Guided by the Experts: Provable Feature Learning Dynamic of Soft-Routed Mixture-of-Experts

Fangshuo Liao[1]
[1]Computer Science Department, Rice University

Anastasios Kyrillidis[1,2]
[2]Ken Kennedy Institute, Rice University

Abstract

Mixture-of-Experts (MoE) architectures have emerged as a cornerstone of modern AI systems. In particular, MoEs route inputs dynamically to specialized experts whose outputs are aggregated through weighted summation. Despite their widespread application, theoretical understanding of MoE training dynamics remains limited to either separate expert-router optimization or only top-1 routing scenarios with carefully constructed datasets. This paper advances MoE theory by providing convergence guarantees for joint training of soft-routed MoE models with non-linear routers and experts in a student-teacher framework. We prove that, with moderate over-parameterization, the student network undergoes a feature learning phase, where the router's learning process is "guided" by the experts, that recovers the teacher's parameters. Moreover, we show that a post-training pruning can effectively eliminate redundant neurons, followed by a provably convergent fine-tuning process that reaches global optimality. To our knowledge, our analysis is the first to bring novel insights in understanding the optimization landscape of the MoE architecture.

1 INTRODUCTION

Mixture-of-Experts (MoE) architectures have become a fundamental building block in modern artificial intelligence systems, enabling significant advances in model capacity without corresponding increases in computational costs (Shazeer et al., 2017; Fedus et al., 2022). At its core, a MoE system treats a complicated task as a combination of multiple simpler tasks, which can be handled efficiently by smaller models. In concept, the conditional routing of the input to different submodules of the MoE allows each sub-module to "specialize" in its own domain, leading to an effective decoupling of the overall task complexity.[1] This approach has achieved remarkable success in large language models (LLMs) (Fedus et al., 2022), computer vision (Riquelme et al., 2021), and multi-modal agentic systems (Mustafa et al., 2022), where conditional computation provides an efficient way to scale model capacity.

The fundamental structure of an MoE layer consists of a set of expert networks (the "experts") and a gating network that determines the contribution of each expert to the final output. While simple, this architecture presents theoretical challenges, particularly regarding the joint optimization of both components. The gating function, typically implemented using a softmax activation, introduces non-convexity that makes the analysis difficult. This is further complicated by the interplay between expert specialization and router assignments: experts must specialize in certain inputs, while the router must correctly identify the correct combination of experts appropriate for each input.

Despite the widespread adoption of MoE architectures in practice, such a theoretical understanding of their optimization dynamic remains limited; see Related Works section. Existing theoretical work has focused on either simplified linear models or has analyzed the experts and gating networks separately; e.g. Li et al. (2025) and Kawata et al. (2025) study the setting where the experts are trained first with the router parameter fixed, followed by a fine-tuning stage of the router itself. While such a setting simplifies the analysis by decoupling the updates of the router and expert parameters, it deviates from the more beneficial sce-

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

[1]In this paper we consider a "classic" MoE instead of the MoE used to increase parameterization without increasing computational costs such as Fedus et al. (2022).

nario, where a joint optimization is applied to handle intricate task combinations (Kong et al., 2025; Zhang et al., 2025a).

A prior work (Chen et al., 2022) studies the joint optimization of the expert and router parameters in a top-1 routed MoE on patched input data, thus reducing the interference between the learning process of the experts in each gradient step. Although top-1 routing has been a popular approach (Fedus et al., 2022), most state-of-the-art language models such as Mixtral 8x7B (Jiang et al., 2024), DeepSeek-V3 (DeepSeek-AI et al., 2025), and Qwen 3 (Yang et al., 2025) uses a top-K routing with $K > 1$, leading to gaps between theory and practice.

In general, it remains open to study the optimization dynamics of MoEs with more than one activated experts, where experts and router are jointly trained. This gap is increasingly significant as MoE architectures become fundamental components in state-of-the-art AI systems. Analysis of these dynamics would not only inform architectural improvements but also provide formal guarantees about model behavior and performance. Moreover, such an analysis could help us understand better Agentic AI systems (Hu et al., 2025; Zhang et al., 2025b,c), where component orchestration mirrors MoE routing mechanisms (Bhatt et al., 2025). These systems must dynamically select appropriate specialized modules (tools, APIs, or reasoning components) based on input context—functionally analogous to expert selection in MoE architectures.

Contributions. Given the difficulty of the task, we focus on the learning of a MoE model with one-layer sigmoid router and non-linear experts over the mean-square-error (MSE) loss in a teacher-student set-up on high dimensional Gaussian input. In particular, we show that, with moderate over-parameterization and under the gradient flow training, the student MoE enjoys a near-perfect recovery of the feature from the teacher model's in a sequential order in $\mathcal{O}\left(\sqrt{d}\right)$ time, where d is the dimensionality of the input. Moreover, after the feature learning stage, a greedy pruning can be applied to remove the unused experts. Lastly, the post-pruning fine-tuning of the student model converges to zero loss.

Notations. Without further specification, we use regular lower-case letters (e.g. a) to denote scalars, bold-face lower-case letters (e.g. $\mathbf{a}$) to denote vectors, and bold-face capital letters (e.g. $\mathbf{A}$) to denote matrices. We use $\mathcal{N}\left(\mu, \sigma^2\right)$ to denote the the Gaussian distribution with mean μ and (co)variance σ^2. For a function $f(x)$, we use $f'(x), f''(x)$ and $f'''(x)$ to denote its first three order derivatives, and $f^{(a)}(x)$ to denote is arbitray ath order derivative. We use $\text{poly}\left(x_1, \ldots, x_n\right)$ to denote the polynomial dependency in terms of $x_1, \ldots, x_n$.

2 RELATED WORKS

Theory of Mixture-of-Experts. From an optimization perspective, Chen et al. (2022) studies the convergence rate of top-1 MoE with CNN experts on patched input data. Chowdhury et al. (2023) studies the patch-level routing under both the setting with a separately trained expert and router, and the setting of pre-trained experts. Chowdhury et al. (2024) shows the pruning effectiveness after fine-tuning a pre-trained MoE model. Kawata et al. (2025) considers the training both a top-1 MoE and a ReLU routed MoE, but under a four-stage training algorithm. Fruytier et al. (2025) studies the convergence of the Expectation-Maximization algorithm for learning MoE. Li et al. (2025) studies the optimization of MoE in a continual learning set-up. From the perspective of sample complexity, Nguyen et al. (2024a,c, 2025) studies the sample complexity of correctly identifying experts for softmax MoE under both the logistic loss and the MSE loss. Nguyen et al. (2024b) shows that sigmoid gated MoEs enjoy a better sample complexity compared with softmax gated MoEs. Other works (Kratsios et al., 2024; Wang and E, 2025) studies MoE under operator learning, and the expressive power of MoEs, respectively. Following the expressivity line of work, Boix-Adsera and Rigollet (2025) studies how the granularity of the experts affects the expressive power of MoEs.

Feature Learning of Neural Networks. As its name suggested, the feature learning framework explores the ability of the neural network to learn intrinsic features of the dataset, which is an ability not present in the traditional Neural Tangent Kernel framework (Jacot et al., 2020; Du et al., 2019). In particular Shi et al. (2022, 2023) studies the hidden-neuron evolution during training, and Damian et al. (2022); Mousavi-Hosseini et al. (2023a) investigates how gradient-based learning discovers the intrinsic low-dimensional subspace of data. Along this line of work Ba et al. (2023); Mousavi-Hosseini et al. (2023b) studies the learning with data sampled from distribution with a spiked covariance matrix. A related line of work studies the feature learning dynamics of two-layer ReLU networks, including the characterization of gradient flow dynamics under orthogonal inputs (Boursier et al., 2022), the phase diagram at infinite width (Luo et al., 2021), quantization of learned features under gradient descent (Maennel et al., 2018), early alignment phenomena and their implications for robustness and generalization (Boursier and Flammarion, 2025; Min et al., 2023), as well as implicit biases of gradient descent leading to feature averaging (Li et al., 2024).

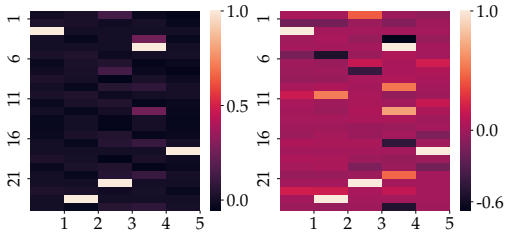


Figure 1: Training MoE in (1) on (3) with $m^\star = 5, m = 25$, and $d = 1000$ with online batch SGD simulating GF on the population loss. Left: alignment values of the router parameters $\bar{\mathbf{v}}_i^\top \bar{\mathbf{v}}_j^\star$. Right: alignment values of the expert parameters $\bar{\mathbf{w}}_i^\top \bar{\mathbf{w}}_j^\star$.

Recently, a popular line of work studies the learning of Gaussian single/multi-index models (Bietti et al., 2022; Lee et al., 2024; Ren et al., 2025; Bietti et al., 2023; Ba et al., 2023; Şimşek et al., 2025). Noticeably, this line of work adopts the Hermite expansion of the nonlinear function to transform the loss objective into a form similar to the tensor decomposition (Ge et al., 2017). In terms of the proof technique, our work is similar to Ren et al. (2025) by utilizing the sharp phase transition that occurs from the high information exponent of the activation function.

3 PRELIMINARY AND SET-UP

Student model. In this paper, we consider the training of a normalized MoE with m experts. In particular, given inputs $\mathbf{x} \in \mathbb{R}^d$, we study the setting of a one-layer router with parameter $\mathbf{V}$, given by $\pi\left(\bar{\mathbf{V}}\mathbf{x}\right)$. Here $\pi(\cdot)$ is an entry-wise sigmoid function, and $\bar{\mathbf{V}}$ denote the row-wise normalized version of $\mathbf{V}$. In short, we have

$$\pi\left(\bar{\mathbf{V}}\mathbf{x}\right)_i := \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) = \pi\left(\frac{\mathbf{v}_i^\top \mathbf{x}}{\|\mathbf{v}_i\|_2}\right); \forall i \in [m]$$

where $\mathbf{v}_i$ is the ith row of $\mathbf{V}$. We consider each expert as a one-layer non-linear function with parameter $\mathbf{w}_i$ given by $\sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) = \sigma\left(\frac{\mathbf{w}_i^\top \mathbf{x}}{\|\mathbf{w}_i\|_2}\right)$. In this paper, we choose $\sigma(a) = a^3 - 3a$ to be the third-order Hermite polynomial. Letting $\boldsymbol{\theta} = \{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^m$, then the student model is given by

$$f(\boldsymbol{\theta}, \mathbf{x}) := \sum_{i=1}^m \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \tag{1}$$

Our choice of the sigmoid router is motivated by Nguyen et al. (2024b) which shows that sigmoid routing is more sample efficient than softmax routing. Moreover, using Hermite polynomial as the activation function has been a popular approach in previous study of the feature learning mechanism of neural networks (Arous et al., 2025). Lastly, the choice of normalizing the weights is also a popular choice in prior works (Wang et al., 2020; Ren et al., 2025).

Data and Teacher Model. The teacher model $f^\star$ we consider has the same structure as in the the student model, but with $m^\star$ experts. In addition, we assume that the parameter of the teacher model's parameters $\bar{\mathbf{v}}_1^\star, \ldots, \bar{\mathbf{v}}_{m^\star}^\star, \bar{\mathbf{w}}_1^\star, \ldots, \bar{\mathbf{w}}_{m^\star}^\star$ satisfy the following assumption:

Assumption 1 (Teacher Orthonormality). *The teacher model's parameters $\bar{\mathbf{v}}_1^\star, \ldots, \bar{\mathbf{v}}_{m^\star}^\star, \bar{\mathbf{w}}_1^\star, \ldots, \bar{\mathbf{w}}_{m^\star}^\star$ form an orthonormal list.*

We consider the input data that come from a standard Gaussian distribution $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and labels are generated by the teacher model according to

$$y = f^\star(\mathbf{x}) := \sum_{i=1}^{m^\star} \pi\left(\bar{\mathbf{v}}_i^{\star\top} \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^{\star\top} \mathbf{x}\right) \tag{2}$$

Intuitively, this set-up implies that the input space $\mathbb{R}^d$ is softly partitioned by the teacher's router. The goal of the student model is to learn both the features $\bar{\mathbf{v}}_i^\star$s that gives a correct partitions, and the features $\bar{\mathbf{w}}_i^\star$s that leads to the effective specialization of experts.

We consider training the student model $f(\boldsymbol{\theta}, \mathbf{x})$ on the population mean-squared error (MSE) loss $\mathcal{L}(\boldsymbol{\theta})$ using gradient flow $\frac{d}{dt}\boldsymbol{\theta}(t) = -\nabla \mathcal{L}(\boldsymbol{\theta}(t))$ over the data distribution defined by the teacher model. To be more specific, the MSE has the form

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{2}\mathbb{E}_{\mathbf{x},y}\left[\left(f(\boldsymbol{\theta}, \mathbf{x}) - y\right)^2\right] \\ &= \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\left(f(\boldsymbol{\theta}, \mathbf{x}) - f^\star(\mathbf{x})\right)^2\right] \end{aligned} \tag{3}$$

We assume that the parameters of the student model are initialized according to the following assumption:

Assumption 2 (Initialization). *The expert weights are initialized as $\mathbf{w}_i(0) \sim \mathcal{N}\left(\mathbf{0}, d^{-1}\mathbf{I}_d\right)$. The router weights are initialized as $\mathbf{v}_i(0) = \hat{\mathbf{v}}_i(0) - \hat{\mathbf{v}}_i(0)^\top \bar{\mathbf{w}}_i(0) \cdot \bar{\mathbf{w}}_i(0)$, where $\hat{\mathbf{v}}_i(0) \sim \mathcal{N}\left(\mathbf{0}, d^{-1}\mathbf{I}_d\right)$, so as to decouple the router and expert weights at initialization.*[2]

Lastly, we make the following assumption on the sigmoid function, which is numerically checked in Appendix E.

Assumption 3. *Let $z_1, z_2 \sim \mathcal{N}(0,1)$ with arbitrary covariance $Cov(z_1, z_2) \in [-1, 1]$, it holds that $\mathbb{E}_{z_1,z_2}\left[\pi'(z_1)\pi^{(3)}(z_2)\right] \leq 0$*

We empirically verify that this setting allows that each of the $m^\star$ experts and corresponding router parameter

[2]The behavior that $\mathbf{v}_i(t)^\top \mathbf{w}_i(t) = 0$ does not hold throughout gradient flow training.

in the teacher model can be recovered by one and only one expert and corresponding router in the student model. According to Figure 1, for each $\bar{\mathbf{v}}_j^\star, \bar{\mathbf{w}}_j^\star$ in the teacher (x-axis), there is one and only one expert and router $\bar{\mathbf{v}}_j, \bar{\mathbf{w}}_j$ that converges to $\bar{\mathbf{v}}_j^\star, \bar{\mathbf{w}}_j^\star$ (lighter color indicates that $\bar{\mathbf{v}}_i^\top \bar{\mathbf{v}}_j^\star$ and $\bar{\mathbf{w}}_i^\top \bar{\mathbf{w}}_j^\star$ are closer to one.)

4 MAIN RESULT: SEQUENTIAL FEATURE LEARNING

In this section, we present the main result of the feature learning phase. Recall the set-up of the student and teacher MoE models in (1) and (2). An ideal feature learning result would be that, for each router-expert pair $(\bar{\mathbf{v}}_i^\star, \bar{\mathbf{w}}_i^\star)$ in the teacher model, there is an exclusive router-expert pair $(\bar{\mathbf{v}}_i, \bar{\mathbf{w}}_i)$ that converges to it. The theorem below states that the matching between the router-expert pair from the teacher model and the router-expert pair from the student model happens in a sequential order.

Theorem 1. *Consider training the MoE model $f(\boldsymbol{\theta}, \mathbf{x})$ in (1) with respect to a teacher model given by (2) using the gradient flow on the population MSE loss in (3). Let $\delta_{\mathbb{P}} \in (0, 1/7)$ be given. If $m \geq \Omega\left(m^\star \log \frac{m^\star}{\delta_{\mathbb{P}}}\right)$ and $d \geq poly\left(m, \delta_{\mathbb{P}}^{-1}\right)$, then there exists an injective mapping $\mathcal{I} : [m^\star] \to [m]$ and time steps $0 \leq T_1 \leq \cdots \leq T_{m^\star} \leq T^\star \leq \mathcal{O}\left(\sqrt{d}\right)$ such that for all $\ell \in [m^\star - 1]$ and $t \in [T_\ell, T_{\ell+1})$, we have that*

- *(Recovered expert-router pairs) $\bar{\mathbf{v}}_{\mathcal{I}(i)}(t)^\top \bar{\mathbf{v}}_i^\star \geq 0.9$ and $\bar{\mathbf{w}}_{\mathcal{I}(i)}(t)^\top \bar{\mathbf{w}}_i^\star \geq 0.9$ for all $i \leq \ell$.*
- *(Unrecovered expert-router pairs) For all $i > \ell$, $\max\left\{\left|\bar{\mathbf{v}}_{\mathcal{I}(i)}(t)^\top \bar{\mathbf{v}}_i^\star\right|, \left|\bar{\mathbf{w}}_{\mathcal{I}(i)}(t)^\top \bar{\mathbf{w}}_i^\star\right|\right\} \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$*

Moreover, for $T^\star \leq t \leq T^\star + \mathcal{O}\left(\frac{\delta_{\mathbb{P}}\sqrt{d}}{m^2}\right)$, we have that

- *(Learned features) $\bar{\mathbf{v}}_{\mathcal{I}(i)}(t)^\top \bar{\mathbf{v}}_i^\star, \bar{\mathbf{w}}_{\mathcal{I}(i)}(T)^\top \bar{\mathbf{w}}_i^\star \geq 1 - \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)$ for all $i \in [m^\star]$*
- *(Unused expert-router pairs) for all $i_1 \in [m] \setminus \mathcal{I}([m^\star]), i_2 \in [m], i_1 \neq i_2$ and $j \in [m^\star]$, the following quantities*

$$\left|\bar{\mathbf{v}}_{i_1}(t)^\top \bar{\mathbf{v}}_{i_2}(t)\right|, \left|\bar{\mathbf{v}}_{i_1}(t)^\top \bar{\mathbf{w}}_{i_1}(t)\right|, \left|\bar{\mathbf{w}}_{i_1}(t)^\top \bar{\mathbf{w}}_{i_2}(t)\right|$$
$$\left|\bar{\mathbf{v}}_{i_1}(t)^\top \bar{\mathbf{w}}_{i_2}(t)\right|, \left|\bar{\mathbf{v}}_{i_2}(t)^\top \bar{\mathbf{w}}_{i_i}(t)\right|$$
$$\left|\bar{\mathbf{v}}_{i_1}(t)^\top \bar{\mathbf{v}}_j^\star\right|, \left|\bar{\mathbf{v}}_{i_1}(t)^\top \bar{\mathbf{w}}_j^\star\right|, \left|\bar{\mathbf{w}}_{i_1}(t)^\top \bar{\mathbf{w}}_j^\star\right|, \left|\bar{\mathbf{w}}_{i_1}(t)^\top \bar{\mathbf{v}}_j^\star\right|$$

are all upper bounded by $\mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$

In particular, Theorem 1 states the result that the MoE training in our set-up undergoes a sequential feature learning phase. As requirements of the theorem, we need m to be as large as $\Omega\left(m^\star \log \frac{m^\star}{\delta_{\mathbb{P}}}\right)$ to ensure that, at initialization, at least one of the router-expert pair from the student model have a good enough alignment for each router-expert pair in the teacher. Also we require d to be polynomially large in terms of m and $\delta_{\mathbb{P}}^{-1}$ to control the interference between the convergence of each router-expert pair, as well as between the convergence of the router parameter and the expert parameter. Due to the polynomial scaling of d in terms of m and $\delta_{\mathbb{P}}^{-1}$, quantities $\mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$ and $\mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)$ are in general small, and $\mathcal{O}\left(\frac{\delta_{\mathbb{P}}\sqrt{d}}{m^2}\right)$ is large.

In the set-up of the theorem, the mapping $\mathcal{I}$ establishes the correct matching between the expert-router pair in the teacher model and the router-expert pair in the student model. Ideally, we expect $\bar{\mathbf{v}}_{\mathcal{I}(i)}$ and $\bar{\mathbf{w}}_{\mathcal{I}(i)}$ to converge to $\bar{\mathbf{v}}_i^\star$ and $\bar{\mathbf{w}}_i^\star$. Two key points of the theorem are outlined below:

Sequential weak recovery. The first part of the theorem states that such convergence happens in a sequential order. By its set-up, T_ℓ denotes the time where the first ℓ pairs of $\left(\bar{\mathbf{v}}_{\mathcal{I}(i)}, \bar{\mathbf{w}}_{\mathcal{I}(i)}\right)$ for $i \leq \ell$ just achieved a weak convergence to $\bar{\mathbf{v}}_i^\star$ and $\bar{\mathbf{w}}_i^\star$ by achieving an inner product with of at least 0.9. In the mean time, before t reaches $T_{\ell+1}$, all the remaining pairs $\left(\bar{\mathbf{v}}_{\mathcal{I}(i)}, \bar{\mathbf{w}}_{\mathcal{I}(i)}\right)$ for $i > \ell$ still stays within a small alignment value with their reference.

Order of Recovery. We remark that the order in which the student's expert-router pairs converge to the teacher's parameters is determined by the alignment values $\mathbf{w}_i(0)^\top \bar{\mathbf{w}}_j^\star$ at initialization, as can be seen from the proof sketch in Section 4.1. Intuitively, a student expert i that achieves a larger initial alignment with a teacher expert j will converge to $\bar{\mathbf{w}}_j^\star$ earlier than a student expert with smaller initial alignment, provided that $\bar{\mathbf{w}}_i$ has not converged to another $\bar{\mathbf{w}}_{j'}^\star$ and no other $\bar{\mathbf{w}}_{i'}$ has converged to $\bar{\mathbf{w}}_j^\star$ yet. While one might expect the norms $\|\mathbf{w}_i(0)\|_2$ to play a role in governing this ordering, their influence is in fact negligible: by concentration, the norms of the randomly initialized weight vectors are tightly concentrated around their expectation, and thus do not meaningfully differentiate the convergence order across experts. In contrast, as shown in Lemma 2, the inner products $\mathbf{w}_i(0)^\top \bar{\mathbf{w}}_j^\star$ vary across pairs and serve as the primary signal that breaks the symmetry among student experts, ultimately determining which student expert specializes to which teacher expert and at what rate.

Near-perfect recovery. The second part of the the-

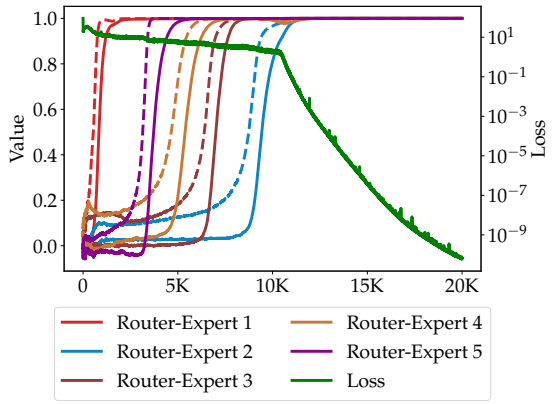


Figure 2: Dynamics of the routers' and experts' alignment value with the teacher's parameter under the same set-up as Figure 1. The green curve denotes the loss value. Except for the green curve, dashed line and solid line of the same color denotes a pair of router and expert alignment value.

orem shows that for any time t that exceeds some $T^\star \leq \mathcal{O}\left(\sqrt{d}\right)$ but stays under $T^\star + \mathcal{O}\left(\frac{\delta_{\mathbb{P}}\sqrt{d}}{m^2}\right)$, the learned features have converged to inner product values of at least $1-\mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}} d^{\frac{3}{2}}}\right)$. In the meantime, all the router and expert parameters in the student model that did not converge to any teacher's parameter must stay nearly orthogonal both to the teacher model's parameter and to each other. In Section 5, we will utilize this property to prove the theoretical guarantee of pruning these unused experts in the student model.

4.1 Guided by the Experts: A Proof Sketch of Theorem 1

In this section, we will discuss the difficulties and techniques arises in the proof of Theorem 1.

Hermite expansion of the loss and gradient. The starting point of our proof relies on the Hermite expansion of non-linear functions to study its property with Gaussian inputs. Let $He_k(x)$ denote the kth-order probabilist's Hermite polynomial. It is known that the set of Hermite polynomials $\{He_k(x)\}_{k=0}^{\infty}$ consists an basis of the square integrable functions under the Gaussian measure. Therefore, we can expand the sigmoid function as

$$\pi(x) = \sum_{k=0}^{\infty} \frac{c_k}{k!} He_k(x); c_k = \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\pi^{(k)}(x)\right]$$

Here c_k's are the Hermite coefficients of $\pi(x)$. Since $\sigma(x) = He_3(x)$, $f(\boldsymbol{\theta}, \mathbf{x})$ and $f^\star(\mathbf{x})$ has the form

$$f(\boldsymbol{\theta}, \mathbf{x}) = \sum_{k=0}^{\infty} \frac{c_k}{k!} He_k\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) He_3\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)$$

$$f^\star(\mathbf{x}) = \sum_{k=0}^{\infty} \frac{c_k}{k!} He_k\left(\bar{\mathbf{v}}_i^{\star\top} \mathbf{x}\right) He_3\left(\bar{\mathbf{w}}_i^{\star\top} \mathbf{x}\right)$$

Carrying this idea to the setting of minimizing the MSE in 3, we notice that $\mathcal{L}(\boldsymbol{\theta})$ can be written as

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}_{\mathbf{x}}\left[f(\boldsymbol{\theta}, \mathbf{x})^2\right] - \mathbb{E}_{\mathbf{x}}\left[f(\boldsymbol{\theta}, \mathbf{x}) f^\star(\mathbf{x})\right] + \frac{1}{2}\mathbb{E}_{\mathbf{x}}\left[f^\star(\mathbf{x})^2\right]$$

where one could notice that the last term does not depend on $\boldsymbol{\theta}$. However, the first two terms involves second-order terms on $f(\boldsymbol{\theta}, \mathbf{x})$ and $f^\star(\mathbf{x})$. As an illustration, we expand the first term as

$$\mathbb{E}_{\mathbf{x}}\left[f(\boldsymbol{\theta}, \mathbf{x})^2\right] = \sum_{i,j=1}^{m}\sum_{k,\ell=0}^{\infty} \frac{c_k c_\ell}{k!\ell!} \mathcal{C}_{k,\ell,3,3}^{i,j}$$

$$\mathcal{C}_{k,\ell,3,3}^{i,j} = \mathbb{E}_{\mathbf{x}}\left[He_k\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) He_\ell\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right) He_3\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) He_3\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\right]$$

Although a large body of prior work has exploited the nice property of Hermite polynomial that $\mathbb{E}_{\mathbf{x}}\left[He_k\left(\mathbf{u}_1^\top \mathbf{x}\right) He_\ell\left(\mathbf{u}_2^\top \mathbf{x}\right)\right] = k!\left(\mathbf{u}_1^\top \mathbf{u}_2\right)^k \mathbb{I}\{k=\ell\}$ for $\mathbf{u}_1, \mathbf{u}_2$ with unit norm, in our setting we have to deal with the expectation of the product of four Hermite polynomial. Our main tool of handling this difficulty is the lemma below.[3]

Lemma 1. *Let $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. For some multi-index $\mathbf{k} \in \mathbb{N}^n$, we define the multi-variate Hermite polynomial as*

$$He_{\mathbf{k}}(\mathbf{x}) = \prod_{i=1}^{n} He_{\mathbf{k}[i]}(\mathbf{x}[i])$$

Then we have that for $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$,

$$\mathbb{E}_{\mathbf{x}}\left[He_{\mathbf{k}}(\mathbf{x})\right] = \left(\prod_{i=1}^{n} \mathbf{k}[i]!\right) \sum_{\mathbf{M}\in\mathcal{S}} \prod_{i,j=1}^{n} \frac{\boldsymbol{\Sigma}[i,j]^{\mathbf{M}[i,j]}}{\mathbf{M}[i,j]!}$$

where the set $\mathcal{S}$ is the set of symmetric matrices $\mathbf{M} \in \mathbb{N}^{n\times n}$ satisfying

$$\mathbf{M}[i,i] = 0; \sum_{j=1}^{n} \mathbf{M}[i,j] = \mathbf{k}[i]; \forall i \in [n]$$

In Lemma 1, each $\mathbf{M} \in \mathcal{S}$ can be considered as the adjacency matrix of a graph with n nodes and integer weights such that the degree of the ith node is

[3] We are not the first to introduce this result. However, we could not find a formal published source that proves the result.

$\mathbf{k}[i]$. Applying Lemma 1 to our case thus only requires to enumerate the graphs of four nodes with degree $(k, \ell, 3, 3)$. With Lemma 1, we are able to derive the form of $\nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta})$ and $\nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta})$, whose exact form are omitted from the main text due to its intricacy.

Gradient flow dynamic of target alignment. Recall that our goal is to show that there is some $\bar{\mathbf{v}}_i, \bar{\mathbf{w}}_i$ that converges to $\bar{\mathbf{v}}_j^\star, \bar{\mathbf{w}}_j^\star$ for $j \in [m^\star]$. Thus, it is intuitive to start with tracking the dynamic of $\bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{v}}_j^\star$ and $\bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{w}}_j^\star$. With gradient flow, we have that

$$\begin{aligned}\frac{d}{dt}\bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{v}}_j^\star &= -\left\|\mathbf{v}_i(t)\right\|_2^{-1} \nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta})^\top \bar{\mathbf{v}}_j^\star \\ \frac{d}{dt}\bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{w}}_j^\star &= -\left\|\mathbf{w}_i(t)\right\|_2^{-1} \nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta})^\top \bar{\mathbf{w}}_j^\star\end{aligned} \tag{4}$$

due to the fact that $\|\mathbf{v}_i\|_2$ and $\|\mathbf{w}_i\|_2$ stays constant during the gradient flow process. Utilizing the fact that $\bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{v}}_j^\star, \bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{w}}_j^\star \ll 1$ when $\bar{\mathbf{v}}_i(t)$ and $\bar{\mathbf{w}}_i(t)$ are near their initialization, we can utilize Lemma 1 to approximate (4) as

$$\begin{aligned}\frac{d}{dt}\bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{v}}_j^\star &\propto \left(\bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{w}}_j^\star\right)^3 \\ \frac{d}{dt}\bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{w}}_j^\star &\propto \left(\bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{w}}_j^\star\right)^2\end{aligned} \tag{5}$$

The approximation above exhibits two interesting behaviors near initialization. First, the improvements in both the router alignment $\bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{v}}_j^\star$ and the expert alignment $\bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{w}}_j^\star$ depends on the current magnitude of the expert alignment. This implies that, once the expert alignment reaches a magnitude of $\Omega(1)$, it will take only constant time for the router and expert alignments to grow to a value near one (perfect alignment). This behavior can be observed in Figure 2 where the router alignment values (solid lines except for the gree on) follow closely as the expert alignment values (dashed lines) increases.

Second, the quadratic dependency in the expert alignment dynamic induces a sharp phase transition where the alignment value starts off slow for a period of time, and suddenly increases with a fast speed (see dashed lines in Figure 2), as studied in Ren et al. (2025). This sharp phase transition is particularly helpful to prevent multiple experts from the student model to converge to the same expert in the teacher model. As an example, consider dynamics $\bar{\mathbf{w}}_1(t)^\top \mathbf{w}_j^\star$ and $\bar{\mathbf{w}}_2(t)^\top \mathbf{w}_j^\star$ with a small different $\Delta \geq 0$ at initialization

$$0 \leq (1+\Delta)\bar{\mathbf{w}}_1(t)^\top \mathbf{w}_j^\star \leq \bar{\mathbf{w}}_2(t)^\top \mathbf{w}_j^\star \leq \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{d}}\right)$$

Solving the ODE in (5) gives that for $i \in \{1, 2\}$

$$\bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{w}}_j^\star \approx \left(\left(\bar{\mathbf{w}}_i(0)^\top \bar{\mathbf{w}}_j^\star\right)^{-1} - t\right)^{-1}$$

Thus, the time T required for $\bar{\mathbf{w}}_2(t)^\top \mathbf{w}_j^\star \geq \frac{1}{2}$ is $T = \left(\bar{\mathbf{w}}_2(0)^\top \bar{\mathbf{w}}_j^\star\right)^{-1} - 2 \leq \left(\bar{\mathbf{w}}_2(0)^\top \bar{\mathbf{w}}_j^\star\right)^{-1}$. However, at time T, one can compute that

$$\begin{aligned}\bar{\mathbf{w}}_1(t)^\top \bar{\mathbf{w}}_j^\star &\leq \left(\left(\bar{\mathbf{w}}_1(0)^\top \bar{\mathbf{w}}_j^\star\right)^{-1} - \left(\bar{\mathbf{w}}_2(0)^\top \bar{\mathbf{w}}_j^\star\right)^{-1}\right)^{-1} \\ &\leq \frac{\bar{\mathbf{w}}_2(0)^\top \bar{\mathbf{w}}_j^\star}{\Delta} \leq \tilde{\mathcal{O}}\left(\frac{\Delta^{-1}}{\sqrt{d}}\right)\end{aligned}$$

When $\sqrt{d} \gg \Delta^{-1}$, we can conclude that $\bar{\mathbf{w}}_2(T)^\top \mathbf{w}_j^\star \geq \frac{1}{2}$ while $\bar{\mathbf{w}}_1(T)^\top \mathbf{w}_j^\star \ll \frac{1}{2}$. This behavior implies that, the expert in the student model that aligns with expert j in the teacher best at initialization will converge to some $\Omega(1)$ quickly while the other experts' alignment remains small. Below, we formalize this dominance determined by the initialization.

Alignment gap at Initialization. We show that there is a set of experts in the student model that aligns with each expert in the teacher model good enough to create a gap compared with other experts in the student model. At a high level, our goal here is to construct the mapping $\mathcal{I}$ in Theorem 1 based on the initialization. Our approach is a greedy forward selection similar to Ren et al. (2025). In particular, we define Define $\mathcal{R}_\ell = \{i_k^\star\}_{k=1}^\ell$ and $\mathcal{C}_\ell = \{j_\ell^\star\}_{k=1}^\ell$ recursively as follows

$$i_{\ell+1}^\star, j_{\ell+1}^\star = \underset{i \in [m] \setminus \mathcal{R}_\ell, j \in [m^\star] \setminus \mathcal{C}_\ell}{\arg\max} \mathbf{w}_i(0)^\top \bar{\mathbf{w}}_j^\star \tag{6}$$

We expect that $\mathcal{I}(j_\ell^\star) = i_\ell^\star$. Namely, we expect $\bar{\mathbf{v}}_{i_\ell^\star}(t)^\top \bar{\mathbf{v}}_{j_\ell^\star}^\star$ and $\bar{\mathbf{w}}_{i_\ell^\star}(t)^\top \bar{\mathbf{w}}_{j_\ell^\star}^\star$ to converge to 1. The index ℓ denotes the order of the sequential convergence. That is, we expect that $\bar{\mathbf{v}}_{i_1^\star}(t)^\top \bar{\mathbf{v}}_{j_1^\star}^\star$ to grow large first, followed by $\bar{\mathbf{v}}_{i_2^\star}(t)^\top \bar{\mathbf{v}}_{j_2^\star}^\star$, etc. Our theorem below shows that at initialization, the pairs $\bar{\mathbf{w}}_{i_\ell^\star}(t)^\top \bar{\mathbf{w}}_{j_\ell^\star}^\star$ has a gap compared with other alignment values.

Lemma 2. *Let $\mathbf{w}_1, \ldots, \mathbf{w}_m \sim \mathcal{N}\left(0, d^{-1}\mathbf{I}_d\right)$ be I.I.D. Gaussian random vectors. Define*

$$\begin{aligned}i_\ell^\star, j_\ell^\star &= \underset{i \in [m] \setminus \mathcal{R}_{\ell-1}, j \in [m^\star] \setminus \mathcal{C}_{\ell-1}}{\arg\max} \mathbf{w}_i[j] \\ \mathcal{R}_\ell &= \{i_k^\star\}_{k=1}^\ell; \mathcal{C}_\ell = \{j_k^\star\}_{k=1}^\ell\end{aligned}$$

Let any $\delta_{\mathbb{P}} \in (0, 1/2)$ be given. Then there exists some absolute constant $\beta_2, \beta_4 > 0$ such that if $m \geq \beta_4 m^\star \log \frac{m^\star}{\delta_{\mathbb{P}}}$, then for $\delta_s = \frac{\beta_2 \delta_{\mathbb{P}}}{m^2}$, with probability at least $1 - 4\delta_{\mathbb{P}}$, it holds that

- *(Row-wise Gap) $\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \geq (1 + 2\delta_s)\mathbf{w}_{i_\ell^\star}[j]$ for all $\ell \in [m^\star]$ and $j \in [m^\star] \setminus \mathcal{C}_\ell$*
- *(Column-wise Gap) $\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \geq (1 + 2\delta_s)\mathbf{w}_i[j_\ell^\star]$ for all $\ell \in [m^\star]$ and $i \in [m] \setminus \mathcal{R}_\ell$*
- *(Threshold Gap) $\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \geq (1 + 2\delta_s)\mathbf{w}_{i_{\ell+1}^\star}[j_{\ell+1}^\star]^2$ for all $\ell \in [m^\star - 1]$*

- *(Magnitude Lower Bound)* $\mathbf{w}_{i_\ell^\star}[j_\ell^\star]^2 \geq \frac{\log m^\star}{d}$ *for all* $\ell \in [m]$

Since the standard Gaussian distribution is rotational invariant, we can the gaps and lower bound shown in Lemma 2 to the initial alignment scores $\bar{\mathbf{w}}_i(0)^\top \bar{\mathbf{w}}_j^\star$. Roughly speaking, the *row-wise gap* facilitates that $\bar{\mathbf{w}}_{i_\ell^\star}$ will not converge to $\bar{\mathbf{w}}_j$s with $j \neq j_\ell^\star$; the *column-wise gap* induces the fact that no $\bar{\mathbf{w}}_i$ will converge to $\bar{\mathbf{w}}_{j_\ell^\star}$ except for $\bar{\mathbf{w}}_{i_\ell^\star}$. Moreover, the *threshold gap* leads to the sequential recovery as stated in Theorem 1. Finally, the *magnitude lower bound* guarantees that at the target alignment values at initialization are not too small for the whole convergence process to be too long.

Induction-based Proof. With the goal of tracking the growth of $\bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{v}}_j^\star$ and $\bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{w}}_j^\star$ in mind, however, we also have to track the "mis-alignments" including $\bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{w}}_j^\star, \bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{v}}_j^\star$ and "self-alignments" $\bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{v}}_j(t), \bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{w}}_j(t), \bar{\mathbf{v}}_i(t)\bar{\mathbf{w}}_j(t)$ due to the complicated form of the gradient, as can be seem from (10) from the Appendix, so that their value does not interrupt with the target dynamics. To this end, our proof is an induction on $\ell \in [m^\star]$ that assumes

- $\bar{\mathbf{v}}_{i_{\ell'}^\star}(t)^\top \bar{\mathbf{v}}_j^\star$ and $\bar{\mathbf{w}}_{i_{\ell'}^\star}(t)^\top \bar{\mathbf{w}}_j^\star$ are close to one for $\ell' < \ell$, i.e., the top-$\ell-1$ router and experts are recovered well while the ℓth router-expert pair still remains not learned.
- The "mis-alignments" and "self-alignments" associated with the recovered router-expert pairs must be small throughout the process.

to show that $\bar{\mathbf{v}}_{i_\ell^\star}(t)^\top \bar{\mathbf{v}}_j^\star$ and $\bar{\mathbf{w}}_{i_\ell^\star}(t)^\top \bar{\mathbf{w}}_j^\star$ converge to a close-to-one value. A formal statement of the inductive hypothesis is provided in Appendix A.1, and the complete proof is provided in Appendix A.

5 PRUNING AND FINE-TUNING

Theorem 1 guarantees that in $\mathcal{O}\left(\sqrt{d}\right)$ time, the student MoE model trained with gradient flow extracts $m^\star$-pairs of near-perfect features from the teacher model. However, recall that the student have an over-parameterization of $m \geq \Omega\left(m^\star \log \frac{m^\star}{\delta_{\mathbb{P}}}\right)$. Despite being moderate, the $\log \frac{m^\star}{\delta_{\mathbb{P}}}$ factor still leads to a large number of excessive parameters. Continuing to train these unused experts together with their corresponding router parameter results in a wast of the computation resource, regardless of whether they can converge to zero output or not. This theoretical insight corresponds with existing empirical works (Lu et al., 2024; Chowdhury et al., 2024; Zhang et al., 2025d) which discovers the existence of redundant experts in pre-trained LLMs.

5.1 Pruning the Redundant Experts

In this section, we adopt a greedy pruning algorithm based on the test loss similar to Lu et al. (2024) to remove the redundant experts, and show that, if we apply the algorithm at $T^\star \leq t \leq T^\star + \mathcal{O}\left(\frac{\sqrt{d}}{\delta_{\mathbb{P}} m^2}\right)$, then we can provably remove all the unused experts and keep all the correctly learned router-expert pairs as stated in Theorem 1. To state the algorithm, we first define the sub-model MoE induced by $\mathcal{S} \subseteq [m]$ as

$$f_{\mathcal{S}}(\boldsymbol{\theta}, \mathbf{x}) = \sum_{i \in [m] \setminus \mathcal{S}} \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)$$

We consider the following pruning procedure that iteratively constructs the pruned set $\mathcal{S}$. In the τth step, we identify an index $r_\tau \in [m] \setminus \mathcal{S}_{\tau-1}$

$$\begin{aligned} r_\tau = \mathop{\arg\min}_{r \in [m] \setminus \mathcal{S}_{\tau-1}} \mathbb{E}_{\mathbf{x}}\left[\left(f_{S_{\tau-1} \cup \{r\}}(\boldsymbol{\theta}, \mathbf{x}) - f^\star(\mathbf{x})\right)^2\right] \\ \mathcal{S}_\tau = \mathcal{S}_{\tau-1} \cup \{r_\tau\} \end{aligned} \tag{7}$$

The procedure will stop when pruning one more expert does not improve the population loss. In particular, we define the stopping step $\tau^\star$ be such that

$$\begin{aligned} \min_{r \in [m]} \mathbb{E}_{\mathbf{x}}\left[\left(f_{S_{\tau^\star} \cup \{r\}}(\boldsymbol{\theta}, \mathbf{x}) - f^\star(\mathbf{x})\right)^2\right] \\ \geq \mathbb{E}_{\mathbf{x}}\left[\left(f_{\mathcal{S}_{\tau^\star}}(\boldsymbol{\theta}, \mathbf{x}) - f^\star(\mathbf{x})\right)^2\right] \end{aligned} \tag{8}$$

For the simplicity of the analysis, we assume that $\mathcal{I}(i) = i$, as reordering the router-expert pairs does not change $f(\boldsymbol{\theta}, \mathbf{x})$. To facilitate the analysis of the pruning procedure, we make the following assumption

Assumption 4. *Let* $\{\mathbf{v}_i\}_{i=1}^m$ *and* $\{\mathbf{w}_i\}_{i=1}^m$ *be the router and expert weights of the MoE model in (1). Let* $\{\bar{\mathbf{v}}_i^\star\}_{i=1}^{m^\star}$ *and* $\{\bar{\mathbf{w}}_i^\star\}_{i=1}^{m^\star}$ *be the router and expert weights of the teacher model in (2). There exists* $\varepsilon \leq o\left(\frac{1}{\sqrt{m}}\right)$ *such that for all* $i \in [m^\star]$ *it holds that*

$$\min\left\{\bar{\mathbf{v}}_i^\top \bar{\mathbf{v}}_i^\star, \bar{\mathbf{w}}_i^\top \bar{\mathbf{w}}_i^\star\right\} \geq 1 - \varepsilon,$$

for all $i_1 \in [m] \setminus [m^\star], i_2 \in [m], i_1 \neq i_2$ *and* $j \in [m^\star]$ *it holds that*

$$\begin{aligned} \left|\bar{\mathbf{v}}_{i_1}^\top \bar{\mathbf{v}}_{i_2}\right|, \left|\bar{\mathbf{w}}_{i_1}^\top \bar{\mathbf{w}}_{i_2}\right|, \left|\bar{\mathbf{v}}_{i_1}^\top \bar{\mathbf{w}}_{i_2}\right|, \left|\bar{\mathbf{v}}_{i_2}^\top \bar{\mathbf{w}}_{i_1}\right|, \left|\bar{\mathbf{v}}_{i_1}^\top \bar{\mathbf{w}}_{i_1}\right| \leq \varepsilon \\ \left|\bar{\mathbf{v}}_i^\top \bar{\mathbf{v}}_j^\star\right|, \left|\bar{\mathbf{w}}_i^\top \bar{\mathbf{w}}_j^\star\right|, \left|\bar{\mathbf{v}}_i^\top \bar{\mathbf{w}}_j^\star\right|, \left|\bar{\mathbf{w}}_i^\top \bar{\mathbf{v}}_j^\star\right| \leq \varepsilon \end{aligned}$$

Theorem 2. *Let* $f_{\mathcal{S}}(\boldsymbol{\theta}, \mathbf{x}), \mathcal{S}_\tau$, *and* $\tau^\star$ *be defined above. If Assumption 4 holds, then we have that*

$$f_{\mathcal{S}_{\tau^\star}}(\boldsymbol{\theta}, \mathbf{x}) = \sum_{i=1}^{m^\star} \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)$$

Theorem 2 states that, after $\tau^\star$ steps of pruning, the resulting model contains the exact $m^\star$ router-expert

pairs with learned features from the teacher model. As a condition of Theorem 2, Assumption 4 is satisfied by Theorem 1 with $\varepsilon = \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right) \leq o\left(\frac{1}{\sqrt{m}}\right)$ since $d \gg m$. This implies that, if we perform the pruning at $T^\star \leq t \leq T^\star + \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$ in the gradient flow process, we are guaranteed to remove all unused router-expert pairs and keep all necessary ones.

Notice that in the pruning procedure we evaluate the model on the population loss. To apply the algorithm in practice, one can effectively approximate the population loss with the sample loss. We use the population loss for the succinctness of the theoretical analysis.

Sketch of Proof. From a high level perspective, our proof relies on the observation that for two nonlinear function $h_1, h_2 : \mathbb{R} \to \mathbb{R}$ and vectors $\mathbf{u}_1, \mathbf{u}_2$ with $\mathbf{u}_1^\top \mathbf{u}_2 \approx 0$, we have that

$$\begin{aligned}&\mathbb{E}_{\mathbf{x}}\left[h_1\left(\mathbf{u}_1^\top \mathbf{x}\right) h_2\left(\mathbf{u}_2^\top \mathbf{x}\right)\right] \\ &\quad \approx \mathbb{E}_{\mathbf{x}}\left[h_1\left(\mathbf{u}_1^\top \mathbf{x}\right)\right]\mathbb{E}_{\mathbf{x}}\left[h_2\left(\mathbf{u}_2^\top \mathbf{x}\right)\right]\end{aligned}$$

Let $q\left(\bar{\mathbf{v}}_i, \bar{\mathbf{w}}_i, \mathbf{x}\right) = \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)$. This allows us to approximate the loss as

$$\begin{aligned}\mathcal{L}\left(\theta\right) \approx \mathbb{E}_{\mathbf{x}}\left[\left(\sum_{i=1}^{m^\star}\left(q\left(\bar{\mathbf{v}}_i, \bar{\mathbf{w}}_i, \mathbf{x}\right) - q\left(\bar{\mathbf{v}}_i^\star, \bar{\mathbf{w}}_i^\star, \mathbf{x}\right)\right)\right)^2\right] \\ + \sum_{i=m^\star+1}^{m} \mathbb{E}_{\mathbf{x}}\left[\pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)^2\right]\mathbb{E}_{\mathbf{x}}\left[\sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)^2\right]\end{aligned}$$

The first term is naturally small due to the fact that $\bar{\mathbf{v}}_i^\top \bar{\mathbf{v}}_i^\star$ and $\bar{\mathbf{w}}_i^\top \bar{\mathbf{w}}_i^\star$ are close to one for $i \in [m^\star]$. The second term involves a summation of positive terms, which depends on the redundant router-expert pairs. Thus, removing each one of these will decrease $\mathcal{L}\left(\boldsymbol{\theta}\right)$. The proof of Theorem 2 is provided in Appendix B.

5.2 Fine-Tuning the Pruned Model

Recall from Theorem 1 that, although the $m^\star$ router-expert pairs in the student model extracted near-perfect features from the teacher model, there is still an $\mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)$ error for each $\bar{\mathbf{v}}_{\mathcal{I}(i)}^\top \bar{\mathbf{v}}_i^\star$ and $\bar{\mathbf{w}}_{\mathcal{I}(i)}^\top \bar{\mathbf{w}}_i^\star$. This results in a non-zero loss even after the pruning in Section 5.1. In this section, we study the convergence guarantee of fine-tuning the pruned model with gradient flow on the population MSE. In particular, we assume that $f\left(\boldsymbol{\theta}, \mathbf{x}\right)$ is the pruned model from Section 5.1 given by

$$f\left(\boldsymbol{\theta}, \mathbf{x}\right) = \sum_{i=1}^{m^\star} \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)$$

and the fine-tuning starts at $\boldsymbol{\theta}(T_0)$ learned from Theorem 1 at time $T^\star \leq T_0 \leq \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)$. We slightly abuse the notation by denoting

$$\boldsymbol{\theta} = [\mathbf{v}_1^\top, \ldots, \mathbf{v}_{m^\star}^\top, \mathbf{w}_1^\top, \ldots, \mathbf{w}_{m^\star}^\top]^\top \in \mathbb{R}^{2m^\star d}$$

Moreover, we denote the normalized version of $\boldsymbol{\theta}$ as

$$\bar{\boldsymbol{\theta}} = [\bar{\mathbf{v}}_1^\top, \ldots, \bar{\mathbf{v}}_{m^\star}^\top, \bar{\mathbf{w}}_1^\top, \ldots, \bar{\mathbf{w}}_{m^\star}^\top]^\top \in \mathbb{R}^{2m^\star d}$$

The following theorem shows the convergence of gradient flow in the fine-tuning phase.

Theorem 3. *Let $\boldsymbol{\theta}(T_0)$ that satisfy $\|\bar{\mathbf{v}}_i - \bar{\mathbf{v}}_i^\star\|_2 \leq \varepsilon$ and $\|\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_i^\star\|_2 \leq \varepsilon$ for some $\varepsilon \leq o\left(\frac{1}{m^{\star 2}}\right)$. Let $C_{S,0} = 2\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\pi\left(x\right)^2\right]$ and $C_{S,1} = 6\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\pi'\left(x\right)^2\right]$. If $C_{S,0} \geq 1.1C_{S,1}$, then there exists some constant $\kappa > 0$ that only depends on the property of the sigmoid function $\pi\left(\cdot\right)$ such that*

$$\left\|\bar{\boldsymbol{\theta}}\left(t + T_0\right) - \boldsymbol{\theta}^\star\right\|_2^2 \leq exp\left(-\frac{\kappa t}{2}\right)\left\|\bar{\boldsymbol{\theta}}(T_0) - \boldsymbol{\theta}^\star\right\|_2^2$$

Under the condition that $\bar{\mathbf{v}}_i$'s and $\bar{\mathbf{w}}_i$'s are ε-close to $\bar{\mathbf{v}}_i^\star$'s and $\bar{\mathbf{w}}_i^\star$'s, Theorem 3 shows a linear convergence rate in terms of the difference between the pruned model's normalized parameters $\bar{\boldsymbol{\theta}}$ and the optimal parameters $\boldsymbol{\theta}^\star$. In this fine-tuning stage, the convergence rate κ is independent of the dimension d or the number of experts $m^\star$. Instead, it only depends on the property of the router's non-linear function $\pi(\cdot)$. Since $\boldsymbol{\theta}\left(T_0\right)$ is given by the learned result in Theorem 1, the condition that $\|\bar{\mathbf{v}}_i - \bar{\mathbf{v}}_i^\star\|_2 \leq \varepsilon$ and $\|\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_i^\star\|_2 \leq \varepsilon$ for some $\varepsilon \leq \mathcal{O}\left(\frac{m}{\sqrt{\delta_{\mathbb{P}}} d^{\frac{1}{4}}}\right) \leq o\left(\frac{1}{m^{\star 2}}\right)$ are automatically satisfied under $d \gg m$, since

$$\|\bar{\mathbf{v}}_i - \bar{\mathbf{v}}_i^\star\|_2^2 = 2 - 2\bar{\mathbf{v}}_i^\top \bar{\mathbf{v}}_i^\star \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$$

The assumption that $C_{S,0} \geq 1.1C_{S,1}$ only depends on the property of $\pi(\cdot)$ and is checked in Appendix E.

Sketch of Proof. Our proof relies on the idea that, near the global minimum, the Hessian matrix is positive definite. In particular, we show that for any vector $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^{2m^\star d}$ such that $\cos\langle \mathbf{u}_1, \mathbf{u}_2\rangle \approx 1$, it holds that $\mathbf{u}_1^\top \nabla^2 \mathcal{L}\left(\boldsymbol{\theta}\right)\mathbf{u}_2 \geq \kappa \|\mathbf{u}_1\|_2 \|\mathbf{u}_2\|$ for some constant $\kappa > 0$ and $\boldsymbol{\theta}$ satisfying $\|\bar{\mathbf{v}}_i - \bar{\mathbf{v}}_i^\star\|_2 \leq \varepsilon$ and $\|\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_i^\star\|_2 \leq \varepsilon$ with $\varepsilon \leq o\left(\frac{1}{m^{\star 2}}\right)$. Theorem 4 in Appendix C provides a formal statement of the result. Based on the positive-definiteness of the Hessian matrix, we leverage the classic convex optimization technique to show that the trajectory never leaves the neighborhood near the global minima, and that the distance to the global minimum converges linearly. The full proof is deferred to Appendix C.

6 CONCLUSION

Under the teacher-student set-up, we study the learning dynamics of the sigmoid-routed MoE with non-

linear experts defined in (1) when trained with gradient flow on the population MSE over high dimensional Gaussian inputs. In particular, our main result is a characterization of the feature learning stage, where proper features of the router-expert pairs are discovered in sequential order, with the expert's recovery leading the router's recovery. At the end of the feature learning stage, we show that a pruning procedure can be conducted to provably remove all the redundant experts and keep all necessary ones. Lastly, we show a linear convergence rate to the global minima for the the post-pruning fine-tuning with gradient flow. To the best of our knowledge, our work is the first to provide theoretical understanding on the joint training guarantee of MoEs with more than one activated experts and a general data assumption. In general, our paper is a further step into understanding the complicated dynamics of MoE training, and leads to the following open problems:

Online SGD and Sample Complexity. Due to the already sophisticated proof, our study is restricted to the setting of gradient flow on the population loss. However, as the main idea of the proof consists of an ODE based dynamic analysis, one could discretize the dynamic and apply martingale-based analysis to extend the theory to online SGD, as in Ren et al. (2025). This extension may lead to a sample complexity bound of learning $m^\star$ experts on d-dimensional data.

Experts with different importance. In our work we considered the teacher's router and expert parameter $\bar{\mathbf{v}}_i^\star$ and $\bar{\mathbf{w}}_i^\star$ to be an orthonormal list. Due to the rotational invariance, this set-up puts equal importance to each router-expert pairs. Future work can investigate the scenario where the ith expert is scaled with a factor of α_i, and study the explicit ordering of the recovered experts in the student model.

Relax the dependency of d on m. Our current theory relies on the fact that $d \gg m$. While in the practical application of MoE we rarely set the number of experts to be larger than the input dimension, in most cases the scale of the two remains relatively the same. A meaningful future direction is to bridge the gap by studying the setting where d is only moderately larger than m.

7 Acknowledgements

This work was supported in part by NSF CAREER Award no. 2145629 and the Ken Kennedy Institute at Rice University. AK also acknowledges support from a Microsoft Research Award and an Amazon Research Award.

References

Gérard Ben Arous, Murat A. Erdogdu, N. Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: Sgd dynamics and scaling laws, 2025. URL https://arxiv.org/abs/2508.03688.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: A spiked random matrix perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HlIAoCHDWW.

Umang Bhatt, Sanyam Kapoor, Mihir Upadhyay, Ilia Sucholutsky, Francesco Quinzan, Katherine M. Collins, Adrian Weller, Andrew Gordon Wilson, and Muhammad Bilal Zafar. When should we orchestrate multiple agents?, 2025. URL https://arxiv.org/abs/2503.13577.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks, 2022. URL https://arxiv.org/abs/2210.15651.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow, 2023. URL https://arxiv.org/abs/2310.19793.

Enric Boix-Adsera and Philippe Rigollet. The power of fine-grained experts: Granularity boosts expressivity in mixture of experts, 2025. URL https://arxiv.org/abs/2505.06839.

Etienne Boursier and Nicolas Flammarion. Early alignment in two-layer networks training is a two-edged sword. *Journal of Machine Learning Research*, 26(183):1–75, 2025.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs. In *Advances in Neural Information Processing Systems*, volume 35, pages 20105–20118, 2022.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding mixture of experts in deep learning, 2022. URL https://arxiv.org/abs/2208.02813.

Mohammed Nowaz Rabbani Chowdhury, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. Patch-level routing in mixture-of-experts is provably sample-efficient for convolutional neural networks, 2023. URL https://arxiv.org/abs/2306.04073.

Mohammed Nowaz Rabbani Chowdhury, Meng Wang, Kaoutar El Maghraoui, Naigang Wang, Pin-Yu Chen, and Christopher Carothers. A provably effective method for pruning experts in fine-tuned sparse

mixture-of-experts, 2024. URL https://arxiv.org/abs/2405.16646.

Alex Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent, 2022. URL https://arxiv.org/abs/2206.15144.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks, 2019. URL https://arxiv.org/abs/1811.03804.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Quentin Fruytier, Aryan Mokhtari, and Sujay Sanghavi. Learning mixtures of experts with em: A mirror descent perspective, 2025. URL https://arxiv.org/abs/2411.06056.

Rong Ge, Jason D. Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design, 2017. URL https://arxiv.org/abs/1711.00501.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems, 2025. URL https://arxiv.org/abs/2408.08435.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2020. URL https://arxiv.org/abs/1806.07572.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Ryotaro Kawata, Kohsei Matsutani, Yuri Kinoshita, Naoki Nishikawa, and Taiji Suzuki. Mixture of experts provably detect and learn the latent cluster structure in gradient-based learning, 2025. URL https://arxiv.org/abs/2506.01656.

Yilun Kong, Guozheng Ma, Qi Zhao, Haoyu Wang, Li Shen, Xueqian Wang, and Dacheng Tao. Mastering massive multi-task reinforcement learning via mixture-of-expert decision transformer, 2025. URL https://arxiv.org/abs/2505.24378.

Anastasis Kratsios, Takashi Furuya, Jose Antonio Lara Benitez, Matti Lassas, and Maarten de Hoop. Mixture of experts soften the curse of dimensionality in operator learning, 2024. URL https://arxiv.org/abs/2404.09101.

Jason D. Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit, 2024. URL https://arxiv.org/abs/2406.01581.

Binghui Li et al. Feature averaging: An implicit bias of gradient descent leading to non-robustness in neural networks. *arXiv preprint arXiv:2410.10322*, 2024.

Hongbo Li, Sen Lin, Lingjie Duan, Yingbin Liang, and Ness B. Shroff. Theory on mixture-of-experts in continual learning, 2025. URL https://arxiv.org/abs/2406.16437.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models, 2024. URL https://arxiv.org/abs/2402.14800.

Tao Luo et al. Phase diagram for two-layer ReLU neural networks at infinite-width limit. *Journal of Machine Learning Research*, 22(71):1–47, 2021.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient Descent Quantizes ReLU network features. *arXiv preprint arXiv:1803.08367*, 2018.

Hancheng Min, Enrique Mallada, and Rene Vidal. Early neuron alignment in two-layer ReLU networks with small initialization. *arXiv preprint arXiv:2307.12851*, 2023.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A. Erdogdu. Neural networks efficiently learn low-dimensional representations with sgd, 2023a. URL https://arxiv.org/abs/2209.14863.

Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A. Erdogdu. Gradient-based feature learning under structured data, 2023b. URL https://arxiv.org/abs/2309.03843.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with LIMoE: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

Huy Nguyen, Pedram Akbarian, TrungTin Nguyen, and Nhat Ho. A general theory for softmax gating multinomial logistic mixture of experts, 2024a. URL https://arxiv.org/abs/2310.14188.

Huy Nguyen, Nhat Ho, and Alessandro Rinaldo. Sigmoid gating is more sample efficient than softmax gating in mixture of experts, 2024b. URL https://arxiv.org/abs/2405.13997.

Huy Nguyen, Nhat Ho, and Alessandro Rinaldo. On least square estimation in softmax gating mixture of experts, 2024c. URL https://arxiv.org/abs/2402.02952.

Huy Nguyen, Nhat Ho, and Alessandro Rinaldo. Convergence rates for softmax gating mixture of experts, 2025. URL https://arxiv.org/abs/2503.03213.

Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D. Lee. Emergence and scaling laws in sgd learning of shallow neural networks, 2025. URL https://arxiv.org/abs/2504.19983.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Zhenmei Shi, Junyi Wei, and Yingyu Liang. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features, 2022. URL https://arxiv.org/abs/2206.01717.

Zhenmei Shi, Junyi Wei, and Yingyu Liang. Provable guarantees for neural networks via gradient feature learning, 2023. URL https://arxiv.org/abs/2310.12408.

Mingze Wang and Weinan E. On the expressive power of mixture-of-experts for structured complex tasks, 2025. URL https://arxiv.org/abs/2505.24205.

Xiang Wang, Chenwei Wu, Jason D. Lee, Tengyu Ma, and Rong Ge. Beyond lazy training for over-parameterized tensor decomposition, 2020. URL https://arxiv.org/abs/2010.11356.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Danyang Zhang, Junhao Song, Ziqian Bi, Yingfang Yuan, Tianyang Wang, Joe Yeong, and Junfeng Hao. Mixture of experts in large language models, 2025a. URL https://arxiv.org/abs/2507.11181.

Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. G-designer: Architecting multi-agent communication topologies via graph neural networks, 2025b. URL https://arxiv.org/abs/2410.11782.

Ruichen Zhang, Mufan Qiu, Zhen Tan, Mohan Zhang, Vincent Lu, Jie Peng, Kaidi Xu, Leandro Z. Agudelo, Peter Qian, and Tianlong Chen. Symbiotic cooperation for web agents: Harnessing complementary strengths of large and small llms, 2025c. URL https://arxiv.org/abs/2502.07942.

Zeliang Zhang, Xiaodong Liu, Hao Cheng, Chenliang Xu, and Jianfeng Gao. Diversifying the expert knowledge for task-agnostic pruning in sparse mixture-of-experts, 2025d. URL https://arxiv.org/abs/2407.09590.

Berfin Şimşek, Amire Bendjeddou, and Daniel Hsu. Learning gaussian multi-index models with gradient flow: Time complexity and directional convergence, 2025. URL https://arxiv.org/abs/2411.08798.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable] **Yes**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] **Yes**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] **Yes**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable] **Yes**
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] **Yes**
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] **Yes**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable] **Yes**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable] **Yes**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable] **No**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] **No**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable] **Not Applicable**
   (b) The license information of the assets, if applicable. [Yes/No/Not Applicable] **Not Applicable**
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable] **Not Applicable**
   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable] **Not Applicable**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] **Not Applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable] **Not Applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable] **Not Applicable**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable] **Not Applicable**

Instructions for Paper Submissions to AISTATS 2026: Supplementary Materials

A Proof of Theorem 1

A.1 Proof Outline

Initialization Property. At initialization, the following property needs to be satisfied.

Condition 1 (Initialization)**.** *At initialization* $\{\mathbf{w}_i(0)\}_{i=1}^m$ *and* $\{\mathbf{v}_i(0)\}_{i=1}^m$ *satisfies*

- $\mathbf{w}_{i_\ell^\star}(0)^\top \bar{\mathbf{w}}_{j_\ell^\star}^\star \geq (1+2\delta_s)\,\mathbf{w}_{i_\ell^\star}(0)^\top \bar{\mathbf{w}}_j^\star$ *for all* $\ell \in [m^\star]$ *and* $j \in [m^\star] \setminus \mathcal{C}_{\ell+1}$.
- $\mathbf{w}_{i_\ell^\star}(0)^\top \bar{\mathbf{w}}_{j_\ell^\star}^\star \geq (1+2\delta_s)\,\mathbf{w}_i(0)^\top \bar{\mathbf{w}}_{j_\ell^\star}^\star$ *for all* $\ell \in [m^\star]$ *and* $i \in [m] \setminus \mathcal{R}_{\ell+1}$.
- $\mathbf{w}_{i_\ell^\star}(0)^\top \bar{\mathbf{w}}_{j_\ell^\star}^\star \geq (1+2\delta_s)\,\mathbf{w}_{i_{\ell+1}^\star}(0)^\top \bar{\mathbf{w}}_{j_{\ell+1}^\star}^\star$ *for all* $\ell \in [m^\star - 1]$.
- $\left(\mathbf{w}_{i_\ell^\star}(0)^\top \bar{\mathbf{w}}_{j_\ell^\star}^\star\right)^2 \geq \frac{\log m^\star}{d}$ *for all* $\ell \in [m^\star]$.
- $\|\mathbf{w}_i\|_2, \|\mathbf{v}_i\|_2 \in [1-\beta_2\delta_s, 1+\beta_2\delta_s]$ *for all* $i \in [m]$.
- $\max\left\{\left(\mathbf{v}_i(0)^\top \mathbf{v}_j^\star\right)^2, \left(\mathbf{v}_i(0)^\top \mathbf{w}_j^\star\right)^2, \left(\mathbf{w}_i(0)^\top \mathbf{v}_j^\star\right)^2\right\} \leq \frac{\beta_3}{d}\log\frac{m}{\delta_\mathbb{P}}$ *for all* $i \in [m], j \in [m^\star]$.
- $\max\left\{\left(\mathbf{v}_i(0)^\top \mathbf{v}_j(0)\right)^2, \left(\mathbf{w}_i(0)^\top \mathbf{w}_j(0)\right)^2, \left(\mathbf{v}_i(0)^\top \mathbf{w}_j(0)\right)^2\right\} \leq \frac{\beta_3}{d}\log\frac{m}{\delta_\mathbb{P}}$ *for all* $i, j \in [m]$ *and* $i \neq j$. *Moreover,* $\mathbf{v}_i(0)^\top \mathbf{w}_i(0) = 0$ *for all* $i \in [m]$.

where $\delta_s = \frac{\beta_1\delta_\mathbb{P}}{mm^\star}$ *for some absolute constant* $\beta_1, \beta_3 > 0$ *and* $\beta_2 \leq o(1)$ *and any* $\delta_\mathbb{P} \in (0, 1/7)$.

By Lemma 2 and Lemma 3, the above condition holds with probability at least $1-7\delta_\mathbb{P}$ as long as $d \geq \frac{\beta_5 m^4}{\delta_\mathbb{P}^2}\log\frac{m}{\delta_\mathbb{P}}$ and $m \geq \beta_4 m^\star \log\frac{m^\star}{\delta_\mathbb{P}}$.

Inductive Hypothesis. Now we are going to show that $\bar{\mathbf{w}}_{i_\ell^\star}(t)^\top \bar{\mathbf{w}}_{j_\ell^\star}^\star$ converges to at least $1-\frac{c}{\sqrt{d}}$ for all $\ell \in [m^\star]$ by induction. To start, we denote the values of interest as follows

$$\begin{gathered}\gamma_{i,j}^{(1)}(t) = \bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{v}}_j^\star; \quad \gamma_{i,j}^{(2)}(t) = \bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{w}}_j^\star; \quad \zeta_{i,j}^{(1)}(t) = \bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{w}}_j^\star; \quad \zeta_{i,j}^{(2)}(t) = \bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{v}}_j^\star \\ I_{i,j}^{(1)}(t) = \bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{v}}_j(t); \quad I_{i,j}^{(2)}(t) = \bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{w}}_j(t); \quad I_{i,j}^{(3)}(t) = \bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{w}}_j(t)\end{gathered} \tag{9}$$

To state the inductive hypothesis, we need the following error bounds.

Definition 1 (Error Bounds)**.** *For each* $\ell \in [m^\star]$*, we define the following:*

$$\varepsilon_{1,\ell}(t) := \max_{i\in[m]\setminus\mathcal{R}_\ell, j\in[m^\star]} \left|\gamma_{i,j}^{(1)}(t)\right|; \varepsilon_{2,\ell}(t) := \max_{j\in[m^\star]\setminus\{j_\ell^\star\}} \left|\gamma_{i_\ell^\star,j}^{(1)}(t)\right|;$$

$$\varepsilon_{3,\ell}(t) := \max_{i\in[m]\setminus\mathcal{R}_\ell, j\in[m^\star]} \left|\gamma_{i,j}^{(2)}(t)\right|; \varepsilon_{4,\ell}(t) := \max_{j\in[m^\star]\setminus\{j_\ell^\star\}} \left|\gamma_{i_\ell^\star,j}^{(2)}(t)\right|; \varepsilon_{5,\ell}(t) := \left|I_{i_\ell^\star,i_\ell^\star}^{(3)}(t)\right|$$

Moreover, we also define the forward error, the backward error, and the aggregated error as

$$\varepsilon_{\mathcal{F},\ell}(t) = \max\left\{\varepsilon_{\mathcal{F},\ell,1}(t), \varepsilon_{\mathcal{F},\ell,2}(t)\right\}; \varepsilon_{\mathcal{B},\ell}^{(1)}(t) = \max\left\{\varepsilon_{\mathcal{B},\ell,1}(t), \varepsilon_{\mathcal{B},\ell,2}(t), \varepsilon_{2,\ell}(t)\right\}$$

$$\varepsilon_{\mathcal{B},\ell}^{(2)}(t) = \max\left\{\varepsilon_{\mathcal{B},\ell}^{(1)}(t), \varepsilon_{1,\ell}(t)\right\}; \varepsilon_{\mathcal{B},\ell}^{(3)}(t) = \max_{i\in[m]\setminus\mathcal{R}_\ell} \left|I_{i,i}^{(3)}(t)\right|$$

$$\hat{\varepsilon}_{\mathcal{A},\ell}^{(1)}(t) = \max\left\{\varepsilon_{4,\ell}(t), \varepsilon_{\mathcal{B},\ell}^{(1)}(t), \varepsilon_{\mathcal{F},\ell}(t)\right\}; \hat{\varepsilon}_{\mathcal{A},\ell}^{(2)}(t) = \max\left\{\hat{\varepsilon}_{\mathcal{A},\ell}^{(1)}(t), \varepsilon_{1,\ell}(t), \varepsilon_{3,\ell}(t)\right\}$$

where $\varepsilon_{\mathcal{F},\ell,1}(t), \varepsilon_{\mathcal{F},\ell,2}(t), \varepsilon_{\mathcal{F},\ell,3}(t)$ *and* $\varepsilon_{\mathcal{B},\ell,1}(t), \varepsilon_{\mathcal{B},\ell,2}(t)$ *are defined as*

$$\varepsilon_{\mathcal{F},\ell,1}(t) := \max_{\ell'\le\ell, j\in[m^\star]\setminus\{j^\star_{\ell'}\}} \max\left\{\left|\gamma^{(1)}_{i^\star_{\ell'},j}(t)\right|, \left|\gamma^{(2)}_{i^\star_{\ell'},j}(t)\right|\right\};$$
$$\varepsilon_{\mathcal{F},\ell,2}(t) := \max_{i\in\mathcal{R}_{\ell-1}, j\in[m^\star]} \max\left\{\left|\zeta^{(1)}_{i,j}(t)\right|, \left|\zeta^{(2)}_{i,j}(t)\right|\right\}$$
$$\varepsilon_{\mathcal{B},\ell,1}(t) := \max_{i\in[m]\setminus\mathcal{R}_{\ell-1}, j\in[m^\star]} \max\left\{\left|\zeta^{(1)}_{i,j}(t)\right|, \left|\zeta^{(2)}_{i,j}(t)\right|\right\}$$
$$\varepsilon_{\mathcal{B},\ell,2}(t) := \max_{i,j\in[m], i\neq j} \max\left\{\left|I^{(1)}_{i,j}(t)\right|, \left|I^{(2)}_{i,j}(t)\right|, \left|I^{(3)}_{i,j}(t)\right|\right\}$$

Lastly, we are going to define the monotonic upper bound of $\hat{\varepsilon}^{(1)}_{\mathcal{A},\ell}(t)$ *and* $\hat{\varepsilon}^{(2)}_{\mathcal{A},\ell}(t)$

$$\varepsilon^{(1)}_{\mathcal{A},\ell}(t) = \sup_{t'\in[0,t]} \hat{\varepsilon}^{(1)}_{\mathcal{A},\ell}(t); \varepsilon^{(2)}_{\mathcal{A},\ell}(t) = \sup_{t'\in[0,t]} \hat{\varepsilon}^{(2)}_{\mathcal{A},\ell}(t)$$

Definition 2 (Recovery Time)**.** *Define the* ξ*-recovery time of* $\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)$*, denoted as* $T_\ell(\xi)$*, as*

$$T_\ell(\xi) = \min\left\{t\ge 0 : \gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \ge \xi\right\}$$

Based on $T_\ell(\xi)$*, we define the* ***constant-recovery time*** *and the* ***near-perfect-recovery time*** *as* $T_{r,\ell} = T_\ell(0.9)$ *and* $T_{p,\ell} = T_\ell\left(1 - \frac{\beta_9 m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)$*, respectively.*

Condition 2 (Inductive Hypothesis)**.** *Let* $\ell \in [m^\star]$*. Then we have that*

- *(Sequential recovery) For all* $t \ge T_{p,\ell-1}$ *such that* $\varepsilon^{(1)}_{\mathcal{A},\ell}(t) \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$ *we have* $\gamma^{(2)}_{i^{\prime\star}_\ell,j^{\prime\star}_\ell}(t) \ge 1 - \frac{\beta_9 m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}$*, and* $\gamma^{(1)}_{i^{\prime\star}_\ell,j^{\prime\star}_\ell}(t) \ge 1 - \frac{\beta_9 m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}$ *for all* $\ell' < \ell$*.*
- *(Error bound of* $\gamma^{(1)}_{i,j}, \gamma^{(2)}_{i,j}$*) For* $t \le T_{p,\ell-1}$*, we have*

$$\max_{i\in[m]\setminus\mathcal{R}_{\ell-1}, j\in[m^\star]} \max\left\{\left|\gamma^{(1)}_{i,j}(t)\right|, \left|\gamma^{(2)}_{i,j}(t)\right| \le \frac{\beta_6 m^2}{\sqrt{d}}\right\}$$

  .
- *(Error bound of remaining items)* $\varepsilon_{\mathcal{F},\ell}(t) \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$ *for all* t *such that* $\varepsilon^{(1)}_{\mathcal{A},\ell}(t) \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$*.*

Here $\beta_6, \beta_9 > 0$ *are some absolute constant.*

The proof proceeds by establishing the inductive hypothesis.

A.2 Initialization Property

Lemma 3. *Let* $\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_m$ *and* $\mathbf{w}_1, \ldots, \mathbf{w}_m$ *be I.I.D. random vectors from* $\mathcal{N}\left(\mathbf{0}, d^{-1}\mathbf{I}_d\right)$*. Define* $\mathbf{v}_i = \left(\mathbf{I} - \frac{1}{\|\mathbf{w}_i\|_2^2}\mathbf{w}_i\mathbf{w}_i^\top\right)\hat{\mathbf{v}}_i$*. Then there exists some absolute constant* $\beta_1, \beta_3, \beta_5 > 0, \beta_2 \le o(1)$*, and* $\delta_{\mathbb{P}} \in (0, 1/3)$ *such that if* $d \ge \frac{\beta_5 m^4}{\delta_{\mathbb{P}}^2}\log\frac{m}{\delta_{\mathbb{P}}}$ *and* $\delta_s = \frac{\beta_1\delta_{\mathbb{P}}}{m^2}$*, with probability at least* $1 - 3\delta_{\mathbb{P}}$ *we have that*

- $\|\mathbf{v}_i\|_2^2, \|\mathbf{w}_i\|_2^2 \in [1 - \beta_2\delta_s, 1 + \beta_2\delta_s]$ *for all* $i \in [m]$*;*
- $\max\left\{\mathbf{v}_i[j]^2, \mathbf{w}_i[j]^2\right\} \le \frac{\beta_3}{d}\log\frac{m}{\delta_{\mathbb{P}}}$ *for all* $i \in [m], j \in [m^\star]$*;*
- $\max\left\{\left(\mathbf{v}_i^\top\mathbf{v}_j\right)^2, \left(\mathbf{w}_i^\top\mathbf{w}_j\right)^2, \left(\mathbf{v}_i^\top\mathbf{w}_j\right)^2\right\} \le \frac{\beta_3}{d}\log\frac{m}{\delta_{\mathbb{P}}}$ *for all* $i, j \in [m]$ *with* $i \neq j$*.*

Proof. Our proof starts with showing the concentration for $\|\hat{\mathbf{v}}_i\|_2^2, \|\mathbf{w}_i\|_2^2$ and $\hat{\mathbf{v}}_i^\top \mathbf{w}_i$, and then moves to the proof of the desired statement.

Concentration of $\|\hat{\mathbf{v}}_i\|_2^2, \|\mathbf{w}_i\|_2^2$. Let $\mathbf{v} \sim \mathcal{N}\left(\mathbf{0}, d^{-1}\mathbf{I}_d\right)$. Then $\mathbf{v}[i] \sim \mathcal{N}\left(0, d^{-1}\right)$. Thus, $d\mathbf{v}[i]^2 \in \texttt{subE}\left(2,2\right)$. Since $\mathbf{v}[i]$'s are I.I.D., we have that

$$d\left\|\mathbf{v}\right\|_2^2 = d\sum_{i=1}^d \mathbf{v}[i]^2 \in \texttt{subE}\left(2d, 2\right)$$

Therefore, by the tail bound of sub-exponential random variables, we have that

$$\Pr\left(\left|\left\|\mathbf{v}\right\|_2^2 - 1\right| \geq \frac{t}{d}\right) = \Pr\left(\left|d\left\|\mathbf{v}\right\|_2^2 - d\mathbb{E}\left[\left\|\mathbf{v}\right\|_2^2\right]\right| \geq t\right) \leq 2\exp\left(-\frac{1}{4}\min\left\{\frac{t^2}{d}, t\right\}\right)$$

We are going to focus on the case where $t \leq d$. In particular, we set $t = \frac{1}{2}\beta_2\delta_s d$. Then we have that

$$\Pr\left(\left|\left\|\mathbf{v}\right\|_2^2 - 1\right| \geq \frac{1}{2}\beta_2\delta_s\right) \leq 2\exp\left(-\frac{1}{16}\beta_2^2\delta_s^2 d\right)$$

Take a union bound over all $i \in [m]$ for $\hat{\mathbf{v}}_i$ and $\mathbf{w}_i$ gives that, with probability at least $1 - 4m\exp\left(-\frac{1}{16}\beta_2^2\delta_s^2 d\right)$, it holds that

$$\|\hat{\mathbf{v}}_i\|_2^2, \|\mathbf{w}_i\|_2^2 \in \left[1 - \frac{1}{2}\beta_2\delta_s, 1 + \frac{1}{2}\beta_2\delta_s\right]$$

Setting $d \geq \frac{\beta_5 m^4}{\delta_{\mathbb{P}}^2}\log\frac{m}{\delta_{\mathbb{P}}} \geq \frac{16}{\beta_2^2\delta_s^2}\log\frac{4m}{\delta_{\mathbb{P}}}$ guarantees that the failing probability is within $\delta_{\mathbb{P}}$.

Concentration of $\hat{\mathbf{v}}_i^\top \mathbf{w}_j, \hat{\mathbf{v}}_i^\top \hat{\mathbf{v}}_j$, and $\mathbf{w}_i^\top \mathbf{w}_j$. Due to the independence between $\hat{\mathbf{v}}_i$'s and $\mathbf{w}_j$'s, we have that

$$\mathbb{E}\left[\hat{\mathbf{v}}_i^\top \mathbf{w}_j\right] = 0; \quad d\hat{\mathbf{v}}_i[\ell]\mathbf{w}_j[\ell] \in \texttt{subE}\left(1,1\right).$$

Thus, we have that $\hat{\mathbf{v}}_i^\top \mathbf{w}_j \in \texttt{subE}\left(d, 1\right)$. Applying the tail bound of sub-exponential random variable gives that

$$\Pr\left(\left|\hat{\mathbf{v}}_i^\top \mathbf{w}_j\right| \geq \frac{t}{d}\right) = \Pr\left(d\left|\hat{\mathbf{v}}_i^\top \mathbf{w}_j\right| \geq t\right) \leq 2\exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{d}, t\right\}\right)$$

The same concentration also holds for $\hat{\mathbf{v}}_i^\top \hat{\mathbf{v}}_j$ and $\mathbf{w}_i^\top \mathbf{w}_j$. Again we are going to focus on the case $t \leq d$. Take a union bound over all $i, j \in [m]$ and $\hat{\mathbf{v}}_i, \mathbf{w}_i$ gives that

$$\Pr\left(\max\left\{\left|\hat{\mathbf{v}}_i^\top \mathbf{w}_j\right|, \left|\hat{\mathbf{v}}_i^\top \hat{\mathbf{v}}_j\right|, \left|\mathbf{w}_i^\top \mathbf{w}_j\right|\right\} \geq \frac{t}{d}; \forall i, j \in [m]\right) \leq 8m^2\exp\left(-\frac{t^2}{2d}\right)$$

Setting the failing probability to $\delta_{\mathbb{P}}$ gives that with probability at least $1 - \delta_{\mathbb{P}}$, we have that

$$\hat{\mathbf{v}}_i^\top \mathbf{w}_j, \hat{\mathbf{v}}_i^\top \hat{\mathbf{v}}_j, \mathbf{w}_i^\top \mathbf{w}_j \in \left[-\frac{1}{\sqrt{d}}\left(\log\frac{8m^2}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}, -\frac{1}{\sqrt{d}}\left(\log\frac{8m^2}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right]$$

Proof of the first statement. Notice that the bound for $\|\mathbf{w}_i\|_2^2$ is already implied by its concentration property. To prove the bound for $\|\mathbf{v}_i\|_2^2$, we write

$$\|\mathbf{v}_i\|_2^2 = \left\|\left(\mathbf{I} - \frac{1}{\|\mathbf{w}_i\|_2^2}\mathbf{w}_i\mathbf{w}_i^\top\right)\hat{\mathbf{v}}_i\right\|_2^2 = \|\hat{\mathbf{v}}_i\|_2^2 - \frac{1}{\|\mathbf{w}\|_i^2}\left(\mathbf{w}_i^\top \hat{\mathbf{v}}_i\right)^2$$

By the concentration property of $\|\mathbf{w}\|_i^2$ and $\mathbf{w}_i^\top \hat{\mathbf{v}}_j$, we have that

$$\frac{1}{\|\mathbf{w}\|_i^2}\left(\mathbf{w}_i^\top \hat{\mathbf{v}}_i\right)^2 \leq \frac{1}{d\left(1 - \frac{1}{2}\beta_s\delta_s\right)}\log\frac{8m^2}{\delta\mathbb{P}}$$

For $d \geq \frac{\beta_5 m^2}{\delta_{\mathbb{P}}} \log \frac{m}{\delta_{\mathbb{P}}} \geq \frac{2}{\beta_2 \delta_s \left(1-\frac{1}{2}\beta_2\delta_s\right)} \log \frac{8m^2}{\delta_{\mathbb{P}}}$, we have that $\frac{1}{\|\mathbf{w}\|_i^2}\left(\mathbf{w}_i^\top \hat{\mathbf{v}}_i\right)^2 \leq \frac{1}{2}\beta_2\delta_s$. Combined with the concentration property of $\hat{\mathbf{v}}_i$, we have that $\|\mathbf{v}_i\|_2^2 \in [1-\beta_2\delta_s, 1+\beta_2\delta_s]$.

Proof of second statement. By the tail bound of Gaussian random variable, we have that for all $z \sim \mathcal{N}(0,1)$, it holds that

$$\Pr\left(z^2 \geq t\right) = 2\Pr\left(z \geq \sqrt{t}\right) \leq \exp\left(-\frac{t}{2}\right)$$

Apply the above to $z = \sqrt{d}\cdot \mathbf{w}_i[j]$ and $\sqrt{d}\cdot\hat{\mathbf{v}}_i[j]$ with a union bound over all $i \in [m]$ and $j \in [m^\star]$ gives

$$\Pr\left(\max_{i\in[m],j\in[m^\star]} \hat{\mathbf{v}}_i[j]^2 \geq \frac{t}{d}; \max_{i\in[m],j\in[m^\star]} \mathbf{w}_i[j]^2 \geq \frac{t}{d}\right) \leq 2mm^\star \exp\left(-\frac{t}{2}\right)$$

Set $t = \frac{\beta_3}{4}\log\frac{m}{\delta_{\mathbb{P}}} \geq 2\log\frac{2mm^\star}{\delta_{\mathbb{P}}}$ gives the desired result for $\mathbf{w}_i[j]^2$s. To bound $\mathbf{v}_i[j]^2$, we notice that

$$|\mathbf{v}_i[j]| \leq |\hat{\mathbf{v}}_i[j]| + \frac{1}{\|\mathbf{w}_i\|_2^2}\left|\mathbf{w}_i^\top \hat{\mathbf{v}}_i \cdot \mathbf{w}_i[j]\right|$$

By the previous bounds, we have that

$$|\hat{\mathbf{v}}_i[j]| \leq \frac{\sqrt{\beta_3}}{2\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right); \frac{1}{\|\mathbf{w}_i\|_2^2}\left|\mathbf{w}_i^\top\hat{\mathbf{v}}_i\cdot\mathbf{w}_i[j]\right| \leq \frac{1}{\sqrt{d}\left(1-\frac{1}{2}\beta_2\delta_s\right)}\left(\log\frac{8m^2}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}} \cdot \frac{\sqrt{\beta_3}}{2\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)$$

With $d \geq \frac{\beta_5 m^4}{\delta_{\mathbb{P}}}\log\frac{m}{\delta_{\mathbb{P}}}$ we can guarantee that the latter is also upper bounded by $\frac{\sqrt{\beta_3}}{2\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)$. Combining the two bounds and square both sides gives the desired result for $\mathbf{v}_i[j]^2$.

Proof of the third statement. The bound of $\mathbf{w}_i^\top\mathbf{w}_j$ is again implied by the concentration we showed above. To show the bound of $\mathbf{v}_i^\top\mathbf{v}_j$, we write

$$\begin{aligned}
\mathbf{v}_i^\top\mathbf{v}_j &= \hat{\mathbf{v}}_i^\top\left(\mathbf{I} - \frac{1}{\|\mathbf{w}_i\|_2^2}\mathbf{w}_i\mathbf{w}_i^\top\right)\left(\mathbf{I} - \frac{1}{\|\mathbf{w}_j\|_2^2}\mathbf{w}_j\mathbf{w}_j^\top\right)\hat{\mathbf{v}}_j \\
&= \hat{\mathbf{v}}_i^\top\hat{\mathbf{v}}_j - \frac{1}{\|\mathbf{w}_i\|_2^2}\mathbf{w}_i^\top\hat{\mathbf{v}}_i\cdot\mathbf{w}_i\top\hat{\mathbf{v}}_j - \frac{1}{\|\mathbf{w}_j\|_2^2}\mathbf{w}_j^\top\hat{\mathbf{v}}_j\cdot\mathbf{w}_j^\top\hat{\mathbf{v}}_i + \frac{1}{\|\mathbf{w}_i\|_2^2\|\mathbf{w}_j\|_2^2}\mathbf{w}_i^\top\hat{\mathbf{v}}_i\cdot\mathbf{w}_j^\top\hat{\mathbf{v}}_j\cdot\mathbf{w}_i^\top\mathbf{w}_j
\end{aligned}$$

By the concentration of the norms and inner-products, we have that

$$\begin{aligned}
\frac{1}{\|\mathbf{w}_i\|_2^2}\left|\mathbf{w}_i^\top\hat{\mathbf{v}}_i\cdot\mathbf{w}_i\top\hat{\mathbf{v}}_j\right|, \frac{1}{\|\mathbf{w}_j\|_2^2}\left|\mathbf{w}_j^\top\hat{\mathbf{v}}_j\cdot\mathbf{w}_j^\top\hat{\mathbf{v}}_i\right| &\leq \frac{1}{d\left(1-\frac{1}{2}\beta_s\delta_s\right)}\log\frac{8m^2}{\delta_{\mathbb{P}}} \\
\frac{1}{\|\mathbf{w}_i\|_2^2\|\mathbf{w}_j\|_2^2}\left|\mathbf{w}_i^\top\hat{\mathbf{v}}_i\cdot\mathbf{w}_j^\top\hat{\mathbf{v}}_j\cdot\mathbf{w}_i^\top\mathbf{w}_j\right| &\leq \frac{1}{d^{\frac{3}{2}}\left(1-\frac{1}{2}\beta_s\delta_s\right)^2}\left(\log\frac{8m^2}{\delta_{\mathbb{P}}}\right)^{\frac{3}{2}}
\end{aligned}$$

With the condition $d \geq \frac{\beta_5 m^2}{\delta_{\mathbb{P}}}\log\frac{m}{\delta_{\mathbb{P}}}$ we have that

$$\frac{1}{d\left(1-\frac{1}{2}\beta_s\delta_s\right)^2}\log\frac{8m^2}{\delta_{\mathbb{P}}} \leq \frac{1}{6\sqrt{d}}\left(\log\frac{8m^2}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}} \leq \frac{\sqrt{\beta_3}}{6\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}} \leq 1$$

Therefore, we can conclude that

$$\frac{1}{\|\mathbf{w}_i\|_2^2}\left|\mathbf{w}_i^\top\hat{\mathbf{v}}_i\cdot\mathbf{w}_i\top\hat{\mathbf{v}}_j\right|, \frac{1}{\|\mathbf{w}_j\|_2^2}\left|\mathbf{w}_j^\top\hat{\mathbf{v}}_j\cdot\mathbf{w}_j^\top\hat{\mathbf{v}}_i\right|, \frac{1}{\|\mathbf{w}_i\|_2^2\|\mathbf{w}_j\|_2^2}\left|\mathbf{w}_i^\top\hat{\mathbf{v}}_i\cdot\mathbf{w}_j^\top\hat{\mathbf{v}}_j\cdot\mathbf{w}_i^\top\mathbf{w}_j\right| \leq \frac{\sqrt{\beta_3}}{6\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}$$

Combined with the bound on $\hat{\mathbf{v}}_i^\top\hat{\mathbf{v}}_j$ gives that

$$\mathbf{v}_i^\top\mathbf{v}_j \in \left[-\frac{\sqrt{\beta_3}}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}, \frac{\sqrt{\beta_3}}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right]$$

Squaring both sides gives the desired result. Lastly, to bound $\mathbf{v}_i^\top \mathbf{w}_j$, we write

$$\mathbf{v}_i^\top \mathbf{w}_j = \mathbf{w}_j^\top \left(\mathbf{I} - \frac{1}{\|\mathbf{w}_i\|_2^2}\mathbf{w}_i\mathbf{w}_i^\top\right)\hat{\mathbf{v}}_i = \hat{\mathbf{v}}_i^\top \mathbf{w}_j - \frac{1}{\|\mathbf{w}_i\|_2^2}\mathbf{w}_i^\top \hat{\mathbf{v}}_i \cdot \mathbf{w}_i^\top \mathbf{w}_j$$

Similar to the above, we have that

$$\frac{1}{\|\mathbf{w}_i\|_2^2}\mathbf{w}_i^\top \hat{\mathbf{v}}_i \cdot \mathbf{w}_i^\top \mathbf{w}_j \le \frac{1}{d\left(1-\frac{1}{2}\beta_s\delta_s\right)}\log\frac{8m^2}{\delta_{\mathbb{P}}} \le \frac{\sqrt{\beta_3}}{6\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}$$

Therefore, we can conclude that

$$\mathbf{v}_i^\top \mathbf{w}_j \in \left[-\frac{\sqrt{\beta_3}}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}, \frac{\sqrt{\beta_3}}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right]$$

Squaring both sides gives the desired result. □

Lemma 4. *[Restatement of Lemma 2] Let $\mathbf{w}_1, \ldots, \mathbf{w}_m \sim \mathcal{N}\left(0, d^{-1}\mathbf{I}_d\right)$ be I.I.D. Gaussian random vectors. Define*

$$i_\ell^\star, j_\ell^\star = \underset{i\in[m]\setminus\mathcal{R}_{\ell-1}, j\in[m^\star]\setminus\mathcal{C}_{\ell-1}}{\arg\max} \mathbf{w}_i[j]; \mathcal{R}_\ell = \{i_k^\star\}_{k=1}^\ell; \mathcal{C}_\ell = \{j_k^\star\}_{k=1}^\ell$$

Let any $\delta_{\mathbb{P}} \in (0, 1/2)$ be given. Then there exists some absolute constant $\beta_2, \beta_4 > 0$ such that if $m \ge \beta_4 m^\star \log\frac{m^\star}{\delta_{\mathbb{P}}}$, then for $\delta_s = \frac{\beta_2\delta_{\mathbb{P}}}{m^2}$, with probability at least $1 - 4\delta_{\mathbb{P}}$, it holds that

- *(Row-wise Gap) $\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \ge (1+2\delta_s)\,\mathbf{w}_{i_\ell^\star}[j]$ for all $\ell \in [m^\star]$ and $j \in [m^\star]\setminus\mathcal{C}_\ell$*
- *(Column-wise Gap) $\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \ge (1+2\delta_s)\,\mathbf{w}_i[j_\ell^\star]$ for all $\ell \in [m^\star]$ and $i \in [m]\setminus\mathcal{R}_\ell$*
- *(Threshold Gap) $\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \ge (1+2\delta_s)\,\mathbf{w}_{i_{\ell+1}^\star}[j_{\ell+1}^\star]^2$ for all $\ell \in [m^\star - 1]$*
- *(Magnitude Lower Bound) $\mathbf{v}_{i_\ell^\star}[j_\ell^\star]^2 \ge \frac{\log m^\star}{d}$ for all $\ell \in [m]$*

Proof. We start by proving an auxiliary result that, with high probability, there are at least $\frac{m}{3}$ out of the m $\mathbf{w}_i[j]$'s that are positive for each $j \in [m]^\star$. Define

$$s_{i,j} = \mathbb{I}\left\{\mathbf{w}_i[j] > 0\right\}; S_j = \sum_{i=1}^m s_{i,j}$$

Due to the symmetry of Gaussian, we have that $s_{i,j} \sim \texttt{Bern}\,(0.5)$ independently. Therefore, $\mathbb{E}[S_j] = \frac{m}{2}$. By Hoeffding's inequality, we have that

$$\Pr\left(S_j - \frac{m}{2} \le -t\right) \le \exp\left(-\frac{2t^2}{m}\right)$$

Setting $t = \frac{m}{6}$ and take a union bound over all $j \in [m^\star]$ gives that

$$\Pr\left(S_j \ge \frac{m}{3}; \forall j \in [m^\star]\right) \le m^\star \exp\left(-\frac{m}{18}\right)$$

Since $m \ge \beta_4 m^\star \log\frac{m^\star}{\delta_{\mathbb{P}}} \ge 18\log\frac{m^\star}{\delta_{\mathbb{P}}}$, with probability at least $1-\delta_{\mathbb{P}}$, we have that at least $\frac{m}{3}$ out of $\{\mathbf{w}_i[j]\}_{i=1}^m$ are positive for all $j \in [m]$.

Proof of the first statement. Let any $i_1, i_2 \in [m]$ and $j_1, j_2 \in [m^\star]$ such that (i_1, j_1) and (i_2, j_2) differ in at least one coordinate. Then we have that $\mathbf{w}_{i_1}[j_1]$ and $\mathbf{w}_{i_2}[j_2]$ are I.I.D. Gaussian random variables in $\mathcal{N}\left(0, d^{-1}\right)$. Therefore, $\frac{\mathbf{w}_{i_1}[j_1]}{\mathbf{w}_{i_2}[j_2]}$ is a standard Cauchy random variable. Given the condition that $\mathbf{w}_{i_1}[j_1], \mathbf{w}_{i_2}[j_2] \ge 0$, we

have that $\mathbf{w}_{i_1}[j_1], \mathbf{w}_{i_2}[j_2]$ are half-Gaussian, and thus $\frac{\mathbf{w}_{i_1}[j_1]}{\mathbf{w}_{i_2}[j_2]}$ is half-Cauchy. Using the CDF of Cauchy random variables, we have that for any $\delta \in (0, 1/2)$

$$\begin{aligned}
&\Pr\left(\mathbf{w}_{i_1}[j_1] \in (\mathbf{w}_{i_2}[j_2], (1+\delta)\,\mathbf{w}_{i_2}[j_2]) \mid \mathbf{w}_{i_1}[j_1], \mathbf{w}_{i_2}[j_2] \geq 0\right) \\
&\quad = \Pr\left(\left|\frac{\mathbf{w}_{i_1}[j_1]}{\mathbf{w}_{i_2}[j_2]}\right| \in (1, 1+\delta)\right) \\
&\quad = 2\Pr\left(\frac{\mathbf{w}_{i_1}[j_1]}{\mathbf{w}_{i_2}[j_2]} \in (1, 1+\delta)\right) \\
&\quad = \frac{2}{\pi}\left(\arctan(1+\delta) - \arctan 1\right) \\
&\quad = \frac{2}{\pi}\arctan\frac{\delta}{2+\delta} \\
&\quad \leq \frac{\delta}{\pi}
\end{aligned}$$

where the last inequality follows from $\arctan x \leq x$ with $x \geq 0$ and $\delta > 0$. To prove the first property, we let $j_1 = j_\ell^\star, j_2 = j$ and $i_1 = i_2 = i_\ell^\star$. Fix any $\ell \in [m^\star]$, we have that

$$\Pr\left(\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \in \left(\mathbf{w}_{i_\ell^\star}[j], (1+\delta)\,\mathbf{w}_{i_\ell^\star}[j]\right); \forall j \in [m^\star] \setminus \mathcal{C}_\ell \text{ s.t. } \mathbf{w}_{i_\ell^\star}[j] \geq 0\right) \leq m^\star\delta$$

Recall that $\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \geq \mathbf{w}_{i_\ell^\star}[j]$ for all $j \in [m^\star] \setminus \mathcal{C}_\ell$. If $\mathbf{w}_{i_\ell^\star}[j] < 0$, since by definition $\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \geq 0$, it must hold that $\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \geq (1+2\delta_s)\mathbf{w}_{i_\ell^\star}[j]$. Take a union bound over $\ell \in [m^\star]$, and set $\delta = 2\delta_s$ with $\delta_s \leq \frac{\delta_{\mathbb{P}}}{2mm^\star}$ gives the first property.

Proof of the second statement. To prove the second property, we set $i_1 = i_\ell^\star, i_2 = i$ and $j_1 = j_2 = j_\ell^\star$. Fix any $\ell \in [m^\star]$, we have that

$$\Pr\left(\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \in (\mathbf{w}_i[j_\ell^\star], (1+\delta)\,\mathbf{w}_i[j_\ell^\star]); \forall i \in [m] \setminus \mathcal{R}_\ell \text{ s.t. } \mathbf{w}_i[j_\ell^\star] \geq 0\right) \leq m\delta$$

Recall that $\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \geq \mathbf{w}_i[j_\ell^\star]$ for all $i \in [m] \setminus \mathcal{R}_\ell$. If $\mathbf{w}_i[j_\ell^\star] < 0$, since by definition $\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \geq 0$, it must hold that $\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \geq (1+2\delta_s)\mathbf{w}_i[j_\ell^\star]$. Take a union bound over $\ell \in [m^\star]$, and set $\delta = 2\delta_s$ with $\delta_s \leq \frac{\delta_{\mathbb{P}}}{2m^{\star 2}}$ gives the second property.

Proof of the third statement. To prove the third property, we set $i_1 = i_\ell^\star, i_2 = i_{\ell+1}^\star$ and $j_1 = j_\ell^\star, j_1 = j_{\ell+1}^\star$. Fix any $\ell \in [m^\star - 1]$, we have that

$$\Pr\left(\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \in \left(\mathbf{w}_{i_{\ell+1}^\star}[j_{\ell+1}^\star], (1+\delta)\,\mathbf{w}_{i_{\ell+1}^\star}[j_{\ell+1}^\star]\right)\right) \leq \delta$$

Recall that $\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \geq \mathbf{w}_{i_{\ell+1}^\star}[j_{\ell+1}^\star] \geq 0$ for all $\ell \in [m^\star - 1]$. Take a union bound over $\ell \in [m^\star - 1]$, and set $\delta = 2\delta_s$ with $\delta_s \leq \frac{\delta_{\mathbb{P}}}{2m^\star}$ gives the third property.

Proof of the fourth statement. To show the last result, we notice that by definition of $i_\ell^\star$, it must holds that

$$\mathbf{w}_{i_\ell^\star}[j_\ell^\star] \geq \mathbf{w}_i[j_\ell^\star]; \quad \forall i \in [m] \setminus \mathcal{R}_\ell$$

This gives that for any $\gamma > 0$

$$\begin{aligned}
\Pr\left(\mathbf{w}_{i_\ell^\star}[j_\ell^\star] < \frac{\gamma}{\sqrt{d}}\right) &\leq \prod_{i \in [m]\setminus\mathcal{R}_\ell} \Pr\left(\mathbf{w}_i[j_\ell^\star] < \frac{\gamma}{\sqrt{d}}\right) \\
&= \prod_{i \in [m]\setminus\mathcal{R}_\ell} \Pr\left(\mathbf{w}_i[j_\ell^\star] < \frac{\gamma}{\sqrt{d}}\right) \\
&= \Pr_{z \sim \mathcal{N}(0,1)}(z < \gamma)^{m-m^\star} \\
&\leq \left(1 - \Pr_{z \sim \mathcal{N}(0,1)}(z \geq \gamma)\right)^{m-m^\star}
\end{aligned}$$

By the tail bound of Gaussian random variable, we have that

$$\Pr_{z\sim\mathcal{N}(0,1)}(z\geq\gamma)\geq\frac{1}{\sqrt{2\pi}\cdot\gamma}\exp\left(-\frac{\gamma^2}{2}\right)$$

Therefore

$$\Pr\left(\mathbf{w}_{i_\ell^\star}[j_\ell^\star]<\frac{\gamma}{\sqrt{d}}\right)\leq\left(1-\frac{1}{\sqrt{2\pi}\cdot\gamma}\exp\left(-\frac{\gamma^2}{2}\right)\right)^{m-m^\star}$$

Take a union bound over all $\ell\in[m^\star]$ gives that

$$\Pr\left(\mathbf{w}_{i_\ell^\star}[j_\ell^\star]^2\geq\frac{\gamma^2}{d}\right)\geq 1-m^\star\left(1-\frac{1}{\sqrt{2\pi}\cdot\gamma}\exp\left(-\frac{\gamma^2}{2}\right)\right)^{m-m^\star}$$

For the failing probability to be upper bounded by $\delta_{\mathbb{P}}$, we simply need

$$m\geq m^\star+\frac{\log\frac{m^\star}{\delta_{\mathbb{P}}}}{\log\left(1-\frac{1}{\sqrt{2\pi}\cdot\gamma}\exp\left(-\frac{\gamma^2}{2}\right)\right)^{-1}}$$

Using $\log\frac{1}{x}\geq 1-x$ for all $x>0$, it suffice to guarantee that

$$m\geq m^\star+\sqrt{2\pi}\cdot\gamma\exp\left(\frac{\gamma^2}{2}\right)\log\frac{m^\star}{\delta_{\mathbb{P}}}$$

Setting $\gamma=(\log m^\star)^{\frac{1}{2}}$ gives the desired result. □

A.3 Hermite Expansion of the Gradient and the Gradient Flow Dynamics

Notice that the population MSE has the following form

$$\mathcal{L}(\boldsymbol{\theta})=\frac{1}{2}\mathbb{E}_{\mathbf{x}}\Big[f(\boldsymbol{\theta},\mathbf{x})^2-2f(\boldsymbol{\theta},\mathbf{x})f(\boldsymbol{\theta}^\star,\mathbf{x})+f(\boldsymbol{\theta}^\star,\mathbf{x})^2\Big]$$

where the last term is independent of $\boldsymbol{\theta}^\star$, and is thus omitted from our analysis. Consider the Hermite expansion of $\pi(\cdot)$ and $\sigma(\cdot)$, respectively

$$\pi(x)=\sum_{k=0}^{\infty}c_k He_k(x);\quad \sigma(x)=He_3(x)$$

where $c_k=\mathbb{E}_{x\sim\mathcal{N}(0,1)}\big[\pi^{(k)}(x)\big]$. Then we have that

$$f(\boldsymbol{\theta},\mathbf{x})^2=\sum_{i,j=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_kc_\ell}{k!\ell!}He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_j\right)$$

$$f(\boldsymbol{\theta},\mathbf{x})f(\boldsymbol{\theta}^\star\mathbf{x})=\sum_{i=1}^{m}\sum_{j=1}^{m^\star}\sum_{k,\ell=0}^{\infty}\frac{c_kc_\ell}{k!\ell!}He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j^\star\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_j^\star\right)$$

This gives that

$$\frac{\partial}{\partial\mathbf{v}_i}f(\boldsymbol{\theta},\mathbf{x})^2=2\sum_{j=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_\ell}{k!\ell!}\cdot He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_j\right)\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\frac{\mathbf{x}}{\|\mathbf{v}_i\|_2}$$

Taking the expectation gives

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}}\left[\frac{\partial}{\partial \mathbf{v}_i} f\left(\boldsymbol{\theta}, \mathbf{x}\right)^2\right] \\
&\quad= \frac{2\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)}{\left\|\mathbf{v}_i\right\|_2}\sum_{j=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_\ell}{k!\ell!}\mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_j\right)\mathbf{x}\big] \\
&\quad= \frac{2\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)}{\left\|\mathbf{v}_i\right\|_2}\sum_{j=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_{\ell+1}}{k!\ell!}\mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_j\right)\big]\bar{\mathbf{v}}_j \\
&\qquad+ \frac{6\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)}{\left\|\mathbf{v}_i\right\|_2}\sum_{j=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_\ell}{k!\ell!}\mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j\right)He_2\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_j\right)\big]\bar{\mathbf{w}}_i \\
&\qquad+ \frac{6\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)}{\left\|\mathbf{v}_i\right\|_2}\sum_{j=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_\ell}{k!\ell!}\mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_2\left(\mathbf{x}^\top\bar{\mathbf{w}}_j\right)\big]\bar{\mathbf{w}}_j
\end{aligned}$$

Similar, we can obtain that

$$\begin{aligned}
&2\mathbb{E}_{\mathbf{x}}\left[\frac{\partial}{\partial \mathbf{v}_i} f\left(\boldsymbol{\theta}, \mathbf{x}\right) f\left(\boldsymbol{\theta}^\star, \mathbf{x}\right)\right] \\
&\quad= \frac{2\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)}{\left\|\mathbf{v}_i\right\|_2}\sum_{j=1}^{m^\star}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_{\ell+1}}{k!\ell!}\mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j^\star\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_j^\star\right)\big]\bar{\mathbf{v}}_j^\star \\
&\qquad+ \frac{6\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)}{\left\|\mathbf{v}_i\right\|_2}\sum_{j=1}^{m^\star}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_\ell}{k!\ell!}\mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j^\star\right)He_2\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_j^\star\right)\big]\bar{\mathbf{w}}_i \\
&\qquad+ \frac{6\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)}{\left\|\mathbf{v}_i\right\|_2}\sum_{j=1}^{m^\star}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_\ell}{k!\ell!}\mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j^\star\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_2\left(\mathbf{x}^\top\bar{\mathbf{w}}_j^\star\right)\big]\bar{\mathbf{w}}_j^\star
\end{aligned}$$

For the gradient of $\mathbf{w}_i$, we can compute that

$$\frac{\partial}{\partial \mathbf{w}_i} f\left(\boldsymbol{\theta}, \mathbf{x}\right)^2 = 6\sum_{j=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_k c_\ell}{k!\ell!}He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j\right)He_2\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_j\right)\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\frac{\mathbf{x}}{\left\|\bar{\mathbf{w}}_i\right\|_2}$$

Taking the expectation gives

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}}\left[\frac{\partial}{\partial \mathbf{w}_i} f\left(\boldsymbol{\theta}, \mathbf{x}\right)^2\right] \\
&\quad= \frac{6\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)}{\left\|\mathbf{w}_i\right\|_2}\sum_{j=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_\ell}{k!\ell!}\mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j\right)He_2\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_j\right)\big]\bar{\mathbf{v}}_i \\
&\qquad+ \frac{6\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)}{\left\|\mathbf{w}_i\right\|_2}\sum_{j=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_k c_{\ell+1}}{k!\ell!}\mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j\right)He_2\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_j\right)\big]\bar{\mathbf{v}}_j \\
&\qquad+ \frac{18\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)}{\left\|\mathbf{w}_i\right\|_2}\sum_{j=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_k c_\ell}{k!\ell!}\mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j\right)He_2\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_2\left(\mathbf{x}^\top\bar{\mathbf{w}}_j\right)\big]\bar{\mathbf{w}}_j
\end{aligned}$$

Similarly, we have

$$
\begin{aligned}
&2\mathbb{E}_{\mathbf{x}}\left[\frac{\partial}{\partial \mathbf{w}_i} f\left(\boldsymbol{\theta}, \mathbf{x}\right) f\left(\boldsymbol{\theta}^\star, \mathbf{x}\right)\right] \\
&\quad= \frac{6\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)}{\left\|\mathbf{w}_i\right\|_2}\sum_{j=1}^{m^\star}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_\ell}{k!\ell!}\mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j^\star\right)He_2\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_j^\star\right)\big]\bar{\mathbf{v}}_i \\
&\qquad+ \frac{6\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)}{\left\|\mathbf{w}_i\right\|_2}\sum_{j=1}^{m^\star}\sum_{k,\ell=0}^{\infty}\frac{c_kc_{\ell+1}}{k!\ell!}\mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j^\star\right)He_2\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_3\left(\mathbf{x}^\top\bar{\mathbf{w}}_j^\star\right)\big]\bar{\mathbf{v}}_j^\star \\
&\qquad+ \frac{18\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)}{\left\|\mathbf{w}_i\right\|_2}\sum_{j=1}^{m^\star}\sum_{k,\ell=0}^{\infty}\frac{c_kc_\ell}{k!\ell!}\mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j^\star\right)He_2\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_2\left(\mathbf{x}^\top\bar{\mathbf{w}}_j^\star\right)\big]\bar{\mathbf{w}}_j^\star
\end{aligned}
$$

This gives that

$$
\begin{aligned}
\mathbb{E}_{\mathbf{x}}\left[\frac{\partial}{\partial \mathbf{v}_i} f\left(\boldsymbol{\theta}, \mathbf{x}\right) f\left(\boldsymbol{\theta}^\star, \mathbf{x}\right)\right]^\top\mathbf{v}_i = \mathbb{E}_{\mathbf{x}}\left[\frac{\partial}{\partial \mathbf{v}_i} f\left(\boldsymbol{\theta}, \mathbf{x}\right)^2\right]^\top\mathbf{v}_i = 0 \\
\mathbb{E}_{\mathbf{x}}\left[\frac{\partial}{\partial \mathbf{w}_i} f\left(\boldsymbol{\theta}, \mathbf{x}\right) f\left(\boldsymbol{\theta}^\star, \mathbf{x}\right)\right]^\top\mathbf{w}_i = \mathbb{E}_{\mathbf{x}}\left[\frac{\partial}{\partial \mathbf{w}_i} f\left(\boldsymbol{\theta}, \mathbf{x}\right)^2\right]^\top\mathbf{w}_i = 0
\end{aligned}
$$

According to the gradient flow dynamics, we can conclude that

$$
\frac{d}{dt}\left\|\mathbf{v}_i(t)\right\|_2^2 = \frac{d}{dt}\left\|\mathbf{w}_i(t)\right\|_2^2 = 0
$$

which implies that the norm of each $\mathbf{v}_i$ and $\mathbf{w}_i$ are fixed at initialization. For the convenience of the analysis, we shall denote

$$
\begin{aligned}
\mathcal{C}_{k,\ell,a,b}^{i,j} &= \mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j\right)He_a\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_b\left(\mathbf{x}^\top\bar{\mathbf{w}}_j\right)\big] \\
\hat{\mathcal{C}}_{k,\ell,a,b}^{i,j} &= \mathbb{E}_{\mathbf{x}}\big[He_k\left(\mathbf{x}^\top\bar{\mathbf{v}}_i\right)He_\ell\left(\mathbf{x}^\top\bar{\mathbf{v}}_j^\star\right)He_a\left(\mathbf{x}^\top\bar{\mathbf{w}}_i\right)He_b\left(\mathbf{x}^\top\bar{\mathbf{w}}_j^\star\right)\big]
\end{aligned}
$$

and that $\left\|\bar{\mathbf{v}}_i(t)\right\|_2 = a_i, \left\|\bar{\mathbf{w}}_i(t)\right\|_2 = b_i$ Then we have that

$$
\begin{aligned}
\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}\right) &= \frac{\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top}{\left\|\mathbf{v}_i\right\|_2}\left(\sum_{r=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_{\ell+1}}{k!\ell!}\mathcal{C}_{k,\ell,3,3}^{i,r}\bar{\mathbf{v}}_r - \sum_{r=1}^{m^\star}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_{\ell+1}}{k!\ell!}\hat{\mathcal{C}}_{k,\ell,3,3}^{i,r}\bar{\mathbf{v}}_r^\star\right) \\
&\quad+ \frac{3\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)}{\left\|\mathbf{v}_i\right\|_2}\left(\sum_{r=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_\ell}{k!\ell!}\mathcal{C}_{k,\ell,2,3}^{i,r} - \sum_{r=1}^{m^\star}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_\ell}{k!\ell!}\hat{\mathcal{C}}_{k,\ell,2,3}^{i,r}\right)\bar{\mathbf{w}}_i \\
&\quad+ \frac{3\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)}{\left\|\mathbf{v}_i\right\|_2}\left(\sum_{r=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_\ell}{k!\ell!}\mathcal{C}_{k,\ell,3,2}^{i,r}\bar{\mathbf{w}}_r - \sum_{r=1}^{m^\star}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_\ell}{k!\ell!}\hat{\mathcal{C}}_{k,\ell,3,2}^{i,r}\bar{\mathbf{w}}_r^\star\right) \\
\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}\right) &= \frac{3\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)}{\left\|\mathbf{w}_i\right\|_2}\left(\sum_{r=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_\ell}{k!\ell!}\mathcal{C}_{k,\ell,2,3}^{i,r} - \sum_{r=1}^{m^\star}\sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_\ell}{k!\ell!}\hat{\mathcal{C}}_{k,\ell,2,3}^{i,r}\right)\bar{\mathbf{v}}_i \\
&\quad+ \frac{3\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)}{\left\|\mathbf{w}_i\right\|_2}\left(\sum_{r=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_kc_{\ell+1}}{k!\ell!}\mathcal{C}_{k,\ell,2,3}^{i,r}\bar{\mathbf{v}}_r - \sum_{r=1}^{m^\star}\sum_{k,\ell=0}^{\infty}\frac{c_kc_{\ell+1}}{k!\ell!}\hat{\mathcal{C}}_{k,\ell,2,3}^{i,r}\bar{\mathbf{v}}_r^\star\right) \\
&\quad+ \frac{9\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)}{\left\|\mathbf{w}_i\right\|_2}\left(\sum_{r=1}^{m}\sum_{k,\ell=0}^{\infty}\frac{c_kc_\ell}{k!\ell!}\mathcal{C}_{k,\ell,2,2}^{i,r}\bar{\mathbf{w}}_r - \sum_{r=1}^{m^\star}\sum_{k,\ell=0}^{\infty}\frac{c_kc_\ell}{k!\ell!}\hat{\mathcal{C}}_{k,\ell,2,2}^{i,r}\bar{\mathbf{w}}_r^\star\right)
\end{aligned}
\tag{10}
$$

In particular, we notice that there there are several quantities that appears in the form of the gradient. We make the following definition for the convenience of the analysis

$$\lambda_{i,j,1} = \sum_{k,\ell=0}^{\infty} \frac{c_{k+1}c_{\ell+1}}{k!\ell!}\mathcal{C}^{i,j}_{k,\ell,3,3}; \quad \hat{\lambda}_{i,j,1} = \sum_{k,\ell=0}^{\infty} \frac{c_{k+1}c_{\ell+1}}{k!\ell!}\hat{\mathcal{C}}^{i,j}_{k,\ell,3,3};$$
$$\lambda_{i,j,2} = \sum_{k,\ell=0}^{\infty} \frac{c_{k+1}c_{\ell}}{k!\ell!}\mathcal{C}^{i,j}_{k,\ell,2,3}; \quad \hat{\lambda}_{i,j,2} = \sum_{k,\ell=0}^{\infty} \frac{c_{k+1}c_{\ell}}{k!\ell!}\hat{\mathcal{C}}^{i,j}_{k,\ell,2,3}$$
$$\lambda_{i,j,3} = \sum_{k,\ell=0}^{\infty} \frac{c_{k}c_{\ell+1}}{k!\ell!}\mathcal{C}^{i,j}_{k,\ell,2,3}; \quad \hat{\lambda}_{i,j,3} = \sum_{k,\ell=0}^{\infty} \frac{c_{k}c_{\ell+1}}{k!\ell!}\hat{\mathcal{C}}^{i,j}_{k,\ell,2,3}$$
$$\lambda_{i,j,4} = \sum_{k,\ell=0}^{\infty} \frac{c_{k+1}c_{\ell}}{k!\ell!}\mathcal{C}^{i,j}_{k,\ell,3,2}; \quad \hat{\lambda}_{i,j,4} = \sum_{k,\ell=0}^{\infty} \frac{c_{k+1}c_{\ell}}{k!\ell!}\hat{\mathcal{C}}^{i,j}_{k,\ell,3,2}$$
$$\lambda_{i,j,5} = \sum_{k,\ell=0}^{\infty} \frac{c_{k}c_{\ell}}{k!\ell!}\mathcal{C}^{i,j}_{k,\ell,2,2}; \quad \hat{\lambda}_{i,j,5} = \sum_{k,\ell=0}^{\infty} \frac{c_{k}c_{\ell}}{k!\ell!}\hat{\mathcal{C}}^{i,j}_{k,\ell,2,2}$$

Recall that our goal is to study the dynamics of the following alignment scores

$$\gamma^{(1)}_{i,j}(t) = \bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{v}}^\star_j; \quad \gamma^{(2)}_{i,j}(t) = \bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{w}}^\star_j; \quad \zeta^{(1)}_{i,j}(t) = \bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{w}}^\star_j; \quad \zeta^{(2)}_{i,j}(t) = \bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{v}}^\star_j$$
$$I^{(1)}_{i,j}(t) = \bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{v}}_j(t); \quad I^{(2)}_{i,j}(t) = \bar{\mathbf{w}}_i(t)^\top \bar{\mathbf{w}}_j(t); \quad I^{(3)}_{i,j}(t) = \bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{w}}_j(t) \tag{11}$$

This allows us to rewrite the gradient as

$$\begin{aligned}
\nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{a_i}\left(\mathbf{I} - \bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\left(\sum_{r=1}^{m}(\lambda_{i,r,1}\bar{\mathbf{v}}_r + 3\lambda_{i,r,2}\bar{\mathbf{w}}_i + 3\lambda_{i,r,4}\bar{\mathbf{w}}_r) - \sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,1}\bar{\mathbf{v}}^\star_r + 3\hat{\lambda}_{i,r,2}\bar{\mathbf{w}}_i + 3\hat{\lambda}_{i,r,4}\bar{\mathbf{w}}^\star_r\right)\right) \\
&= \frac{1}{a_i}\sum_{r=1}^{m}(\lambda_{i,r,1}\bar{\mathbf{v}}_r + 3\lambda_{i,r,2}\bar{\mathbf{w}}_i + 3\lambda_{i,r,4}\bar{\mathbf{w}}_r) - \frac{1}{a_i}\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,1}\bar{\mathbf{v}}^\star_r + 3\hat{\lambda}_{i,r,2}\bar{\mathbf{w}}_i + 3\hat{\lambda}_{i,r,4}\bar{\mathbf{w}}^\star_r\right) \\
&\quad - \frac{1}{a_i}\sum_{r=1}^{m}\left(\lambda_{i,r,1}I^{(1)}_{i,r} + 3\lambda_{i,r,2}I^{(3)}_{i,i} + 3\lambda_{i,r,4}I^{(3)}_{i,r}\right)\bar{\mathbf{v}}_i \\
&\quad + \frac{1}{a_i}\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,1}\gamma^{(1)}_{i,r} + 3\hat{\lambda}_{i,r,2}I^{(3)}_{i,i} + 3\hat{\lambda}_{i,r,4}\zeta^{(1)}_{i,r}\right)\bar{\mathbf{v}}_i \\
\nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta}) &= \frac{3}{b_i}\left(\mathbf{I} - \bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\left(\sum_{r=1}^{m}(\lambda_{i,r,2}\bar{\mathbf{v}}_i + \lambda_{i,r,3}\bar{\mathbf{v}}_r + 3\lambda_{i,r,5}\bar{\mathbf{w}}_r) - \sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,2}\bar{\mathbf{v}}_i + \hat{\lambda}_{i,r,3}\bar{\mathbf{v}}^\star_r + 3\hat{\lambda}_{i,r,5}\bar{\mathbf{w}}^\star_r\right)\right) \\
&= \frac{3}{b_i}\sum_{r=1}^{m}(\lambda_{i,r,2}\bar{\mathbf{v}}_i + \lambda_{i,r,3}\bar{\mathbf{v}}_r + 3\lambda_{i,r,5}\bar{\mathbf{w}}_r) - \frac{3}{b_i}\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,2}\bar{\mathbf{v}}_i + \hat{\lambda}_{i,r,3}\bar{\mathbf{v}}^\star_r + 3\hat{\lambda}_{i,r,5}\bar{\mathbf{w}}^\star_r\right) \\
&\quad - \frac{3}{b_i}\sum_{r=1}^{m}\left(\lambda_{i,r,2}I^{(3)}_{i,i} + \lambda_{i,r,3}I^{(3)}_{r,i} + 3\lambda_{i,r,5}I^{(2)}_{r,i}\right)\bar{\mathbf{w}}_i \\
&\quad + \frac{3}{b_i}\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,2}I^{(3)}_{i,i} + \hat{\lambda}_{i,r,3}\zeta^{(2)}_{i,r} + 3\hat{\lambda}_{i,r,5}\gamma^{(2)}_{i,r}\right)\bar{\mathbf{w}}_i
\end{aligned}$$

Recall that our goal is to study the dynamics in (9). By the gradient flow dynamic, we have that

$$\frac{d}{dt}\gamma^{(1)}_{i,j}(t) = -\frac{1}{a_i}\cdot\nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top\bar{\mathbf{v}}^\star_j; \quad \frac{d}{dt}\gamma^{(2)}_{i,j}(t) = -\frac{1}{b_i}\cdot\nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top\bar{\mathbf{w}}^\star_j$$
$$\frac{d}{dt}\zeta^{(1)}_{i,j}(t) = -\frac{1}{a_i}\cdot\nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top\bar{\mathbf{w}}^\star_j; \quad \frac{d}{dt}\zeta^{(2)}_{i,j}(t) = -\frac{1}{b_i}\cdot\nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top\bar{\mathbf{v}}^\star_j$$

Moreover, we also have that

$$\frac{d}{dt}I_{i,j}^{(1)}(t) = -\frac{1}{a_i}\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_j - \frac{1}{a_j}\nabla_{\mathbf{v}_j}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_i$$
$$\frac{d}{dt}I_{i,j}^{(2)}(t) = -\frac{1}{b_i}\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}_j - \frac{1}{b_j}\nabla_{\mathbf{w}_j}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}_i$$
$$\frac{d}{dt}I_{i,j}^{(3)}(t) = -\frac{1}{a_i}\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}_j - \frac{1}{b_j}\nabla_{\mathbf{w}_j}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_i$$

Therefore, we should consider inner product between the gradient and the vectors $\bar{\mathbf{v}}_i, \bar{\mathbf{w}}_i, \bar{\mathbf{v}}_j^\star, \bar{\mathbf{w}}_j^\star$ above, which will give us exactly eight terms to analyze. The inner product between $\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)$ and $\bar{\mathbf{v}}_j^\star, \bar{\mathbf{w}}_j^\star$ can be written as

$$\begin{aligned}
-\frac{1}{a_i}\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_j^\star = {} & \frac{1}{a_i^2}\left(\hat{\lambda}_{i,j,1}(t) - \sum_{r=1}^{m}\lambda_{i,r,1}(t)\gamma_{r,j}^{(1)}(t)\right) \\
& - \frac{3}{a_i^2}\left(\sum_{r=1}^{m}\left(\lambda_{i,r,2}(t)\zeta_{i,j}^{(2)}(t) + \lambda_{i,r,4}\zeta_{r,j}^{(2)}(t)\right) - \sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,2}(t)\zeta_{i,j}^{(2)}(t)\right) \\
& + \frac{1}{a_i^2}\sum_{r=1}^{m}\left(\lambda_{i,r,1}(t)I_{i,r}^{(1)}(t) + 3\lambda_{i,r,2}(t)I_{i,i}^{(3)}(t) + 3\lambda_{i,r,4}(t)I_{i,r}^{(3)}(t)\right)\gamma_{i,j}^{(1)}(t) \\
& - \frac{1}{a_i^2}\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,1}(t)\gamma_{i,r}^{(1)}(t) + 3\hat{\lambda}_{i,r,2}(t)I_{i,i}^{(3)}(t) + 3\hat{\lambda}_{i,r,4}(t)\zeta_{i,r}^{(1)}(t)\right)\gamma_{i,j}^{(1)}(t) \\
-\frac{1}{a_i}\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}_j^\star = {} & \frac{3}{a_i^2}\left(\hat{\lambda}_{i,j,4}(t) - \sum_{r=1}^{m}\lambda_{i,r,4}(t)\gamma_{r,j}^{(2)}(t)\right) \\
& - \frac{1}{a_i^2}\left(\sum_{r=1}^{m}\left(\lambda_{i,r,1}(t)\zeta_{r,j}^{(1)}(t) + 3\lambda_{i,r,2}(t)\gamma_{i,j}^{(2)}(t)\right) - 3\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,2}(t)\gamma_{i,j}^{(2)}(t)\right) \\
& + \frac{1}{a_i^2}\sum_{r=1}^{m}\left(\lambda_{i,r,1}(t)I_{i,r}^{(1)}(t) + 3\lambda_{i,r,2}(t)I_{i,i}^{(3)}(t) + 3\lambda_{i,r,4}(t)I_{i,r}^{(3)}(t)\right)\zeta_{i,j}^{(1)}(t) \\
& - \frac{1}{a_i^2}\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,1}(t)\gamma_{i,r}^{(1)}(t) + 3\hat{\lambda}_{i,r,2}(t)I_{i,i}^{(3)}(t) + 3\hat{\lambda}_{i,r,4}\zeta_{i,r}^{(1)}(t)\right)\zeta_{i,j}^{(1)}(t)
\end{aligned}$$

The inner product between $\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)$ and $\bar{\mathbf{v}}_j^\star, \bar{\mathbf{w}}_j^\star$ can be written as

$$
\begin{aligned}
-\frac{1}{b_i}\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}_j^\star = \frac{9}{b_i^2}&\left(\hat{\lambda}_{i,j,5}(t) - \sum_{r=1}^{m}\lambda_{i,r,5}(t)\gamma_{r,j}^{(2)}(t)\right)\\
&- \frac{3}{b_i^2}\left(\sum_{r=1}^{m}\left(\lambda_{i,r,2}(t)\zeta_{i,j}^{(1)}(t) + \lambda_{i,r,3}(t)\zeta_{r,j}^{(1)}(t)\right) - \sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,2}(t)\zeta_{i,j}^{(1)}(t)\right)\\
&+ \frac{3}{b_i^2}\sum_{r=1}^{m}\left(\lambda_{i,r,2}(t)I_{i,i}^{(3)}(t) + \lambda_{i,r,3}(t)I_{r,i}^{(3)}(t) + 3\lambda_{i,r,5}(t)I_{r,i}^{(2)}(t)\right)\gamma_{i,j}^{(2)}(t)\\
&- \frac{3}{b_i^2}\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,2}(t)I_{i,i}^{(3)}(t) + \hat{\lambda}_{i,r,3}(t)\zeta_{i,r}^{(2)}(t) + 3\hat{\lambda}_{i,r,5}(t)\gamma_{i,r}^{(2)}(t)\right)\gamma_{i,j}^{(2)}(t)\\
-\frac{1}{b_i}\cdot\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_j^\star = \frac{3}{b_i^2}&\left(\hat{\lambda}_{i,j,3}(t) - \sum_{r=1}^{m}\lambda_{i,r,3}(t)\gamma_{r,j}^{(1)}(t)\right)\\
&- \frac{3}{b_i^2}\left(\sum_{r=1}^{m}\left(\lambda_{i,r,2}(t)\gamma_{i,j}^{(1)}(t) + 3\lambda_{i,r,5}(t)\zeta_{r,j}^{(2)}(t)\right) - \sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,2}(t)\gamma_{i,j}^{(1)}(t)\right)\\
&+ \frac{3}{b_i^2}\sum_{r=1}^{m}\left(\lambda_{i,r,2}(t)I_{i,i}^{(3)}(t) + \lambda_{i,r,3}(t)I_{r,i}^{(3)}(t) + 3\lambda_{i,r,5}(t)I_{r,i}^{(2)}(t)\right)\zeta_{i,j}^{(2)}(t)\\
&- \frac{3}{b_i^2}\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,2}(t)I_{i,i}^{(3)}(t) + \hat{\lambda}_{i,r,3}(t)\zeta_{i,r}^{(2)}(t) + 3\hat{\lambda}_{i,r,5}(t)\gamma_{i,r}^{(2)}(t)\right)\zeta_{i,j}^{(2)}(t)
\end{aligned}
$$

The inner product between $\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)$ and $\bar{\mathbf{v}}_j, \bar{\mathbf{w}}_j$ can be written as

$$
\begin{aligned}
-\frac{1}{a_i}\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_j = \frac{1}{a_i^2}&\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,1}(t)\gamma_{j,r}^{(1)}(t) + 3\hat{\lambda}_{i,r,2}(t)I_{j,i}^{(3)}(t) + 3\hat{\lambda}_{i,r,4}(t)\zeta_{j,r}^{(1)}(t)\right)\\
&- \frac{1}{a_i^2}\sum_{r=1}^{m}\left(\lambda_{i,r,1}(t)I_{r,j}^{(1)}(t) + 3\lambda_{i,r,2}(t)I_{j,i}^{(3)}(t) + 3\lambda_{i,r,4}(t)I_{j,r}^{(3)}(t)\right)\\
&+ \frac{1}{a_i^2}\sum_{r=1}^{m}\left(\lambda_{i,r,1}(t)I_{i,r}^{(1)}(t) + 3\lambda_{i,r,2}(t)I_{i,i}^{(3)}(t) + 3\lambda_{i,r,4}(t)I_{i,r}^{(3)}(t)\right)I_{i,j}^{(1)}(t)\\
&- \frac{1}{a_i^2}\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,1}(t)\gamma_{i,r}^{(1)}(t) + 3\hat{\lambda}_{i,r,2}(t)I_{i,i}^{(3)}(t) + 3\hat{\lambda}_{i,r,4}(t)\zeta_{i,r}^{(1)}(t)\right)I_{i,j}^{(1)}(t)\\
-\frac{1}{a_i}\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}_j = \frac{1}{a_i^2}&\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,1}(t)\zeta_{j,r}^{(2)} + 3\hat{\lambda}_{i,r,2}(t)I_{i,j}^{(2)} + 3\hat{\lambda}_{i,r,4}(t)\gamma_{j,r}^{(2)}(t)\right)\\
&- \frac{1}{a_i^2}\sum_{r=1}^{m}\left(\lambda_{i,r,1}(t)I_{r,j}^{(3)}(t) + 3\lambda_{i,r,2}(t)I_{i,j}^{(2)}(t) + 3\lambda_{i,r,4}(t)I_{r,j}^{(2)}(t)\right)\\
&+ \frac{1}{a_i^2}\sum_{r=1}^{m}\left(\lambda_{i,r,1}(t)I_{i,r}^{(1)}(t) + 3\lambda_{i,r,2}(t)I_{i,i}^{(3)}(t) + 3\lambda_{i,r,4}(t)I_{i,r}^{(3)}(t)\right)I_{i,j}^{(3)}(t)\\
&- \frac{1}{a_i^2}\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,1}(t)\gamma_{i,r}^{(1)}(t) + 3\hat{\lambda}_{i,r,2}(t)I_{i,i}^{(3)}(t) + 3\hat{\lambda}_{i,r,4}(t)\zeta_{i,r}^{(1)}(t)\right)I_{i,j}^{(3)}(t)
\end{aligned}
$$

Lastly, the inner product between $\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)$ and $\bar{\mathbf{v}}_j, \bar{\mathbf{w}}_j$ can be written as

$$
\begin{aligned}
-\frac{1}{b_i}\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_j =& \frac{3}{b_j^2}\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,2}(t)I_{i,j}^{(1)}(t) + \hat{\lambda}_{i,r,3}(t)\gamma_{j,r}^{(1)}(t) + 3\hat{\lambda}_{i,r,5}(t)\zeta_{j,r}^{(1)}(t)\right)\\
&- \frac{3}{b_j^2}\sum_{r=1}^{m}\left(\lambda_{i,r,2}(t)I_{i,j}^{(1)}(t) + \lambda_{i,r,3}(t)I_{j,r}^{(1)}(t) + 3\lambda_{i,r,5}(t)I_{j,r}^{(3)}(t)\right)\\
&+ \frac{3}{b_j^2}\sum_{r=1}^{m}\left(\lambda_{i,r,2}(t)I_{i,i}^{(3)}(t) + \lambda_{i,r,3}(t)I_{r,i}^{(3)}(t) + 3\lambda_{i,r,5}(t)I_{r,i}^{(2)}(t)\right)I_{i,j}^{(3)}(t)\\
&- \frac{3}{b_j^2}\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,2}(t)I_{i,i}^{(3)}(t) + \hat{\lambda}_{i,r,3}(t)\zeta_{i,r}^{(2)}(t) + 3\hat{\lambda}_{i,r,5}(t)\gamma_{i,r}^{(2)}(t)\right)I_{i,j}^{(3)}(t)\\
-\frac{1}{b_i}\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}_j =& \frac{3}{b_i^2}\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,2}(t)I_{i,j}^{(3)}(t) + \hat{\lambda}_{i,r,3}(t)\zeta_{j,r}^{(2)}(t) + 3\hat{\lambda}_{i,r,5}(t)\gamma_{j,r}^{(2)}(t)\right)\\
&- \frac{3}{b_i^2}\sum_{r=1}^{m}\left(\lambda_{i,r,2}(t)I_{i,j}^{(3)}(t) + \lambda_{i,r,3}(t)I_{r,j}^{(3)}(t) + 3\lambda_{i,r,5}(t)I_{r,j}^{(2)}(t)\right)\\
&+ \frac{3}{b_i^2}\sum_{r=1}^{m}\left(\lambda_{i,r,2}(t)I_{i,i}^{(3)}(t) + \lambda_{i,r,3}(t)I_{r,i}^{(3)}(t) + 3\lambda_{i,r,5}(t)I_{r,i}^{(2)}(t)\right)I_{i,j}^{(2)}(t)\\
&- \frac{3}{b_i^2}\sum_{r=1}^{m^\star}\left(\hat{\lambda}_{i,r,2}(t)I_{i,i}^{(3)}(t) + \hat{\lambda}_{i,r,3}(t)\zeta_{i,r}^{(2)}(t) + 3\hat{\lambda}_{i,r,5}(t)\gamma_{i,r}^{(2)}(t)\right)I_{i,j}^{(2)}(t)
\end{aligned}
$$

A.4 Approximating the Gradient Flow Dynamics

In order to understand the property of the GF induced dynamics given in the previous section, we need to first compute $\lambda_{i,j,\ell}$s and $\hat{\lambda}_{i,j,\ell}$s. The following lemma provides such property.

Lemma 5. *Fix $i \in [m], j \in [m^\star]$ and $t \geq 0$. If for any $\left|\zeta_{i,j}^{(1)}(t)\right|, \left|\zeta_{i,j}^{(1)}(t)\right|, \left|I_{i,i}^{(3)}(t)\right| \leq \delta_r$ for some $\delta_r > 0$, then we have that*

$$
\begin{aligned}
\hat{\lambda}_{i,j,1}(t) &= 6\sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}\gamma_{i,j}^{(1)}(t)^k\gamma_{i,j}^{(2)}(t)^3 \pm \mathcal{O}\left(\delta_r^2\right)\gamma_{i,j}^{(2)}(t)^2 \pm \mathcal{O}\left(\delta_r^4\right)\\
\hat{\lambda}_{i,j,2}(t) &= 6\sum_{k=1}^{\infty}\frac{c_{k+2}c_k}{k!}\gamma_{i,j}^{(1)}(t)^k\gamma_{i,j}^{(2)}(t)^2\zeta_{i,j}^{(1)}(t) \pm \mathcal{O}\left(\delta_r^3\right)\\
\hat{\lambda}_{i,j,3}(t) &= 6\sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}\gamma_{i,j}^{(1)}(t)^k\gamma_{i,j}^{(2)}(t)^2\zeta_{i,j}^{(1)}(t) \pm \mathcal{O}\left(\delta_r^3\right)\\
\hat{\lambda}_{i,j,4}(t) &= 6\sum_{k=1}^{\infty}\frac{c_{k+2}c_k}{k!}\gamma_{i,j}^{(1)}(t)^k\gamma_{i,j}^{(2)}(t)^2 I_{i,i}^{(3)}(t) + 6\sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}\gamma_{i,j}^{(1)}(t)^k\gamma_{i,j}^{(2)}(t)^2\zeta_{i,j}^{(2)}(t) \pm \mathcal{O}\left(\delta_r^3\right)\\
\hat{\lambda}_{i,j,5}(t) &= 2\sum_{k=0}^{\infty}\frac{c_k^2}{k!}\gamma_{i,j}^{(1)}(t)^k\gamma_{i,j}^{(2)}(t)^2 \pm \mathcal{O}\left(\delta_r^2\right)\gamma_{i,j}^{(2)}(t) \pm \mathcal{O}\left(\delta_r^4\right)
\end{aligned}
$$

Proof. We are going to Lemma 21 with $\mathbf{v}_1 = \bar{\mathbf{v}}_i(t), \mathbf{v}_2 = \bar{\mathbf{v}}_j^\star$ and $\mathbf{w}_1 = \bar{\mathbf{w}}_i(t), \mathbf{w}_2 = \bar{\mathbf{w}}_j^\star$. In this case, we have that $\mathbf{v}_1^\top\mathbf{v}_2 = \gamma_{i,j}^{(1)}(t)$ and $\mathbf{w}_1^\top\mathbf{w}_2 = \gamma_{i,j}^{(2)}(t)$. We start with $\hat{\lambda}_{i,j,1}(t)$. By definition, we have that

$$
\hat{\lambda}_{i,j,1}(t) = \sum_{k,\ell=0}^{\infty}\frac{c_{k+1}c_{\ell+1}}{k!\ell!}\mathbb{E}\left[He_k\left(\bar{\mathbf{v}}_i(t)^\top\mathbf{x}\right)He_\ell\left(\bar{\mathbf{v}}_j^{\star\top}\mathbf{x}\right)He_3\left(\bar{\mathbf{w}}_i(t)^\top\mathbf{x}\right)He_3\left(\bar{\mathbf{w}}_j^{\star\top}\mathbf{x}\right)\right]
$$

Here, invoking Lemma 21 with $h_k = c_{k+1}, h'_\ell = c_{\ell+1}$ gives

$$
\hat{\lambda}_{i,j,1}(t) = 6\sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}\gamma_{i,j}^{(1)}(t)^k\gamma_{i,j}^{(2)}(t)^3 \pm \mathcal{O}\left(\delta_r^2\right)\gamma_{i,j}^{(2)}(t)^2 \pm \mathcal{O}\left(\delta_r^4\right)
$$

For $\hat{\lambda}_{i,j,5}(t)$, by definition we have

$$\hat{\lambda}_{i,j,5}(t) = \sum_{k,\ell=0}^{\infty} \frac{c_k c_\ell}{k!\ell!} \mathbb{E}\left[He_k\left(\bar{\mathbf{v}}_i(t)^\top \mathbf{x}\right) He_\ell\left(\bar{\mathbf{v}}_j^{\star\top} \mathbf{x}\right) He_2\left(\bar{\mathbf{w}}_i(t)^\top \mathbf{x}\right) He_2\left(\bar{\mathbf{w}}_j^{\star\top} \mathbf{x}\right)\right]$$

Invoking Lemma 21 with $h_k = c_k, h'_\ell = c_\ell$ gives

$$\hat{\lambda}_{i,j,5}(t) = 2\sum_{k=0}^{\infty} \frac{c_k^2}{k!} \gamma_{i,j}^{(1)}(t)^k \gamma_{i,j}^{(2)}(t)^2 \pm \mathcal{O}\left(\delta_r^2\right) \gamma_{i,j}^{(2)}(t) \pm \mathcal{O}\left(\delta_r^4\right)$$

For $\hat{\lambda}_{i,j,2}(t)$, by definition we have

$$\hat{\lambda}_{i,j,2}(t) = \sum_{k,\ell=0}^{\infty} \frac{c_{k+1} c_\ell}{k!\ell!} \mathbb{E}\left[He_k\left(\bar{\mathbf{v}}_i(t)^\top \mathbf{x}\right) He_\ell\left(\bar{\mathbf{v}}_j^{\star\top} \mathbf{x}\right) He_2\left(\bar{\mathbf{w}}_i(t)^\top \mathbf{x}\right) He_3\left(\bar{\mathbf{w}}_j^{\star\top} \mathbf{x}\right)\right]$$

Here, invoking Lemma 21 with $h_k = c_{k+1}, h'_\ell = c_\ell$, and noticing that $\bar{\mathbf{v}}_j^{\star\top} \bar{\mathbf{w}}_j = 0$, gives

$$\hat{\lambda}_{i,j,2}(t) = 6\sum_{k=0}^{\infty} \frac{c_{k+2} c_k}{k!} \gamma_{i,j}^{(1)}(t)^k \gamma_{i,j}^{(2)}(t)^2 \zeta_{i,j}^{(1)}(t) \pm \mathcal{O}\left(\delta_r^3\right)$$

Noticing that $c_2 = 0$ gives that

$$\hat{\lambda}_{i,j,2}(t) = 6\sum_{k=1}^{\infty} \frac{c_{k+2} c_k}{k!} \gamma_{i,j}^{(1)}(t)^k \gamma_{i,j}^{(2)}(t)^2 \zeta_{i,j}^{(1)}(t) \pm \mathcal{O}\left(\delta_r^3\right)$$

For $\hat{\lambda}_{i,j,3}(t)$, by definition, we have that

$$\hat{\lambda}_{i,j,3}(t) = \sum_{k,\ell=0}^{\infty} \frac{c_k c_{\ell+1}}{k!\ell!} \mathbb{E}\left[He_k\left(\bar{\mathbf{v}}_i(t)^\top \mathbf{x}\right) He_\ell\left(\bar{\mathbf{v}}_j^{\star\top} \mathbf{x}\right) He_2\left(\bar{\mathbf{w}}_i(t)^\top \mathbf{x}\right) He_3\left(\bar{\mathbf{w}}_j^{\star\top} \mathbf{x}\right)\right]$$

Invoking Lemma 21 with $h_k = c_k, h'_\ell = c_{\ell+1}$, and noticing that $\bar{\mathbf{v}}_j^{\star\top} \bar{\mathbf{w}}_j = 0$, gives

$$\hat{\lambda}_{i,j,3}(t) = 6\sum_{k=0}^{\infty} \frac{c_{k+1}^2}{k!} \gamma_{i,j}^{(1)}(t)^k \gamma_{i,j}^{(2)}(t)^2 \zeta_{i,j}^{(1)}(t) \pm \mathcal{O}\left(\delta_r^3\right)$$

Lastly, for $\hat{\lambda}_4$, by definition we have

$$\hat{\lambda}_{i,j,4}(t) = \sum_{k,\ell=0}^{\infty} \frac{c_{k+1} c_\ell}{k!\ell!} \mathbb{E}\left[He_k\left(\bar{\mathbf{v}}_i(t)^\top \mathbf{x}\right) He_\ell\left(\bar{\mathbf{v}}_j^{\star\top} \mathbf{x}\right) He_3\left(\bar{\mathbf{w}}_i(t)^\top \mathbf{x}\right) He_2\left(\bar{\mathbf{w}}_j^{\star\top} \mathbf{x}\right)\right]$$

Therefore, we need to consider $\mathbf{v}_1 = \bar{\mathbf{v}}_i(t), \mathbf{v}_2 = \bar{\mathbf{v}}_j^\star$ and $\mathbf{w}_1 = \bar{\mathbf{w}}_j^\star, \mathbf{w}_2 = \bar{\mathbf{w}}_i(t)$. Moreover, we need to set $h_k = c_{k+1}$ and $h'_\ell = c_\ell$. In this case, $\mathbf{v}_1^\top \mathbf{w}_2 = I_{i,i}^{(3)}(t)$ and $\mathbf{v}_2^\top \mathbf{w}_2 = \zeta_{i,j}^{(2)}(t)$. Therefore

$$\hat{\lambda}_{i,j,4}(t) = 6\sum_{k=0}^{\infty} \frac{c_{k+2} c_k}{k!} \gamma_{i,j}^{(1)}(t)^k \gamma_{i,j}^{(2)}(t)^2 I_{i,i}^{(3)}(t) + 6\sum_{k=0}^{\infty} \frac{c_{k+1}^2}{k!} \gamma_{i,j}^{(1)}(t)^k \gamma_{i,j}^{(2)}(t)^2 \zeta_{i,j}^{(2)}(t) \pm \mathcal{O}\left(\delta_r^3\right)$$

Noticing that $c_2 = 0$ gives that

$$\hat{\lambda}_{i,j,4}(t) = 6\sum_{k=1}^{\infty} \frac{c_{k+2} c_k}{k!} \gamma_{i,j}^{(1)}(t)^k \gamma_{i,j}^{(2)}(t)^2 I_{i,i}^{(3)}(t) + 6\sum_{k=0}^{\infty} \frac{c_{k+1}^2}{k!} \gamma_{i,j}^{(1)}(t)^k \gamma_{i,j}^{(2)}(t)^2 \zeta_{i,j}^{(2)}(t) \pm \mathcal{O}\left(\delta_r^3\right)$$

□

Lemma 6. *Fix $i, j \in [m]$ and $t \geq 0$. If for any $\left|I_{i,j}^{(1)}(t)\right|, \left|I_{i,j}^{(2)}(t)\right|, \left|I_{i,j}^{(3)}(t)\right| \leq \delta_r$ when $i \neq j$, and $\left|I_{i,i}^{(3)}(t)\right| \leq \delta_p \leq \mathcal{O}(\delta_r)$, then we have that*

$$\lambda_{i,j,1}(t) = \begin{cases} \pm\mathcal{O}\left(\delta_r^3\right) & \text{if } i \neq j \\ 6\sum_{k=0}^{\infty} \frac{c_{k+1}^2}{k!} \pm \mathcal{O}\left(\delta_p^2\right) & \text{if } i = j \end{cases}$$

$$\lambda_{i,j,2}(t) = \begin{cases} \pm\mathcal{O}\left(\delta_r^3\right) & \text{if } i \neq j \\ 6C_{S,2} I_{i,i}^{(3)}(t) \pm \mathcal{O}\left(\delta_p^3\right) & \text{if } i = j \end{cases}$$

$$\lambda_{i,j,3}(t) = \begin{cases} \pm\mathcal{O}\left(\delta_r^3\right) & \text{if } i \neq j \\ 6C_{S,2} I_{i,i}^{(3)}(t) \pm \mathcal{O}\left(\delta_p^3\right) & \text{if } i = j \end{cases}$$

$$\lambda_{i,j,4}(t) = \begin{cases} \pm\mathcal{O}\left(\delta_r^3\right) & \text{if } i \neq j \\ 6C_{S,2} I_{i,i}^{(3)}(t) \pm \mathcal{O}\left(\delta_p^3\right) & \text{if } i = j \end{cases}$$

$$\lambda_{i,j,5}(t) = \begin{cases} \pm\mathcal{O}\left(\delta_r^2\right) & \text{if } i \neq j \\ 2\sum_{k=0}^{\infty} \frac{c_k^2}{k!} \pm \mathcal{O}\left(\delta_p^2\right) & \text{if } i = j \end{cases}$$

Here $C_{S,2} = \sum_{k=0}^{\infty} \frac{c_{k+1}^2 + c_k c_{k+2}}{k!}$.

Proof. We are going to use Lemma 21 with $\mathbf{v}_1 = \bar{\mathbf{v}}_i(t), \mathbf{v}_2 = \bar{\mathbf{v}}_j(t)$ and $\mathbf{w}_1 = \bar{\mathbf{w}}_i(t), \mathbf{w}_2 = \bar{\mathbf{w}}_j(t)$. In this case, we have that $\mathbf{v}_1^\top \mathbf{v}_2 = I_{i,j}^{(1)}(t)$, $\mathbf{w}_1^\top \mathbf{w}_2 = I_{i,j}^{(2}(t)$, and $\mathbf{v}_1^\top \mathbf{w}_2 = I_{i,j}^{(3)}(t), \mathbf{v}_2^\top \mathbf{w}_1 = I_{j,i}^{(3)}(t)$. Now, for $\lambda_{i,j,1}$, by definition, we have that

$$\lambda_{i,j,1}(t) = \sum_{k,\ell=0}^{\infty} \frac{c_{k+1} c_{\ell+1}}{k!\ell!} \mathbb{E}\left[He_k\left(\bar{\mathbf{v}}_i(t)^\top \mathbf{x}\right) He_\ell\left(\bar{\mathbf{v}}_j(t)^\top \mathbf{x}\right) He_3\left(\bar{\mathbf{w}}_i(t)^\top \mathbf{x}\right) He_3\left(\bar{\mathbf{w}}_j(t)^\top \mathbf{x}\right)\right]$$

Invoking Lemma 21 with $h_k = c_{k+1}, h'_\ell = c_{\ell+1}$ gives

$$\lambda_{i,j,1}(t) = 6\sum_{k=0}^{\infty} \frac{c_{k+1}^2}{k!} I_{i,j}^{(1)}(t)^k I_{i,j}^{(2)}(t)^3 \pm \mathcal{O}(\delta_r) I_{i,j}^{(2)}(t)^2 \pm \mathcal{O}\left(\delta_r^4\right)$$

Since $\left|I_{i,j}^{(2)}\right| \leq \delta_r$, we have that

$$\lambda_{i,j,1}(t) = \pm\mathcal{O}\left(\delta_r^3\right)$$

In the special case where $i = j$, we have that $I_{i,j}^{(1)} = I_{i,j}^{(2)} = 1$, and $\mathbf{v}_1^\top \mathbf{w}_1 = \mathbf{v}_1^\top \mathbf{w}_2 = \mathbf{v}_2^\top \mathbf{w}_1 = \mathbf{v}_2^\top \mathbf{w}_2 = I_{i,i}^{(3)}$. Therefore

$$\lambda_{i,j,1}(t) = 6\sum_{k=0}^{\infty} \frac{c_{k+1}^2}{k!} \pm \mathcal{O}\left(\delta_p^2\right)$$

For $\lambda_{i,j,5}$, by definition, we have that

$$\lambda_{i,j,5}(t) = \sum_{k,\ell=0}^{\infty} \frac{c_k c_\ell}{k!\ell!} \mathbb{E}\left[He_k\left(\bar{\mathbf{v}}_i(t)^\top \mathbf{x}\right) He_\ell\left(\bar{\mathbf{v}}_j(t)^\top \mathbf{x}\right) He_2\left(\bar{\mathbf{w}}_i(t)^\top \mathbf{x}\right) He_2\left(\bar{\mathbf{w}}_j(t)^\top \mathbf{x}\right)\right]$$

Invoking Lemma 21 with $h_k = c_k, h'_\ell = c_k$ gives

$$\lambda_{i,j,5}(t) = 2\sum_{k=0}^{\infty} \frac{c_k^2}{k!} I_{i,j}^{(1)}(t)^k I_{i,j}^{(2)}(t)^2 \pm \mathcal{O}\left(\delta_r^2\right) = \pm\mathcal{O}\left(\delta_r^2\right)$$

In the case where $i = j$, we have that

$$\lambda_{i,j,5}(t) = 2\sum_{k=0}^{\infty} \frac{c_k^2}{k!} \pm \mathcal{O}\left(\delta_p^2\right)$$

For $\lambda_{i,j,2}$, by definition, we have that

$$\lambda_{i,j,2}(t) = \sum_{k,\ell=0}^{\infty} \frac{c_{k+1}c_\ell}{k!\ell!}\mathbb{E}\big[He_k\left(\bar{\mathbf{v}}_i(t)^\top \mathbf{x}\right) He_\ell\left(\bar{\mathbf{v}}_j(t)^\top \mathbf{x}\right) He_2\left(\bar{\mathbf{w}}_i(t)^\top \mathbf{x}\right) He_3\left(\bar{\mathbf{w}}_j(t)^\top \mathbf{x}\right)\big]$$

Invoking Lemma 21 with $h_k = c_{k+1}, h'_\ell = c_\ell$ gives

$$\lambda_{i,j,2}(t) = 6\sum_{k=0}^{\infty} \frac{c_{k+2}c_k}{k!} I_{i,j}^{(1)}(t)^k I_{i,j}^{(2)}(t)^2 I_{i,j}^{(3)}(t) + 6\sum_{k=0}^{\infty} \frac{c_{k+1}^2}{k!} I_{i,j}^{(1)}(t)^k I_{i,j}^{(2)}(t)^2 I_{j,j}^{(3)}(t) \pm \mathcal{O}\left(\delta_r^3\right) = \pm\mathcal{O}\left(\delta_r^3\right)$$

In the case where $i = j$, we have that $I_{i,j}^{(1)}(t) = I_{i,j}^{(2)}(t) = 1$. Therefore

$$\lambda_{i,j,2}(t) = 6I_{i,i}^{(3)}(t)\sum_{k=0}^{\infty} \frac{c_k c_{k+2} + c_{k+1}^2}{k!} \pm \mathcal{O}\left(\delta_p^3\right)$$

For $\lambda_{i,j,3}$, by definition, we have that

$$\lambda_{i,j,3}(t) = \sum_{k,\ell=0}^{\infty} \frac{c_k c_{\ell+1}}{k!\ell!}\mathbb{E}\big[He_k\left(\bar{\mathbf{v}}_i(t)^\top \mathbf{x}\right) He_\ell\left(\bar{\mathbf{v}}_j(t)^\top \mathbf{x}\right) He_2\left(\bar{\mathbf{w}}_i(t)^\top \mathbf{x}\right) He_3\left(\bar{\mathbf{w}}_j(t)^\top \mathbf{x}\right)\big]$$

Invoking Lemma 21 with $h_k = c_k, h'_\ell = c_{\ell+1}$ gives

$$\lambda_{i,j,3}(t) = 6\sum_{k=0}^{\infty} \frac{c_{k+1}^2}{k!} I_{i,j}^{(1)}(t)^k I_{i,j}^{(2)}(t)^2 I_{i,j}^{(3)}(t) + 6\sum_{k=0}^{\infty} \frac{c_{k+2}c_k}{k!} I_{i,j}^{(1)}(t)^k I_{i,j}^{(2)}(t)^2 I_{j,j}^{(3)}(t) \pm \mathcal{O}\left(\delta_r^3\right) = \pm\mathcal{O}\left(\delta_r^3\right)$$

In the case where $i = j$, we have that $I_{i,j}^{(1)}(t) = I_{i,j}^{(2)}(t) = 1$. Therefore

$$\lambda_{i,j,3}(t) = 6I_{i,i}^{(3)}(t)\sum_{k=0}^{\infty} \frac{c_k c_{k+2} + c_{k+1}^2}{k!} \pm \mathcal{O}\left(\delta_p^3\right)$$

Lastly, for $\lambda_{i,j,4}$, we have that

$$\lambda_{i,j,4}(t) = \sum_{k,\ell=0}^{\infty} \frac{c_{k+1}c_\ell}{k!\ell!}\mathbb{E}\big[He_k\left(\bar{\mathbf{v}}_i(t)^\top \mathbf{x}\right) He_\ell\left(\bar{\mathbf{v}}_j(t)^\top \mathbf{x}\right) He_3\left(\bar{\mathbf{w}}_i(t)^\top \mathbf{x}\right) He_2\left(\bar{\mathbf{w}}_j(t)^\top \mathbf{x}\right)\big]$$

Here we need to apply Lemma 21 with $\mathbf{v}_1 = \bar{\mathbf{v}}_i(t), \mathbf{v}_2 = \bar{\mathbf{v}}_j(t), \mathbf{w}_1 = \bar{\mathbf{w}}_j(t), \mathbf{w}_2 = \bar{\mathbf{w}}_i(t)$ and $h_k = c_{k+1}, h'_\ell = c_\ell$. This gives that

$$\lambda_{i,j,4}(t) = 6\sum_{k=0}^{\infty} \frac{c_{k+2}c_k}{k!} I_{i,j}^{(1)}(t)^k I_{i,j}^{(2)}(t)^2 I_{i,i}^{(1)}(t) + 6\sum_{k=0}^{\infty} \frac{c_{k+1}^2}{k!} I_{i,j}^{(1)}(t)^k I_{i,j}^{(2)}(t)^2 I_{j,i}^{(3)}(t) \pm \mathcal{O}\left(\delta_r^3\right) = \pm\mathcal{O}\left(\delta_r^3\right)$$

In the case where $i = j$, we have that $I_{i,j}^{(1)}(t) = I_{i,j}^{(2)}(t) = 1$. Therefore

$$\lambda_{i,j,4}(t) = 6I_{i,i}^{(3)}(t)\sum_{k=0}^{\infty} \frac{c_k c_{k+2} + c_{k+1}^2}{k!} \pm \mathcal{O}\left(\delta_p^3\right)$$

□

With the above lemmas that studies $\lambda_{i,j,\ell}$s and $\hat{\lambda}_{i,j,\ell}$s, we are ready to analyze the dynamics of $\gamma_{i,j}^{(1)}, \gamma_{i,j}^{(2)}, \zeta_{i,j}^{(1)}, \zeta_{i,j}^{(2)}$, and $I_{i,j}^{(1)}, I_{i,j}^{(2)}, I_{i,j}^{(3)}$. In particular, we will fix any $\ell \in [m^\star]$, and assumes the inductive hypothesis.

Lemma 7. *Let $\varepsilon_{\mathcal{A},\ell}^{(2)}(t), \varepsilon_{\mathcal{B},\ell}^{(3)}(t)$, and $\varepsilon_{5,\ell}(t)$ be defined in Definition 1 with $\varepsilon_{5,\ell}(t) \leq \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)$. Then the gradient alignment $\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{v}}_j^\star$ and $\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{w}}_j^\star$ satisfies*

$$\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{v}}_j^\star = -\frac{1}{a_i}\cdot\begin{cases}\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\\ -\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\gamma_{i,j}^{(1)}(t) \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 & \\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i=i_\ell^\star, j\in[m^\star]\setminus\mathcal{C}_\ell\\ \left(1-\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)^2\right)\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t) \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 & \\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i=i_\ell^\star, j=j_\ell^\star\end{cases}$$

$$\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{w}}_j^\star = -\frac{9}{b_i}\cdot\begin{cases}\hat{\lambda}_{i,j,5}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \\ \quad\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\\ \hat{\lambda}_{i,j,5}(t) - \lambda_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\gamma_{i,j}^{(2)}(t) & \\ \quad\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) & \\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t) + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i=i_\ell^\star, j\in[m^\star]\setminus\mathcal{C}_\ell\\ \left(1-\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\right)\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t) \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) & \\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t) + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i=i_\ell^\star, j=j_\ell^\star\end{cases}$$

and in particular, for the case $\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{w}}_{j_{\ell'}^\star}^\star$, we have that

$$\begin{aligned}\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{w}}_{j_{\ell'}^\star}^\star = {}& \frac{9}{b_i}\lambda_{i,i_{\ell'}^\star,5}(t)\gamma_{i_{\ell'}^\star,j_{\ell'}^\star}^{(2)}(t)\\ & -\frac{9}{b_i}\cdot\begin{cases}\hat{\lambda}_{i,j,5}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\\ \hat{\lambda}_{i,j,5}(t) - \lambda_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\gamma_{i,j}^{(2)}(t) & \\ \quad\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 & \text{if } i=i_\ell^\star\\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \end{cases}\end{aligned}$$

Further more, the mis-alignment terms $\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{w}}_j^\star$ and $\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{v}}_j^\star$ satisfies

$$\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{w}}_j^\star = -\frac{1}{a_i}\cdot\begin{cases}\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\\ \hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i,j}^{(2)}(t) - \hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\zeta_{i,j}^{(1)}(t) & \\ \quad -36C_{S,2}\gamma_{i,j}^{(2)}(t)I_{i,i}^{(3)}(t) & \text{if } i=i_\ell^\star, j\neq j_\ell^\star\\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t) + \varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2\right) & \\ 3\hat{\lambda}_{i_\ell^\star,j_\ell^\star,4}(t) + \hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) - \hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\zeta_{i_\ell^\star,j_\ell^\star}^{(1)}(t) & \\ \quad -36C_{S,2}\gamma_{i,j}^{(2)}(t)I_{i,i}^{(3)}(t) & \text{if } i=i_\ell^\star, j=j_\ell^\star\\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t) + \varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2\right) & \end{cases}$$

$$\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{v}}_j^\star = \frac{3}{b_i}\cdot\begin{cases}\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\\ \hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i,j}^{(1)}(t) - 3\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\zeta_{i,j}^{(2)}(t) & \\ \quad -12C_{S,2}\gamma_{i,j}^{(2)}(t)I_{i,i}^{(3)}(t) & \text{if } i=i_\ell^\star, j\neq j_\ell^\star\\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t) + \varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2\right) & \\ \hat{\lambda}_{i_\ell^\star,j_\ell^\star,3}(t) + \hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t) - 3\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\zeta_{i_\ell^\star,j_\ell^\star}^{(2)}(t) & \\ \quad -12C_{S,2}\gamma_{i,j}^{(2)}(t)I_{i,i}^{(3)}(t) & \text{if } i=i_\ell^\star, j=j_\ell^\star\\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t) + \varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2\right) & \end{cases}$$

The self-alignments $\nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta}(t))\,\bar{\mathbf{v}}_j, \nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta}(t))\,\bar{\mathbf{w}}_j$, in the case of $i \neq j$, are given by

$$\nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top \bar{\mathbf{v}}_j = -\frac{1}{a_i}\cdot\begin{cases}\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\\ \hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{j,j_\ell^\star}^{(1)}(t) - \hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)I_{i,j}^{(1)}(t) & \\ \quad\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 & \text{if } i = i_\ell^\star\\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \end{cases}$$

$$\nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top \bar{\mathbf{w}}_j = -\frac{9}{b_i}\cdot\begin{cases}\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i,j\in[m]\setminus\mathcal{R}_\ell\\ \hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{j,j_\ell^\star}^{(2)}(t) - \hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)I_{i,j}^{(2)}(t) & \\ \quad\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 & \text{if } i = i_\ell^\star\\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \\ \pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) & \\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } j = i_\ell^\star\end{cases}$$

The self-alignments $\nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta}(t))\,\bar{\mathbf{w}}_j, \nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta}(t))\,\bar{\mathbf{v}}_j$, in the case of $i \neq j$, are given by

$$\nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top \bar{\mathbf{w}}_j = -\frac{1}{a_i}\cdot\begin{cases}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\left(\zeta_{j,j_\ell^\star}^{(2)}(t) - \gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)I_{i,j}^{(3)}(t)\right) & \\ \quad\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 & \text{if } i = i_\ell^\star\\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \\ \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\end{cases}$$

$$\nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top \bar{\mathbf{v}}_j = -\frac{9}{b_i}\cdot\begin{cases}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\left(\zeta_{j,j_\ell^\star}^{(1)}(t) - \gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)I_{i,j}^{(3)}(t)\right) & \\ \quad\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 & \text{if } i = i_\ell^\star\\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \\ \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\end{cases}$$

Lastly, the self-alignments $\nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta}(t))\,\bar{\mathbf{w}}_j, \nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta}(t))\,\bar{\mathbf{v}}_j$, in the case of $i = j \in [m]\setminus\mathcal{R}_{\ell-1}$, are given by

$$\begin{aligned}\nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top \bar{\mathbf{w}}_j = &-\frac{1}{a_i}\left(\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\zeta_{j,j_\ell^\star}^{(2)}(t) + 3\hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t) + 3\hat{\lambda}_{i_\ell^\star,j_\ell^\star,4}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\right)\mathbb{I}\{i = i_\ell^\star\}\\ &-\frac{1}{a_i}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\mathbb{I}\{i = i_\ell^\star\} - \frac{36}{a_i}C_{S,2}I_{i,i}^{(3)}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\\ \nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top \bar{\mathbf{v}}_j = &-\frac{3}{b_i}\cdot\left(3\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\zeta_{j,j_\ell^\star}^{(1)}(t) + \hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t) + \hat{\lambda}_{i_\ell^\star,j_\ell^\star,3}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\right)\mathbb{I}\{i = i_\ell^\star\}\\ &-\frac{9}{b_i}\cdot\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\mathbb{I}\{i = i_\ell^\star\} - 36C_{S,2}I_{i,i}^{(3)}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\end{aligned}$$

Proof. In the following of the proof we will assume that $i \in [m]\setminus\mathcal{R}_{\ell-1}$ and $j \in [m^\star]$. For $\hat{\lambda}_{i,j,1},\dots\hat{\lambda}_{i,j,5}$, we apply Lemma 5 with

$$\delta_r = \varepsilon_{\mathcal{B},\ell}^{(1)}(t);\ \text{for } i\in[m]\setminus\mathcal{R}_{\ell-1}, j\in[m^\star];$$

For $\lambda_{i,j,1},\dots,\lambda_{i,j,5}$, we apply

$$\delta_r = \begin{cases}\varepsilon_{\mathcal{B},\ell}^{(1)}(t);\ \text{if } i,j\in[m]\setminus\mathcal{R}_{\ell-1}, i\neq j\\ \varepsilon_{\mathcal{F},\ell}(t);\ \text{if } i\in\mathcal{R}_{\ell-1}\vee j\in\mathcal{R}_{\ell-1}, i\neq j\end{cases} \leq \varepsilon_{\mathcal{A},\ell}^{(1)}(t);\quad \delta_p = \begin{cases}\varepsilon_{\mathcal{B},\ell}^{(3)}(t) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\\ \varepsilon_{5,\ell}(t) & \text{if } i = i_\ell^\star\end{cases}$$

For $\hat{\lambda}_{i_{\ell'}^\star,j,1},\dots,\hat{\lambda}_{i_{\ell'}^\star,j,5}$ for $\ell' < \ell$, we have that

$$\delta_r = \varepsilon_{\mathcal{F},\ell}(t) \leq \varepsilon_{\mathcal{A},\ell}^{(1)}(t);$$

For $\lambda_{i^\star_{\ell'},i^\star_{\ell'},1},\ldots,\lambda_{i^\star_{\ell'},i^\star_{\ell'},5}$, we have that

$$\delta_r = \delta_p = \varepsilon_{\mathcal{F},\ell}(t) \leq \varepsilon^{(1)}_{\mathcal{A},\ell}(t)$$

Moreover, we also have that

$$\left|\gamma^{(1)}_{i,j}(t)\right|, \left|\gamma^{(2)}_{i,j}(t)\right| \leq \begin{cases} \varepsilon^{(2)}_{\mathcal{A},\ell}(t) & \text{if } i \neq i^\star_\ell \\ \varepsilon^{(1)}_{\mathcal{A},\ell}(t) & \text{if } i = i^\star_\ell \end{cases} \forall j \in [m^\star], (i,j) \neq (i^\star_{\ell'}, j^\star_{\ell'}) \forall \ell' \leq \ell$$

$$\left|\zeta^{(1)}_{i,j}(t)\right|, \left|\zeta^{(2)}_{i,j}(t)\right| \leq \varepsilon^{(1)}_{\mathcal{A},\ell}(t); \forall i \in [m], j \in [m^\star]$$

$$\left|I^{(1)}_{i,j}(t)\right|, \left|I^{(2)}_{i,j}(t)\right|, \left|I^{(3)}_{i,j}(t)\right| \leq \varepsilon^{(1)}_{\mathcal{A},\ell}(t); \forall i,j \in [m], i \neq j$$

$$\left|I^{(3)}_{i,i}(t)\right| \leq \begin{cases} \varepsilon_{\mathcal{F},\ell}(t) & \text{if } i \in \mathcal{R}_{\ell-1} \\ \varepsilon_{5,\ell}(t) & \text{if } i = i^\star_\ell \\ \varepsilon^{(3)}_{\mathcal{B},\ell}(t) & \text{if } i \in [m] \setminus \mathcal{R}_\ell \end{cases}$$

This gives that for all $j \in [m^\star]$ and such that $(i,j) \neq (i^\star_\ell, j^\star_\ell)$

$$\hat{\lambda}_{i,j,1}(t) \leq \begin{cases} \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right) & \text{if } i \neq i^\star_\ell \\ \mathcal{O}\left(\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right) & \text{if } i = i^\star_\ell \end{cases}; \hat{\lambda}_{i,j,5}(t) \leq \begin{cases} \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^2\right) & \text{if } i \neq i^\star_\ell \\ \mathcal{O}\left(\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\right) & \text{if } i = i^\star_\ell \end{cases}$$

We will analyze each dynamic separately. However, we should notice some common terms that appears in the dynamics.

Common terms. To start, let's tackle some common terms in the dynamics we are interest in. In particular, we have that for $i \in [m] \setminus \mathcal{R}_{\ell-1}$,

$$\sum_{r=1}^m \lambda_{i,r,1}(t) I^{(1)}_{i,r}(t) = \lambda_{i,i,1}(t) \pm \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right)$$

$$\sum_{r=1}^m \lambda_{i,r,2}(t) I^{(3)}_{i,i}(t) = \pm \begin{cases} \mathcal{O}\left(\varepsilon^{(3)}_{\mathcal{B},\ell}(t)^2 + m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right) & \text{if } i \in [m] \setminus \mathcal{R}_\ell \\ \mathcal{O}\left(\varepsilon_{5,\ell}(t)^2 + m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right) & \text{if } i = i^\star_\ell \end{cases}$$

$$\sum_{r=1}^m \lambda_{i,r,4}(t) I^{(3)}_{i,r}(t) = \pm \begin{cases} \mathcal{O}\left(\varepsilon^{(3)}_{\mathcal{B},\ell}(t)^2 + m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right) & \text{if } i \in [m] \setminus \mathcal{R}_\ell \\ \mathcal{O}\left(\varepsilon_{5,\ell}(t)^2 + m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right) & \text{if } i = i^\star_\ell \end{cases}$$

$$\sum_{r=1}^{m^\star} \hat{\lambda}_{i,r,1}(t) \gamma^{(t)}_{i,r}(t) = \begin{cases} \lambda_{i^\star_\ell, j^\star_\ell, 1}(t) \gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t) \pm \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right) & \text{if } i = i^\star_\ell \\ \pm\mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right) & \text{if } i \in [m] \setminus \mathcal{R}_\ell \end{cases}$$

$$\sum_{r=1}^{m^\star} \hat{\lambda}_{i,r,2}(t) I^{(3)}_{i,i}(t) = \begin{cases} \pm\mathcal{O}\left(\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\right) \gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)^2 \pm \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right) & \text{if } i = i^\star_\ell \\ \pm\mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2 \varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right) & \text{if } i \in [m] \setminus \mathcal{R}_\ell \end{cases}$$

$$\sum_{r=1}^{m^\star} \hat{\lambda}_{i,r,4}(t) \zeta^{(1)}_{i,r,}(t) = \begin{cases} \pm\mathcal{O}\left(\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\right) \gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)^2 \pm \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right) & \text{if } i = i^\star_\ell \\ \pm\mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2 \varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right) & \text{if } i \in [m] \setminus \mathcal{R}_\ell \end{cases}$$

Therefore, we have that

$$\begin{aligned}
&\sum_{r=1}^{m}\lambda_{i,r,1}(t)I_{i,r}^{(1)}(t)+3\sum_{r=1}^{m}\lambda_{i,r,2}(t)I_{i,i}^{(3)}(t)+3\sum_{r=1}^{m}\lambda_{i,r,4}(t)I_{i,r}^{(3)}(t)\\
&\qquad=\lambda_{i,i,1}\pm\begin{cases}\mathcal{O}\left(\varepsilon_{5,\ell}(t)^2+m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i=i_\ell^\star\\ \mathcal{O}\left(\varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2+m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\end{cases}\\
&\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,1}(t)\gamma_{i,r}^{(t)}(t)+3\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,2}(t)I_{i,i}^{(3)}(t)+\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,4}(t)\zeta_{i,r,}^{(1)}(t)\\
&\qquad=\begin{cases}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i=i_\ell^\star\\ \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\end{cases}
\end{aligned}\tag{12}$$

Moreover, we can also compute that

$$\begin{aligned}
&\sum_{r=1}^{m}\lambda_{i,r,2}(t)I_{i,i}^{(3)}(t)=\pm\begin{cases}\mathcal{O}\left(\varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2+m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\\ \mathcal{O}\left(\varepsilon_{5,\ell}(t)^2+m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i=i_\ell^\star\end{cases}\\
&\sum_{r=1}^{m}\lambda_{i,r,3}(t)I_{r,i}^{(3)}(t)=\begin{cases}\mathcal{O}\left(\varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2+m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\\ \mathcal{O}\left(\varepsilon_{5,\ell}(t)^2+m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i=i_\ell^\star\end{cases}\\
&\sum_{r=1}^{m}\lambda_{i,r,5}(t)I_{r,i}^{(2)}(t)=\lambda_{i,i,5}(t)\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)\\
&\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,2}(t)I_{i,i}^{(3)}(t)=\begin{cases}\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i=i_\ell^\star\\ \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\end{cases}\\
&\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,3}(t)\zeta_{i,r}^{(2)}(t)=\begin{cases}\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i=i_\ell^\star\\ \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\end{cases}\\
&\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,5}(t)\gamma_{i,r}^{(2)}(t)=\begin{cases}\lambda_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i=i_\ell^\star\\ \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\end{cases}
\end{aligned}$$

This gives that

$$\begin{aligned}
&\sum_{r=1}^{m}\lambda_{i,r,2}(t)I_{i,i}^{(3)}(t)+\sum_{r=1}^{m}\lambda_{i,r,3}(t)I_{r,i}^{(3)}(t)+3\sum_{r=1}^{m}\lambda_{i,r,5}(t)I_{r,i}^{(2)}(t)\\
&\qquad=\lambda_{i,i,5}(t)\pm\begin{cases}\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3+\varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\\ \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3+\varepsilon_{5,\ell}(t)^2\right) & \text{if } i=i_\ell^\star\end{cases}\\
&\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,2}(t)I_{i,i}^{(3)}(t)+\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,3}(t)\zeta_{i,r}^{(2)}(t)+\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,5}(t)\gamma_{i,r}^{(2)}(t)\\
&\qquad=\begin{cases}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i=i_\ell^\star\\ \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\end{cases}
\end{aligned}\tag{13}$$

Now we are ready to analyze the dynamics.

Analysis of $\nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta}(t))\,\bar{\mathbf{v}}_j^\star$. To analyze $\nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta}(t))\,\bar{\mathbf{v}}_j^\star$, we first compute the following quantities for $i\in[m]\setminus$

$\mathcal{R}_{\ell-1}$

$$\hat{\lambda}_{i,j,1}(t) = \begin{cases} \pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \\ \pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i = i_\ell^\star, j \neq j_\ell^\star \end{cases};$$

$$\sum_{r=1}^m \lambda_{i,r,1}(t)\gamma_{r,j}^{(1)}(t) = \lambda_{i,i,1}(t)\gamma_{i,j}^{(1)}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$

$$\sum_{r=1}^m \lambda_{i,r,2}(t)\zeta_{i,j}^{(2)}(t) = \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) \pm \begin{cases} \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \\ \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i = i_\ell^\star \end{cases}$$

$$\sum_{r=1}^m \lambda_{i,r,4}(t)\zeta_{r,j}^{(2)}(t) = \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) \pm \begin{cases} \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \\ \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i = i_\ell^\star \end{cases}$$

$$\sum_{r=1}^{m^\star} \hat{\lambda}_{i,r,2}(t)\zeta_{i,j}^{(2)}(t) = \begin{cases} \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i = i_\ell^\star \\ \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \end{cases}$$

Therefore, combining with (12), and noticing that the term $\lambda_{i,i,1}(t)\gamma_{i,j}^{(1)}(t)$ cancels out, we have that

$$\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{v}}_j^\star = -\frac{1}{a_i}\cdot\begin{cases} \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \\ -\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\gamma_{i,j}^{(1)}(t) \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 & \\ \qquad \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i = i_\ell^\star, j \in [m^\star]\setminus\mathcal{C}_\ell \\ \left(1-\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)^2\right)\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t) \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 & \\ \qquad \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i = i_\ell^\star, j = j_\ell^\star \end{cases}$$

Analysis of $\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{w}}_j^\star$. To analyze $\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{w}}_j^\star$, we first compute the following quantities

$$\sum_{r=1}^m \lambda_{i,r,5}(t)\gamma_{r,j}^{(2)}(t) = \lambda_{i,i,5}(t)\gamma_{i,j}^{(5)}(t) + \lambda_{i,i_{\ell'}^\star,5}(t)\gamma_{i_{\ell'}^\star,j_{\ell'}^\star}^{(2)}(t)\mathbb{I}\left\{j = j_{\ell'}^\star, \ell' < \ell\right\}$$
$$\pm \begin{cases} \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\mathbb{I}\left\{j = j_\ell^\star\right\} & \text{if } i \in [m]\setminus\mathcal{R}_\ell \\ \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right) & \text{if } i = i_\ell^\star \end{cases}$$

$$\sum_{r=1}^m \lambda_{i,r,2}(t)\zeta_{i,j}^{(1)}(t) = \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) \pm \begin{cases} \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \\ \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i = i_\ell^\star \end{cases}$$

$$\sum_{r=1}^m \lambda_{i,r,3}(t)\zeta_{r,j}^{(1)}(t) = \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) \pm \begin{cases} \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \\ \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i = i_\ell^\star \end{cases}$$

$$\sum_{r=1}^{m^\star} \hat{\lambda}_{i,r,2}(t)\zeta_{i,j}^{(1)}(t) = \begin{cases} \pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) & \text{if } i = i_\ell^\star \\ \pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \end{cases}$$

Therefore, combining with (13), and noticing that the term $\lambda_{i,i,5}(t)\gamma_{i,j}^{(2)}(t)$ cancels out, we have that

$$\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{w}}_{j_{\ell'}^\star}^\star = \frac{9}{b_i}\lambda_{i,i_{\ell'}^\star,5}(t)\gamma_{i_{\ell'}^\star,j_{\ell'}^\star}^{(2)}(t)$$
$$-\frac{9}{b_i}\cdot\begin{cases} \hat{\lambda}_{i,j,5}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \\ \hat{\lambda}_{i,j,5}(t) - \lambda_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\gamma_{i,j}^{(2)}(t) & \\ \qquad \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 & \text{if } i = i_\ell^\star \\ \qquad \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t) + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \end{cases}$$

Moreover, we have that

$$\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{w}}_j^\star = -\frac{9}{b_i}\cdot\begin{cases}\hat{\lambda}_{i,j,5}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \\ \qquad \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \\ \hat{\lambda}_{i,j,5}(t) - \lambda_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\gamma_{i,j}^{(2)}(t) & \\ \qquad \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) & \text{if } i = i_\ell^\star, j \in [m^\star]\setminus\mathcal{C}_\ell \\ \qquad \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t) + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)\varepsilon_{5,\ell}(t)\right) & \\ \left(1-\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\right)\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t) \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) & \\ \qquad \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t) + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i = i_\ell^\star, j = j_\ell^\star\end{cases}$$

Analysis of $\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{w}}_j^\star$. To analyze $\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{w}}_j^\star$, we first compute the following quantities for $i \in [m]\setminus\mathcal{R}_{\ell-1}$:

$$\hat{\lambda}_{i,j,4}(t) = \begin{cases}\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \\ \pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i = i_\ell^\star, j \neq j_\ell^\star\end{cases}$$

$$\sum_{r=1}^m \lambda_{i,r,4}(t)\gamma_{r,j}^{(2)}(t) = 6C_{S,2}\gamma_{i,j}^{(2)}(t)I_{i,i}^{(3)}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)$$

$$\sum_{r=1}^m \lambda_{i,r,1}(t)\zeta_{r,j}^{(1)}(t) = \lambda_{i,i,1}(t)\zeta_{i,j}^{(1)}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$

$$\sum_{r=1}^m \lambda_{i,r,2}(t)\gamma_{i,j}^{(2)}(t) = 6C_{S,2}\gamma_{i,j}^{(2)}(t)I_{i,i}^{(3)}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)$$

$$\sum_{r=1}^{m^\star} \hat{\lambda}_{i,r,2}(t)\gamma_{i,j}^{(2)}(t) = \begin{cases}\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \\ \hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i,j}^{(2)}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i = i_\ell^\star\end{cases}$$

Combining with (12) and noticing that the term $\lambda_{i,i,1}(t)\zeta_{i,j}^{(1)}(t)$ cancels out, we have that

$$\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{w}}_j^\star = -\frac{1}{a_i}\cdot\begin{cases}\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \\ \hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i,j}^{(2)}(t) - \hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\zeta_{i,j}^{(1)}(t) & \\ \qquad - 36C_{S,2}\gamma_{i,j}^{(2)}(t)I_{i,i}^{(3)}(t) & \text{if } i = i_\ell^\star, j \neq j_\ell^\star \\ \qquad \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t) + \varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2\right) & \\ 3\hat{\lambda}_{i_\ell^\star,j_\ell^\star,4}(t) + \hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) - \hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\zeta_{i_\ell^\star,j_\ell^\star}^{(1)}(t) & \\ \qquad - 36C_{S,2}\gamma_{i,j}^{(2)}(t)I_{i,i}^{(3)}(t) & \text{if } i = i_\ell^\star, j = j_\ell^\star \\ \qquad \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t) + \varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2\right) & \end{cases}$$

Analysis of $\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{v}}_j^\star$. To analyze $\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{v}}_j^\star$, we first compute the following for $i\in[m]\setminus\mathcal{R}_{\ell-1}$:

$$\hat{\lambda}_{i,j,3}(t)=\begin{cases}\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\\ \pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i=i_\ell^\star, j\neq j_\ell^\star\end{cases}$$

$$\sum_{r=1}^m\lambda_{i,r,3}(t)\gamma_{r,j}^{(1)}(t)=6C_{S,2}\gamma_{i,j}^{(1)}(t)I_{i,i}^{(3)}(t)\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)$$

$$\sum_{r=1}^m\lambda_{i,r,2}(t)\gamma_{i,j}^{(1)}(t)=6C_{S,2}\gamma_{i,j}^{(1)}(t)I_{i,i}^{(3)}(t)\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)$$

$$\sum_{r=1}^m\lambda_{i,r,5}(t)\zeta_{r,j}^{(2)}(t)=\lambda_{i,i,5}(t)\zeta_{i,j}^{(2)}(t)\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$

$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,2}(t)\gamma_{i,j}^{(1)}(t)=\begin{cases}\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\\ \hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i,j}^{(1)}(t)\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) & \text{if } i=i_\ell^\star\end{cases}$$

Combining with (13) and noticing that the term $\lambda_{i,i,5}\zeta_{i,j}^{(2)}(t)$ cancels out, we have that

$$\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\bar{\mathbf{v}}_j^\star=\frac{3}{b_i}\cdot\begin{cases}\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3+\varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\\ \hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i,j}^{(1)}(t)-3\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\zeta_{i,j}^{(2)}(t)\\ \quad-12C_{S,2}\gamma_{i,j}^{(2)}(t)I_{i,i}^{(3)}(t) & \text{if } i=i_\ell^\star, j\neq j_\ell^\star\\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t)+\varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2\right)\\ \hat{\lambda}_{i_\ell^\star,j_\ell^\star,3}(t)+\hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)-3\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\zeta_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\\ \quad-12C_{S,2}\gamma_{i,j}^{(2)}(t)I_{i,i}^{(3)}(t) & \text{if } i=i_\ell^\star, j=j_\ell^\star\\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t)+\varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2\right)\end{cases}$$

Analysis of $\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top\bar{\mathbf{v}}_j$. To analyze $\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top\bar{\mathbf{v}}_j$, we first compute that for $i,j\in[m]\setminus\mathcal{R}_{\ell-1}, i\neq j$

$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,1}(t)\gamma_{j,r}^{(1)}(t)=\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{j,j_\ell^\star}^{(1)}(t)\mathbb{I}\left\{i=i_\ell^\star\right\}\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$

$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,2}(t)I_{j,i}^{(3)}(t)=\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\mathbb{I}\left\{i=i_\ell^\star\right\}\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$

$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,4}(t)\zeta_{j,r}^{(1)}(t)=\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\mathbb{I}\left\{i=i_\ell^\star\right\}\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$

$$\sum_{r=1}^m\lambda_{i,r,2}(t)I_{j,i}^{(3)}(t)=\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\pm\begin{cases}\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i\neq i_\ell^\star\\ \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i=i_\ell^\star\end{cases}$$

$$\sum_{r=1}^m\lambda_{i,r,4}(t)I_{j,r}^{(3)}(t)=\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\pm\begin{cases}\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i\neq i_\ell^\star\\ \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i=i_\ell^\star\end{cases}$$

Combining with (12) and noticing that the term $\lambda_{i,i,1}(t)I_{i,j}^{(1)}(t)$ and the term $\lambda_{j,j,1}(t)I_{i,j}^{(1)}(t)$ cancels out, we have

$$\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top\bar{\mathbf{v}}_j=-\frac{1}{a_i}\cdot\begin{cases}\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3+\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i\in[m]\setminus\mathcal{R}_\ell\\ \hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{j,j_\ell^\star}^{(1)}(t)-\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)I_{i,j}^{(1)}(t)\\ \quad\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 & \text{if } i=i_\ell^\star\\ \quad\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3+\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right)\end{cases}$$

Analysis of $\nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top \bar{\mathbf{w}}_j$. To analyze $\nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top \bar{\mathbf{w}}_j$, we first compute that for $i,j \in [m] \setminus \mathcal{R}_{\ell-1}, i \neq j$

$$\sum_{r=1}^{m^\star} \hat{\lambda}_{i,r,2}(t) I_{i,j}^{(3)}(t) = \pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\mathbb{I}\{i=i_\ell^\star\} \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$
$$\sum_{r=1}^{m^\star} \hat{\lambda}_{i,r,3}(t) \zeta_{j,r}^{(2)}(t) = \pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\mathbb{I}\{i=i_\ell^\star\} \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$
$$\sum_{r=1}^{m^\star} \hat{\lambda}_{i,r,5}(t) \gamma_{j,r}^{(2)}(t) = \hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}\gamma_{j,j_\ell^\star}^{(2)}(t)\mathbb{I}\{i=i_\ell^\star\} \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\mathbb{I}\{j=i_\ell^\star\} \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$
$$\sum_{r=1}^{m} \lambda_{i,r,2}(t) I_{i,j}^{(3)}(t) = \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) \pm \begin{cases} \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i \neq i_\ell^\star \\ \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i = i_\ell^\star \end{cases}$$
$$\sum_{r=1}^{m} \lambda_{i,r,3}(t) I_{r,j}^{(3)}(t) = \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) \pm \begin{cases} \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i \neq i_\ell^\star \\ \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i = i_\ell^\star \end{cases}$$
$$\sum_{r=1}^{m} \lambda_{i,r,5}(t) I_{r,j}^{(2)}(t) = \lambda_{i,i,5}(t) I_{i,j}^{(2)}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$

Combining with (13) and noticing that the term $\lambda_{i,i,5}(t)I_{i,j}^{(2)}(t)$ and the term $\lambda_{j,j,5}(t)I_{i,j}^{(2)}(t)$ cancels out, we have

$$\nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top \bar{\mathbf{w}}_j = -\frac{9}{b_i}\cdot \begin{cases} \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i,j \in [m]\setminus\mathcal{R}_\ell \\ \hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{j,j_\ell^\star}^{(2)}(t) - \hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) I_{i,j}^{(2)}(t) & \\ \qquad \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 & \text{if } i = i_\ell^\star \\ \qquad \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \\ \pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) & \\ \qquad \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } j = i_\ell^\star \end{cases}$$

Analysis of $\nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top \bar{\mathbf{w}}_j$. The analysis of $\nabla_{\mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top \bar{\mathbf{w}}_j$ will be separated into two cases. First, regardless of the cases, we have that

$$\sum_{r=1}^{m^\star} \hat{\lambda}_{i,r,1}(t)\zeta_{j,r}(t) = \hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\zeta_{j,j_\ell^\star}^{(2)}(t)\mathbb{I}\{i=i_\ell^\star\} \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$
$$\sum_{r=1}^{m} \lambda_{i,r,1}(t) I_{r,j}^{(3)}(t) = \lambda_{i,i,1}(t) I_{i,j}^{(3)}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$

We first analyze the case $i=j$, and then we dive into $i \neq j$. In the case where $i=j$, we have that

$$\sum_{r=1}^{m^\star} \hat{\lambda}_{i,r,2}(t) I_{i,j}^{(2)}(t) = \hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\mathbb{I}\{i=i_\ell^\star\} \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$
$$\sum_{r=1}^{m^\star} \hat{\lambda}_{i,r,4}(t)\gamma_{j,r}^{(2)}(t) = \hat{\lambda}_{i_\ell^\star,j_\ell^\star,4}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\mathbb{I}\{i=i_\ell^\star\} \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$
$$\sum_{r=1}^{m} \lambda_{i,r,2}(t) I_{i,j}^{(2)}(t) = 6C_{S,2} I_{i,i}^{(3)}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$
$$\sum_{r=1}^{m} \lambda_{i,r,4}(t) I_{r,j}^{(2)}(t) = 6C_{S,2} I_{i,i}^{(3)}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$

Thus, for the case $i = j$, we have that

$$\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}_j = -\frac{1}{a_i}\left(\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\zeta^{(2)}_{j,j_\ell^\star}(t) + 3\hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t) + 3\hat{\lambda}_{i_\ell^\star,j_\ell^\star,4}(t)\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)\right)\mathbb{I}\left\{i = i_\ell^\star\right\}$$
$$- \frac{1}{a_i}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma^{(1)}_{i_\ell^\star,j_\ell^\star}(t)\mathbb{I}\left\{i = i_\ell^\star\right\} - \frac{36}{a_i}C_{S,2}I^{(3)}_{i,i}(t) \pm \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)$$

Now, for the case $i \neq j$, we can compute that

$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,2}(t)I^{(2)}_{i,j}(t) = \pm\mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^2\right)\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)^2\mathbb{I}\left\{i = i_\ell^\star\right\} \pm \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)$$
$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,4}(t)\gamma^{(2)}_{j,r}(t) = \pm\mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^2\right)\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)^2\mathbb{I}\left\{i = i_\ell^\star\right\} \pm \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)$$
$$\sum_{r=1}^{m}\lambda_{i,r,2}(t)I^{(2)}_{i,j}(t) = \pm\mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right) \pm \begin{cases}\mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\varepsilon^{(3)}_{\mathcal{B},\ell}(t)\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell\\ \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i = i_\ell^\star\end{cases}$$
$$\sum_{r=1}^{m}\lambda_{i,r,4}(t)I^{(2)}_{r,j}(t) = \pm\mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right) \pm \begin{cases}\mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\varepsilon^{(3)}_{\mathcal{B},\ell}(t)\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell\\ \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i = i_\ell^\star\end{cases}$$

Therefore, combining with (12) gives

$$\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}_j = -\frac{1}{a_i}\cdot\begin{cases}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\left(\zeta^{(2)}_{j,j_\ell^\star}(t) - \gamma^{(1)}_{i_\ell^\star,j_\ell^\star}(t)I^{(3)}_{i,j}(t)\right)\\ \qquad \pm\mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^2\right)\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)^2 & \text{if } i = i_\ell^\star\\ \qquad \pm\mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(2)}_{\mathcal{A},\ell}(t)\varepsilon_{5,\ell}(t)\right)\\ \pm\mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(2)}_{\mathcal{A},\ell}(t)\varepsilon^{(3)}_{\mathcal{B},\ell}(t)\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell\end{cases}$$

Analysis of $\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top\bar{\mathbf{v}}_j$. The analysis of $\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top\bar{\mathbf{v}}_j$ will also be separated into two cases. First, regardless of the cases, we have that

$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,5}(t)\zeta^{(1)}_{j,r}(t) = \hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\zeta^{(1)}_{j,j_\ell^\star}(t)\mathbb{I}\left\{i = i_\ell^\star\right\} \pm \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)$$
$$\sum_{r=1}^{m}\lambda_{i,r,5}(t)I^{(3)}_{j,r}(t) = \lambda_{i,i,5}(t)I^{(3)}_{j,i}(t) \pm \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^2\right)$$

We first analyze the case $i = j$, and then we dive into $i \neq j$. In the case where $i = j$, we have that

$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,2}(t)I^{(1)}_{j,i}(t) = \hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\mathbb{I}\left\{i = i_\ell^\star\right\} \pm \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)$$
$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,3}(t)\gamma^{(1)}_{j,r}(t) = \hat{\lambda}_{i_\ell^\star,j_\ell^\star,3}(t)\gamma^{(1)}_{i_\ell^\star,j_\ell^\star}(t)\mathbb{I}\left\{i = i_\ell^\star\right\} \pm \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)$$
$$\sum_{r=1}^{m}\lambda_{j,r,2}(t)I^{(1)}_{i,j}(t) = 6C_{S,2}I^{(3)}_{i,i}(t) \pm \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)$$
$$\sum_{r=1}^{m}\lambda_{j,r,3}(t)I^{(1)}_{i,r}(t) = 6C_{S,2}I^{(3)}_{i,i}(t) \pm \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)$$

Thus, for the case $i = j$, we have that

$$\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top\bar{\mathbf{v}}_j = -\frac{3}{b_i}\cdot\left(3\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\zeta^{(1)}_{j,j_\ell^\star}(t) + \hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t) + \hat{\lambda}_{i_\ell^\star,j_\ell^\star,3}(t)\gamma^{(1)}_{i_\ell^\star,j_\ell^\star}(t)\right)\mathbb{I}\left\{i = i_\ell^\star\right\}$$
$$- \frac{9}{b_i}\cdot\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)\mathbb{I}\left\{i = i_\ell^\star\right\} - 36C_{S,2}I^{(3)}_{i,i}(t) \pm \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^2\right)$$

Now, for the case $i \neq j$, we can compute that

$$\sum_{r=1}^{m^\star} \hat{\lambda}_{i,r,2}(t) I_{j,i}^{(1)}(t) = \pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\mathbb{I}\{i=i_\ell^\star\} \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$
$$\sum_{r=1}^{m^\star} \hat{\lambda}_{i,r,3}(t) \gamma_{j,r}^{(1)}(t) = \pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\mathbb{I}\{i=i_\ell^\star\} \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$
$$\sum_{r=1}^{m} \lambda_{i,r,2}(t) I_{j,i}^{(1)}(t) = \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) \pm \begin{cases} \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \\ \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i = i_\ell^\star \end{cases}$$
$$\sum_{r=1}^{m} \lambda_{i,r,3}(t) I_{j,r}^{(1)}(t) = \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) \pm \begin{cases} \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \\ \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \text{if } i = i_\ell^\star \end{cases}$$

Therefore, combining with (12) and (13) gives

$$\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_j = -\frac{9}{b_i}\cdot\begin{cases} \hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\left(\zeta_{j,j_\ell^\star}^{(1)}(t) - \gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) I_{i,j}^{(3)}(t)\right) & \\ \qquad \pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 & \text{if } i = i_\ell^\star \\ \qquad \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t)\right) & \\ \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right) & \text{if } i \in [m]\setminus\mathcal{R}_\ell \end{cases}$$

□

A.5 Establishing the Inductive Hypothesis: Phase 1

Starting from this section, we assume that for a fixed $\ell \in [m^\star]$, the inductive hypothesis holds. That is, the condition of Lemma 7 holds. Then we shall analyze the convergence for that ℓ to prove the inductive hypothesis and establish convergence. Notice that, by the statement of the inductive hypothesis, the case $\ell = 1$ naturally satisfies it, thus requiring no additional proof. Therefore, we focus on the case of a general fixed ℓ, and the proof will be divided into two phase. Phase 1 (this section) show that there exist some $T^\star$ such that $\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(T^\star)$ and $\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(T^\star)$ are some constant close to 1, while $\varepsilon_{\mathcal{A},\ell}^{(2)}(t), \varepsilon_{\mathcal{B},\ell}^{(3)}(t)$, and $\varepsilon_{5,\ell}(t)$ are $\mathcal{O}\left(\frac{m^2}{\delta_\mathbb{P}\sqrt{d}}\right)$.

We denote the following time

$$\begin{aligned} T_c &:= \min\left\{t \geq 0 : A_1(t) \wedge A_2(t) \wedge A_3(t)\right\} \\ A_1(t) &= \left\{\min\left\{\varepsilon_{\mathcal{B},\ell}^{(2)}(t), \varepsilon_{3,\ell}(t), \varepsilon_{4,\ell}(t)\right\} > \frac{\beta_6 m^2}{\delta_\mathbb{P}\sqrt{d}}\right\}; \quad A_2(t) = \left\{\varepsilon_{\mathcal{B},\ell}^{(3)}(t) \geq \beta_6 m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right\} \\ A_3(t) &= \left\{\varepsilon_{5,\ell}(t) \geq \beta_6\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{B},\ell}^{(1)}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\right)\right\} \end{aligned} \tag{14}$$

By the initialization property, we have that

$$\varepsilon_{\mathcal{B},\ell}^{(1)}(t), \varepsilon_{3,\ell}(t), \varepsilon_{4,\ell}(t) \leq \frac{\beta_3}{\sqrt{d}}\left(\log\frac{m}{\delta_\mathbb{P}}\right)^{\frac{1}{2}}; \quad \varepsilon_{\mathcal{B},\ell}^{(3)}(0) = \varepsilon_{5,\ell}(0) = 0$$

Therefore, $T_c > 0$. Moreover, for all $t \leq T_c$, by the inductive hypothesis, we have that

$$\varepsilon_{\mathcal{B},\ell}^{(1)}(t), \varepsilon_{3,\ell}(t), \varepsilon_{4,\ell}(t) \leq \frac{\beta_6 m^2}{\delta_\mathbb{P}\sqrt{d}} \Rightarrow \varepsilon_{\mathcal{F},\ell}(t) \leq \mathcal{O}\left(\frac{m^2}{\delta_\mathbb{P}\sqrt{d}}\right)$$

which implies that $\varepsilon_{\mathcal{A},\ell}^{(2)}(t) \leq \mathcal{O}\left(\frac{m^2}{\delta_\mathbb{P}\sqrt{d}}\right)$. Also, by definition we have that $\varepsilon_{\mathcal{A},\ell}^{(1)}(t) \leq \varepsilon_{\mathcal{A},\ell}^{(2)}(t)$.

Growth of $\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)$. We will start with analyzing $\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)$. For any $\xi \in (0,1)$, recall the definition of $T_\ell(\xi)$ in Definition 2. We should notice that, although by definition we have $\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(0) > 0$, it is not always the case that

$\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(0)$ will also be positive. Therefore, we also need to control the "negativeness" of $\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(0)$ when analyzing the growth of $\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)$. To do this, we present the following lemma.

Lemma 8. *Suppose that the inductive hypothesis in Condition 2 and the initialization condition in Condition 1 holds. Let T_c be defined in (14) and $T(\xi)$ in Definition 2. If $T_c \geq T_\ell\left(\frac{1}{2}\right)$, then we have that*

$$\frac{d}{dt}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \geq 0; \lambda_{i^\star_\ell,j^\star_\ell,5}(t) \geq 0; \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t) \geq -\frac{\beta_6}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}; \forall t \leq \min\{T_c, T(\xi)\}$$

$$\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t+T_0) \geq \left(\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(T_0)^{-1} - \frac{18}{b^2_{i^\star_\ell}}\left(c_0^2\left(1-\xi^2\right) - \mathcal{O}\left(\frac{m^7}{\delta^3_{\mathbb{P}}\sqrt{d}}\right)\right)t\right)^{-1}$$

$$T_\ell\left(\xi\right) \leq \frac{b^2_{i^\star_\ell}}{18}\left(c_0^2\left(1-\xi\right) - \mathcal{O}\left(\frac{m^7}{\delta^3_{\mathbb{P}}\sqrt{d}}\right)\right)^{-1}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(0)^{-1}$$

for all ξ such that $(1-\xi^2)^{-1} \leq \mathcal{O}(1)$.

Proof. To start, we lower bound the time-derivative of $\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)$. Let T_c' be defined as

$$T_c' = \min\left\{t \geq 0 : \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t) < -\frac{\beta_6}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right\}$$

Since $\left(\mathbf{v}_{i^\star_\ell}^\top \bar{\mathbf{v}}^\star_j\right) \leq \frac{\beta_3}{d}\log\frac{m}{\delta_{\mathbb{P}}}$ and $\left\|\mathbf{v}_{i^\star_\ell}\right\|_2 \geq 1-\beta_2\delta_s$ for some $\beta_3 > 0, \beta_2 \leq o(1)$ and $\delta_s \leq \mathcal{O}\left(\frac{1}{m^2}\right)$, we have that $\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(0)^2 \leq \frac{2\beta_3}{d}\log\frac{m}{\delta_{\mathbb{P}}}$. For $\beta_6 \geq 4\sqrt{\beta_3}$, we must have that $T_c' > 0$. By Lemma 30, we have that for all $t \leq T_c'$, it holds that

$$\sum_{k=0}^{\infty}\frac{c_k^2}{k!}\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)^k \in c_0^2 \pm \mathcal{O}\left(\frac{1}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right); \quad \sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)^k \in c_1^2 \pm \mathcal{O}\left(\frac{1}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right)$$

By Lemma 7, we have that

$$\begin{aligned}\frac{d}{dt}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) &= -\frac{1}{b_{i^\star_\ell}}\nabla_{\mathbf{w}^\star_{i^\star_\ell}}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}^\star_{j^\star_\ell} \\ &= \frac{9}{b^2_{i^\star_\ell}}\left(1-\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,5}(t) \pm \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^2\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \pm \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(2)}_{\mathcal{A},\ell}(t)\varepsilon_{5,\ell}(t)\right)\end{aligned}$$

Therefore, with the form of $\hat{\lambda}_{i^\star_\ell,j^\star_\ell,5}$ from Lemma 5, we have that

$$\begin{aligned}\frac{d}{dt}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) &= \frac{18}{b^2_{i^\star_\ell}}\left(1-\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2\sum_{k=0}^{\infty}\frac{c_k^2}{k!}\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)^k \pm \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(2)}_{\mathcal{A},\ell}(t)\varepsilon_{5,\ell}(t)\right) \\ &\quad \pm \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^2\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \\ &\geq \frac{18}{b^2_{i^\star_\ell}}\left(1-\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2 - \mathcal{O}\left(\frac{m^4}{\delta^2_{\mathbb{P}}\sqrt{d}}\right)\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2\left(c_0^2 - \mathcal{O}\left(\frac{1}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right)\right) \\ &\quad - \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(2)}_{\mathcal{A},\ell}(t)\varepsilon_{5,\ell}(t)\right) \\ &\geq \frac{18}{b^2_{i^\star_\ell}}\left(1-\xi^2 - \mathcal{O}\left(\frac{m^4}{\delta^2_{\mathbb{P}}\sqrt{d}}\right)\right)\left(c_0^2 - \mathcal{O}\left(\frac{1}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right)\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2 - \mathcal{O}\left(\frac{m^7}{\delta^3_{\mathbb{P}}d^{\frac{3}{2}}}\right) \\ &\geq \frac{18}{b^2_{i^\star_\ell}}\left(c_0^2\left(1-\xi^2\right) - \mathcal{O}\left(\frac{m^4}{\delta^2_{\mathbb{P}}\sqrt{d}}\right)\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2 - \mathcal{O}\left(\frac{m^7}{\delta^3_{\mathbb{P}}d^{\frac{3}{2}}}\right)\end{aligned} \tag{15}$$

Let ξ be given such that $\left(1-\xi^2\right)^{-1} \leq \mathcal{O}(1)$ By the lower bound that $\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(0)^2 \geq \frac{\log m^\star}{d}$, we have that

$$\frac{d}{dt}\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(0) \geq \beta_{\text{temp}}\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(0)^2 - \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right) \geq 0$$

for $d \geq \beta_5 \delta_{\mathbb{P}}^6 m^{16}$ for some $\beta_{\text{temp}} > 0$. This shows that $\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)^2 \geq \frac{\log m^\star}{d}$ for all $t \leq \min\{T_c, T_c', T_\ell(\xi)\}$ for any $\left(1-\xi^2\right)^{-1} \leq \mathcal{O}(1)$, and thus $\frac{d}{dt}\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t) \geq 0$ for the sane choice of t. This shows the first property. Moreover, by Lemma 5,

$$\begin{aligned}
\hat{\lambda}_{i^\star_\ell, j^\star_\ell, 5}(t) &= 2\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)^2 \sum_{k=0}^{\infty} \frac{c_k^2}{k!}\gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t)^k \pm \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^2\right)\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)^2 \pm \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^4\right) \\
&\geq 2\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)^2 \left(c_0^2 - \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2\sqrt{d}}\right)\right) - \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right) \\
&\geq 0
\end{aligned}$$

as $d \geq \beta_5 \delta_{\mathbb{P}}^6 m^{16}$. This shows the second property. To lower bound $\gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t)$, we first write that, by Lemma 7

$$\begin{aligned}
\frac{d}{dt}\gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t) &= -\frac{1}{a_i}\nabla_{\mathbf{v}_{i^\star_\ell}}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}^\star_{j^\star_\ell} \\
&= \frac{1}{a_i^2}\left(1-\gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t)^2\right)\hat{\lambda}_{i^\star_\ell, j^\star_\ell, 1}(t) \pm \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^2\right)\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)^2 \pm \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(2)}_{\mathcal{A},\ell}(t)\varepsilon_{5,\ell}(t)\right)
\end{aligned}$$

Therefore, we could notice that, since

$$\sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}\gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t)^k \geq c_1^2 - \mathcal{O}\left(\frac{1}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right) \geq 0$$

for $d \geq \beta_5 m^{14} \geq \beta_5 \log\frac{m}{\delta_{\mathbb{P}}}$[4], we must have that for all $t \leq \min\{T_c, T_c', T_\ell(\xi)\}$ for any $\left(1-\xi^2\right)^{-1} \leq \mathcal{O}(1)$, either $\gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t) \geq \frac{1}{2}$, or

$$\begin{aligned}
\frac{d}{dt}\gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t) &\geq 6\left(1-\gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t)^2\right)\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)^3 \sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}\gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t)^k - \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(2)}_{\mathcal{A},\ell}(t)\varepsilon_{5,\ell}(t)\right) \\
&\quad - \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^2\right)\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)^2 \\
&\geq c_1^2\left(\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t) - \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^2\right)\right)\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)^2 - \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right) \\
&\geq -\mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)
\end{aligned} \tag{16}$$

where the last inequality follows from the fact that

$$\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^2 \leq \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right) \leq \left(\frac{\log m^\star}{d}\right)^{\frac{1}{2}} \leq \gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(0) \leq \gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)$$

Therefore, for all $t \leq \min\{T_c, T_c', T_\ell(\xi)\}$ for any $\left(1-\xi^2\right)^{-1} \leq \mathcal{O}(1)$ we must have that

$$\gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t) \geq \gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(0) - \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)\cdot t \geq -\frac{\sqrt{2\beta_4}}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}$$

[4] since we require $m \geq \frac{\beta_4 \log m^\star}{\delta_{\mathbb{P}}}$

Choosing $\beta_6 \geq 4\sqrt{\beta_4}$ gives that

$$T_{c'} \geq \min\left\{T_c, T_\ell(\xi), \Omega\left(\frac{d}{m^7}\right)\right\}$$

Now we are going to lower bound $T_\ell(\xi)$. When $d \geq \beta_5 m^{15}$,

$$\frac{d}{dt}\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t) \geq \frac{18}{b^2_{i^\star_\ell}}\left(c_0^2\left(1-\xi^2\right) - \mathcal{O}\left(\frac{m^7}{\delta^3_{\mathbb{P}}\sqrt{d}}\right)\right)\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)^2$$

Solving the differential equation gives that

$$\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t+T_0) \geq \left(\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(T_0)^{-1} - \frac{18}{b^2_{i^\star_\ell}}\left(c_0^2\left(1-\xi^2\right) - \mathcal{O}\left(\frac{m^7}{\delta^3_{\mathbb{P}}\sqrt{d}}\right)\right)t\right)^{-1}$$

for any $T_0 \geq 0$ and $t + T_0 \leq T_c$. This gives that, if $T(\xi) \leq T_c$, then

$$\begin{aligned} T_\ell(\xi) &\leq \frac{b^2_{i^\star_\ell}}{18}\left(c_0^2\left(1-\xi\right) - \mathcal{O}\left(\frac{m^7}{\delta^3_{\mathbb{P}}\sqrt{d}}\right)\right)^{-1}\left(\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(0)^{-1} - \zeta^{-1}\right) \\ &\leq \frac{b^2_{i^\star_\ell}}{18}\left(c_0^2\left(1-\xi\right) - \mathcal{O}\left(\frac{m^7}{\delta^3_{\mathbb{P}}\sqrt{d}}\right)\right)^{-1}\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(0)^{-1} \end{aligned}$$

This shows the last two properties. Moreover, we have that $T_\ell(\frac{1}{2}) \leq \mathcal{O}\left(\left(\frac{d}{\log m^\star}\right)^{\frac{1}{2}}\right)$. However, at $T_\ell\left(\frac{1}{2}\right)$, by (16), we have that

$$\frac{d}{dt}\gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t) \geq c_1^2\left(\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t) - \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^2\right)\right)\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)^2 - \mathcal{O}\left(\frac{m^7}{\delta^3_{\mathbb{P}} d^{\frac{3}{2}}}\right) \geq 0$$

which implies that $\gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t) \geq -\frac{\beta_6}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}$ for all $t \leq \min\left\{T_c, T_\ell(\xi)\right\}$ for all ξ such that $(1-\xi^2)^{-1} \leq \mathcal{O}(1)$, as long as $T\left(\frac{1}{2}\right) \leq T_c$. □

Upper bounding $\gamma^{(2)}_{i,j}(t)$ for $i \in [m] \setminus \mathcal{R}_{\ell-1}$ and $j \in [m^\star] \setminus \mathcal{C}_{\ell-1}$ with $(i,j) \neq (i^\star_\ell, j^\star_\ell)$. Here we are going to show that the growth of $\gamma^{(2)}_{i,j}(t)$ with $i \in [m] \setminus \mathcal{R}_{\ell-1}, j \in [m^\star] \setminus \mathcal{C}_{\ell-1}$ and $(i,j) \neq (i^\star_\ell, j^\star_\ell)$ is slow in terms of when $\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)$ reaches ξ, $\gamma^{(2)}_{i,j}(t)$ is no bigger than $\mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$.

Lemma 9. *Suppose that the inductive hypothesis in Condition 2, and the initialization condition in Condition 1. Let T_c be defined in (14) and $T_\ell(\xi)$ in Definition 2. Then there exists some constant $\beta_7 > 0$ such that for all $t \leq \min\left\{T_c, T\left(\frac{\delta^2_{\mathbb{P}}}{d^{\frac{5}{2}}}\right) + \frac{\beta_7\delta_{\mathbb{P}}\sqrt{d}}{m^2}\right\}$, we have that $\left|\gamma^{(2)}_{i,j}(t)\right| \leq O\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$ for all $(i,j) \neq (i^\star_\ell, j^\star_\ell)$.*

Proof. For $t \leq T_c$, by Lemma 7, we write out the dynamic of $\gamma^{(2)}_{i,j}(t)$ as

$$\begin{aligned} \frac{d}{dt}\gamma^{(2)}_{i,j}(t) &= -\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}^\star_j \\ &= \frac{18}{b_i^2}\sum_{k=0}^{\infty}\frac{c_k^2}{k!}\gamma^{(1)}_{i,j}(t)^k\gamma^{(2)}_{i,j}(t)^2 - \frac{9}{b_i^2}\cdot\hat{\lambda}_{i^\star_\ell, j^\star_\ell, 5}\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)\gamma^{(2)}_{i,j}(t)\mathbb{I}\left\{i = i^\star_\ell\right\} \\ &\quad \pm \mathcal{O}\left(\frac{m^7}{\delta^3_{\mathbb{P}} d^{\frac{3}{2}}}\right) \pm \mathcal{O}\left(\frac{m^4}{\delta^2_{\mathbb{P}} d}\right)\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)\mathbb{I}\left\{i = i^\star_\ell\right\} \end{aligned}$$

As shown in Lemma 8, $\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \geq 0, \hat{\lambda}_{i^\star_\ell,j^\star_\ell,5}(t) \geq 0$ for all $t \leq T_c$. Therefore, we have that

$$\begin{aligned}
\frac{d}{dt}\left|\gamma^{(2)}_{i,j}(t)\right| &\leq \frac{18}{b_i^2}\left|\sum_{k=0}^{\infty}\frac{c_k^2}{k!}\gamma^{(1)}_{i,j}(t)^k\right| \cdot \left|\gamma^{(2)}_{i,j}(t)\right|^2 - \frac{9}{b_i^2}\cdot\hat{\lambda}_{i^\star_\ell,j^\star_\ell,5}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)\left|\gamma^{(2)}_{i,j}(t)\right|\mathbb{I}\{i=i^\star_\ell\} \\
&\quad + \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right) + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \\
&\leq \frac{18}{b_i^2}\sum_{k=0}^{\infty}\frac{c_k^2}{k!}\left|\gamma^{(1)}_{i,j}(t)\right|^k\left|\gamma^{(2)}_{i,j}(t)\right|^2 + \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right) + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \\
&\leq \frac{18}{b_i^2}\left(c_0^2 + \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)\right)\left|\gamma^{(2)}_{i,j}(t)\right|^2 + \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right) + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)
\end{aligned}$$

where the last inequality is because $\left|\gamma^{(1)}_{i,j}(t)\right| \leq \varepsilon_1(t) \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$ for $t \leq T_c$. For any $t \leq T\left(\frac{\delta_{\mathbb{P}}^2}{d^{\frac{2}{5}}}\right)$, we must have that

$$\mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \leq \mathcal{O}\left(\frac{m^4}{d^{\frac{7}{5}}}\right)$$

Therefore, for any $t \leq T\left(\frac{\delta_{\mathbb{P}}^2}{d^{\frac{2}{5}}}\right)$, we have

$$\frac{d}{dt}\left|\gamma^{(2)}_{i,j}(t)\right| \leq \frac{18}{b_i^2}\left(c_0^2 + \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)\right)\left|\gamma^{(2)}_{i,j}(t)\right|^2 + \mathcal{O}\left(\frac{m^4}{d^{\frac{7}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)$$

Notice that $\left|\gamma^{(2)}_{i,j}(t)\right|$ must be upper bounded by $\hat{\gamma}^{(2)}_{i,j}(t)$ given by

$$\frac{d}{dt}\hat{\gamma}^{(2)}_{i,j}(t) = \frac{18}{b_i^2}\left(c_0^2 + \mathcal{O}\left(\frac{m^2}{\sqrt{d}}\right)\right)\hat{\gamma}^{(2)}_{i,j}(t)^2 + \mathcal{O}\left(\frac{m^4}{d^{\frac{7}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right); \quad \hat{\gamma}^{(2)}_{i,j}(0) = \left|\gamma^{(2)}_{i,j}(0)\right|$$

Observe that for any $t \leq T_c$, $\hat{\gamma}^{(2)}_{i,j}(t)$ grows monotonically as $\hat{\gamma}^{(2)}_{i,j}(0)$ grows. By the initialization property, we have that $\hat{\gamma}^{(2)}_{i,j}(0) = \left|\gamma^{(2)}_{i,j}(0)\right| \leq \frac{\sqrt{\beta_3}}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}$. We can observe that $\hat{\gamma}^{(2)}_{i,j}(t)$ increases as $\hat{\gamma}^{(2)}_{i,j}(t)$ becomes larger. Thus, it suffice to consider $\hat{\gamma}^{(2)}_{i,j}(0) = \frac{\sqrt{\beta_3}}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}$. In this case,

$$\frac{d}{dt}\hat{\gamma}^{(2)}_{i,j}(t) \leq \frac{18}{b_i^2}\left(c_0^2 + \mathcal{O}\left(\frac{m^4}{d^{\frac{2}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)\right)\hat{\gamma}^{(2)}_{i,j}(t)^2$$

Solving the differential equation gives that

$$\left|\gamma^{(2)}_{i,j}(t)\right| \leq \hat{\gamma}^{(2)}_{i,j}(t) \leq \left(\left|\gamma^{(2)}_{i,j}(0)\right|^{-1} - \frac{18}{b_i^2}\left(c_0^2 + \mathcal{O}\left(\frac{m^4}{d^{\frac{2}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)\right)t\right)^{-1}$$

Recall that $(1+\delta_s)^2 b_i^2 \gamma_{i,j}^{(2)}(0)^2 \le b_{i_\ell^\star}\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(0)^2$ by the initialization property. Thus, $t \le T\left(\frac{\delta_{\mathbb{P}}^2}{d^{\frac{2}{5}}}\right)$, it holds that

$$\begin{aligned}
\left|\gamma_{i,j}^{(2)}(t)\right| &\le \left(\left|\gamma_{i,j}^{(2)}(0)\right|^{-1} - \frac{18}{b_i^2}\left(c_0^2 + \mathcal{O}\left(\frac{m^4}{d^{\frac{2}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)\right) T\left(\frac{m^{\frac{3}{2}}}{d^{\frac{1}{4}}}\right)\right)^{-1} \\
&\le \left(\left|\gamma_{i,j}^{(2)}(0)\right|^{-1} - \frac{b_{i_\ell^\star}^2}{b_i^2}\left(c_0^2 + \mathcal{O}\left(\frac{m^4}{d^{\frac{2}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)\right)\left(c_0^2\left(1 - \frac{\delta_{\mathbb{P}}^2}{d^{\frac{2}{5}}}\right) - \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}\sqrt{d}}\right)\right)^{-1}\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(0)^{-1}\right)^{-1} \\
&\le \left(\left|\gamma_{i,j}^{(2)}(0)\right|^{-1} - \frac{b_{i_\ell^\star}^2}{b_i^2}\cdot\frac{c_0^2 + \mathcal{O}\left(\frac{m^4}{d^{\frac{2}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)}{c_0^2 - \mathcal{O}\left(\frac{\delta_{\mathbb{P}^2}}{d^{\frac{2}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)}\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(0)^{-1}\right)^{-1} \\
&\le \left(\frac{b_i}{b_{i_\ell^\star}}(1+\delta_s)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(0)^{-1} - \frac{b_{i_\ell^\star}^2}{b_i^2}\cdot\frac{c_0^2 + \mathcal{O}\left(\frac{m^4}{d^{\frac{2}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)}{c_0^2 - \mathcal{O}\left(\frac{\delta_{\mathbb{P}^2}}{d^{\frac{2}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)}\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(0)^{-1}\right)^{-1} \\
&\le \frac{b_{i_\ell^\star}}{b_i}\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(0)\left(1 + \delta_s - \frac{b_{i_\ell^\star}^3}{b_i^3}\cdot\frac{c_0^2 + \mathcal{O}\left(\frac{m^4}{d^{\frac{2}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)}{c_0^2 - \mathcal{O}\left(\frac{\delta_{\mathbb{P}^2}}{d^{\frac{2}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)}\right)^{-1}
\end{aligned}$$

By the initialization property, we have that $b_i, b_{i_\ell^\star} \in [1-\beta_2\delta_s, 1+\beta_2\delta_s]$ for $\beta_2 \le o(1)$. Therefore, we have that

$$\begin{aligned}
\frac{b_{i_\ell^\star}^3}{b_i^3}\cdot\frac{c_0^2 + \mathcal{O}\left(\frac{m^4}{d^{\frac{2}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)}{c_0^2 - \mathcal{O}\left(\frac{\delta_{\mathbb{P}^2}}{d^{\frac{2}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)} &\le \left(\frac{1+\beta_2\delta_s}{1-\beta_2\delta_s}\right)^3\frac{c_0^2 + \mathcal{O}\left(\frac{m^4}{d^{\frac{2}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)}{c_0^2 - \mathcal{O}\left(\frac{\delta_{\mathbb{P}^2}}{d^{\frac{2}{5}}} + \frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)} \\
&\le \left(\frac{1+\beta_2\delta_s}{1-\beta_2\delta_s}\right)^4 \\
&\le 1 + \frac{1}{2}\delta_s
\end{aligned}$$

where the second inequality is due to $d \ge \beta_5 m^{16}$ and the last inequality due to $\beta_2 \le o(1)$. Therefore

$$\left|\gamma_{i,j}^{(2)}(t)\right| \le \frac{b_{i_\ell^\star}}{b_i}\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(0)\cdot\frac{2}{\delta_s} \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$$

Let $T_1 = T\left(\frac{m^{\frac{3}{2}}}{d^{\frac{1}{4}}}\right)$. Then for $t \ge T_1$, the dynamic of $\gamma_{i,j}^{(2)}(t)$ is upper bounded by

$$\frac{d}{dt}\hat{\gamma}_{i,j}^{(2)}(t) \le \frac{18}{b_i^2}\left(c_0^2 + \mathcal{O}\left(\frac{m^2}{\sqrt{d}}\right)\right)\left|\gamma_{i,j}^{(2)}(t)\right|^2 + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)$$

Let T_2 be the smallest $t \ge 0$ such that $\hat{\gamma}_{i,j}^{(2)}(t) \ge 2\hat{\gamma}_{i,j}^{(2)}(T_1)$. Then for all $t \le T_2$, we have $\frac{d}{dt}\hat{\gamma}_{i,j}^{(2)}(t) \le \mathcal{O}\left(\frac{m^4}{d}\right)$. Therefore

$$\hat{\gamma}_{i,j}^{(2)}(t+T_1) \le \hat{\gamma}_{i,j}^{(2)}(T_1) + t\cdot\max_{t\le T_2}\frac{d}{dt}\hat{\gamma}_{i,j}^{(2)}(t) \le \hat{\gamma}_{i,j}^{(2)}(T_1) + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)t$$

Thus, we must have that $T_2 \ge \mathcal{O}\left(\frac{\delta_{\mathbb{P}}\sqrt{d}}{m^2}\right) + T_1$. Therefore, we can conclude that there exists some constant $\beta_7 > 0$ such that for all $t \le T\left(\frac{\delta_{\mathbb{P}}^2}{d^{\frac{2}{5}}}\right) + \frac{\beta_7\delta_{\mathbb{P}}\sqrt{d}}{m^2}$ and $t \le T_c$, we have that

$$\left|\gamma_{i,j}^{(2)}(t)\right| \le O\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$$

for all $(i,j) \ne (i_\ell^\star, j_\ell^\star)$. □

Upper bounding $\gamma_{i,j}^{(2)}(t)$ for $i \in [m] \setminus \mathcal{R}_{\ell-1}, j \in \mathcal{C}_{\ell-1}$. In this section, we show that the alignment of $\bar{\mathbf{w}}_i$ with previously recovered components $\bar{\mathbf{w}}_j^\star$ must be small.

Lemma 10. *Suppose that the inductive hypothesis in Condition 2 and the initialization condition in Condition 1 holds. Let T_c be defined in (14) and $T_\ell(\xi)$ in Definition 2. Then we have that for all $t \leq \min\left\{T_c, T_\ell(\xi) + \mathcal{O}\left(\frac{\delta_{\mathbb{P}}\sqrt{d}}{m^2}\right), \mathcal{O}\left(\frac{\delta_{\mathbb{P}}^2 d^{\frac{3}{4}}}{m^5}\right)\right\}$ for any ξ such that $(1-\xi)^{-1} \leq \mathcal{O}(1)$, it holds that*

$$\left|\gamma_{i,j_{\ell'}^\star}^{(2)}(t)\right| \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right); \forall i \in [m] \setminus \mathcal{R}_{\ell-1}, \ell' < \ell$$

Proof. By the inductive hypothesis, we have that $\gamma_{i_{\ell'}^\star, j_{\ell'}^\star}^{(2)}(t) \geq 1 - \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)$. Fix any $i \in [m] \setminus \mathcal{R}_{\ell-1}$. By Lemma 7, we have that

$$\begin{aligned}
\frac{d}{dt}\gamma_{i,j_\ell^\star}^{(2)}(t) &= -\frac{1}{b_i}\nabla_{\mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta}(t))^\top \bar{\mathbf{w}}_{j_{\ell'}^\star}^\star \\
&= \frac{9}{b_i^2}\left(\hat{\lambda}_{i,j_{\ell'}^\star,5}(t) - \lambda_{i,i_{\ell'}^\star,5}(t)\gamma_{i_{\ell'}^\star,j_{\ell'}^\star}^{(2)}(t)\right) - \frac{9}{b_i^2}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\gamma_{i,j_{\ell'}^\star}^{(2)}(t)\mathbb{I}\{i = i_\ell^\star\} \\
&\quad \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t) + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right)
\end{aligned}$$

Due to the same reasoning as in the previous lemma, we have that $\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) \geq 0$. Therefore, for all $t \leq T_c$, we have that

$$\begin{aligned}
\frac{d}{dt}\left|\gamma_{i,j_{\ell'}^\star}^{(2)}(t)\right| &\leq \frac{9}{b_i^2}\left|\hat{\lambda}_{i,j_{\ell'}^\star,5}(t) - \lambda_{i,i_\ell^\star,5}(t)\gamma_{i_{\ell'}^\star,j_{\ell'}^\star}^{(2)}(t)\right| + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right) \\
&\leq \frac{9c_0^2}{b_i^2}\left|\gamma_{i,j_{\ell'}}^{(2)}(t)^2 - I_{i,i_{\ell'}^\star}^{(2)}(t)^2\gamma_{i_{\ell'}^\star,j_{\ell'}^\star}^{(2)}(t)\right| + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right) \\
&\leq \frac{9c_0^2}{b_i^2}\left(1 - \gamma_{i_{\ell'}^\star,j_{\ell'}^\star}^{(2)}(t)\right)\gamma_{i,j_{\ell'}}^{(2)}(t)^2 + \frac{9c_0^2}{b_i^2}\gamma_{i_{\ell'}^\star,j_{\ell'}^\star}^{(2)}(t)\left|\gamma_{i,j_{\ell'}}^{(2)}(t)^2 - I_{i,i_{\ell'}^\star}^{(2)}(t)^2\right| \\
&\qquad + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)
\end{aligned}$$

Recall that $\gamma_{i,j_{\ell'}^\star}^{(2)}(t) = \bar{\mathbf{v}}_i^\top \bar{\mathbf{v}}_{j_{\ell'}^\star}^\star$ and $I_{i,i_{\ell'}^\star} = \bar{\mathbf{v}}_i^\top \bar{\mathbf{v}}_{i_{\ell'}^\star}^\star$. Therefore,

$$\begin{aligned}
\left|\gamma_{i,j_{\ell'}^\star}^{(2)}(t)^2 - I_{i,i_{\ell'}^\star}^{(2)}(t)^2\right| &= \left|\gamma_{i,j_{\ell'}^\star}^{(2)}(t) + I_{i,i_{\ell'}^\star}^{(2)}(t)\right| \cdot \left|\gamma_{i,j_{\ell'}^\star}^{(2)}(t) - I_{i,i_{\ell'}^\star}^{(2)}(t)\right| \\
&\leq 2\left|\gamma_{i,j_{\ell'}^\star}^{(2)}(t)\right| \cdot \left|\gamma_{i,j_{\ell'}}^{(2)}(t) - I_{i,i_{\ell'}^\star}^{(2)}(t)\right| + \left(\gamma_{i,j_{\ell'}}^{(2)}(t) - I_{i,i_{\ell'}^\star}^{(2)}(t)\right)^2 \\
&\leq 2\left|\gamma_{i,j_{\ell'}^\star}^{(2)}(t)\right| \cdot \left|\bar{\mathbf{v}}_i^\top\left(\bar{\mathbf{v}}_{j_{\ell'}^\star}^\star - \bar{\mathbf{v}}_{i_{\ell'}^\star}\right)\right| + \left(\bar{\mathbf{v}}_i^\top\left(\bar{\mathbf{v}}_{j_{\ell'}^\star}^\star - \bar{\mathbf{v}}_{i_{\ell'}^\star}\right)\right)^2 \\
&\leq 2\left|\gamma_{i,j_{\ell'}^\star}^{(2)}(t)\right| \cdot \left\|\bar{\mathbf{v}}_{j_{\ell'}^\star}^\star - \bar{\mathbf{v}}_{i_{\ell'}^\star}\right\|_2 + \left\|\bar{\mathbf{v}}_{j_{\ell'}^\star}^\star - \bar{\mathbf{v}}_{i_{\ell'}^\star}\right\|_2^2 \\
&\leq \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d^{\frac{3}{4}}}\right)\left|\gamma_{i,j_{\ell'}^\star}^{(2)}(t)\right| + \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)
\end{aligned}$$

Moreover, since $\gamma^{(2)}_{i^\star_{\ell'},j^\star_{\ell'}}(t) \geq 1 - \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)$, for $T_{p,\ell-1} \leq t \leq T_c$, we must have that

$$\begin{aligned}
\frac{d}{dt}\left|\gamma^{(2)}_{i,j^\star_{\ell'}}(t)\right| &\leq \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d^{\frac{3}{4}}}\right)\left|\gamma^{(2)}_{i,j^\star_{\ell'}}(t)\right| + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2 + \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right) \\
&\leq \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d^{\frac{3}{4}}}\right)\left|\gamma^{(2)}_{i,j^\star_{\ell'}}(t)\right| + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2 + \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right) \\
&\leq \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)\frac{d}{dt}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) + \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{5}{4}}}\right)
\end{aligned}$$

where the last inequality follows from Lemma 8. This gives that

$$\begin{aligned}
\left|\gamma^{(2)}_{i,j^\star_{\ell'}}(t+T_{p,\ell-1})\right| &\leq \left|\gamma^{(2)}_{i,j^\star_{\ell'}}(T_{p,\ell-1})\right| + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) + \mathcal{O}\left(\frac{m^7 t}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right) \\
&\leq \left|\gamma^{(2)}_{i,j^\star_{\ell'}}(T_{p,\ell-1})\right| + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d} + \frac{m^7 t}{\delta_{\mathbb{P}}^3 d^{\frac{5}{4}}}\right) \\
&\leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}} + \frac{m^7 t}{\delta_{\mathbb{P}}^3 d^{\frac{5}{4}}}\right)
\end{aligned}$$

where the last step follows from the inductive hypothesis. Further requiring that $t \leq \mathcal{O}\left(\frac{\delta_{\mathbb{P}}^2 d^{\frac{3}{4}}}{m^5}\right)$ keeps $\left|\gamma^{(2)}_{i,j^\star_{\ell'}}(t+T_{p,\ell-1})\right| \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$. When $\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \geq \xi$, we can still obtain that

$$\frac{d}{dt}\left|\gamma^{(2)}_{i,j^\star_{\ell'}}(t)\right| \leq \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)$$

Thus, for all $t \leq \min\left\{T_\ell(\xi) + \mathcal{O}\left(\frac{\delta_{\mathbb{P}}\sqrt{d}}{m^2}\right), \mathcal{O}\left(\frac{\delta_{\mathbb{P}}^2 d^{\frac{3}{4}}}{m^5}\right)\right\}$ we can guarantee that $\left|\gamma^{(2)}_{i,j^\star_{\ell'}}(t+T_{p,\ell-1})\right| \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$. □

Bounding $\varepsilon_{1,\ell}(t), \varepsilon_{2,\ell}(t)$, and $\varepsilon^{(1)}_{\mathcal{B},\ell}(t)$. Here we are going to upper bound $\varepsilon_{1,\ell}(t), \varepsilon_{2,\ell}(t)$ and $\varepsilon^{(1)}_{\mathcal{B},\ell}(t)$. In particular, we are going to analyze $\gamma^{(1)}_{i,j}(t)$ for $i \in [m] \setminus \mathcal{R}_{\ell-1}$ and $j \in [m^\star]$ where $(i,j) \neq i^\star_\ell, j^\star_\ell$, $\zeta^{(1)}_{i,j}, \zeta^{(2)}_{i,j}$ for $i \in [m] \setminus \mathcal{R}_{\ell-1}$ and $j \in [m^\star]$, and also $I^{(1)}_{i,j}, I^{(2)}_{i,j}, I^{(3)}_{i,j}$ for $i, j \in [m]$ and $i \neq j$.

Lemma 11. *Suppose that the inductive hypothesis in Condition 2 and the initialization condition in Condition 1 holds. Let T_c be defined in (14) and $T(\xi)$ in Definition 2. Then there exists some constant $\beta_7 > 0$ such that for all $\beta_8 \leq \mathcal{O}(1)$, for all $t \leq \min\left\{T_c, T\left(\frac{\delta_{\mathbb{P}}^2}{d^{\frac{5}{2}}}\right) + \frac{\beta_7 \delta_{\mathbb{P}}\sqrt{d}}{m}, T(\xi)\right\} + \beta_8$, we have that $\varepsilon^{(2)}_{\mathcal{A},\ell} \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$. Moreover, for all $t \leq \min\left\{\mathcal{O}\left(\frac{\delta_{\mathbb{P}}^2 d}{m^5}\right), T_c\right\}$ we shall have that*

$$\begin{aligned}
&\max\left\{\left|\gamma^{(1)}_{i,j}(t)\right|, \left|\zeta^{(1)}_{i,j}(t)\right|, \left|\zeta^{(2)}_{i,j}(t)\right|\right\} \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right); \forall i \in [m] \setminus \mathcal{R}_\ell, j \in [m^\star] \\
&\max\left\{\left|I^{(1)}_{i,j}(t)\right|, \left|I^{(2)}_{i,j}(t)\right|, \left|I^{(3)}_{i,j}(t)\right|\right\} \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right); \forall i, j \in [m] \setminus \mathcal{R}_\ell, i \neq j
\end{aligned}$$

Proof. Throughout the proof, we will relax the upper bound in terms of $\varepsilon^{(1)}_{\mathcal{A},\ell}(t)$ into the upper bound in terms of $\varepsilon^{(2)}_{\mathcal{A},\ell}(t)$.

Bounding $\zeta^{(1)}_{i,j}(t), \zeta^{(2)}_{i,j}(t)$. First, we are going to derive some rough estimation for $\zeta^{(1)}_{i,j}(t), \zeta^{(1)}_{i,j}(t)$ for $i \in [m] \setminus$

$\mathcal{R}_{\ell-1}, j \in [m^\star]$. By Lemma 7, we have that

$$\begin{aligned}\frac{d}{dt}\left|\zeta_{i,j}^{(1)}(t)\right| &= -\frac{1}{a_i}\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}_j^\star \cdot \operatorname{sign}\left(\zeta_{i,j}^{(1)}(t)\right)\\ &\le \frac{3}{a_i^2}\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,4}\right| + \frac{1}{a_i^2}\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\right| - \hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\left|\zeta_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\right|\mathbb{I}\left\{i=i_\ell^\star\right\}\\ &\quad + \frac{36}{a_i^2}\left|C_{S,2}\gamma_{i,j}^{(2)}(t)I_{i,i}^{(3)}\right| + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2 + \varepsilon_{5,\ell}(t)^2\right)\end{aligned}$$

Diving into the details of the first three terms, we have that

$$\begin{aligned}\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,4}(t)\right| &\le 6\left|I_{i_\ell^\star,i_\ell^\star}^{(3)}(t)\right|\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\sum_{k=0}^{\infty}\frac{c_{k+2}c_k}{k!}\left|\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\right|^k\\ &\quad + 6\left|\zeta_{i,j}^{(2)}(t)\right|\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}\left|\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\right|^k + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\\ &\le \mathcal{O}\left(\varepsilon_{\mathcal{B},\ell}^{(1)}(t) + \varepsilon_{5,\ell}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\\ \left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\right| &\le 6\left|\zeta_{i,j}^{(1)}(t)\right|\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^3\sum_{k=0}^{\infty}\frac{c_kc_{k+2}}{k!}\left|\gamma_{i_\ell^\star,j_{\ell^\star}}^{(1)}(t)\right|^k + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\\ &\le \mathcal{O}\left(\varepsilon_{\mathcal{B},\ell}^{(1)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\\ -\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\left|\zeta_{i,j}^{(1)}(t)\right| &\le 6\left|\zeta_{i,j}^{(1)}(t)\right|\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^3\sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}\left|\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\right|^k\\ &\quad + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\\ &\le \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2 + \varepsilon_{\mathcal{B},\ell}^{(1)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\end{aligned}$$

Also, we notice that $\left|\gamma_{i,j}^{(2)}(t)\right| \le \mathcal{O}\left(\varepsilon_{5,\ell}(t) + \varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right)$. Applying the fact that $a_i^{-1} \le \mathcal{O}(1)$, we have that

$$\begin{aligned}\frac{d}{dt}\left|\zeta_{i,j}^{(1)}(t)\right| &\le \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2 + \varepsilon_{\mathcal{B},\ell}^{(1)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{B},\ell}^{(3)}(t) + \varepsilon_{5,\ell}(t)\right)\\ &\le \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2 + \varepsilon_{\mathcal{B},\ell}^{(1)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\end{aligned}$$

Similarly, for $\zeta_{i,j}^{(2)}(t)$, by Lemma 7, we have that

$$\begin{aligned}\frac{d}{dt}\left|\zeta_{i,j}^{(2)}(t)\right| &= -\frac{1}{b_i}\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_j^\star \cdot \operatorname{sign}\left(\zeta_{i,j}^{(2)}(t)\right)\\ &\le \frac{3}{b_i^2}\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,3}(t)\right| + \frac{3}{b_i^2}\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\right| - 3\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\left|\zeta_{i,j}^{(2)}(t)\right|\mathbb{I}\left\{i=i_\ell^\star\right\}\\ &\quad + \frac{36}{b_i^2}\left|C_{S,2}\gamma_{i,j}^{(1)}(t)I_{i,i}^{(3)}(t)\right| + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{B},\ell}^{(3)}(t)^2 + \varepsilon_{5,\ell}(t)^2\right)\end{aligned}$$

The third term is apparently negative due to the fact that $\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t) \ge 0, \gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) \ge 0$. For the first two, we have that

$$\begin{aligned}\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,3}(t)\right| &\le 6\left|\zeta_{i,j}^{(1)}(t)\right|\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}\left|\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\right|^k + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\\ &\le \mathcal{O}\left(\varepsilon_{\mathcal{B},\ell}^{(1)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\\ \left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\right| &\le 6\left|\zeta_{i,j}^{(1)}(t)\right|\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\sum_{k=0}^{\infty}\frac{c_{k+2}c_k}{k!}\left|\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\right|^{k+1} + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\\ &\le \mathcal{O}\left(\varepsilon_{\mathcal{B},\ell}^{(1)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\end{aligned}$$

As before, we have that $\left|I_{i,i}^{(3)}(t)\right| \leq \mathcal{O}\left(\varepsilon_{\mathcal{B},\ell}^{(3)}(t) + \varepsilon_{5,\ell}(t)\right)$ Therefore, we have that for $t \leq T_c$,

$$\frac{d}{dt}\left|\zeta_{i,j}^{(2)}(t)\right| \leq \mathcal{O}\left(\varepsilon_{\mathcal{B},\ell}^{(1)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$

Bounding $I_{i,j}^{(1)}(t), I_{i,j}^{(2)}(t)$, and $I_{i,j}^{(3)}(t)$ for $i,j \in [m]\setminus\mathcal{R}_{\ell-1}$, and $i \neq j$. Next, we will look at $I_{i,j}^{(1)}$. By Lemma 7, we have that

$$\begin{aligned}
\frac{d}{dt}\left|I_{i,j}^{(1)}(t)\right| &= -\operatorname{sign}\left(I_{i,j}^{(1)}(t)\right)\left(\frac{1}{a_i}\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_j + \frac{1}{a_j}\nabla_{\mathbf{v}_j}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_i\right) \\
&\leq \frac{1}{a_i^2}\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\right|\left(\left|\gamma_{i,j_\ell^\star}^{(1)}(t)\right| + \left|\gamma_{j,j_\ell^\star}^{(1)}(t)\right|\right) - \frac{1}{a_i^2}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\left|I_{i,j}^{(1)}(t)\right|\mathbb{I}\{i = i_\ell^\star \vee j = i_\ell^\star\} \\
&\quad + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{5,\ell}(t) + \varepsilon_{\mathcal{A},\ell}^{(2)}(t)\varepsilon_{\mathcal{B},\ell}^{(3)}(t)\right)
\end{aligned}$$

Since $i \neq j$, for the first term we have that

$$\begin{aligned}
\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\right|\left(\left|\gamma_{i,j_\ell^\star}^{(1)}(t)\right| + \left|\gamma_{j,j_\ell^\star}^{(1)}(t)\right|\right) &\leq \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\right| \\
&\leq \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)
\end{aligned}$$

For the second term, we have that

$$\begin{aligned}
-\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\left|I_{i,j}^{(1)}(t)\right| &\leq 6\left|I_{i,j}^{(1)}(t)\right|\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^3\sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}\left|\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\right|^k \\
&\quad + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) \\
&\leq \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)
\end{aligned}$$

Therefore, we have that

$$\frac{d}{dt}\left|I_{i,j}^{(1)}(t)\right| \leq \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$

For $I_{i,j}^{(2)}(t)$, we have that

$$\begin{aligned}
\frac{d}{dt}\left|I_{i,j}^{(2)}(t)\right| &= -\operatorname{sign}\left(I_{i,j}^{(2)}(t)\right)\left(\frac{1}{b_i}\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}_j + \frac{1}{b_j}\nabla_{\mathbf{w}_j}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)\right) \\
&\leq \frac{9}{b_i^2}\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\right|\left(\left|\gamma_{i,j_\ell^\star}^{(2)}(t)\right| + \left|\gamma_{j,j_\ell^\star}^{(2)}(t)\right|\right) - \frac{9}{b_i^2}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\left|I_{i,j}^{(2)}(t)\right|\mathbb{I}\{i = i_\ell^\star \vee j = i_\ell^\star\} \\
&\quad + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) \\
&\leq \frac{9}{b_i^2}\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\right|\left(\left|\gamma_{i,j_\ell^\star}^{(2)}(t)\right| + \left|\gamma_{j,j_\ell^\star}^{(2)}(t)\right|\right) + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)
\end{aligned}$$

since $\lambda_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) \geq 0$. For the first term, we have that

$$\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\right|\left(\left|\gamma_{i,j_\ell^\star}^{(2)}(t)\right| + \left|\gamma_{j,j_\ell^\star}^{(2)}(t)\right|\right) \leq \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\right| \leq \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$

Therefore, we have that

$$\frac{d}{dt}\left|I_{i,j}^{(2)}(t)\right| \leq \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$

For $I_{i,j}^{(3)}$, by Lemma 7, we also have that

$$\begin{aligned}\frac{d}{dt}\left|I_{i,j}^{(3)}(t)\right| &= -\text{sign}\left(\frac{1}{a_i}\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}_j + \frac{1}{b_j}\nabla_{\mathbf{w}_j}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_i\right)\\ &\le \frac{1}{a_i^2}\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\zeta_{j,j_\ell^\star}^{(2)}(t)\right| - \frac{1}{a_i^2}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\left|I_{i,j}^{(3)}(t)\right|\mathbb{I}\left\{i=i_\ell^\star\right\}\\ &\quad + \frac{9}{b_j^2}\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\zeta_{i,j_\ell^\star}^{(1)}(t)\right| - \frac{9}{b_j^2}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\left|I_{i,j}^{(3)}(t)\right|\mathbb{I}\left\{j=i_\ell^\star\right\}\\ &\quad + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\end{aligned}$$

where for the first, second, and third term, we have

$$\begin{aligned}\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\zeta_{j,j_\ell^\star}^{(2)}(t)\right| &\le \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\\ \left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\zeta_{j,j_\ell^\star}^{(1)}(t)\right| &\le \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\\ -\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\left|I_{i,j}^{(3)}(t)\right| &\le 6\left|I_{i,j}^{(3)}(t)\right|\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^3\sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}\left|\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\right|^k\\ &\quad + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\\ &\le \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\end{aligned}$$

Noticing that $\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t), \gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) \ge 0$ for all $t \le T_c$, we have that

$$\frac{d}{dt}\left|I_{i,j}^{(3)}(t)\right| \le \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)$$

Bounding $\gamma_{i,j}^{(1)}(t)$. Finally, for $\gamma_{i,j}^{(1)}(t)$, in the case of $(i,j) \ne (i_\ell^\star, j_\ell^\star)$, we have that

$$\begin{aligned}\frac{d}{dt}\left|\gamma_{i,j}^{(1)}(t)\right| &= -\frac{1}{a_i}\text{sign}\left(\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_j^\star\right)\\ &\le -\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\left|\gamma_{i,j}^{(1)}(t)\right|\mathbb{I}\left\{i=i_\ell^\star\right\} + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)\end{aligned}$$

where since $(i,j) \ne (i_\ell^\star, j_\ell^\star)$, we must have that

$$-\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\left|\gamma_{i,j}^{(1)}(t)\right| \le \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$

Therefore

$$\frac{d}{dt}\left|\gamma_{i,j}^{(1)}(t)\right| \le \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$

Bounding $I_{i,j}^{(1)}(t), I_{i,j}^{(2)}(t)$, and $I_{i,j}^{(3)}(t)$ for $i,j \in [m] \setminus \mathcal{R}_{\ell-1}$ for $i \in \mathcal{R}_{\ell-1}$ or $j \in \mathcal{R}_{\ell-1}$. Recall the definition of $I_{i,j}^{(1)}(t), I_{i,j}^{(2)}(t)$, and $I_{i,j}^{(3)}(t)$

$$I_{i,j}^{(1)}(t) = \bar{\mathbf{v}}_i(t)^\top\bar{\mathbf{v}}_j(t); I_{i,j}^{(2)}(t) = \bar{\mathbf{w}}_i(t)^\top\bar{\mathbf{w}}_j(t); I_{i,j}^{(3)}(t) = \bar{\mathbf{v}}_i(t)^\top\bar{\mathbf{w}}_j(t)$$

For the sake of convenience, we let $i \in \mathcal{R}_{\ell-1}$ and study $I_{i,j}^{(1)}(t), I_{i,j}^{(2)}(t)$, and $I_{i,j}^{(3)}(t), I_{j,i}^{(3)}(t)$. Since $i \in \mathcal{R}_{\ell-1}$, there exists $\ell' < \ell$ such that $i = i_{\ell'}^\star$. Therefore, for $I_{i,j}^{(1)}(t)$, when $t \ge T_{p,\ell-1}$, we have that

$$\begin{aligned}\left|I_{i,j}^{(1)}(t)\right| &= \left|\bar{\mathbf{v}}_j(t)^\top\bar{\mathbf{v}}_{j_{\ell'}^\star}^\star + \bar{\mathbf{v}}_j(t)^\top\left(\bar{\mathbf{v}}_{i_{\ell'}^\star}(t) - \bar{\mathbf{v}}_{j_{\ell'}^\star}^\star\right)\right|\\ &\le \left|\gamma_{j,j_{\ell'}^\star}^{(1)}(t)\right| + \left\|\bar{\mathbf{v}}_{i_{\ell'}^\star}(t) - \bar{\mathbf{v}}_{j_{\ell'}^\star}(t)\right\|_2\\ &\le \left|\gamma_{j,j_{\ell'}^\star}^{(1)}(t)\right| + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d^{\frac{3}{4}}}\right)\end{aligned} \tag{17}$$

for all $t \geq T_{p,\ell-1}$. Similarly, for $I_{i,j}^{(2)}(t)$, when $t \geq T_{p,\ell-1}$, we have that

$$\begin{aligned}\left|I_{i,j}^{(2)}(t)\right| &= \left|\bar{\mathbf{w}}_j(t)^\top \bar{\mathbf{w}}_{j_{\ell'}^\star}^\star + \bar{\mathbf{w}}_j(t)^\top \left(\bar{\mathbf{w}}_{i_{\ell'}^\star}(t) - \bar{\mathbf{w}}_{j_{\ell'}^\star}^\star\right)\right| \\ &\leq \left|\gamma_{j,j_{\ell'}^\star}^{(2)}(t)\right| + \left\|\bar{\mathbf{w}}_{i_{\ell'}^\star}(t) - \bar{\mathbf{w}}_{j_{\ell'}^\star}(t)\right\|_2 \\ &\leq \left|\gamma_{j,j_{\ell'}^\star}^{(2)}(t)\right| + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d^{\frac{3}{4}}}\right)\end{aligned} \tag{18}$$

For $I_{i,j}^{(3)}(t)$ and $I_{j,i}^{(3)}(t)$, when $t \geq T_{p,\ell-1}$, we have that

$$\begin{aligned}\left|I_{i,j}^{(3)}(t)\right| &= \left|\bar{\mathbf{w}}_j(t)^\top \bar{\mathbf{v}}_{j_{\ell'}^\star}^\star + \bar{\mathbf{w}}_j(t)^\top \left(\bar{\mathbf{v}}_{i_{\ell'}^\star}(t) - \bar{\mathbf{v}}_{j_{\ell'}^\star}^\star\right)\right| \\ &\leq \left|\zeta_{j,j_{\ell'}^\star}^{(2)}(t)\right| + \left\|\bar{\mathbf{v}}_{i_{\ell'}^\star}(t) - \bar{\mathbf{v}}_{j_{\ell'}^\star}(t)\right\|_2 \\ &\leq \left|\zeta_{j,j_{\ell'}^\star}^{(2)}(t)\right| + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d^{\frac{3}{4}}}\right) \\ \left|I_{j,i}^{(3)}(t)\right| &= \left|\bar{\mathbf{v}}_i(t)^\top \bar{\mathbf{w}}_{j_{\ell'}^\star}^\star + \bar{\mathbf{v}}_i(t)^\top \left(\bar{\mathbf{w}}_{i_{\ell'}^\star}(t) - \bar{\mathbf{w}}_{j_{\ell'}^\star}^\star\right)\right| \\ &\leq \left|\zeta_{i,j_{\ell'}^\star}^{(1)}(t)\right| + \left\|\bar{\mathbf{w}}_{i_{\ell'}^\star}(t) - \bar{\mathbf{w}}_{j_{\ell'}^\star}(t)\right\|_2 \\ &\leq \left|\zeta_{i,j_{\ell'}^\star}^{(1)}(t)\right| + \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d^{\frac{3}{4}}}\right)\end{aligned} \tag{19}$$

Moreover, by the inductive hypothesis, we have that for $t \leq T_{p,\ell-1}$

$$\left|I_{i,j}^{(1)}(t)\right|, \left|I_{i,j}^{(2)}(t)\right|, \left|I_{i,j}^{(3)}(t)\right|, \left|I_{j,i}^{(3)}(t)\right| \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$$

Gathering the results. Defining $\hat{\varepsilon}_\ell(t)$ to be the maximum of $\left|\zeta_{i,j}^{(1)}(t)\right|, \left|\zeta_{i,j}^{(2)}(t)\right|, \left|\gamma_{i,j}^{(1)}(t)\right|$ for i,j bounded above excluding $(i,j) = (i_\ell^\star, j_\ell^\star$ for $\gamma_{i,j}^{(1)}(t)$, and $\left|I_{i,j}^{(1)}(t)\right|, \left|I_{i,j}^{(1)}(t)\right|, \left|I_{i,j}^{(1)}(t)\right|$ for i,j bounded above excluding $i \in \mathcal{R}_{\ell-1}$ or $j \in \mathcal{R}_{\ell-1}$. Gathering the results, we have that

$$\frac{d}{dt}\hat{\varepsilon}_\ell(t) \leq \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$

for $t \leq T_c$. In the meantime, by Lemma 9 we have that

$$\left|\gamma_{i,j}^{(2)}(t)\right| \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right); \forall t \leq \min\left\{T_c, T\left(\frac{\delta_{\mathbb{P}}^2}{d^{\frac{2}{5}}}\right) + \frac{\beta_7 \delta_{\mathbb{P}}\sqrt{d}}{m^2}\right\}$$

where $i \in [m] \setminus \mathcal{R}_{\ell-1}, j \in [m^\star] \setminus \mathcal{C}_{\ell-1}$ and $(i,j) \neq (i_\ell^\star, j_\ell^\star)$. Therefore, we have that

$$\begin{aligned}\frac{d}{dt}\hat{\varepsilon}_\ell(t) &\leq \mathcal{O}\left(\hat{\varepsilon}_\ell(t) + \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) \\ &\leq \mathcal{O}\left(\hat{\varepsilon}_\ell(t) + \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)\right)\frac{d}{dt}\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right) \\ &\leq \mathcal{O}\left(\hat{\varepsilon}_\ell(t) + \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)\right)\frac{d}{dt}\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) + \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)\end{aligned}$$

where the last inequality follows from Lemma 8. Solving the differential equation gives

$$\hat{\varepsilon}_\ell(t) \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)\hat{\varepsilon}_\ell(0)\exp\left(\mathcal{O}\left(\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\right)\right) + \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)\exp\left(\mathcal{O}\left(\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\right)\right)\cdot t$$

Imposing $\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \le 1$ gives the desired result.

Fine-Grained Result for $i \in [m] \setminus \mathcal{R}_\ell$. By Lemma 7, we have that for all $i \in [m] \setminus \mathcal{R}_\ell$ and $j \in [m^\star]$, it holds that

$$\max\left\{\left|\gamma^{(1)}_{i,j}(t)\right|, \left|\zeta^{(1)}_{i,j}(t)\right|, \left|\zeta^{(2)}_{i,j}(t)\right|\right\} \le \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(2)}_{\mathcal{A},\ell}(t)\varepsilon^{(3)}_{\mathcal{B},\ell}(t)\right) \le \mathcal{O}\left(\frac{m^7}{\delta^2_{\mathbb{P}} d^{\frac{3}{2}}}\right)$$

For $i, j \in [m] \setminus \mathcal{R}_\ell$ with $i \neq j$, we have that

$$\max\left\{\left|I^{(1)}_{i,j}(t)\right|, \left|I^{(2)}_{i,j}(t)\right|, \left|I^{(3)}_{i,j}(t)\right|\right\} \le \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(2)}_{\mathcal{A},\ell}(t)\varepsilon^{(3)}_{\mathcal{B},\ell}(t)\right) \le \mathcal{O}\left(\frac{m^7}{\delta^3_{\mathbb{P}} d^{\frac{3}{2}}}\right)$$

Thus, for all $t \le \mathcal{O}\left(\frac{\delta^2_{\mathbb{P}} d}{m^5}\right)$ we shall have that

$$\max\left\{\left|\gamma^{(1)}_{i,j}(t)\right|, \left|\zeta^{(1)}_{i,j}(t)\right|, \left|\zeta^{(2)}_{i,j}(t)\right|\right\} \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right); \forall i \in [m] \setminus \mathcal{R}_\ell, j \in [m^\star]$$
$$\max\left\{\left|I^{(1)}_{i,j}(t)\right|, \left|I^{(2)}_{i,j}(t)\right|, \left|I^{(3)}_{i,j}(t)\right|\right\} \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right); \forall i, j \in [m] \setminus \mathcal{R}_\ell, i \neq j$$

□

Bounding $I^{(3)}_{i,i}(t)$. Here we are going to upper bound $I^{(3)}_{i,i}(t)$ for $i \in [m] \setminus \mathcal{R}_{\ell-1}$

Lemma 12. *Suppose that the inductive hypothesis in Condition 2, and the initialization condition in Condition 1. Let T_c be defined in (14) and $T(\xi)$ in Definition 2. Then for all $t \le T_c$, we have that*

$$\left|I^{(3)}_{i,i}(t)\right| \le \begin{cases} \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right) & \text{if } i \neq i^\star_\ell \\ \mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2 + \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right) & \text{if } i = i^\star_\ell \end{cases}$$

Proof. By Lemma 7, we have that, in the case where $i \neq i^\star_\ell$

$$\begin{aligned}\frac{d}{dt}\left|I^{(3)}_{i,i}(t)\right| &= -\text{sign}\left(I^{(3)}_{i,i}(t)\right)\left(\frac{1}{a_i}\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}_j + \frac{1}{b_i}\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_j\right) \\ &\le -36C_{S,2}\left(\frac{1}{a_i^2}+\frac{1}{b_i^2}\right)\left|I^{(3)}_{i,i}(t)\right| + \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)\end{aligned}$$

Since $\varepsilon^{(2)}_{\mathcal{A},\ell}(t)$ is monotonic non-decreasin, we must have that $\left|I^{(3)}_{i,i}(t)\right| \le \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)$ for all $t \ge 0$. Similarly, in the case $i = i^\star_\ell$, we have that

$$\begin{aligned}\frac{d}{dt}\left|I^{(3)}_{i,i}(t)\right| &= -\text{sign}\left(I^{(3)}_{i,i}(t)\right)\left(\frac{1}{a_i}\nabla_{\mathbf{v}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{w}}_j + \frac{1}{b_i}\nabla_{\mathbf{w}_i}\mathcal{L}\left(\boldsymbol{\theta}(t)\right)^\top \bar{\mathbf{v}}_j\right) \\ &= \frac{1}{a_i^2}\left|\hat{\lambda}_{i^\star_\ell,j^\star_\ell,1}(t)\zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t)\right| + \frac{9}{b_i^2}\left|\hat{\lambda}_{i^\star_\ell,j^\star_\ell,5}(t)\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\right| + 3\left|\hat{\lambda}_{i^\star_\ell,j^\star_\ell,2}(t)\right|\left(\frac{1}{a_i^2}+\frac{1}{b_i^2}\right) \\ &\quad + \frac{3}{a_i^2}\text{sign}\left(I^{(3)}_{i,i}(t)\right)\hat{\lambda}_{i^\star_\ell,j^\star_\ell,4}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) + \frac{3}{b_i^2}\left|\hat{\lambda}_{i^\star_\ell,j^\star_\ell,3}(t)\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\right| \\ &\quad - 36C_{S,2}\left(\frac{1}{a_i^2}+\frac{1}{b_i^2}\right)\left|I^{(3)}_{i,i}(t)\right| - \frac{1}{a_i^2}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,1}(t)\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}\left|I^{(3)}_{i,i}(t)\right| \\ &\quad - \frac{9}{b_i^2}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,5}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}\left|I^{(3)}_{i,i}(t)\right| + \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)\end{aligned}$$

As in the previous proof, we have that

$$\begin{aligned}\left|\hat{\lambda}_{i^\star_\ell,j^\star_\ell,3}(t)\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\right| &\le 6\left|\zeta^{(1)}_{i,j}(t)\right|\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2\sum_{k=0}^{\infty}\frac{c_k c_{k+2}}{k!}\left|\gamma^{(1)}_{i^\star_\ell,j_{\ell^\star}}(t)\right|^{k+1} + \mathcal{O}\left(\varepsilon_1(t)^3\right) \\ &\le \mathcal{O}\left(\varepsilon^{(1)}_{\mathcal{B},\ell}(t)\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2 + \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)\end{aligned}$$

The second and third term requires a more careful analysis. First, by the definition of $\hat{\lambda}_{i_\ell^\star,j_\ell^\star,4}$, we have that

$$\begin{aligned}
&\hat{\lambda}_{i_\ell^\star,j_\ell^\star,4}(t)\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)\cdot \operatorname{sign}\left(I^{(3)}_{i,i}(t)\right)\\
&\quad\le 18\left|I^{(3)}_{i,i}(t)\right|\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)^3\sum_{k=0}^{\infty}\frac{c_{k+2}c_k}{k!}\gamma^{(1)}_{i_\ell^\star,j_\ell^\star}(t)^k+18\left|\zeta^{(2)}_{i,j}(t)\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)^3\sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}\gamma^{(1)}_{i_\ell^\star,j_\ell^\star}(t)^k\right|\\
&\quad\le 18\left|I^{(3)}_{i,i}(t)\right|\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)^3\sum_{k=0}^{\infty}\frac{c_{k+2}c_k}{k!}\gamma^{(1)}_{i_\ell^\star,j_\ell^\star}(t)^k+\mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right)\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)^3\\
&\quad\le \mathcal{O}\left(\frac{1}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right)\left|I^{(3)}_{i,i}(t)\right|+\mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right)\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)^3
\end{aligned}$$

Moreover, by the definition of $\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}$, we have that

$$\begin{aligned}
-\lambda_{i_\ell^\star,j_\ell^\star,1}(t)\gamma^{(1)}_{i_\ell^\star,j_\ell^\star}&=-6\sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}\gamma^{(1)}_{i_\ell^\star,j_\ell^\star}(t)^{k+1}\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)^3\\
&=-6\gamma^{(1)}_{i_\ell^\star,j_\ell^\star}(t)\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)^3\left(c_1^2\pm\mathcal{O}\left(\frac{1}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right)\right)\\
&\le\mathcal{O}\left(\frac{1}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right)
\end{aligned}$$

Recalling that $\lambda_{i_\ell^\star,j_\ell^\star,5}(t)\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}\ge 0$ for all t, we have that

$$\begin{aligned}
\frac{d}{dt}\left|I^{(3)}_{i,i}(t)\right|&\le-\left(36C_{S,2}\left(\frac{1}{a_i^2}+\frac{1}{b_i^2}\right)-\mathcal{O}\left(\frac{1}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right)\right)\left|I^{(3)}_{i,i}(t)\right|\\
&\quad+\mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right)\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)^2+\mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)\\
&\le-35C_{S,2}\left(\frac{1}{a_i^2}+\frac{1}{b_i^2}\right)\left|I^{(3)}_{i,i}(t)\right|+\mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right)\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)^2+\mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)
\end{aligned}$$

Since $\varepsilon^{(2)}_{\mathcal{A},\ell}(t)$ is monotonically non-decreasing, we must have that $I^{(3)}_{i,i}\le\mathcal{O}\left(\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right)\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)^2+\mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right)$ for all $t\ge 0$. □

With the above preparation work, we are ready to derive the result for phase 1 convergence.

Lemma 13. *Suppose that the inductive hypothesis in Condition 2 and the initialization condition in Condition 1 holds. Let T_c be defined in (14) and $T(\xi)$ be defined in Definition 2. Then there exists constant $\beta_7\ge 0$ such that for all ξ satisfying $(1-\xi^2)^{-1}\le\mathcal{O}(1)$ and all constant $\beta_8\ge 0$, it holds that $T_c\ge\min\left\{T\left(\frac{\delta_{\mathbb{P}}^2}{d^{\frac{2}{5}}}\right)+\frac{\beta_7\delta_{\mathbb{P}}\sqrt{d}}{m^2},T(\xi)\right\}+\beta_8$, and there exists $T_1\le\min\left\{T_c,\mathcal{O}\left(\sqrt{d}\right)\right\}$ such that for all $T_1\le t\le T_c$ we have that $\gamma^{(1)}_{i_\ell^\star,j_\ell^\star}(t),\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)\ge 0.9$.*

Proof. By Lemma 10, Lemma 11 and Lemma 12, we have that $T_c\ge\min\left\{T\left(\frac{\delta_{\mathbb{P}}}{d^{\frac{2}{5}}}\right)+\frac{\beta_7\delta_{\mathbb{P}}\sqrt{d}}{m},T(\xi)\right\}+\beta_8$. Thus, it remains to show that there exists some $T_1\le T_c$ such that $\gamma^{(1)}_{i_\ell^\star,j_\ell^\star}(t),\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}(t)\ge 0.9$. To start, we choose $\xi=\frac{1}{2}$. By Lemma 8 we have that

$$\begin{aligned}
\frac{1}{2}&=\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}\left(T\left(\frac{1}{2}\right)\right)\\
&\ge\left(\gamma^{(2)}_{i_\ell^\star,j_\ell^\star}\left(T\left(\frac{\delta_{\mathbb{P}}}{d^{\frac{2}{5}}}\right)\right)^{-1}-\frac{18}{b_{i_\ell^\star}}\left(c_0^2\left(1-\frac{1}{4}\right)-\mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3\sqrt{d}}\right)\right)\left(T\left(\frac{1}{2}\right)-T\left(\frac{\delta_{\mathbb{P}}}{d^{\frac{2}{5}}}\right)\right)\right)^{-1}
\end{aligned}$$

This gives that

$$T\left(\frac{1}{2}\right) \leq \mathcal{O}\left(\frac{d^{\frac{2}{5}}}{\delta_{\mathbb{P}}^2}\right) + T\left(\frac{\delta_{\mathbb{P}}}{d^{\frac{2}{5}}}\right) \leq T\left(\frac{\delta_{\mathbb{P}}}{d^{\frac{2}{5}}}\right) + \frac{\beta_7 \delta_{\mathbb{P}}\sqrt{d}}{m}$$

when $d \geq \delta_{\mathbb{P}}^{10} m^5$. Thus, it suffice to show that a $T_1 \leq T\left(\frac{1}{2}\right) + \beta_8$ achieves $\gamma_{i_\ell^\star, j_\ell^\star}^{(1)}(t), \gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t) \geq 0.9$ for some constant β_8. Notice that this choice of T_1 satisfies $T_1 \leq \mathcal{O}\left(\sqrt{d}\right)$ That is, if we define

$$T_{1,1} = \min\left\{t \geq 0 : \gamma_{i_\ell^\star, j_\ell^\star}^{(1)}(t) \geq 0.9\right\}; T_{1,2} = \min\left\{t \geq 0 : \gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t) \geq 0.9\right\}$$

Then it suffice to show that $T_{1,1}, T_{1,2} \leq T\left(\frac{1}{2}\right) + \beta_8$. By Lemma 7, we have that for $T\left(\frac{1}{2}\right) \leq t \leq \min\left\{T_{1,2}, T\left(\frac{1}{2}\right) + \beta_8\right\}$, it holds that

$$\begin{aligned}
\frac{d}{dt}\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t) &\geq \gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)^2\left(1 - \gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)^2\right)\sum_{k=0}^{\infty}\frac{c_k^2}{k!}\gamma_{i_\ell^\star, j_\ell^\star}^{(1)}(t)^k - \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right) \\
&\geq \gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)^2\left(1 - \gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)^2\right)\left(c_0^2 - \mathcal{O}\left(\frac{1}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right)\right) - \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right) \\
&\geq \frac{c_0^2}{2}\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)^2\left(1 - \gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)^2\right) - \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)
\end{aligned}$$

This shows that $\frac{d}{dt}\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t) \geq 0$ as long as $1 - \gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)^2 \geq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$. Thus, we can conclude that $\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t) \geq \frac{1}{2}$ for all $T\left(\frac{1}{2}\right) \leq t \leq \min\left\{T_{1,2}, T\left(\frac{1}{2}\right) + \beta_8\right\}$. For the dynamic of $\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)$, this implies that for $T\left(\frac{1}{2}\right) \leq t \leq \min\left\{T_{1,2}, T\left(\frac{1}{2}\right) + \beta_8\right\}$

$$\frac{d}{dt}\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t) \geq \frac{c_0^2}{42} - \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right) \geq \frac{c_0^2}{50}$$

Thus $\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}\left(t + T\left(\frac{1}{2}\right)\right) \geq \gamma_{i_\ell^\star, j_\ell^\star}^{(2)}\left(T\left(\frac{1}{2}\right)\right) + \frac{c_0^2}{50} \cdot t$, which implies that $T_{1,2} \leq T\left(\frac{1}{2}\right) + \frac{20}{c_0^2} \leq T\left(\frac{1}{2}\right) + \beta_8$ for some $\beta_8 \geq 0$. Next, for the dynamic of $\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)$, we have that

$$\begin{aligned}
\frac{d}{dt}\gamma_{i_\ell^\star, j_\ell^\star}^{(1)}(t) &\geq 6\left(1 - \gamma_{i_\ell^\star, j_\ell^\star}^{(1)}(t)^2\right)\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)^3\sum_{k=0}^{\infty}\frac{c_k^2}{k!}\gamma_{i_\ell^\star, j_\ell^\star}^{(1)}(t)^k - \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right) \\
&\geq \frac{3}{4}\left(1 - \gamma_{i_\ell^\star, j_\ell^\star}^{(1)}(t)^2\right)\left(c_1^2 - \mathcal{O}\left(\frac{1}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right)\right) - \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right) \\
&\geq \frac{c_1^2}{8}
\end{aligned}$$

Thus $\gamma_{i_\ell^\star, j_\ell^\star}^{(1)}\left(t + T\left(\frac{1}{2}\right)\right) \geq \gamma_{i_\ell^\star, j_\ell^\star}^{(1)}\left(T\left(\frac{1}{2}\right)\right) + \frac{c_1^2}{8} \cdot t$, which implies that $T_{1,1} \leq T\left(\frac{1}{2}\right) + \frac{16}{c_1^2}$ since $\gamma_{i_\ell^\star, j_\ell^\star}^{(1)}\left(T\left(\frac{1}{2}\right)\right) \geq -\mathcal{O}\left(\frac{1}{\sqrt{d}}\left(\log\frac{m}{\delta_{\mathbb{P}}}\right)^{\frac{1}{2}}\right)$ by Lemma 8. Therefore, we can conclude that $T_{1,1}, T_{1,2} \leq T\left(\frac{1}{2}\right) + \beta_8$ for some constant β_8. Now, we consider the dynamic of $\gamma_{i_\ell^\star, j_\ell^\star}^{(1)}(t)$ and $\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)$ for t $geq \max\{T_{1,1}, T_{1,2}\}$. As before, we have that

$$\begin{aligned}
\frac{d}{dt}\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t) &\geq \frac{c_1^2}{2}\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)^3\left(1 - \gamma_{i_\ell^\star, j_\ell^\star}^{(1)}(t)^2\right) - \mathcal{O}\left(\frac{m^2}{\delta_{\delta_{\mathbb{P}}\sqrt{d}}}\right) \\
\frac{d}{dt}\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t) &\geq \frac{c_0^2}{2}\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)^2\left(1 - \gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)^2\right) - \mathcal{O}\left(\frac{m^2}{\delta_{\delta_{\mathbb{P}}\sqrt{d}}}\right)
\end{aligned}$$

We can observe that $\frac{d}{dt}\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t), \frac{d}{dt}\gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t)$ if

$$\gamma_{i_\ell^\star, j_\ell^\star}^{(1)}(t), \gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t) \leq 1 - \mathcal{O}\left(\frac{m^2}{\delta_{\delta_{\mathbb{P}}\sqrt{d}}}\right)$$

Thus, for all $T_1 \leq t \leq T_c$ we have that $\gamma_{i_\ell^\star, j_\ell^\star}^{(1)}(t), \gamma_{i_\ell^\star, j_\ell^\star}^{(2)}(t) \geq 0.9$ □

A.6 Establishing the Inductive Hypothesis: Phase 2

A.7 Phase 2: Growth to Near-Perfect Alignment.

In this section, our goal is to show complete the inductive hypothesis.

The decay of $\left|\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\right|, \left|\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\right|$ and $\left|I^{(3)}_{i^\star_\ell,j^\star_\ell}(t)\right|$.

Lemma 14. *Suppose that the inductive hypothesis in Condition 2, the initialization condition in Condition 1, and the condition on the sigmoid function in Assumption 3 holds. Then for $t \geq T_1$, where T_1 is defined in Lemma 13, we have that*

$$\max\left\{\left|\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\right|, \left|\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\right|, \left|I^{(3)}_{i^\star_\ell,i^\star_\ell}(t)\right|\right\} \leq \begin{cases} \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2 + \frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right) & \text{for } T_1 \leq t \leq T_c \\ \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2 + \frac{m^7}{\delta^3_{\mathbb{P}} d^{\frac{3}{2}}}\right) & \text{for } T_1 + \mathcal{O}\left(\log d\right) \leq t \leq T_c \end{cases}$$

Proof. To ease the analysis in this section, we are going to define

$$Q_0(t) = \gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2 \sum_{k=0}^{\infty} \frac{c_k^2}{k!} \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)^k;$$
$$Q_1(t) = \gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2 \sum_{k=0}^{\infty} \frac{c_{k+1}^2}{k!} \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)^k;$$
$$Q_2(t) = \gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2 \sum_{k=0}^{\infty} \frac{c_k c_{k+2}}{k!} \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)^k$$

By Lemma 5, we have that

$$\hat{\lambda}_{i^\star_\ell,j^\star_\ell,1}(t) = 6\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_1(t) \pm \mathcal{O}\left(\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2 \pm \mathcal{O}\left(\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right)$$
$$\hat{\lambda}_{i^\star_\ell,j^\star_\ell,2}(t) = 6\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t)Q_2(t) \pm \mathcal{O}\left(\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right)$$
$$\hat{\lambda}_{i^\star_\ell,j^\star_\ell,3}(t) = 6\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t)Q_1(t) \pm \mathcal{O}\left(\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right)$$
$$\hat{\lambda}_{i^\star_\ell,j^\star_\ell,4}(t) = 6I^{(3)}_{i^\star_\ell,j^\star_\ell}(t)Q_2(t) + 6\zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_1(t) \pm \mathcal{O}\left(\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right)$$
$$\hat{\lambda}_{i^\star_\ell,j^\star_\ell,5}(t) = 2Q_0(t) \pm \mathcal{O}\left(\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \pm \mathcal{O}\left(\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right)$$

We start with a detailed analysis of $I^{(3)}_{i^\star_\ell,j^\star_\ell}(t)$. Recall that

$$\begin{aligned}
\frac{d}{dt}I^{(3)}_{i^\star_\ell,j^\star_\ell}(t) =& \frac{1}{a_i^2}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,1}(t)\zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t) + \frac{9}{b_i^2}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,5}(t)\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t) + 3\hat{\lambda}_{i^\star_\ell,j^\star_\ell,2}(t)\left(\frac{1}{a_i^2}+\frac{1}{b_i^2}\right) \\
&+ \frac{3}{a_i^2}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,4}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) + \frac{3}{b_i^2}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,3}(t)\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t) \\
&- 36C_{S,2}\left(\frac{1}{a_i^2}+\frac{1}{b_i^2}\right)I^{(3)}_{i^\star_\ell,i^\star_\ell}(t) - \frac{1}{a_i^2}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,1}(t)\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)I^{(3)}_{i^\star_\ell,i^\star_\ell}(t) \\
&- \frac{9}{b_i^2}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,5}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)I^{(3)}_{i^\star_\ell,i^\star_\ell}(t) \pm \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right) \\
=& \frac{6}{a_i^2}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_1(t)\zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t) + \frac{18}{b_i^2}Q_0(t)\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t) + 18Q_2(t)\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\left(\frac{1}{a_i^2}+\frac{1}{b_i^2}\right) \\
&+ \frac{18}{a_i^2}Q_2(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)I^{(3)}_{i^\star_\ell,j^\star_\ell}(t) + \frac{18}{a_i^2}Q_1(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)\zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t) + \frac{18}{b_i^2}\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)Q_1(t)\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t) \\
&- 36C_{S,2}\left(\frac{1}{a_i^2}+\frac{1}{b_i^2}\right)I^{(3)}_{i^\star_\ell,i^\star_\ell}(t) - \frac{6}{a_i^2}\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_1(t)I^{(3)}_{i^\star_\ell,i^\star_\ell}(t) \\
&- \frac{18}{b_i^2}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_0(t)I^{(3)}_{i^\star_\ell,i^\star_\ell}(t) \pm \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right)
\end{aligned}$$

Next, we will look into $\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t), \zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t)$. In particular, for $\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t)$, we have that

$$\begin{aligned}
\frac{d}{dt}\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t) &= \frac{3}{a_i^2}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,4} + \frac{1}{a_i^2}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,2}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) - \frac{1}{a_i^2}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,1}(t)\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t) \\
&\quad - \frac{36}{a_i^2}C_{S,2}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)I^{(3)}_{i^\star_\ell,j^\star_\ell}(t) \pm \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right) \\
&= \frac{18}{a_i^2}Q_2(t)I^{(3)}_{i^\star_\ell,j^\star_\ell}(t) + \frac{18}{a_i^2}Q_1(t)\zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t) + \frac{6}{a_i^2}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_2(t)\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t) \\
&\quad - \frac{6}{a_i^2}\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_1(t)\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t) - \frac{36}{a_i^2}C_{S,2}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)I^{(3)}_{i^\star_\ell,j^\star_\ell}(t) \pm \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right)
\end{aligned}$$

Similarly, for $\zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t)$, we have that

$$\begin{aligned}
\frac{d}{dt}\zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t) &= \frac{3}{b_i^2}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,3}(t) + \frac{3}{b_i^2}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,2}(t)\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t) - \frac{9}{b_i^2}\hat{\lambda}_{i^\star_\ell,j^\star_\ell,5}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)\zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \\
&\quad - \frac{36}{b_i^2}C_{S,2}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)I^{(3)}_{i^\star_\ell,j^\star_\ell}(t) \pm \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right) \\
&= \frac{18}{b_i^2}Q_1(t)\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t) + \frac{18}{b_i^2}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_2(t)\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t) - \frac{18}{b_i^2}Q_0(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)\zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \\
&\quad - \frac{36}{b_i^2}C_{S,2}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)I^{(3)}_{i^\star_\ell,j^\star_\ell}(t) \pm \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right)
\end{aligned}$$

Then we can write $\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t), \zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t)$, and $I^{(3)}_{i^\star_\ell,j^\star_\ell}(t)$ into a system given by

$$\begin{aligned}
\frac{d}{dt}I^{(3)}_{i^\star_\ell,j^\star_\ell}(t) &= -\alpha_1 I^{(3)}_{i^\star_\ell,j^\star_\ell}(t) + \iota_1\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t) + \rho_1\zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \pm \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right) \\
\frac{d}{dt}\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t) &= -\alpha_2 I^{(3)}_{i^\star_\ell,j^\star_\ell}(t) - \iota_2\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t) + \rho_2\zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \pm \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right) \\
\frac{d}{dt}\zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t) &= -\alpha_3 I^{(3)}_{i^\star_\ell,j^\star_\ell}(t) + \iota_3\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t) + \rho_3\zeta^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \pm \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right)
\end{aligned}$$

where

$$\begin{aligned}
\alpha_1 &= 36C_{S,2}\left(\frac{1}{a_i^2}+\frac{1}{b_i^2}\right) + \frac{6}{a_i^2}\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_1(t) + \frac{18}{b_i^2}\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)Q_1(t) - \frac{18}{a_i^2}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_2(t) \\
\alpha_2 &= \frac{36}{a_i^2}C_{S,2}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) - \frac{18}{a_i^2}Q_2(t); \quad \alpha_3 = \frac{36}{b_i^2}C_{S,2}\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \\
\iota_1 &= \frac{18}{b_i^2}\left(Q_0(t) + \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)Q_1(t)\right) + 18\left(\frac{1}{a_i^2}+\frac{1}{b_i^2}\right)Q_2(t); \\
\iota_2 &= \frac{6}{a_i^2}\left(\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_1(t) - \gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_2(t)\right) \\
\iota_3 &= \frac{18}{b_i^2}\left(Q_1(t) + \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)Q_2(t)\right); \quad \rho_1 = \frac{24}{a_i^2}\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)Q_1(t) \\
\rho_2 &= \frac{18}{a_i^2}Q_1(t); \quad \rho_3 = \frac{18}{b_i^2}Q_0(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)
\end{aligned}$$

□

By Lemma 28, we first need to check that $\iota_2\rho_3 \geq \iota_3\rho_2$

$$\begin{aligned}
\iota_2\rho_3 - \iota_3\rho_2 &= \frac{108}{a_i^2b_i^2}\left(\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_1(t) - \gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_2(t)\right)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)Q_0(t) \\
&\quad - \frac{324}{a_i^2b_i^2}\left(Q_1(t) + \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)Q_2(t)\right)Q_1(t)
\end{aligned}$$

By Lemma 29, we have that $Q_2(t) \le 0$ for all $\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t), \gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \ge 0$. Therefore

$$\iota_2\rho_3 - \iota_3\rho_2 \ge \frac{108}{a_i^2 b_i^2} Q_1(t) \left(\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^2 Q_0(t) - 3Q_1(t)\right)$$

Thus the condition holds if $Q_0(t) \ge 4.12 Q_1(t)$ given that $\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t), \gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \ge 0.9$, which translates to

$$\sum_{k=0}^{\infty} \frac{c_k^2}{k!} \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)^k \ge 4.12 \sum_{k=0}^{\infty} \frac{c_{k+1}^2}{k!} \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)^k$$

where we can use

$$\sum_{k=0}^{\infty} \frac{c_k^2}{k!} \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)^k \ge c_0^2 = 0.25$$

and noticing that $\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t)^{-2} Q_1(t) \le \sum_{k=0}^{\infty} \frac{c_{k+1}^2}{k!} = \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\pi'(x)^2\right] \le 0.05$. Then, we need to check that

$$\alpha_1^2\iota_2 + \alpha_1^2\rho_3 + \alpha_1\alpha_2\iota_1 + \alpha_1\alpha_3\rho_1 + \alpha_1\iota_2^2 + \alpha_2\iota_1\iota_2 + \alpha_1\rho_3^2 + \alpha_3\rho_1\rho_3 > \alpha_2\iota_3\rho_1 + \alpha_3\iota_1\rho_2$$

To start, we notice that

$$\alpha_1\alpha_2 - \alpha_3\rho_2 = \frac{36}{a_i^2} C_{S,2} \gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \left(\alpha_1 - \frac{a_i^2}{b_i^2}\cdot\rho_2\right) = \frac{36}{a_i^2} C_{S,2} \gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \left(\alpha_1 - \frac{18}{b_i^2} Q_1(t)\right) > 0$$

when $\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t), \gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \ge 0.9$. Thus, we have that $\alpha_1\alpha_2\iota_1 > \alpha_3\iota_1\rho_2$. Lastly, we also notice that

$$\begin{aligned}
\iota_1\iota_2 - \iota_3\rho_1 \ge{}& \frac{108}{a_i^2 b_i^2} \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \left(Q_0(t) + \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t) Q_1(t)\right) Q_1(t) \\
&+ \frac{108}{a_i^2} \left(\frac{1}{a_i^2} + \frac{1}{b_i^2}\right) \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) Q_1(t) Q_2(t) - \frac{432}{a_i^2 b_i^2} \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t) Q_1(t)^2 \\
\ge{}& \frac{108}{a_i^2 b_i^2} \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t) Q_1(t) \left(\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) Q_0(t) + \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) Q_1(t) - 4Q_1(t)\right) \\
&+ \frac{108}{a_i^2} \left(\frac{1}{a_i^2} + \frac{1}{b_i^2}\right) \gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) Q_1(t) Q_2(t)
\end{aligned}$$

which can be numerically verified as $\gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \ge 0.9$. Therefore, by Lemma 28 we have that

$$\begin{aligned}
&\max\left\{\left|\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(T_1+t)\right|, \left|\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(T_1+t)\right|, \left|I^{(3)}_{i^\star_\ell,i^\star_\ell}(T_1+t)\right|\right\} \\
&\quad\le e^{-\Omega(t)} \left(\left|\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(T_1)\right| + \left|\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(T_1)\right| + \left|I^{(3)}_{i^\star_\ell,i^\star_\ell}(T_1)\right|\right) + \mathcal{O}\left(m\varepsilon^{(2)}_{\mathcal{A},\ell}(t)^3\right) \\
&\quad\le e^{-\Omega(t)} \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right) + \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\varepsilon^{(2)}_{\mathcal{A},\ell}(t)\right)
\end{aligned}$$

Thus, for $t \ge T_1 + \mathcal{O}(\log d)$ we shall have that

$$\max\left\{\left|\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\right|, \left|\zeta^{(1)}_{i^\star_\ell,j^\star_\ell}(t)\right|, \left|I^{(3)}_{i^\star_\ell,i^\star_\ell}(t)\right|\right\} \le \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2 + \frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)$$

The same holds for $\zeta^{(1)}_{i^\star_\ell,j}(t), \zeta^{(1)}_{i^\star_\ell,j}(t), I^{(3)}_{i,i}(t)$ as their growth is upper bounded by the above.

Bounding $\zeta^{(1)}_{i^\star_\ell,j}(t), \zeta^{(2)}_{i^\star_\ell,j}(t)$ and $\gamma^{(1)}_{i^\star_\ell,j}(t), \gamma^{(1)}_{i^\star_\ell,j}(t)$ for $j \in [m^\star] \setminus \{j^\star_\ell\}$

Lemma 15. *Suppose that the inductive hypothesis in Condition 2 and the initialization condition in Condition 1 hold. Then for all $t \ge 0$ we have that*

$$\begin{aligned}
\left|\zeta^{(1)}_{i^\star_\ell,j}(t)\right|, \left|\zeta^{(2)}_{i^\star_\ell,j}(t)\right| &\le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}} + m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\right); \forall j \in [m^\star] \\
\left|\gamma^{(1)}_{i^\star_\ell,j}(t)\right|, \left|\gamma^{(2)}_{i^\star_\ell,j}(t)\right| &\le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}} + m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\right); \forall j \ne j^\star_\ell
\end{aligned}$$

Proof. For $j \in [m^\star] \setminus \{j_\ell^\star\}$, we write out the dynamic of $\zeta_{i_\ell^\star,j}^{(1)}(t), \zeta_{i_\ell^\star,j}^{(2)}(t)$ from Lemma 7

$$\frac{d}{dt}\left|\zeta_{i_\ell^\star,j}^{(1)}(t)\right| = -\frac{1}{a_{i_\ell^\star}^2}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\left|\zeta_{i_\ell^\star,j}^{(1)}(t)\right| + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)$$
$$\frac{d}{dt}\left|\zeta_{i_\ell^\star,j}^{(2)}(t)\right| = -\frac{9}{b_{i_\ell^\star}^2}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\left|\zeta_{i_\ell^\star,j}^{(2)}(t)\right| + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)$$

where we applied the upper bound that $\left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i_\ell^\star,j}^{(1)}(t)\right|, \left|\hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\gamma_{i_\ell^\star,j}^{(2)}(t)\right| \le \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)$ and $\left|\gamma_{i_\ell^\star,j}^{(2)}(t)I_{i_\ell^\star,i_\ell^\star}^{(3)}(t)\right| \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$ from Lemma 14. Since both $\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)$ and $\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)$ are positive, we have that $\left|\zeta_{i_\ell^\star,j}^{(1)}(t)\right|$ and $\left|\zeta_{i_\ell^\star,j}^{(2)}(t)\right|$ enjoys an exponential decay up to $\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)$. Recall that at the end of phase 1 we have $\left|\zeta_{i_\ell^\star,j}^{(1)}(t)\right|, \left|\zeta_{i_\ell^\star,j}^{(2)}(t)\right| \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$. Therefore, we can conclude that

$$\left|\zeta_{i_\ell^\star,j}^{(1)}(t)\right|, \left|\zeta_{i_\ell^\star,j}^{(2)}(t)\right| \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}} + m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)$$

Now, we focus on $\gamma_{i_\ell^\star,j}^{(1)}(t)$ and $\gamma_{i_\ell^\star,j}^{(2)}(t)$ for $j \neq j_\ell^\star$. In particular, by Lemma 7, we have that

$$\frac{d}{dt}\gamma_{i_\ell^\star,j}^{(1)}(t) = -\frac{1}{a_{i_\ell^\star}^2}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\left|\gamma_{i_\ell^\star,j}^{(1)}(t)\right| + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)$$
$$\frac{d}{dt}\gamma_{i_\ell^\star,j}^{(2)}(t) = -\frac{9}{b_{i_\ell^\star}^3}\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\left|\gamma_{i_\ell^\star,j}^{(2)}(t)\right| + \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)$$

Since both $\hat{\lambda}_{i_\ell^\star,j_\ell^\star,1}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)$ and $\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)$ are positive, we have that $\left|\gamma_{i_\ell^\star,j}^{(1)}(t)\right|$ and $\left|\gamma_{i_\ell^\star,j}^{(2)}(t)\right|$ for $j \neq j_\ell^\star$ enjoys an exponential decay up to $\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)$. Thus, we can conclude that

$$\left|\gamma_{i_\ell^\star,j}^{(1)}(t)\right|, \left|\gamma_{i_\ell^\star,j}^{(2)}(t)\right| \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}} + m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right); \forall j \neq j_\ell^\star$$

□

Phase 2 Growth of $\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t), \gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)$. In this section, we show that $\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)$ continues growing up to at least $1 - \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)$.

Lemma 16. *Suppose that the inductive hypothesis in Condition 2 and the initialization condition in Condition 1 hold. Then for all $t \ge T_1 + \mathcal{O}(\log d)$, where T_1 is defined in Lemma 13, we have that*

$$\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t), \gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t) \ge 1 - \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3 + \frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)$$

Proof. To start, we need to perform a more fine-grained analysis of the dynamic of $\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)$. Recall from the proof of Lemma 7 we have that for $(i,j) = (i_\ell^\star, j_\ell^\star)$

$$\sum_{r=1}^{m}\lambda_{i,r,5}(t)\gamma_{r,j}^{(2)}(t) = \lambda_{i,i,5}(t)\gamma_{i,j}^{(5)}(t) \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\varepsilon_{\mathcal{A},\ell}^{(2)}(t)\right)$$
$$\sum_{r=1}^{m}\lambda_{i,r,2}(t)\zeta_{i,j}^{(1)}(t) = \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)\varepsilon_{5,\ell}(t)\right)$$
$$\sum_{r=1}^{m}\lambda_{i,r,3}(t)\zeta_{r,j}^{(1)}(t) = \pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right) \pm \mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)\varepsilon_{5,\ell}(t)\right)$$
$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,2}(t)\zeta_{i,j}^{(1)}(t) = \pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2 \pm \mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$

and also

$$\sum_{r=1}^{m}\lambda_{i,r,2}(t)I_{i,i}^{(3)}(t)=\pm\mathcal{O}\left(\varepsilon_{5,\ell}(t)^2+m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$
$$\sum_{r=1}^{m}\lambda_{i,r,3}(t)I_{r,i}^{(3)}(t)=\pm\mathcal{O}\left(\varepsilon_{5,\ell}(t)^2+m\varepsilon_{\mathcal{A},\ell}^{(2)}(t)^3\right)$$
$$\sum_{r=1}^{m}\lambda_{i,r,5}(t)I_{r,i}^{(2)}(t)=\lambda_{i,i,5}(t)\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$
$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,2}(t)I_{i,i}^{(3)}(t)=\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$
$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,3}(t)\zeta_{i,r}^{(2)}(t)=\pm\mathcal{O}\left(\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$
$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i,r,5}(t)\gamma_{i,r}^{(2)}(t)=\lambda_{i_\ell^\star,j_\ell^\star,5}(t)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$

In this case, we need to refine the bound that

$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i_\ell^\star,r,2}(t)\zeta_{i_\ell^\star,j_\ell^\star}^{(1)}(t)=\hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)\zeta_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$
$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i_\ell^\star,r,2}(t)I_{i_\ell^\star,i_\ell^\star}^{(3)}(t)=\hat{\lambda}_{i_\ell^\star,j_\ell^\star,2}(t)I_{i_\ell^\star,i_\ell^\star}^{(3)}(t)\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$
$$\sum_{r=1}^{m^\star}\hat{\lambda}_{i_\ell^\star,r,3}(t)\zeta_{i_\ell^\star,r}^{(2)}(t)=\hat{\lambda}_{i_\ell^\star,j_\ell^\star,3}(t)\zeta_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\pm\mathcal{O}\left(m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$

Denote $\hat{\varepsilon}_\ell(t)=\max\left\{\left|\zeta_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\right|,\left|\zeta_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\right|,\left|I_{i_\ell^\star,i_\ell^\star}^{(3)}(t)\right|\right\}$. Then we have that

$$\left|\sum_{r=1}^{m^\star}\hat{\lambda}_{i_\ell^\star,r,2}(t)\zeta_{i_\ell^\star,j_\ell^\star}^{(1)}(t)\right|\le\mathcal{O}\left(\hat{\varepsilon}_\ell(t)^2+m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$
$$\left|\sum_{r=1}^{m^\star}\hat{\lambda}_{i_\ell^\star,r,2}(t)I_{i_\ell^\star,i_\ell^\star}^{(3)}(t)\right|\le\mathcal{O}\left(\hat{\varepsilon}_\ell(t)^2+m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$
$$\left|\sum_{r=1}^{m^\star}\hat{\lambda}_{i_\ell^\star,r,3}(t)\zeta_{i_\ell^\star,r}^{(2)}(t)\right|\le\mathcal{O}\left(\hat{\varepsilon}_\ell(t)^2+m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$

Noticing that $\varepsilon_{5,\ell}(t)\le\hat{\varepsilon}_\ell(t)$, we have that

$$\frac{d}{dt}\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\ge\left(1-\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\right)\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)-\mathcal{O}\left(\hat{\varepsilon}_\ell(t)^2+m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)$$

By Lemma 21 we have that

$$\hat{\lambda}_{i_\ell^\star,j_\ell^\star,5}(t)\ge2\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)\sum_{k=0}^{\infty}\frac{c_k^2}{k!}\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)^k-\mathcal{O}\left(\hat{\varepsilon}_\ell(t)^2\right)$$

This gives that

$$\begin{aligned}\frac{d}{dt}\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)&\ge2\left(1-\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\right)\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\sum_{k=0}^{\infty}\frac{c_k^2}{k!}\gamma_{i_\ell^\star,j_\ell^\star}^{(1)}(t)^k-\mathcal{O}\left(\hat{\varepsilon}_\ell(t)^2+m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)\\&\ge c_0^2\left(1-\gamma_{i_\ell^\star,j_\ell^\star}^{(2)}(t)^2\right)-\mathcal{O}\left(\hat{\varepsilon}_\ell(t)^2+m\varepsilon_{\mathcal{A},\ell}^{(1)}(t)^3\right)\end{aligned}$$

Let $\hat{T}$ be the first time when $\hat{\varepsilon}_\ell(t) \le \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3 + \frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)$. Recall that for all $t \le \hat{T}$ we have $\hat{\varepsilon}_\ell(t) \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$. Therefore, we have that either $\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t) \ge 1 - \mathcal{O}\left(\frac{m^4}{\delta_{\mathbb{P}}^2 d}\right)$ or $\frac{d}{dt}\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t) \ge 0$. Thus, we must have that at $\hat{T}$

$$\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(\hat{T}) \ge \gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(T_1) \ge 0.9$$

For all $t \ge \hat{T}$, we have that

$$\frac{d}{dt}\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t) \ge c_0^2\left(1 - \gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)^2\right) - \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3 + \frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)$$

Before $\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)$ first reaches $\mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3 + \frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right)$, we have that

$$\frac{d}{dt}\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t) \ge \frac{c_0^2}{2}\left(1 - \gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t)^2\right)$$

This gives that

$$\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t + \hat{T}) \ge \frac{\left(1 + \gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(\hat{T})\right)\exp\left(\frac{c_0^2 t}{2}\right) - 1 + \gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(\hat{T})}{\left(1 + \gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(\hat{T})\right)\exp\left(\frac{c_0^2 t}{2}\right) + 1 - \gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(\hat{T})} \ge \frac{\exp\left(\frac{c_0^2 t}{2}\right) - 0.1}{\exp\left(\frac{c_0^2 t}{2}\right) + 0.1}$$

Thus, for $t \ge \mathcal{O}(\log d)$ we have that $\gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t + \hat{T}) \ge 1 - \mathcal{O}\left(\frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}} + m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3\right)$ and stays at that magnitude. We conclude the proof by noticing that $\hat{T} \le T_1 + \mathcal{O}(\log d)$ by Lemma 14, and the same analysis holds for $\gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t)$. □

With all the preparation work, we are ready to state the lemma for phase 2 convergence.

Lemma 17. *Suppose that the inductive hypothesis in Condition 2, the initialization condition in Condition 1, and the condition on the sigmoid function in Assumption 3 holds. Let T_c be defined in (14). Then for all $t \ge T_1$, where T_1 is defined in Lemma 13, we have that*

$$\gamma^{(1)}_{i^\star_\ell, j^\star_\ell}(t), \gamma^{(2)}_{i^\star_\ell, j^\star_\ell}(t) \ge \begin{cases} 0.9 & \text{for } t \ge T_1 \\ 1 - \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3 + \frac{m^7}{\delta_{\mathbb{P}}^3 d^{\frac{3}{2}}}\right) & \text{for } t \ge T_1 + \mathcal{O}(\log d) \end{cases}$$

Moreover, for all $t \ge T_1$, we can upper bound $\left|\zeta^{(1)}_{i^\star_\ell, j}(t)\right|, \left|\zeta^{(2)}_{i^\star_\ell, j}(t)\right|$ for $j \in [m^\star]$ and $\left|\gamma^{(1)}_{i^\star_\ell, j}(t)\right|, \left|\gamma^{(2)}_{i^\star_\ell, j}(t)\right|$ for $j \in [m^\star] \setminus \{j^\star_\ell\}$ by

$$\left|\zeta^{(1)}_{i^\star_\ell, j}(t)\right|, \left|\zeta^{(2)}_{i^\star_\ell, j}(t)\right| \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}} + m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\right); \forall j \in [m^\star]$$
$$\left|\gamma^{(1)}_{i^\star_\ell, j}(t)\right|, \left|\gamma^{(2)}_{i^\star_\ell, j}(t)\right| \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}} + m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\right); \forall j \ne j^\star_\ell$$

Proof. The first part of the proof simply follows from a combination of Lemma 14, Lemma 15, and Lemma 16. □

A.8 Formalizing the Proof of Theorem 1

Now we are ready to prove the theorem for gradient flow.

Proof of Theorem 1. By Lemma 17, we have that

$$\left|\zeta^{(1)}_{i^\star_\ell, j}(t)\right|, \left|\zeta^{(2)}_{i^\star_\ell, j}(t)\right| \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}} + m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\right); \forall j \in [m^\star]$$
$$\left|\gamma^{(1)}_{i^\star_\ell, j}(t)\right|, \left|\gamma^{(2)}_{i^\star_\ell, j}(t)\right| \le \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}} + m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\right); \forall j \ne j^\star_\ell$$

By the definition of $\varepsilon^{(1)}_{\mathcal{A},\ell}(t)$, if $\varepsilon^{(1)}_{\mathcal{A},\ell}(t) \geq \varepsilon^{(2)}_{\mathcal{A},\ell+1}(t)$, then we must have that $\left|\zeta^{(1)}_{i^\star_\ell,j}(t)\right|, \left|\zeta^{(2)}_{i^\star_\ell,j}(t)\right|$ and $\left|\gamma^{(1)}_{i^\star_\ell,j}(t)\right|, \left|\gamma^{(2)}_{i^\star_\ell,j}(t)\right|$ is greater than $\varepsilon^{(2)}_{\mathcal{A},\ell+1}(t)$. In this case, we must have that $\varepsilon^{(1)}_{\mathcal{A},\ell}(t)$ is dominated by $\left|\zeta^{(1)}_{i^\star_\ell,j}(t)\right|, \left|\zeta^{(2)}_{i^\star_\ell,j}(t)\right|$ and $\left|\gamma^{(1)}_{i^\star_\ell,j}(t)\right|, \left|\gamma^{(2)}_{i^\star_\ell,j}(t)\right|$. For $t \leq T_c$, we can thus conclude that $\left|\zeta^{(1)}_{i^\star_\ell,j}(t)\right|, \left|\zeta^{(2)}_{i^\star_\ell,j}(t)\right|$ and $\left|\gamma^{(1)}_{i^\star_\ell,j}(t)\right|, \left|\gamma^{(2)}_{i^\star_\ell,j}(t)\right|$ must stay below $\mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$. Otherwise, we will have that $\varepsilon^{(1)}_{\mathcal{A},\ell}(t) \leq \varepsilon^{(2)}_{\mathcal{A},\ell+1}(t)$ and thus

$$\left|\zeta^{(1)}_{i^\star_\ell,j}(t)\right|, \left|\zeta^{(2)}_{i^\star_\ell,j}(t)\right| \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}} + m\varepsilon^{(1)}_{\mathcal{A},\ell+1}(t)^3 + \varepsilon^{(1)}_{\mathcal{A},\ell+1}(t)^2\right); \forall j \in [m^\star]$$
$$\left|\gamma^{(1)}_{i^\star_\ell,j}(t)\right|, \left|\gamma^{(2)}_{i^\star_\ell,j}(t)\right| \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}} + m\varepsilon^{(1)}_{\mathcal{A},\ell+1}(t)^3 + \varepsilon^{(1)}_{\mathcal{A},\ell+1}(t)^2\right); \forall j \neq j^\star_\ell$$

which implies that for all t such that $\varepsilon^{(1)}_{\mathcal{A},\ell}(t) \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$, it holds that

$$\left|\zeta^{(1)}_{i^\star_\ell,j}(t)\right|, \left|\zeta^{(2)}_{i^\star_\ell,j}(t)\right| \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right); \forall j \in [m^\star]$$
$$\left|\gamma^{(1)}_{i^\star_\ell,j}(t)\right|, \left|\gamma^{(2)}_{i^\star_\ell,j}(t)\right| \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right); \forall j \neq j^\star_\ell$$

Combining with the inductive hypothesis that $\varepsilon_{\mathcal{F},\ell}(t) \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$ gives that $\varepsilon_{\mathcal{F},\ell+1}(t) \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$ when $\varepsilon^{(1)}_{\mathcal{A},\ell+1}(t) \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$. This gives the third inductive hypothesis. For the same reasoning, by Lemma 17, we can also have the first inductive hypothesis. Lastly, the second inductive hypothesis follows from Lemma 9.

Based on the first and second inductive hypothesis, we can conclude bullet point 1-3 in Theorem 1. To see the last statement, we recall from Lemma 17 that

$$\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t), \gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \geq 1 - \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3 + \frac{m^7}{\delta^3_{\mathbb{P}} d^{\frac{3}{2}}}\right) \text{ for } t \geq T_1 + \mathcal{O}(\log d)$$

Let $T^\star \geq T_1 + \mathcal{O}(\log d)$ to be as small as possible for all ℓ. Then $T^\star \leq \mathcal{O}\left(\sqrt{d} + \log d\right) \leq \mathcal{O}\left(\sqrt{d}\right)$ by Lemma 13. By definition of $\varepsilon^{(1)}_{\mathcal{A},\ell}(t)$, we have that $\varepsilon^{(1)}_{\mathcal{A},\ell}(t) \leq \varepsilon^{(1)}_{\mathcal{A},\ell+1}(t)$ for all $\ell \in [m^\star - 1]$. Thus, $\varepsilon^{(1)}_{\mathcal{A},\ell}(t) \leq \varepsilon^{(1)}_{\mathcal{A},m^\star}(t)$ for all $\ell \in [m^\star]$. This gives that

$$\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t), \gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \geq 1 - \mathcal{O}\left(m\varepsilon^{(1)}_{\mathcal{A},m^\star}(t)^3 + \frac{m^7}{\delta^3_{\mathbb{P}} d^{\frac{3}{2}}}\right) \text{ for } t \geq T^\star$$

Thus, it remains to bound $\varepsilon^{(1)}_{\mathcal{A},m^\star}(t)$. By definition, $\varepsilon^{(1)}_{\mathcal{A},m^\star}(t)$ depends on $\gamma^{(1)}_{i^\star_\ell,j}(t), \gamma^{(1)}_{i^\star_\ell,j}(t)$ for all $\ell \in [m^\star]$ and $j \in [m^\star] \setminus \{j^\star_\ell\}$, $\zeta^{(1)}_{i,j}(t), \zeta^{(2)}_{i,j}(t)$ for all $i \in [m]$ and $j \in [m^\star]$, and $I^{(1)}_{i,j}(t), I^{(2)}_{i,j}(t), I^{(3)}_{i,j}(t)$ for all $i, j \in [m]$ with $i \neq j$. Recall that by Lemma 17, we have that for all $t \geq T_1$, it holds that

$$\left|\zeta^{(1)}_{i^\star_\ell,j}(t)\right|, \left|\zeta^{(2)}_{i^\star_\ell,j}(t)\right| \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}} + m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\right); \forall j \in [m^\star]$$
$$\left|\gamma^{(1)}_{i^\star_\ell,j}(t)\right|, \left|\gamma^{(2)}_{i^\star_\ell,j}(t)\right| \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}} + m\varepsilon^{(1)}_{\mathcal{A},\ell}(t)^3 + \varepsilon^{(1)}_{\mathcal{A},\ell}(t)^2\right); \forall j \neq j^\star_\ell$$

For $\zeta^{(1)}_{i,j}(t), \zeta^{(2)}_{i,j}(t)$ and $I^{(1)}_{i,j}(t), I^{(2)}_{i,j}(t), I^{(3)}_{i,j}(t)$ with $i \in [m] \setminus \mathcal{R}_{m^\star}$, we obtain from Lemma 11 that the above are bounded by $\mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$ for all $t \leq \mathcal{O}\left(\frac{\delta^2_{\mathbb{P}} d}{m^5}\right)$. Lastly, for $I^{(1)}_{i,j}(t), I^{(2)}_{i,j}(t), I^{(3)}_{i,j}(t)$ with $i \in \mathcal{R}_{m^\star}$ or $j \in \mathcal{R}_{m^\star}$, we can obtain from (17), (18), and (19) that

$$\max\left\{\left|I^{(1)}_{i,j}(t)\right|, \left|I^{(2)}_{i,j}(t)\right|, \left|I^{(3)}_{i,j}(t)\right|\right\} \leq \max\left\{\left|\gamma^{(1)}_{j,j^\star_{\ell'}}(t)\right|, \left|\gamma^{(2)}_{j,j^\star_{\ell'}}(t)\right|, \left|\zeta^{(1)}_{j,j^\star_{\ell'}}(t)\right|, \left|\zeta^{(1)}_{j,j^\star_{\ell'}}(t)\right|\right\} + \mathcal{O}\left(\frac{m^4}{\delta^2_{\mathbb{P}} d^{\frac{3}{4}}}\right)$$

As $\left|\zeta^{(1)}_{j,j^\star_{\ell'}}(t)\right|, \left|\zeta^{(1)}_{j,j^\star_{\ell'}}(t)\right|$ are bounded above, we just need to look into $\left|\gamma^{(1)}_{j,j^\star_{\ell'}}(t)\right|, \left|\gamma^{(2)}_{j,j^\star_{\ell'}}(t)\right|$. which are bounded by $\mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$ as shown in Lemma 10 when $t \leq \left\{T_{m^\star}(\xi) + \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right), \mathcal{O}\left(\frac{\delta^2_{\mathbb{P}} d^{\frac{3}{4}}}{m^5}\right)\right\}$. Combining all the bounds above we can conclude that $\varepsilon^{(1)}_{\mathcal{A},m^\star}(t) \leq \mathcal{O}\left(\frac{m^2}{\delta_{\mathbb{P}}\sqrt{d}}\right)$ for all $t \leq \min\left\{T_{m^\star}(\xi) + \mathcal{O}\left(\frac{\delta_{\mathbb{P}}\sqrt{d}}{m^2}\right)\right\}$. Thus, we can obtain that

$$\gamma^{(1)}_{i^\star_\ell,j^\star_\ell}(t), \gamma^{(2)}_{i^\star_\ell,j^\star_\ell}(t) \geq 1 - \mathcal{O}\left(\frac{m^7}{\delta^3_{\mathbb{P}} d^{\frac{3}{2}}}\right); \forall T^\star \leq t \leq T^\star + \mathcal{O}\left(\frac{\delta_{\mathbb{P}}\sqrt{d}}{m^2}\right)$$

□

B Proof of Theorem 2

Proof of Theorem 2. We simply need to show that under the stopping criteria (8) the procedure in (7) satisfies that $r_\tau \in [m] \setminus [m^\star]$ for all $\tau \in [\tau^\star]$ and that $\tau^\star = m - m^\star$. This is done in **Part 1** and **Part 2** below. Before we start, we define the loss over the pruned model as

$$\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}}\Big[\left(f_{\mathcal{S}}(\boldsymbol{\theta}, \mathbf{x}) - f^\star(\mathbf{x})\right)^2\Big]$$

and we write the target pruned model as

$$\hat{f}(\boldsymbol{\theta}, \mathbf{x}) = \sum_{i=1}^{m^\star} \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)$$

For the convenience, we also denote $h_i(\mathbf{x}) := \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)$ and $h^\star_i(\mathbf{x}) := \pi\left(\bar{\mathbf{v}}_i^{\star\top} \mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^{\star\top} \mathbf{x}\right)$.

Part 1. Assume that $\mathcal{S}_{\tau-1} \subseteq [m] \setminus [m^\star]$ and $|\mathcal{S}_{\tau-1}| \leq m - m^\star$. Let $r_\tau \in [m] \setminus [m^\star]$ and $r'_\tau \in [m^\star]$. Let $\mathcal{S}_\tau = \mathcal{S}_{\tau-1} \cup \{r_\tau\}$ and $\mathcal{S}'_\tau = \mathcal{S}_{\tau-1} \cup \{r'_\tau\}$. Moreover, let $\mathcal{S}^\perp = [m] \setminus [m^\star] \setminus \mathcal{S}_\tau$. Then we have that

$$\begin{aligned}
\mathcal{L}_{\mathcal{S}_\tau}(\boldsymbol{\theta}) - \mathcal{L}_{\mathcal{S}'_\tau}(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{x}}\Big[\left(f_{\mathcal{S}_\tau}(\boldsymbol{\theta}, \mathbf{x}) - f^\star(\mathbf{x})\right)^2 - \left(f_{\mathcal{S}'_\tau}(\boldsymbol{\theta}, \mathbf{x}) - f^\star(\mathbf{x})\right)^2\Big] \\
&= \mathbb{E}_{\mathbf{x}}\Big[\left(f_{\mathcal{S}_\tau}(\boldsymbol{\theta}, \mathbf{x}) + f_{\mathcal{S}'_\tau}(\boldsymbol{\theta}, \mathbf{x}) - 2f^\star(\mathbf{x})\right)\left(f_{\mathcal{S}_\tau}(\boldsymbol{\theta}, \mathbf{x}) - f_{\mathcal{S}'_\tau}(\boldsymbol{\theta}, \mathbf{x})\right)\Big] \\
&= \mathbb{E}_{\mathbf{x}}\Big[\left(f_{\mathcal{S}_\tau}(\boldsymbol{\theta}, \mathbf{x}) + f_{\mathcal{S}'_\tau}(\boldsymbol{\theta}, \mathbf{x}) - 2\hat{f}(\boldsymbol{\theta}, \mathbf{x})\right)\left(f_{\mathcal{S}_\tau}(\boldsymbol{\theta}, \mathbf{x}) - f_{\mathcal{S}'_\tau}(\boldsymbol{\theta}, \mathbf{x})\right)\Big] \\
&\quad \underbrace{- 2\,\mathbb{E}_{\mathbf{x}}\Big[\left(\hat{f}(\boldsymbol{\theta}, \mathbf{x}) - f^\star(\mathbf{x})\right)\left(f_{\mathcal{S}_\tau}(\boldsymbol{\theta}, \mathbf{x}) - f_{\mathcal{S}'_\tau}(\boldsymbol{\theta}, \mathbf{x})\right)\Big]}_{\mathcal{T}_1} \\
&= \mathbb{E}_{\mathbf{x}}\left[\left(2\sum_{i\in\mathcal{S}^\perp} h_i(\mathbf{x}) + h_{r_\tau}(\mathbf{x}) - h_{r'_\tau}(\mathbf{x})\right)\left(h_{r'_\tau}(\mathbf{x}) - h_{r_\tau}(\mathbf{x})\right)\right] - 2\mathcal{T}_1 \\
&= -\mathbb{E}_{\mathbf{x}}\Big[\left(h_{r'_\tau}(\mathbf{x}) - h_{r_\tau}(\mathbf{x})\right)^2\Big] - 2\mathcal{T}_1 + \underbrace{2\,\mathbb{E}_{\mathbf{x}}\left[\sum_{i\in\mathcal{S}^\perp} h_i(\mathbf{x})\left(h_{r'_\tau}(\mathbf{x}) - h_{r_\tau}(\mathbf{x})\right)\right]}_{\mathcal{T}_2} \\
&\leq -\mathbb{E}_{\mathbf{x}}\Big[\left(h_{r'_\tau}(\mathbf{x}) - h_{r_\tau}(\mathbf{x})\right)^2\Big] + 2|\mathcal{T}_1| + 2|\mathcal{T}_2|
\end{aligned}$$

It suffice to upper bound $|\mathcal{T}_1|$ and $|\mathcal{T}_2|$, and lower bound $\mathbb{E}_{\mathbf{x}}\Big[\left(h_{r'_\tau}(\mathbf{x}) - h_{r_\tau}(\mathbf{x})\right)^2\Big]$. To start, the lower bound can be derived as

$$\mathbb{E}_{\mathbf{x}}\Big[\left(h_{r'_\tau}(\mathbf{x}) - h_{r_\tau}(\mathbf{x})\right)^2\Big] = \mathbb{E}_{\mathbf{x}}\Big[h_{r'_\tau}(\mathbf{x})^2\Big] + \mathbb{E}_{\mathbf{x}}\Big[h_{r_\tau}(\mathbf{x})^2\Big] - 2\mathbb{E}_{\mathbf{x}}\Big[h_{r_\tau}(\mathbf{x})\, h_{r'_\tau}(\mathbf{x})\Big] \geq 12\sum_{k=0}^{\infty}\frac{c_k^2}{k!} - \mathcal{O}\left(\varepsilon^2\right) \tag{20}$$

where the last inequality follows from Lemma 22. Next, for $|\mathcal{T}_1|$, we have that

$$\begin{aligned}\mathcal{T}_1 &= \mathbb{E}_{\mathbf{x}}\Big[\Big(\hat{f}\left(\boldsymbol{\theta},\mathbf{x}\right)-f^\star\left(\mathbf{x}\right)\Big)\left(h_{r_\tau}\left(\mathbf{x}\right)-h_{r'_\tau}\left(\mathbf{x}\right)\right)\Big]\\ &= \mathbb{E}_{\mathbf{x}}\left[h_{r_\tau}\left(\mathbf{x}\right)\sum_{i=1}^{m^\star}h_i\left(\mathbf{x}\right)\right]-\mathbb{E}_{\mathbf{x}}\left[h_{r_\tau}\left(\mathbf{x}\right)\sum_{i=1}^{m^\star}h_i^\star\left(\mathbf{x}\right)\right]-\mathbb{E}_{\mathbf{x}}\left[h'_{r_\tau}\left(\mathbf{x}\right)\sum_{i=1,i\neq r'_\tau}^{m^\star}h_i\left(\mathbf{x}\right)\right]\\ &\quad+\mathbb{E}_{\mathbf{x}}\left[h'_{r_\tau}\left(\mathbf{x}\right)\sum_{i=1,i\neq r'_\tau}^{m^\star}h_i^\star\left(\mathbf{x}\right)\right]+\mathbb{E}_{\mathbf{x}}\Big[h_{r'_\tau}\left(\mathbf{x}\right)^2-h_{r'_\tau}\left(\mathbf{x}\right)h^\star_{r'_\tau}\left(\mathbf{x}\right)\Big]\\ &= \mathbb{E}_{\mathbf{x}}\Big[h_{r'_\tau}\left(\mathbf{x}\right)^2-h_{r'_\tau}\left(\mathbf{x}\right)h^\star_{r'_\tau}\left(\mathbf{x}\right)\Big]\pm\mathcal{O}\left(m^\star\varepsilon^4\right)\end{aligned}\tag{21}$$

By Lemma 22, since $r'_\tau\in[m^\star]$, we have that

$$\mathbb{E}_{\mathbf{x}}\Big[h_{r'_\tau}\left(\mathbf{x}\right)h^\star_{r'_\tau}\left(\mathbf{x}\right)\Big]=6\mathbb{E}_{\mathbf{x}}\big[h_{r'_\tau}(\mathbf{x})^2\big]\pm\mathcal{O}\left(\varepsilon\right)\tag{22}$$

Thus, we have that

$$|\mathcal{T}_1|\leq\mathcal{O}\left(\varepsilon+m^\star\varepsilon^4\right)\tag{23}$$

For $|\mathcal{T}_2|$, we notice that $r_\tau\in\mathcal{S}_\tau$ and $r'_\tau\in[m^\star]$. Therefore, $r_\tau,r'_\tau\notin\mathcal{S}^\perp$. Thus

$$|\mathcal{T}_2|\leq\sum_{i\in\mathcal{S}^\perp}\left|\mathbb{E}_{\mathbf{x}}\big[h_i\left(\mathbf{x}\right)h_{r_\tau}\left(\mathbf{x}\right)\big]\right|+\sum_{i\in\mathcal{S}^\perp}\left|\mathbb{E}_{\mathbf{x}}\big[h_i\left(\mathbf{x}\right)h_{r'_\tau}\left(\mathbf{x}\right)\big]\right|\leq\mathcal{O}\left(m\varepsilon^4\right)\tag{24}$$

Combining (20), (23), and (24) gives that

$$\mathcal{L}_{\mathcal{S}_\tau}\left(\boldsymbol{\theta}\right)-\mathcal{L}_{\mathcal{S}'_\tau}\left(\boldsymbol{\theta}\right)\leq-12\sum_{k=0}^{\infty}\frac{c_k^2}{k!}+\mathcal{O}\left(\varepsilon+m\varepsilon^4\right)\leq0$$

when $\varepsilon\leq o\left(\frac{1}{\sqrt{m}}\right)$.

Part 2. Assume that $\tau^\star<m-m^\star$. We show that $\mathcal{L}_{\mathcal{S}_{\tau^\star}}\left(\boldsymbol{\theta}\right)\geq\mathcal{L}_{\mathcal{S}_{\tau^\star}}\left(\boldsymbol{\theta}+1\right)$ by letting $\mathcal{S}_{\tau^\star+1}=\mathcal{S}_{\tau^\star}\cup\{r^\star\}$ where $r^\star\in[m]\setminus[m^\star]$. Similar to before, let $\mathcal{S}^\perp=[m]\setminus[m^\star]\setminus\mathcal{S}_{\tau^\star+1}$, we have that

$$\begin{aligned}\mathcal{L}_{\mathcal{S}_{\tau^\star}}\left(\boldsymbol{\theta}\right)-\mathcal{L}_{\mathcal{S}_{\tau^\star}}\left(\boldsymbol{\theta}+1\right) &= \mathbb{E}_{\mathbf{x}}\Big[\left(f_{\mathcal{S}_{\tau^\star}}\left(\boldsymbol{\theta},\mathbf{x}\right)-f^\star\left(\mathbf{x}\right)\right)^2-\left(f_{\mathcal{S}_{\tau^\star+1}}\left(\boldsymbol{\theta},\mathbf{x}\right)-f^\star\left(\mathbf{x}\right)\right)^2\Big]\\ &= \mathbb{E}_{\mathbf{x}}\big[\left(f_{\mathcal{S}_{\tau^\star}}\left(\boldsymbol{\theta},\mathbf{x}\right)+f_{\mathcal{S}_{\tau^\star+1}}\left(\boldsymbol{\theta},\mathbf{x}\right)-2f^\star\left(\mathbf{x}\right)\right)\left(f_{\mathcal{S}_{\tau^\star}}\left(\boldsymbol{\theta},\mathbf{x}\right)-f_{\mathcal{S}_{\tau^\star+1}}\left(\boldsymbol{\theta},\mathbf{x}\right)\right)\big]\\ &= \mathbb{E}_{\mathbf{x}}\big[h_{r^\star}\left(\mathbf{x}\right)\left(f_{\mathcal{S}_{\tau^\star}}\left(\boldsymbol{\theta},\mathbf{x}\right)+f_{\mathcal{S}_{\tau^\star+1}}\left(\boldsymbol{\theta},\mathbf{x}\right)-2f^\star\left(\mathbf{x}\right)\right)\big]\\ &= \mathbb{E}_{\mathbf{x}}\Big[h_{r^\star}\left(\mathbf{x}\right)\Big(f_{\mathcal{S}_{\tau^\star}}\left(\boldsymbol{\theta},\mathbf{x}\right)+f_{\mathcal{S}_{\tau^\star+1}}\left(\boldsymbol{\theta},\mathbf{x}\right)-2\hat{f}\left(\boldsymbol{\theta},\mathbf{x}\right)\Big)\Big]\\ &\quad-\underbrace{2\,\mathbb{E}_{\mathbf{x}}\Big[h_{r^\star}\left(\mathbf{x}\right)\Big(\hat{f}\left(\boldsymbol{\theta},\mathbf{x}\right)-f^\star\left(\mathbf{x}\right)\Big)\Big]}_{\mathcal{T}_1}\\ &= \mathbb{E}_{\mathbf{x}}\Big[h_{r^\star}\left(\mathbf{x}\right)^2\Big]+\underbrace{2\,\mathbb{E}_{\mathbf{x}}\left[h_{r^\star}\left(\mathbf{x}\right)\sum_{i\in\mathcal{S}^\perp}h_i\left(\mathbf{x}\right)\right]}_{\mathcal{T}_2}-2\mathcal{T}_1\\ &\geq \mathbb{E}_{\mathbf{x}}\Big[h_{r^\star}\left(\mathbf{x}\right)^2\Big]-2\left|\mathcal{T}_1\right|-2\left|\mathcal{T}_2\right|\end{aligned}$$

As before, we have that $\mathbb{E}_{\mathbf{x}}\Big[h_{r^\star}\left(\mathbf{x}\right)^2\Big]\geq6\sum_{k=0}^{\infty}\frac{c_k^2}{k!}-\mathcal{O}\left(\varepsilon^2\right)$ and $|\mathcal{T}_2|\leq\mathcal{O}\left(m\varepsilon^4\right)$. For $\mathcal{T}_1$, we have that

$$\begin{aligned}|\mathcal{T}_1| &= \left|\mathbb{E}_{\mathbf{x}}\Big[h_{r^\star}\left(\mathbf{x}\right)\Big(\hat{f}\left(\boldsymbol{\theta},\mathbf{x}\right)-f^\star\left(\mathbf{x}\right)\Big)\Big]\right|\\ &\leq \sum_{i=1}^{m^\star}\left|\mathbb{E}_{\mathbf{x}}\big[h_{r^\star}\left(\mathbf{x}\right)h_i\left(\mathbf{x}\right)\big]\right|+\sum_{i=1}^{m^\star}\left|\mathbb{E}_{\mathbf{x}}\big[h_{r^\star}\left(\mathbf{x}\right)h_i^\star\left(\mathbf{x}\right)\big]\right|\\ &\leq \mathcal{O}\left(m\varepsilon^4\right)\end{aligned}$$

Thus, we have that

$$\mathcal{L}_{\mathcal{S}_{\tau^\star}}(\boldsymbol{\theta}) - \mathcal{L}_{\mathcal{S}_{\tau^\star}}(\boldsymbol{\theta}+1) \geq 6\sum_{k=0}^{\infty}\frac{c_k^2}{k!} - \mathcal{O}\left(\varepsilon^2 + m\varepsilon^4\right) \geq 0$$

when $\varepsilon \leq \mathcal{O}\left(\frac{1}{\sqrt{m}}\right)$. This shows that $\tau^\star \geq m - m^\star$. Next, we assume that $\tau^\star > m - m^\star$. Then $r_{m-m^\star+1} \in [m^\star]$. We show that $\mathcal{L}_{\mathcal{S}_{m-m^\star}}(\boldsymbol{\theta}) \leq \mathcal{L}_{\mathcal{S}_{m-m^\star+1}}(\boldsymbol{\theta})$. As before, let . Notice that by **Part 1**, $f_{\mathcal{S}_{m-m^\star}}(\boldsymbol{\theta}, \mathbf{x}) = \hat{f}(\boldsymbol{\theta}, \mathbf{x})$. Then we have that

$$\begin{aligned}
\mathcal{L}_{\mathcal{S}_{m-m^\star}}(\boldsymbol{\theta}) - \mathcal{L}_{\mathcal{S}_{m-m^\star+1}}(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{x}}\Big[h_{r_{m-m^\star+1}}(\mathbf{x})\left(f_{\mathcal{S}_{m-m^\star}}(\boldsymbol{\theta}, \mathbf{x}) + f_{\mathcal{S}_{m-m^\star+1}}(\boldsymbol{\theta}, \mathbf{x}) - 2f^\star(\mathbf{x})\right)\Big] \\
&= 2\mathbb{E}_{\mathbf{x}}\Big[h_{r_{m-m^\star+1}}(\mathbf{x})\left(\hat{f}(\boldsymbol{\theta}, \mathbf{x}) - f^\star(\mathbf{x})\right)\Big] - \mathbb{E}_{\mathbf{x}}\Big[h_{r_{m-m^\star+1}}(\mathbf{x})^2\Big]
\end{aligned}$$

As before, we have that $\mathbb{E}_{\mathbf{x}}\Big[h_{r_{m-m^\star+1}}(\mathbf{x})^2\Big] \geq 6\sum_{k=0}^{\infty}\frac{c_k^2}{k!} - \mathcal{O}\left(\varepsilon^2\right)$. It remains to upper bound the first term. In particular, we have that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}\Big[h_{r_{m-m^\star+1}}(\mathbf{x})\left(\hat{f}(\boldsymbol{\theta}, \mathbf{x}) - f^\star(\mathbf{x})\right)\Big] &\leq \mathbb{E}_{\mathbf{x}}\Big[h_{r_{m-m^\star+1}}(\mathbf{x})^2 - h_{r_{m-m^\star+1}}(\mathbf{x})\, h^\star_{r_{m-m^\star+1}}(\mathbf{x})\Big] \\
&\quad + \sum_{i \neq r_{m-m^\star+1}} \mathbb{E}_{\mathbf{x}}\Big[h_{r_{m-m^\star+1}}(\mathbf{x})\left(h_i(\mathbf{x}) - h_i^\star(\mathbf{x})\right)\Big] \\
&\leq \mathcal{O}\left(\varepsilon + m^\star\varepsilon^4\right)
\end{aligned}$$

Thus, we can conclude that

$$\mathcal{L}_{\mathcal{S}_{m-m^\star}}(\boldsymbol{\theta}) - \mathcal{L}_{\mathcal{S}_{m-m^\star+1}}(\boldsymbol{\theta}) \leq -6\sum_{k=0}^{\infty}\frac{c_k^2}{k!} + \mathcal{O}\left(\varepsilon + m\varepsilon^4\right) \leq 0$$

when $\varepsilon \leq \mathcal{O}\left(\frac{1}{\sqrt{m}}\right)$. This shows that $\tau^\star = m - m^\star$, which finishes the proof. □

C Proof of Theorem 3

We will analyze the Hessian in a small region near the global minima $\boldsymbol{\theta}^\star = \{(\bar{\mathbf{v}}_i^\star, \bar{\mathbf{w}}_i^\star)\}_{i=1}^{m^\star}$. To do this, we utilize the following second-order Stein's lemma.

Lemma 18. *Let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ and $g, h : \mathbb{R} \to \mathbb{R}$. Then we have that*

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[g\left(\mathbf{v}^\top\mathbf{x}\right)h\left(\mathbf{w}^\top\mathbf{x}\right)\mathbf{x}\mathbf{x}^\top\right] &= \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[g\left(\mathbf{v}^\top\mathbf{x}\right)h\left(\mathbf{w}^\top\mathbf{x}\right)\right]\mathbf{I}_d \\
&\quad + \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[g'\left(\mathbf{v}^\top\mathbf{x}\right)h'\left(\mathbf{w}^\top\mathbf{x}\right)\right]\left(\mathbf{v}\mathbf{w}^\top + \mathbf{v}\mathbf{w}^\top\right) \\
&\quad + \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[g''\left(\mathbf{v}^\top\mathbf{x}\right)h\left(\mathbf{w}^\top\mathbf{x}\right)\right]\mathbf{v}\mathbf{v}^\top \\
&\quad + \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[g\left(\mathbf{v}^\top\mathbf{x}\right)h''\left(\mathbf{w}^\top\mathbf{x}\right)\right]\mathbf{w}\mathbf{w}^\top
\end{aligned}$$

The proof of Lemma 18 follows by applying Stein's lemma twice. In particular, we shall prove the following result

Theorem 4. *Let $\boldsymbol{\theta} = \{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{m^\star}$ be the parameter of the MoE, let $\alpha_1, \ldots, \alpha_{m^\star}, \beta_1, \ldots \beta_{m^\star} \geq \Omega(1)$, and let $\mathbf{u}_1, \ldots, \mathbf{u}_{m^\star}, \mathbf{q}_1, \ldots, \mathbf{q}_{m^\star} \in \mathbb{R}^d$ be any set of vectors such that $\|(\mathbf{I} - \bar{\mathbf{v}}_i\bar{\mathbf{v}}_i)\mathbf{u}_i\|_2^2 \geq \mathcal{Q}^\star\|\mathbf{u}_i\|_2^2$ and $\|(\mathbf{I} - \bar{\mathbf{w}}_i\bar{\mathbf{w}}_i)\mathbf{q}_i\|_2^2 \geq \mathcal{Q}^\star\|\mathbf{q}_i\|_2^2$ for some $\mathcal{Q}^\star > 0$ for all $i \in [m^\star]$. If $\boldsymbol{\theta}$ also satisfies that $\|\bar{\mathbf{v}}_i - \bar{\mathbf{v}}_i^\star\|_2, \|\bar{\mathbf{w}}_i - \bar{\mathbf{w}}_i^\star\|_2 \leq \frac{\varepsilon}{2}$ for some $\varepsilon \leq o\left(\frac{N_{\min}\mathcal{Q}^\star}{m^{\star 2}}\right)$, and $\frac{C_{S,0}}{C_{S,1}} \geq \frac{N_{\max}^2(1+\beta_i)^2}{N_{\min}^2\mathcal{Q}^{\star 2}\alpha_i^2}$, then we have that*

$$\begin{bmatrix} \alpha_1\mathbf{u}_1 \\ \vdots \\ \alpha_{m^\star}\mathbf{u}_{m^\star} \\ \beta_1\mathbf{q}_1 \\ \vdots \\ \beta_{m^\star}\mathbf{q}_{m^\star} \end{bmatrix}^\top \nabla^2\mathcal{L}(\boldsymbol{\theta}) \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_{m^\star} \\ \mathbf{q}_1 \\ \vdots \\ \mathbf{q}_{m^\star} \end{bmatrix} \geq N_{\min}\mathcal{Q}^\star\kappa\sum_{i=1}^{m^\star}\left(\|\mathbf{u}_i\|_2^2 + \|\mathbf{q}_i\|_2^2\right)$$

for some constant $\kappa > 0$.

Proof. **Form of Hessian.** Here we are going to compute $\frac{\partial^2}{\partial \mathbf{v}_i \partial \mathbf{v}_j}\mathcal{L}(\boldsymbol{\theta})$, $\frac{\partial^2}{\partial \mathbf{w}_i \partial \mathbf{w}_j}\mathcal{L}(\boldsymbol{\theta})$, and $\frac{\partial^2}{\partial \mathbf{v}_i \partial \mathbf{w}_j}\mathcal{L}(\boldsymbol{\theta})$. Recall that the gradient takes the form

$$\frac{\partial}{\partial \mathbf{v}_i}\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{\|\mathbf{v}_i\|_2}\left(\mathbf{I}_d - \bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\mathbb{E}_{\mathbf{x}}\left[\left(f(\boldsymbol{\theta},\mathbf{x}) - f^\star(\mathbf{x})\right)\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\mathbf{x}\right]$$
$$\frac{\partial}{\partial \mathbf{w}_i}\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{\|\mathbf{w}_i\|_2}\left(\mathbf{I}_d - \bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\mathbb{E}_{\mathbf{x}}\left[\left(f(\boldsymbol{\theta},\mathbf{x}) - f^\star(\mathbf{x})\right)\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\mathbf{x}\right]$$

Therefore

$$\begin{aligned}
\frac{\partial^2}{\partial \mathbf{v}_i \partial \mathbf{v}_j}\mathcal{L}(\boldsymbol{\theta}) =& \frac{1}{\|\mathbf{v}_i\|_2\|\mathbf{v}_j\|_2}\left(\mathbf{I} - \bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\underbrace{\mathbb{E}_{\mathbf{x}}\left[\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\pi'\left(\bar{\mathbf{v}}_j^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_j^\top\mathbf{x}\right)\mathbf{x}\mathbf{x}^\top\right]}_{\mathcal{T}_{i,j,1}}\left(\mathbf{I} - \bar{\mathbf{v}}_j\bar{\mathbf{v}}_j^\top\right)\\
&+ \frac{\mathbb{I}\{i=j\}}{\|\mathbf{v}_i\|_2^2}\left(\mathbf{I} - \bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\underbrace{\mathbb{E}_{\mathbf{x}}\left[\left(f(\boldsymbol{\theta},\mathbf{x}) - f^\star(\mathbf{x})\right)\pi''\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\mathbf{x}\mathbf{x}^\top\right]}_{\mathcal{T}_{i,j,2}}\left(\mathbf{I} - \bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\\
&- \frac{\mathbb{I}\{i=j\}}{\|\mathbf{v}_i\|_2^2}\left(\mathbf{I} - \bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\underbrace{\mathbb{E}_{\mathbf{x}}\left[\left(f(\boldsymbol{\theta},\mathbf{x}) - f^\star(\mathbf{x})\right)\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\bar{\mathbf{v}}_i^\top\mathbf{x}\right]}_{\mathcal{T}_{i,j,3}}\\
&- \frac{\mathbb{I}\{i=j\}}{\|\mathbf{v}_i\|_2^2}\underbrace{\mathbb{E}_{\mathbf{x}}\left[\left(f(\boldsymbol{\theta},\mathbf{x}) - f^\star(\mathbf{x})\right)\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\left(\bar{\mathbf{v}}_i\mathbf{x}^\top + \mathbf{x}\bar{\mathbf{v}}_i^\top\right)\right]}_{\mathcal{T}_{i,j,4}}\\
\frac{\partial^2}{\partial \mathbf{w}_i \partial \mathbf{w}_j}\mathcal{L}(\boldsymbol{\theta}) =& \frac{1}{\|\mathbf{w}_i\|_2\|\mathbf{w}_j\|_2}\left(\mathbf{I} - \bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\underbrace{\mathbb{E}_{\mathbf{x}}\left[\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\pi\left(\bar{\mathbf{v}}_j^\top\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_j^\top\mathbf{x}\right)\mathbf{x}\mathbf{x}^\top\right]}_{\mathcal{T}_{i,j,5}}\left(\mathbf{I} - \bar{\mathbf{w}}_j\bar{\mathbf{w}}_j^\top\right)\\
&+ \frac{\mathbb{I}\{i=j\}}{\|\mathbf{w}_i\|_2^2}\left(\mathbf{I} - \bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\underbrace{\mathbb{E}_{\mathbf{x}}\left[\left(f(\boldsymbol{\theta},\mathbf{x}) - f^\star(\mathbf{x})\right)\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma''\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\mathbf{x}\mathbf{x}^\top\right]}_{\mathcal{T}_{i,j,6}}\left(\mathbf{I} - \bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\\
&- \frac{\mathbb{I}\{i=j\}}{\|\mathbf{w}_i\|_2^2}\left(\mathbf{I} - \bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\underbrace{\mathbb{E}_{\mathbf{x}}\left[\left(f(\boldsymbol{\theta},\mathbf{x}) - f^\star(\mathbf{x})\right)\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\bar{\mathbf{w}}_i^\top\mathbf{x}\right]}_{\mathcal{T}_{i,j,7}}\\
&- \frac{\mathbb{I}\{i=j\}}{\|\mathbf{w}_i\|_2^2}\underbrace{\mathbb{E}_{\mathbf{x}}\left[\left(f(\boldsymbol{\theta},\mathbf{x}) - f^\star(\mathbf{x})\right)\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\left(\bar{\mathbf{w}}_i\mathbf{x}^\top + \mathbf{x}\bar{\mathbf{w}}_i^\top\right)\right]}_{\mathcal{T}_{i,j,8}}\\
\frac{\partial^2}{\partial \mathbf{v}_i \partial \mathbf{w}_j}\mathcal{L}(\boldsymbol{\theta}) =& \frac{1}{\|\mathbf{v}_i\|_2\|\mathbf{w}_j\|_2}\left(\mathbf{I} - \bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\underbrace{\mathbb{E}_{\mathbf{x}}\left[\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\pi\left(\bar{\mathbf{v}}_j^\top\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_j^\top\mathbf{x}\right)\mathbf{x}\mathbf{x}^\top\right]}_{\mathcal{T}_{i,j,9}}\left(\mathbf{I} - \bar{\mathbf{w}}_j\bar{\mathbf{w}}_j^\top\right)\\
&+ \frac{\mathbb{I}\{i=j\}}{\|\mathbf{v}_i\|_2\|\mathbf{w}_i\|_2}\left(\mathbf{I} - \bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\underbrace{\mathbb{E}_{\mathbf{x}}\left[\left(f(\boldsymbol{\theta},\mathbf{x}) - f^\star(\mathbf{x})\right)\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\mathbf{x}\mathbf{x}^\top\right]}_{\mathcal{T}_{i,j,10}}\left(\mathbf{I} - \bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)
\end{aligned}$$

For the convenience of the analysis, we define $C_{S,0} = 2\sum_{k=0}^{\infty}\frac{c_k^2}{k!}$ and $C_{S,1} = 6\sum_{k=0}^{\infty}\frac{c_{k+1}^2}{k!}$. Our next lemma controls the magnitudes of these blocks.

Bounding $\mathcal{T}_{i,j,2}, \mathcal{T}_{i,j,6}$ and $\mathcal{T}_{i,j,10}$. By Lemma 18, we have that

$$\begin{aligned}
\mathcal{T}_{i,j,2} =& \mathbb{E}_{\mathbf{x}}\left[\left(f(\boldsymbol{\theta},\mathbf{x}) - f^\star(\mathbf{x})\right)\pi''\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right]\mathbf{I}_d\\
&+ \mathbb{E}_{\mathbf{x}}\left[\left(f(\boldsymbol{\theta},\mathbf{x}) - f^\star(\mathbf{x})\right)\nabla_{\mathbf{x}}^2\left(\pi''\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right)\right]\\
&+ \mathbb{E}_{\mathbf{x}}\left[\nabla_{\mathbf{x}}^2\left(f(\boldsymbol{\theta},\mathbf{x}) - f^\star(\mathbf{x})\right)\pi''\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right]\\
&+ \mathbb{E}_{\mathbf{x}}\left[\nabla_{\mathbf{x}}\left(f(\boldsymbol{\theta},\mathbf{x}) - f^\star(\mathbf{x})\right)\nabla_{\mathbf{x}}\left(\pi''\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right)^\top\right]
\end{aligned}$$

Similarly, we have that

$$\begin{aligned}\mathcal{T}_{i,j,6} = \mathbb{E}_{\mathbf{x}}\big[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\pi\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma''\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\big]\mathbf{I}_d \\ + \mathbb{E}_{\mathbf{x}}\big[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\nabla_{\mathbf{x}}^2\left(\pi\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma''\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\right)\big] \\ + \mathbb{E}_{\mathbf{x}}\big[\nabla_{\mathbf{x}}^2\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\pi\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma''\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\big] \\ + \mathbb{E}_{\mathbf{x}}\Big[\nabla_{\mathbf{x}}\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\nabla_{\mathbf{x}}\left(\pi\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma''\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\right)^{\top}\Big]\end{aligned}$$

Also, we have that

$$\begin{aligned}\mathcal{T}_{i,j,10} = \mathbb{E}_{\mathbf{x}}\big[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\pi'\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\big]\mathbf{I}_d \\ + \mathbb{E}_{\mathbf{x}}\big[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\nabla_{\mathbf{x}}^2\left(\pi'\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\right)\big] \\ + \mathbb{E}_{\mathbf{x}}\big[\nabla_{\mathbf{x}}^2\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\pi'\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\big] \\ + \mathbb{E}_{\mathbf{x}}\Big[\nabla_{\mathbf{x}}\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\nabla_{\mathbf{x}}\left(\pi'\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\right)^{\top}\Big]\end{aligned}$$

Thus, we can apply Lemma 23 to obtain that

$$\left\|\mathcal{T}_{i,j,2}\right\|_2, \left\|\mathcal{T}_{i,j,6}\right\|_2, \left\|\mathcal{T}_{i,j,10}\right\|_2 \leq \mathcal{O}\left(m^{\star}\varepsilon\right)$$

Bounding $\mathcal{T}_{i,j,3}$ and $\mathcal{T}_{i,j,7}$ By Stein's Lemma, we have that

$$\begin{aligned}\mathcal{T}_{i,j,3} &= \mathbb{E}_{\mathbf{x}}\big[\bar{\mathbf{v}}_i^{\top}\nabla_{\mathbf{x}}\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\pi'\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\big] \\ &\quad + \mathbb{E}_{\mathbf{x}}\big[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\bar{\mathbf{v}}_i^{\top}\nabla_{\mathbf{x}}\left(\pi'\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\right)\big] \\ \mathcal{T}_{i,j,7} &= \mathbb{E}_{\mathbf{x}}\big[\bar{\mathbf{w}}_i^{\top}\nabla_{\mathbf{x}}\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\pi\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\big] \\ &\quad + \mathbb{E}_{\mathbf{x}}\big[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\bar{\mathbf{w}}_i^{\top}\nabla_{\mathbf{x}}\left(\pi\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\right)\big]\end{aligned}$$

Therefore, we have that

$$\begin{aligned}\left|\mathcal{T}_{i,j,3}\right| &\leq \left\|\mathbb{E}_{\mathbf{x}}\big[\nabla_{\mathbf{x}}\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\pi'\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\big]\right\|_2 \\ &\quad + \left\|\mathbb{E}_{\mathbf{x}}\big[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\nabla_{\mathbf{x}}\left(\pi'\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\right)\big]\right\|_2 \\ \left|\mathcal{T}_{i,j,7}\right| &\leq \left\|\mathbb{E}_{\mathbf{x}}\big[\nabla_{\mathbf{x}}\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\pi\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\big]\right\|_2 \\ &\quad + \left\|\mathbb{E}_{\mathbf{x}}\big[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\nabla_{\mathbf{x}}\left(\pi\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\right)\big]\right\|_2\end{aligned}$$

Applying Lemma 23 gives that

$$\left|\mathcal{T}_{i,j,3}\right|, \left|\mathcal{T}_{i,j,7}\right| \leq \mathcal{O}\left(m^{\star}\varepsilon\right)$$

Bounding $\mathcal{T}_{i,j,4}$ and $\mathcal{T}_{i,j,8}$. By the structure of $\mathcal{T}_{i,j,4}$ and $\mathcal{T}_{i,j,8}$, we have that

$$\begin{aligned}\left\|\mathcal{T}_{i,j,4}\right\|_2 &\leq 2\left\|\mathbb{E}_{\mathbf{x}}\big[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\pi'\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\big]\mathbf{x}\right\|_2 \\ &\leq 2\left\|\mathbb{E}_{\mathbf{x}}\big[\nabla_{\mathbf{x}}\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\pi'\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\big]\right\|_2 \\ &\quad + 2\left\|\mathbb{E}_{\mathbf{x}}\big[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\nabla_{\mathbf{x}}\left(\pi'\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\right)\big]\right\|_2\end{aligned}$$

Similarly, for $\mathcal{T}_{i,j,8}$, we have that

$$\begin{aligned}\left\|\mathcal{T}_{i,j,8}\right\|_2 &\leq 2\left\|\mathbb{E}_{\mathbf{x}}\big[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\pi\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\big]\mathbf{x}\right\|_2 \\ &\leq 2\left\|\mathbb{E}_{\mathbf{x}}\big[\nabla_{\mathbf{x}}\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\pi\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\big]\right\|_2 \\ &\quad + 2\left\|\mathbb{E}_{\mathbf{x}}\big[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^{\star}\left(\mathbf{x}\right)\right)\nabla_{\mathbf{x}}\left(\pi\left(\bar{\mathbf{v}}_i^{\top}\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^{\top}\mathbf{x}\right)\right)\big]\right\|_2\end{aligned}$$

Applying Lemma 23 gives that

$$\left\|\mathcal{T}_{i,j,4}\right\|_2, \left\|\mathcal{T}_{i,j,8}\right\|_2 \leq \mathcal{O}\left(m^\star \varepsilon\right)$$

Bounding $\mathcal{T}_{i,j,1}, \mathcal{T}_{i,j,5}, \mathcal{T}_{i,j,10}$ for $i \neq j$. By Lemma 18, we have that

$$\begin{aligned}
\mathcal{T}_{i,j,1} = \mathbb{E}_{\mathbf{x}}\big[& \pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \pi'\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\big] \mathbf{I}_d \\
& + \mathbb{E}_{\mathbf{x}}\big[\nabla_{\mathbf{x}}^2 \left(\pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\right) \pi'\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\big] \\
& + \mathbb{E}_{\mathbf{x}}\big[\pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \nabla_{\mathbf{x}}^2 \left(\pi'\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\right)\big] \\
& + \mathbb{E}_{\mathbf{x}}\Big[\nabla_{\mathbf{x}} \left(\pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\right) \nabla_{\mathbf{x}} \left(\pi'\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\right)^\top \Big] \\
\mathcal{T}_{i,j,5} = \mathbb{E}_{\mathbf{x}}\big[& \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \pi\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\big] \mathbf{I}_d \\
& + \mathbb{E}_{\mathbf{x}}\big[\nabla_{\mathbf{x}}^2 \left(\pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\right) \pi\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\big] \\
& + \mathbb{E}_{\mathbf{x}}\big[\pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \nabla_{\mathbf{x}}^2 \left(\pi\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\right)\big] \\
& + \mathbb{E}_{\mathbf{x}}\Big[\nabla_{\mathbf{x}} \left(\pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\right) \nabla_{\mathbf{x}} \left(\pi\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\right)^\top \Big] \\
\mathcal{T}_{i,j,10} = \mathbb{E}_{\mathbf{x}}\big[& \pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \pi\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\big] \mathbf{I}_d \\
& + \mathbb{E}_{\mathbf{x}}\big[\nabla_{\mathbf{x}}^2 \left(\pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\right) \pi\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\big] \\
& + \mathbb{E}_{\mathbf{x}}\big[\pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \nabla_{\mathbf{x}}^2 \left(\pi\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\right)\big] \\
& + \mathbb{E}_{\mathbf{x}}\Big[\nabla_{\mathbf{x}} \left(\pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\right) \nabla_{\mathbf{x}} \left(\pi\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\right)^\top \Big]
\end{aligned}$$

Applying Lemma 24 gives that

$$\left\|\mathcal{T}_{i,j,1}\right\|_2, \left\|\mathcal{T}_{i,j,5}\right\|_2, \left\|\mathcal{T}_{i,j,10}\right\|_2 \leq \mathcal{O}\left(\varepsilon\right)$$

Bounding $\mathcal{T}_{i,i,1}, \mathcal{T}_{i,i,5}, \mathcal{T}_{i,i,10}$. By Lemma 18, we have that

$$\begin{aligned}
\mathcal{T}_{i,i,1} = & \mathbb{E}_{\mathbf{x}}\Big[\pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)^2 \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)^2 \Big] \mathbf{I}_d + 4\mathbb{E}_{\mathbf{x}}\Big[\pi''\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)^2 \sigma'\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)^2 \Big] \left(\bar{\mathbf{v}}_i \bar{\mathbf{w}}_i^\top + \bar{\mathbf{w}}_i \bar{\mathbf{v}}_i^\top\right) \\
& + 2\mathbb{E}_{\mathbf{x}}\Big[\left(\pi''\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)^2 + \pi'''\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)^2 \Big] \bar{\mathbf{v}}_i \bar{\mathbf{v}}_i^\top \\
& + 2\mathbb{E}_{\mathbf{x}}\Big[\left(\sigma'\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)^2 + \sigma''\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\right) \pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)^2 \Big] \bar{\mathbf{w}}_i \bar{\mathbf{w}}_i^\top \\
\mathcal{T}_{i,i,5} = & \mathbb{E}_{\mathbf{x}}\Big[\pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)^2 \sigma'\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)^2 \Big] \mathbf{I}_d + 4\mathbb{E}_{\mathbf{x}}\Big[\pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)^2 \sigma''\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)^2 \Big] \left(\bar{\mathbf{v}}_i \bar{\mathbf{w}}_i^\top + \bar{\mathbf{w}}_i \bar{\mathbf{v}}_i^\top\right) \\
& + 2\mathbb{E}_{\mathbf{x}}\Big[\left(\pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)^2 + \pi''\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)\right) \sigma'\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)^2 \Big] \bar{\mathbf{v}}_i \bar{\mathbf{v}}_i^\top \\
& + 2\mathbb{E}_{\mathbf{x}}\Big[\left(\sigma''\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)^2 + \sigma'''\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\right) \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)^2 \Big] \bar{\mathbf{w}}_i \bar{\mathbf{w}}_i^\top \\
\mathcal{T}_{i,i,10} = & \mathbb{E}_{\mathbf{x}}\big[\pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\big] \mathbf{I}_d \\
& + \mathbb{E}_{\mathbf{x}}\Big[\pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)^2 \sigma'\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)^2 \Big] \left(\bar{\mathbf{v}}_i \bar{\mathbf{w}}_i^\top + \bar{\mathbf{w}}_i \bar{\mathbf{v}}_i^\top\right) \\
& + \mathbb{E}_{\mathbf{x}}\big[\pi''\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma''\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\big] \left(\bar{\mathbf{v}}_i \bar{\mathbf{w}}_i^\top + \bar{\mathbf{w}}_i \bar{\mathbf{v}}_i^\top\right) \\
& + \mathbb{E}_{\mathbf{x}}\Big[\pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)^2 \sigma''\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\Big] \left(\bar{\mathbf{v}}_i \bar{\mathbf{w}}_i^\top + \bar{\mathbf{w}}_i \bar{\mathbf{v}}_i^\top\right) \\
& + \mathbb{E}_{\mathbf{x}}\Big[\pi''\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)^2 \Big] \left(\bar{\mathbf{v}}_i \bar{\mathbf{w}}_i^\top + \bar{\mathbf{w}}_i \bar{\mathbf{v}}_i^\top\right) \\
& + \mathbb{E}_{\mathbf{x}}\Big[\left(\pi'''\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)^2 + 2\pi''\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)\right) \sigma'\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \Big] \bar{\mathbf{v}}_i \bar{\mathbf{v}}_i^\top \\
& + \mathbb{E}_{\mathbf{x}}\Big[\left(2\sigma''\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \sigma'\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)^2 + \sigma'''\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) \sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\right) \pi'\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) \Big] \bar{\mathbf{w}}_i \bar{\mathbf{w}}_i^\top
\end{aligned}$$

Thus, we have that

$$\begin{aligned}
&\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\mathcal{T}_{i,i,1}\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\\
&\quad=\mathbb{E}_{\mathbf{x}}\Big[\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)^2\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)^2\Big]\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\\
&\qquad+2\mathbb{E}_{\mathbf{x}}\Big[\Big(\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)^2+\sigma''\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\Big)\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)^2\Big]\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\\
&\quad=\mathbb{E}_{\mathbf{x}}\Big[\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)^2\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)^2\Big]\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\\
&\qquad+2\mathbb{E}_{\mathbf{x}}\Big[\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)^2\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)^2\Big]\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)+\hat{\mathcal{T}}_{i,i,1}\\
&\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\mathcal{T}_{i,i,5}\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\\
&\quad=\mathbb{E}_{\mathbf{x}}\Big[\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)^2\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)^2\Big]\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\\
&\qquad+2\mathbb{E}_{\mathbf{x}}\Big[\Big(\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)^2+\pi''\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\Big)\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)^2\Big]\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)
\end{aligned}$$

with $\left\|\hat{\mathcal{T}}_{i,i,1}\right\|_2\leq\mathcal{O}\left(\varepsilon\right)$. This gives that

$$\begin{aligned}
\mathbf{u}^\top\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\mathcal{T}_{i,i,1}\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\mathbf{u}&\geq\mathbb{E}_{\mathbf{x}}\Big[\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)^2\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)^2\Big]\left\|\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\mathbf{u}\right\|_2^2-\mathcal{O}\left(\varepsilon\right)\\
\mathbf{u}^\top\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\mathcal{T}_{i,i,5}\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\mathbf{u}&\geq\mathbb{E}_{\mathbf{x}}\Big[\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)^2\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)^2\Big]\left\|\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\mathbf{u}\right\|_2^2
\end{aligned}$$

Imposing the condition that $\left\|\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\mathbf{u}\right\|_2^2\geq\mathcal{Q}^\star\left\|\mathbf{u}\right\|_2^2$ gives that

$$\begin{aligned}
\mathbf{u}^\top\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\mathcal{T}_{i,i,1}\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\mathbf{u}&\geq\mathcal{Q}^\star\mathbb{E}_{\mathbf{x}}\Big[\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)^2\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)^2\Big]\left\|\mathbf{u}\right\|_2^2-\mathcal{O}\left(\varepsilon\right)\\
\mathbf{u}^\top\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\mathcal{T}_{i,i,5}\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\mathbf{u}&\geq\mathcal{Q}^\star\mathbb{E}_{\mathbf{x}}\Big[\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)^2\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)^2\Big]\left\|\mathbf{u}\right\|_2^2
\end{aligned}$$

For $\mathcal{T}_{i,i,10}$, we first notice that

$$\begin{aligned}
\left|\mathbb{E}_{\mathbf{x}}\left[\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right]\right|&\leq\mathcal{O}\left(\varepsilon\right)\\
\left|\mathbb{E}_{\mathbf{x}}\left[\pi''\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma''\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right]\right|&\leq\mathcal{O}\left(\varepsilon\right)\\
\left|\mathbb{E}_{\mathbf{x}}\Big[\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)^2\sigma''\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\Big]\right|&\leq\mathcal{O}\left(\varepsilon\right)\\
\left|\mathbb{E}_{\mathbf{x}}\Big[\Big(\pi'''\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)^2+2\pi''\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\Big)\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\Big]\right|&\leq\mathcal{O}\left(\varepsilon\right)\\
\left|\mathbb{E}_{\mathbf{x}}\Big[\Big(2\sigma''\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)^2+\sigma'''\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\Big)\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\Big]\right|&\leq\mathcal{O}\left(\varepsilon\right)
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
&\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\mathcal{T}_{i,i,10}\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\\
&\quad=\mathbb{E}_{\mathbf{x}}\Big[\Big(\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)^2+\pi''\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\Big)\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)^2\Big]\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\bar{\mathbf{v}}_i\bar{\mathbf{w}}_i^\top\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)
\end{aligned}$$

This implies that

$$\begin{aligned}
&\mathbf{u}_1^\top\left(\mathbf{I}-\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\mathcal{T}_{i,i,10}\left(\mathbf{I}-\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\mathbf{u}_2\\
&\qquad\leq\mathbb{E}_{\mathbf{x}}\Big[\Big(\pi'\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)^2+\pi''\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\Big)\sigma'\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)^2\Big]\left\|\mathbf{u}_1\right\|_2\left\|\mathbf{u}_2\right\|_2
\end{aligned}$$

□

Now, we are ready to prove the fine-tuninig convergence.

Proof of Theorem 3. By the mean-value theorem, we have that

$$\nabla \mathcal{L}\left(\boldsymbol{\theta}\right) = \nabla \mathcal{L}\left(\boldsymbol{\theta}^\star\right) + \nabla^2 \mathcal{L}\left(\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right) = \nabla^2 \mathcal{L}\left(\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^\star\right)$$

for some $\hat{\boldsymbol{\theta}} \in [\boldsymbol{\theta}, \boldsymbol{\theta}^\star]$. The gradient flow dynamic implies that

$$\begin{aligned}
\frac{d}{dt}\left\|\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}^\star\right\|_2^2 &= \left\langle \bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}^\star, \frac{d}{dt}\bar{\boldsymbol{\theta}}(t)\right\rangle \\
&= \left\langle \bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}^\star, \mathbf{N}_{\boldsymbol{\theta}(t)}\frac{d}{dt}\boldsymbol{\theta}(t)\right\rangle \\
&= -\left\langle \mathbf{N}_{\boldsymbol{\theta}(t)}\left(\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}^\star\right), \nabla \mathcal{L}\left(\boldsymbol{\theta}\right)\right\rangle \\
&= -\left\langle \mathbf{N}_{\boldsymbol{\theta}(t)}\left(\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}^\star\right), \nabla^2 \mathcal{L}\left(\hat{\boldsymbol{\theta}}\right)\left(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\right)\right\rangle
\end{aligned}$$

Notice that $\mathbf{N}_{\boldsymbol{\theta}(t)}\left(\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}^\star\right)$ takes the form

$$\mathbf{N}_{\boldsymbol{\theta}(t)}\left(\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}^\star\right) = \begin{bmatrix} \|\mathbf{v}_1\|_2^{-1}\bar{\mathbf{v}}_1 \\ \vdots \\ \|\mathbf{v}_{m^\star}\|_2^{-1}\bar{\mathbf{v}}_{m^\star} \\ \|\mathbf{w}_1\|_2^{-1}\bar{\mathbf{w}}_1 \\ \vdots \\ \|\bar{\mathbf{w}}_{m^\star}\|_2^{-1}\bar{\mathbf{w}}_{m^\star} \end{bmatrix}$$

Thus, we are going to apply Theorem 4 with $N_{\min} = 1 - o(1)\frac{\delta_{\mathbb{P}}}{m^2}, N_{\max} = 1 + o(1)\frac{\delta_{\mathbb{P}}}{m^2}$, and $\alpha_i = \|\mathbf{v}_i\|_2^{-1} \geq \frac{m^2}{m^2 + o(1)\delta_{\mathbb{P}}}, \beta_i = \|\mathbf{w}_i\|_2^{-1} \leq \frac{m^2}{m^2 - o(1)\delta_{\mathbb{P}}}$. This leads to the condition that $\varepsilon \leq o\left(\frac{\mathcal{Q}^{\star 2}}{m^{\star 2}}\right)$ and $\frac{C_{S,0}}{C_{S,1}} \geq \frac{1.05}{\mathcal{Q}^{\star 2}}$. Under such condition, by Theorem 4, we have that

$$\frac{d}{dt}\left\|\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}^\star\right\|_2^2 \leq -\left(1 - o(1)\frac{\delta_{\mathbb{P}}}{m^2}\right)\mathcal{Q}^\star \kappa \left\|\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}^\star\right\|_2^2$$

This shows that $\left\|\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}^\star\right\|_2^2$ decreases monotonically. To find $\mathcal{Q}^\star$, we notice that

$$\begin{aligned}
\left\|\left(\mathbf{I} - \bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\left(\bar{\mathbf{v}}_i(t) - \bar{\mathbf{v}}_i^\star\right)\right\|_2^2 &= \left\langle \bar{\mathbf{v}}_i(t) - \bar{\mathbf{v}}_i^\star, \left(\mathbf{I} - \bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\right)\left(\bar{\mathbf{v}}_i(t) - \bar{\mathbf{v}}_i^\star\right)\right\rangle \\
&= \left\|\bar{\mathbf{v}}_i(t) - \bar{\mathbf{v}}_i^\star\right\|_2^2 - \left(\bar{\mathbf{v}}_i^\top\left(\bar{\mathbf{v}}_i(t) - \bar{\mathbf{v}}_i^\star\right)\right)^2 \\
&= \left\|\bar{\mathbf{v}}_i(t) - \bar{\mathbf{v}}_i^\star\right\|_2^2 - \left(1 - \bar{\mathbf{v}}_i^\top\bar{\mathbf{v}}_i^\star\right)^2 \\
&= \left\|\bar{\mathbf{v}}_i(t) - \bar{\mathbf{v}}_i^\star\right\|_2^2 - \frac{1}{4}\left\|\bar{\mathbf{v}}_i(t) - \bar{\mathbf{v}}_i^\star\right\|_2^4 \\
&= \left(1 - \frac{1}{4}\left\|\bar{\mathbf{v}}_i(t) - \bar{\mathbf{v}}_i^\star\right\|_2^2\right)\left\|\bar{\mathbf{v}}_i(t) - \bar{\mathbf{v}}_i^\star\right\|_2^2
\end{aligned}$$

Similarly, we can obtain that

$$\left\|\left(\mathbf{I} - \bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top\right)\left(\bar{\mathbf{w}}_i(t) - \bar{\mathbf{w}}_i^\star\right)\right\|_2^2 = \left(1 - \frac{1}{4}\left\|\bar{\mathbf{w}}_i(t) - \bar{\mathbf{w}}_i^\star\right\|_2^2\right)\left\|\bar{\mathbf{w}}_i(t) - \bar{\mathbf{w}}_i^\star\right\|_2^2$$

This gives that $\mathcal{Q}^\star = 1 - \frac{1}{4}\max_{i\in[m^\star]}\max\left\{\left\|\bar{\mathbf{v}}_i(0) - \bar{\mathbf{v}}_i^\star\right\|_2^2, \left(\bar{\mathbf{w}}_i(0) - \bar{\mathbf{w}}_i^\star\right)\right\} = 1 - \mathcal{O}\left(\varepsilon\right)$. Thus, the condition that $\varepsilon \leq o\left(\frac{1}{m^{\star 2}}\right)$ and $C_{S,0} \geq 1.1 C_{S,1}$ suffice. This gives us that

$$\frac{d}{dt}\left\|\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}^\star\right\|_2^2 \leq -\left(1 - o(1)\frac{\delta_{\mathbb{P}}}{m^2}\right)\left(1 - \mathcal{O}\left(\varepsilon\right)\right)\kappa\left\|\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}^\star\right\|_2^2 \leq \frac{\kappa}{2}\left\|\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}^\star\right\|_2^2$$

Solving the ODE gives that

$$\left\|\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}^\star\right\|_2^2 \leq \exp\left(-\frac{\kappa t}{2}\right)\left\|\bar{\boldsymbol{\theta}}(0) - \boldsymbol{\theta}^\star\right\|_2^2$$

□

D Auxiliary Results

D.1 Hermite Polynomials

Lemma 19 (Restatement of Lemma 1)**.** *Let* $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$*. For some multi-index* $\mathbf{k} \in \mathbb{N}^n$*, we define the multi-variate Hermite polynomial as*

$$He_{\mathbf{k}}(\mathbf{x}) = \prod_{i=1}^{n} He_{\mathbf{k}[i]}(\mathbf{x}[i])$$

Then we have that

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}[He_{\mathbf{k}}(\mathbf{x})] = \left(\prod_{i=1}^{n} \mathbf{k}[i]!\right) \sum_{\mathbf{M}\in\mathcal{S}} \prod_{i,j=1}^{n} \frac{\mathbf{\Sigma}[i,j]^{\mathbf{M}[i,j]}}{\mathbf{M}[i,j]!}$$

where the set $\mathcal{S}$ *is defined by*

$$\mathcal{S} = \left\{ \mathbf{M} \in \mathbb{N}^{n\times n} : \mathbf{M} = \mathbf{M}^\top, \sum_{j=1}^{n} \mathbf{M}[i,j] = \mathbf{k}[i], \mathbf{M}[i,i] = 0; \forall i \in [n], \right\}$$

Proof. Consider the generating function of Hermite polynomials

$$\exp\left(xt - \frac{t^2}{2}\right) = \sum_{k=0}^{\infty} \frac{He_k(x)}{k!} \cdot t^k$$

Let $x_1, \ldots, x_n \sim \mathcal{N}(0,1)$, then for all $\{t_i\}_{i=1}^n$ we have

$$\begin{aligned}
\exp\left(\sum_{i=1}^{n}\left(x_i t_i - \frac{t_i^2}{2}\right)\right) &= \prod_{i=1}^{n} \exp\left(x_i t_i - \frac{t_i^2}{2}\right) \\
&= \prod_{i=1}^{n}\left(\sum_{k=0}^{\infty} \frac{He_k(x_i)}{k!} \cdot t_i^k\right) \\
&= \sum_{\mathbf{k}\in\mathbb{N}^n}\left(\prod_{i=1}^{n} \frac{He_{k[i]}(x_i)}{k[i]!}\right) \cdot \mathbf{t}^{\mathbf{k}}
\end{aligned}$$

where $\mathbf{k} \in \mathbb{N}^n$ is the multi-index. On the other hand, if we write $x_i = \mathbf{u}_i^\top \mathbf{x}$ for some $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{u}_i$ satisfying $\|\mathbf{u}_i\|_2 = 1$, then we have

$$\begin{aligned}
\mathbb{E}_{x_i\sim\mathcal{N}(0,1)}\left[\exp\left(\sum_{i=1}^{n}\left(x_i t_i - \frac{t_i^2}{2}\right)\right)\right] &= \mathbb{E}_{x_i\sim\mathcal{N}(0,1)}\left[\exp\left(\sum_{i=1}^{n}\left(\mathbf{u}_i^\top \mathbf{x} \cdot t_i - \frac{t_i^2}{2}\right)\right)\right] \\
&= \mathbb{E}_{x_i\sim\mathcal{N}(0,1)}\left[\exp\left(\mathbf{x}^\top\left(\sum_{i=1}^{n} t_i \mathbf{u}_i\right)\right)\right] \exp\left(-\frac{1}{2}\sum_{i=1}^{n} t_i^2\right)
\end{aligned}$$

Since $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we must have that $\mathbf{x}^\top\left(\sum_{i=1}^n t_i \mathbf{u}_i\right) \sim \mathcal{N}\left(\mathbf{0}, \left\|\sum_{i=1}^n t_i \mathbf{u}_i\right\|_2^2\right)$. By the moment-generating function of Gaussian random variable we have that

$$\mathbb{E}_{x_i\sim\mathcal{N}(0,1)}\left[\exp\left(\mathbf{x}^\top\left(\sum_{i=1}^{n} t_i \mathbf{u}_i\right)\right)\right] = \exp\left(\frac{1}{2}\left\|\sum_{i=1}^{n} t_i \mathbf{u}_i\right\|_2^2\right)$$

Thus, we have that

$$\mathbb{E}_{x_i\sim\mathcal{N}(0,1)}\left[\exp\left(\sum_{i=1}^{n}\left(x_i t_i - \frac{t_i^2}{2}\right)\right)\right] = \exp\left(\frac{1}{2}\left(\left\|\sum_{i=1}^{n} t_i \mathbf{u}_i\right\|_2^2 - \sum_{i=1}^{n} t_i^2\right)\right) = \exp\left(\frac{1}{2}\sum_{i\neq j} \mathbf{u}_i^\top \mathbf{u}_j t_i t_j\right)$$

Applying Taylor's expansion gives

$$\exp\left(\frac{1}{2}\sum_{i\neq j}\mathbf{u}_i^\top\mathbf{u}_j t_i t_j\right)=\sum_{\ell=0}^{\infty}\frac{1}{\ell!}\left(\sum_{i<j}\mathbf{u}_i^\top\mathbf{u}_j t_i t_j\right)^\ell$$

Combining the results gives

$$\sum_{\mathbf{k}\in\mathbb{N}^d}\mathbb{E}_{x_i\sim\mathcal{N}(0,1)}\left[\prod_{i=1}^n He_{\mathbf{k}[i]}(x_i)\right]\prod_{i=1}^n\frac{t_i^{\mathbf{k}[i]}}{\mathbf{k}[i]!}=\sum_{\ell=0}^{\infty}\frac{1}{\ell!}\left(\sum_{i<j}\mathbf{u}_i^\top\mathbf{u}_j t_i t_j\right)^\ell$$

We intend to find out the cofficients of term $\prod_{i=1}^n t_i^{\mathbf{k}[i]}$ on the right-hand side. Notice that such term must only appears for term with ℓ satisfying $2\ell=\|\mathbf{k}\|_1$. By the multinomial theorem we have that

$$\left(\sum_{i\neq j}\mathbf{u}_i^\top\mathbf{u}_j t_i t_j\right)^\ell=\sum_{\mathbf{k}':\|\mathbf{k}'\|_1=\ell}\ell!\cdot\prod_{i<j}\frac{\left(\mathbf{u}_i^\top\mathbf{u}_j\right)^{\mathbf{k}'[i,j]}}{\mathbf{k}'[i,j]!}t_i^{\mathbf{k}'[i,j]}t_j^{\mathbf{k}'[i,j]}$$

Therefore, we must have that

$$\mathbb{E}_{x_i\sim\mathcal{N}(0,1)}\left[\prod_{i=1}^n He_{\mathbf{k}[i]}(x_i)\right]=\left(\prod_{i=1}^n\mathbf{k}[i]!\right)\sum_{\mathbf{M}\in\mathcal{S}}\prod_{i<j}\frac{\mathbb{E}\big[x_i x_j\big]^{\mathbf{M}[i,j]}}{\mathbf{M}[i,j]!}$$

where $\mathcal{S}$ is given by

$$\mathcal{S}=\left\{\mathbf{M}\in\mathbb{N}^{n\times n}:\mathtt{Tr}\left(\mathbf{M}\right)=0;\forall i\in[n],\sum_{j=1}^n\mathbf{M}[i,j]=\mathbf{k}[i]\right\}$$

Using vector notations, we have that

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\big[He_{\mathbf{k}}(\mathbf{x})\big]=\left(\prod_{i=1}^n\mathbf{k}[i]!\right)\sum_{\mathbf{M}\in\mathcal{S}}\prod_{i<j}\frac{\mathbf{\Sigma}[i,j]^{\mathbf{M}[i,j]}}{\mathbf{M}[i,j]!}$$

□

Lemma 20 (Parseval's Identity)**.** *Let $f(x)$ be given such that $\mathbb{E}_{x\sim\mathcal{N}(0,1)}\big[f^{(\ell)}(x)^2\big]\leq B$, and let c_k be the kth Hermite coefficient of $f(x)$. Then we have that*

$$c_{k+\ell}^2\leq B\cdot k!;\quad\forall k\geq 0$$

Proof. Taking the Hermite expansion of $f^{(\ell)}$ gives

$$f^{(\ell)}=\sum_{k=0}^{\infty}\frac{c_k'}{k!}He_k(x);\quad c_k'=\mathbb{E}_{x\sim\mathcal{N}(0,1)}\Big[f^{(\ell+k)}(x)\Big]=c_{k+\ell}$$

Therefore, we have that

$$\mathbb{E}_{x\sim\mathcal{N}(0,1)}\Big[f^{(\ell)}(x)^2\Big]=\sum_{k=0}^{\infty}\frac{c_k'^{\,2}}{k!}=\sum_{k=0}^{\infty}\frac{c_{k+\ell}}{k!}\leq B$$

This implies that $c_{k+\ell}^2\leq B\cdot k!$ since $c_{k+\ell}^2\geq 0$ for all k. □

Lemma 21. *Let $\mathbf{v}_1,\mathbf{v}_2,\mathbf{w}_1,\mathbf{w}_2\in\mathbb{R}^d$ be vectors of unit norm such that*

$$\max\left\{\left|\mathbf{v}_1^\top\mathbf{w}_1\right|,\left|\mathbf{v}_1^\top\mathbf{w}_2\right|,\left|\mathbf{v}_2^\top\mathbf{w}_1\right|,\left|\mathbf{v}_2^\top\mathbf{w}_2\right|\right\}\leq\delta_r$$

with some $\delta_r\in(0,1)$. Let $\{h_k\}_{k=0}^\infty,\{h_k'\}_{k=0}^\infty$ be two sequences of real numbers such that

$$\sum_{k=0}^{\infty}\frac{h_{k+a}h_{k+b}'}{k!}\leq\mathcal{O}(1);\forall a+b\leq 6,a,b\in\mathbb{N}\cup\{0\}\tag{25}$$

Then we have that

$$\sum_{k,\ell=0}^{\infty} \frac{h_k h'_\ell}{k!\ell!} \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top \mathbf{x}\right) He_\ell\left(\mathbf{v}_2^\top \mathbf{x}\right) He_3\left(\mathbf{w}_1^\top \mathbf{x}\right) He_3\left(\mathbf{w}_2^\top \mathbf{x}\right)\right] \tag{26}$$
$$= 6\sum_{k=0}^{\infty} \frac{h_k h'_k}{k!}\left(\mathbf{v}_1^\top \mathbf{v}_2\right)^k \left(\mathbf{w}_1^\top \mathbf{w}_2\right)^3 \pm \mathcal{O}\left(\delta_r^2 \left(\mathbf{w}_1^\top \mathbf{w}_2\right)^2 + \delta_r^4\right)$$

$$\sum_{k,\ell=0}^{\infty} \frac{h_k h'_\ell}{k!\ell!} \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top \mathbf{x}\right) He_\ell\left(\mathbf{v}_2^\top \mathbf{x}\right) He_2\left(\mathbf{w}_1^\top \mathbf{x}\right) He_2\left(\mathbf{w}_2^\top \mathbf{x}\right)\right] \tag{27}$$
$$= 2\sum_{k=0}^{\infty} \frac{h_k h_{k'}}{k!}\left(\mathbf{v}_1^\top \mathbf{v}_2\right)^k \left(\mathbf{w}_1^\top \mathbf{w}_2\right)^2 \pm \mathcal{O}\left(\delta_r^2 \cdot \left|\mathbf{w}_1^\top \mathbf{w}_2\right| + \delta_r^4\right)$$

$$\sum_{k,\ell=0}^{\infty} \frac{h_k h'_\ell}{k!\ell!} \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top \mathbf{x}\right) He_\ell\left(\mathbf{v}_2^\top \mathbf{x}\right) He_2\left(\mathbf{w}_1^\top \mathbf{x}\right) He_3\left(\mathbf{w}_2^\top \mathbf{x}\right)\right]$$
$$= 6\sum_{k=0}^{\infty} \frac{h_{k+1} h'_k}{k!}\left(\mathbf{v}_1^\top \mathbf{v}_2\right)^k \left(\mathbf{w}_1^\top \mathbf{w}_2\right)^2 \mathbf{v}_1^\top \mathbf{w}_2 \tag{28}$$
$$+ 6\sum_{k=0}^{\infty} \frac{h_k h'_{k+1}}{k!}\left(\mathbf{v}_1^\top \mathbf{v}_2\right)^k \left(\mathbf{w}_1^\top \mathbf{w}_2\right)^2 \mathbf{v}_2^\top \mathbf{w}_2 \pm \mathcal{O}\left(\delta_r^3\right).$$

Proof. The general idea of proving this lemma is to use Lemma 1. In particular, in our case we have that

$$\mathbf{\Sigma}[i,j] = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{w}_1, \mathbf{w}_2]^\top [\mathbf{v}_1, \mathbf{v}_2, \mathbf{w}_1, \mathbf{w}_2] \in \mathbb{R}^{4\times 4}$$

For the convenience of the analysis, we denote

$$\gamma_1 = \mathbf{v}_1^\top \mathbf{v}_2; \gamma_2 = \mathbf{w}_1^\top \mathbf{w}_2; \zeta_{ij} = \mathbf{v}_i^\top \mathbf{w}_j$$

By the assumption, we have that $|\zeta_{ij}| \le \delta_r$.

Proof of (26). We start from the first equation. In particular we need to study

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top \mathbf{x}\right) He_\ell\left(\mathbf{v}_2^\top \mathbf{x}\right) He_3\left(\mathbf{w}_1^\top \mathbf{x}\right) He_3\left(\mathbf{w}_2^\top \mathbf{x}\right)\right]$$

By Lemma 1, we need to enumerate all $\mathcal{S}$, which is essentially the symmetric matrix $\mathbf{M}$ with zero diagonal and non-negative entries whose row-sum equal to the vector $[k, \ell, 3, 3]$. This is equivalent to construct a weighted graph with four nodes and node degree $[k, \ell, 3, 3]$. Thus, it suffice to consider cases $k = \ell, |k-\ell| = 2, |k-\ell| = 4$, and $|k-\ell| = 6$. Due to symmetry between k and ℓ, we will first study the case $k \ge \ell$ and switch the indices to obtain all cases.

Case $k = \ell$. The node $\mathbf{w}_1$ and $\mathbf{w}_2$ has a total degree of 6, therefore, the pair of node $\mathbf{v}_1$ and $\mathbf{v}_2$ can have outgoing degree at most 6. Thus, the condition can be broken down into $\mathbf{M}[1,2] \in \{k, k-1, k-2, k-3\}$. When $\mathbf{M}[1,2] = k$, we have that $\mathbf{M}[3,4] = 3$, and all other edges 0. When $\mathbf{M}[1,2] = k-1$, we have that $\mathbf{M}[3,4] = 2$ and either $(\mathbf{M}[1,3], \mathbf{M}[2,4]) = (1,1)$ or $(\mathbf{M}[1,4], \mathbf{M}[2,3]) = (1,1)$. When $\mathbf{M}[1,2] = k-2$, we have that $\mathbf{M}[3,4] = 1$. Here we can have $(\mathbf{M}[1,3], \mathbf{M}[2,4]) = (2,2), (\mathbf{M}[1,4], \mathbf{M}[2,3]) = (2,2)$, or $(\mathbf{M}[1,3], \mathbf{M}[2,4], \mathbf{M}[1,4], \mathbf{M}[2,3]) = (1,1,1,1)$. When $\mathbf{M}[1,2] = k-3$, then $\mathbf{M}[3,4] = 0$. Thus we have $(\mathbf{M}[1,3], \mathbf{M}[2,4]) = (3,3), (\mathbf{M}[1,4], \mathbf{M}[2,3]) = (3,3)$ or $(\mathbf{M}[1,3], \mathbf{M}[2,4], \mathbf{M}[1,4], \mathbf{M}[2,3]) = (1,1,2,2)$, or $(\mathbf{M}[1,3], \mathbf{M}[2,4], \mathbf{M}[1,4], \mathbf{M}[2,3]) = (2,2,1,1)$. Plugging the possibilities into Lemma 1 gives that, under the case $k = \ell$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top \mathbf{x}\right) He_\ell\left(\mathbf{v}_2^\top \mathbf{x}\right) He_3\left(\mathbf{w}_1^\top \mathbf{x}\right) He_3\left(\mathbf{w}_2^\top \mathbf{x}\right)\right] \\
&\quad = 6k!\gamma_1^k\gamma_2^3 + 18P(k,1)k!\gamma_1^{k-1}\gamma_2^2\left(\zeta_{11}\zeta_{22} + \zeta_{12}\zeta_{21}\right) \\
&\qquad + 9P(k,2)k!\gamma_1^{k-2}\gamma_2\left(\zeta_{11}^2\zeta_{22}^2 + 4\zeta_{11}\zeta_{12}\zeta_{21}\zeta_{22} + \zeta_{12}^2\zeta_{21}^2\right) \\
&\qquad + P(k,3)k!\gamma_1^{k-3}\left(\zeta_{11}^3\zeta_{22}^3 + \zeta_{12}^3\zeta_{21}^3 + 9\zeta_{11}^2\zeta_{12}\zeta_{21}\zeta_{22}^2 + 9\zeta_{11}\zeta_{12}^2\zeta_{21}^2\zeta_{22}\right) \\
&\quad = 6k!\gamma_1^k\gamma_2^3 \pm \mathcal{O}\left(\delta_r^2\gamma_2^2\right)\cdot k!P(\ell,1) \pm \mathcal{O}\left(\delta_r^4\right)\cdot k!\left(P(\ell,2) + P(\ell,3)\right)
\end{aligned}$$

where $P(k,a) = \frac{k!}{(k-a)!}$ if $k \geq a$ and $P(k,a) = 0$ if $k < a$ represents the permutation number.

Case $k = \ell + 2$. In this case we have that $\mathbf{M}[1,2] \in \{\ell, \ell-1, \ell-2\}$. If $\mathbf{M}[1,2] = \ell$, then $\mathbf{M}[3,4] = 2$. Here we have that $(\mathbf{M}[1,3], \mathbf{M}[1,4]) = (1,1)$. If $\mathbf{M}[1,2] = \ell - 1$, then $\mathbf{M}[3,4] = 1$. Here we have $(\mathbf{M}[1,3], \mathbf{M}[1,4], \mathbf{M}[2,4]) = (2,1,1)$ or $(\mathbf{M}[1,3], \mathbf{M}[1,4], \mathbf{M}[2,3]) = (1,2,1)$. If $\mathbf{M}[1,2] = \ell - 2$, then $\mathbf{M}[3,4] = 0$. Here we have $(\mathbf{M}[1,3], \mathbf{M}[1,4], \mathbf{M}[2,4]) = (3,1,2)$ or $(\mathbf{M}[1,3], \mathbf{M}[1,4], \mathbf{M}[2,3]) = (1,3,2)$ or $(\mathbf{M}[1,3], \mathbf{M}[1,4], \mathbf{M}[2,3], \mathbf{M}[2,4]) = (2,2,1,1)$. Gathering all possibilities gives that, under the case $k = \ell + 2$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top \mathbf{x}\right) He_\ell\left(\mathbf{v}_2^\top \mathbf{x}\right) He_3\left(\mathbf{w}_1^\top \mathbf{x}\right) He_3\left(\mathbf{w}_2^\top \mathbf{x}\right)\right] \\
&\quad = 18k!\gamma_1^\ell \gamma_2^2 \zeta_{11}\zeta_{12} + 18P(\ell,1)k!\gamma_1^{\ell-1}\gamma_2\zeta_{11}\zeta_{12}\left(\zeta_{11}\zeta_{22} + \zeta_{12}\zeta_{21}\right) \\
&\qquad + 3P(\ell,2)k!\gamma_1^{\ell-2}\zeta_{11}\zeta_{12}\left(\zeta_{11}^2\zeta_{22}^2 + \zeta_{12}^2\zeta_{21}^2\right) \\
&\qquad + 9P(\ell,2)k!\gamma_1^{\ell-2}\zeta_{11}^2\zeta_{12}^2\zeta_{21}\zeta_{22} \\
&\quad = \pm\mathcal{O}\left(\delta_r^2\gamma_2^2\right)\cdot k! \pm \mathcal{O}\left(\delta_r^4\right)\cdot k!\left(P(\ell,1) + P(\ell,2)\right)
\end{aligned}$$

Case $k = \ell + 4$. In this case we have that $\mathbf{M}[1,2] \in \{\ell, \ell-1\}$. If $\mathbf{M}[1,2] = \ell$, then $\mathbf{M}[3,4] = 1$. Here we have that $(\mathbf{M}[1,3], \mathbf{M}[1,4]) = (2,2)$. If $\mathbf{M}[1,2] = \ell - 1$, then $\mathbf{M}[3,4] = 0$. Here we have that $(\mathbf{M}[1,3], \mathbf{M}[1,4], \mathbf{M}[2,4]) = (3,2,1)$ or $(\mathbf{M}[1,3], \mathbf{M}[1,4], \mathbf{M}[2,3]) = (2,3,1)$. Gathering all possibilities gives that, under the case $k = \ell + 4$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top \mathbf{x}\right) He_\ell\left(\mathbf{v}_2^\top \mathbf{x}\right) He_3\left(\mathbf{w}_1^\top \mathbf{x}\right) He_3\left(\mathbf{w}_2^\top \mathbf{x}\right)\right] \\
&\quad = 9k!\gamma_1^\ell\gamma_2\zeta_{11}^2\zeta_{12}^2 + 3P(\ell,1)k!\gamma_1^{\ell-1}\zeta_{11}^2\zeta_{12}^2\left(\zeta_{11}\zeta_{22} + \zeta_{21}\zeta_{12}\right) \\
&\quad = \pm\mathcal{O}\left(\delta_r^4\right)\cdot k!\left(1 + P(\ell,1)\right)
\end{aligned}$$

Case $k = \ell + 6$. In this case we have that $\mathbf{M}[1,2] = \ell, \mathbf{M}[3,4] = 0$ and $\mathbf{M}[1,3] = \mathbf{M}[1,4] = 3$. Thus, if $k = \ell + 6$, we have that

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top \mathbf{x}\right) He_\ell\left(\mathbf{v}_2^\top \mathbf{x}\right) He_3\left(\mathbf{w}_1^\top \mathbf{x}\right) He_3\left(\mathbf{w}_2^\top \mathbf{x}\right)\right] = k!\gamma_1^\ell\gamma_2\zeta_{11}^3\zeta_{12}^3 = \pm\mathcal{O}\left(\delta_r^4\right)\cdot k!$$

Putting things together, we have that

$$\begin{aligned}
&\sum_{k,\ell=0}^{\infty}\frac{h_k h'_\ell}{k!\ell!}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top \mathbf{x}\right) He_\ell\left(\mathbf{v}_2^\top \mathbf{x}\right) He_3\left(\mathbf{w}_1^\top \mathbf{x}\right) He_3\left(\mathbf{w}_2^\top \mathbf{x}\right)\right] \\
&\quad = 6\sum_{k=0}^{\infty}\frac{h_k h'_k}{k!}\gamma_1^k\gamma_2^3 \pm \mathcal{O}\left(\delta_r^2\right)\sum_{k=0}^{\infty}\frac{h_k h'_k}{k!} \pm \mathcal{O}\left(\delta_r^2\gamma_2^2\right)\sum_{k=0}^{\infty}\frac{h_{k+1}h'_{k+1}}{k!} \pm \mathcal{O}\left(\delta_r^4\right)\sum_{k=0}^{\infty}\frac{c_{k+2}c'_{k+2}}{k!} \\
&\qquad \pm \mathcal{O}\left(\delta_r^4\right)\sum_{k=0}^{\infty}\frac{h_{k+3}h'_{k+3}}{k!} \pm \mathcal{O}\left(\delta_r^4\right)\sum_{k=0}^{\infty}\frac{h_k h'_{k+2} + h'_k h_{k+2}}{k!} \pm \mathcal{O}\left(\delta_r^4\right)\sum_{k=0}^{\infty}\frac{h_{k+1}h'_{k+3} + h'_{k+1}h_{k+3}}{k!} \\
&\qquad \pm \mathcal{O}\left(\delta_r^4\right)\sum_{k=0}^{\infty}\frac{h_{k+2}h'_{k+4} + h'_{k+2}h_{k+4}}{k!} \pm \mathcal{O}\left(\delta_r^4\right)\sum_{k=0}^{\infty}\frac{h_k h'_{k+4} + c'_k c_{k+4}}{k!} \\
&\qquad \pm \mathcal{O}\left(\delta_r^4\right)\sum_{k=0}^{\infty}\frac{h_{k+1}h'_{k+5} + h'_{k+1}h_{k+5}}{k!} \pm \mathcal{O}\left(\delta_r^4\right)\sum_{k=0}^{\infty}\frac{h_k h'_{k+6} + h'_k h_{k+6}}{k!} \\
&\quad = 6\sum_{k=0}^{\infty}\frac{h_k h'_k}{k!}\gamma_1^k\gamma_2^3 \pm \mathcal{O}\left(\delta_r^2\gamma_2^2 + \delta_r^4\right)
\end{aligned}$$

Proof of (27). Similar to before, the combination of k, ℓ can be $k = \ell$, $|k-\ell| = 2$, or $|k-\ell| = 4$ due to the total degree of $\mathbf{w}_1$ and $\mathbf{w}_2$ is 4. We study the case $k \geq \ell$.

Case $k = \ell$. In this case, we have $\mathbf{M}[1,2] \in \{\ell, \ell-1, \ell-2\}$. If $\mathbf{M}[1,2] = \ell$, then $\mathbf{M}[3,4] = 2$ and all other edges are 0. If $\mathbf{M}[1,2] = \ell - 1$, then $\mathbf{M}[3,4] = 1$, and either $(\mathbf{M}[1,3], \mathbf{M}[2,4]) = (1,1)$ or $(\mathbf{M}[1,4], \mathbf{M}[2,3]) = (1,1)$. If $\mathbf{M}[1,2] = \ell - 2$, then $\mathbf{M}[3,4] = 0$. Here we can have $(\mathbf{M}[1,3], \mathbf{M}[2,4]) = (2,2)$ or $(\mathbf{M}[1,4], \mathbf{M}[2,3]) = (2,2)$ or $(\mathbf{M}[1,3], \mathbf{M}[1,4], \mathbf{M}[2,3], \mathbf{M}[2,4]) = (1,1,1,1)$. Gathering all possibilities gives that, under the case $k = \ell$, we

have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top\mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top\mathbf{x}\right)He_2\left(\mathbf{w}_1^\top\mathbf{x}\right)He_2\left(\mathbf{w}_2^\top\mathbf{x}\right)\right]\\
&\quad= 2k!\gamma_1^k\gamma_2^2 + 4P(\ell,1)k!\gamma_1^{k-1}\gamma_2\left(\zeta_{11}\zeta_{22}+\zeta_{12}\zeta_{21}\right)\\
&\qquad+ P(\ell,2)k!\gamma_1^{k-2}\left(\zeta_{11}^2\zeta_{22}^2+\zeta_{12}^2\zeta_{21}^2+4\zeta_{11}\zeta_{12}\zeta_{21}\zeta_{22}\right)\\
&\quad= 2k!\gamma_1^k\gamma_2^2 \pm \mathcal{O}\left(\delta_r^2\gamma_2\right)\cdot k!P(\ell,1)\pm\mathcal{O}\left(\delta_r^4\right)\cdot k!P(\ell,2)
\end{aligned}$$

In the case $\zeta_{22}=0$, we denote $\hat{\zeta}=\max\left\{|\zeta_{21}|,|\zeta_{12}|\right\}$ we have that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top\mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top\mathbf{x}\right)He_2\left(\mathbf{w}_1^\top\mathbf{x}\right)He_2\left(\mathbf{w}_2^\top\mathbf{x}\right)\right]\\
&\quad= 2k!\gamma_1^k\gamma_2^2 + 4P(\ell,1)k!\gamma_1^{k-1}\gamma_2\zeta_{12}\zeta_{21} + P(\ell,2)k!\gamma_1^{k-2}\zeta_{12}^2\zeta_{21}^2\\
&\quad= 2k!\gamma_1^k\gamma_2^2 \pm \mathcal{O}\left(\hat{\zeta}^2\gamma_2\right)\cdot k!P(\ell,1)\pm\mathcal{O}\left(\hat{\zeta}^4\right)\cdot k!P(\ell,2)
\end{aligned}$$

Case $k=\ell+2$. In this case we have $\mathbf{M}[1,2]\in\{\ell,\ell-1\}$. If $\mathbf{M}[1,2]=\ell$, then $\mathbf{M}[3,4]=1$ and $(\mathbf{M}[1,3],\mathbf{M}[1,4])=(1,1)$. If $\mathbf{M}[1,2]=\ell-1$, then $\mathbf{M}[3,4]=0$ and either $(\mathbf{M}[1,3],\mathbf{M}[1,4],(\mathbf{M})[2,3])=(1,2,1)$ or $(\mathbf{M}[1,3],\mathbf{M}[1,4],(\mathbf{M})[2,4])=(2,1,1)$. Gathering all possibilities gives that, under the case $k=\ell+2$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top\mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top\mathbf{x}\right)He_2\left(\mathbf{w}_1^\top\mathbf{x}\right)He_2\left(\mathbf{w}_2^\top\mathbf{x}\right)\right]\\
&\quad= 4k!\gamma_1^\ell\gamma_2\zeta_{11}\zeta_{12} + 2P(\ell,1)k!\gamma_1^{\ell-1}\zeta_{11}\zeta_{12}\left(\zeta_{11}\zeta_{22}+\zeta_{12}\zeta_{21}\right)\\
&\quad= \pm\mathcal{O}\left(\delta_r^2\gamma_2\right)\cdot k! \pm \mathcal{O}\left(\delta_r^4\right)\cdot k!P(\ell,1)
\end{aligned}$$

In the case where $\zeta_{22}=0$, we have that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top\mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top\mathbf{x}\right)He_2\left(\mathbf{w}_1^\top\mathbf{x}\right)He_2\left(\mathbf{w}_2^\top\mathbf{x}\right)\right]\\
&\quad= 4k!\gamma_1^\ell\gamma_2\zeta_{11}\zeta_{12} + 2P(\ell,1)k!\gamma_1^{\ell-1}\zeta_{11}\zeta_{12}^2\zeta_{21}\\
&\quad= \pm\mathcal{O}\left(\delta_r\hat{\zeta}\gamma_2\right)\cdot k! \pm \mathcal{O}\left(\hat{\zeta}^3\right)\cdot k!P(\ell,1)
\end{aligned}$$

Case $k=\ell+4$. In this case we can only have $\mathbf{M}[1,2]=\ell$, $\mathbf{M}[3,4]=0$, and $\mathbf{M}[1,3]=\mathbf{M}[1,4]=2$. Thus if $k=\ell+4$, we have

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top\mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top\mathbf{x}\right)He_2\left(\mathbf{w}_1^\top\mathbf{x}\right)He_2\left(\mathbf{w}_2^\top\mathbf{x}\right)\right] = k!\gamma_1^\ell\zeta_{11}^2\zeta_{12}^2 = \pm\mathcal{O}\left(\delta_r^4\right)\cdot k!$$

In the case where $\zeta_{22}=0$, we have that

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top\mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top\mathbf{x}\right)He_2\left(\mathbf{w}_1^\top\mathbf{x}\right)He_2\left(\mathbf{w}_2^\top\mathbf{x}\right)\right] = k!\gamma_1^\ell\zeta_{11}^2\zeta_{12}^2 = \pm\mathcal{O}\left(\delta_r^2\hat{\zeta}^2\right)\cdot k!$$

Putting things together, we have that

$$\begin{aligned}
&\sum_{k,\ell=0}^{\infty}\frac{h_k h'_\ell}{k!\ell!}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top\mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top\mathbf{x}\right)He_2\left(\mathbf{w}_1^\top\mathbf{x}\right)He_2\left(\mathbf{w}_2^\top\mathbf{x}\right)\right]\\
&\quad= 2\sum_{k=0}^{\infty}\frac{h_k h_{k'}}{k!}\gamma_1^k\gamma_2^2 \pm \mathcal{O}\left(\delta_r^2\gamma_2\right)\sum_{k=0}^{\infty}\frac{h_{k+1}h'_{k+1}}{k!} \pm \mathcal{O}\left(\delta_r^4\right)\sum_{k=0}^{\infty}\frac{h_{k+2}h_{k+2}^2}{k!}\\
&\qquad\pm \mathcal{O}\left(\delta_r^2\gamma_2\right)\sum_{k=0}^{\infty}\frac{h_k h_{k+2}+h'_k h_{k+2}}{k!} \pm \mathcal{O}\left(\delta_r^4\right)\sum_{k=0}^{\infty}\frac{h_{k+1}h_{k+3}+h'_{k+1}h'_{k+3}}{k!}\\
&\qquad\pm \mathcal{O}\left(\delta_r^4\right)\sum_{k=0}^{\infty}\frac{h_k h_{k+4}+h'_k h_{k+4}}{k!}\\
&\quad= 2\sum_{k=0}^{\infty}\frac{h_k h_{k'}}{k!}\gamma_1^k\gamma_2^2 \pm \mathcal{O}\left(\delta_r^2\gamma_2+\delta_r^4\right)
\end{aligned}$$

In the case where $\zeta_{22}=0$, we have that

$$\begin{aligned}
&\sum_{k,\ell=0}^{\infty}\frac{h_k h'_\ell}{k!\ell!}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top\mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top\mathbf{x}\right)He_2\left(\mathbf{w}_1^\top\mathbf{x}\right)He_2\left(\mathbf{w}_2^\top\mathbf{x}\right)\right]\\
&\quad=2\sum_{k=0}^{\infty}\frac{h_k h_{k'}}{k!}\gamma_1^k\gamma_2^2\pm\mathcal{O}\left(\hat{\zeta}^2\gamma_2\right)\sum_{k=0}^{\infty}\frac{h_{k+1}h'_{k+1}}{k!}\pm\mathcal{O}\left(\hat{\zeta}^4\right)\sum_{k=0}^{\infty}\frac{h_{k+2}h^2_{k+2}}{k!}\\
&\qquad\pm\mathcal{O}\left(\delta_r\hat{\zeta}\gamma_2\right)\sum_{k=0}^{\infty}\frac{h_k h'_{k+2}}{k!}\pm\mathcal{O}\left(\hat{\zeta}^3\right)\sum_{k=0}^{\infty}\frac{h_{k+1}h'_{k+3}}{k!}\\
&\qquad\pm\mathcal{O}\left(\delta_r^2\hat{\zeta}^2\right)\sum_{k=0}^{\infty}\frac{h_k h'_{k+4}}{k!}\\
&\quad=2\sum_{k=0}^{\infty}\frac{h_k h_{k'}}{k!}\gamma_1^k\gamma_2^2\pm\mathcal{O}\left(\delta_r\hat{\zeta}\gamma_2+\delta_r\hat{\zeta}_r^2\right)
\end{aligned}$$

Proof of (28). We notice that in this case k,ℓ must satisfy $|k-\ell|\in\{1,3,5\}$. Similar to before, we assume that $k\geq\ell$.

Case $k=\ell+1$. In this case $\mathbf{M}[1,2]\in\{\ell,\ell-1,\ell-2\}$. If $\mathbf{M}[1,2]=\ell$, then $\mathbf{M}[3,4]=2$ and $\mathbf{M}[1,4]=1$. If $\mathbf{M}[1,2]=\ell-1$, then $\mathbf{M}[3,4]=1$, and either $(\mathbf{M}[1,3],\mathbf{M}[1,4],\mathbf{M}[2,4])=(1,1,1)$ or $(\mathbf{M}[1,4],\mathbf{M}[2,3])=(2,1)$. If $\mathbf{M}[1,2]=\ell-2$, then $\mathbf{M}[3,4]=0$. Here we have $(\mathbf{M}[1,4],\mathbf{M}[2,3])=(3,2)$ or $(\mathbf{M}[1,3],\mathbf{M}[1,4],\mathbf{M}[2,3],\mathbf{M}[2,4])=(1,2,1,1)$ or $(\mathbf{M}[1,3],\mathbf{M}[1,4],\mathbf{M}[2,4])=(2,1,2)$. Gathering all possibilities gives that, under the case $k=\ell+1$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top\mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top\mathbf{x}\right)He_2\left(\mathbf{w}_1^\top\mathbf{x}\right)He_3\left(\mathbf{w}_2^\top\mathbf{x}\right)\right]\\
&\quad=6k!\gamma_1^\ell\gamma_2^2\zeta_{12}+12P(\ell,1)k!\gamma_1^{\ell-1}\gamma_2\zeta_{12}\left(2\zeta_{11}\zeta_{22}+\zeta_{12}\zeta_{21}\right)\\
&\qquad+P(\ell,2)k!\gamma_1^{\ell-2}\zeta_{12}\left(\zeta_{12}^2\zeta_{21}^2+3\zeta_{11}^2\zeta_{22}^2+6\zeta_{11}\zeta_{12}\zeta_{21}\zeta_{22}\right)\\
&\quad=6k!\gamma_1^\ell\gamma_2^2\zeta_{12}\pm\mathcal{O}\left(\delta_r^3\right)\left(P(\ell,1)+P(\ell,2)\right)
\end{aligned}$$

Case $k=\ell+3$. In this case we have $\mathbf{M}[1,2]\in\{\ell,\ell-1\}$. If $\mathbf{M}[1,2]=\ell$, then $\mathbf{M}[3,4]=1$, and $(\mathbf{M}[1,3],\mathbf{M}[1,4])=(1,2)$. If $\mathbf{M}[1,2]=\ell-1$, then $\mathbf{M}[3,4]=0$, and we have $(\mathbf{M}[1,3],\mathbf{M}[1,4],\mathbf{M}[2,3])=(1,3,1)$ or $(\mathbf{M}[1,3],\mathbf{M}[1,4],\mathbf{M}[2,4])=(2,2,1)$. Gathering all possibilities gives that, under the case $k=\ell+3$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top\mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top\mathbf{x}\right)He_2\left(\mathbf{w}_1^\top\mathbf{x}\right)He_3\left(\mathbf{w}_2^\top\mathbf{x}\right)\right]\\
&\quad=6k!\gamma_1^\ell\gamma_2\zeta_{11}\zeta_{12}^2+6P(\ell,1)k!\gamma_1^{\ell-1}\zeta_{11}\zeta_{12}^2\left(3\zeta_{21}+2\zeta_{22}\right)\\
&\quad=\pm\mathcal{O}\left(\delta_r^3\right)\left(1+P(\ell,1)\right)
\end{aligned}$$

Case $k=\ell+5$. In this case we must have that $\mathbf{M}[1,2]=\ell,\mathbf{M}[3,4]=0$, and $\mathbf{M}[1,3]=2,\mathbf{M}[1,4]=3$. Thus

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top\mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top\mathbf{x}\right)He_2\left(\mathbf{w}_1^\top\mathbf{x}\right)He_3\left(\mathbf{w}_2^\top\mathbf{x}\right)\right]=k!\gamma_1^\ell\zeta_{11}^2\zeta_{12}^3=\pm\mathcal{O}\left(\delta_r^5\right)$$

Putting things together gives

$$\begin{aligned}
&\sum_{k,\ell=0}^{\infty}\frac{h_k h'_\ell}{k!\ell!}\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[He_k\left(\mathbf{v}_1^\top\mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top\mathbf{x}\right)He_2\left(\mathbf{w}_1^\top\mathbf{x}\right)He_3\left(\mathbf{w}_2^\top\mathbf{x}\right)\right]\\
&\quad=6\sum_{k=0}^{\infty}\frac{h_{k+1}h'_k}{k!}\gamma_1^k\gamma_2^2\zeta_{12}+6\sum_{k=0}^{\infty}\frac{h_k h'_{k+1}}{k!}\gamma_1^k\gamma_2^2\zeta_{22}\pm\mathcal{O}\left(\delta_r^3\right)\sum_{k=0}^{\infty}\frac{h_{k+1}h'_{k+2}+h'_{k+1}h_{k+2}}{k!}\\
&\qquad\pm\mathcal{O}\left(\delta_r^3\right)\sum_{k=0}^{\infty}\frac{h_{k+2}h'_{k+3}+h'_{k+2}h_{k+3}}{k!}\pm\mathcal{O}\left(\delta_r^3\right)\sum_{k=0}^{\infty}\frac{h_k h'_{k+3}+h'_k h_{k+3}}{k!}\\
&\qquad\pm\mathcal{O}\left(\delta_r^3\right)\sum_{k=0}^{\infty}\frac{h_{k+1}h'_{k+4}+h'_{k+1}h_{k+4}}{k!}\pm\mathcal{O}\left(\delta_r^3\right)\sum_{k=0}^{\infty}\frac{h_k h'_{k+5}+h'_k h_{k+5}}{k!}\\
&\quad=6\sum_{k=0}^{\infty}\frac{h_{k+1}h'_k}{k!}\gamma_1^k\gamma_2^2\zeta_{12}+6\sum_{k=0}^{\infty}\frac{h_k h'_{k+1}}{k!}\gamma_1^k\gamma_2^2\zeta_{22}\pm\mathcal{O}\left(\delta_r^3\right)
\end{aligned}$$

□

Lemma 22. *Let* $h_i(\mathbf{x}) = \pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)$ *and* $h_j(\mathbf{x}) = \pi\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)$. *Suppose that*

$$\max\left\{\left|\bar{\mathbf{v}}_i^\top \bar{\mathbf{w}}_i\right|, \left|\bar{\mathbf{v}}_j^\top \bar{\mathbf{w}}_j\right|, \left|\bar{\mathbf{v}}_i^\top \bar{\mathbf{w}}_j\right|, \left|\bar{\mathbf{v}}_j^\top \bar{\mathbf{w}}_i\right|\right\} \le \delta_r$$

then we have that

$$\mathbb{E}_{\mathbf{x}}\Big[h_i(\mathbf{x})^2\Big] = 6\sum_{k=0}^{\infty}\frac{c_k^2}{k!} \pm \mathcal{O}\left(\delta_r^2\right)$$

If it holds that $\left|\bar{\mathbf{w}}_i^\top \bar{\mathbf{w}}_j\right| \le \delta_r$, *then we have that*

$$\mathbb{E}_{\mathbf{x}}\big[h_i(\mathbf{x})\, h_j(\mathbf{x})\big] = \pm\mathcal{O}\left(\delta_r^3\right);$$

If it holds that $\left|\bar{\mathbf{w}}_i^\top \bar{\mathbf{w}}_j\right| \le \delta_r$, *then we have that*

$$\mathbb{E}_{\mathbf{x}}\big[h_i(\mathbf{x})\, h_j(\mathbf{x})\big] = 6\mathbb{E}_{\mathbf{x}}\big[h_i(\mathbf{x})^2\big] \pm \mathcal{O}\left(\delta_r\right)$$

Proof. Adopting the Hermite expansion, by Lemma 21 we have that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}\big[h_i(\mathbf{x})\, h_j(\mathbf{x})\big] &= \sum_{k,\ell=0}^{\infty}\frac{c_k c_\ell}{k!\ell!}\mathbb{E}_{\mathbf{x}}\big[He_k\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right) He_\ell\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right) He_3\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\big] \\
&= 6\left(\bar{\mathbf{w}}_i^\top \bar{\mathbf{w}}_j\right)^3\sum_{k=0}^{\infty}\frac{c_k^2}{k!}\left(\bar{\mathbf{v}}_i^\top \bar{\mathbf{v}}_j\right)^k \pm \mathcal{O}\left(\delta_r^2\left(\bar{\mathbf{w}}_i^\top \bar{\mathbf{w}}_j\right)^2 + \delta_r^4\right)
\end{aligned}$$

Thus, in the case where $i \neq j$ and $\left|\bar{\mathbf{w}}_i^\top \bar{\mathbf{w}}_j\right| \le \delta_r$, we have that

$$\mathbb{E}_{\mathbf{x}}\big[h_i(\mathbf{x})\, h_j(\mathbf{x})\big] = \pm\mathcal{O}\left(\delta_r^3\right)$$

On the other hand, if $i = j$, then we have that

$$\mathbb{E}_{\mathbf{x}}\Big[h_i(\mathbf{x})^2\Big] = 6\sum_{k=0}^{\infty}\frac{c_k^2}{k!} \pm \mathcal{O}\left(\delta_r^2\right)$$

If $\bar{\mathbf{w}}_i^\top \bar{\mathbf{w}}_j \ge 1 - \delta_r$ and $\bar{\mathbf{v}}_i^\top \bar{\mathbf{v}}_j \ge 1 - \delta$, then we have that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}\big[h_i(\mathbf{x})\, h_j(\mathbf{x})\big] &= 6\left(1 - 3\delta_r\right)\sum_{k=0}^{\infty}\frac{c_k^2}{k!}\left(\bar{\mathbf{v}}_i^\top \bar{\mathbf{v}}_j\right)^k \pm \mathcal{O}\left(\delta_r^2\right) \\
&= 6\left(1 - 3\delta_r\right)\mathbb{E}_{\mathbf{x}}\big[\pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)\pi\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right)\big] \pm \mathcal{O}\left(\delta_r^2\right)
\end{aligned}$$

Applying the Taylor's expansion gives that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}\big[\pi\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)\pi\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right)\big] &= \mathbb{E}_{z_1, z_2 \sim \mathcal{N}(0,1), \mathrm{Cov}(z_1, z_2) = 1 - \delta_r}\big[\pi\left(z_1\right)\pi\left(z_2\right)\big] \\
&= \mathbb{E}_{z\sim\mathcal{N}(0,1)}\Big[\pi\left(z_1\right)^2\Big] - \frac{1}{2}\mathrm{Var}\left(\pi'\left(z_1\right)\right)\delta_r \pm \mathcal{O}\left(\delta_r\right) \\
&\ge \mathbb{E}_{z\sim\mathcal{N}(0,1)}\Big[\pi\left(z_1\right)^2\Big] \pm \mathcal{O}\left(\delta_r\right)
\end{aligned}$$

Thus, we have that

$$\mathbb{E}_{\mathbf{x}}\big[h_i(\mathbf{x})\, h_j(\mathbf{x})\big] = 6\mathbb{E}_{\mathbf{x}}\big[h_i(\mathbf{x})^2\big] \pm \mathcal{O}\left(\delta_r\right)$$

□

Lemma 23. *Let* $\boldsymbol{\theta}$ *satisfy that* $\left\|\bar{\mathbf{v}}_i^\top - \bar{\mathbf{v}}_i^\star\right\|_2^2, \left\|\bar{\mathbf{w}}_i^\top - \bar{\mathbf{w}}_i^\star\right\|_2^2 \le \varepsilon$, $\left|\bar{\mathbf{v}}_i^\top \bar{\mathbf{v}}_j\right|, \left|\bar{\mathbf{w}}_i^\top \bar{\mathbf{w}}_j\right|, \left|\bar{\mathbf{v}}_i^\top \bar{\mathbf{v}}_j^\star\right|, \left|\bar{\mathbf{w}}_i^\top \bar{\mathbf{w}}_j^\star\right| \le \varepsilon$ *for all* $i \neq j$, *and* $\left|\bar{\mathbf{v}}_i^\top \bar{\mathbf{w}}_j\right|, \left|\bar{\mathbf{w}}_i^\top \bar{\mathbf{v}}_j\right|, \left|\bar{\mathbf{v}}_i^\top \bar{\mathbf{w}}_j^\star\right|, \left|\bar{\mathbf{w}}_i^\top \bar{\mathbf{v}}_j^\star\right| \le \varepsilon$ *for all* i, j. *Then the following holds:*

- $\left|\mathbb{E}_{\mathbf{x}}\big[\left(f\left(\boldsymbol{\theta}, \mathbf{x}\right) - f^\star\left(\mathbf{x}\right)\right)\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)\sigma^{(b)}\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\big]\right| \le \mathcal{O}\left(m^\star \varepsilon\right)$

- $\left\|\mathbb{E}_{\mathbf{x}}\left[\nabla_{\mathbf{x}}^2\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^\star\left(\mathbf{x}\right)\right)\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b)}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right]\right\|_2\leq\mathcal{O}\left(m^\star\varepsilon\right)$
- $\left\|\mathbb{E}_{\mathbf{x}}\left[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^\star\left(\mathbf{x}\right)\right)\nabla_{\mathbf{x}}^2\left(\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b)}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right)\right]\right\|_2\leq\mathcal{O}\left(m^\star\varepsilon\right)$
- $\left\|\mathbb{E}_{\mathbf{x}}\left[\nabla_{\mathbf{x}}\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^\star\left(\mathbf{x}\right)\right)\nabla_{\mathbf{x}}\left(\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b)}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right)\right]\right\|_2\leq\mathcal{O}\left(m^\star\varepsilon\right)$
- $\left\|\mathbb{E}_{\mathbf{x}}\left[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^\star\left(\mathbf{x}\right)\right)\nabla_{\mathbf{x}}\left(\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b)}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right)\right]\right\|_2\leq\mathcal{O}\left(m^\star\varepsilon\right)$
- $\left\|\mathbb{E}_{\mathbf{x}}\left[\nabla_{\mathbf{x}}\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^\star\left(\mathbf{x}\right)\right)\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b)}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right]\right\|_2\leq\mathcal{O}\left(m^\star\varepsilon\right)$

Proof. We first write out the gradient with respect to $\mathbf{x}$

$$\begin{aligned}
\nabla_{\mathbf{x}}f\left(\boldsymbol{\theta},\mathbf{x}\right)&=\sum_{j=1}^{m^\star}\pi'\left(\bar{\mathbf{v}}_j^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_j^\top\mathbf{x}\right)\bar{\mathbf{v}}_j+\sum_{j=1}^{m^\star}\pi\left(\bar{\mathbf{v}}_j^\top\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_j^\top\mathbf{x}\right)\bar{\mathbf{w}}_j\\
\nabla_{\mathbf{x}}f^\star\left(\mathbf{x}\right)&=\sum_{j=1}^{m^\star}\pi'\left(\bar{\mathbf{v}}_j^{\star\top}\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_j^{\star\top}\mathbf{x}\right)\bar{\mathbf{v}}_j^\star+\sum_{j=1}^{m^\star}\pi\left(\bar{\mathbf{v}}_j^{\star\top}\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_j^{\star\top}\mathbf{x}\right)\bar{\mathbf{w}}_j^\star\\
\nabla_{\mathbf{x}}^2f\left(\boldsymbol{\theta},\mathbf{x}\right)&=\sum_{j=1}^{m^\star}\pi'\left(\bar{\mathbf{v}}_j^\top\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_j^\top\mathbf{x}\right)\left(\bar{\mathbf{v}}_j\bar{\mathbf{w}}_j^\top+\bar{\mathbf{w}}_j\bar{\mathbf{v}}_j^\top\right)\\
&\quad+\sum_{j=1}^{m^\star}\pi''\left(\bar{\mathbf{v}}_j^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_j^\top\mathbf{x}\right)\bar{\mathbf{v}}_j\bar{\mathbf{v}}_j^\top+\sum_{j=1}^{m^\star}\pi\left(\bar{\mathbf{v}}_j^\top\mathbf{x}\right)\sigma''\left(\bar{\mathbf{w}}_j^\top\mathbf{x}\right)\bar{\mathbf{w}}_j\bar{\mathbf{w}}_j^\top\\
\nabla_{\mathbf{x}}^2f^\star\left(\mathbf{x}\right)&=\sum_{j=1}^{m^\star}\pi'\left(\bar{\mathbf{v}}_j^{\star\top}\mathbf{x}\right)\sigma'\left(\bar{\mathbf{w}}_j^{\star\top}\mathbf{x}\right)\left(\bar{\mathbf{v}}_j^\star\bar{\mathbf{w}}_j^{\star\top}+\bar{\mathbf{w}}_j^\star\bar{\mathbf{v}}_j^{\star\top}\right)\\
&\quad+\sum_{j=1}^{m^\star}\pi''\left(\bar{\mathbf{v}}_j^{\star\top}\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_j^{\star\top}\mathbf{x}\right)\bar{\mathbf{v}}_j^\star\bar{\mathbf{v}}_j^{\star\top}+\sum_{j=1}^{m^\star}\pi\left(\bar{\mathbf{v}}_j^{\star\top}\mathbf{x}\right)\sigma''\left(\bar{\mathbf{w}}_j^{\star\top}\mathbf{x}\right)\bar{\mathbf{w}}_j^\star\bar{\mathbf{w}}_j^{\star\top}
\end{aligned}$$

Moreover, we also have that

$$\begin{aligned}
\nabla_{\mathbf{x}}\left(\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b)}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right)&=\pi^{(a+1)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b)}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\bar{\mathbf{v}}_i+\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b+1)}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\bar{\mathbf{w}}_i\\
\nabla_{\mathbf{x}}^2\left(\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b)}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right)&=\pi^{(a+1)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b+1)}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\left(\bar{\mathbf{v}}_i\bar{\mathbf{w}}_i^\top+\bar{\mathbf{w}}_i\bar{\mathbf{v}}_i^\top\right)\\
&\quad+\pi^{(a+2)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b)}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\bar{\mathbf{v}}_i\bar{\mathbf{v}}_i^\top\\
&\quad+\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b+2)}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\bar{\mathbf{w}}_i\bar{\mathbf{w}}_i^\top
\end{aligned}$$

Therefore, for the last two bounds, we have

$$\begin{aligned}
\left\|\mathbb{E}_{\mathbf{x}}\left[\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^\star\left(\mathbf{x}\right)\right)\nabla_{\mathbf{x}}\left(\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b)}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right)\right]\right\|_2&=\mathcal{E}_{1,1}+\mathcal{E}_{2,1}\\
\left\|\mathbb{E}_{\mathbf{x}}\left[\nabla_{\mathbf{x}}\left(f\left(\boldsymbol{\theta},\mathbf{x}\right)-f^\star\left(\mathbf{x}\right)\right)\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b)}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right]\right\|_2&=\mathcal{E}_{1,2}+\mathcal{E}_{2,2}
\end{aligned}$$

where $\mathcal{E}_{1,1}$ and $\mathcal{E}_{1,2}$ are a summation of two terms in the form

$$\left|\mathbb{E}_{\mathbf{x}}\left[\left(\pi\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)-\pi\left(\bar{\mathbf{v}}_i^{\star\top}\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^{\star\top}\mathbf{x}\right)\right)\pi^{(a')}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b')}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right]\right|$$

Thus, by Taylor expansion, we obtain that $\mathcal{E}_{1,1},\mathcal{E}_{1,2}\leq\mathcal{O}\left(\varepsilon\right)$. Moreover, $\mathcal{E}_{1,2}$ and $\mathcal{E}_{2,2}$ are a summation of $4m^\star$ terms of the form

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbf{x}}\left[\pi\left(\bar{\mathbf{v}}_j^\top\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_j^\top\mathbf{x}\right)\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b')}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right]\right|\\
&\left|\mathbb{E}_{\mathbf{x}}\left[\pi\left(\bar{\mathbf{v}}_j^{\star\top}\mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_j^{\star\top}\mathbf{x}\right)\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top\mathbf{x}\right)\sigma^{(b')}\left(\bar{\mathbf{w}}_i^\top\mathbf{x}\right)\right]\right|
\end{aligned}$$

By Lemma 1, we have that $\mathcal{E}_{2,1}, \mathcal{E}_{2,2} \leq \mathcal{O}(m^\star \varepsilon)$. This gives the last two property. For the rest of the property, we can apply similar strategy to decompose the objective in terms of

$$\left|\mathbb{E}_{\mathbf{x}}\Big[\left(\pi^{(a_0)}\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)\sigma^{(b_0)}\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right) - \pi\left(\bar{\mathbf{v}}_i^{\star\top} \mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_i^{\star\top} \mathbf{x}\right)\right)\pi^{(a')}\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)\sigma^{(b')}\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\Big]\right|$$

which can be upper bounded by Taylor expansion, and

$$\left|\mathbb{E}_{\mathbf{x}}\Big[\pi\left(\bar{\mathbf{v}}_j^\top \mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_j^\top \mathbf{x}\right)\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)\sigma^{(b')}\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\Big]\right|$$
$$\left|\mathbb{E}_{\mathbf{x}}\Big[\pi\left(\bar{\mathbf{v}}_j^{\star\top} \mathbf{x}\right)\sigma\left(\bar{\mathbf{w}}_j^{\star\top} \mathbf{x}\right)\pi^{(a)}\left(\bar{\mathbf{v}}_i^\top \mathbf{x}\right)\sigma^{(b')}\left(\bar{\mathbf{w}}_i^\top \mathbf{x}\right)\Big]\right|$$

which can be upper bounded by lemma 25. Since there are in total $\mathcal{O}(m^\star)$ terms for each quantity, we can conclude the desired result. □

Lemma 24. *Let $\mathbf{v}_1, \mathbf{v}_2, \mathbf{w}_1, \mathbf{w}_2$ be unit vectors satisfying that any two of the four have an inner product with magnitude less than ε. Then for $a_1, a_2, b_1, b_2 > 0$ with $b_1 + b_2 \leq 3$, the following holds*

- $\left|\mathbb{E}_{\mathbf{x}}\left[\pi^{(a_1)}\left(\mathbf{v}_1^\top \mathbf{x}\right)\sigma^{(b_1)}\left(\mathbf{w}_1^\top \mathbf{x}\right)\pi^{(a_2)}\left(\mathbf{v}_2^\top \mathbf{x}\right)\sigma^{(b_2)}\left(\mathbf{w}_2^\top \mathbf{x}\right)\right]\right| \leq \mathcal{O}(\varepsilon)$
- $\left\|\mathbb{E}_{\mathbf{x}}\left[\nabla_{\mathbf{x}}^2\left(\pi^{(a_1)}\left(\mathbf{v}_1^\top \mathbf{x}\right)\sigma^{(b_1)}\left(\mathbf{w}_1^\top \mathbf{x}\right)\right)\pi^{(a_2)}\left(\mathbf{v}_2^\top \mathbf{x}\right)\sigma^{(b_2)}\left(\mathbf{w}_2^\top \mathbf{x}\right)\right]\right\|_2 \leq \mathcal{O}(\varepsilon)$
- $\left\|\mathbb{E}_{\mathbf{x}}\left[\pi^{(a_1)}\left(\mathbf{v}_1^\top \mathbf{x}\right)\sigma^{(b_1)}\left(\mathbf{w}_1^\top \mathbf{x}\right)\nabla_{\mathbf{x}}^2\left(\pi^{(a_2)}\left(\mathbf{v}_2^\top \mathbf{x}\right)\sigma^{(b_2)}\left(\mathbf{w}_2^\top \mathbf{x}\right)\right)\right]\right\|_2 \leq \mathcal{O}(\varepsilon)$
- $\left\|\mathbb{E}_{\mathbf{x}}\left[\nabla_{\mathbf{x}}\left(\pi^{(a_1)}\left(\mathbf{v}_1^\top \mathbf{x}\right)\sigma^{(b_1)}\left(\mathbf{w}_1^\top \mathbf{x}\right)\right)\nabla_{\mathbf{x}}\left(\pi^{(a_2)}\left(\mathbf{v}_2^\top \mathbf{x}\right)\sigma^{(b_2)}\left(\mathbf{w}_2^\top \mathbf{x}\right)\right)\right]\right\|_2 \leq \mathcal{O}(\varepsilon)$

Proof. The first quantity is directly bounded by applying Lemma 25. For the rest, we write out the form of the gradients with respect to $\mathbf{x}$ as

$$\begin{aligned}
\nabla_{\mathbf{x}}\left(\pi^{(a)}\left(\mathbf{v}^\top \mathbf{x}\right)\sigma^{(b)}\left(\mathbf{w}^\top\right)\right) &= \pi^{(a+1)}\left(\mathbf{v}^\top \mathbf{x}\right)\sigma^{(b)}\left(\mathbf{w}^\top\right)\mathbf{v} + \pi^{(a)}\left(\mathbf{v}^\top \mathbf{x}\right)\sigma^{(b+1)}\left(\mathbf{w}^\top\right)\mathbf{w} \\
\nabla_{\mathbf{x}}^2\left(\pi^{(a)}\left(\mathbf{v}^\top \mathbf{x}\right)\sigma^{(b)}\left(\mathbf{w}^\top\right)\right) &= \pi^{(a+2)}\left(\mathbf{v}^\top \mathbf{x}\right)\sigma^{(b)}\left(\mathbf{w}^\top\right)\mathbf{v}\mathbf{v}^\top + \pi^{(a)}\left(\mathbf{v}^\top \mathbf{x}\right)\sigma^{(b+2)}\left(\mathbf{w}^\top\right)\mathbf{w}\mathbf{w}^\top \\
&\quad + \pi^{(a+1)}\left(\mathbf{v}^\top \mathbf{x}\right)\sigma^{(b+1)}\left(\mathbf{w}^\top\right)\left(\mathbf{v}\mathbf{w}^\top + \mathbf{w}\mathbf{v}^\top\right)
\end{aligned}$$

Therefore, for each of the rest property, it can be written in terms of a summation of terms of the form

$$\mathbb{E}_{\mathbf{x}}\Big[\pi^{(a_1')}\left(\mathbf{v}_1^\top \mathbf{x}\right)\sigma^{(b_1')}\left(\mathbf{w}_1^\top \mathbf{x}\right)\pi^{(a_2')}\left(\mathbf{v}_2^\top \mathbf{x}\right)\sigma^{(b_2')}\left(\mathbf{w}_2^\top \mathbf{x}\right)\Big]$$

Since $b_1 + b_2 \leq 3$, taking twice derivative gives $b_1' + b_2' \leq 5$. Therefore, applying Lemma 25 gives that all the rest terms are upper bounded by $\mathcal{O}(\varepsilon)$. □

Lemma 25. *Let $\mathbf{v}_1, \mathbf{v}_2, \mathbf{w}_1, \mathbf{w}_2$ be unit vectors such that any two of the four have an inner product with magnitude upper bounded by ε. Then we have that for any a_1, a_2, b_1, b_2 such that $b_1 + b_2 \leq 5$*

$$\left|\mathbb{E}_{\mathbf{x}}\Big[\pi^{(a_1)}\left(\mathbf{v}_1^\top \mathbf{x}\right)\pi^{(a_2)}\left(\mathbf{v}_2^\top \mathbf{x}\right)\sigma^{(b_1)}\left(\mathbf{w}_1^\top \mathbf{x}\right)\sigma^{(b_2)}\left(\mathbf{w}_2^\top \mathbf{x}\right)\Big]\right| \leq \mathcal{O}(\varepsilon)$$

Proof. Taking the Hermite expansion

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}}\Big[\pi^{(a_1)}\left(\mathbf{v}_1^\top \mathbf{x}\right)\pi^{(a_2)}\left(\mathbf{v}_2^\top \mathbf{x}\right)\sigma^{(b_1)}\left(\mathbf{w}_1^\top \mathbf{x}\right)\sigma^{(b_2)}\left(\mathbf{w}_2^\top \mathbf{x}\right)\Big] \\
&\quad = \sum_{k,\ell=0}^{\infty} \frac{c_{k+a_1} c_{\ell+a_2}}{k!\ell!}\mathbb{E}_{\mathbf{x}}\left[He_k\left(\mathbf{v}_1^\top \mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top \mathbf{x}\right)He_{3-b_1}\left(\mathbf{w}_1^\top \mathbf{x}\right)He_{3-b_2}\left(\mathbf{w}_2^\top \mathbf{x}\right)\right]
\end{aligned}$$

We could observe that at least one of $3 - b_1$ and $3 - b_2$ is nonzero. Therefore, by Lemma 1, we have that $\mathbb{E}_{\mathbf{x}}\left[He_k\left(\mathbf{v}_1^\top \mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top \mathbf{x}\right)He_{3-b_1}\left(\mathbf{w}_1^\top \mathbf{x}\right)He_{3-b_2}\left(\mathbf{w}_2^\top \mathbf{x}\right)\right]$ is a polynomial with lowest degree at most 1. Therefore, we have that

$$\left|\mathbb{E}_{\mathbf{x}}\left[He_k\left(\mathbf{v}_1^\top \mathbf{x}\right)He_\ell\left(\mathbf{v}_2^\top \mathbf{x}\right)He_{3-b_1}\left(\mathbf{w}_1^\top \mathbf{x}\right)He_{3-b_2}\left(\mathbf{w}_2^\top \mathbf{x}\right)\right]\right| \leq \mathcal{O}(\varepsilon)$$

Moreover, this quantity is nonzero only when $k - \ell \leq 6 - b_1 - b_2$. Thus, by the boundedness of the Hermite coefficients, we can conclude that

$$\left|\mathbb{E}_{\mathbf{x}}\left[\pi^{(a_1)}\left(\mathbf{v}_1^\top \mathbf{x}\right)\pi^{(a_2)}\left(\mathbf{v}_2^\top \mathbf{x}\right)\sigma^{(b_1)}\left(\mathbf{w}_1^\top \mathbf{x}\right)\sigma^{(b_2)}\left(\mathbf{w}_2^\top \mathbf{x}\right)\right]\right| \leq \mathcal{O}\left(\varepsilon\right)$$

□

D.2 Other Auxiliary Results

Lemma 26. *Let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$, and define $f\left(\mathbf{v}\right) = \frac{1}{\|\mathbf{v}\|_2}\left(\mathbf{I} - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2}\right)\mathbf{w}$. Then we have that*

$$\mathcal{J}f\left(\mathbf{v}\right) = -\frac{\mathbf{v}^\top \mathbf{w}}{\|\mathbf{v}\|_2^3}\left(\mathbf{I} - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2}\right) - \frac{1}{\|\mathbf{v}\|_2^3}\left(\mathbf{v}\mathbf{w}^\top + \mathbf{w}\mathbf{v}^\top\right)$$

Lemma 27. *Let $f(x) = \frac{1}{1+e^{-x}}$ be the sigmoid function. Then we have that*

- *$f'''(x)^2 \leq 1$ for all $x \in \mathbb{R}$*
- *$\mathbb{E}_{x\sim\mathcal{N}(0,1)}\big[f(x)\big] = \frac{1}{2}$*
- *$\mathbb{E}_{x\sim\mathcal{N}(0,1)}\big[f^{(2k)}(x)\big] = 0$ for all $k \geq 1$*

Proof. Using simple calculations, we can obtain that

$$f'(x) = f(x)(1 - f(x))$$

This gives that

$$f''(x) = f'(x)(1 - f(x)) - f(x)f'(x) = f'(x)(1 - 2f(x))$$

Thus, $f'''(x)$ can be written as

$$\begin{aligned} f'''(x) &= f''(x)(1 - 2f(x)) - 2f'(x)^2 \\ &= f'(x)(1 - 2f(x))^2 - 2f'(x)^2 \\ &= f'(x)\left(1 - 4f(x) + 4f(x)^2 - 2f'(x)\right) \\ &= f'(x)\left(1 - 6f(x) + 6f(x)^2\right) \end{aligned}$$

One the range $[0, 1]$, the function $1 - 6y + 6y^2$ takes extremes at $y = 0, \frac{1}{2}, 1$. At $y = 0$ and $y = 1$, we have $1 - 6y + 6y^2 = 1$. At $y = \frac{1}{2}$, we have that $1 - 6y + 6y^2 = -\frac{1}{2}$. Thus, we can conclude that $\left|1 - 6y + 6y^2\right| \leq 1$ for all $y \in [0, 1]$. Moreover, we have that $f'(x) = f(x)(1 - f(x)) \in [0, 1]$, since $f(x) \in [0, 1]$. Therefore, we can conclude that $|f'''(x)| \leq 1$, which implies the first property. To prove the second, we notice that

$$f(-x) = \frac{1}{1 + e^x} = \frac{e^{-x}}{1 + e^{-x}} = 1 - f(x)$$

Therefore, due to the symmetry of Gaussian distribution, we have that

$$\mathbb{E}_{x\sim\mathcal{N}(0,1)}\big[f(x)\big] = \mathbb{E}_{x\sim\mathcal{N}(0,1)}\big[f(-x)\big] = 1 - \mathbb{E}_{x\sim\mathcal{N}(0,1)}\big[f(x)\big]$$

This gives that $\mathbb{E}_{x\sim\mathcal{N}(0,1)}\big[f(x)\big] = \frac{1}{2}$. To prove the third property, we notice that

$$f'(-x) = f(-x)(1 - f(-x)) = (1 - f(x))f(x) = f'(x)$$

which shows that $f'(-x)$ is even. Therefore, $f^{(2k)}(x)$ are odd functions for all $k \geq 1$. This implies the third property. □

Lemma 28. *Consider function $f(x), g(x), h(x)$ given by the ODE system*

$$\begin{aligned} f'(x) &= -a_1 f(x) + b_1 g(x) + c_1 h(x) + p \\ g'(x) &= -a_2 f(x) - b_2 g(x) + c_2 h(x) + p \\ h'(x) &= -a_3 f(x) + b_3 g(x) - c_3 h(x) + p \end{aligned}$$

for some $a_1, a_2, a_3, b_1, b_2, b_3, c_1, c_2, c_3 > 0$. If $b_2 c_3 \geq b_3 c_2$ and

$$a_1^2 b_2 + a_1^2 c_3 + a_1 a_2 b_1 + a_1 a_3 c_1 + a_1 b_2^2 + a_2 b_1 b_2 + a_1 c_3^2 + a_3 c_1 c_3 \geq a_2 b_3 c_1 + a_3 b_1 c_2$$

then we have that

$$\max\left\{|f(x)|, |g(x)|, |h(x)|\right\} \leq e^{-\Omega(x)} \left(|f(0)| + |g(0)| + |h(0)|\right) + \mathcal{O}\left(p\right)$$

for all $x \geq 0$

Proof. Let $\mathbf{A} \in \mathbb{R}^{3\times 3}$ be given by

$$\mathbf{A} = \begin{bmatrix} -a_1 & b_1 & c_1 \\ -a_2 & -b_2 & c_2 \\ -a_3 & b_3 & -c_3 \end{bmatrix}$$

Then we have that

$$\begin{bmatrix} f'(x) \\ g'(x) \\ h'(x) \end{bmatrix} = \mathbf{A} \begin{bmatrix} f(x) \\ g(x) \\ h(x) \end{bmatrix} + p\mathbf{1}$$

Solving the system gives

$$\begin{bmatrix} f(x) \\ g(x) \\ h(x) \end{bmatrix} = e^{\mathbf{A}x} \begin{bmatrix} f(0) \\ g(0) \\ h(0) \end{bmatrix} + p\mathbf{A}^{-1}\left(e^{\mathbf{A}x} - \mathbf{I}\right)\mathbf{1}$$

Let λ be the eigenvalue of A with the largest real part. Then we have that

$$\left\| \begin{bmatrix} f(x) \\ g(x) \\ h(x) \end{bmatrix} \right\|_2 \leq e^{\lambda x} \left\| \begin{bmatrix} f(0) \\ g(0) \\ h(0) \end{bmatrix} \right\| + \mathcal{O}\left(p\right)$$

Thus, it suffice to show that all eigenvalues of $\mathbf{A}$ has negative real parts. To do this, we write out the characteristic polynomial of $\mathbf{A}$ as

$$P(y) = y^3 - \mathrm{Tr}\left(\mathbf{A}\right) y^2 + \frac{1}{2}\left(\mathrm{Tr}\left(\mathbf{A}\right)^2 - \mathrm{Tr}\left(\mathbf{A}^2\right)\right) y - \det\left(\mathbf{A}\right)$$

By the Routh-Hurwitz criteria, it suffice to show that

$$\mathrm{Tr}\left(\mathbf{A}\right) < 0, \det\left(\mathbf{A}\right) < 0, \frac{1}{2}\mathrm{Tr}\left(\mathbf{A}\right)\left(\mathrm{Tr}\left(\mathbf{A}\right)^2 - \mathrm{Tr}\left(\mathbf{A}^2\right)\right) < \det\left(\mathbf{A}\right)$$

With the form of $\mathbf{A}$, we obtain that

$$\begin{aligned} \mathrm{Tr}\left(\mathbf{A}\right) &= -\left(a_1 + b_2 + c_3\right) \\ \frac{1}{2}\mathrm{Tr}\left(\mathbf{A}\right)\left(\mathrm{Tr}\left(\mathbf{A}\right)^2 - \mathrm{Tr}\left(\mathbf{A}^2\right)\right) &= a_1 b_2 + a_1 c_3 + b_2 c_3 + a_2 b_1 + a_3 c_1 - b_3 c_2 \\ \det\left(\mathbf{A}\right) &= -a_1 b_2 c_3 - a_2 b_3 c_1 - a_3 b_1 c_2 - a_3 b_2 c_1 - a_2 b_1 c_3 - a_1 b_3 c_2 \end{aligned}$$

Thus, it is easy to see that $\mathrm{Tr}\left(\mathbf{A}\right) < 0, \det\left(\mathbf{A}\right) < 0$. It remains to show that $\frac{1}{2}\mathrm{Tr}\left(\mathbf{A}\right)\left(\mathrm{Tr}\left(\mathbf{A}\right)^2 - \mathrm{Tr}\left(\mathbf{A}^2\right)\right) <$

$\det(\mathbf{A})$. This is equivalent to show that $S \geq 0$ with

$$\begin{aligned}
S &= \det(\mathbf{A}) - \frac{1}{2}\mathrm{Tr}(\mathbf{A})\left(\mathrm{Tr}(\mathbf{A})^2 - \mathrm{Tr}(\mathbf{A}^2)\right) \\
&= (a_1+b_2+c_3)(a_1b_2+a_1c_3+b_2c_3+a_2b_1+a_3c_1-b_3c_2) \\
&\quad -(a_1b_2c_3+a_2b_3c_1+a_3b_1c_2+a_3b_2c_1+a_2b_1c_3+a_1b_3c_2) \\
&= a_1^2b_2+a_1^2c_3+a_1b_2c_3+a_1a_2b_1+a_1a_3c_1-a_1b_3c_2+a_1b_2^2+a_1b_2c_3+b_2^2c_3+a_2b_1b_2 \\
&\quad a_3b_2c_1-b_2b_3c_2+a_1b_2c_3+a_1c_3^2+b_2c_3^2+a_2b_1c_3+a_3c_1c_3-b_3c_2c_3 \\
&\quad -a_1b_2c_3-a_2b_3c_1-a_3b_1c_2-a_3b_2c_1-a_2b_1c_3-a_1b_3c_2 \\
&= a_1^2b_2+a_1^2c_3+2a_1b_2c_3+a_1a_2b_1+a_1a_3c_1-2a_1b_3c_2+a_1b_2^2+b_2^2c_3+a_2b_1b_2 \\
&\quad -b_2b_3c_2+a_1c_3^2+b_2c_3^2+a_3c_1c_3-b_3c_2c_3-a_2b_3c_1-a_3b_1c_2
\end{aligned}$$

If $b_2c_3 \geq b_3c_2$, then we have that

$$a_1b_2c_3 \geq a_1b_3c_2; \quad b_2c_3^2 \geq b_3c_2c_3; \quad b_2^2c_3 \geq b_2b_3c_2$$

This gives that

$$S \geq a_1^2b_2+a_1^2c_3+a_1a_2b_1+a_1a_3c_1+a_1b_2^2+a_2b_1b_2+a_1c_3^2+a_3c_1c_3-a_2b_3c_1-a_3b_1c_2$$

□

Lemma 29. *Let c_k be the kth order Hermite coefficient of $\pi(x)$. Let $z_1, z_2 \sim \mathcal{N}(0,1)$ with $Cov(z_1,z_2)=\rho$. If $\mathbb{E}_{z_1,z_2}[\pi'(z_1)\pi'''(z_2)] \leq 0$ for all ρ, then we have that*

$$\sum_{k=0}^{\infty} \frac{c_kc_{k+2}}{k!}\gamma^k \leq 0; \forall \gamma > 0$$

Proof. Notice that, by taking the Hermite expansion of $\pi(x)$ and $\pi''(x)$, we have that for $z_1, z_2 \sim \mathcal{N}(0,1)$ with $\mathrm{Cov}(z_1,z_2)=\gamma$

$$\mathbb{E}_{z_1,z_2}[\pi(z_1)\pi''(z_2)] = \sum_{k=0}^{\infty}\frac{c_kc_{k+2}}{k!}\gamma^k$$

Let $f(\gamma) = \mathbb{E}_{z_1,z_2}[\pi(z_1)\pi''(z_2)]$. Then by Price's Theorem we have that

$$f'(\gamma) = \mathbb{E}_{z_1,z_2}[\pi'(z_1)\pi'''(z_2)] \leq 0$$

Moreover, at $\gamma = 0$, we have that

$$f(0) = \mathbb{E}_{z_1}[\pi(z_1)]\mathbb{E}_{z_2}[\pi''(z_2)] = 0$$

where the last equality is due to Lemma 27. Therefore, we can conclude that

$$\sum_{k=0}^{\infty}\frac{c_kc_{k+2}}{k!}\gamma^k = f(\gamma) \leq f(0) = 0$$

□

Lemma 30. *Let $x \in (-1,1)$, and consider $S = \sum_{k=0}^{\infty}\frac{c_k^2}{k!}x^k$. If $c_k \leq \sqrt{B\cdot k!}$, then we have that*

$$c_0^2 - \frac{|x|}{1-|x|} \leq S \leq c_0^2 + \frac{|x|}{1-|x|}$$

Proof. We write S as

$$S = c_0^2 + \sum_{k=1}^{\infty}\frac{c_k^2}{k!}x^k$$

Notice that

$$\left|\sum_{k=1}^{\infty}\frac{c_k^2}{k!}x^k\right| \leq \sum_{k=1}^{\infty}\frac{c_k^2}{k!}|x|^k \leq B\sum_{k=1}^{\infty}|x|^k = \frac{B|x|}{1-|x|}$$

Therefore, we can conclude that

$$c_0^2 - \frac{B|x|}{1-|x|} \leq S \leq c_0^2 + \frac{B|x|}{1-|x|}$$

□

Lemma 31. *Let c_k denote the kth order Hermite coefficient of $\pi(\cdot)$ such that $c_2 = 0$ and $c_{k+3} \leq \sqrt{B \cdot k!}$ for all $k \geq 0$. Let $\gamma \in (b, 1]$. If $\mathbb{E}_{z_1,z_2\sim\mathcal{N}(0,1),Cov(z_1,z_2)=\rho}\left[\pi'(z_1)\pi^{(3)}(z_2)\right] \leq 0$ for all $\rho \in [-1,1]$, then we have that*

$$\sum_{k=0}^{\infty}\frac{c_k c_{k+2}}{k!}\gamma^k \leq \frac{|b|}{1-|b|}$$

Proof. Notice that

$$\mathbb{E}_{z_1,z_2}[\pi(z_1)\pi''(z_2)] = \mathbb{E}_{z_1,z_2}\left[\left(\sum_{k=0}^{\infty}\frac{c_k}{k!}He_k(z_1)\right)\left(\sum_{k=0}^{\infty}\frac{c_{k+2}}{k!}He_k(z_1)\right)\right] = \sum_{k=0}^{\infty}\frac{c_k c_{k+2}}{k!}$$

Therefore, we can define

$$h(\rho) = \sum_{k=0}^{\infty}\frac{c_k c_{k+2}}{k!}\rho^k = \mathbb{E}_{z_1,z_2}[\pi(z_1)\pi''(z_2)]\rho^k$$

By Price's Theorem, we have that

$$h'(\rho) = \mathbb{E}_{z_1,z_2}\left[\pi'(z_1)\pi^{(3)}(z_2)\right] \leq 0$$

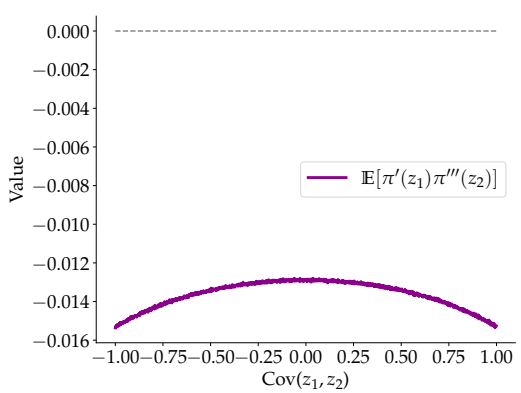
Therefore, $h(\rho)$ decreases monotonically, which implies that

$$h(\rho) = \sum_{k=0}^{\infty}\frac{c_k c_{k+2}}{k!}\rho^k \leq h(b) = b\sum_{k=0}^{\infty}\frac{c_{k+1}c_{k+3}}{k!}b^k \leq \sum_{k=1}^{\infty}|b|^k \leq \frac{|b|}{1-|b|}$$

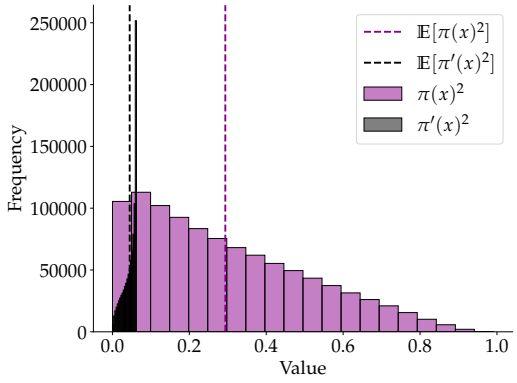
□

E Plotting Sigmoid Property

In this section, we plot the simulation result of the properties of the sigmoid function. Figure 3a is generated by taking 10^5 samples of correlated Gaussian random variables for each covariance value $\rho \in [-1, 1]$. Figure 3b is generated by taking 10^6 samples of standard Gaussian random variables.



(a) Value of $\mathbb{E}_{z_1,z_2\sim\mathcal{N}(0,1)}\left[\pi'(z_1)\pi'''(z_2)\right]$ for different values of Cov (z_1, z_2). As in the figure, all values are negative.



(b) Values of $\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\pi(x)^2\right]$ and $\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\pi'(x)^2\right]$. Figures shows that $\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\pi(x)^2\right] \geq 1.1\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\pi'(x)^2\right]$.

Figure 3: Plot of the property of $\pi(x)$