(\pi x), \quad x \in [-1, 1]. \end{aligned} \tag{31}$$

In practice, we usually use neural networks with multiple layers (e.g., 2-hidden layers) and accelerated gradient descent algorithms (e.g., Adam (Kingma & Ba, 2015)). We uniformly sample 300 interior data points and 100 on each boundary (totally $0.6K$ samples). A neural network with 2 hidden layers and the ReLU^3 activation function is adopted as a surrogate to the solution. We use the widely adopted initialization (e.g., the LecunUniform (Klambauer et al., 2017)) and default hyperparameters for the Adam optimizer. Here, we denote m' the numbers of parameters for the neural network and the relative error is defined as $\text{err} = \sqrt{\frac{\sum_{i=1}^n (\phi(\boldsymbol{x}_i) - y_i)^2}{\sum_{i=1}^n y_i^2}}$, where $\{(t_i, x_i, y_i)\}_{i=1}^n$ are testing samples with $y_i = u(t_i, x_i)$ and $\boldsymbol{x}_i = (t_i, x_i)$. The curves of the loss and the relative error versus iterations for different m' are plotted in Figure 1. Figure 1a (also for Figure 2a) shows that we get lower loss when m' is larger. Although our theory is only applicable for shallow neural networks and the classic gradient descent algorithm, the numerical results are still consistent with our theory that the over-parameterization helps gradient descent find the global optima. Moreover, the generalization error (i.e., the relative error for the prediction) also decreases with the loss, even though the neural network is over-parameterized, as is displayed in Figure 1b and Figure 2b.

6. Conclusion

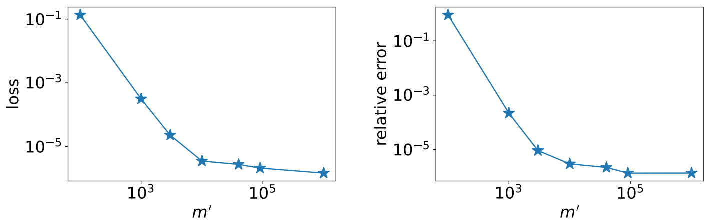
In this paper, we have shown that the gradient flow and the gradient descent find the global optima of the loss function when using two-layer PINNs to solve second-order linear PDEs. It provides theoretical insights into the phenomenon that one can achieve very low empirical loss by gradient descent methods in practical applications. Besides the simple pedagogical example, we further extend our results for a wider class of second-order PDEs and some common ini-



(a) Loss

(b) Relative error.

Figure 1. loss and relative error versus iterations for different parameter sizes m' (1-D heat equation).



(a) Loss

(b) Relative error

Figure 2. Loss and the relative error versus the parameter size m' (1-D heat equation).

tialization methods. There are some future works. Firstly, the extension of our theory to multi-layer neural networks. The main idea could be the positive definiteness of Gram matrices, but details might be more tedious and complicated. Secondly, the generalization of PINNs using the Lipschitzness (Fournier & Guillin, 2015), Rademacher complexity (Bartlett & Mendelson, 2002; E et al., 2020) and Hölder regularization (Shin et al., 2020), and the optimal size of PINNs balancing the convergence and the generalization, which is an unsolved and still open question in the field.

Acknowledgement

We thank all anonymous reviewers and chair members for their helpful comments. Michael K. Ng is supported by HKRGC GRF 12300519, 17201020 and 17300021, HKRGC CRF C1013-21GF and C7004-21GF, and Joint NSFC and RGC N-HKU769/21.

References

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Cai, S., Wang, Z., Wang, S., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks for heat transfer problems. *Journal of Heat Transfer*, 143(6), 2021.

Cai, S., Mao, Z., Wang, Z., Yin, M., and Karniadakis, G. E. Physics-informed neural networks (PINNs) for fluid mechanics: A review. *Acta Mechanica Sinica*, pp. 1–12, 2022.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

Chatterjee, S. Convergence of gradient descent for deep neural networks. *arXiv preprint arXiv:2203.16462*, 2022.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

Dwivedi, V. and Srinivasan, B. Physics informed extreme learning machine (PIELM)–a rapid method for the numerical solution of partial differential equations. *Neurocomputing*, 391:96–118, 2020.

E, W., Ma, C., Wojtowytsch, S., and Wu, L. Towards a mathematical understanding of neural network-based machine learning: What we know and what we don't. *arXiv preprint arXiv:2009.10713*, 2020.

Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

Gu, Y., Yang, H., and Zhou, C. Selectnet: Self-paced learning for high-dimensional partial differential equations. *Journal of Computational Physics*, 441:110444, 2021.

He, J., Li, L., Xu, J., and Zheng, C. ReLU deep neural networks and linear finite elements. *Journal of Computational Mathematics*, 38(3):502–527, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on Imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *International Conference for Learning Representations (ICLR)*, 2015.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. Self-normalizing neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Li, X. and Orabona, F. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 983–992. PMLR, 2019.

Liu, D. C. and Nocedal, J. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

Luo, T. and Yang, H. Two-layer neural networks for partial differential equations: Optimization and generalization theory. *arXiv preprint arXiv:2006.15733*, 2020.

Mao, Z., Jagtap, A. D., and Karniadakis, G. E. Physics-informed neural networks for high-speed flows. *Computer Methods in Applied Mechanics and Engineering*, 360:112789, 2020.

Meng, X., Babaee, H., and Karniadakis, G. E. Multi-fidelity Bayesian neural networks: Algorithms and applications. *Journal of Computational Physics*, 438:110361, 2021.

Nesterov, Y. and Polyak, B. T. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Pang, G., D'Elia, M., Parks, M., and Karniadakis, G. E. nPINNs: nonlocal physics-informed neural networks for a parametrized nonlocal universal Laplacian operator. Algorithms and applications. *Journal of Computational Physics*, 422:109760, 2020.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. 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.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of Adam and beyond. *International Conference for Learning Representations*, 2019.

Shin, Y., Darbon, J., and Karniadakis, G. E. On the convergence of physics informed neural networks for linear second-order elliptic and parabolic type PDEs. *Communications in Computational Physics*, 28(5):2042–2074, 2020.

Siegel, J. W. and Xu, J. High-order approximation rates for shallow neural networks with cosine and ReLUk activation functions. *Applied and Computational Harmonic Analysis*, 58:1–26, 2022.

Sirignano, J. and Spiliopoulos, K. DGM: A deep learning algorithm for solving partial differential equations. *Journal of Computational Physics*, 375:1339–1364, 2018.

Soltanolkotabi, M. Learning ReLUs via gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Soltanolkotabi, M., Javanmard, A., and Lee, J. D. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Soudry, D. and Carmon, Y. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

Wang, S., Yu, X., and Perdikaris, P. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022.

Xie, B., Liang, Y., and Song, L. Diverse neural network learns true target functions. In *The 20th International Conference on Artificial Intelligence and Statistics*, pp. 1216–1224. PMLR, 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

A. Some Preparatory Works

A.1. Computation

After defining the third derivative of the activation function ReLU^3 in Equation (7) and Equation (8), we have

$$\begin{aligned}
\frac{\partial \phi}{\partial x_i}(\boldsymbol{x}_p; \boldsymbol{w}, \boldsymbol{a}) &= \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \cdot \sigma'(\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot w_{ri} \\
&= \frac{1}{\sqrt{m}} \sum_{r=1}^{m} 3a_r \cdot (\boldsymbol{w}_r^\top \boldsymbol{y}_p)^2 \cdot w_{ri} \cdot \mathbb{I}(\boldsymbol{w}_r^\top \boldsymbol{y}_p > 0), \\
\frac{\partial^2 \phi}{\partial x_i^2}(\boldsymbol{x}_p; \boldsymbol{w}, \boldsymbol{a}) &= \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \cdot \sigma''(\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot w_{ri}^2 \\
&= \frac{1}{\sqrt{m}} \sum_{r=1}^{m} 6a_r \cdot (\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot w_{ri}^2 \cdot \mathbb{I}(\boldsymbol{w}_r^\top \boldsymbol{y}_p > 0), \\
\sum_{i=1}^{d} \frac{\partial^2 \phi}{\partial x_i^2}(\boldsymbol{x}_p; \boldsymbol{w}, \boldsymbol{a}) &= \frac{1}{\sqrt{m}} \sum_{r=1}^{m} 6a_r \cdot (\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot \left(\sum_{i=1}^{d} w_{ri}^2 \right) \cdot \mathbb{I}(\boldsymbol{w}_r^\top \boldsymbol{y}_p > 0).
\end{aligned} \tag{32}$$

Moreover,

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{w}_r} \left(\frac{\partial \phi}{\partial x_0}(\boldsymbol{x}_p; \boldsymbol{w}, \boldsymbol{a}) \right) \\
=&\frac{1}{\sqrt{m}} \cdot a_r \cdot \mathbb{I}(\boldsymbol{w}_r^\top \boldsymbol{y}_p > 0) \cdot \left(\sigma'(\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot \boldsymbol{e}_0 + w_{r0} \cdot \sigma''(\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot \boldsymbol{y}_p \right) \\
=&\frac{1}{\sqrt{m}} \cdot a_r \cdot \mathbb{I}(\boldsymbol{w}_r^\top \boldsymbol{y}_p > 0) \cdot \left(3(\boldsymbol{w}_r^\top \boldsymbol{y}_p)^2 \cdot \boldsymbol{e}_0 + 6w_{r0} \cdot (\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot \boldsymbol{y}_p \right),
\end{aligned} \tag{33}$$

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{w}_r} \left(\frac{\partial^2 \phi}{\partial x_i^2}(\boldsymbol{x}_p; \boldsymbol{w}, \boldsymbol{a}) \right) \\
=&\frac{1}{\sqrt{m}} \cdot a_r \cdot \mathbb{I}(\boldsymbol{w}_r^\top \boldsymbol{y}_p > 0) \cdot \left(2w_{ri} \cdot \sigma''(\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot \boldsymbol{e}_i + \sigma'''(\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot w_{ri}^2 \cdot \boldsymbol{y}_p \right) \\
=&\frac{1}{\sqrt{m}} \cdot a_r \cdot \mathbb{I}(\boldsymbol{w}_r^\top \boldsymbol{y}_p > 0) \cdot \left(12w_{ri} \cdot (\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot \boldsymbol{e}_i + 6w_{ri}^2 \cdot \boldsymbol{y}_p \right), \quad i \neq 0,
\end{aligned} \tag{34}$$

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{w}_r} \phi(\tilde{\boldsymbol{x}}_k; \boldsymbol{w}, \boldsymbol{a}) &= \frac{1}{\sqrt{m}} \cdot a_r \cdot \mathbb{I}(\boldsymbol{w}_r^\top \tilde{\boldsymbol{y}}_k > 0) \cdot \sigma'(\boldsymbol{w}_r^\top \tilde{\boldsymbol{y}}_k) \cdot \tilde{\boldsymbol{y}}_k \\
&= \frac{1}{\sqrt{m}} \cdot a_r \cdot \mathbb{I}(\boldsymbol{w}_r^\top \tilde{\boldsymbol{y}}_k > 0) \cdot \left(3(\boldsymbol{w}_r^\top \tilde{\boldsymbol{y}}_k)^2 \cdot \tilde{\boldsymbol{y}}_k \right),
\end{aligned} \tag{35}$$

$$\frac{\partial}{\partial a_r} \left(\frac{\partial \phi}{\partial x_0}(\boldsymbol{x}_p; \boldsymbol{w}, \boldsymbol{a}) \right) = \frac{3}{\sqrt{m}} \cdot (\boldsymbol{w}_r^\top \boldsymbol{y}_p)^2 \cdot w_{r0} \cdot \mathbb{I}(\boldsymbol{w}_r^\top \boldsymbol{y}_p > 0), \tag{36}$$

$$\frac{\partial}{\partial a_r} \left(\frac{\partial^2 \phi}{\partial x_i^2}(\boldsymbol{x}_p; \boldsymbol{w}, \boldsymbol{a}) \right) = \frac{6}{\sqrt{m}} \cdot (\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot w_{ri}^2 \cdot \mathbb{I}(\boldsymbol{w}_r^\top \boldsymbol{y}_p > 0), \tag{37}$$

and

$$\frac{\partial}{\partial a_r} \phi(\tilde{\boldsymbol{x}}_k; \boldsymbol{w}, \boldsymbol{a}) = \frac{1}{\sqrt{m}} \cdot (\boldsymbol{w}_r^\top \tilde{\boldsymbol{y}}_k)^3 \cdot \mathbb{I}(\boldsymbol{w}_r^\top \tilde{\boldsymbol{y}}_k > 0), \tag{38}$$

where $\boldsymbol{e}_i \in \mathbb{R}^{d+2}$ is the base vector whose i-th element is 1 and others are zero. According to (33)-(38), the objective is first-order differentiable almost everywhere but is not second-order differentiable. Therefore, general convergence analysis that relies on smoothness may not make sense in such a setting.

A.2. The Derivation of Gradient Flow

We continue from the beginning of Section 3 that

$$
\begin{aligned}
\frac{ds_p(\boldsymbol{w}(t),\boldsymbol{a}(t))}{dt} &= \sum_{r=1}^{m}\left\langle \frac{\partial s_p(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}_r}, \frac{d\boldsymbol{w}_r(t)}{dt}\right\rangle + \sum_{r=1}^{m}\frac{\partial s_p(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial a_r}\cdot\frac{da_r(t)}{dt}\\
=& -\sum_{q=1}^{n_1} s_q(\boldsymbol{w}(t),\boldsymbol{a}(t))\cdot\sum_{r=1}^{m}\left\langle \frac{\partial s_p(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}_r}\right\rangle\\
&-\sum_{l=1}^{n_2} h_l(\boldsymbol{w}(t),\boldsymbol{a}(t))\cdot\sum_{r=1}^{m}\left\langle \frac{\partial s_p(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}_r}, \frac{\partial h_l(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}_r}\right\rangle\\
&-\sum_{q=1}^{n_1} s_q(\boldsymbol{w}(t),\boldsymbol{a}(t))\cdot\sum_{r=1}^{m}\frac{\partial s_p(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial a_r}\cdot\frac{\partial s_q(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial a_r}\\
&-\sum_{l=1}^{n_2} h_l(\boldsymbol{w}(t),\boldsymbol{a}(t))\cdot\sum_{r=1}^{m}\frac{\partial s_p(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial a_r}\cdot\frac{\partial h_l(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial a_r},
\end{aligned}
\tag{39}
$$

and

$$
\begin{aligned}
\frac{dh_k(\boldsymbol{w}(t))}{dt} &= \sum_{r=1}^{m}\left\langle \frac{\partial h_k(\boldsymbol{w}(t))}{\partial \boldsymbol{w}_r}, \frac{d\boldsymbol{w}_r(t)}{dt}\right\rangle + \sum_{r=1}^{m}\frac{\partial h_k(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial a_r}\cdot\frac{da_r(t)}{dt}\\
=& -\sum_{q=1}^{n_1} s_q(\boldsymbol{w}(t),\boldsymbol{a}(t))\cdot\sum_{r=1}^{m}\left\langle \frac{\partial h_k(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}_r}\right\rangle\\
&-\sum_{l=1}^{n_2} h_l(\boldsymbol{w}(t),\boldsymbol{a}(t))\cdot\sum_{r=1}^{m}\left\langle \frac{\partial h_k(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}_r}, \frac{\partial h_l(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}_r}\right\rangle\\
&-\sum_{q=1}^{n_1} s_q(\boldsymbol{w}(t),\boldsymbol{a}(t))\cdot\sum_{r=1}^{m}\frac{\partial h_k(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial a_r}\cdot\frac{\partial s_q(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial a_r}\\
&-\sum_{l=1}^{n_2} h_l(\boldsymbol{w}(t),\boldsymbol{a}(t))\cdot\sum_{r=1}^{m}\frac{\partial h_k(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial a_r}\cdot\frac{\partial h_l(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial a_r}.
\end{aligned}
\tag{40}
$$

Then, we have

$$
\begin{aligned}
&\frac{d}{dt}\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t))\\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t))\end{bmatrix}\\
=& -\begin{bmatrix} \boldsymbol{S}(\boldsymbol{w}(t),\boldsymbol{a}(t))+\widetilde{\boldsymbol{S}}(\boldsymbol{w}(t),\boldsymbol{a}(t)) & \boldsymbol{Q}(\boldsymbol{w}(t),\boldsymbol{a}(t))+\widetilde{\boldsymbol{Q}}(\boldsymbol{w}(t),\boldsymbol{a}(t))\\ \boldsymbol{Q}(\boldsymbol{w}(t),\boldsymbol{a}(t))^\top+\widetilde{\boldsymbol{Q}}(\boldsymbol{w}(t),\boldsymbol{a}(t))^\top & \boldsymbol{H}(\boldsymbol{w}(t),\boldsymbol{a}(t))+\widetilde{\boldsymbol{H}}(\boldsymbol{w}(t),\boldsymbol{a}(t))\end{bmatrix}\\
&\cdot\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t))\\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t))\end{bmatrix}\\
=& -\Big(\boldsymbol{G}(\boldsymbol{w}(t),\boldsymbol{a}(t))+\widetilde{\boldsymbol{G}}(\boldsymbol{w}(t),\boldsymbol{a}(t))\Big)\cdot\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t))\\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t))\end{bmatrix},
\end{aligned}
\tag{41}
$$

where $\boldsymbol{S}(\boldsymbol{w},\boldsymbol{a}) = [S_{pq}(\boldsymbol{w},\boldsymbol{a})]$, $\widetilde{\boldsymbol{S}}(\boldsymbol{w},\boldsymbol{a}) = \left[\widetilde{S}_{pq}(\boldsymbol{w},\boldsymbol{a})\right]$, $\boldsymbol{Q}(\boldsymbol{w},\boldsymbol{a}) = [Q_{pl}(\boldsymbol{w},\boldsymbol{a})]$, $\widetilde{\boldsymbol{Q}}(\boldsymbol{w},\boldsymbol{a}) = \left[\widetilde{Q}_{pl}(\boldsymbol{w},\boldsymbol{a})\right]$, $\boldsymbol{H}(\boldsymbol{w}) = [H_{kl}(\boldsymbol{w})]$ and $\widetilde{\boldsymbol{H}}(\boldsymbol{w},\boldsymbol{a}) = \left[\widetilde{H}_{kl}(\boldsymbol{w},\boldsymbol{a})\right]$ are defined as

$$\begin{aligned}
S_{pq}(\boldsymbol{w},\boldsymbol{a}) &= \sum_{r=1}^{m} \left\langle \frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r} \right\rangle = \left\langle \frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}}, \frac{\partial s_q(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}} \right\rangle, \\
\widetilde{S}_{pq}(\boldsymbol{w},\boldsymbol{a}) &= \sum_{r=1}^{m} \frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial a_r} \cdot \frac{\partial s_q(\boldsymbol{w},\boldsymbol{a})}{\partial a_r} = \left\langle \frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{a}}, \frac{\partial s_q(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{a}} \right\rangle, \\
Q_{pl}(\boldsymbol{w},\boldsymbol{a}) &= \sum_{r=1}^{m} \left\langle \frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r}, \frac{\partial h_l(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r} \right\rangle = \left\langle \frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}}, \frac{\partial h_l(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}} \right\rangle, \\
\widetilde{Q}_{pl}(\boldsymbol{w},\boldsymbol{a}) &= \sum_{r=1}^{m} \frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial a_r} \cdot \frac{\partial h_l(\boldsymbol{w},\boldsymbol{a})}{\partial a_r} = \left\langle \frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{a}}, \frac{\partial h_l(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{a}} \right\rangle, \\
H_{kl}(\boldsymbol{w},\boldsymbol{a}) &= \sum_{r=1}^{m} \left\langle \frac{\partial h_k(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r}, \frac{\partial h_l(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r} \right\rangle = \left\langle \frac{\partial h_k(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}}, \frac{\partial h_l(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}} \right\rangle, \\
\widetilde{H}_{kl}(\boldsymbol{w},\boldsymbol{a}) &= \sum_{r=1}^{m} \frac{\partial h_k(\boldsymbol{w},\boldsymbol{a})}{\partial a_r} \cdot \frac{\partial h_l(\boldsymbol{w},\boldsymbol{a})}{\partial a_r} = \left\langle \frac{\partial h_k(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{a}}, \frac{\partial h_l(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{a}} \right\rangle.
\end{aligned}$$

Here, $\boldsymbol{G}(\boldsymbol{w},\boldsymbol{a})$ and $\widetilde{\boldsymbol{G}}(\boldsymbol{w},\boldsymbol{a})$ are the Gram matrices for the dynamics, defined as

$$\boldsymbol{G}(\boldsymbol{w},\boldsymbol{a}) = \boldsymbol{D}^\top \cdot \boldsymbol{D} \tag{42}$$

and

$$\widetilde{\boldsymbol{G}}(\boldsymbol{w},\boldsymbol{a}) = \widetilde{\boldsymbol{D}}^\top \cdot \widetilde{\boldsymbol{D}} \tag{43}$$

respectively, where

$$\boldsymbol{D} = \begin{bmatrix} \frac{\partial s_1(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}} & \cdots & \frac{\partial s_{n_1}(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}} & \frac{\partial h_1(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}} & \cdots & \frac{\partial h_{n_2}(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}} \end{bmatrix}$$

and

$$\widetilde{\boldsymbol{D}} = \begin{bmatrix} \frac{\partial s_1(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{a}} & \cdots & \frac{\partial s_{n_1}(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{a}} & \frac{\partial h_1(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{a}} & \cdots & \frac{\partial h_{n_2}(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{a}} \end{bmatrix}.$$

Note that $\widetilde{\boldsymbol{G}}(\boldsymbol{w},\boldsymbol{a})$ is independent of $\boldsymbol{a}$, but we keep the variable $\boldsymbol{a}$ here for the consistent symbol format with $\boldsymbol{G}(\boldsymbol{w},\boldsymbol{a})$.

B. Technical Proofs for Section 3

B.1. Proof of Lemma 3.2

Let

$$\varphi(\boldsymbol{x};\boldsymbol{w}) := \frac{\partial}{\partial \boldsymbol{a}}\left(\sqrt{\frac{1}{n_1}}\left(\frac{\partial \phi}{\partial x_0}(\boldsymbol{x};\boldsymbol{w},\boldsymbol{a}) - \sum_{i=1}^{d}\frac{\partial^2 \phi}{\partial x_i^2}(\boldsymbol{x};\boldsymbol{w},\boldsymbol{a}) - f(\boldsymbol{x})\right)\right), \tag{44}$$

and $\varphi(\boldsymbol{x};\boldsymbol{w}) := [\varphi_1(\boldsymbol{x};\boldsymbol{w}_1)\ \cdots\ \varphi_m(\boldsymbol{x};\boldsymbol{w}_m)]^\top$. Using (5), we obtain

$$\varphi_r(\boldsymbol{x};\boldsymbol{w}_r) = \frac{3}{\sqrt{mn_1}}\cdot \mathbb{I}(\boldsymbol{w}_r^\top \boldsymbol{y} > 0)\cdot\left((\boldsymbol{w}_r^\top \boldsymbol{y})\cdot\left((\boldsymbol{w}_r^\top \boldsymbol{y})\cdot w_{r0} - 2\sum_{i=1}^{d} w_{ri}^2\right)\right). \tag{45}$$

Similarly, let

$$\psi(\boldsymbol{x};\boldsymbol{w}) = [\psi_1(\boldsymbol{x};\boldsymbol{w}_1)\ \cdots\ \psi_m(\boldsymbol{x};\boldsymbol{w}_m)]^\top := \frac{\partial}{\partial \boldsymbol{a}}\left(\sqrt{\frac{\nu}{n_2}}\left(\phi(\boldsymbol{x};\boldsymbol{w},\boldsymbol{a}) - g(\boldsymbol{x})\right)\right),$$

where

$$\psi_r(\boldsymbol{x};\boldsymbol{w}_r) = \sqrt{\frac{\nu}{mn_2}}\cdot \mathbb{I}(\boldsymbol{w}_r^\top \boldsymbol{y} > 0)\cdot(\boldsymbol{w}_r^\top \boldsymbol{y})^3. \tag{46}$$

To prove the positive definiteness of the matrix $\widetilde{\boldsymbol{G}}^\infty$, it suffices to show that vectors $\varphi(\boldsymbol{x}_1;\boldsymbol{w}),\ldots,\varphi(\boldsymbol{x}_{n_1};\boldsymbol{w})$, $\psi(\tilde{\boldsymbol{x}}_1;\boldsymbol{w}),\ldots,\psi(\tilde{\boldsymbol{x}}_{n_2};\boldsymbol{w})$ are linearly independent. Suppose that there exist some constants $\alpha_1,\ldots,\alpha_{n_1}$ and $\beta_1,\ldots,\beta_{n_2}$ such that

$$\alpha_1\varphi(\boldsymbol{x}_1;\boldsymbol{w}) + \cdots + \alpha_{n_1}\varphi(\boldsymbol{x}_{n_1};\boldsymbol{w}) + \beta_1\psi(\tilde{\boldsymbol{x}}_1;\boldsymbol{w}) + \cdots + \beta_{n_2}\psi(\tilde{\boldsymbol{x}}_{n_2};\boldsymbol{w}) = 0,$$

for almost all $\boldsymbol{w}\in\mathbb{R}^{m(d+2)}$. Denote $I_p = \{\tilde{\boldsymbol{w}}\in\mathbb{R}^{d+2} : \tilde{\boldsymbol{w}}^\top \boldsymbol{y}_p = 0\}$ and $J_k = \{\tilde{\boldsymbol{w}}\in\mathbb{R}^{d+2} : \tilde{\boldsymbol{w}}^\top \tilde{\boldsymbol{y}}_k = 0\}$, for $p = 1,\ldots,n_1$ and $k = 1,\ldots,n_2$. Since all samples in $\{\boldsymbol{y}_p\}_{p=1}^{n_1}\bigcup\{\tilde{\boldsymbol{y}}_k\}_{k=1}^{n_2}$ are not parallel by Proposition 3.1, then

$$I_p \not\subset \left(\bigcup_{q\neq p} I_q\right)\bigcup\left(\bigcup_{l} J_l\right),$$

which implies that there exists $\boldsymbol{z}\in I_p$ such that $\boldsymbol{z}\notin\left(\bigcup_{q\neq p} I_q\right)\bigcup\left(\bigcup_l J_l\right)$ and $z_i\neq 0$ for all $0\le i\le d$. Because sets I_q and J_l are closed, there exists a small enough radius $\gamma_0>0$ such that $\mathcal{B}(\boldsymbol{z},\gamma_0)\bigcap I_q=\emptyset$ and $\mathcal{B}(\boldsymbol{z},\gamma_0)\bigcap J_l=\emptyset$ for all $q\neq p$ and l. Moreover, $\varphi_r(\boldsymbol{x}_q;\cdot)$ is continuous in $\mathcal{B}(\boldsymbol{z},\gamma_0)$ for all $q\neq p$. For any $\gamma\le\gamma_0$, we define $\mathcal{B}_\gamma^+ = \mathcal{B}(\boldsymbol{z},\gamma)\bigcap\{\tilde{\boldsymbol{w}}\in\mathbb{R}^{d+2} : \tilde{\boldsymbol{w}}^\top \boldsymbol{y}_p > 0\}$ and $\mathcal{B}_\gamma^- = \mathcal{B}(\boldsymbol{z},\gamma)\bigcap\{\tilde{\boldsymbol{w}}\in\mathbb{R}^{d+2} : \tilde{\boldsymbol{w}}^\top \boldsymbol{y}_p < 0\}$. Note that $\varphi_r(\boldsymbol{x}_q;\boldsymbol{w}_r)$ and $\varphi_r(\tilde{\boldsymbol{x}}_k;\boldsymbol{w}_r)$ are polynomials of $\boldsymbol{w}_r$ (thus are $\mathcal{C}^\infty$ smooth) on $\mathcal{B}_\gamma^+$ and $\mathcal{B}_\gamma^-$. Then, using the Lebesgue differentiation theorem, we have

$$\lim_{\gamma\to 0^+}\frac{1}{\mu(\mathcal{B}_\gamma^+)}\int_{\mathcal{B}_\gamma^+}\frac{\partial}{\partial\tilde{\boldsymbol{z}}}\varphi_r(\boldsymbol{x}_q;\tilde{\boldsymbol{z}})\mathrm{d}\tilde{\boldsymbol{z}} - \frac{1}{\mu(\mathcal{B}_\gamma^-)}\int_{\mathcal{B}_\gamma^-}\frac{\partial}{\partial\tilde{\boldsymbol{z}}}\varphi_r(\boldsymbol{x}_q;\tilde{\boldsymbol{z}})\mathrm{d}\tilde{\boldsymbol{z}} = \boldsymbol{0},\ \text{for } q\neq p, \tag{47}$$

$$\lim_{\gamma\to 0^+}\frac{1}{\mu(\mathcal{B}_\gamma^+)}\int_{\mathcal{B}_\gamma^+}\frac{\partial}{\partial\tilde{\boldsymbol{z}}}\psi_r(\tilde{\boldsymbol{x}}_l;\tilde{\boldsymbol{z}})\mathrm{d}\tilde{\boldsymbol{z}} - \frac{1}{\mu(\mathcal{B}_\gamma^-)}\int_{\mathcal{B}_\gamma^-}\frac{\partial}{\partial\tilde{\boldsymbol{z}}}\psi_r(\tilde{\boldsymbol{x}}_l;\tilde{\boldsymbol{z}})\mathrm{d}\tilde{\boldsymbol{z}} = \boldsymbol{0}, \tag{48}$$

$$\begin{aligned}
&\lim_{\gamma\to 0^+}\frac{1}{\mu(\mathcal{B}_\gamma^+)}\int_{\mathcal{B}_\gamma^+}\frac{\partial}{\partial\tilde{\boldsymbol{z}}}\varphi_r(\boldsymbol{x}_p;\tilde{\boldsymbol{z}})\mathrm{d}\tilde{\boldsymbol{z}} \qquad (49)\\
= &\lim_{\gamma\to 0^+}\frac{1}{\mu(\mathcal{B}_\gamma^+)}\int_{\mathcal{B}_\gamma^+}\frac{3}{\sqrt{mn_1}}\cdot\frac{\partial}{\partial\tilde{\boldsymbol{z}}}\left((\tilde{\boldsymbol{z}}^\top \boldsymbol{y}_p)\cdot\left((\tilde{\boldsymbol{z}}^\top \boldsymbol{y}_p)\cdot\tilde{z}_0 - 2\sum_{i=1}^{d}\tilde{z}_i^2\right)\right)\mathrm{d}\tilde{\boldsymbol{z}}\\
= &\frac{3}{\sqrt{mn_1}}\cdot\left(-2\sum_{i=1}^{d} z_i^2\right)\cdot\boldsymbol{y}_p,
\end{aligned}$$

and

$$\lim_{\gamma\to 0^+}\frac{1}{\mu(\mathcal{B}_\gamma^-)}\int_{\mathcal{B}_\gamma^-}\frac{\partial}{\partial\tilde{\boldsymbol{z}}}\varphi_r(\boldsymbol{x}_p)(\tilde{\boldsymbol{z}})d\tilde{\boldsymbol{z}} = \lim_{\gamma\to 0^+}\frac{1}{\mu(\mathcal{B}_\gamma^-)}\int_{\mathcal{B}_\gamma^-}\frac{\partial}{\partial\tilde{\boldsymbol{z}}}0\ \ d\tilde{\boldsymbol{z}} = \boldsymbol{0}. \tag{50}$$

Therefore,

$$\begin{aligned}\mathbf{0} \;&=\; \lim_{\gamma\to 0^+}\frac{1}{\mu(\mathcal{B}_\gamma^+)}\int_{\mathcal{B}_\gamma^+}\sum_{q=1}^{n_1}\alpha_q\cdot\frac{\partial}{\partial\tilde{\boldsymbol{z}}}\varphi_r(\boldsymbol{x}_q;\tilde{\boldsymbol{z}})+\sum_{l=1}^{n_2}\beta_l\cdot\frac{\partial}{\partial\tilde{\boldsymbol{z}}}\psi_r(\tilde{\boldsymbol{x}}_l;\tilde{\boldsymbol{z}})\mathrm{d}\tilde{\boldsymbol{z}}-\\ &\quad\;\lim_{\gamma\to 0^-}\frac{1}{\mu(\mathcal{B}_\gamma^-)}\int_{\mathcal{B}_\gamma^+}\sum_{q=1}^{n_1}\alpha_q\cdot\frac{\partial}{\partial\tilde{\boldsymbol{z}}}\varphi_r(\boldsymbol{x}_q;\tilde{\boldsymbol{z}})+\sum_{l=1}^{n_2}\beta_l\cdot\frac{\partial}{\partial\tilde{\boldsymbol{z}}}\psi_r(\tilde{\boldsymbol{x}}_l;\tilde{\boldsymbol{z}})\mathrm{d}\tilde{\boldsymbol{z}} \\ &=\; \alpha_p\cdot\lim_{\gamma\to 0^+}\frac{1}{\mu(\mathcal{B}_\gamma^+)}\int_{\mathcal{B}_\gamma^+}\frac{\partial}{\partial\tilde{\boldsymbol{z}}}\varphi_r(\boldsymbol{x}_p;\tilde{\boldsymbol{z}})\mathrm{d}\tilde{\boldsymbol{z}}=\alpha_p\cdot\frac{3}{\sqrt{mn_1}}\cdot\left(-2\sum_{i=1}^{d}z_i^2\right)\cdot\boldsymbol{y}_p,\end{aligned} \tag{51}$$

which implies that $\alpha_p=0$ for all $p\in[n_1]$ since $\sum_{i=1}^d z_i^2\neq 0$.

On the other hand, we can prove $\beta_k=0$ for all $k\in[n_2]$ by similar argument, where the terms $\int_{\mathcal{B}_\gamma^+}\frac{\partial}{\partial\tilde{\boldsymbol{z}}}\varphi_r(\boldsymbol{x};\tilde{\boldsymbol{z}})\mathrm{d}\tilde{\boldsymbol{z}}$ in (47)-(50) are replaced with $\int_{\mathcal{B}_\gamma^+}\frac{\partial^3}{\partial\tilde{z}_i^3}\psi_r(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{z}})\mathrm{d}\tilde{\boldsymbol{z}}$. Note that $\widetilde{\boldsymbol{G}}^\infty$ is independent of m, since $\boldsymbol{w}_r$ are independent for all $r\in[m]$.

B.2. Proof of Lemma 3.5

Observe that

$$\widetilde{S}_{pq}(\boldsymbol{w}(0))=\sum_{r=1}^m\frac{\partial s_p(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial a_r}\cdot\frac{\partial s_q(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial a_r},$$

with $\boldsymbol{w}_r(0)\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_{d+2})$, for $r\in[m]$. We have

$$\begin{aligned}\frac{\partial s_p(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial a_r}&=\frac{3}{\sqrt{mn_1}}\cdot\mathbb{I}(\boldsymbol{w}_r(0)^\top\boldsymbol{y}_p>0)\\ &\quad\cdot\left((\boldsymbol{w}_r(0)^\top\boldsymbol{y}_p)\cdot\left((\boldsymbol{w}_r(0)^\top\boldsymbol{y}_p)\cdot w_{r0}(0)-2\sum_{i=1}^d w_{ri}(0)^2\right)\right),\end{aligned} \tag{52}$$

and

$$\left|\frac{\partial s_p(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial a_r}\right|\leq\frac{9}{\sqrt{mn_1}}\cdot\|\boldsymbol{w}_r(0)\|_2^3. \tag{53}$$

Let the random variable X_r be defined as

$$X_r=m\cdot\frac{\partial s_p(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial a_r}\cdot\frac{\partial s_q(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial a_r},$$

then

$$\widetilde{S}_{pq}(\boldsymbol{w}(0),\boldsymbol{a}(0))-\widetilde{S}_{pq}^\infty=\frac{1}{m}\sum_{r=1}^m X_r-\mathbb{E}X_r,$$

and

$$|X_r|\leq\frac{81}{n_1}\cdot\|\boldsymbol{w}_r(0)\|_2^6,$$

where the expectation is taken over $\boldsymbol{w}_r(0)\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_{d+2})$. Similarly, we have

$$\frac{\partial h_k(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial a_r}=\sqrt{\frac{\nu}{mn_2}}\cdot\mathbb{I}(\boldsymbol{w}_r(0)^\top\tilde{\boldsymbol{y}}_k>0)\cdot(\boldsymbol{w}_r(0)^\top\tilde{\boldsymbol{y}}_k)^3, \tag{54}$$

and

$$\left|\frac{\partial h_k(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial a_r}\right|\leq\sqrt{\frac{\nu}{mn_2}}\cdot\|\boldsymbol{w}_r(0)\|_2^3. \tag{55}$$

Let the random variable Y_r be defined as

$$Y_r=m\cdot\frac{\partial s_p(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial a_r}\cdot\frac{\partial h_l(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial a_r},$$

then

$$\widetilde{Q}_{pl}(\boldsymbol{w}(0), \boldsymbol{a}(0)) - \widetilde{Q}_{pl}^{\infty} = \frac{1}{m}\sum_{r=1}^{m} Y_r - \mathbb{E}Y_r,$$

and

$$|Y_r| \leq 9\sqrt{\frac{\nu}{n_1 n_2}} \cdot \|\boldsymbol{w}_r(0)\|_2^6,$$

where the expectation is taken over $\boldsymbol{w}_r(0) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d+2})$. Let the random variable Z_r be defined as

$$Z_r = m \cdot \frac{\partial h_k(\boldsymbol{w}(0), \boldsymbol{a}(0))}{\partial a_r}, \frac{\partial h_l(\boldsymbol{w}(0), \boldsymbol{a}(0))}{\partial a_r},$$

then

$$\widetilde{H}_{kl}(\boldsymbol{w}(0), \boldsymbol{a}(0)) - \widetilde{H}_{kl}^{\infty} = \frac{1}{m}\sum_{r=1}^{m} Z_r - \mathbb{E}Z_r,$$

and

$$|Z_r| \leq \frac{\nu}{n_2} \cdot \|\boldsymbol{w}_r(0)\|_2^6,$$

where the expectation is taken over $\boldsymbol{w}_r(0) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d+2})$. Then there exists a universal positive constant c_0 such that

$$\max\{|X_r|, |Y_r|, |Z_r|\} \leq c_0 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot \|\boldsymbol{w}_r(0)\|_2^6.$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left(\max\{|X_r|, |Y_r|, |Z_r|\} \geq R'\right) &\leq \mathbb{P}\left(c_0 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \cdot \|\boldsymbol{w}_r(0)\|_2^6 \geq R'\right) \\
&= \mathbb{P}\left(\|\boldsymbol{w}_r(0)\|_2^2 \geq \left(\frac{R' \cdot n_1 n_2}{c_0(n_1+n_2)}\right)^{1/3}\right) \\
&\leq (d+2) \cdot \mathbb{P}_{z\sim\mathcal{N}(0,1)}\left(|z| \geq \left(\frac{R' \cdot n_1 n_2}{c_0 (d+2)^3 \cdot (n_1+n_2)}\right)^{1/6}\right) \\
&\leq (d+2) \cdot \exp\left(-\frac{1}{2} \cdot \left(\frac{R' \cdot n_1 n_2}{c_0 (d+2)^3 \cdot (n_1+n_2)}\right)^{1/3}\right).
\end{aligned}$$

Then, with a probability of at least $1 - \frac{\delta}{2m}$,

$$\max\{|X_r|, |Y_r|, |Z_r|\} \leq R' =: 4c_0(d+2)^3 \cdot \frac{n_1+n_2}{n_1 n_2} \cdot \left(\log\frac{2m(d+2)}{\delta}\right)^3.$$

Furthermore, with probability of at least $1 - \frac{\delta}{2}$,

$$\max\{|X_r|, |Y_r|, |Z_r|\} \leq R' =: 4c_0(d+2)^3 \cdot \frac{n_1+n_2}{n_1 n_2} \cdot \left(\log\frac{2m(d+2)}{\delta}\right)^3,$$

for all $r \in [m]$. By the Hoeffding's inequality, we have

$$\begin{aligned}
\mathbb{P}\left(\left|\widetilde{S}_{pq}(\boldsymbol{w}(0), \boldsymbol{a}(0)) - \widetilde{S}_{pq}^{\infty}\right| > \epsilon\right) &= \mathbb{P}\left(\left|\frac{1}{m}\sum_{r=1}^{m} X_r - \mathbb{E}X_r\right| > \epsilon\right) \leq 2\exp\left(-\frac{m\cdot\epsilon^2}{2R'^2}\right), \\
\mathbb{P}\left(\left|\widetilde{Q}_{pl}(\boldsymbol{w}(0), \boldsymbol{a}(0)) - \widetilde{Q}_{pl}^{\infty}\right| > \epsilon\right) &= \mathbb{P}\left(\left|\frac{1}{m}\sum_{r=1}^{m} Y_r - \mathbb{E}Y_r\right| > \epsilon\right) \leq 2\exp\left(-\frac{m\cdot\epsilon^2}{2R'^2}\right), \qquad (56) \\
\mathbb{P}\left(\left|\widetilde{H}_{kl}(\boldsymbol{w}(0), \boldsymbol{a}(0)) - \widetilde{H}_{kl}^{\infty}\right| > \epsilon\right) &= \mathbb{P}\left(\left|\frac{1}{m}\sum_{r=1}^{m} Z_r - \mathbb{E}Z_r\right| > \epsilon\right) \leq 2\exp\left(-\frac{m\cdot\epsilon^2}{2R'^2}\right),
\end{aligned}$$

and therefore, with probability of at least $1 - \frac{\delta}{2(n_1+n_2)^2}$,

$$\left|\widetilde{S}_{pq}(\boldsymbol{w}(0)) - \widetilde{S}_{pq}^{\infty}\right| = \left|\frac{1}{m}\sum_{r=1}^{m} X_r - \mathbb{E}X_r\right| \le \sqrt{\frac{2R'^2}{m} \cdot \log\frac{4(n_1+n_2)^2}{\delta}}, \tag{57}$$

$$\left|\widetilde{Q}_{pl}(\boldsymbol{w}(0)) - \widetilde{Q}_{pl}^{\infty}\right| = \left|\frac{1}{m}\sum_{r=1}^{m} Y_r - \mathbb{E}Y_r\right| \le \sqrt{\frac{2R'^2}{m} \cdot \log\frac{4(n_1+n_2)^2}{\delta}}, \tag{58}$$

and

$$\left|\widetilde{H}_{kl}(\boldsymbol{w}(0)) - \widetilde{H}_{kl}^{\infty}\right| = \left|\frac{1}{m}\sum_{r=1}^{m} Z_r - \mathbb{E}Z_r\right| \le \sqrt{\frac{2R'^2}{m} \cdot \log\frac{4(n_1+n_2)^2}{\delta}}. \tag{59}$$

Therefore,

$$\begin{aligned}\sum_{p,q=1}^{n_1}\left|\widetilde{S}_{pq}(\boldsymbol{w}(0),\boldsymbol{a}(0)) - \widetilde{S}_{pq}^{\infty}\right| + 2\cdot\sum_{p=1}^{n_1}\sum_{l=1}^{n_2}\left|\widetilde{Q}_{pl}(\boldsymbol{w}(0),\boldsymbol{a}(0)) - \widetilde{Q}_{pl}^{\infty}\right| \\ + \sum_{k,l=1}^{n_2}\left|\widetilde{H}_{kl}(\boldsymbol{w}(0),\boldsymbol{a}(0)) - \widetilde{H}_{kl}^{\infty}\right| \\ \le (n_1+n_2)^2\cdot\sqrt{\frac{2R'^2}{m}\cdot\log\frac{4(n_1+n_2)^2}{\delta}}\end{aligned} \tag{60}$$

holds, with probability of at least $1 - \frac{\delta}{2}$. When m is large enough, such that

$$\sqrt{\frac{2R'^2}{m}\cdot\log\frac{4(n_1+n_2)^2}{\delta}} \le \frac{\widetilde{\lambda}_0}{4(n_1+n_2)^2}, \tag{61}$$

then

$$\begin{aligned}&\left\|\widetilde{\boldsymbol{G}}(\boldsymbol{w}(0),\boldsymbol{a}(0)) - \widetilde{\boldsymbol{G}}^{\infty}\right\|_2 \le \left\|\widetilde{\boldsymbol{G}}(\boldsymbol{w}(0),\boldsymbol{a}(0)) - \widetilde{\boldsymbol{G}}^{\infty}\right\|_F \\ &\le \sum_{p,q=1}^{n_1}\left|\widetilde{S}_{pq}(\boldsymbol{w}(0),\boldsymbol{a}(0)) - \widetilde{S}_{pq}^{\infty}\right| + 2\cdot\sum_{p=1}^{n_1}\sum_{l=1}^{n_2}\left|\widetilde{Q}_{pl}(\boldsymbol{w}(0),\boldsymbol{a}(0)) - \widetilde{Q}_{pl}^{\infty}\right| \\ &\quad + \sum_{k,l=1}^{n_2}\left|\widetilde{H}_{kl}(\boldsymbol{w}(0),\boldsymbol{a}(0)) - \widetilde{H}_{kl}^{\infty}\right| \\ &\le (n_1+n_2)^2\cdot\frac{\widetilde{\lambda}_0}{4(n_1+n_2)^2} = \frac{\widetilde{\lambda}_0}{4}.\end{aligned} \tag{62}$$

Here, the condition (61) implies that

$$m \ge \frac{32(n_1+n_2)^2\cdot R'^2}{\widetilde{\lambda}_0^2}\cdot\log\frac{4(n_1+n_2)^2}{\delta} = \widetilde{\Omega}\left(\frac{(n_1+n_2)^4}{(n_1n_2)^2\widetilde{\lambda}_0^2}\cdot\left(\log\frac{1}{\delta}\right)^7\right). \tag{63}$$

We can conduct similar steps for $\boldsymbol{G}(\boldsymbol{w}(0),\boldsymbol{a}(0))$. With the probability of at least $1-\delta$ over the initialization, we have

$$\left\|\boldsymbol{G}(\boldsymbol{w}(0),\boldsymbol{a}(0)) - \widetilde{\boldsymbol{G}}^{\infty}\right\|_2 \le \left\|\boldsymbol{G}(\boldsymbol{w}(0),\boldsymbol{a}(0)) - \widetilde{\boldsymbol{G}}^{\infty}\right\|_F \le \frac{\lambda_0}{4},$$

if

$$m \ge \widetilde{\Omega}\left(\frac{(n_1+n_2)^4}{(n_1n_2)^2\lambda_0^2}\cdot\left(\log\frac{1}{\delta}\right)^5\right).$$

B.3. Proof of Lemma 3.6

By the property of Gaussian variables, we have

$$\begin{aligned}\mathbb{P}\left(\|\boldsymbol{w}_r(0)\|_2 \geq R'\right) = \mathbb{P}\left(\|\boldsymbol{w}_r(0)\|_2^2 \geq R'^2\right) &\leq (d+2)\cdot \mathbb{P}_{z\sim\mathcal{N}(0,1)}\left(|z| \geq \frac{R'}{\sqrt{d+2}}\right)\\ &\leq (d+2)\exp\left(-\frac{1}{2}\cdot\frac{R'^2}{d+2}\right),\end{aligned}$$

where the first inequality holds since $\|\boldsymbol{w}_r(0)\|_2^2 \geq R'^2$ implies that there exists at least one element of $\boldsymbol{w}_r(0)$ such that $w_{ri}(0)^2 \geq \frac{R'^2}{d+2}$. Then, with a probability of at least $1-\frac{\delta}{4m}$,

$$\|\boldsymbol{w}_r(0)\|_2 \leq R' =: \sqrt{2(d+2)\cdot\log\left(\frac{4m(d+2)}{\delta}\right)}, \tag{64}$$

and (64) holds for all $r\in[m]$ with probability of at least $1-\frac{\delta}{4}$. Without the loss of generality, we assume that $R_w \leq R'$, therefore, $\|\tilde{\boldsymbol{w}}_r\|_2 \leq \|\tilde{\boldsymbol{w}}_r - \boldsymbol{w}_r(0)\|_2 + \|\boldsymbol{w}_r(0)\|_2 \leq 2R'$. Moreover, $|\tilde{a}_r| \leq |a_r(0)| + |\tilde{a}_r - a_r(0)| \leq 1 + R_a \leq 2$. In the next part, we only consider $\tilde{\boldsymbol{w}}_r$ and $\tilde{a}_r$ that are bounded by $2R'$ and 2, respectively.

We first discuss the error bound for $|S_{pq}(\tilde{\boldsymbol{w}}) - S_{pq}(\boldsymbol{w}(0))|$. Case 1: $\mathbb{I}(\boldsymbol{w}_r(0)^\top\boldsymbol{y}_p > 0, \boldsymbol{w}_r(0)^\top\boldsymbol{y}_q > 0) = \mathbb{I}(\tilde{\boldsymbol{w}}_r^\top\boldsymbol{y}_p > 0, \tilde{\boldsymbol{w}}_r^\top\boldsymbol{y}_q > 0) = 1$ for all $\|\tilde{\boldsymbol{w}}_r - \boldsymbol{w}_r(0)\|_2 \leq R_w$. Denote

$$F_{pq}(\boldsymbol{w}_r, a_r) = \left\langle \frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r}\right\rangle,$$

then

$$\left\langle \frac{\partial s_p(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial \boldsymbol{w}_r}\right\rangle = F_{pq}(\boldsymbol{w}_r(0), a_r(0)),$$

and

$$\left\langle \frac{\partial s_p(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})}{\partial \boldsymbol{w}_r}\right\rangle = F_{pq}(\tilde{\boldsymbol{w}}_r, \tilde{a}_r).$$

Note that $F_{pq}(\boldsymbol{w}_r, a_r)$ is a polynomial of $\boldsymbol{w}_r$ with degree 4, then there exist universal constants $c_1 > 0$ and $c_2 > 0$, such that

$$\left\|\frac{\partial F_{pq}(\boldsymbol{w}_r, a_r)}{\partial \boldsymbol{w}_r}\right\|_2 \leq \frac{c_1}{mn_1}\cdot\|\boldsymbol{w}_r\|_2^3,$$

and

$$\left\|\frac{\partial F_{pq}(\boldsymbol{w}_r, a_r)}{\partial a_r}\right\|_2 \leq \frac{c_2}{mn_1}\cdot\|\boldsymbol{w}_r\|_2^4.$$

Therefore, by the mean value theorem, we have

$$\begin{aligned}&\left|\left\langle \frac{\partial s_p(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})}{\partial \boldsymbol{w}_r}\right\rangle - \left\langle \frac{\partial s_p(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial \boldsymbol{w}_r}\right\rangle\right|\\ =&\left|F_{pq}(\tilde{\boldsymbol{w}}_r,\tilde{a}_r) - F_{pq}(\boldsymbol{w}_r(0),a_r(0))\right|\\ \leq&\left|F_{pq}(\tilde{\boldsymbol{w}}_r,\tilde{a}_r) - F_{pq}(\tilde{\boldsymbol{w}}_r,a_r(0)) + F_{pq}(\tilde{\boldsymbol{w}}_r,a_r(0)) - F_{pq}(\boldsymbol{w}_r(0),a_r(0))\right|\\ \leq&\left|F_{pq}(\tilde{\boldsymbol{w}}_r,\tilde{a}_r) - F_{pq}(\tilde{\boldsymbol{w}}_r,a_r(0))\right| + \left|F_{pq}(\tilde{\boldsymbol{w}}_r,a_r(0)) - F_{pq}(\boldsymbol{w}_r(0),a_r(0))\right|\\ \leq&\frac{c_2}{mn_1}\cdot\left(2R'\right)^4\cdot\|a_r - a_r(0)\|_2 + \frac{c_1}{mn_1}\cdot\left(2R'\right)^3\cdot\|\tilde{\boldsymbol{w}}_r - \boldsymbol{w}_r(0)\|_2\\ \leq&\frac{c_1}{mn_1}\cdot\left(2R'\right)^3\cdot R_w + \frac{c_2}{mn_1}\cdot\left(2R'\right)^4\cdot R_a.\end{aligned} \tag{65}$$

Case 2: $\mathbb{I}(\boldsymbol{w}_r(0)^\top\boldsymbol{y}_p > 0, \boldsymbol{w}_r(0)^\top\boldsymbol{y}_q > 0) \neq \mathbb{I}(\tilde{\boldsymbol{w}}_r^\top\boldsymbol{y}_p > 0, \tilde{\boldsymbol{w}}_r^\top\boldsymbol{y}_q > 0)$ for some $\|\tilde{\boldsymbol{w}}_r - \boldsymbol{w}_r(0)\|_2 \leq R_w$. Without the loss of generality, we assume that $\mathbb{I}(\boldsymbol{w}_r(0)^\top\boldsymbol{y}_p > 0, \boldsymbol{w}_r(0)^\top\boldsymbol{y}_q > 0) = 0$ and $\mathbb{I}(\tilde{\boldsymbol{w}}_r^\top\boldsymbol{y}_p > 0, \tilde{\boldsymbol{w}}_r^\top\boldsymbol{y}_q > 0) = 1$, denoted as the event E_2, then it happens only if $|\boldsymbol{w}_r(0)^\top \boldsymbol{y}_p| < R_w$ or $|\boldsymbol{w}_r(0)^\top \boldsymbol{y}_q| < R_w$. Here,

$$\begin{aligned}
&\mathbb{P}\left(E_2\right) \\
\leq &\mathbb{P}\left(|\boldsymbol{w}_r(0)^\top \boldsymbol{y}_p| < R_w \text{ or } |\boldsymbol{w}_r(0)^\top \boldsymbol{y}_q| < R_w\right) \\
\leq &\mathbb{P}\left(|\boldsymbol{w}_r(0)^\top \boldsymbol{y}_p| < R_w\right) + \mathbb{P}\left(|\boldsymbol{w}_r(0)^\top \boldsymbol{y}_q| < R_w\right) \\
= &2 \cdot \mathbb{P}_{z\sim\mathcal{N}(0,1)}\left(|z| \leq R_w\right) \\
\leq &\frac{4R_w}{\sqrt{2\pi}},
\end{aligned}$$

and furthermore,

$$\mathbb{P}(E_2|\ \|\boldsymbol{w}_r(0)\|_2 \leq R') = \frac{\mathbb{P}(\text{case } 2, \|\boldsymbol{w}_r(0)\|_2 \leq R')}{\mathbb{P}(\|\boldsymbol{w}_r(0)\|_2 \leq R')} \leq \frac{\mathbb{P}\left(\text{case } 2\right)}{1-\delta} \leq \frac{8R_w}{\sqrt{2\pi}}, \tag{66}$$

where the last inequality holds if we assume that $\delta < \frac{1}{2}$. Moreover,

$$\begin{aligned}
&\left|\left\langle \frac{\partial s_p(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})}{\partial \boldsymbol{w}_r}\right\rangle - \left\langle \frac{\partial s_p(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial \boldsymbol{w}_r}\right\rangle\right| \\
=&\left|\left\langle \frac{\partial s_p(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})}{\partial \boldsymbol{w}_r}\right\rangle\right| \\
=&\left|F_{pq}(\tilde{\boldsymbol{w}}_r, \tilde{a}_r)\right| \\
\leq&\frac{c_3}{mn_1} \cdot \left(2R'\right)^4,
\end{aligned} \tag{67}$$

for a universal constant $c_3 > 0$, if $\|\tilde{\boldsymbol{w}}_r\|_2 \leq 2R'$ and $\tilde{a}_r \leq 2$.

Combining (65), (66) with (67), we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{w}_r(0),a_r(0)}\left(\sup_{\substack{\|\tilde{\boldsymbol{w}}_r - \boldsymbol{w}_r(0)\|_2 \leq R_w \\ |\tilde{a}_r - a_r(0)|_2 \leq R_a}} \left|\left\langle \frac{\partial s_p(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})}{\partial \boldsymbol{w}_r}\right\rangle - \right.\right. \\
&\qquad \left.\left.\left\langle \frac{\partial s_p(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial \boldsymbol{w}_r}\right\rangle\right| \;\middle|\; \|\boldsymbol{w}_r(0)\|_2 \leq R' \right) \\
\leq\ & \frac{c_1}{mn_1} \cdot \left(2R'\right)^3 \cdot R_w + \frac{c_2}{mn_1} \cdot \left(2R'\right)^4 \cdot R_a + \frac{8R_w}{\sqrt{2\pi}} \cdot \frac{c_3}{mn_1} \cdot \left(2R'\right)^4 \\
\leq\ & \frac{c_4}{mn_1} \cdot R'^4 \cdot \left(R_w + R_a\right),
\end{aligned} \tag{68}$$

where the last inequality holds for a universal constant $c_4 > 0$ and with $R' > 1$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{w}(0)}\left(\sup_{\substack{\|\tilde{\boldsymbol{w}}_r - \boldsymbol{w}_r(0)\|_2 \leq R_w \\ |\tilde{a}_r - a_r(0)|_2 \leq R_a}} \left|S_{pq}(\boldsymbol{w}(0),\boldsymbol{a}(0)) - S_{pq}(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})\right| \;\middle|\; \|\boldsymbol{w}_r(0)\|_2 \leq R',\ \ r \in [m] \right) \\
\leq\ & \sum_{r=1}^{m} \mathbb{E}_{\boldsymbol{w}_r(0)}\left(\sup_{\substack{\|\tilde{\boldsymbol{w}}_r - \boldsymbol{w}_r(0)\|_2 \leq R_w \\ |\tilde{a}_r - a_r(0)|_2 \leq R_a}} \left|\left\langle \frac{\partial s_p(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})}{\partial \boldsymbol{w}_r}\right\rangle - \right.\right. \\
&\qquad \left.\left.\left\langle \frac{\partial s_p(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial \boldsymbol{w}_r}, \frac{\partial s_q(\boldsymbol{w}(0),\boldsymbol{a}(0))}{\partial \boldsymbol{w}_r}\right\rangle\right| \;\middle|\; \|\boldsymbol{w}_r(0)\|_2 \leq R' \right) \\
\leq\ & \frac{c_4}{n_1} \cdot R'^4 \cdot \left(R_w + R_a\right)
\end{aligned} \tag{69}$$

and thus

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{w}(0)}\left(\sup_{\substack{\|\tilde{\boldsymbol{w}}_r-\boldsymbol{w}_r(0)\|_2\le R_w\\ |\tilde{a}_r-a_r(0)|_2\le R_a}} \sum_{p,q=1}^{n_1}|S_{pq}(\boldsymbol{w}(0),\boldsymbol{a}(0))-S_{pq}(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})| \;\middle|\; \|\boldsymbol{w}_r(0)\|_2\le R',\;\; r\in[m]\right)\\
\le\;& \sum_{p,q=1}^{n_1}\mathbb{E}_{\boldsymbol{w}(0)}\left(\sup_{\substack{\|\tilde{\boldsymbol{w}}_r-\boldsymbol{w}_r(0)\|_2\le R_w\\ |\tilde{a}_r-a_r(0)|_2\le R_a}} |S_{pq}(\boldsymbol{w}(0),\boldsymbol{a}(0))-S_{pq}(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})| \;\middle|\; \|\boldsymbol{w}_r(0)\|_2\le R'\;\; r\in[m]\right)\\
\le\;& c_4\cdot n_1\cdot R'^4\cdot(R_w+R_a)\,.
\end{aligned}\tag{70}$$

Similarly, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{w}(0)}\left(\sup_{\substack{\|\tilde{\boldsymbol{w}}_r-\boldsymbol{w}_r(0)\|_2\le R_w\\ |\tilde{a}_r-a_r(0)|_2\le R_a}} \sum_{p=1}^{n_1}\sum_{l=1}^{n_2}|Q_{pl}(\boldsymbol{w}(0),\boldsymbol{a}(0))-Q_{pl}(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})| \;\middle|\; \|\boldsymbol{w}_r(0)\|_2\le R',\;\; r\in[m]\right)\\
\le\;& c_5\cdot\sqrt{n_1 n_2}\cdot R'^4\cdot(R_w+R_a)\,.
\end{aligned}\tag{71}$$

and

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{w}(0)}\left(\sup_{\substack{\|\tilde{\boldsymbol{w}}_r-\boldsymbol{w}_r(0)\|_2\le R_w\\ |\tilde{a}_r-a_r(0)|_2\le R_a}} \sum_{k,l=1}^{n_2}|H_{kl}(\boldsymbol{w}(0),\boldsymbol{a}(0))-H_{kl}(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})| \;\middle|\; \|\boldsymbol{w}_r(0)\|_2\le R',\;\; r\in[m]\right)\\
\le\;& c_6\cdot n_2\cdot R'^4\cdot(R_w+R_a)\,.
\end{aligned}\tag{72}$$

for some universal constants $c_5>0$ and $c_6>0$. Moreover, let

$$\begin{aligned}
&e(\boldsymbol{w},\tilde{\boldsymbol{w}},\boldsymbol{a},\tilde{\boldsymbol{a}})\\
=\;& \sum_{p,q=1}^{n_1}|S_{pq}(\boldsymbol{w},\boldsymbol{a})-S_{pq}(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})|+2\sum_{p=1}^{n_1}\sum_{l=1}^{n_2}|Q_{pl}(\boldsymbol{w},\boldsymbol{a})-Q_{pl}(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})|\\
&+\sum_{k,l=1}^{n_1}|H_{kl}(\boldsymbol{w},\boldsymbol{a})\;-H_{kl}(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})|\,,
\end{aligned}$$

then

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{w}(0)}\left(\sup_{\substack{\|\tilde{\boldsymbol{w}}_r-\boldsymbol{w}_r(0)\|_2\le R_w\\ |\tilde{a}_r-a_r(0)|_2\le R_a}} e(\boldsymbol{w}(0),\tilde{\boldsymbol{w}},\boldsymbol{a}(0),\tilde{\boldsymbol{a}}) \;\middle|\; \|\boldsymbol{w}_r(0)\|_2\le R',\;\; r\in[m]\right)\\
\le\;& c_4\cdot n_1\cdot R'^4\cdot(R_w+R_a)+2c_5\cdot\sqrt{n_1n_2}\cdot R'^4\cdot(R_w+R_a)+c_6\cdot n_2\cdot R'^4\cdot(R_w+R_a)\\
\le\;& c_0\cdot(n_1+n_2)\cdot R'^4\cdot(R_w+R_a)\,,
\end{aligned}\tag{73}$$

for a universal constant $c_0>0$. By Markov's inequality, with a probability of at least $1-\frac{\delta}{4}$,

$$\sup_{\substack{\|\tilde{\boldsymbol{w}}_r-\boldsymbol{w}_r(0)\|_2\le R_w\\ |\tilde{a}_r-a_r(0)|_2\le R_a}} e(\boldsymbol{w}(0),\tilde{\boldsymbol{w}},\boldsymbol{a}(0),\tilde{\boldsymbol{a}})\le\frac{4c_0\cdot(n_1+n_2)\cdot R'^4\cdot(R_w+R_a)}{\delta},$$

if $\|\boldsymbol{w}_r(0)\|_2 \leq R'$, for all $r \in [m]$. Therefore, with a probability of at least $1 - \frac{\delta}{2}$, we have

$$\sup_{\substack{\|\tilde{\boldsymbol{w}}_r - \boldsymbol{w}_r(0)\|_2 \leq R_w \\ |\tilde{a}_r - a_r(0)|_2 \leq R_a}} e(\boldsymbol{w}(0), \tilde{\boldsymbol{w}}, \boldsymbol{a}(0), \tilde{\boldsymbol{a}}) \leq \frac{4c_0 \cdot (n_1 + n_2) \cdot R'^4 \cdot (R_w + R_a)}{\delta},$$

with $\boldsymbol{w}_r(0)$ are i.i.d. sampled from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d+2})$ for all $r \in [m]$. Furthermore, if

$$\frac{2c_0 \cdot (n_1 + n_2) \cdot R'^4 \cdot (R_w + R_a)}{\delta} \leq \frac{\lambda_0}{4},$$

and equivalently

$$R_w + R_a = \widetilde{\mathcal{O}}\left(\frac{\lambda_0 \cdot \delta}{(n_1 + n_2) \cdot (\log m)^2}\right), \tag{74}$$

then

$$\begin{aligned} \|\boldsymbol{G}(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{a}}) - \boldsymbol{G}(\boldsymbol{w}(0), \boldsymbol{a}(0))\|_2 &\leq \|\boldsymbol{G}(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{a}}) - \boldsymbol{G}(\boldsymbol{w}(0), \boldsymbol{a}(0))\|_F \\ &\leq e(\boldsymbol{w}(0), \tilde{\boldsymbol{w}}, \boldsymbol{a}(0), \tilde{\boldsymbol{a}}) \leq \frac{\lambda_0}{4}. \end{aligned}$$

We can similarly develop the error bound for $|\widetilde{S}_{pq}(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{a}}) - \widetilde{S}_{pq}(\boldsymbol{w}(0), \boldsymbol{a}(0))|$. Denote

$$\begin{aligned} &\widetilde{F}_{pq}(\boldsymbol{w}_r, a_r) \\ &= \left\langle \frac{\partial s_p(\boldsymbol{w}_r, a_r)}{\partial a_r}, \frac{\partial s_q(\boldsymbol{w}_r, a_r)}{\partial a_r} \right\rangle \\ &= \frac{9}{mn_1} \cdot (\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot (\boldsymbol{w}_r^\top \boldsymbol{y}_q) \cdot \left((\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot w_{r0} - 2\sum_{i=1}^d w_{ri}^2\right) \cdot \left((\boldsymbol{w}_r^\top \boldsymbol{y}_q) \cdot w_{r0} - 2\sum_{i=1}^d w_{ri}^2\right), \end{aligned} \tag{75}$$

which is a polynomial of $\boldsymbol{w}_r$ with degree 6. Then we similarly have

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{w}_r(0), a_r(0)}\Bigg(\sup_{\substack{\|\tilde{\boldsymbol{w}}_r - \boldsymbol{w}_r(0)\|_2 \leq R_w \\ |a_r - a_r(0)|_2 \leq R_a}} \left|\left\langle \frac{\partial s_p(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{a}})}{\partial a_r}, \frac{\partial s_q(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{a}})}{\partial a_r} \right\rangle - \right. \\ &\left.\left\langle \frac{\partial s_p(\boldsymbol{w}(0), \boldsymbol{a}(0))}{\partial a_r}, \frac{\partial s_q(\boldsymbol{w}(0), \boldsymbol{a}(0))}{\partial a_r} \right\rangle\right| \;\Bigg|\; \|\boldsymbol{w}_r(0)\|_2 \leq R' \Bigg) \\ \leq\;& \frac{c_7}{mn_1} \cdot R'^6 \cdot R_w, \end{aligned} \tag{76}$$

where the last inequality holds for a universal constant $c_7 > 0$ and with $R' > 1$. Therefore, with a probability of at least $1 - \frac{\delta}{2}$,

$$\left\|\widetilde{\boldsymbol{G}}(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{a}}) - \widetilde{\boldsymbol{G}}(\boldsymbol{w}(0), \boldsymbol{a}(0))\right\|_2 \leq \frac{\widetilde{\lambda}_0}{4},$$

if

$$R_w = \widetilde{\mathcal{O}}\left(\frac{\widetilde{\lambda}_0 \cdot \delta}{(n_1 + n_2) \cdot (\log m)^3}\right). \tag{77}$$

B.4. Proof of Lemma 3.7

Let

$$s_p(\boldsymbol{w}, \boldsymbol{a}) = \sum_{r=1}^m s_{pr}(\boldsymbol{w}_r, a_r) - \sqrt{\frac{1}{n_1}} f(\boldsymbol{x}_p)$$

and

$$h_k(\boldsymbol{w},\boldsymbol{a}) = \sum_{r=1}^{m} h_{kr}(\boldsymbol{w}_r) - \sqrt{\frac{1}{n_2}} g(\tilde{\boldsymbol{x}}_k),$$

where

$$s_{pr}(\boldsymbol{w}_r, a_r) = \frac{1}{\sqrt{mn_1}} \cdot \left(3a_r \cdot (\boldsymbol{w}_r^\top \boldsymbol{y}_p)^2 \cdot w_{r0} - 6a_r \cdot (\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot \sum_{i=1}^{d} w_{ri}^2 \right) \cdot \mathbb{I}(\boldsymbol{w}_r^\top \boldsymbol{y}_p > 0)$$

and

$$h_{kr}(\boldsymbol{w}_r, a_r) = \sqrt{\frac{\nu}{mn_2}} \cdot a_r \cdot (\boldsymbol{w}_r^\top \tilde{\boldsymbol{y}}_k)^3 \cdot \mathbb{I}(\boldsymbol{w}_r^\top \tilde{\boldsymbol{y}}_k > 0).$$

Then using $\mathbb{E} a_r(0) = 0$, we have

$$\begin{aligned}
& \mathbb{E}_{\boldsymbol{w}(0),\boldsymbol{a}(0)} \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix} \right\|_2^2 \\
= \; & \mathbb{E}_{\boldsymbol{w}(0),\boldsymbol{a}(0)} \sum_{p=1}^{n_1} s_p(\boldsymbol{w}(0),\boldsymbol{a}(0))^2 + \sum_{k=1}^{n_2} h_k(\boldsymbol{w}(0),\boldsymbol{a}(0))^2 \\
= \; & \mathbb{E}_{\boldsymbol{w}(0),\boldsymbol{a}(0)} \sum_{p=1}^{n_1} \sum_{r=1}^{m} s_{pr}(\boldsymbol{w}_r(0), a_r(0))^2 + \sum_{k=1}^{n_2} \sum_{r=1}^{m} h_{kr}(\boldsymbol{w}_r(0), a_r(0))^2 \\
& + \frac{1}{n_1} \sum_{p=1}^{n_1} f(\boldsymbol{x}_p)^2 + \frac{1}{n_2} \sum_{k=1}^{n_2} g(\tilde{\boldsymbol{x}}_k)^2 \le c,
\end{aligned}$$

where the universal constant $c > 0$ are independent of m, n_1 and n_2. Therefore, by Markov's inequality, we have with a probability of at least $1 - \delta$ over the initialization,

$$\left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix} \right\|_2^2 = \mathcal{O}\left(\frac{1}{\delta}\right).$$

B.5. Some Useful Lemmas and Proof of Theorem 3.8

The Proof of Theorem 3.8 consists of some lemmas. we will depict them one by one.

Lemma B.1. *If for all $0 \le \tau \le t$, $\lambda_{\min}(\boldsymbol{G}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau)) + \widetilde{\boldsymbol{G}}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))) \ge \frac{\lambda_0 + \widetilde{\lambda}_0}{2}$, then*

$$\left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau)) \\ \boldsymbol{h}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau)) \end{bmatrix} \right\|_2^2 \le \exp\left(-\left(\lambda_0 + \widetilde{\lambda}_0\right) \cdot \tau \right) \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix} \right\|_2^2. \tag{78}$$

Proof.

$$\begin{aligned}
& \frac{d}{d\tau} \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau)) \\ \boldsymbol{h}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau)) \end{bmatrix} \right\|_2^2 \\
= & -2 \left[\boldsymbol{s}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))^\top, \boldsymbol{h}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))^\top \right] \cdot \left(\boldsymbol{G}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau)) + \widetilde{\boldsymbol{G}}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau)) \right) \cdot \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau)) \\ \boldsymbol{h}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau)) \end{bmatrix} \qquad (79) \\
\le & -\left(\lambda_0 + \widetilde{\lambda}_0\right) \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau)) \\ \boldsymbol{h}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau)) \end{bmatrix} \right\|_2^2,
\end{aligned}$$

which completes the proof. □

Lemma B.2. *If $m = \Omega\left(\frac{1}{\left(\lambda_0 + \widetilde{\lambda}_0\right)^2} \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix} \right\|_2^2 \cdot \left(\frac{R'^4}{R_w^2} + \frac{R'^6}{R_a^2} \right) \right)$, $|a_r(\tau)| \le 2$, $\|\boldsymbol{w}_r(\tau)\|_2 \le 2R'$, $\lambda_{\min}(\boldsymbol{G}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))) \ge \frac{\lambda_0}{2}$ and $\lambda_{\min}(\widetilde{\boldsymbol{G}}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))) \ge \frac{\widetilde{\lambda}_0}{2}$ for all $0 \le \tau \le t$, then $\|\boldsymbol{w}_r(\tau) - \boldsymbol{w}_r(0)\|_2 \le R_w$ and $|a_r(\tau) - a_r(0)| \le R_a$, for all $r \in [m]$ and $0 \le \tau \le t$, where R_w, R_a and R' are defined in Lemma 3.6 and its proof in Appendix B.3.*

Proof. Note that

$$\left\|\frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r}\right\|_2 \leq \frac{1}{\sqrt{mn_1}}\cdot|a_r|\cdot 27\|\boldsymbol{w}_r\|_2^2 \leq c_0\cdot\frac{1}{\sqrt{mn_1}}\cdot R'^2$$

and

$$\left\|\frac{\partial h_k(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r}\right\|_2 \leq 3\sqrt{\frac{\nu}{mn_2}}\cdot|a_r|\cdot\|\boldsymbol{w}_r\|_2^2 \leq c_0\cdot\frac{1}{\sqrt{mn_2}}\cdot R'^2,$$

for a universal constant $c_0 > 0$, since $|a_r| \leq 2$. Then by Equation (14), we have

$$\begin{aligned}
&\left\|\frac{d}{d\tau}\boldsymbol{w}_r(\tau)\right\|_2\\
=&\left\|\sum_{p=1}^{n_1} s_p(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\cdot\frac{\partial s_p(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))}{\partial \boldsymbol{w}_r}+\sum_{k=1}^{n_2} h_k(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\cdot\frac{\partial h_k(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))}{\partial \boldsymbol{w}_r}\right\|_2\\
\leq&\sum_{p=1}^{n_1}|s_p(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))|\cdot\left\|\frac{\partial s_p(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))}{\partial \boldsymbol{w}_r}\right\|_2+\sum_{k=1}^{n_2}|h_k(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))|\cdot\left\|\frac{\partial h_k(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))}{\partial \boldsymbol{w}_r}\right\|_2\\
\leq&\frac{1}{\sqrt{m}}\cdot c_0\cdot R'^2\cdot\|\boldsymbol{s}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\|_2+\frac{1}{\sqrt{m}}\cdot c_0\cdot R'^2\cdot\|\boldsymbol{h}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\|_2\\
\leq&\sqrt{\frac{2}{m}}c_0\cdot R'^2\cdot\left\|\begin{bmatrix}\boldsymbol{s}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\\ \boldsymbol{h}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\end{bmatrix}\right\|_2\\
\leq&\sqrt{\frac{2}{m}}c_0\cdot R'^2\cdot\exp\left(-\frac{\lambda_0+\widetilde{\lambda}_0}{2}\cdot\tau\right)\cdot\left\|\begin{bmatrix}\boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0))\\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{bmatrix}\right\|_2,
\end{aligned}\tag{80}$$

and

$$\begin{aligned}
\|\boldsymbol{w}_r(\tau)-\boldsymbol{w}_r(0)\|_2 &\leq \int_0^\tau\left\|\frac{d}{d\upsilon}\boldsymbol{w}_r(\upsilon)\right\|_2 d\upsilon\\
&\leq\frac{2\sqrt{2}c_0}{\sqrt{m}}\cdot R'^2\cdot\frac{1}{\lambda_0+\widetilde{\lambda}_0}\cdot\left\|\begin{bmatrix}\boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0))\\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{bmatrix}\right\|_2.
\end{aligned}\tag{81}$$

Moreover,

$$\left|\frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial a_r}\right| \leq \frac{9}{\sqrt{mn_1}}\cdot\|\boldsymbol{w}_r\|_2^3 \leq \frac{c_1}{\sqrt{mn_1}}\|\boldsymbol{w}_r\|_2^3$$

and

$$\left|\frac{\partial h_k(\boldsymbol{w},\boldsymbol{a})}{\partial a_r}\right| \leq \sqrt{\frac{\nu}{mn_2}}\cdot\|\boldsymbol{w}_r\|_2^3 \leq \frac{c_1}{\sqrt{mn_2}}\|\boldsymbol{w}_r\|_2^3,$$

for a universal constant $c_1 > 0$. Then by Equation (15), we have

$$\begin{aligned}
&\left|\frac{d}{d\tau}a_r(\tau)\right|\\
=&\left|\sum_{p=1}^{n_1} s_p(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\cdot\frac{\partial s_p(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))}{\partial a_r}+\sum_{k=1}^{n_2} h_k(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\cdot\frac{\partial h_k(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))}{\partial a_r}\right|\\
\leq&\sum_{p=1}^{n_1}|s_p(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))|\cdot\left|\frac{\partial s_p(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))}{\partial a_r}\right|+\sum_{k=1}^{n_2}|h_k(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))|\cdot\left|\frac{\partial h_k(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))}{\partial a_r}\right|\\
\leq&\frac{1}{\sqrt{m}}\cdot c_1\cdot R'^3\cdot\|\boldsymbol{s}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\|_2+\frac{1}{\sqrt{m}}\cdot c_1\cdot R'^3\cdot\|\boldsymbol{h}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\|_2\\
\leq&\frac{\sqrt{2}c_1}{\sqrt{m}}\cdot R'^3\cdot\left\|\begin{pmatrix}\boldsymbol{s}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\\ \boldsymbol{h}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\end{pmatrix}\right\|_2\\
\leq&\frac{\sqrt{2}c_0}{\sqrt{m}}\cdot R'^3\cdot\exp\left(-\frac{\lambda_0+\widetilde{\lambda}_0}{2}\cdot\tau\right)\cdot\left\|\begin{bmatrix}\boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0))\\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{bmatrix}\right\|_2,
\end{aligned}\tag{82}$$

and

$$\begin{aligned}|a_r(\tau)-a_r(0)| &\le \int_0^\tau \left|\frac{d}{dv}a_r(v)\right|_2 dv \\ &\le \frac{2\sqrt{2}c_0}{\sqrt{m}}\cdot R'^3\cdot\frac{1}{\lambda_0+\widetilde{\lambda}_0}\cdot\left\|\left[\begin{array}{c}\boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0))\\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{array}\right]\right\|_2.\end{aligned} \tag{83}$$

If m is large enough such that

$$\frac{2\sqrt{2}c_0}{\sqrt{m}}\cdot R'^2\cdot\frac{1}{\lambda_0+\widetilde{\lambda}_0}\cdot\left\|\left[\begin{array}{c}\boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0))\\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{array}\right]\right\|_2\le R_w$$

and

$$\frac{2\sqrt{2}c_0}{\sqrt{m}}\cdot R'^3\cdot\frac{1}{\lambda_0+\widetilde{\lambda}_0}\cdot\left\|\left[\begin{array}{c}\boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0))\\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{array}\right]\right\|_2\le R_a,$$

or equivalently

$$m=\Omega\left(\frac{1}{\left(\lambda_0+\widetilde{\lambda}_0\right)^2}\cdot\left(\frac{R'^4}{R_w^2}+\frac{R'^6}{R_a^2}\right)\cdot\left\|\left[\begin{array}{c}\boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0))\\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{array}\right]\right\|_2^2\right),$$

we have

$$\|\boldsymbol{w}_r(\tau)-\boldsymbol{w}_r(0)\|_2\le R_w \quad \text{and} \quad |a_r(\tau)-a_r(0)|\le R_a,$$

for all $0\le\tau\le t$. □

The proof consists of four parts.

Firstly, the initialized Gram matrix $\boldsymbol{G}(\boldsymbol{w}(0),\boldsymbol{a}(0))$ and $\widetilde{\boldsymbol{G}}(\boldsymbol{w}(0),\boldsymbol{a}(0))$ are positive definite. By Lemma 3.5 in the paper, if $m=\widetilde{\Omega}\left(\frac{(n_1+n_2)^4}{(n_1n_2)^2\cdot\left(\min\{\lambda_0,\widetilde{\lambda}_0\}\right)^2}\cdot\left(\log\frac{1}{\delta}\right)^7\right)$, then with probability of at least $1-\frac{\delta}{3}$, the initialized Gram matrices $\boldsymbol{G}(\boldsymbol{w}(0),\boldsymbol{a}(0))$ and $\widetilde{\boldsymbol{G}}(\boldsymbol{w}(0),\boldsymbol{a}(0))$ satisfy

$$\lambda_{\min}(\boldsymbol{G}(\boldsymbol{w}(0),\boldsymbol{a}(0)))\ge\frac{3}{4}\lambda_0.$$

and

$$\lambda_{\min}\left(\widetilde{\boldsymbol{G}}\left(\boldsymbol{w}(0),\boldsymbol{a}(0)\right)\right)\ge\frac{3}{4}\widetilde{\lambda}_0.$$

Secondly, the initialized loss is bounded. Lemma 3.7 shows that with probability of at least $1-\frac{\delta}{3}$ over the initialization of $\boldsymbol{w}_r(0)$ and $a_r(0)$ for all $r\in[m]$, the following holds

$$\left\|\left[\begin{array}{c}\boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0))\\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{array}\right]\right\|_2^2=\mathcal{O}\left(\frac{1}{\delta}\right)$$

Thirdly, Gram matrices $\boldsymbol{G}(\boldsymbol{w},\boldsymbol{a})$ and $\widetilde{\boldsymbol{G}}(\boldsymbol{w},\boldsymbol{a})$ are continuous at $\boldsymbol{w}(0)$ and $\boldsymbol{a}(0)$, as is shown in Lemma 3.6. With a probability of at least $1-\frac{\delta}{3}$, if the radius $R_w=\widetilde{\mathcal{O}}\left(\frac{\min\{\lambda_0,\widetilde{\lambda}_0\}\cdot\delta}{(n_1+n_2)\cdot(\log m)^3}\right)$ and $R_a=\widetilde{\mathcal{O}}\left(\frac{\min\{\lambda_0,\widetilde{\lambda}_0\}\cdot\delta}{(n_1+n_2)\cdot(\log m)^2}\right)$, then

$$\|\boldsymbol{G}(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})-\boldsymbol{G}(\boldsymbol{w}(0),\boldsymbol{a}(0))\|_2\le\frac{\lambda_0}{4}$$

and

$$\|\widetilde{\boldsymbol{G}}(\tilde{\boldsymbol{w}},\tilde{\boldsymbol{a}})-\widetilde{\boldsymbol{G}}(\boldsymbol{w}(0),\boldsymbol{a}(0))\|_2\le\frac{\widetilde{\lambda}_0}{4},$$

for all $\|\tilde{\boldsymbol{w}}_r - \boldsymbol{w}_r(0)\|_2 \le R_w$, $|\tilde{a}_r - a_r(0)| \le R_a \le 1$ and $r \in [m]$. It implies that Gram matrices in the neighborhood of $\boldsymbol{w}(0)$ and $\boldsymbol{a}(0)$ are still positive definite, i.e., $\lambda_{\min}(\boldsymbol{G}(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{a}})) \ge \frac{\lambda_0}{2}$ and $\lambda_{\min}(\widetilde{\boldsymbol{G}}(\tilde{\boldsymbol{w}}, \tilde{a})) \ge \frac{\widetilde{\lambda}_0}{2}$. Moreover,

$$\|\boldsymbol{w}_r(0)\|_2 \le R' =: \sqrt{2(d+2) \cdot \log\left(\frac{4m(d+2)}{\delta}\right)} = \mathcal{O}\left(\sqrt{\log\left(\frac{m}{\delta}\right)}\right), \tag{84}$$

holds for all $r \in [m]$.

Finally, $\boldsymbol{w}_r(t)$ and $a_r(t)$ will not go out of the ball $\mathcal{B}(\boldsymbol{w}_r(0), R_w)$ and $\mathcal{B}(a_r(0), R_a)$ respectively, for all $r \in [m]$. Without the loss of generality, we assume that $R' \ge R_w$, and thus $\|\boldsymbol{w}_r(\tau)\|_2 \le 2R'$, if $\boldsymbol{w}_r(\tau)$ stays in the ball $\mathcal{B}(\boldsymbol{w}_r(0), R_w)$. Lemma B.2 shows that if

$$\begin{aligned} m &= \Omega\left(\frac{1}{\left(\lambda_0 + \widetilde{\lambda}_0\right)^2} \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0)) \end{bmatrix} \right\|_2^2 \cdot \left(\frac{R'^4}{R_w^2} + \frac{R'^6}{R_a^2}\right)\right) \\ &= \widetilde{\Omega}\left(\frac{(n_1+n_2)^2}{\left(\lambda_0 + \widetilde{\lambda}_0\right)^2 \cdot \left(\min\{\lambda_0, \widetilde{\lambda}_0\}\right)^2 \cdot \delta^3}\right), \end{aligned} \tag{85}$$

then we have $\|\boldsymbol{w}_r(t) - \boldsymbol{w}_r(0)\|_2 \le R_w$, $|\tilde{a}_r - a_r(0)| \le R_a$, $\lambda_{\min}(\boldsymbol{G}(\boldsymbol{w}(t), \boldsymbol{a}(t))) \ge \frac{\lambda_0}{2}$ and $\lambda_{\min}(\widetilde{\boldsymbol{G}}(\boldsymbol{w}(t), \boldsymbol{a}(t))) \ge \frac{\widetilde{\lambda}_0}{2}$, for all $t > 0$ and $r \in [m]$. Furthermore, we have

$$\mathcal{L}(\boldsymbol{w}(t), \boldsymbol{a}(t)) \le \exp\left(-\left(\lambda_0 + \widetilde{\lambda}_0\right) \cdot t\right) \cdot \mathcal{L}(\boldsymbol{w}(0), \boldsymbol{a}(0)),$$

for all $t > 0$, by Lemma B.1.

C. Technical Proofs for Section 4

C.1. Proof for Lemma 4.1

Note that

$$\left\|\frac{\partial s_p(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))}{\partial \boldsymbol{w}_r}\right\|_2 \le c_0 \cdot \frac{1}{\sqrt{mn_1}} \cdot R^2 \quad \text{and} \quad \left\|\frac{\partial h_k(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))}{\partial \boldsymbol{w}_r}\right\|_2 \le c_0 \cdot \frac{1}{\sqrt{mn_2}} \cdot R^2,$$

for a universal constant $c_0 > 0$ and $|a_r| \le 2$. Then,

$$\begin{aligned}
\|\boldsymbol{w}_r(t+1) - \boldsymbol{w}_r(0)\|_2 &\le \eta \cdot \sum_{\tau=0}^{t} \left\|\frac{\partial \mathcal{L}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))}{\partial \boldsymbol{w}_r}\right\|_2 \\
&\le \eta \cdot \sum_{\tau=0}^{t} \left(\sum_{p=1}^{n_1} |s_p(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))| \cdot \left\|\frac{\partial}{\partial \boldsymbol{w}_r} s_p(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\right\|_2 + \right. \\
&\qquad \left. \sum_{k=1}^{n_2} |h_k(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))| \cdot \left\|\frac{\partial}{\partial \boldsymbol{w}_r} h_k(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\right\|_2 \right) \\
&\le \eta \cdot \sum_{\tau=0}^{t} \frac{2c_0}{\sqrt{m}} \cdot R^2 \cdot \left(1 - \eta \cdot \frac{\lambda_0 + \widetilde{\lambda}_0}{2}\right)^{\tau/2} \cdot \left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix}\right\|_2 \\
&\le 8c_0 \cdot \frac{R^2}{\sqrt{m}} \cdot \frac{1}{\lambda_0 + \widetilde{\lambda}_0} \cdot \left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix}\right\|_2 := R_w.
\end{aligned}$$

Furthermore, note that

$$\left|\frac{\partial s_p(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))}{\partial a_r}\right| \le \frac{c_1}{\sqrt{mn_1}} R^3 \quad \text{and} \quad \left|\frac{\partial h_k(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))}{\partial a_r}\right| \le \frac{c_1}{\sqrt{mn_2}} R^3,$$

for a universal constant $c_1 > 0$. Then,

$$\begin{aligned}
|a_r(t+1) - a_r(0)| &\le \eta \cdot \sum_{\tau=0}^{t} \left\|\frac{\partial \mathcal{L}(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))}{\partial a_r}\right\|_2 \\
&\le \eta \cdot \sum_{\tau=0}^{t} \left(\sum_{p=1}^{n_1} |s_p(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))| \cdot \left|\frac{\partial}{\partial a_r} s_p(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\right| + \right. \\
&\qquad \left. \sum_{k=1}^{n_2} |h_k(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))| \cdot \left|\frac{\partial}{\partial a_r} h_k(\boldsymbol{w}(\tau),\boldsymbol{a}(\tau))\right| \right) \\
&\le \eta \cdot \sum_{\tau=0}^{t} \frac{2c_1}{\sqrt{m}} \cdot R^3 \cdot \left(1 - \eta \cdot \frac{\lambda_0 + \widetilde{\lambda}_0}{2}\right)^{\tau/2} \cdot \left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix}\right\|_2 \\
&\le 8c_1 \cdot \frac{R^3}{\sqrt{m}} \cdot \frac{1}{\lambda_0 + \widetilde{\lambda}_0} \cdot \left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix}\right\|_2 := R_a.
\end{aligned}$$

C.2. Proof for Lemma 4.2

By the property of Gaussian variables, we have

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{w}_r(0)\|_2 \ge R'\right) = \mathbb{P}\left(\|\boldsymbol{w}_r(0)\|_2^2 \ge R'^2\right) &\le (d+2) \cdot \mathbb{P}_{z\sim\mathcal{N}(0,1)}\left(|z| \ge \frac{R'}{\sqrt{d+2}}\right) \\
&\le (d+2)\exp\left(-\frac{1}{2} \cdot \frac{R'^2}{d+2}\right).
\end{aligned}$$

Then, with probability of at least $1 - \frac{\delta}{2m}$,

$$\|\boldsymbol{w}_r(0)\|_2 \le R' := \sqrt{2(d+2) \cdot \log\left(\frac{2m(d+2)}{\delta}\right)}, \tag{86}$$

holds for each $r \in [m]$. Hence, with probability of at least $1 - \frac{\delta}{2}$, $\|\boldsymbol{w}_r(0)\|_2 \le R'$ holds for all $r \in [m]$.

Note that $\boldsymbol{w}_r$ (also for a_r, $r \in [m]$) in $s_p(\boldsymbol{w}, \boldsymbol{a})$ and $h_k(\boldsymbol{w}, \boldsymbol{a})$ can be separated (because of the formulation Equation (5) for $\phi(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{a})$) that

$$s_p(\boldsymbol{w}, \boldsymbol{a}) = \sum_{r=1}^{m} \zeta_p(\boldsymbol{w}_r, a_r) \quad \text{and} \quad h_k(\boldsymbol{w}, \boldsymbol{a}) = \sum_{r=1}^{m} \xi_k(\boldsymbol{w}_r, a_r), \tag{87}$$

where

$$\begin{aligned} & \zeta_p(\boldsymbol{w}_r, a_r) \\ = \ & \frac{3a_r}{\sqrt{mn_1}} \cdot \left((\boldsymbol{w}_r^\top \boldsymbol{y}_p)^2 \cdot w_{ri} - 2 \cdot (\boldsymbol{w}_r^\top \boldsymbol{y}_p) \cdot \left(\sum_{i=1}^{d} w_{ri}^2 \right) \right) \cdot \mathbb{I}\left(\boldsymbol{w}_r^\top \boldsymbol{y}_p > 0\right) - \frac{1}{m\sqrt{n_1}} \cdot f(\boldsymbol{x}_p) \end{aligned}$$

and $\xi_k(\boldsymbol{w}_r, a_r) = \sqrt{\frac{\nu}{mn_2}} \cdot \left(\boldsymbol{w}_r^\top \tilde{\boldsymbol{y}}_k\right)^3 \cdot \mathbb{I}\left(\boldsymbol{w}_r^\top \tilde{\boldsymbol{y}}_k > 0\right) - \frac{1}{m\sqrt{n_2}} \cdot g(\tilde{\boldsymbol{x}}_k)$. In the next part, we discuss six potential cases for $\boldsymbol{w}_r(t+1)$ and $\boldsymbol{w}_r(t)$.

Case 1.1: $\mathbb{I}\left(\boldsymbol{w}_r(t+1)^\top \boldsymbol{y}_p > 0\right) = \mathbb{I}\left(\boldsymbol{w}_r(t)^\top \boldsymbol{y}_p > 0\right) = 1$. Then

$$\begin{aligned} & \zeta_p(\boldsymbol{w}_r(t+1), a_r(t+1)) - \zeta_p(\boldsymbol{w}_r(t), a_r(t)) \\ = \ & \left\langle \frac{\partial \zeta_p(\boldsymbol{w}_r(t), a_r(t))}{\partial \boldsymbol{w}_r}, \boldsymbol{w}_r(t+1) - \boldsymbol{w}_r(t) \right\rangle + \frac{\partial \zeta_p(\boldsymbol{w}_r(t), a_r(t))}{\partial a_r} \cdot (a_r(t+1) - a_r(t)) + \chi_{pr}(t), \end{aligned}$$

where

$$|\chi_{pr}(t)| \le \frac{c_2}{\sqrt{mn_1}} \cdot \left(R' \cdot \widetilde{R}_w^2 + R'^2 \cdot \widetilde{R}_w \cdot \widetilde{R}_a \right), \tag{88}$$

for a universal constant $c_2 > 0$, since $\frac{\partial^2 \zeta_p(\boldsymbol{w}_r, a_r)}{\partial w_{ri}^2}$ is a polynomial of $\boldsymbol{w}_r$ with degree 1, $\frac{\partial^2 \zeta_p(\boldsymbol{w}_r, a_r)}{\partial w_{ri} \partial a_r}$ is a polynomial of $\boldsymbol{w}_r$ with degree 2 and $\frac{\partial^2 \zeta_p(\boldsymbol{w}_r, a_r)}{\partial a_r^2} = 0$.

Case 1.2: $\mathbb{I}\left(\boldsymbol{w}_r(t+1)^\top \boldsymbol{y}_p > 0\right) \neq \mathbb{I}\left(\boldsymbol{w}_r(t)^\top \boldsymbol{y}_p > 0\right)$. Without the loss of generality, we assume that $\mathbb{I}\left(\boldsymbol{w}_r(t+1)^\top \boldsymbol{y}_p > 0\right) = 0$ and $\mathbb{I}\left(\boldsymbol{w}_r(t)^\top \boldsymbol{y}_p > 0\right) = 1$, denoted as the event E_1. Then it happens only if $\left|\boldsymbol{w}_r(0)^\top \boldsymbol{y}_p\right| < R_w$, with

$$\mathbb{P}\left(E_1\right) \le \mathbb{P}\left(\left|\boldsymbol{w}_r(0)^\top \boldsymbol{y}_p\right| < R_w\right) = \mathbb{P}_{z \sim \mathcal{N}(0,1)}\left(|z| < R_w\right) \le \frac{2R_w}{\sqrt{2\pi}}.$$

Furthermore,

$$\mathbb{P}(E_1 | \, \|\boldsymbol{w}_r(0)\|_2 \le R') = \frac{\mathbb{P}(E_1, \|\boldsymbol{w}_r(0)\|_2 \le R')}{\mathbb{P}(\|\boldsymbol{w}_r(0)\|_2 \le R')} \le \frac{\mathbb{P}\left(E_1\right)}{1-\delta} \le \frac{8R_w}{\sqrt{2\pi}}, \tag{89}$$

where the last inequality holds if we assume that $\delta < \frac{1}{2}$. Let the set $\mathcal{R}_p^t(\boldsymbol{w}(0))$ be defined as

$$\mathcal{R}_p^t(\boldsymbol{w}(0)) = \left\{ r \in [m] : \mathbb{I}\left(\boldsymbol{w}_r(t+1)^\top \boldsymbol{y}_p > 0\right) \neq \mathbb{I}\left(\boldsymbol{w}_r(t)^\top \boldsymbol{y}_p > 0\right) \right\}.$$

Then, we have

$$\mathbb{E}_{\boldsymbol{w}(0)} \sum_{p=1}^{n_1} \left|\mathcal{R}_p^t(\boldsymbol{w}(0))\right| \le \sum_{p=1}^{n_1} \frac{8R_w}{\sqrt{2\pi}} \cdot m = \frac{8R_w \cdot mn_1}{\sqrt{2\pi}}.$$

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

$$\sum_{p=1}^{n_1} \left|\mathcal{R}_p^t(\boldsymbol{w}(0))\right| \le \frac{8R_w \cdot mn_1}{\sqrt{2\pi}\delta}.$$

Here, $\mathbb{I}\left(\boldsymbol{w}_r(t+1)^\top \boldsymbol{y}_p > 0\right) \neq \mathbb{I}\left(\boldsymbol{w}_r(t)^\top \boldsymbol{y}_p > 0\right)$ implies that

$$\left|\boldsymbol{w}_r(t)^\top \boldsymbol{y}_p\right| = \left| \left(\boldsymbol{w}_r(t) - \boldsymbol{w}_r(t+1)\right)^\top \boldsymbol{y}_p + \boldsymbol{w}_r(t+1)^\top \boldsymbol{y}_p \right| \le \left\|\boldsymbol{w}_r(t) - \boldsymbol{w}_r(t+1)\right\|_2. \tag{90}$$

Let

$$\begin{aligned}&\zeta_p(\boldsymbol{w}_r(t+1),a_r(t+1))-\zeta_p(\boldsymbol{w}_r(t),a_r(t))\\=\ &\left\langle\frac{\partial\zeta_p(\boldsymbol{w}_r(t),a_r(t))}{\partial\boldsymbol{w}_r},\boldsymbol{w}_r(t+1)-\boldsymbol{w}_r(t)\right\rangle+\frac{\partial\zeta_p(\boldsymbol{w}_r(t),a_r(t))}{\partial a_r}\cdot(a_r(t+1)-a_r(t))+\chi_{pr}(t),\end{aligned}$$

then

$$\begin{aligned}|\chi_{pr}(t)|\ &\leq\ \left|\left\langle\frac{\partial\zeta_p(\boldsymbol{w}_r(t),a_r(t))}{\partial\boldsymbol{w}_r},\boldsymbol{w}_r(t+1)-\boldsymbol{w}_r(t)\right\rangle\right|+\\&\quad\left|\frac{\partial\zeta_p(\boldsymbol{w}_r(t),a_r(t))}{\partial a_r}\cdot(a_r(t+1)-a_r(t))\right|+|\zeta_p(\boldsymbol{w}_r(t),a_r(t))|\\&\leq\ \frac{c_3}{\sqrt{mn_1}}\cdot\left(R'^2\cdot\widetilde{R}_w+R'^3\cdot\widetilde{R}_a\right),\end{aligned}\tag{91}$$

where the last inequality holds for a universal constant $c_3>0$ due to (90).

Case 1.3: $\mathbb{I}\left(\boldsymbol{w}_r(t+1)^\top\boldsymbol{y}_p>0\right)=\mathbb{I}\left(\boldsymbol{w}_r(t)^\top\boldsymbol{y}_p>0\right)=0$. Therefore, we easily obtain that

$$\begin{aligned}&\zeta_p(\boldsymbol{w}_r(t+1),a_r(t+1))-\zeta_p(\boldsymbol{w}_r(t),a_r(t))\\=\ &\left\langle\frac{\partial\zeta_p(\boldsymbol{w}_r(t),a_r(t))}{\partial\boldsymbol{w}_r},\boldsymbol{w}_r(t+1)-\boldsymbol{w}_r(t)\right\rangle+\frac{\partial\zeta_p(\boldsymbol{w}_r(t),a_r(t))}{\partial a_r}\cdot(a_r(t+1)-a_r(t))=0,\end{aligned}\tag{92}$$

i.e., $\chi_{pr}(t)=0$.

Case 2.1: $\mathbb{I}\left(\boldsymbol{w}_r(t+1)^\top\tilde{\boldsymbol{y}}_k>0\right)=\mathbb{I}\left(\boldsymbol{w}_r(t)^\top\tilde{\boldsymbol{y}}_k>0\right)=1$. Then

$$\begin{aligned}\xi_k(\boldsymbol{w}_r(t+1),a_r(t+1))-\xi_k(\boldsymbol{w}_r(t),a_r(t))=&\left\langle\frac{\partial\xi_k(\boldsymbol{w}_r(t),a_r(t))}{\partial\boldsymbol{w}_r},\boldsymbol{w}_r(t+1)-\boldsymbol{w}_r(t)\right\rangle\\&+\frac{\partial\xi_k(\boldsymbol{w}_r(t),a_r(t))}{\partial a_r}\cdot(a_r(t+1)-a_r(t))+\tilde{\chi}_{kr}(t),\end{aligned}$$

where

$$|\tilde{\chi}_{kr}(t)|\leq\frac{c_4}{\sqrt{mn_2}}\cdot\left(R'\cdot\widetilde{R}_w^2+R'^2\cdot\widetilde{R}_w\cdot\widetilde{R}_a\right),\tag{93}$$

for a universal constant $c_4>0$, since $\frac{\partial^2\xi_k(\boldsymbol{w}_r,a_r)}{\partial w_{ri}^2}$ is a first order polynomial of $\boldsymbol{w}_r$, $\frac{\partial^2\zeta_p(\boldsymbol{w}_r,a_r)}{\partial w_{ri}\partial a_r}$ is a polynomial of $\boldsymbol{w}_r$ with degree 2 and $\frac{\partial^2\zeta_p(\boldsymbol{w}_r,a_r)}{\partial a_r^2}=0$.

Case 2.2: $\mathbb{I}\left(\boldsymbol{w}_r(t+1)^\top\tilde{\boldsymbol{y}}_k>0\right)\neq\mathbb{I}\left(\boldsymbol{w}_r(t)^\top\tilde{\boldsymbol{y}}_k>0\right)$. Without the loss of generality, we assume that $\mathbb{I}\left(\boldsymbol{w}_r(t+1)^\top\tilde{\boldsymbol{y}}_k>0\right)=0$ and $\mathbb{I}\left(\boldsymbol{w}_r(t)^\top\tilde{\boldsymbol{y}}_k>0\right)=1$, denoted as E_2. Then it happens only if $\left|\boldsymbol{w}_r(0)^\top\tilde{\boldsymbol{y}}_k\right|<R_w$, with

$$\mathbb{P}\left(E_2\right)\leq\mathbb{P}\left(\left|\boldsymbol{w}_r(0)^\top\tilde{\boldsymbol{y}}_k\right|<R_w\right)=\mathbb{P}_{z\sim\mathcal{N}(0,1)}\left(|z|<R_w\right)\leq\frac{2R_w}{\sqrt{2\pi}}.$$

Furthermore,

$$\mathbb{P}(E_2|\,\|\boldsymbol{w}_r(0)\|_2\leq R')=\frac{\mathbb{P}(E_2,\|\boldsymbol{w}_r(0)\|_2\leq R')}{\mathbb{P}(\|\boldsymbol{w}_r(0)\|_2\leq R')}\leq\frac{\mathbb{P}\left(E_2\right)}{1-\delta}\leq\frac{8R_w}{\sqrt{2\pi}},\tag{94}$$

where the last inequality holds if we assume that $\delta<\frac{1}{2}$. Let the set $\widetilde{\mathcal{R}}_k^t(\boldsymbol{w}(0))$ be defined as

$$\widetilde{\mathcal{R}}_k^t(\boldsymbol{w}(0))=\left\{r\in[m]:\mathbb{I}\left(\boldsymbol{w}_r(t+1)^\top\tilde{\boldsymbol{y}}_k>0\right)=0\text{ and }\mathbb{I}\left(\boldsymbol{w}_r(t)^\top\tilde{\boldsymbol{y}}_k>0\right)=1\right\}.$$

Then, we have

$$\mathbb{E}_{\boldsymbol{w}(0)}\sum_{k=1}^{n_2}\left|\widetilde{\mathcal{R}}_k^t(\boldsymbol{w}(0))\right|\leq\sum_{k=1}^{n_2}\frac{8R_w}{\sqrt{2\pi}}\cdot m=\frac{8R_w\cdot mn_2}{\sqrt{2\pi}}.$$

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

$$\sum_{k=1}^{n_2}\left|\widetilde{\mathcal{R}}_k^t(\boldsymbol{w}(0))\right| \leq \frac{8R_w\cdot mn_2}{\sqrt{2\pi}\delta}.$$

Here, $\mathbb{I}\left(\boldsymbol{w}_r(t+1)^\top\tilde{\boldsymbol{y}}_k>0\right)\neq\mathbb{I}\left(\boldsymbol{w}_r(t)^\top\tilde{\boldsymbol{y}}_k>0\right)$ implies that

$$\left|\boldsymbol{w}_r(t)^\top\tilde{\boldsymbol{x}}_k\right| = \left|\left(\boldsymbol{w}_r(t)-\boldsymbol{w}_r(t+1)\right)^\top\tilde{\boldsymbol{y}}_k+\boldsymbol{w}_r(t+1)^\top\tilde{\boldsymbol{y}}_k\right| \leq \left\|\boldsymbol{w}_r(t)-\boldsymbol{w}_r(t+1)\right\|_2. \tag{95}$$

Let

$$\begin{aligned}
&\xi_k(\boldsymbol{w}_r(t+1),a_r(t+1))-\xi_k(\boldsymbol{w}_r(t),a_r(t))\\
=\;&\left\langle\frac{\partial\xi_k(\boldsymbol{w}_r(t),a_r(t))}{\partial\boldsymbol{w}_r},\boldsymbol{w}_r(t+1)-\boldsymbol{w}_r(t)\right\rangle+\frac{\partial\xi_k(\boldsymbol{w}_r(t),a_r(t))}{\partial a_r}\cdot(a_r(t+1)-a_r(t))+\tilde{\chi}_{kr}(t),
\end{aligned}$$

then

$$\begin{aligned}
|\tilde{\chi}_{kr}(t)| &\leq \left|\left\langle\frac{\partial\xi_k(\boldsymbol{w}_r(t),a_r(t))}{\partial\boldsymbol{w}_r},\boldsymbol{w}_r(t+1)-\boldsymbol{w}_r(t)\right\rangle\right| \\
&+ \left|\frac{\partial\xi_k(\boldsymbol{w}_r(t),a_r(t))}{\partial a_r}\cdot(a_r(t+1)-a_r(t))\right|+|\xi_k(\boldsymbol{w}_r(t),a_r(t))| \\
&\leq \frac{c_5}{\sqrt{mn_1}}\cdot\left(R'^2\cdot\widetilde{R}_w+R'^3\cdot\widetilde{R}_a\right),
\end{aligned} \tag{96}$$

where the last inequality holds for a universal constant $c_5>0$ since (95).

Case 2.3: $\mathbb{I}\left(\boldsymbol{w}_r(t+1)^\top\tilde{\boldsymbol{y}}_k>0\right)=\mathbb{I}\left(\boldsymbol{w}_r(t)^\top\tilde{\boldsymbol{y}}_k>0\right)=0$. Therefore, we easily obtain that

$$\begin{aligned}
&\xi_k(\boldsymbol{w}_r(t+1),a_r(t+1))-\xi_k(\boldsymbol{w}_r(t),a_r(t))\\
=\;&\left\langle\frac{\partial\xi_k(\boldsymbol{w}_r(t),a_r(t))}{\partial\boldsymbol{w}_r},\boldsymbol{w}_r(t+1)-\boldsymbol{w}_r(t)\right\rangle+\frac{\partial\xi_k(\boldsymbol{w}_r(t),a_r(t))}{\partial a_r}\cdot(a_r(t+1)-a_r(t))=0,
\end{aligned} \tag{97}$$

i.e., $\tilde{\chi}_{kr}(t)=0$.

Let $\boldsymbol{\chi}(t)=\left[\chi_1(t)\ \cdots\ \chi_{n_1}(t)\right]^\top$ and $\tilde{\boldsymbol{\chi}}(t)=\left[\tilde{\chi}_1(t)\ \cdots\ \tilde{\chi}_{n_2}(t)\right]^\top$. Combining with above six cases, we have

$$\begin{aligned}
&\left\|\begin{pmatrix}\boldsymbol{\chi}(t)\\ \tilde{\boldsymbol{\chi}}(t)\end{pmatrix}\right\|_2 \\
\leq\;& \sqrt{n_1}\cdot m\cdot\frac{c_2}{\sqrt{mn_1}}\cdot\left(R'\cdot\widetilde{R}_w^2+R'^2\cdot\widetilde{R}_w\cdot\widetilde{R}_a\right)+\sqrt{n_2}\cdot m\cdot\frac{c_4}{\sqrt{mn_2}}\cdot\left(R'\cdot\widetilde{R}_w^2+R'^2\cdot\widetilde{R}_w\cdot\widetilde{R}_a\right)+\\
&\sum_{p=1}^{n_1}\left|\mathcal{R}_p^t(\boldsymbol{w}(0))\right|\cdot\frac{c_3}{\sqrt{mn_1}}\cdot\left(R'^2\cdot\widetilde{R}_w+R'^3\cdot\widetilde{R}_a\right)+\sum_{k=1}^{n_2}\left|\widetilde{\mathcal{R}}_k^t(\boldsymbol{w}(0))\right|\cdot\frac{c_5}{\sqrt{mn_1}}\cdot\left(R'^2\cdot\widetilde{R}_w+R'^3\cdot\widetilde{R}_a\right)\\
\leq\;& \tilde{c}_0\cdot\eta\cdot\left(\frac{\sqrt{n_1+n_2}}{\delta\cdot(\lambda_0+\widetilde{\lambda}_0)\cdot\sqrt{m}}\right)R'^8\cdot\left\|\begin{bmatrix}\boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t))\\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t))\end{bmatrix}\right\|_2\cdot\left\|\begin{bmatrix}\boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0))\\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0))\end{bmatrix}\right\|_2+\\
&\tilde{c}_1\cdot\eta^2\cdot\frac{R'^7}{\sqrt{m}}\cdot\left\|\begin{bmatrix}\boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t))\\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t))\end{bmatrix}\right\|_2^2,
\end{aligned} \tag{98}$$

for some universal constants $\tilde{c}_0>0$ and $\tilde{c}_1>0$ and with the assumption that $R'\geq 1$.

C.3. Proof for Lemma 4.3

By (86) and the Hoeffding's inequality, we have

$$\mathbb{P}\left(\frac{1}{m}\sum_{r=1}^m\|\boldsymbol{w}_r(0)\|_2^4-\mathbb{E}\|\boldsymbol{w}_r\|_2^4>\epsilon\right)\leq\exp\left(-\frac{m\cdot\epsilon^2}{2R'^8}\right).$$

Taking $\epsilon = 1$, with the probability of at least $1 - \frac{\delta}{2}$, we have

$$\frac{1}{m}\sum_{r=1}^{m}\|\boldsymbol{w}_r(0)\|_2^4 \leq \mathbb{E}\|\boldsymbol{w}_r\|_2^4 + 1,$$

if $m = \widetilde{\Omega}\left(\left(\log\frac{1}{\delta}\right)^5\right)$. Now we let R_w is small enough such that

$$\begin{aligned}
& \frac{1}{m}\sum_{r=1}^{m}\|\boldsymbol{w}_r(t+1)\|_2^4 = \frac{1}{m}\sum_{r=1}^{m}\|\boldsymbol{w}_r(t+1) - \boldsymbol{w}_r(0) + \boldsymbol{w}_r(0)\|_2^4 \\
\leq \quad & \frac{1}{m}\sum_{r=1}^{m}\|\boldsymbol{w}_r(t+1) - \boldsymbol{w}_r(0)\|_2^4 + 4\|\boldsymbol{w}_r(t+1) - \boldsymbol{w}_r(0)\|_2^3 \cdot \|\boldsymbol{w}_r(0)\|_2 + \\
& 6\|\boldsymbol{w}_r(t+1) - \boldsymbol{w}_r(0)\|_2^2 \cdot \|\boldsymbol{w}_r(0)\|_2^2 + 4\|\boldsymbol{w}_r(t) - \boldsymbol{w}_r(0)\|_2 \cdot \|\boldsymbol{w}_r(0)\|_2^3 + \|\boldsymbol{w}_r(0)\|_2^4 \\
\leq \quad & R_w \cdot \left(R_w^3 + 4R_w^2 R' + 6R_w R'^2 + 4R'^3\right) + C_0 \leq 2C_0.
\end{aligned} \tag{99}$$

Here the last inequality requires that $R_w \cdot \left(R_w^3 + 4R_w^2 R' + 6R_w R'^2 + 4R'^3\right) \leq C_0$ and using Lemma 4.1, we need $m = \widetilde{\Omega}\left(\frac{\left(\log(\frac{1}{\delta})\right)^4}{(\lambda_0+\widetilde{\lambda}_0)^2} \cdot \left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix}\right\|_2^2\right)$. Similarly, if $m = \widetilde{\Omega}\left(\frac{\left(\log(\frac{1}{\delta})\right)^6}{(\lambda_0+\widetilde{\lambda}_0)^2} \cdot \left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix}\right\|_2^2\right)$, then $\frac{1}{m}\sum_{r=1}^{m}\|\boldsymbol{w}_r(t+1)\|_2^6 \leq 2C_1$.

C.4. Proof for Theorem 4.5

Note that there exists a universal constant $c_1 > 0$ such that

$$\left\|\frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r}\right\|_2 \leq \frac{c_1}{\sqrt{mn_1}} \cdot \|\boldsymbol{w}_r\|_2^2, \quad \left\|\frac{\partial h_k(\boldsymbol{w},\boldsymbol{a})}{\partial \boldsymbol{w}_r}\right\|_2 \leq \frac{c_1}{\sqrt{mn_2}} \cdot \|\boldsymbol{w}_r\|_2^2,$$

and

$$\left|\frac{\partial s_p(\boldsymbol{w},\boldsymbol{a})}{\partial a_r}\right| \leq \frac{c_1}{\sqrt{mn_1}} \cdot \|\boldsymbol{w}_r\|_2^3, \quad \left|\frac{\partial h_k(\boldsymbol{w},\boldsymbol{a})}{\partial a_r}\right| \leq \frac{c_1}{\sqrt{mn_2}} \cdot \|\boldsymbol{w}_r\|_2^3,$$

if $|a_r| \leq 2$, for all $r \in [m]$. Assume that the result (27) holds for $\tau = 0, \cdots, t$, we then further prove that it also holds for $\tau = t+1$. Therefore, (27) holds for all $t \in \mathbb{N}$ by induction. Recall that $\mathcal{L}(\boldsymbol{w}(t),\boldsymbol{a}(t)) = \frac{1}{2}\left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \end{bmatrix}\right\|_2^2$, then we have

$$\begin{aligned}
& \left\|\frac{\partial \mathcal{L}(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}_r}\right\|_2 \\
& \leq \sum_{p=1}^{n_1}|s_p(\boldsymbol{w}(t),\boldsymbol{a}(t))| \cdot \left\|\frac{\partial s_p(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}_r}\right\|_2 + \sum_{k=1}^{n_2}|h_k(\boldsymbol{w}(t),\boldsymbol{a}(t))| \cdot \left\|\frac{\partial h_k(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial \boldsymbol{w}_r}\right\|_2 \\
& \leq \sqrt{\frac{2}{m}}c_1 \cdot \|\boldsymbol{w}_r(t)\|_2^2 \cdot \left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \end{bmatrix}\right\|_2 \leq \sqrt{\frac{2}{m}}c_1 \cdot \|\boldsymbol{w}_r(t)\|_2^2 \cdot \left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix}\right\|_2,
\end{aligned} \tag{100}$$

and

$$\begin{aligned}
& \left|\frac{\partial \mathcal{L}(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial a_r}\right| \\
& \leq \sum_{p=1}^{n_1}s_p(\boldsymbol{w}(t),\boldsymbol{a}(t)) \cdot \left|\frac{\partial s_p(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial a_r}\right| + \sum_{k=1}^{n_2}h_k(\boldsymbol{w}(t),\boldsymbol{a}(t)) \cdot \left|\frac{\partial h_k(\boldsymbol{w}(t),\boldsymbol{a}(t))}{\partial a_r}\right| \\
& \leq \frac{2c_1 \cdot \|\boldsymbol{w}_r(t)\|_2^3}{\sqrt{m}} \cdot \left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t),\boldsymbol{a}(t)) \end{bmatrix}\right\|_2 \leq \frac{2c_1 \cdot \|\boldsymbol{w}_r(t)\|_2^3}{\sqrt{m}} \cdot \left\|\begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0),\boldsymbol{a}(0)) \end{bmatrix}\right\|_2.
\end{aligned} \tag{101}$$

Then, with the probability of at least $1 - \frac{\delta}{3}$, we have

$$\begin{aligned}
&\left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t+1), \boldsymbol{a}(t+1)) \\ \boldsymbol{h}(\boldsymbol{w}(t+1), \boldsymbol{a}(t+1)) \end{bmatrix} - \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t), \boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t), \boldsymbol{a}(t)) \end{bmatrix} \right\|_2 \\
\leq& \sqrt{n_1} \cdot \frac{c_1 \cdot \sqrt{2C_0}}{\sqrt{n_1}} \cdot \|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\|_2 + \sqrt{n_1} \cdot \frac{c_1 \cdot \sqrt{2C_1}}{\sqrt{n_1}} \cdot \|\boldsymbol{a}(t+1) - \boldsymbol{a}(t)\|_2 \\
&+ \sqrt{n_2} \cdot \frac{c_1 \cdot \sqrt{2C_0}}{\sqrt{n_2}} \cdot \|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\|_2 + \sqrt{n_2} \cdot \frac{c_1 \cdot \sqrt{2C_1}}{\sqrt{n_2}} \cdot \|\boldsymbol{a}(t+1) - \boldsymbol{a}(t)\|_2 \\
\leq& 2c_1 \cdot \sqrt{2C_0} \cdot \|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\|_2 + 2c_1 \cdot \sqrt{2C_1} \cdot \|\boldsymbol{a}(t+1) - \boldsymbol{a}(t)\|_2 \\
\leq& c_2 \cdot \eta \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t), \boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t), \boldsymbol{a}(t)) \end{bmatrix} \right\|_2,
\end{aligned}$$

if

$$m = \widetilde{\Omega}\left(\frac{\left(\log(\frac{1}{\delta})\right)^6}{(\lambda_0 + \widetilde{\lambda}_0)^2} \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0), \boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0), \boldsymbol{a}(0)) \end{bmatrix} \right\|_2^2 \right) \text{ and } m = \widetilde{\Omega}\left(\left(\log \frac{1}{\delta} \right)^7 \right).$$

Here, the first inequality holds with the probability of at least $1 - \frac{\delta}{6}$ (by Lemma 4.3), because of the mean value theorem with (100) and (101). The third inequality comes from Lemma 4.4 with the probability of at least $1 - \frac{\delta}{6}$. Moreover, according to Lemma 4.2, if

$$m = \widetilde{\Omega}\left(\frac{(n_1 + n_2)}{\delta^2 \cdot (\lambda_0 + \widetilde{\lambda}_0)^2} \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0), \boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0), \boldsymbol{a}(0)) \end{bmatrix} \right\|_2^2 \right),$$

then with the probability of at least $1 - \frac{\delta}{6}$, we have

$$\left\| \begin{pmatrix} \boldsymbol{\chi}(t) \\ \tilde{\boldsymbol{\chi}}(t) \end{pmatrix} \right\|_2 \leq c_3 \cdot (\eta + \eta^2) \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(t), \boldsymbol{a}(t)) \\ \boldsymbol{h}(\boldsymbol{w}(t), \boldsymbol{a}(t)) \end{bmatrix} \right\|_2,$$

for a small universal constant $c_3 > 0$. Moreover, by Lemma 3.5, with the probability of at least $1 - \frac{\delta}{6}$, the initialized Gram matrices are positive definite, i.e., $\lambda_{\min}(\boldsymbol{G}(\boldsymbol{w}(0), \boldsymbol{a}(0))) \geq \frac{3}{4}\lambda_0$ and $\lambda_{\min}\left(\widetilde{\boldsymbol{G}}\left(\boldsymbol{w}(0), \boldsymbol{a}(0)\right)\right) \geq \frac{3}{4}\widetilde{\lambda}_0$, if

$$m = \widetilde{\Omega}\left(\frac{(n_1 + n_2)^4}{(n_1 n_2)^2 \cdot \left(\min\{\lambda_0, \widetilde{\lambda}_0\}\right)^2} \cdot \left(\log \frac{1}{\delta} \right)^7 \right).$$

To guarantee the positive definiteness of Gram matrices $\boldsymbol{G}(\boldsymbol{w}(t), \boldsymbol{a}(t))$ and $\widetilde{\boldsymbol{G}}(\boldsymbol{w}(t), \boldsymbol{a}(t))$, i.e.,

$$\lambda_{\min}\left(\boldsymbol{G}(\boldsymbol{w}(t), \boldsymbol{a}(t)) + \widetilde{\boldsymbol{G}}(\boldsymbol{w}(t), \boldsymbol{a}(t)) \right) \geq \frac{\lambda_0 + \widetilde{\lambda}_0}{2},$$

R_w and R_a in Lemma 4.1 should satisfies conditions in Lemma 3.6 (with the probability of at least $1 - \frac{\delta}{6}$), therefore, we require

$$m = \widetilde{\Omega}\left(\frac{(n_1 + n_2)^2}{\left(\lambda_0 + \widetilde{\lambda}_0\right)^2 \cdot \left(\min\{\lambda_0, \widetilde{\lambda}_0\}\right)^2 \cdot \delta^2} \cdot \left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0), \boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0), \boldsymbol{a}(0)) \end{bmatrix} \right\|_2^2 \right).$$

Note that Lemma 3.7 shows that $\left\| \begin{bmatrix} \boldsymbol{s}(\boldsymbol{w}(0), \boldsymbol{a}(0)) \\ \boldsymbol{h}(\boldsymbol{w}(0), \boldsymbol{a}(0)) \end{bmatrix} \right\|_2^2 = \mathcal{O}(\frac{1}{\delta})$, with the probability of at least $1 - \frac{\delta}{6}$.

To simplify the formulation and improve the readability, we slightly change some notations here, i.e., $\boldsymbol{s}(\boldsymbol{w}(t), \boldsymbol{a}(t)) := \boldsymbol{s}^t$

and $\boldsymbol{G}(\boldsymbol{w}(t), \boldsymbol{a}(t)) := \boldsymbol{G}^t$ (similarly for $\boldsymbol{h}^t$ and $\widetilde{\boldsymbol{G}}^t$). Combining with aforementioned results, we have

$$\begin{aligned}
&\left\|\begin{bmatrix} \boldsymbol{s}^{t+1} \\ \boldsymbol{h}^{t+1} \end{bmatrix}\right\|_2^2 = \left\|\begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix} + \left(\begin{bmatrix} \boldsymbol{s}^{t+1} \\ \boldsymbol{h}^{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}\right)\right\|_2^2 \\
&= \left\|\begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}\right\|_2^2 + 2\begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}^\top \cdot \left(\begin{bmatrix} \boldsymbol{s}^{t+1} \\ \boldsymbol{h}^{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}\right) + \left\|\begin{bmatrix} \boldsymbol{s}^{t+1} \\ \boldsymbol{h}^{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}\right\|_2^2 \\
&= \left\|\begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}\right\|_2^2 - 2\eta \cdot \begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}^\top \cdot \left(\boldsymbol{G}^t + \widetilde{\boldsymbol{G}}^t\right) \cdot \begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix} \\
&\quad - 2\eta \cdot \begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}^\top \cdot \begin{bmatrix} \boldsymbol{\chi}^{t} \\ \tilde{\boldsymbol{\chi}}^{t} \end{bmatrix} + \left\|\begin{bmatrix} \boldsymbol{s}^{t+1} \\ \boldsymbol{h}^{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}\right\|_2^2 \\
&\leq \left(1 - 2\eta \cdot \frac{\lambda_0 + \widetilde{\lambda_0}}{2}\right) \cdot \left\|\begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}\right\|_2^2 + 2\eta \cdot \left\|\begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}\right\|_2 \cdot \left\|\begin{bmatrix} \boldsymbol{\chi}^{t} \\ \tilde{\boldsymbol{\chi}}^{t} \end{bmatrix}\right\|_2 \\
&\quad + \left\|\begin{bmatrix} \boldsymbol{s}^{t+1} \\ \boldsymbol{h}^{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}\right\|_2^2 \\
&\leq \left(1 - 2\eta \cdot \frac{\lambda_0 + \widetilde{\lambda_0}}{2}\right) \cdot \left\|\begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}\right\|_2^2 + 2\eta \cdot c_3(\eta + \eta^2) \cdot \left\|\begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}\right\|_2^2 \\
&\quad + c_2^2 \cdot \eta^2 \cdot \left\|\begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}\right\|_2^2 \\
&\leq \left(1 - \eta \cdot \frac{\lambda_0 + \widetilde{\lambda_0}}{2}\right) \cdot \left\|\begin{bmatrix} \boldsymbol{s}^{t} \\ \boldsymbol{h}^{t} \end{bmatrix}\right\|_2^2,
\end{aligned} \tag{102}$$

where the last inequality holds when $\eta = \mathcal{O}\left(\lambda_0 + \widetilde{\lambda}_0\right)$ such that $2\eta \cdot c_3(\eta + \eta^2) + c_2^2\eta^2 \leq \eta \cdot \frac{\lambda_0 + \widetilde{\lambda}_0}{2}$.