

SINGULAR PERTURBATIONS AND HIERARCHICAL LEARNING IN TWO-LAYER NEURAL NETWORKS

CÉDRIC GERBELOT

Unité de Mathématiques Pures et Appliquées (UMPA), ENS Lyon

JEAN-CHRISTOPHE MOURRAT

Unité de Mathématiques Pures et Appliquées (UMPA), ENS Lyon and CNRS

ABSTRACT. We study the population gradient flow of an infinitely wide two-layer neural network learning a misspecified single-index model in high dimension. The two layers are optimized jointly, with a perturbative parameter tuning the relative training speed between the first and second layer. This setting was considered by Berthier, Montanari and Zhou in [13], who conjectured a hierarchical learning scenario with explicit timescales as the second layer is trained faster than the first. In this paper, we prove that the constant and linear components of the hidden link function are indeed recovered within the predicted timescales, at sharp explicit thresholds. We then analyze the onset of learning of the quadratic component and show that the components learned at earlier stages continue to influence the dynamics in an essential way. Our proof is based on quantitative approximation results for singularly perturbed flows evolving near a manifold defined by integral constraints. At a phenomenological level, we also show that the empirical measure of the weights displays singular behaviour when reaching the quadratic component of the hidden link, with a small fraction of neurons growing significantly while the remaining ones rearrange to preserve the components already learned.

1. INTRODUCTION

In the effort to understand the success of gradient-based optimization methods for deep learning [11], it is crucial to find models that are amenable to theoretical analysis while retaining practical relevance. Among such models, two-layer neural networks remain a significant technical challenge despite a flurry of recent contributions: in the absence of convexity, it is difficult to find analytically tractable approximations of the dynamical system induced by a gradient flow on the parameters. In that regard, several approaches have appeared to go beyond the kernel regime [37, 26] which, despite the appealing simplicity of the ensuing (linear) dynamics, falls short of explaining the ability of neural networks to perform feature learning [21, 55, 28]. Notably, kernel-based analyses of neural network training dynamics do not show how these architectures adapt to latent low-dimensional structure in the data, thereby avoiding the curse of dimensionality [6, 57, 29]. We summarize in what follows, without claim of exhaustivity, two of these approaches that motivate the present work.

On the one hand, the series of papers [44, 19, 48] proposes to study infinitely wide two-layer neural networks and shows that, upon an appropriate scaling of the network and parameter initialization, the evolution of the empirical measure of the weights may be described by a non-linear transport equation corresponding to a Wasserstein gradient flow on a convex functional. Although this approach allows one to provide significant generalization guarantees in certain settings [20], the arguments often rely on infinite training time (see also [54]), and quantitative convergence results still largely rely on Polyak-Lojasiewicz assumptions, see e.g. [18].

On the other hand, a vast body of work studies the high-dimensional dynamics of various instances of gradient methods on simple problems, with random design data models displaying latent low-dimensional structure. Examples include single [56, 9, 15, 13] and multi-index regression [33, 23, 5, 2, 10, 24, 14],

E-mail addresses: cedric.gerbelot-barrillon@ens-lyon.fr, jean-christophe.mourrat@ens-lyon.fr.

Date: July 14, 2026.

JCM acknowledges the support of the ERC MSCA grant SLOHD (101203974).

as well as binary and multiclass classification [47, 7, 30]. These approaches rely on reducing the high-dimensional dynamics to a low-dimensional (Euclidean) dynamical system capturing both the effect of empirical noise and population terms. Postponing a more detailed discussion to Section 4, we mention for now that this has led to quantitative descriptions of a number of phenomena observed in practice such as hierarchical learning [4], where the neural network learns functions of increasing complexity on different timescales, with associated plateaus in the training and test error.

In the case of single index regression, if the link function is known, a single neuron is sufficient to learn the target function (i.e. the hidden index), and the gradient dynamics can be reduced to a one dimensional ordinary differential equation (ODE). The time for a parameter vector under gradient flow to achieve non-trivial correlation with the hidden index from an uninformative start is then governed by the local curvature of the loss landscape near the initial condition [27, 8, 9]. This, in turn, allows a sharp analysis of gradient methods with no oracle modification to the algorithm. When the activation function is not known however, a (fully parametrized) two-layer neural network is required to solve the corresponding semiparametric problem, and the ensuing effective dynamics is much harder to analyze. This has motivated the use of oracle-based proof methods such as layer-wise training [25, 15, 40], or meta-algorithms operating directly in an appropriate Hilbert space [46]. Thus, understanding the joint learning dynamics of a two-layer neural network trained with gradient descent to solve a semiparametric regression task is still unresolved.

Recently, the work [13] took a significant step towards answering this problem. Blending elements from both the approaches described above, the authors study gradient flow on an infinitely wide two-layer neural network, in the high-dimensional limit, and in a regime where the second layer is trained faster than the first. The relative speed is controlled by a perturbative parameter that allows one to interpolate between the unchanged gradient flow and layerwise training. Focusing on the population limit, they derive solutions to the effective dynamics using a matched asymptotic expansion argument, a non-rigorous method from the theory of singularly perturbed dynamical systems [35]. These computations lead to a precise description of the hierarchical learning process for the constant and linear components of the unknown activation, while progressively recovering the hidden index, with explicit expressions for the corresponding timescales. An additional conjecture is then established for the higher order components based on neglecting the terms associated to the ones learned previously and solving simplified dynamics that do not require a perturbative analysis.

In this paper, we build a framework to rigorously study a family of singularly perturbed dynamical systems that includes the learning dynamics of two-layer neural networks, providing mathematical grounding for the results of [13] regarding the constant and linear components of the hidden activation along the way. We then provide an explicit representation for the learning dynamics of the non-linear components, showing that the terms associated to the previously learned ones cannot be neglected. At a phenomenological level, beyond the constant and linear components, the empirical measure of the weights displays singular behaviour as it moves towards learning the next components while retaining the information acquired at earlier stages of the learning process. At a technical level, we provide a general methodology to construct approximations of dynamical systems evolving near target sets defined by integral constraints, and to obtain time dependent bounds that adapt to the hierarchical nature of the dynamics for the validity of these approximations.

2. SETTING

Our setting is the same as that of [13], and we adopt the same notations. For the sake of clarity, we will not reproduce their main results explicitly or enumerate the assumptions required for them to hold. Rather, we will give an informal description of only the elements required for our results to be stated. We refer the reader to their paper for more precise statements, and to the second paragraph of Section 4 for further discussion on this matter.

Consider the joint law $P_{x,y}$ on $\mathbb{R}^d \times \mathbb{R}$ defined by the conditional model

$$y = \varphi(\langle u_*, x \rangle), \quad (2.1)$$

where $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ is an unknown link function, x is a random vector in $\mathbb{R}^d$ with i.i.d. standard normal entries and $u_* \in \mathbb{S}^{d-1}$ is a fixed, unknown vector on the unit sphere. We learn this target function using

a two-layer neural network parametrized as follows

$$f(x, a, u) = \frac{1}{m} \sum_{i=1}^m a_i \sigma(\langle u_i, x \rangle), \quad a_1, \dots, a_m \in \mathbb{R}, \quad u_1, \dots, u_m \in \mathbb{S}^{d-1}, \quad (2.2)$$

and where $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is an activation function. The parameters of this network are trained with gradient flow on the population loss

$$\mathcal{R}(a, u) = \frac{1}{2} \mathbb{E} \left\{ (y - f(x, a, u))^2 \right\}, \quad (2.3)$$

where we denote $(a, u) = (a_1, \dots, a_m, u_1, \dots, u_m)$ the vectors gathering the external and internal weights. The gradient flow is considered on the product manifold $(\mathbb{R} \times \mathbb{S}^{d-1})^m$, so that a spherical gradient is considered for the internal weights, and an Euclidean gradient for the external ones, leading to the system of ordinary differential equations (ODEs)

$$\partial_t a_i = -m \partial_{a_i} \mathcal{R}(a, u), \quad (2.4)$$

$$\partial_t u_i = -m (I_d - u_i u_i^\top) \nabla_{u_i} \mathcal{R}(a, u). \quad (2.5)$$

Additionally, the relative training speed between the first and second layer weights is tuned with a perturbative parameter $\varepsilon > 0$, understood as small, as follows

$$\partial_t a_i = -\frac{m}{\varepsilon} \partial_{a_i} \mathcal{R}(a, u), \quad (2.6)$$

$$\partial_t u_i = -m (I_d - u_i u_i^\top) \nabla_{u_i} \mathcal{R}(a, u). \quad (2.7)$$

Thus, as ε becomes small, the second layer weights are tuned progressively faster than the first layer ones. At initialization, the vectors a, u are assumed to be drawn i.i.d. and independently respectively from a measure P_A , assumed to be absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}$, and the invariant measure on the d -dimensional unit sphere, denoted $\mathcal{U}(\mathcal{S}^{d-1})$. The activations σ, ϕ are assumed to be square-integrable with respect to the standard Gaussian measure. The coefficients of their decompositions in Hermite polynomials are respectively denoted $\{\sigma_k\}_{k=0}^\infty$ and $\{\varphi_k\}_{k=0}^\infty$. Letting, for each $1 \leq i \leq m$, $s_i = \langle w_i, u_* \rangle$, the authors of [13] show that, when taking m, d to infinity, the dynamical system (2.6)-(2.7) can be reduced to a set of autonomous ODEs on the variables $\{(a_i, s_i)\}_{i=1}^m$. At initialization, the overlaps $\{s_i\}_{i=1}^m$ are then i.i.d. according to the distribution of the push-forward of $\mathcal{U}(\mathcal{S}^{d-1})$ through the Euclidean projection on u_* , denoted P_S . In turn, these variables may be reindexed using a probability space $(\Omega, \mathcal{F}, \rho)$ and random processes $a(t, \omega), s(t, \omega)$ (for $\omega \in \Omega$ and $t \geq 0$), respectively taking values in $\mathbb{R}$ and $[-1, 1]$, and following the evolution

$$\partial_t a(t, \omega) = \frac{1}{\varepsilon} \sum_{k=0}^\infty \sigma_k s(t, \omega)^k \left(\varphi_k - \sigma_k \int a(t, \nu) s(t, \nu)^k d\rho(\nu) \right), \quad (2.8)$$

$$\partial_t s(t, \omega) = a(t, \omega) (1 - s(t, \omega)^2) \sum_{k=1}^\infty k \sigma_k s(t, \omega)^{k-1} \left(\varphi_k - \sigma_k \int a(t, \nu) s(t, \nu)^k d\rho(\nu) \right), \quad (2.9)$$

which can be understood as a gradient flow on the cost function

$$\frac{1}{2} \sum_{k=0}^\infty \left(\varphi_k - \sigma_k \int a(\omega) s(\omega)^k d\rho(\omega) \right)^2. \quad (2.10)$$

The initial condition $a(0, \omega), s(0, \omega)$ is a pair of measurable maps such that pushing ρ through them recovers $P_A \otimes P_S$, and, since P_S converges weakly to a point mass at 0 as d goes to infinity, we have $s(0, \omega) = 0$. Also, if we assume that ρ is the uniform measure on $[0, 1]$, $a(0, \omega)$ is simply the inverse cumulative density function (CDF) of P_A . We refer the reader to Section 4.3 and 4.4 of [13], and to the references therein, for a more detailed discussion on the links between this formulation and the one usually found in the mean-field theory of neural networks [43, 19, 48]. The evolution (2.8)-(2.9) will be the main topic of our study, and, in particular, we seek to understand whether or not the dynamics reaches points where the terms $(\varphi_k - \sigma_k \int a(\omega) s(\omega)^k d\rho(\omega))^2$ are small, for increasing values of k .

Notations. We will denote the L^∞ norm on $[0, 1]$ (or other subsets of the real line) by $\|\cdot\|$, and the Euclidean norm on $\mathbb{R}^n$ (for any integer n) by $\|\cdot\|_2$. Also, for any piecewise continuous, integrable function f on Ω , we denote $\bar{f} = \int_\Omega f(\omega) d\rho(\omega)$, and $f_\perp = f - \bar{f}$. We will sometimes abbreviate the centered, standard Gaussian measure with μ . Finally, for any real number x , we let $(x)_+ = \max(0, x)$.

Remark 2.1. In what follows, various initial value problems related to the evolution (2.8)-(2.9) will be considered, always for initial conditions around which the vector fields defining the dynamics are locally Lipschitz continuous. Thus, we will often omit to justify the existence and uniqueness of maximal solutions, which immediately follow from the Cauchy-Lipschitz theorem.

3. MAIN RESULTS

This section gathers the results that are the most meaningful to describe the learning dynamics in an intuitive way. They are consequences of more general technical results that can be found in Section 7 and Appendix B. Our assumptions are as follows:

Assumption 1.

- The function $a(0, \cdot)$, which will be denoted a_{init} , is piecewise continuous from $[0, 1]$ to $\mathbb{R}$ and there exists a constant M such that $\|a(0, \cdot)\| \leq M$.
- $s(0, \cdot) = 0$.
- The functions σ, σ' and φ are in $L^2(\mu)$. For each $k \geq 0$, $\sigma_k > 0$ and $\varphi_k > 0$.

We will sometimes refer to the parameters $M, \{\sigma_k\}_{k=0}^\infty, \{\varphi_k\}_{k=0}^\infty$ as the problem *data*. Until further notice, all the following statements concern the solution $a(t, \cdot), s(t, \cdot)$ to the initial value problem defined by the evolution (2.8)-(2.9), with any initial condition such that Assumption 1 is verified.

3.1. LEARNING THE CONSTANT AND LINEAR COMPONENTS

Our first main result shows that the timescale on which the constant component (the coefficient φ_0) is learned to a precision close to ε is of order $\frac{\varepsilon \log(\frac{1}{\varepsilon})}{\sigma_0^2}$.

Proposition 3.1 (Learning the constant component). *There exists a positive constant C_0 such that, for any $\eta \in (0, 1)$ and ε sufficiently small,*

$$\inf \left\{ t \geq 0 : \left| \bar{a}(t) - \frac{\varphi_0}{\sigma_0} \right| \leq \varepsilon^\eta + C_0 \varepsilon \log\left(\frac{1}{\varepsilon}\right) \right\} \leq \frac{\left(\eta \log\left(\frac{1}{\varepsilon}\right) + \log(|\bar{a}(0) - \frac{\varphi_0}{\sigma_0}|) \right)_+}{\sigma_0^2} \varepsilon \quad (3.1)$$

The next result characterizes the timescale required to learn the linear component (the coefficient φ_1). The first part of the result shows that learning φ_1 up to a precision of order $o(1)$ is arbitrarily close to $\frac{1}{4\sigma_1\varphi_1} \sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$. The second part shows that, on a longer timescale of order $\mathcal{O}(\sqrt{\varepsilon} \log(\frac{1}{\varepsilon}))$, the coefficient is learned up to a precision of $\varepsilon^{1/4}$.

Proposition 3.2 (Learning the linear component). *For any constant $c_1 > 0$, $\alpha \in (0, 1)$ and ε sufficiently small,*

$$\inf \left\{ t \geq 0 : \bar{a}s(t) = (1 - \alpha) \frac{\varphi_1}{\sigma_1} \right\} \leq \left(\frac{1}{4\sigma_1\varphi_1} + c_1 \right) \sqrt{\varepsilon} \log\left(\frac{1}{\varepsilon}\right). \quad (3.2)$$

Furthermore, there exists a constant $C_1 > \frac{1}{4\sigma_1\varphi_1}$ such that, for ε sufficiently small,

$$\inf \left\{ t \geq 0 : \bar{a}s(t) = (1 - \varepsilon^{1/4}) \frac{\varphi_1}{\sigma_1} \right\} \leq C_1 \sqrt{\varepsilon} \log\left(\frac{1}{\varepsilon}\right). \quad (3.3)$$

Propositions 3.1 and 3.2 are corollaries of the following more general result, that provides a quantitative approximation to the solution of the dynamical system (2.8)-(2.9) on timescales up to order $\mathcal{O}(\sqrt{\varepsilon} \log(\frac{1}{\varepsilon}))$. To state this result, we begin by defining the idealized system around which our approximation is constructed. Let $\bar{b} : \mathbb{R}_+ \rightarrow \mathbb{R}$ be the function

$$\bar{b}(t) = \frac{\varphi_0}{\sigma_0} + \left(\bar{a}(0) - \frac{\varphi_0}{\sigma_0} \right) \exp\left(-\frac{\sigma_0^2 t}{\varepsilon}\right). \quad (3.4)$$

$\bar{b}$ is the solution to the ODE obtained for the quantity $\bar{a}$ if only the term corresponding to the coefficient φ_0 is conserved in the dynamical system (2.8)-(2.9). Similarly, with $\sigma'_1 = \sigma_1 \int_0^1 a_{init,\perp}^2(\nu) d\nu$, let β, γ be the solutions to the dynamical system

$$\partial_t \beta(t) = \frac{\sigma_1}{\varepsilon} \gamma(t) (\varphi_1 - \sigma'_1 \beta(t) \gamma(t)), \quad (3.5)$$

$$\partial_t \gamma(t) = \sigma_1 \beta(t) (\varphi_1 - \sigma'_1 \beta(t) \gamma(t)), \quad (3.6)$$

with initial condition $\beta(0) = 1$ and $\gamma(0) = 0$. This system can be integrated (almost) explicitly, and β, γ are studied in detail in Appendix A. Now, define $b(t, \cdot) = \beta(t) a_{init,\perp}(\cdot)$ and $g(t, \cdot) = \gamma(t) a_{init,\perp}(\cdot)$. The functions b, g are the solution to the dynamical system (2.8)-(2.9) if only the terms corresponding to the coefficient φ_1 are kept. The following theorem provides a quantitative approximation result for $a(t, \cdot), s(t, \cdot)$ in the form of tubes with time-dependent radii around the curves defined by $\bar{b}, b, g$.

Theorem 3.3. *There exists a constant D such that, for any $c \in [0, \frac{1}{8D}]$ and time $(t_\varepsilon)_{\varepsilon>0}$ of the form*

$$t_\varepsilon = \left(\frac{1}{4\sigma_1\varphi_1} + c \right) \sqrt{\varepsilon} \log\left(\frac{1}{\varepsilon}\right), \quad (3.7)$$

there exists a constant C such that, for any $t \in [0, t_\varepsilon]$ and ε sufficiently small,

$$|\bar{a} - \bar{b}|(t) \leq Ct + C\varepsilon\beta^2(t), \quad |\bar{s}(t)| \leq Ct, \quad (3.8)$$

$$\|(a_\perp - b)(t, \cdot)\| + \varepsilon^{-1/2} \|(s_\perp - g)(t, \cdot)\| \leq C\varepsilon^{\frac{1}{4}-cD} \beta(t). \quad (3.9)$$

Furthermore, there exists a constant $\delta \in (0, \frac{1}{8})$ such that, for any constant $K \geq \frac{1}{4\sigma_1\varphi_1} + \frac{1}{8D}$ and any $t \in [t_\varepsilon, K\sqrt{\varepsilon} \log(\frac{1}{\varepsilon})]$, and ε sufficiently small, (3.8) holds and

$$\|(a_\perp - b)(t, \cdot)\| + \varepsilon^{-1/2} \|(s_\perp - g)(t, \cdot)\| \leq C\varepsilon^\delta \beta(t). \quad (3.10)$$

Remark 3.4. One can show that, for all times, $\gamma(t) \leq \sqrt{\varepsilon}\beta(t)$, justifying the sole use of β to state the bounds on both $a_\perp$ and $s_\perp$ in Theorem 3.3.

Remark 3.5. On timescales of order $\mathcal{O}(\varepsilon \log(\frac{1}{\varepsilon}))$, the terms depending on $\beta(t)$ are negligible, so that the bounds relevant for the constant component are linear in time. As we progress towards the linear component, the dependence in $\beta(t)$ becomes relevant, and we rigorously recover the corresponding timescales from [13] down to the constants.

3.2. LEARNING THE QUADRATIC COMPONENT

In the remainder of this section, we will focus on the following simplification of the dynamical system (2.8)-(2.9):

$$\partial_t a(t, \omega) = \frac{\sigma_0}{\varepsilon} (\varphi_0 - \sigma_0 \bar{a}(t)) + \frac{\sigma_1}{\varepsilon} s(t, \omega) (\varphi_1 - \sigma_1 \bar{a}s(t)) + \frac{\sigma_2 \varphi_2}{\varepsilon} s^2(t, \omega), \quad (3.11)$$

$$\partial_t s(t, \omega) = a(t, \omega) \sigma_1 (\varphi_1 - \sigma_1 \bar{a}s(t)) + 2\sigma_2 \varphi_2 a(t, \omega) s(t, \omega). \quad (3.12)$$

We now turn to the learning process of the quadratic part of φ , namely the coefficient φ_2 . A core difficulty in this case is the absence of a clear decomposition of the solution to (2.8)-(2.9), mainly due to the previously learned coefficients φ_0, φ_1 , the contribution of which in (2.8)-(2.9) does not vanish once they have been learned. Thus, we determine the correct idealized system to study the learning dynamics on a relevant timescale (in powers of ε). Such a timescale is identified in Lemma 7.4 by studying the solution to (2.8)-(2.9) when only the terms corresponding to φ_2 are kept, showing that whatever approximation is obtained should hold for times of order at least $\varepsilon^{1/4}$. We begin by describing the family of initial conditions that we will consider, as prescribed by the terminal values of the pair $a(t, \cdot), s(t, \cdot)$ in Theorem 3.3 (and its corollaries). In particular, upon setting a time of the form $t_0 = K\sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$, which we will choose as our origin, in Corollary 7.3, we obtain that there exist positive constants C and δ such that

$$|\varphi_0 - \sigma_0 \bar{a}(t_0)| \leq C\sqrt{\varepsilon}, \quad |\varphi_1 - \sigma_1 \bar{a}s(t_0)| \leq C\varepsilon^{1/4}, \quad |\bar{s}(t_0)| \leq C\sqrt{\varepsilon} \log\left(\frac{1}{\varepsilon}\right), \quad (3.13)$$

$$\|a_\perp(t_0, \cdot) - \left(\frac{\varphi_1}{\sigma'_1}\right)^{1/2} \varepsilon^{-1/4} a_{init,\perp}(\cdot)\| + \varepsilon^{-1/2} \|s_\perp(t_0, \cdot) - \left(\frac{\varphi_1}{\sigma'_1}\right)^{1/2} \varepsilon^{1/4} a_{init,\perp}(\cdot)\| \leq C\varepsilon^{-1/4+\delta}. \quad (3.14)$$

Owing to the initial scale of a, s prescribed above (respectively $\varepsilon^{-1/4}$ and $\varepsilon^{1/4}$), and the fact that the relevant timescale on which we seek to work is of order $\varepsilon^{1/4}$, it is convenient to consider the following change in time and scale to state our approximation result:

$$\hat{a}(t, \omega) = \varepsilon^{1/4} a(\varepsilon^{1/4} t, \omega), \quad (3.15)$$

$$\hat{s}(t, \omega) = \varepsilon^{-1/4} s(\varepsilon^{1/4} t, \omega). \quad (3.16)$$

Our goal now becomes to study the pair $\hat{a}, \hat{s}$ on a timescale of order one. Now, as done for the constant and linear component, we define our idealized dynamics, in this case as the solution to the system

$$\partial_t z_1(t, \omega) = \sigma_0 \alpha_0(t) + \sigma_1 \alpha_1(t) z_2(t, \omega) + \sigma_2 \varphi_2 z_2^2(t, \omega), \quad (3.17)$$

$$\partial_t z_2(t, \omega) = z_1(t, \omega) (\sigma_1 \alpha_1(t) + 2\sigma_2 \varphi_2 z_2(t, \omega)), \quad (3.18)$$

with any initial condition verifying $\bar{z}_1(t_0) = \varepsilon^{1/4} \frac{\varphi_0}{\sigma_0}$ and $\bar{z}_1 \bar{z}_2(t_0) = \frac{\varphi_1}{\sigma_1}$, and where the coefficients $\alpha_0(t), \alpha_1(t)$ are given by

$$\alpha_0(t) = -\frac{1}{\sigma_0} \left(\sigma_1 \alpha_1(t) \bar{z}_2(t) + \sigma_2 \varphi_2 \bar{z}_2^2(t) \right), \quad (3.19)$$

$$\alpha_1(t) = -\frac{2\sigma_2 \varphi_2 \bar{z}_1^2 \bar{z}_2 + \sigma_2 \varphi_2 \bar{z}_2^3 - \sigma_2 \varphi_2 \bar{z}_2^2 \bar{z}_2}{\sigma_1 \bar{z}_1^2 + \sigma_1 (\bar{z}_2^2 - \bar{z}_2^2)}, \quad (3.20)$$

and respectively ensure that

$$\partial_t \bar{z}_1(t) = 0, \quad \text{and} \quad \partial_t \bar{z}_1 \bar{z}_2(t) = 0. \quad (3.21)$$

The following result is a consequence of the more general Proposition B.1 (and of Corollary B.2), which provides an approximation result for families of singularly perturbed flows evolving near certain target sets defined by integral constraints.

Proposition 3.6. *Let $(a(t, \cdot), s(t, \cdot))$ denote the solution to (3.11)-(3.12) with any initial condition $a(t_0, \cdot), s(t_0, \cdot)$ verifying the bounds (3.13)-(3.14). We then have the following results.*

(a) *There exist continuous functions $z_1(t_0, \cdot), z_2(t_0, \cdot)$ and a constant C_0 such that*

$$\bar{z}_1(t_0) = \varepsilon^{1/4} \frac{\varphi_0}{\sigma_0}, \quad \bar{z}_2(t_0) = 0, \quad \bar{z}_1 \bar{z}_2(t_0) = \frac{\varphi_1}{\sigma_1}, \quad (3.22)$$

$$\|\hat{a}(t_0, \cdot) - z_1(t_0, \cdot)\| + \|\hat{s}(t_0, \cdot) - z_2(t_0, \cdot)\| \leq C_0 \varepsilon^{1/4} \log\left(\frac{1}{\varepsilon}\right). \quad (3.23)$$

(b) *Let $(z_1(t, \cdot), z_2(t, \cdot))$ be the solution to (3.17)-(3.18) with initial condition $(z_1(t_0, \cdot), z_2(t_0, \cdot))$. For any constant $R > 0$, consider the hitting time*

$$T_R = \inf\{t \geq t_0 : (\|\hat{a}(t, \cdot)\| + \|\hat{s}(t, \cdot)\|) \vee (\|z_1(t, \cdot)\| + \|z_2(t, \cdot)\|) \geq R\}. \quad (3.24)$$

Then there exist constants C, K depending only on the problem data and R such that, for any fixed time horizon T independent of ε and any $t_0 \leq t \leq T_R \wedge T$,

$$\|(\hat{a}(t, \cdot), \hat{s}(t, \cdot)) - (z_1(t, \cdot), z_2(t, \cdot))\|_2 \leq \left(C_0 \varepsilon^{1/4} \log\left(\frac{1}{\varepsilon}\right) + C \varepsilon^{1/8} (t - t_0) \right) \exp(K(t - t_0)). \quad (3.25)$$

At an intuitive level, the perturbative parameter ε allows to pin the dynamics in a neighborhood of the sets defined by the previously learned components. While the auxiliary dynamical system (3.17)-(3.18) is difficult to study quantitatively, it precisely characterizes the effect of the previously learned coefficients on the remainder of the learning dynamics. In particular, it shows that the effect of these coefficients cannot be ignored, and that, upon reaching the quadratic term, the dynamics (2.8)-(2.9) is not equivalent to the one obtained by keeping only the terms related to the coefficient φ_2 . Through simulations, we show that the maps $a(t, \omega)$ and $s(t, \omega)$ become highly singular, as the corresponding empirical measures evolve towards learning the quadratic component, but must also retain values of $\bar{a}, \bar{s}$ respectively close to $\frac{\varphi_0}{\sigma_0}, \frac{\varphi_1}{\sigma_1}$.

Remark 3.7. A straightforward consequence of the bounds for the initial condition at Eq.(3.13)-(3.14) is that there exists a constant D depending only on $\varphi_0, \varphi_1, \sigma_0, \sigma_1, M$ such that

$$(\|\hat{a}(t_0, \cdot)\| + \|\hat{s}(t_0, \cdot)\|) \vee (\|z_1(t_0, \cdot)\| + \|z_2(t_0, \cdot)\|) \leq D. \quad (3.26)$$

Thus, Proposition 3.6(b) becomes informative when choosing $R > D$, and describes the evolution of $\hat{a}, \hat{s}$ (or equivalently a, s) at least until they reach a large, order one multiple of their initial scale. The meaning

of this hitting time and the required scale for the pair a, s to learn the coefficient φ_2 are discussed further in Subsection 7.2.

Remark 3.8. While the error bound in Eq.(3.25) is exponential in time, we argue that this does not constitute a technical bottleneck. The error is governed by how close the original system (3.11)-(3.12) evolves to the sets defined by the learned coefficients φ_0, φ_1 , and by the possibility to construct initial conditions that exactly satisfy those constraints while remaining close to the true initial condition, as done in Proposition 3.6(a). This is formalized in Lemma 7.5, of which Proposition 3.6 is a straightforward consequence. More precisely, as soon as the exponential factor in Eq.(3.25) becomes larger than a chosen threshold, we may restart the argument, provided the resulting error is negligible compared to $\hat{a}, \hat{s}$ (here they are of order one and the error of order $\varepsilon^{1/8}$). The local nature of the bounds in Proposition B.1 also allows one to adapt the approximation bound (3.25) to whatever scale is reached by $\hat{a}, \hat{s}$. However, this method rests on a quantitative understanding of the auxiliary dynamics (3.17)-(3.18), which remains challenging.

4. FURTHER RELATED WORKS

Hierarchical learning. Hierarchical learning in two-layer neural networks is an active area of research within machine learning theory, with a significant part of the effort directed towards analytically tractable models that reproduce this phenomenon. Building on insights gathered on the well-specified single index problem [56, 9, 51] and the analysis of a single step of gradient descent [23, 5], several works have given quantitative descriptions of hierarchical learning, taking multi-index models as a benchmark problem. Using layerwise training, the work [2] provides sample complexity guarantees for the recovery of the indices with online stochastic gradient descent (SGD) and data uniformly supported on the hypercube. The plateaus in training and test error are associated with the incremental recovery of the support of the index, with dimension-dependent timescales governed by the degree (in the sense of a polynomial decomposition) of the components of the link function carrying subspaces spanned by the hidden multi-index. Similar guarantees are proposed in [24] for Gaussian data with large gradient steps and correspondingly large data batches, also with layerwise training. Finally [14] gives a richer description of the same hierarchical learning phenomenon using the continuous time and population limit of a two-stage algorithm combining successive nonparametric regression steps with gradient steps. The phenomenology described in the present paper is different: the incremental learning is not associated with the latent subspace being supported on components of increasing degree, but rather with the successive learning of the components themselves while non-trivial correlation with the hidden index is achieved as soon as the linear component is learned, and continues to increase thereafter. This simultaneous fitting of the coefficients of φ and progressive recovery of the hidden index presents a core technical difficulty absent from the aforementioned works. Finally, other models for incremental learning have been studied, notably diagonal linear networks [55, 45, 12] and deep linear networks [50, 38]. The resulting dynamics are significantly different from that considered in the present paper, and the models are not directly comparable.

Mean-field theory of neural networks. Following the identification of the mean field limit for infinitely wide neural networks in [19, 44, 48], several works have proposed quantitative analysis of the resulting trajectories for models trained on synthetic data involving low-dimensional latent representations, or some degree of symmetry. The work [1] studies the recovery of multi-index models with input data uniformly distributed on the hypercube. Using the invariances of the model, the authors reduce the mean field gradient flow (i.e. the transport equation for the empirical measure of the weights) to a dimension free one and provide guarantees for the recovery of the support of the hidden indices. For two-layer networks with ReLU activation, a similar use of symmetries is proposed in [34] which establishes invariance properties of the resulting gradient flow and provides exponential convergence guarantees in a setting reducing to a linear model. Also, the works [53, 3] combine the approaches of [49, 33] with quantitative bounds for the validity of the mean-field limit from [43] to obtain low-dimensional representations (or dimension-free ones at the level of distributions) of the trajectories of online SGD for wide two-layer neural networks. None of these works provide quantitative convergence guarantees or a detailed analysis of how a two-layer neural network solves a semi-parametric problem when trained with gradient flow. We note that the bounds from [13] (see Corollary 1 and Proposition 3), based on propagation of chaos results from [43], are insufficient to show that the mean-field limit is valid on the timescales describing the behaviour of the evolution (2.8)-(2.9) outside of regimes where the ambient dimension and number

of neurons diverge exponentially with $\frac{1}{\varepsilon}$. In the context of machine learning, uniform-in-time bounds for propagation of chaos in single layer neural networks have recently been proposed in [41, 32, 31], with [41] carrying out a complete convergence analysis of the resulting dynamics for a specific instance of a single index target function. Their results do not apply directly to our setting, which involves a two-layer neural network and a singular perturbation. We refer the reader to the related works section of the latter papers for a more detailed discussion on propagation of chaos for neural networks, and leave the development of methods leading to uniform-in-time bounds for our setting to later work.

The two-timescale regime. Two-timescale algorithms and their connections to singular ordinary differential equations appear in [16], and have since then been used in stochastic approximation [39], reinforcement learning [22] and optimization [36]. In the context of two-layer neural networks, the paper [42] provides a detailed study of the training dynamics of a specific instance of a two-layer neural network solving a univariate non-parametric regression problem. By only tuning the second layer weights and biases, the authors provide quantitative convergence guarantees for the two-timescale population gradient flow learning a piecewise constant function, with explicit bounds on the required number of neurons and relative speed parameter ε for the convergence to hold. In particular, their work shows experimentally that the gradient flow may fail to converge without a timescale separation between the training of the outer layer weights and biases. On the one hand, our study is asymptotic in the number of neurons and, although explicit constants for the parameter ε can be extracted from our proof method, we do not track them. On the other hand, our target function is high-dimensional and more complex, with the learning architecture more general and closer to practical ones.

5. SIMULATIONS

Letting h_k denote the k -th normalized Hermite polynomial, we illustrate the results from Section 3 through simulations with the following activation and target link function.

$$\sigma(x) = 0.4h_0(x) + 0.5h_1(x) + 0.2h_2(x), \quad (5.1)$$

$$\varphi(x) = h_0(x) + 2h_1(x) + h_2(x). \quad (5.2)$$

We begin by verifying the separation of timescales for the learning of each of the components $\varphi_0, \varphi_1, \varphi_2$ with simulations reminiscent of those proposed in [13], before turning to the approximation results presented in Theorem 3.3 and Proposition 3.6. In all figures, we choose $\Omega = [0, 1]$ and ρ the corresponding uniform measure, along with $a_{init}(\omega) = -1 + 2\omega$, the inverse CDF of the uniform measure on $[-1, 1]$.

5.1. LEARNING THE CONSTANT AND LINEAR COMPONENTS

Figures 1, 2 and 3 illustrate the results of Section 3 for the constant and linear components. For clarity of exposition, we simulate the dynamical system (2.8)-(2.9) with only the first two coefficients of the above activation and link function.

5.2. LEARNING THE QUADRATIC COMPONENT

Figures 4, 5 and 6 illustrate the results of Section 3 for the quadratic component. In Figure 5, we compare the rescaled maps $\hat{a}(t, \omega), \hat{s}(t, \omega)$ defined at Eq.(3.15)-(3.16), that obey the evolution

$$\partial_t \hat{a}(t, \omega) = \frac{\sigma_0}{\sqrt{\varepsilon}} \left(\varphi_0 - \sigma_0 \varepsilon^{-1/4} \hat{a}(t) \right) + \frac{\sigma_1}{\varepsilon^{1/4}} \hat{s}(t, \omega) (\varphi_1 - \sigma_1 \hat{a} \hat{s}(t)) + \sigma_2 \varphi_2 \hat{s}^2(t, \omega), \quad (5.3)$$

$$\partial_t \hat{s}(t, \omega) = \hat{a}(t, \omega) \frac{\sigma_1}{\varepsilon^{1/4}} (\varphi_1 - \sigma_1 \hat{a} \hat{s}(t)) + 2\sigma_2 \varphi_2 \hat{a}(t, \omega) \hat{s}(t, \omega), \quad (5.4)$$

the idealized maps $z_1(t, \omega), z_2(t, \omega)$, solutions to the dynamical system Eq.(3.17)-(3.18), and the maps $\hat{a}_q(t, \omega), \hat{s}_q(t, \omega)$ corresponding to the solutions to Eq.(2.8)-(2.9) if only the coefficients σ_2, φ_2 are kept (i.e. neglecting the effect of the learned coefficients φ_0, φ_1), with the change in time and scale defined at (3.15)-(3.16). Namely, $\hat{a}_q(t, \omega), \hat{s}_q(t, \omega)$ are solutions to the dynamical system

$$\partial_t \hat{a}_q(t, \omega) = \sigma_2 \varphi_2 \hat{s}_q^2(t, \omega), \quad (5.5)$$

$$\partial_t \hat{s}_q(t, \omega) = 2\sigma_2 \varphi_2 \hat{a}_q(t, \omega) \hat{s}_q(t, \omega). \quad (5.6)$$

The initial conditions for these maps are constructed by solving Eq.(2.8)-(2.9) until the first two coefficients are learned to a chosen precision. To summarize these maps, we compare the quantities $\overline{\hat{a}\hat{s}^2}, \overline{\hat{a}_q\hat{s}_q^2}$ and $\overline{z_1 z_2^2}$, which govern the learning of the coefficient φ_2 . Note that, owing to the change in time and

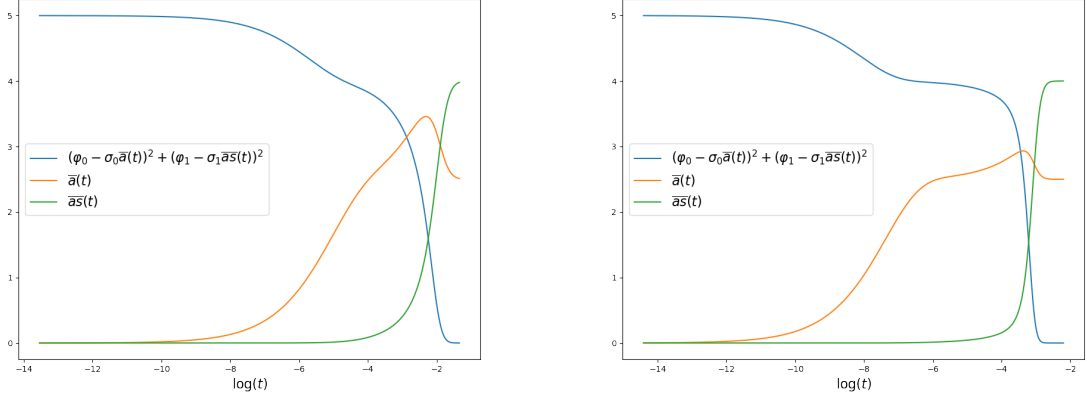


FIGURE 1. Separation of timescales for the constant and linear components: evolution of the averaged quantities $\bar{a}$, $\bar{a}s$, and cost function for the first two coefficients φ_0, φ_1 as a function of time on a logarithmic scale, for $\varepsilon = 10^{-3}$ (left) and $\varepsilon = 10^{-4}$ (right).

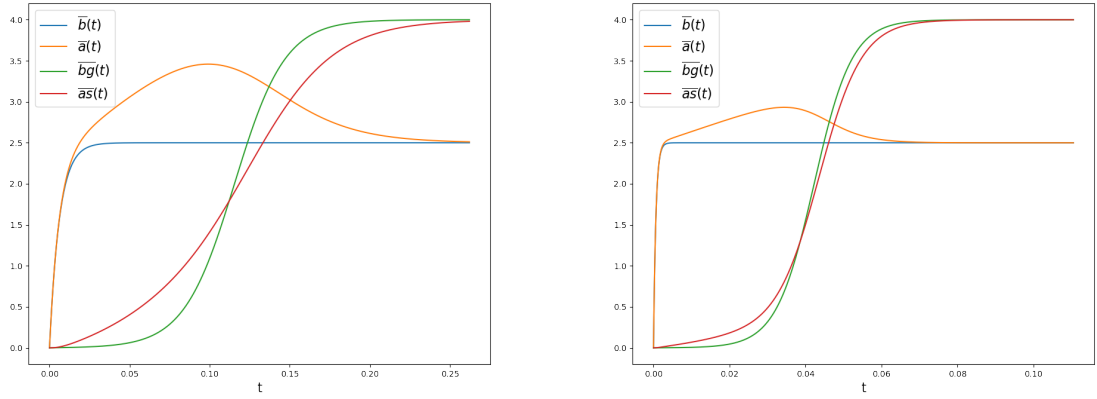


FIGURE 2. Illustration of Theorem 3.3: comparison of $\bar{a}$ (orange curve) with $\bar{b}$ (blue curve), and of $\bar{a}s$ (red curve) with $\bar{b}g$ (green curve), for $\varepsilon = 10^{-3}$ (left) and $\varepsilon = 10^{-4}$ (right). On a timescale of order $\varepsilon \log(1/\varepsilon)$, the blue and orange curves are nearly indistinguishable, and, when moving to a timescale of order $\sqrt{\varepsilon} \log(1/\varepsilon)$, the error becomes larger, as predicted by the bound (3.8) (see also Remark 3.5). The curves collapse at longer times owing to the absence of higher-order terms in the simulation. Overall the relative error between the curves decreases with ε .

scale, the target value for these quantities is $\varepsilon^{-1/4} \frac{\varphi_2}{\sigma_2}$. Finally, in Figure 6, we show the evolution of the maps $z_1(t, \omega)$ and $z_2(t, \omega)$, from the initial condition described above. Due to our choice of (Ω, ρ) , this initial condition corresponds to multiples (the terminal values of β, γ) of $a_{init} = -1 + 2\omega$.

6. OUTLINE OF THE PROOF

The purpose of this section is to give an informal description of the proof method. It revolves around the identification, at various timescales, of an idealized dynamical system that may be (almost) integrated explicitly and captures the phenomenology of the original one. Owing to the initial condition, we expect to be able to successively approximate the evolution (2.8)-(2.9) by the terms corresponding to each harmonic at appropriate timescales. Two major difficulties appear: firstly, obtaining a quantitative control on such approximations and the timescales on which they are valid; and secondly, summarizing the effect of

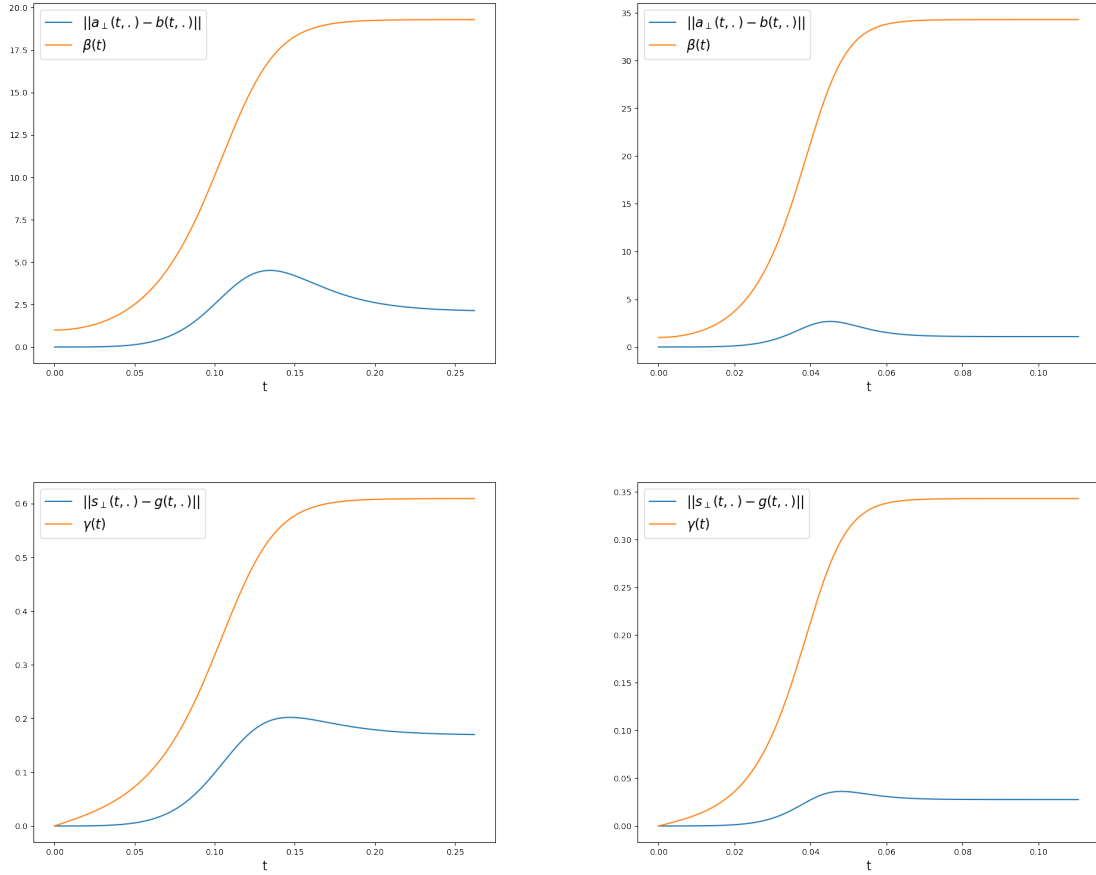


FIGURE 3. Illustration of Theorem 3.3: comparison of $\|a_{\perp}(t, \cdot) - b(t, \cdot)\|$ with $\beta(t)$ (top two figures), and $\|s_{\perp}(t, \cdot) - g(t, \cdot)\|$ with $\gamma(t)$ (bottom two figures), for $\varepsilon = 10^{-3}$ (left panel) and $\varepsilon = 10^{-4}$ (right panel). The approximation errors roughly behave as small multiples of $\beta(t)$, as predicted by the bound (3.9) (see also Remark 3.4), with the error decreasing with ε .

coefficients once they have been learned. Indeed, we cannot expect that, once a certain harmonic has been learned, its effect becomes completely negligible as the system progresses to the next one. As we will see below, when studying the coefficients corresponding to the constant and linear part of φ , an explicit decomposition of the solution allows one to circumvent this second issue. Therefore we postpone dealing with this difficulty to Subsection 6.2.

6.1. CONSTRUCTING TUBES AROUND IDEALIZED SYSTEMS

Let us now focus on the first difficulty described above. A naive idea is to introduce sequences of thresholds for the variables a, s (or related quantities such as their mean or supremum over Ω), and use them to approximate (2.8)-(2.9) around the leading order terms up to the first hitting times of these thresholds. While this strategy can lead to preliminary results, it becomes quickly laborious as more complex dynamics are considered. A more efficient solution is to introduce tubes of varying radii around the curves defined by the idealized system, and show that the original system remains in those tubes on relevant timescales. We will now outline how to determine the radii of those tubes, namely using the knowledge obtained from integrating the idealized system, before formalizing this discussion in Section 7.

In what follows, D will denote a constant that depends only on $\{\varphi_k, \sigma_k\}_{k \geq 0}$ and M , and whose exact

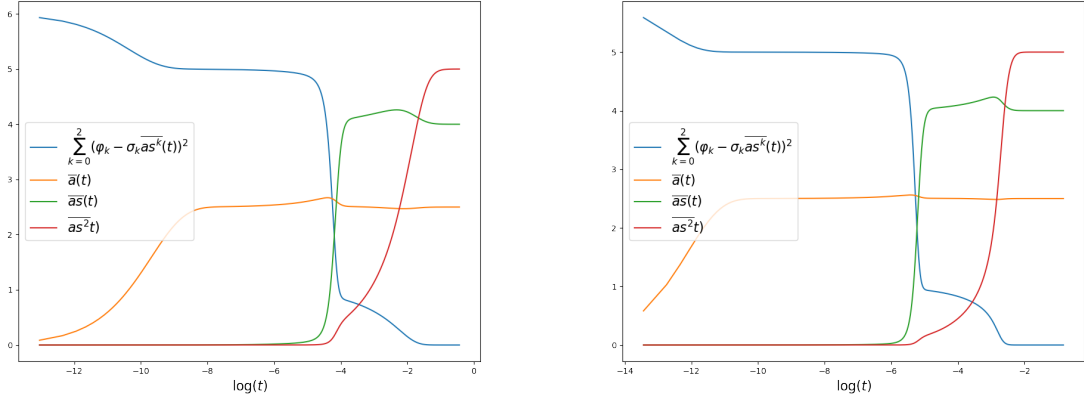


FIGURE 4. Separation of timescales for the constant, linear and quadratic component: evolution of the quantities $\bar{a}, \bar{a}s, \bar{a}s^2$, along with the cost function for the first three coefficients $\varphi_0, \varphi_1, \varphi_2$, as a function of time on a logarithmic scale, for $\varepsilon = 10^{-5}$ (left) and $\varepsilon = 10^{-6}$ (right). Note that, to better observe the separation, smaller values of ε must be used than for the first two components.

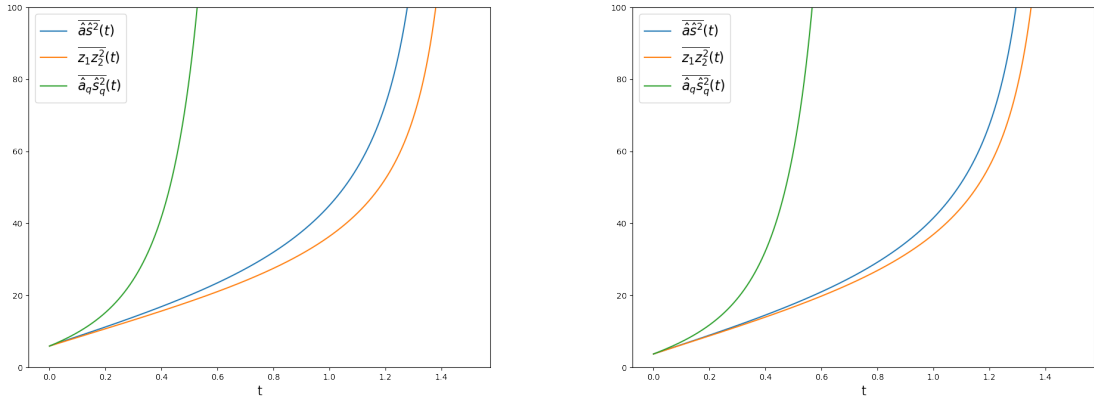


FIGURE 5. Illustration of Proposition 3.6: comparison of the evolution of $\bar{\hat{a}}\hat{s}^2$ (blue curve), $\bar{\hat{a}}_q\hat{s}_q^2$ (green curve) and $\overline{z_1 z_2^2}$ (orange curve), for $\varepsilon = 10^{-5}$ (left) and $\varepsilon = 10^{-6}$ (right). The explosion time for the quantity $\bar{\hat{a}}_q\hat{s}_q^2$ is significantly smaller than that of $\bar{\hat{a}}\hat{s}^2$, whose evolution is much closer to that of the idealized quantity $\overline{z_1 z_2^2}$. The approximation becomes more accurate as ε decreases.

value may change from line to line. We begin by noting that, near initialization,

$$|\partial_t \bar{a}(t) - \frac{\sigma_0}{\varepsilon} (\varphi_0 - \sigma_0 \bar{a}(t))| \leq \frac{D}{\varepsilon} \|s(t, \cdot)\|, \quad (6.1)$$

so that we expect $\bar{a}(t)$ to remain close to the solution to the ODE

$$\begin{cases} \frac{d}{dt} \bar{b}(t) = \frac{\sigma_0}{\varepsilon} (\varphi_0 - \sigma_0 \bar{b}(t)), \\ \bar{b}(0) = \bar{a}(0), \end{cases} \quad (6.2)$$

namely

$$\bar{b}(t) = \frac{\varphi_0}{\sigma_0} + \left(\bar{a}(0) - \frac{\varphi_0}{\sigma_0} \right) \exp\left(-\frac{\sigma_0^2 t}{\varepsilon}\right). \quad (6.3)$$

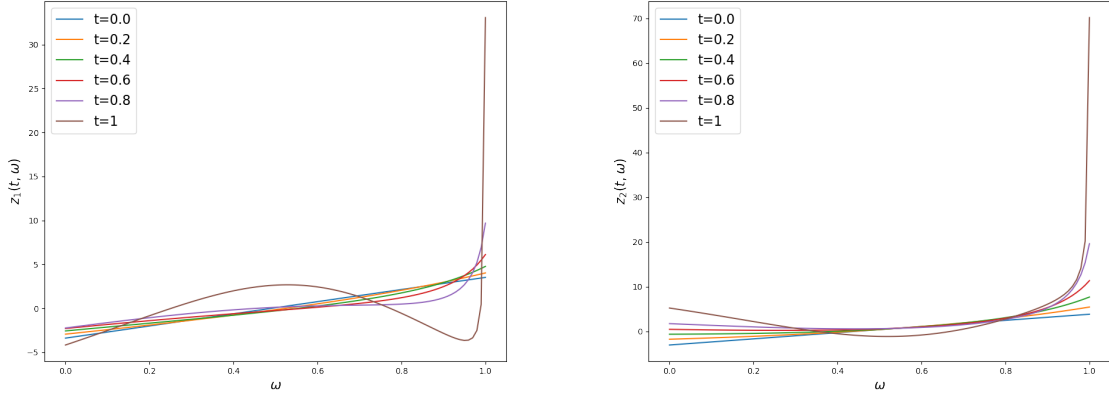


FIGURE 6. Evolution of the maps $z_1(t, \cdot), z_2(t, \cdot)$, in the setting of Figure 5, with $\varepsilon = 10^{-6}$, for fractions of the total running time of Figure 5. A small fraction of neurons diverge towards learning the coefficient φ_2 , while the remaining neurons adapt to the change to conserve the constraints on $\bar{z}_1, \bar{z}_1 \bar{z}_2$.

This function immediately gives the required timescale to learn the constant component, provided we establish a sufficiently sharp control of $\bar{a}$ around $\bar{b}$. Upon rewriting the bound (6.1) as

$$\left| \partial_t (\bar{a}(t) - \bar{b}(t)) + \frac{\sigma_0^2}{\varepsilon} (\bar{a}(t) - \bar{b}(t)) \right| \leq \frac{D}{\varepsilon} \|s(t, \cdot)\|, \quad (6.4)$$

we obtain the first tube for $\bar{a}$ around $\bar{b}$, the radius of which will depend on our control over $\|s(t, \cdot)\|$. Noting that we also have

$$\partial_t \|s(t, \cdot)\| \leq D \|a(t, \cdot)\|, \quad (6.5)$$

$$\partial_t \|a_\perp(t, \cdot)\| \leq \frac{D}{\varepsilon} \|s(t, \cdot)\|, \quad (6.6)$$

near initialization, we may close this system of inequalities to obtain an upper bounding flow on $\|s(t, \cdot)\|$ and conclude with respect to the learning process of the coefficient φ_0 . Instead of doing so, we seek to refine our approximation to also include the learning process of φ_1 . Upon writing the evolution of $a_\perp, s_\perp$, we find:

$$\left| \partial_t a_\perp - \frac{\sigma_1}{\varepsilon} s_\perp (\varphi_1 - \sigma_1 \bar{a} \bar{s}) \right| \leq \frac{D}{\varepsilon} \|s_\perp(t, \cdot)\|^2, \quad (6.7)$$

$$\left| \partial_t s_\perp - \sigma_1 a_\perp (\varphi_1 - \sigma_1 \bar{a} \bar{s}) \right| \leq D \|a_\perp(t, \cdot)\| \|s_\perp(t, \cdot)\|. \quad (6.8)$$

Assuming that $\bar{a}$ remains close to $\bar{b}$, $\bar{s}$ will remain of order Dt (with t a fractional power of ε) until $\bar{a} \bar{s}$ becomes of order one, which is necessary to learn the coefficient φ_1 . Thus, the term $\bar{a} \bar{s}$ should be well approximated by $\bar{a}_\perp \bar{s}_\perp$, and the functions $a_\perp, s_\perp$ should remain close to the solutions to the system

$$\partial_t b = \frac{\sigma_1}{\varepsilon} g(\varphi_1 - \sigma_1 \bar{b} \bar{g}), \quad (6.9)$$

$$\partial_t g = \sigma_1 b(\varphi_1 - \sigma_1 \bar{b} \bar{g}), \quad (6.10)$$

with initial condition $b(0, \cdot) = a_{init, \perp}(\cdot)$ and $g(0, \cdot) = 0$, which can be solved analytically. As before, we can use this solution to define tubes for $a_\perp, s_\perp$ respectively around the curves b, g , and combine them with the tube for $\bar{a}$ around $\bar{b}$ along with the upper bound for $\bar{s}$ (i.e. a tube around 0). The radii of the tubes are then determined by plugging the idealized solutions $\bar{b}, b, g$ in the higher order terms of the evolution equations (2.8)-(2.9), making it possible to complete the proof upon introducing an appropriate stopping time to control the validity of the approximation. These steps are formalized in Subsection 7.1.

6.2. APPROXIMATING THE LEARNING DYNAMICS NEAR TARGET SETS

As we move to the quadratic component, there is no clear decomposition of the solution to (2.8)-(2.9) that leads to a simpler auxiliary dynamical system to study. Indeed, as shown by the bounds (3.13)-(3.14), once the first two coefficients have been learned, the terms $\varphi_0 - \sigma_0 \bar{a}$, $s(\varphi_1 - \sigma_1 \bar{a} \bar{s})$ in (2.8) and $a(\varphi_1 - \sigma_1 \bar{a} \bar{s})$ in (2.9) are respectively of the same order of magnitude as the terms a^2 and as that drive the dynamics towards learning the coefficient φ_2 . There is, however, no straightforward way to show that one dominates the other. We will show instead that the learning dynamics are well approximated by an auxiliary system in which the integral constraints defined by the coefficients φ_0, φ_1 are exactly verified, and are enforced through time dependent coefficients akin to Lagrange multipliers. To provide some understanding of this procedure, let us look at a simplified example. With f, g smooth vector fields on $\mathbb{R}^d$ and $h : \mathbb{R}^d \rightarrow \mathbb{R}_+$ a smooth function, consider the following dynamical system

$$\frac{d}{dt} z_\varepsilon(t) = \frac{h(z_\varepsilon(t))}{\varepsilon} g(z_\varepsilon(t)) + f(z_\varepsilon(t)), \quad (6.11)$$

with any initial condition $z_\varepsilon(0)$ such that there exists a non-trivial time interval $[0, T_0]$ on which a solution exists. Assume that there exists a time T_1 and a continuous, positive, increasing function $u(\varepsilon)$ such that $u(0) = 0$ and such that, for any $0 \leq t \leq T_1$, $h(z_\varepsilon(t)) \leq u(\varepsilon)$. We claim that, for any t and δ such that $t + \delta \leq T_0 \wedge T_1$, the increment $z_\varepsilon(t + \delta) - z_\varepsilon(t)$ is close to the tangent space at $z_\varepsilon(t)$ of the embedded submanifold

$$\mathcal{S} = \{z \in \mathbb{R}^d : h(z) = 0\}. \quad (6.12)$$

Indeed, a second order Taylor expansion of $h(z_\varepsilon(t + \delta))$ around t gives

$$|\langle \nabla h(z_\varepsilon(t)), z_\varepsilon(t + \delta) - z_\varepsilon(t) \rangle| \leq 2u(\varepsilon) + \mathcal{O}(\delta^2). \quad (6.13)$$

Using this inequality and appropriately expanding $z_\varepsilon(t + \delta)$ around $z_\varepsilon(t)$ and, in turn, the flow Eq.(6.11), one may show that the considered solution to (6.11) is close to that of the Cauchy problem

$$\begin{cases} \frac{d}{dt} z(t) = \alpha(t) g(z(t)) + f(z(t)), \\ z(0) = z_0, \end{cases} \quad (6.14)$$

where z_0 is such that $h(z_0) = 0$ (and is sufficiently close to $z_\varepsilon(0)$), and where the coefficient $\alpha(t)$ is given by

$$\alpha(t) = -\frac{\langle \nabla h(z(t)), f(z(t)) \rangle}{\langle \nabla h(z(t)), g(z(t)) \rangle}, \quad (6.15)$$

and imposes that $\frac{d}{dt} h(z(t)) = 0$. This operation is reminiscent of the orthogonal projection on a tangent space used to compute gradients on smooth manifolds, see e.g. [17], but the projection is skewed in the direction of the vector field g . This intuition is formalized by Proposition B.1 in a more general setup (not specific to neural networks), which we specialize to the evolution (2.8)-(2.9) in Corollary B.2. This ultimately leads to Proposition 3.6 upon combining the aforementioned approximation result with the control over the quantities $|\varphi_0 - \sigma_0 \bar{a}|$ and $|\varphi_1 - \sigma_1 \bar{a} \bar{s}|$ established in Theorem 3.3. We leave the construction of tubes, used to prove Theorem 3.3, around the idealized system identified in Proposition 3.6, for future work.

Remark 6.1. As a benefit of working with increments, the proof of Proposition B.1 does not require to estimate derivatives of the quantity $h(z_\varepsilon(t))$, for which we only have the control $h(z_\varepsilon(t)) \leq u(\varepsilon)$; these derivatives may display singular behaviour.

Remark 6.2. We note that a Lagrange multiplier appears in the matched asymptotic expansion argument of [13] when describing the learning dynamics of the coefficient φ_1 , to enforce the constraint $\bar{a} = \frac{\varphi_0}{\sigma_0}$. While the proof of Theorem 3.3 can also be carried out using the approximation result from Proposition B.1 and the iteration argument outlined in Remark 3.8, the decomposition $a = \bar{a} + a_\perp$ and $s = \bar{s} + s_\perp$ naturally leads to the same auxiliary dynamics, allowing for a more direct proof. This, however, is only possible for the coefficients φ_0, φ_1 , while Proposition B.1 allows one to identify the correct auxiliary system at any order.

7. PROOFS

We begin with the constant and linear components, based on the construction of tubes outlined in Section 6.1.

7.1. LEARNING THE CONSTANT AND LINEAR COMPONENTS

In this section, the definitions of b, g, γ, β are those from Appendix A, and $\bar{b}$ is given by (3.4).

Theorem 7.1. *There exists a constant $D < +\infty$ such that, for any $c \in [0, \frac{1}{8D}]$ and time $(t_\varepsilon)_{\varepsilon>0}$ of the form*

$$t_\varepsilon = \left(\frac{1}{4\sigma_1\varphi_1} + c \right) \sqrt{\varepsilon} \log\left(\frac{1}{\varepsilon}\right), \quad (7.1)$$

there exists a constant $C < +\infty$ such that the following holds for every ε sufficiently small. For every $t \leq t_\varepsilon$, we have

$$|\bar{a} - \bar{b}|(t) \leq Ct + C\varepsilon\beta^2(t), \quad |\bar{s}(t)| \leq Ct, \quad (7.2)$$

$$\|(a_\perp - b)(t, \cdot)\| + \varepsilon^{-1/2}\|(s_\perp - g)(t, \cdot)\| \leq C\varepsilon^{\frac{1}{4}-cD}\beta(t). \quad (7.3)$$

Proof. Let us define the hitting time

$$T := \inf \left\{ t \geq 0 : \|a\| + \varepsilon^{-1/2}\|s\| \geq C\beta \quad \text{or} \quad \|a_\perp - b\| + \varepsilon^{-1/2}\|s_\perp - g\| \geq \varepsilon^\eta\beta \right\}, \quad (7.4)$$

where C and η are constants to be determined in the course of the proof. Throughout the argument, we write $D = D(\text{data}) < +\infty$ for a large constant whose value may change from place to place; we stress in particular that the constant D does not depend on the choice of the constant C . Although we will not always state it explicitly, all the derivations in this proof are performed for times $t \leq T \wedge t_\varepsilon$ only.

Controlling $\bar{a}, \bar{s}$. Recall that

$$\partial_t \bar{a} = \frac{1}{\varepsilon} \sum_{k=0}^{\infty} \sigma_k \bar{s}^k \left(\varphi_k - \sigma_k \bar{a} \bar{s}^k \right). \quad (7.5)$$

Then, note that, on the considered time interval

$$\|a\| + \varepsilon^{-1/2}\|s\| \leq C\beta \leq CD\varepsilon^{-1/4}, \quad (7.6)$$

where we used the fact that $\beta(t)$ is increasing and that $\beta(t_\varepsilon) \leq D\varepsilon^{-1/4}$. Thus,

$$\left| \partial_t \bar{a} - \left(\frac{\sigma_0}{\varepsilon} (\varphi_0 - \sigma_0 \bar{a}) + \frac{\sigma_1 \bar{s}}{\varepsilon} (\varphi_1 - \sigma_1 \bar{a} \bar{s}) \right) \right| \leq \left| \frac{1}{\varepsilon} \sum_{k=2}^{\infty} \sigma_k \bar{s}^k \left(\varphi_k - \sigma_k \bar{a} \bar{s}^k \right) \right|. \quad (7.7)$$

But then, for any $k \geq 2$, $\bar{a} \bar{s}^k \leq \bar{a} \bar{s}^2 \leq C^3 D^3 \varepsilon^{1/4}$, so that, for ε sufficiently small,

$$\frac{1}{\varepsilon} \sum_{k=2}^{\infty} \sigma_k \bar{s}^k \left(\varphi_k - \sigma_k \bar{a} \bar{s}^k \right) \leq C^2 D \beta^2. \quad (7.8)$$

Owing to the definition of $\bar{b}$, we may write

$$|\partial_t (\bar{a} - \bar{b}) + \frac{\sigma_0}{\varepsilon} (\bar{a} - \bar{b})| \leq \frac{\sigma_1 |\bar{s}(t)|}{\varepsilon} |\varphi_1 - \sigma_1 \bar{a} \bar{s}| + C^2 D \beta, \quad (7.9)$$

which, upon noting that $\bar{a} \bar{s} \leq C^2 \sqrt{\varepsilon} \beta^2 \leq C^2 D$, leads to

$$|\partial_t (\bar{a} - \bar{b}) + \frac{\sigma_0}{\varepsilon} (\bar{a} - \bar{b})| \leq \frac{C^2 D |\bar{s}|}{\varepsilon} + C^2 D \beta^2. \quad (7.10)$$

Integrating the above and using the fact that β is increasing, we find

$$|\bar{a}(t) - \bar{b}(t)| \leq e^{-\frac{\sigma_0}{\varepsilon} t} \int_0^t e^{\frac{\sigma_0}{\varepsilon} u} \left(\frac{C^2 D |\bar{s}(u)|}{\varepsilon} + C^2 D \beta^2(u) \right) du, \quad (7.11)$$

$$\leq e^{-\frac{\sigma_0}{\varepsilon} t} \left(\frac{C^2 D |\bar{S}(t)|}{\varepsilon} + C^2 D \beta^2(t) \right) \int_0^t e^{\frac{\sigma_0}{\varepsilon} u} du, \quad (7.12)$$

where we introduced $\bar{S}(t) = \sup_{r \leq t} |\bar{s}(r)|$. Finally, we obtain

$$|\bar{a}(t) - \bar{b}(t)| \leq C^2 D (\bar{S}(t) + \varepsilon \beta^2(t)). \quad (7.13)$$

Turning to $\bar{s}$, we write

$$\partial_t \bar{s} = \bar{a} \sigma_1 (\varphi_1 - \sigma_1 \bar{a} \bar{s}) + \sum_{k=2}^{\infty} k \sigma_k \bar{a} \bar{s}^{k-1} \left(\varphi_k - \sigma_k \bar{a} \bar{s}^k \right) - \sum_{k=1}^{\infty} k \sigma_k \bar{a} \bar{s}^{k+1} \left(\varphi_k - \sigma_k \bar{a} \bar{s}^k \right). \quad (7.14)$$

Recalling that $\overline{as^2} \leq C^3 D^3 \varepsilon^{1/4}$, and that $\sum_{k=1}^{\infty} k \sigma_k < \infty$, we find that, for ε sufficiently small

$$|\partial_t \bar{s} - \bar{a} \sigma_1 (\varphi_1 - \sigma_1 \bar{a} \bar{s})| \leq C^2 D \sqrt{\varepsilon} \beta^2 + C^3 D \varepsilon^{1/4}, \quad (7.15)$$

so that

$$|\partial_t \bar{s}| \leq C^2 D (|\bar{a}| + 1). \quad (7.16)$$

Using the decomposition $|\bar{a}| \leq |\bar{b}| + |\bar{a} - \bar{b}|$ along with the bound Eq.(7.13), we find

$$|\bar{a}| \leq C^2 D (\bar{S} + \varepsilon \beta^2 + 1), \quad (7.17)$$

which, recalling that $\beta^2 \leq \frac{D}{\sqrt{\varepsilon}}$, leads to

$$\partial_t \bar{s} \leq C^4 D (\bar{S} + 1). \quad (7.18)$$

This, along with Lemma C.1, implies that

$$\partial_t \bar{S} \leq C^4 D (\bar{S} + 1), \quad (7.19)$$

and

$$\bar{S}(t) \leq C^4 D t. \quad (7.20)$$

Finally, using the above bound in Eq.(7.13), we find

$$|\bar{a}(t) - \bar{b}(t)| \leq C^4 D (t + \varepsilon \beta^2(t)). \quad (7.21)$$

Controlling $a_{\perp}, s_{\perp}$. We now write

$$\partial_t a_{\perp} = \frac{1}{\varepsilon} \sigma_1 s_{\perp} (\varphi_1 - \sigma_1 \bar{a} \bar{s}) + \frac{1}{\varepsilon} \sum_{k=2}^{\infty} \sigma_k (s^k - \bar{s}^k) (\varphi_k - \sigma_k \bar{a} \bar{s}^k), \quad (7.22)$$

and

$$\begin{aligned} \partial_t s_{\perp} &= \sigma_1 a_{\perp} (\varphi_1 - \sigma_1 \bar{a} \bar{s}) + \sum_{k=2}^{\infty} k \sigma_k (a s^{k-1} - \bar{a} \bar{s}^{k-1}) (\varphi_k - \sigma_k \bar{a} \bar{s}^k) \\ &\quad - \sum_{k=1}^{\infty} k \sigma_k (a s^{k+1} - \bar{a} \bar{s}^{k+1}) (\varphi_k - \sigma_k \bar{a} \bar{s}^k). \end{aligned} \quad (7.23)$$

Now note that $\bar{a} \bar{s} = \overline{a_{\perp} s_{\perp}} + \bar{a} \bar{s}$, so that

$$\partial_t a_{\perp} = \frac{1}{\varepsilon} \sigma_1 s_{\perp} (\varphi_1 - \sigma_1 \overline{a_{\perp} s_{\perp}}) + e_1(t), \quad (7.24)$$

$$\partial_t s_{\perp} = \sigma_1 a_{\perp} (\varphi_1 - \sigma_1 \overline{a_{\perp} s_{\perp}}) + e_2(t), \quad (7.25)$$

where

$$e_1(t) = -\frac{1}{\varepsilon} \sigma_1^2 s_{\perp} \bar{a} \bar{s} + \frac{1}{\varepsilon} \sum_{k=2}^{\infty} \sigma_k (s^k - \bar{s}^k) (\varphi_k - \sigma_k \bar{a} \bar{s}^k), \quad (7.26)$$

$$e_2(t) = -\sigma_1^2 a_{\perp} \bar{a} \bar{s} + \sum_{k=2}^{\infty} k \sigma_k (a s^{k-1} - \bar{a} \bar{s}^{k-1}) (\varphi_k - \sigma_k \bar{a} \bar{s}^k) - \sum_{k=1}^{\infty} k \sigma_k (a s^{k+1} - \bar{a} \bar{s}^{k+1}) (\varphi_k - \sigma_k \bar{a} \bar{s}^k). \quad (7.27)$$

To establish upper bounds on e_1, e_2 , we use (7.13) and (7.20) along with the definition of T to find

$$s_{\perp} \bar{a} \bar{s} \leq C \sqrt{\varepsilon} \beta C^4 D (t + \varepsilon \beta^2(t)) C^4 D t, \quad (7.28)$$

$$a_{\perp} \bar{a} \bar{s} \leq C \beta C^4 D (t + \varepsilon \beta^2(t)) C^4 D t. \quad (7.29)$$

Similarly, for any $t \leq T$ and ε sufficiently small,

$$\left| \frac{1}{\varepsilon} \sum_{k=2}^{\infty} \sigma_k (s^k - \bar{s}^k) (\varphi_k - \sigma_k \bar{a} \bar{s}^k) \right| \leq C^2 D \beta^2, \quad (7.30)$$

$$\left| \sum_{k=2}^{\infty} k \sigma_k (a s^{k-1} - \bar{a} \bar{s}^{k-1}) (\varphi_k - \sigma_k \bar{a} \bar{s}^k) \right| \leq C^2 D \sqrt{\varepsilon} \beta^2, \quad (7.31)$$

$$\left| \sum_{k=1}^{\infty} k \sigma_k (a s^{k+1} - \bar{a} \bar{s}^{k+1}) (\varphi_k - \sigma_k \bar{a} \bar{s}^k) \right| \leq C^3 D \varepsilon \beta^3. \quad (7.32)$$

Now, since $t \leq D\sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$ and $\beta(t) \leq D\varepsilon^{-1/4}$, we reach that, for ε sufficiently small,

$$|e_1(t)| + \varepsilon^{-1/2}|e_2(t)| \leq C^2 D\beta^2(t). \quad (7.33)$$

To linearize $a_\perp, s_\perp$ around b, g , we introduce $u := a_\perp - b$ and $v = s_\perp - g$, so that

$$\overline{a_\perp s_\perp} = \overline{bg} + \overline{ug} + \overline{bv} + \overline{uv}, \quad (7.34)$$

and

$$\partial_t a_\perp = \frac{1}{\varepsilon} \sigma_1 s_\perp (\varphi_1 - \sigma_1 \overline{bg}) - \frac{1}{\varepsilon} \sigma_1^2 s_\perp (\overline{ug} + \overline{bv} + \overline{uv}) + e_1(t), \quad (7.35)$$

$$\partial_t s_\perp = \sigma_1 a_\perp (\varphi_1 - \sigma_1 \overline{bg}) - \sigma_1^2 a_\perp (\overline{ug} + \overline{bv} + \overline{uv}) + e_2(t). \quad (7.36)$$

Owing to Proposition A.1, we know that $\|g\| \leq D\sqrt{\varepsilon}\beta$ and $\|b\| \leq D\beta$, so that

$$|\overline{ug} + \overline{bv}| \leq D\beta (\|u\|\sqrt{\varepsilon} + \|v\|), \quad (7.37)$$

and, letting $Z = \|u\| + \varepsilon^{-1/2}\|v\|$,

$$|\overline{ug} + \overline{bv}| \leq D\beta\sqrt{\varepsilon}Z. \quad (7.38)$$

Now,

$$\partial_t(u + \varepsilon^{-1/2}v) = \frac{1}{\varepsilon} \sigma_1 v (\varphi_1 - \sigma_1 \overline{bg}) - \frac{1}{\varepsilon} \sigma_1^2 s_\perp (\overline{ug} + \overline{bv} + \overline{uv}) + e_1(t) \quad (7.39)$$

$$+ \varepsilon^{-1/2} (\sigma_1 u (\varphi_1 - \sigma_1 \overline{bg}) - \sigma_1^2 a_\perp (\overline{ug} + \overline{bv} + \overline{uv}) + e_2(t)). \quad (7.40)$$

or equivalently

$$\partial_t(u + \varepsilon^{-1/2}v) \leq \frac{\sigma_1}{\sqrt{\varepsilon}}(u + \varepsilon^{-1/2}v)(\varphi_1 - \sigma_1 \overline{bg}) - \sigma_1^2 \left(\frac{1}{\sqrt{\varepsilon}}a_\perp + \frac{1}{\varepsilon}s_\perp \right) (\overline{ug} + \overline{bv} + \overline{uv}) + e_1(t) + \varepsilon^{-1/2}e_2(t). \quad (7.41)$$

Decomposing $a_\perp, s_\perp$ further (to eliminate the constant C in leading order terms), we note that $\|a_\perp\| \leq \|b\| + \|u\|$ and $\|s_\perp\| \leq \|g\| + \|v\|$. Using the bound Eq.(7.38), along with the fact that $\frac{\|b\|}{\sqrt{\varepsilon}} + \frac{\|g\|}{\varepsilon} \leq \frac{D\beta}{\sqrt{\varepsilon}}$, we reach

$$\partial_t Z \leq \frac{\sigma_1}{\sqrt{\varepsilon}} Z (\varphi_1 - \sigma_1 \overline{bg}) + \sigma_1^2 D\beta^2 Z + \sigma_1^2 D\beta\sqrt{\varepsilon}Z \left(\frac{\|u\|}{\sqrt{\varepsilon}} + \frac{\|v\|}{\varepsilon} \right) \quad (7.42)$$

$$+ \sigma_1^2 \varepsilon^{-1/2} C\beta(\|u\|\|v\|) + e_1(t) + \varepsilon^{-1/2}e_2(t). \quad (7.43)$$

Using the threshold on u, v defined in T , and the bounds obtained on $e_1(t), e_2(t)$, we find

$$\partial_t Z \leq \frac{\sigma_1}{\sqrt{\varepsilon}} (\varphi_1 - \sigma_1 \overline{bg}) Z + \sigma_1^2 D\beta^2 Z + \sigma_1^2 D\beta^2 Z \varepsilon^\eta + \sigma_1^2 \varepsilon^{2\eta} C\beta^3 + C^2 D\beta^2. \quad (7.44)$$

Thus, as long as $\eta \geq \frac{1}{8}$, we find

$$\partial_t Z \leq \frac{\sigma_1}{\sqrt{\varepsilon}} (\varphi_1 - \sigma_1 \overline{bg}) Z + D\beta^2 Z + C^2 D\beta^2. \quad (7.45)$$

Letting $t_\varepsilon^\circ = \frac{1}{4\sigma_1\varphi_1} \sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$, we begin by considering the interval $[0, t_\varepsilon^\circ]$, for which it will be sufficient to study the differential inequality

$$\partial_t Z \leq \frac{\sigma_1\varphi_1}{\sqrt{\varepsilon}} Z + D\beta^2 Z + C^2 D\beta^2. \quad (7.46)$$

Defining

$$\lambda_\varepsilon(t) = \int_0^t \left(\frac{\sigma_1\varphi_1}{\sqrt{\varepsilon}} + D\beta^2(u) \right) du, \quad (7.47)$$

we thus have

$$Z(t) \leq \exp(\lambda_\varepsilon(t)) \int_0^t \exp(-\lambda_\varepsilon(u)) C^2 D\beta^2(u) du. \quad (7.48)$$

Then,

$$\lambda_\varepsilon(t) - \frac{\sigma_1\varphi_1}{\sqrt{\varepsilon}} t \leq D \int_0^t \exp\left(\frac{2\sigma_1\varphi_1 u}{\sqrt{\varepsilon}}\right) du \quad (7.49)$$

$$\leq D\sqrt{\varepsilon} \exp\left(\frac{2\sigma_1\varphi_1 t}{\sqrt{\varepsilon}}\right). \quad (7.50)$$

Now,

$$\int_0^t \exp(-\lambda_\varepsilon(u)) \beta^2(u) du \leq \int_0^t \exp(-\lambda_\varepsilon(u)) \exp\left(\frac{2\sigma_1\varphi_1 u}{\sqrt{\varepsilon}}\right) du \quad (7.51)$$

$$\leq \int_0^t \exp\left(\frac{\sigma_1\varphi_1 u}{\sqrt{\varepsilon}}\right) du \quad (7.52)$$

$$\leq D\sqrt{\varepsilon} \exp\left(\frac{\sigma_1\varphi_1 t}{\sqrt{\varepsilon}}\right). \quad (7.53)$$

We thus reach

$$Z(t) \leq C^2 D\sqrt{\varepsilon} \exp\left(\frac{2\sigma_1\varphi_1}{\sqrt{\varepsilon}}t + D\sqrt{\varepsilon} \exp\left(\frac{2\sigma_1\varphi_1}{\sqrt{\varepsilon}}t\right)\right). \quad (7.54)$$

Noting that, on the considered time interval, $\frac{1}{D} \exp\left(\frac{\sigma_1\varphi_1 t}{\sqrt{\varepsilon}}\right) \leq \beta(t) \leq D \exp\left(\frac{\sigma_1\varphi_1 t}{\sqrt{\varepsilon}}\right)$, we finally reach

$$Z(t) \leq C^2 D\beta(t)\sqrt{\varepsilon} \exp\left(\frac{\sigma_1\varphi_1 t}{\sqrt{\varepsilon}}\right) \quad (7.55)$$

$$\leq C^2 D\varepsilon^{1/4} \beta(t). \quad (7.56)$$

Thus, as long as $\eta \in [\frac{1}{8}, \frac{1}{4}]$, the proof is complete on this time interval. We now turn to the time interval $[t_\varepsilon^\circ, \tilde{t}_\varepsilon^\circ]$ for some $\tilde{t}_\varepsilon^\circ \geq t_\varepsilon^\circ$ to be set later. Going back to the differential inequality Eq.(7.45), we introduce

$$\Gamma_\varepsilon(t) = \int_0^t \left(\frac{\sigma_1}{\sqrt{\varepsilon}}(\varphi_1 - \sigma_1 \overline{bg})(u) + D\beta^2(u) \right) du, \quad (7.57)$$

so that

$$Z(t) \leq \exp(\Gamma_\varepsilon(t)) \int_0^t \exp(-\Gamma_\varepsilon(u)) C^2 D\beta^2(u) du. \quad (7.58)$$

Beginning with an upper bound on $\Gamma_\varepsilon(t)$, we write

$$\Gamma_\varepsilon(t) = \int_0^{t_\varepsilon^\circ} \left(\frac{\sigma_1}{\sqrt{\varepsilon}}(\varphi_1 - \sigma_1 \overline{bg})(u) + D\beta^2(u) \right) du + \int_{t_\varepsilon^\circ}^t \left(\frac{\sigma_1}{\sqrt{\varepsilon}}(\varphi_1 - \sigma_1 \overline{bg})(u) + D\beta^2(u) \right) du \quad (7.59)$$

$$= \Gamma_\varepsilon(t_\varepsilon^\circ) + \Gamma_\varepsilon(t) - \Gamma_\varepsilon(t_\varepsilon^\circ). \quad (7.60)$$

Using the results proven for the interval $[0, t_\varepsilon^\circ]$, we have

$$\Gamma_\varepsilon(t_\varepsilon^\circ) \leq \frac{1}{4} \log\left(\frac{1}{\varepsilon}\right) + D. \quad (7.61)$$

Then, note that for any $t \in [t_\varepsilon^\circ, \infty)$, $\beta(t) \leq D\varepsilon^{-1/4}$. Furthermore, Corollary A.2 gives that

$$0 \leq \varphi_1 - \sigma_1 \overline{bg} \leq \frac{D}{\sqrt{\varepsilon}} \exp\left(-\frac{2\sigma_1\varphi_1}{\sqrt{\varepsilon}}t\right) + D\varepsilon. \quad (7.62)$$

Thus,

$$\Gamma_\varepsilon(t) - \Gamma_\varepsilon(t_\varepsilon^\circ) \leq \frac{D}{\varepsilon} \int_{t_\varepsilon^\circ}^t \exp\left(-\frac{2\sigma_1\varphi_1}{\sqrt{\varepsilon}}u\right) du + \frac{D}{\sqrt{\varepsilon}}(t - t_\varepsilon^\circ) \quad (7.63)$$

$$\leq D + \frac{D}{\sqrt{\varepsilon}}(t - t_\varepsilon^\circ), \quad (7.64)$$

and we reach the upper bound

$$\exp(\Gamma_\varepsilon(t)) \leq D\varepsilon^{-1/4} \exp\left(\frac{D}{\sqrt{\varepsilon}}(t - t_\varepsilon^\circ)\right). \quad (7.65)$$

Turning to the integral term in Eq.(7.58), we begin with the term

$$\int_0^{t_\varepsilon^\circ} \exp(-\Gamma_\varepsilon(u)) C^2 D\beta^2(u) du \leq C^2 D \int_0^{t_\varepsilon^\circ} \exp\left(\int_0^u \frac{\sigma_1}{\sqrt{\varepsilon}}(\varphi_1 + \sigma_1 \overline{bg})(r) dr\right) du \quad (7.66)$$

$$\leq C^2 D \int_0^{t_\varepsilon^\circ} \exp\left(\frac{\sigma_1\varphi_1}{\sqrt{\varepsilon}}u + \frac{\sigma_1\sigma'_1}{\sqrt{\varepsilon}} \int_0^u \beta(r)\gamma(r) dr\right) du \quad (7.67)$$

$$\leq C^2 D \int_0^{t_\varepsilon^\circ} \exp\left(\frac{\sigma_1\varphi_1}{\sqrt{\varepsilon}}u + \frac{\sigma_1\sigma'_1}{4} \int_0^u \exp\left(\frac{2\sigma_1\varphi_1}{\sqrt{\varepsilon}}r\right) dr\right) du, \quad (7.68)$$

where we used Corollary A.2 in the third line. Since $u \leq t_\varepsilon^\circ$, it holds that $\int_0^u \exp\left(\frac{2\sigma_1\varphi_1}{\sqrt{\varepsilon}}r\right) dr \leq D$, and we reach

$$\int_0^{t_\varepsilon^\circ} \exp(-\Gamma_\varepsilon(u)) C^2 D \beta^2(u) du \leq C^2 D \varepsilon^{1/4}. \quad (7.69)$$

Continuing with the term $\int_{t_\varepsilon^\circ}^t \exp(-\Gamma_\varepsilon(u)) C^2 D \beta^2(u) du$, we require a lower bound on $\Gamma_\varepsilon(u)$, for $t \in [t_\varepsilon^\circ, \infty)$. We then write

$$\Gamma_\varepsilon(t_\varepsilon^\circ) \geq \frac{\sigma_1\varphi_1}{\sqrt{\varepsilon}} t_\varepsilon^\circ - \frac{\sigma_1}{\sqrt{\varepsilon}} \sigma_1' \int_0^{t_\varepsilon^\circ} \beta(u) \gamma(u) du \quad (7.70)$$

$$\geq \frac{1}{4} \log\left(\frac{1}{\varepsilon}\right) - D \int_0^{t_\varepsilon^\circ} \exp\left(\frac{2\sigma_1\varphi_1}{\sqrt{\varepsilon}}u\right) du \quad (7.71)$$

$$\geq \frac{1}{4} \log\left(\frac{1}{\varepsilon}\right) - D, \quad (7.72)$$

where we used Corollary A.2 again in the second line. Then, noting that, for any $t \in [t_\varepsilon^\circ, \infty)$, $\beta(t) \geq \frac{1}{D} \varepsilon^{-1/4}$, we may write

$$\Gamma_\varepsilon(t) - \Gamma_\varepsilon(t_\varepsilon^\circ) \geq \frac{1}{D} \int_{t_\varepsilon^\circ}^t \beta^2(r) dr \quad (7.73)$$

$$\geq \frac{1}{D} \varepsilon^{-1/2} (t - t_\varepsilon^\circ). \quad (7.74)$$

We thus reach

$$\int_{t_\varepsilon^\circ}^t \exp(-\Gamma_\varepsilon(u)) C^2 D \beta^2(u) du \leq C^2 D \varepsilon^{-1/2} \exp(-\Gamma_\varepsilon(t_\varepsilon^\circ)) \int_{t_\varepsilon^\circ}^t \exp\left(-\frac{1}{D\sqrt{\varepsilon}}(u - t_\varepsilon^\circ)\right) du \quad (7.75)$$

$$\leq C^2 D \varepsilon^{1/4}. \quad (7.76)$$

Combining the bounds obtained at Eq.(7.65),(7.69),(7.75), and recalling that, for $t \in [t_\varepsilon^\circ, \infty)$, $\beta(t) \geq \frac{1}{D} \varepsilon^{-1/4}$, we obtain

$$Z(t) \leq C^2 D \beta(t) \varepsilon^{1/4} \exp\left(\frac{D}{\sqrt{\varepsilon}}(t - t_\varepsilon^\circ)\right). \quad (7.77)$$

We now set $\tilde{t}_\varepsilon^\circ = t_\varepsilon^\circ + c\sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$ for some small constant c to be determined in such a way that the above bound becomes

$$Z(t) \leq C^2 D \beta(t) \varepsilon^{1/4 - cD}. \quad (7.78)$$

Thus, as long as $cD \leq \frac{1}{8}$, the proof is complete for the time interval $[0, \tilde{t}_\varepsilon^\circ]$. $\square$

We now turn to a stability result regarding the approximation guarantees established in Theorem 7.1. A core limitation in the proof of the latter comes from the term $D\beta^2 Z$ in the differential inequality (7.45), which comes from controlling $|\bar{a}\bar{s} - \bar{b}\bar{g}|$. This term grows unbounded when, in fact, the difference $|\bar{a}\bar{s} - \bar{b}\bar{g}|$ becomes arbitrarily small at times of order $K\sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$, for K sufficiently large.

Theorem 7.2. *Let $t = t_\varepsilon^\circ = \frac{1}{4\sigma_1\varphi_1} \sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$. There exists a positive constant δ depending only on the problem data such that, for any $K \geq \frac{1}{4\sigma_1\varphi_1}$ independent of ε and time horizon $t_\varepsilon = K\sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$, there exists a constant C depending only on the problem data and K such that, for ε sufficiently small and any $t \in [t_\varepsilon^\circ, t_\varepsilon]$,*

$$|\bar{a} - \bar{b}|(t) \leq Ct + C\varepsilon\beta^2(t), \quad |\bar{s}(t)| \leq Ct, \quad (7.79)$$

$$\|(a_\perp - b)(t, \cdot)\| + \varepsilon^{-1/2} \|(s_\perp - g)(t, \cdot)\| \leq C\varepsilon^\delta \beta(t). \quad (7.80)$$

Proof. Using the notations of Theorem 7.1, we consider the hitting time

$$T := \inf \left\{ t \geq 0 : \|a\| + \varepsilon^{-1/2} \|s\| \geq C\beta \quad \text{or} \quad |\varphi_1 - \sigma_1 \bar{a}\bar{s}| \geq C \right\}, \quad (7.81)$$

for some constant C to be set later. First, note that the argument used to establish the inequalities

$$|\bar{a} - \bar{b}|(t) \leq Ct + C\varepsilon\beta^2(t), \quad (7.82)$$

$$|\bar{s}(t)| \leq Ct, \quad (7.83)$$

in the proof of Theorem 7.1 is valid for any $0 \leq t \leq T$. Thus, for any $0 \leq t \leq t_\varepsilon \wedge T$, we find

$$\left| \frac{\varphi_0}{\sigma_0} - \bar{a}(t) \right| \leq KC\sqrt{\varepsilon} \log\left(\frac{1}{\varepsilon}\right) + C\sqrt{\varepsilon}. \quad (7.84)$$

We then write

$$a\partial_t s = a^2(1-s^2) \sum_{k=1}^{\infty} k\sigma_k s^{k-1} \left(\varphi_k - \sigma_k \overline{as^k} \right) \quad (7.85)$$

$$s\partial_t a = \frac{1}{\varepsilon} \sum_{k=0}^{\infty} \sigma_k s^{k+1} \left(\varphi_k - \sigma_k \overline{as^k} \right), \quad (7.86)$$

so that

$$\left| \partial_t \overline{as} - \frac{1}{\varepsilon} \sigma_0 \bar{s} (\varphi_0 - \sigma_0 \bar{a}) - \left(\overline{a^2} + \frac{1}{\varepsilon} \overline{s^2} \right) \sigma_1 (\varphi_1 - \sigma_1 \overline{as}) \right| \leq C^3 D \sqrt{\varepsilon} \beta^3, \quad (7.87)$$

and thus

$$\left| \partial_t \overline{as} - \left(\overline{a^2} + \frac{1}{\varepsilon} \overline{s^2} \right) \sigma_1 (\varphi_1 - \sigma_1 \overline{as}) \right| \leq C^3 D \sqrt{\varepsilon} \beta^3 + DKC^2 \log^2\left(\frac{1}{\varepsilon}\right). \quad (7.88)$$

Recalling that there exists a constant D depending only on the problem data such that, for any $t \geq 0$, $\beta(t) \leq D\varepsilon^{-1/4}$, we find

$$\left| \partial_t (\varphi_1 - \sigma_1 \overline{as})^2 + \sigma_1^2 \left(\overline{a^2} + \frac{1}{\varepsilon} \overline{s^2} \right) (\varphi_1 - \sigma_1 \overline{as})^2 \right| \leq C^3 D \varepsilon^{-1/4} |\varphi_1 - \sigma_1 \overline{as}|. \quad (7.89)$$

Allowing the exact value of D to change from line to line in what follows, recall from Proposition A.1 that $\beta(t_\varepsilon^\circ) \geq \frac{1}{D}\varepsilon^{-1/4}$. This, along with Theorem 7.1, gives the existence of a (strictly positive) constant c such that, for any t in the interval $[t_\varepsilon^\circ, t_\varepsilon^\circ + c\sqrt{\varepsilon} \log(\frac{1}{\varepsilon})]$ and ε sufficiently small,

$$\overline{a^2} + \frac{1}{\varepsilon} \overline{s^2} \geq \frac{1}{D} \varepsilon^{-1/2}, \quad (7.90)$$

giving, on the same time interval,

$$\partial_t (\varphi_1 - \sigma_1 \overline{as})^2 \leq -\frac{1}{D\sqrt{\varepsilon}} (\varphi_1 - \sigma_1 \overline{as})^2 + C^3 D \varepsilon^{-1/4} |\varphi_1 - \sigma_1 \overline{as}|. \quad (7.91)$$

Lemma C.2 started at $t = t_\varepsilon^\circ$ then gives

$$|\varphi_1 - \sigma_1 \overline{as}| \leq \varphi_1 \exp\left(-\frac{(t - t_\varepsilon^\circ)}{D\sqrt{\varepsilon}}\right) + C^3 D^2 \varepsilon^{1/4} \left(1 - \exp\left(-\frac{(t - t_\varepsilon^\circ)}{D\sqrt{\varepsilon}}\right)\right). \quad (7.92)$$

Now, note that Eq.(7.22),(7.23) and the bounds at (7.30)-(7.32) are valid for any $0 \leq t \leq T$, so that we may write

$$\partial_t a_\perp = \frac{1}{\varepsilon} \sigma_1 s_\perp (\varphi_1 - \sigma_1 \overline{as}) + e_1(t) \quad (7.93)$$

$$\partial_t s_\perp = \sigma_1 a_\perp (\varphi_1 - \sigma_1 \overline{as}) + e_2(t), \quad (7.94)$$

for some continuous functions $e_1(t), e_2(t)$ which verify $\sup_{t \leq T \wedge t_\varepsilon} |e_1(t)| + \varepsilon^{-1/2} |e_2(t)| \leq C^2 D \beta^2$. Then, adding and subtracting $\bar{b}g$,

$$\partial_t a_\perp = \frac{\sigma_1}{\varepsilon} s_\perp (\varphi_1 - \sigma_1 \bar{b}g) - \frac{\sigma_1^2}{\varepsilon} s_\perp (\overline{as} - \bar{b}g) + e_1(t), \quad (7.95)$$

$$\partial_t s_\perp = \sigma_1 a_\perp (\varphi_1 - \sigma_1 \bar{b}g) - \sigma_1^2 a_\perp (\overline{as} - \bar{b}g) + e_2(t). \quad (7.96)$$

Recalling the definition $Z = \|u\| + \varepsilon^{-1/2} \|v\|$, with $u = a_\perp - b$ and $v = s_\perp - g$, we have, for any $0 \leq t \leq T$,

$$\partial_t Z \leq \frac{\sigma_1}{\sqrt{\varepsilon}} ((\varphi_1 - \sigma_1 \bar{b}g) + \sigma_1 |\overline{as} - \bar{b}g|) Z + \frac{D}{\sqrt{\varepsilon}} \beta |\overline{as} - \bar{b}g| + e_1 + \varepsilon^{-1/2} e_2. \quad (7.97)$$

At this point, we have established that the bounds (7.92) and (7.97) are valid on the time interval $[t_\varepsilon^\circ, t_\varepsilon^\circ + c\sqrt{\varepsilon} \log(\frac{1}{\varepsilon})]$. Thus, at time $\bar{t}_\varepsilon^\circ = t_\varepsilon^\circ + c\sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$, the following estimate holds:

$$|\varphi_1 - \sigma_1 \overline{as}| \leq \varphi_1 \varepsilon^{c/D} + C^3 D \varepsilon^{1/4}, \quad (7.98)$$

and, recalling that Corollary A.2 implies the bound $0 \leq \varphi_1 - \sigma_1 \bar{b}g \leq \frac{D}{\sqrt{\varepsilon}} \exp(-\frac{2\sigma_1 \varphi_1}{\sqrt{\varepsilon}} t) + D\varepsilon$, we find that

$$|\varphi_1 - \sigma_1 \bar{b}g| \leq D\varepsilon^{2\sigma_1 \varphi_1 c \wedge 1}. \quad (7.99)$$

We conclude that there exists an exponent $\delta > 0$, depending only on the problem data, such that, at $t = \bar{t}_\varepsilon^\circ$,

$$|\bar{a}s - \bar{b}g| \leq D\varepsilon^\delta. \quad (7.100)$$

Let us now introduce the hitting time

$$T' = \inf \left\{ t \geq \bar{t}_\varepsilon^\circ : \bar{a}^2(t) + \frac{1}{\varepsilon} \bar{s}^2(t) \leq \frac{1}{D} \varepsilon^{-1/2} \right\}. \quad (7.101)$$

Clearly $T' > \bar{t}_\varepsilon^\circ$, and the inequality (7.92) remains valid on the time interval $[\bar{t}_\varepsilon^\circ, T' \wedge T \wedge t_\varepsilon]$. Noting that the r.h.s. of the inequality (7.92) is decreasing for any $t \geq \bar{t}_\varepsilon^\circ$, we find that, for any t in the time interval $[\bar{t}_\varepsilon^\circ, T' \wedge T \wedge t_\varepsilon]$ and ε sufficiently small,

$$\partial_t Z \leq D\varepsilon^{-1/2+\delta} Z + \frac{D}{\varepsilon^{3/4}} \left(\varphi_1 \exp\left(-\frac{t}{D\sqrt{\varepsilon}}\right) + C^3 D^2 \varepsilon^{1/4} \right). \quad (7.102)$$

Integrating the above, we find that, for any $t \in [\bar{t}_\varepsilon^\circ, T' \wedge T \wedge t_\varepsilon]$,

$$Z(t) \leq Z(\bar{t}_\varepsilon^\circ) \exp\left(\varepsilon^{-1/2+\delta}(t - \bar{t}_\varepsilon^\circ)\right) + \frac{D}{\varepsilon^{3/4}} \int_{\bar{t}_\varepsilon^\circ}^t \exp\left(\varepsilon^{-1/2+\delta}(t - s)\right) \left(D \exp\left(-\frac{s}{D\sqrt{\varepsilon}}\right) + C^3 D^2 \varepsilon^{1/4} \right) ds,$$

and, recalling from Theorem 7.1 that $Z(\bar{t}_\varepsilon^\circ) \leq C\varepsilon^{1/8} \beta(\bar{t}_\varepsilon^\circ)$,

$$\begin{aligned} Z(t) &\leq C\varepsilon^{1/8} \beta(\bar{t}_\varepsilon^\circ) \exp\left(\frac{\varepsilon^\delta}{\sqrt{\varepsilon}} t\right) + \frac{D}{\varepsilon^{3/4}} \exp\left(\frac{\varepsilon^\delta}{\sqrt{\varepsilon}} t\right) \left(D^2 \sqrt{\varepsilon} \left[\exp\left(-\frac{\bar{t}_\varepsilon^\circ}{D\sqrt{\varepsilon}}\right) - \exp\left(-\frac{t}{D\sqrt{\varepsilon}}\right) \right] \right) \\ &\quad + \frac{C^3 D^3}{\varepsilon^{1/2}} \exp\left(\frac{\varepsilon^\delta}{\sqrt{\varepsilon}} t\right) \left(\frac{\sqrt{\varepsilon}}{\varepsilon^\delta} \left[\exp\left(-\frac{\varepsilon^\delta \bar{t}_\varepsilon^\circ}{\sqrt{\varepsilon}}\right) - \exp\left(-\frac{\varepsilon^\delta t}{\sqrt{\varepsilon}}\right) \right] \right). \end{aligned}$$

Thus, for any $K > 0$ independent of ε and ε sufficiently small, we find

$$Z(t_\varepsilon) \leq C\varepsilon^{1/8} \beta(\bar{t}_\varepsilon^\circ) + D\varepsilon^{-1/4+\delta} + C^3 D^3 \varepsilon^{-\delta}, \quad (7.103)$$

$$\leq D\varepsilon^\delta \beta(t_\varepsilon), \quad (7.104)$$

where we have used that $\frac{1}{D}\varepsilon^{-1/4} \leq \beta(t) \leq D\varepsilon^{-1/4}$ for any $t \in [t_\varepsilon^\circ, \infty)$. From this bound, we deduce that the desired bounds hold for any t in $[\bar{t}_\varepsilon^\circ, T' \wedge T \wedge t_\varepsilon]$, and, by definition of Z , that $T \wedge T' > K\sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$, concluding the proof. $\square$

As a corollary, we obtain the following result on the stability of the learned components φ_0, φ_1 , which also removes the $\log(\frac{1}{\varepsilon})$ factor in the bound (7.84).

Corollary 7.3. *There exist constants C, δ and $K^\circ > \frac{1}{4\sigma_1 \varphi_1}$ independent of ε , such that, for any $K > K^\circ$, $t_\varepsilon = K\sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$, $t \in [K^\circ \sqrt{\varepsilon} \log(\frac{1}{\varepsilon}), t_\varepsilon]$ and ε sufficiently small,*

$$|\varphi_0 - \sigma_0 \bar{a}(t)| \leq C\sqrt{\varepsilon}, \quad |\varphi_1 - \sigma_1 \bar{a}s(t)| \leq C\varepsilon^{1/4}, \quad |\bar{s}(t)| \leq Ct, \quad (7.105)$$

$$\|a_\perp(t, \cdot) - \left(\frac{\varphi_1}{\sigma_1'}\right)^{1/2} \varepsilon^{-1/4} a_{init, \perp}(\cdot)\| + \varepsilon^{-1/2} \|s_\perp(t, \cdot) - \left(\frac{\varphi_1}{\sigma_1'}\right)^{1/2} \varepsilon^{1/4} a_{init, \perp}(\cdot)\| \leq C\varepsilon^{-1/4+\delta}. \quad (7.106)$$

Proof. With T the same stopping time as the one introduced in the proof of Theorem 7.1, we have the upper bound (7.92), valid for any $t_\varepsilon^\circ \leq t \leq T \wedge t_\varepsilon$. Thus, the part of the statement regarding $|\varphi_1 - \sigma_1 \bar{a}s|$ is true for some time $K_1 \sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$, with $\frac{1}{4\sigma_1 \varphi_1} < K_1 < K$. Now, for any $t \in [K_1 \sqrt{\varepsilon} \log(\frac{1}{\varepsilon}), t_\varepsilon]$, $|\varphi_1 - \sigma_1 \bar{a}s| \leq C\varepsilon^{1/4}$ and $|\bar{s}(t)| \leq CK_1 \sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$, we find

$$|\partial_t \bar{a} - \frac{\sigma_0}{\varepsilon} (\varphi_0 - \sigma_0 \bar{a}(t))| \leq CDK_1 \varepsilon^{-1/4} \log(\frac{1}{\varepsilon}) + C^2 D \varepsilon^{-1/2}, \quad (7.107)$$

so that, for ε sufficiently small,

$$\partial_t (\varphi_0 - \sigma_0 \bar{a})^2 \leq -\frac{\sigma_0}{\varepsilon} (\varphi_0 - \sigma_0 \bar{a})^2 + \frac{C^2 D}{\sqrt{\varepsilon}} |\varphi_0 - \sigma_0 \bar{a}|. \quad (7.108)$$

Lemma C.2 started at $t = K_1\sqrt{\varepsilon}\log(\frac{1}{\varepsilon})$ then gives

$$|\varphi_0 - \sigma_0\bar{a}(t)| \leq C\sqrt{\varepsilon}\log(\frac{1}{\varepsilon}) \exp\left(-\frac{\sigma_0}{2\varepsilon}\left(t - K_1\sqrt{\varepsilon}\log(\frac{1}{\varepsilon})\right)\right) + C^2D\sqrt{\varepsilon}\left(1 - \exp\left(-\frac{\sigma_0}{2\varepsilon}\left(t - K_1\sqrt{\varepsilon}\log(\frac{1}{\varepsilon})\right)\right)\right), \quad (7.109)$$

implying the result for $|\varphi_0 - \sigma_0\bar{a}|$, for any $K^\circ > K_1$ independent of ε . Finally, choosing any $K^\circ \geq \frac{1}{2\sigma_1\varphi_1}$, Corollary A.2 implies that there exists a constant D depending only on $\varphi_1, \sigma_1, \sigma'_1, M$ such that, for any $t \geq K^\circ\sqrt{\varepsilon}\log(\frac{1}{\varepsilon})$,

$$\|b(t, \cdot) - \left(\frac{\varphi_1}{\sigma'_1}\right)^{1/2} \varepsilon^{-1/4} a_{init, \perp}(\cdot)\| + \varepsilon^{-1/2} \|g(t, \cdot) - \left(\frac{\varphi_1}{\sigma'_1}\right)^{1/2} \varepsilon^{1/4} a_{init, \perp}(\cdot)\| \leq D\varepsilon^{1/4}, \quad (7.110)$$

implying the last part of the statement. $\square$

Theorem 3.3 is a direct consequence of Theorem 7.1 and 7.2. We now give the proofs of Proposition 3.1 and 3.2.

Proof of Proposition 3.1. We begin by noting that, for any $\eta \in (0, 1)$,

$$|\bar{b}(t) - \frac{\varphi_0}{\sigma_0}| \leq \varepsilon^\eta \iff \exp\left(\frac{\sigma_0^2}{\varepsilon}t\right) \geq \frac{|\bar{a}(0) - \frac{\varphi_0}{\sigma_0}|}{\varepsilon^\eta}, \quad (7.111)$$

so that

$$\inf\left\{t \geq 0 : |\bar{b}(t) - \frac{\varphi_0}{\sigma_0}| \leq \varepsilon^\eta\right\} \leq \frac{\left(\eta \log(\frac{1}{\varepsilon}) + \log(|\bar{a}(0) - \frac{\varphi_0}{\sigma_0}|)\right)_+}{\sigma_0^2} \varepsilon. \quad (7.112)$$

Letting t_0 denote the right-hand-side of the above inequality, this time horizon falls within the interval of validity of Theorem 7.1, so there exists a constant C depending only on the problem data such that

$$|\bar{a} - \bar{b}|(t_0) \leq Ct_0 + C\varepsilon\beta^2(t_0). \quad (7.113)$$

Using the bounds on $\beta(t)$ given at Proposition A.1 and keeping the leading order terms in ε concludes the proof. $\square$

Proof of Proposition 3.2. First note that, for all $t \geq 0$,

$$|\bar{a}\bar{s}(t) - \bar{b}g(t)| \leq |\bar{a}_\perp \bar{s}_\perp(t) - \bar{b}g(t)| + |\bar{a}(t)\bar{s}(t)|, \\ \leq \|a_\perp(t, \cdot)\| \|s_\perp(t, \cdot) - g(t, \cdot)\| + \|g(t, \cdot)\| \|a_\perp(t, \cdot) - b(t, \cdot)\| + |\bar{a}(t)| |\bar{s}(t)|. \quad (7.114)$$

Owing to Theorem 7.1 and 7.2, for any constant K and $0 \leq t \leq K\sqrt{\varepsilon}\log(\frac{1}{\varepsilon})$, for ε sufficiently small, there exists a constant C depending only on the problem data and K , and a constant $\delta \in (0, \frac{1}{4})$ depending on the problem data, such that

$$|\bar{a}(t)| \leq C, \quad |\bar{s}(t)| \leq C\sqrt{\varepsilon}\log(\frac{1}{\varepsilon}), \quad (7.115)$$

$$\|a_\perp(t, \cdot)\| \leq C\beta(t), \quad \|g(t, \cdot)\| \leq C\sqrt{\varepsilon}\beta(t), \quad (7.116)$$

$$\|s_\perp(t, \cdot) - g(t, \cdot)\| \leq C\varepsilon^{1/2} + \delta\beta(t), \quad \|a_\perp(t, \cdot) - b(t, \cdot)\| \leq C\varepsilon^\delta\beta(t). \quad (7.117)$$

In what follows, we will allow C to change from line to line (as long as it only depends on the problem data and K). Now, recall from Proposition A.1 that there exists a constant D depending only on the problem data such that, for all $t \geq 0$, $\beta(t) \leq D\varepsilon^{-1/4}$, so that we reach, for any $0 \leq t \leq K\sqrt{\varepsilon}\log(\frac{1}{\varepsilon})$,

$$|\bar{a}\bar{s}(t) - \bar{b}g(t)| \leq C\varepsilon^\delta. \quad (7.118)$$

Using Corollary A.2 at time $t_1 = \left(\frac{1}{4\sigma_1\varphi_1} + c_1\right)\sqrt{\varepsilon}\log(\frac{1}{\varepsilon})$, we find that

$$\left|\frac{\varphi_1}{\sigma_1} - \bar{b}g(t_1)\right| \leq C\varepsilon^{2\sigma_1\varphi_1c_1}. \quad (7.119)$$

We conclude that

$$|\bar{a}\bar{s}(t_1) - \frac{\varphi_1}{\sigma_1}| \leq C\varepsilon^{\delta \wedge 2\sigma_1\varphi_1c_1}, \quad (7.120)$$

giving the first part of Theorem 3.3. The second part is a direct consequence of Corollary 7.3. $\square$

7.2. LEARNING THE QUADRATIC COMPONENT

We now turn to the learning process for the term φ_2 , with the initial condition prescribed by Theorem 7.2 and Corollary 7.3. Recall that we study the following truncated version of the system (2.8)-(2.9):

$$\partial_t a(t, \omega) = \frac{\sigma_0}{\varepsilon} (\varphi_0 - \sigma_0 \bar{a}(t)) + \frac{\sigma_1}{\varepsilon} s(t, \omega) (\varphi_1 - \sigma_1 \bar{a}(t)) + \frac{\sigma_2 \varphi_2}{\varepsilon} s^2(t, \omega), \quad (7.121)$$

$$\partial_t s(t, \omega) = a(t, \omega) \sigma_1 (\varphi_1 - \sigma_1 \bar{a}(t)) + 2\sigma_2 \varphi_2 a(t, \omega) s(t, \omega). \quad (7.122)$$

Choosing our starting time at $t_0 = K^\circ \sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$, Corollary 7.3 implies that the first two terms in the r.h.s. of (7.121) are of the same order of magnitude as the third, and the first term in the r.h.s. of (7.122) is of the same order as the second. Thus, we cannot neglect them, and will use Proposition B.1 to determine the correct idealized system. To obtain sufficiently sharp bounds for the error terms inherent to the approximation result of Proposition B.1, we will need to make changes in time and scale. In turn, to better understand the timescale on which we need to control the solution to (7.121)-(7.122), we first study the following idealized dynamical system

$$\partial_t a_q(t, \omega) = \frac{\sigma_2 \varphi_2}{\varepsilon} s_q^2(t, \omega), \quad (7.123)$$

$$\partial_t s_q(t, \omega) = 2\sigma_2 \varphi_2 a_q(t, \omega) s_q(t, \omega), \quad (7.124)$$

with any initial condition $a_q(t_0, \cdot), s_q(t_0, \cdot)$ such that, for some constant C ,

$$|\varphi_0 - \sigma_0 \bar{a}_q(t_0)| \leq C\sqrt{\varepsilon}, \quad |\varphi_1 - \sigma_1 \bar{a}_q s_q(t_0)| \leq C\varepsilon^{1/4}, \quad |\bar{s}_q(t_0)| \leq C\sqrt{\varepsilon} \log(\frac{1}{\varepsilon}), \quad (7.125)$$

$$\|a_{q,\perp}(t_0, \cdot) - \left(\frac{\varphi_1}{\sigma'_1}\right)^{1/2} \varepsilon^{-1/4} a_{init,\perp}(\cdot)\| + \varepsilon^{-1/2} \|s_{q,\perp}(t_0, \cdot) - \left(\frac{\varphi_1}{\sigma'_1}\right)^{1/2} \varepsilon^{1/4} a_{init,\perp}(\cdot)\| \leq C\varepsilon^{-1/4+\delta}. \quad (7.126)$$

Note that the Cauchy-Lipschitz theorem ensures the existence of a time $T_q > t_0$ such that, for all $\omega \in \Omega$, there exists a continuously differentiable solution to (7.123)-(7.124) on $[t_0, T_q]$. The following result shows that a significant fraction of neurons change scale on a timescale of order $\varepsilon^{1/4}$.

Lemma 7.4. *Let $t_0 = K^\circ \sqrt{\varepsilon} \log(\frac{1}{\varepsilon})$, with K° as in Corollary 7.3. Then, for any $t_0 \leq t \leq T_q$, the solution to (7.123)-(7.124) with initial condition verifying (7.125)-(7.126) verifies*

$$a_q^2(t, \omega) - a_q(t_0, \omega)^2 = \frac{1}{2\varepsilon} (s_q^2(t, \omega) - s_q(t_0, \omega)^2), \quad (7.127)$$

Furthermore, there exists a constant K depending only on the problem data such that, for any $c \in (0, 1)$, ε sufficiently small and any ω in the set

$$\mathcal{U}(c) = \left\{ \omega : a_{init,\perp}(\omega) \geq \frac{K}{c} \varepsilon^{1/4} \right\}, \quad (7.128)$$

$a_q(t, \omega)$ verifies

$$\frac{a_q(t_0, \omega)}{1 - (1 - c)\sigma_2 \varphi_2 a_q(t_0, \omega)(t - t_0)} \leq a_q(t, \omega) \leq \frac{a_q(t_0, \omega)}{1 - (2 + c)\sigma_2 \varphi_2 a_q(t_0, \omega)(t - t_0)}. \quad (7.129)$$

Proof. For any $t_0 \leq t \leq T_q$,

$$\partial_t a_q(t, \omega)^2 = \frac{2\sigma_2 \varphi_2}{\varepsilon} a_q(t, \omega) s_q^2(t, \omega), \quad (7.130)$$

$$\partial_t s_q(t, \omega)^2 = 4\sigma_2 \varphi_2 a_q(t, \omega) s_q^2(t, \omega), \quad (7.131)$$

so that,

$$a_q(t, \omega)^2 = \frac{1}{2\varepsilon} s_q(t, \omega)^2 - \frac{1}{2\varepsilon} s_q(t_0, \omega)^2 + a_q(t_0, \omega)^2. \quad (7.132)$$

Now, recall that, owing to Theorem 7.2 and Corollary 7.3, there exist constants C, δ independent of ε such that

$$|\bar{a}_q(t_0) - \frac{\varphi_0}{\sigma_0}| \leq C\sqrt{\varepsilon}, \quad |\bar{s}_q(t_0)| \leq CK^\circ \sqrt{\varepsilon} \log(\frac{1}{\varepsilon}), \quad (7.133)$$

$$\|(a_{q,\perp} - b)(t_0, \cdot)\| + \varepsilon^{-1/2} \|(s_{q,\perp} - g)(t_0, \cdot)\| \leq C\varepsilon^\delta \beta(t_0). \quad (7.134)$$

To compare $s_q(t_0, \omega)$ and $a_q(t_0, \omega)$, we write

$$a_q(t_0, \omega) = b(t_0, \omega) + a_{q,\perp}(t_0, \omega) - b(t_0, \omega) + \overline{a_q}(t_0), \quad (7.135)$$

$$s_q(t_0, \omega) = g(t_0, \omega) + s_{q,\perp}(t_0, \omega) - g(t_0, \omega) + \overline{s_q}(t_0), \quad (7.136)$$

so that, allowing C to change from line to line,

$$\|a_q(t_0, \cdot) - \frac{\varphi_0}{\sigma_0} - \beta(t_0)a_{init,\perp}(\cdot)\| \leq C\varepsilon^\delta \beta(t_0) + C\sqrt{\varepsilon}, \quad (7.137)$$

$$\|s_q(t_0, \cdot) - \gamma(t_0)a_{init,\perp}(\cdot)\| \leq C\sqrt{\varepsilon} \log\left(\frac{1}{\varepsilon}\right) + C\varepsilon^{1/2+\delta} \beta(t_0). \quad (7.138)$$

Choosing any $K^\circ > \frac{1}{4\sigma_1\varphi_1}$, Proposition A.1 then gives that

$$\theta(t_0) = \frac{1}{2} \log\left(\frac{4\varphi_1}{\sigma_1'\sqrt{\varepsilon}} + o\left(\frac{1}{\sqrt{\varepsilon}}\right)\right), \quad (7.139)$$

so that there exists a constant D depending only on $\sigma_1, \varphi_1, \sigma_1'$ such that

$$|\beta(t_0) - \frac{1}{2}e^{\theta(t_0)}| \leq D\varepsilon^{1/4} \quad \text{and} \quad |\gamma(t_0) - \frac{\sqrt{\varepsilon}}{2}e^{\theta(t_0)}| \leq D\varepsilon^{3/4}. \quad (7.140)$$

Combining the bounds (7.137), (7.138) with the above, keeping the leading order in ε and absorbing the constant D in C , we find that

$$\|s_q(t_0, \cdot) - \sqrt{\varepsilon}a_q(t_0, \cdot)\| \leq C\varepsilon^{3/4}, \quad (7.141)$$

which, upon recalling Eq.(7.132), implies in turn

$$\|a_q(t, \cdot)^2 - \frac{1}{2\varepsilon}s_q(t, \cdot)^2 - \frac{1}{2}a_q(t_0, \cdot)^2\| \leq C\left\|\frac{s_q(t_0, \cdot)}{\sqrt{\varepsilon}} - a_q(t_0, \cdot)\right\|\left\|\frac{s_q(t_0, \cdot)}{\sqrt{\varepsilon}} + a_q(t_0, \cdot)\right\|, \quad (7.142)$$

$$\leq C. \quad (7.143)$$

Then, for any $\omega \in \Omega$, we may bound the evolution of $a_q(t, \omega)$ as follows

$$2\sigma_2\varphi_2a_q^2(t, \omega) - \sigma_2\varphi_2a_q(t_0, \omega)^2 - C \leq \partial_t a_q(t, \omega) \leq 2\sigma_2\varphi_2a_q^2(t, \omega) - \sigma_2\varphi_2a_q(t_0, \omega)^2 + C, \quad (7.144)$$

and, since $a_q(t, \omega)$ is increasing for any ω , $a_q(t, \omega)^2$ is increasing for any ω in the set

$$\mathcal{A} = \{\omega : a_q(t_0, \omega) \geq 0\}. \quad (7.145)$$

We may then write, for any $\omega \in \mathcal{A}$,

$$\sigma_2\varphi_2a_q^2(t, \omega) - C \leq \partial_t a_q(t, \omega) \leq 2\sigma_2\varphi_2a_q^2(t, \omega) + C. \quad (7.146)$$

Thus, for ε sufficiently small, the upper and lower bounds at (7.129) hold for any ω in the set

$$\mathcal{B} = \left\{\omega : a_q(t_0, \omega)^2 \geq \frac{C}{c\sigma_2\varphi_2}\right\}. \quad (7.147)$$

The bound (7.137) then ensures that $\mathcal{U}(c) \subset \mathcal{A} \cap \mathcal{B}$, concluding the proof. $\square$

Now, note that, to learn the coefficient φ_2 , we require the quantity $\overline{as^2}$ to be of order one, which, owing to Eq.(7.127), requires a_q to be of order $\varepsilon^{-1/3}$ and s_q of order $\varepsilon^{1/6}$. The bounds from Eq.(7.129) then show that the required time for a_q (or s_q) to make any significant change, e.g. reach a large multiple of its initial condition or the required scale $\varepsilon^{-1/3}$ will be of order $\varepsilon^{1/4}$. Hence, we would like to control the solution to (7.121)-(7.122) on a timescale of order $\varepsilon^{1/4}$. The following lemma shows that the appropriate idealized system to study is the one introduced in Section 3 at Eq.(3.17)-(3.18):

$$\partial_t z_1(t, \omega) = \sigma_0\alpha_0(t) + \sigma_1\alpha_1(t)z_2(t, \omega) + \sigma_2\varphi_2z_2^2(t, \omega), \quad (7.148)$$

$$\partial_t z_2(t, \omega) = z_1(t, \omega) (\sigma_1\alpha_1(t) + 2\sigma_2\varphi_2z_2(t, \omega)), \quad (7.149)$$

with

$$\alpha_0(t) = -\frac{1}{\sigma_0} \left(\sigma_1\alpha_1(t)\overline{z_2}(t) + \sigma_2\varphi_2\overline{z_2^2}(t) \right), \quad (7.150)$$

$$\alpha_1(t) = -\frac{2\sigma_2\varphi_2\overline{z_1^2z_2} + \sigma_2\varphi_2\overline{z_2^3} - \sigma_2\varphi_2\overline{z_2^2z_2}}{\sigma_1\overline{z_1^2} + \sigma_1(\overline{z_2^2} - \overline{z_2^2})}. \quad (7.151)$$

We also recall that, by definition of α_0, α_1 , any solution to the above verifies

$$\partial_t \overline{z_1}(t) = 0, \quad \text{and} \quad \partial_t \overline{z_1z_2}(t) = 0. \quad (7.152)$$

To obtain a sharper control on the error incurred by the application of Proposition B.1, we recall the following change in time and scale introduced in Section 3:

$$\hat{a}(t, \omega) = \varepsilon^{1/4} a(\varepsilon^{1/4} t, \omega), \quad (7.153)$$

$$\hat{s}(t, \omega) = \varepsilon^{-1/4} s(\varepsilon^{1/4} t, \omega), \quad (7.154)$$

which leads us to study the evolution of $\hat{a}, \hat{s}$ on a timescale of order one.

Lemma 7.5. *Let $(a(t, \cdot), s(t, \cdot))$ denote the solution to (7.121)-(7.122) with any initial condition $a(t_0, \cdot), s(t_0, \cdot)$ verifying the bounds (7.125)-(7.126). We then have the following results.*

- *There exist continuous functions $z_1(t_0, \cdot), z_2(t_0, \cdot)$ and constants C_1, C_2 such that*

$$\overline{z_1}(t_0) = \varepsilon^{1/4} \frac{\varphi_0}{\sigma_0}, \quad \overline{z_2}(t_0) = 0, \quad \overline{z_1 z_2}(t_0) = \frac{\varphi_1}{\sigma_1}, \quad (7.155)$$

$$\|\hat{a}(t_0, \cdot) - z_1(t_0, \cdot)\| \leq C_1 \varepsilon^{3/4}, \quad \|\hat{s}(t_0, \cdot) - z_2(t_0, \cdot)\| \leq C_2 \varepsilon^{1/4} \log\left(\frac{1}{\varepsilon}\right). \quad (7.156)$$

Furthermore, there exists a constant D depending only on the problem data such that

$$(\|\hat{a}(t_0, \cdot)\| + \|\hat{s}(t_0, \cdot)\|) \vee (\|z_1(t_0, \cdot)\| + \|z_2(t_0, \cdot)\|) \leq D. \quad (7.157)$$

- *Let $(z_1(t, \cdot), z_2(t, \cdot))$ be the solution to (7.148)-(7.149) with initial condition $(z_1(t_0, \cdot), z_2(t_0, \cdot))$. For any constant $C > D$, consider the hitting time*

$$T_1 = \inf\{t \geq t_0 : (\|\hat{a}(t, \cdot)\| + \|\hat{s}(t, \cdot)\|) \vee (\|z_1(t, \cdot)\| + \|z_2(t, \cdot)\|) \geq C\}. \quad (7.158)$$

Then there exist constants C_5, C_6 depending only on the problem data and C such that, for any fixed time horizon T independent of ε and any $t_0 \leq t \leq T_1 \wedge T$,

$$\|z_\varepsilon(t, \cdot) - z(t, \cdot)\| \leq \left(\|z_\varepsilon(t_0, \cdot) - z(t_0, \cdot)\| + C_5 \varepsilon^{1/8} (t - t_0) \right) \exp(C_6 (t - t_0)). \quad (7.159)$$

Proof. We begin by constructing an appropriate initial condition $z(t_0, \cdot) = (z_1(t_0, \cdot), z_2(t_0, \cdot))$. Recalling that $\hat{a}(t_0, \cdot) = \varepsilon^{1/4} a(t_0, \cdot)$ and $\hat{s}(t_0, \cdot) = \varepsilon^{-1/4} s(t_0, \cdot)$, we have

$$|\overline{\hat{a}}(t_0) - \varepsilon^{1/4} \frac{\varphi_0}{\sigma_0}| \leq C_2 \varepsilon^{3/4}, \quad (7.160)$$

$$|\overline{\hat{a}\hat{s}}(t_0) - \frac{\varphi_1}{\sigma_1}| \leq C_2 \varepsilon^{1/4}. \quad (7.161)$$

Then, we choose

$$z_1(t_0, \cdot) = \hat{a}_\perp(t_0, \cdot) + \varepsilon^{1/4} \frac{\varphi_0}{\sigma_0}, \quad (7.162)$$

$$z_2(t_0, \cdot) = \frac{\varphi_1}{\sigma_1 \overline{\hat{a}_\perp \hat{s}_\perp}(t_0)} \hat{s}_\perp(0, \cdot), \quad (7.163)$$

so that

$$\overline{z_1}(t_0) = \varepsilon^{1/4} \frac{\varphi_0}{\sigma_0}, \quad (7.164)$$

$$\overline{z_1 z_2}(t_0) = \frac{\varphi_1}{\sigma_1}, \quad (7.165)$$

and

$$\|z_1(t_0, \cdot) - \hat{a}(t_0, \cdot)\| \leq C_2 \varepsilon^{3/4}. \quad (7.166)$$

Also,

$$\|z_2(t_0, \cdot) - \hat{s}(t_0, \cdot)\| \leq \frac{|\varphi_1 - \sigma_1 \overline{\hat{a}\hat{s}}(t_0)|}{\sigma_1 \overline{\hat{a}\hat{s}}(t_0)} \|\hat{s}_\perp(t_0, \cdot)\| + |\overline{\hat{s}}(t_0)|, \quad (7.167)$$

which, upon recalling that there exists $\delta > 0$ independent of ε such that,

$$\|\hat{s}_\perp(t_0, \cdot) - \left(\frac{\varphi_1}{\sigma_1'}\right)^{1/2} a_{init, \perp}(\cdot)\| \leq C \varepsilon^\delta \quad \text{and} \quad |\overline{\hat{s}}(t_0)| \leq C \varepsilon^{1/4} \log\left(\frac{1}{\varepsilon}\right), \quad (7.168)$$

implies that there exists a constant C' such that

$$\|z_2(t_0, \cdot) - \hat{s}(t_0, \cdot)\| \leq C' \varepsilon^{1/4} \log\left(\frac{1}{\varepsilon}\right). \quad (7.169)$$

Now, note that, owing to Theorem 7.2, there exists a constant D depending only on $\sigma_1, \varphi_1, \sigma'_1, M$ such that, for ε sufficiently small,

$$\|a(t_0, \cdot)\| + \varepsilon^{-1/2} \|s(t_0, \cdot)\| \leq D\varepsilon^{-1/4}. \quad (7.170)$$

Let T_1 be the hitting time

$$T_1 = \inf\{t \geq t_0 : \|a(t, \cdot)\| + \varepsilon^{-1/2} \|s(t, \cdot)\| \geq C_1 \varepsilon^{-1/4}\}, \quad (7.171)$$

for any constant $C_1 > D$ (that does not depend on ε), and let T be a time horizon independent of ε . Then, for any $t_0 \leq t \leq T\varepsilon^{1/4} \wedge T_1$ consider the following change in time and scale in the dynamical system (7.121)-(7.122):

$$\hat{a}(t, \omega) = \varepsilon^{1/4} a(\varepsilon^{1/4} t, \omega), \quad (7.172)$$

$$\hat{s}(t, \omega) = \varepsilon^{-1/4} s(\varepsilon^{1/4} t, \omega). \quad (7.173)$$

Thus, we are now interested in controlling the evolution of the pair $\hat{a}, \hat{s}$ on a timescale of order one, i.e. independent of ε . The evolution (7.121)-(7.122) then becomes

$$\partial_t \hat{a}(t, \omega) = \frac{\sigma_0}{\sqrt{\varepsilon}} (\varphi_0 - \sigma_0 \bar{a}(t)) + \frac{\sigma_1}{\varepsilon^{1/4}} \hat{s}(t, \omega) (\varphi_1 - \sigma_1 \bar{a}\bar{s}(t)) + \sigma_2 \varphi_2 \hat{s}^2(t, \omega), \quad (7.174)$$

$$\partial_t \hat{s}(t, \omega) = \hat{a}(t, \omega) \frac{\sigma_1}{\varepsilon^{1/4}} (\varphi_1 - \sigma_1 \bar{a}\bar{s}(t)) + 2\sigma_2 \varphi_2 \hat{a}(t, \omega) \hat{s}(t, \omega), \quad (7.175)$$

and we may rewrite T_1 as

$$T_1 = \inf\{t \geq t_0 : \|\hat{a}(t, \cdot)\| + \|\hat{s}(t, \cdot)\| \geq C_1\}. \quad (7.176)$$

Now, an argument similar to the one employed to prove the upper bound Eq.(7.92) in the proof of Theorem 7.2 and Corollary 7.3 shows that there exist constants C_3, C_4 depending only on the problem data and C_1 such that, for any $t_0 \leq t \leq T_1$,

$$|\varphi_0 - \sigma_0 \bar{a}(t)| \leq C_3 \sqrt{\varepsilon}, \quad \text{and} \quad |\varphi_1 - \sigma_1 \bar{a}\bar{s}(t)| \leq C_4 \varepsilon^{1/4}. \quad (7.177)$$

We now note that, by construction,

$$\|z_1(t_0, \cdot)\| + \|z_2(t_0, \cdot)\| \leq D, \quad (7.178)$$

and introduce the hitting time

$$T_2 = \inf\{t \geq t_0 : \|z_1(t, \cdot)\| + \|z_2(t, \cdot)\| \geq C_1\}. \quad (7.179)$$

Letting

$$z(t, \omega) = (z_1(t, \omega), z_2(t, \omega)), \quad (7.180)$$

$$z_\varepsilon(t, \omega) = (\hat{a}(t, \omega), \hat{s}(t, \omega)), \quad (7.181)$$

we apply Corollary B.2 to obtain that there exist constants C_5, C_6 independent of ε such that, for any $t_0 \leq t \leq T_1 \wedge T_2 \wedge T$,

$$\|z_\varepsilon(t, \cdot) - z(t, \cdot)\|_2 \leq \left(\|z_\varepsilon(t_0, \cdot) - z(t_0, \cdot)\|_2 + C_5 \varepsilon^{1/8} (t - t_0) \right) \exp(C_6 (t - t_0)), \quad (7.182)$$

concluding the proof. Proposition 3.6 is an immediate consequence of Lemma 7.5. $\square$

APPENDIX A. AUXILIARY RESULTS ON IDEALIZED SYSTEMS

A.1. IDEALIZED SYSTEMS FOR THE CONSTANT AND LINEAR COMPONENTS

Consider the dynamical system

$$\partial_t b(t, \omega) = \frac{\sigma_1}{\varepsilon} g(t, \omega) (\varphi_1 - \sigma_1 \bar{b}\bar{g}(t)), \quad (A.1)$$

$$\partial_t g(t, \omega) = \sigma_1 b(t, \omega) (\varphi_1 - \sigma_1 \bar{b}\bar{g}(t)), \quad (A.2)$$

with initial condition $b(0, \cdot) = a_{init, \perp}(\cdot)$ and $g(0, \cdot) = 0$. For these dynamics, we have that b and g will remain proportional to $a_{init, \perp}(\cdot)$ for all times, so that $b(t, \cdot) = \beta(t) a_{init, \perp}(\cdot)$ and $g(t, \cdot) = \gamma(t) a_{init, \perp}(\cdot)$, where $\beta(t), \gamma(t)$ are the solutions to the dynamical system

$$\partial_t \beta(t) = \frac{\sigma_1}{\varepsilon} \gamma(t) (\varphi_1 - \sigma'_1 \beta(t) \gamma(t)), \quad (A.3)$$

$$\partial_t \gamma(t) = \sigma_1 \beta(t) (\varphi_1 - \sigma'_1 \beta(t) \gamma(t)), \quad (A.4)$$

where we have introduced $\sigma'_1 = \sigma_1 \int_0^1 a_{init,\perp}^2(\nu) d\nu$, and with $\beta(0) = 1$ and $\gamma(0) = 0$. The solution to the system (A.3)-(A.4) is characterized in the following proposition.

Proposition A.1. *For every $t \geq 0$, we have the representation*

$$\beta(t) = \cosh \theta(t), \quad \gamma(t) = \sqrt{\varepsilon} \sinh \theta(t), \quad (\text{A.5})$$

where $\theta(t)$ is the solution to the ODE

$$\partial_t \theta = \frac{\sigma_1 \varphi_1}{\sqrt{\varepsilon}} - \frac{\sigma_1 \sigma'_1}{2} \sinh 2\theta(t), \quad (\text{A.6})$$

with $\theta(0) = 0$. Furthermore, for every $t \geq 0$, we have the upper and lower bounds

$$l(t) \leq \theta(t) \leq u(t), \quad (\text{A.7})$$

where

$$l(t) = \frac{\sigma_1 \varphi_1}{\sqrt{\varepsilon}} t - \frac{1}{2} \log \left(1 + \frac{\sigma'_1 \sqrt{\varepsilon}}{4\varphi_1} \left(e^{\frac{2\sigma_1 \varphi_1}{\sqrt{\varepsilon}} t} - 1 \right) \right), \quad (\text{A.8})$$

$$u(t) = \left(\frac{\sigma_1 \varphi_1}{\sqrt{\varepsilon}} + \frac{\sigma_1 \sigma'_1}{4} \right) t - \frac{1}{2} \log \left(1 + \frac{\sigma'_1 \sqrt{\varepsilon}}{4\varphi_1} \left(\frac{1}{1 + \frac{\sigma'_1 \sqrt{\varepsilon}}{4\varphi_1}} \right) \left(e^{\left(\frac{2\sigma_1 \varphi_1}{\sqrt{\varepsilon}} + \frac{\sigma_1 \sigma'_1}{2} \right) t} - 1 \right) \right). \quad (\text{A.9})$$

Proof. A direct calculation using (A.3)-(A.4) shows that

$$\partial_t (\beta^2(t) - \varepsilon^{-1} \gamma^2(t)) = 2\beta(t) \partial_t \beta(t) - 2\varepsilon^{-1} \gamma(t) \partial_t \gamma(t) \quad (\text{A.10})$$

$$= 0, \quad (\text{A.11})$$

and, since $\beta(0) = 1$ and $\gamma(0) = 0$, we find that $\beta^2 - \varepsilon^{-1} \gamma^2 = 1$ for all times. This allows us to reparametrize the solution using a hyperbolic angle $\theta(t)$ such that $\beta(t) = \cosh \theta(t)$ and $\gamma(t) = \sqrt{\varepsilon} \sinh \theta(t)$. Substituting this into (A.3), we find

$$(\sinh \theta(t)) \partial_t \theta = \varepsilon^{-1} \sigma_1 (\sqrt{\varepsilon} \sinh \theta(t)) (\varphi_1 - \sigma'_1 \cosh \theta(t) \sqrt{\varepsilon} \sinh \theta(t)). \quad (\text{A.12})$$

Dividing by $\sinh(\theta)$ (for $t > 0$) and using the identity $2 \sinh \theta \cosh \theta = \sinh 2\theta$, we obtain the scalar ODE (A.6).

Now, the latter implies that, for all times,

$$\frac{\sigma_1 \varphi_1}{\sqrt{\varepsilon}} - \frac{\sigma_1 \sigma'_1}{4} e^{2\theta(t)} \leq \partial_t \theta(t) \leq \frac{\sigma_1 \varphi_1}{\sqrt{\varepsilon}} - \frac{\sigma_1 \sigma'_1}{4} (e^{2\theta(t)} - 1), \quad (\text{A.13})$$

so that

$$l(t) \leq \theta(t) \leq u(t), \quad (\text{A.14})$$

where l, u are respectively the solutions of

$$\partial_t l(t) = \frac{\sigma_1 \varphi_1}{\sqrt{\varepsilon}} - \frac{\sigma_1 \sigma'_1}{4} e^{2l(t)}, \quad l(0) = 0, \quad (\text{A.15})$$

$$\partial_t u(t) = \frac{\sigma_1 \varphi_1}{\sqrt{\varepsilon}} - \frac{\sigma_1 \sigma'_1}{4} (e^{2u(t)} - 1), \quad u(0) = 0. \quad (\text{A.16})$$

Focusing on $l(t)$, let us introduce $\tilde{l}(t) = l(t) - \frac{\sigma_1 \varphi_1}{\sqrt{\varepsilon}} t$, so that $\tilde{l}(t)$ solves

$$\partial_t \tilde{l}(t) = -\frac{\sigma_1 \sigma'_1}{4} e^{2\tilde{l}(t)} e^{\frac{2\sigma_1 \varphi_1}{\sqrt{\varepsilon}} t}, \quad \tilde{l}(0) = 0. \quad (\text{A.17})$$

To solve this ODE, we define

$$F(x) = - \int_0^x e^{-2y} dy \quad (\text{A.18})$$

$$= \frac{1}{2} (e^{-2x} - 1), \quad (\text{A.19})$$

$$G(t) = \int_0^t \frac{\sigma_1 \sigma'_1}{4} e^{\frac{2\sigma_1 \varphi_1}{\sqrt{\varepsilon}} s} ds \quad (\text{A.20})$$

$$= \frac{\sigma'_1 \sqrt{\varepsilon}}{8\varphi_1} \left(e^{\frac{2\sigma_1 \varphi_1}{\sqrt{\varepsilon}} t} - 1 \right), \quad (\text{A.21})$$

so that

$$\tilde{l}(t) = F^{-1}(G(t)) = -\frac{1}{2} \log \left(1 + \frac{\sigma'_1 \sqrt{\varepsilon}}{4\varphi_1} \left(e^{\frac{2\sigma_1 \varphi_1 t}{\sqrt{\varepsilon}}} - 1 \right) \right), \quad (\text{A.22})$$

yielding the expression for $l(t)$. A similar procedure applied to $\tilde{u}(t) = u(t) - \left(\frac{\sigma_1 \varphi_1}{\sqrt{\varepsilon}} + \frac{\sigma_1 \sigma'_1}{4} \right) t$ gives the expression for $u(t)$. $\square$

Corollary A.2. *For every $t \geq 0$, we have*

$$|\beta(t)\gamma(t)| \leq \frac{\sqrt{\varepsilon}}{4} \exp \left(\frac{2\sigma_1 \varphi_1 t}{\sqrt{\varepsilon}} \right), \quad (\text{A.23})$$

as well as

$$0 \leq \varphi_1 - \sigma'_1 \beta(t)\gamma(t) \leq \varphi_1 \left(\frac{4\varphi_1 - \sigma'_1 \sqrt{\varepsilon}}{4\varphi_1 + \sigma'_1 \sqrt{\varepsilon} \left(\exp \left(\frac{2\sigma_1 \varphi_1 t}{\sqrt{\varepsilon}} \right) - 1 \right)} \right) + \frac{\sigma'_1 \sqrt{\varepsilon}}{4} \left(\frac{\sigma'_1 \sqrt{\varepsilon}}{4\varphi_1} + \exp \left(-\frac{2\sigma_1 \varphi_1 t}{\sqrt{\varepsilon}} \right) \right). \quad (\text{A.24})$$

Proof. The first result is immediate from Eq.(A.6). For the second one, we write

$$\varphi_1 - \sigma'_1 \beta(t)\gamma(t) \leq \varphi_1 - \sigma'_1 \sqrt{\varepsilon} \cosh(l(t)) \sinh(l(t)) \quad (\text{A.25})$$

$$= \varphi_1 - \frac{\sigma'_1 \sqrt{\varepsilon}}{2} \sinh(2l(t)) \quad (\text{A.26})$$

$$\leq \varphi_1 - \frac{\sigma'_1 \sqrt{\varepsilon}}{4} \left(\frac{\exp \left(\frac{2\sigma_1 \varphi_1 t}{\sqrt{\varepsilon}} \right)}{1 + \frac{\sigma'_1 \sqrt{\varepsilon}}{4\varphi_1} \left(\exp \left(\frac{2\sigma_1 \varphi_1 t}{\sqrt{\varepsilon}} \right) - 1 \right)} - \frac{1 + \frac{\sigma'_1 \sqrt{\varepsilon}}{4\varphi_1} \left(\exp \left(\frac{2\sigma_1 \varphi_1 t}{\sqrt{\varepsilon}} \right) - 1 \right)}{\exp \left(\frac{2\sigma_1 \varphi_1 t}{\sqrt{\varepsilon}} \right)} \right), \quad (\text{A.27})$$

$$(\text{A.28})$$

implying the desired. $\square$

APPENDIX B. AN APPROXIMATION RESULT FOR FLOWS NEAR TARGET SETS

Let $\mathcal{O}$ be an open set in $\mathbb{R}^d$, m be a fixed positive integer, $\{h_i\}_{i=1}^m$ be a collection of twice continuously differentiable functions on $\mathcal{O}$ valued in $\mathbb{R}_+$, and $\{g_i\}_{i=1}^m, f$ be a collection of continuously differentiable vector fields on $\mathcal{O}$. In what follows, we let $\mathcal{S}$ denote the set

$$\mathcal{S} = \left\{ z \in \mathcal{C}(\Omega, \mathbb{R}^d) : \forall 1 \leq i \leq m, \int_{\Omega} h_i(z(\omega)) d\rho(\omega) = 0 \right\}. \quad (\text{B.1})$$

For each $1 \leq i \leq m$, let $v_i : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be a continuous, non-decreasing, positive function such that $v_i(0) = 0$. For any $\varepsilon > 0$, consider the (infinite dimensional) system of ordinary differential equations

$$\partial_t z_\varepsilon(t, \omega) = \sum_{i=1}^m \frac{g_i(z_\varepsilon(t, \omega))}{v_i(\varepsilon)} \int_{\Omega} h_i(z_\varepsilon(t, \nu)) d\rho(\nu) + f(z_\varepsilon(t, \omega)), \quad (\text{B.2})$$

$$z_\varepsilon(0, \omega) = z_{\varepsilon,0}(\omega) \in \mathcal{C}_c(\Omega, \mathbb{R}^d). \quad (\text{B.3})$$

Since, for any $\omega \in \Omega$, the right-hand-side of (B.2) is locally Lipschitz on $\mathcal{O}$, the Cauchy-Lipschitz theorem guarantees the existence and uniqueness of a local solution $z_\varepsilon(t, \omega) \in \mathcal{C}^1([0, T_0(\omega)], \mathbb{R}^d)$ to the initial value problem (B.2)-(B.3), for some time horizon $T_0(\omega)$. In what follows, we will denote $T_0 = \inf_{\omega \in \Omega} T_0(\omega)$. For any $u \in \mathcal{C}(\Omega, \mathbb{R}^d)$, we let $\mathbf{G}(u) \in \mathbb{R}^{m \times m}$ and $\mathbf{F}(u) \in \mathbb{R}^m$ respectively denote the matrix and vector whose entries are given by

$$[\mathbf{G}_{ij}(u)]_{1 \leq i, j \leq m} = \int_{\Omega} \langle g_j(u(\omega)), \nabla h_i(u(\omega)) \rangle d\rho(\omega), \quad (\text{B.4})$$

$$[\mathbf{F}_i(u)]_{1 \leq i \leq m} = \int_{\Omega} \langle f(u(\omega)), \nabla h_i(u(\omega)) \rangle d\rho(\omega). \quad (\text{B.5})$$

Assume that there exists a non-empty subset $\mathcal{U}$ of $\mathcal{C}(\Omega, \mathbb{R}^d)$ such that, for any $u \in \mathcal{U}$, the matrix $\mathbf{G}(u)$ is full-rank. Finally, for any $u \in \mathcal{U}$, we denote $\boldsymbol{\alpha}(u)$ the solution to the linear system

$$\mathbf{G}(u)\boldsymbol{\alpha}(u) + \mathbf{F}(u) = 0. \quad (\text{B.6})$$

With these definitions in hand, consider the auxiliary dynamical system

$$\partial_t z(t, \omega) = \sum_{i=1}^m \alpha_i(t) g_i(z(t, \omega)) + f(z(t, \omega)), \quad (\text{B.7})$$

$$z(0, \omega) = z_0(\omega) \in \mathcal{C}_c(\Omega, \mathbb{R}^d) \cap \mathcal{U}, \quad (\text{B.8})$$

where we have respectively abbreviated $\mathbf{G}(z(t, \cdot)), \boldsymbol{\alpha}(z(t, \cdot)), \mathbf{F}(z(t, \cdot))$ with $\mathbf{G}(t), \boldsymbol{\alpha}(t), \mathbf{F}(t)$. Letting $T_{\mathcal{U}} = \inf\{t \geq 0 : z(t, \cdot) \notin \mathcal{U}\}$, the Cauchy-Lipschitz theorem once again guarantees the existence, for each ω , of a time horizon $\bar{T}_0(\omega)$ such that there exists a unique local solution $z(t, \omega) \in \mathcal{C}^1([0, \bar{T}_0(\omega) \wedge T_{\mathcal{U}}], \mathbb{R}^d)$. In what follows, we will denote $\bar{T}_0 = \inf_{\omega \in \Omega} \bar{T}_0(\omega)$. A short computation then shows that the coefficients $\alpha_i(t)$ impose the constraints

$$z(t, \cdot) \in S, \quad (\text{B.9})$$

$$\partial_t z(t, \omega) \in f(z(t, \omega)) + \text{Vec}(\{g_i(z(t, \omega))\}_{i=1}^m), \quad \forall \omega \in \Omega. \quad (\text{B.10})$$

We seek to bound the ℓ_2 distance $\|z_\varepsilon(t, \omega) - z(t, \omega)\|_2$, for each $\omega \in \Omega$, on a time horizon $T_{\mathcal{S}}$ such that $z_\varepsilon(t, \cdot)$ is close to the set $\mathcal{S}$ in the following sense: for each $1 \leq i \leq m$, we assume that there exists a continuous, non-decreasing, positive function u_i such that $u_i(0) = 0$ and, for any $t \leq T_{\mathcal{S}}$:

$$\int_{\Omega} h_i(z_\varepsilon(t, \omega)) d\rho(\omega) \leq u_i(\varepsilon). \quad (\text{B.11})$$

For any $z_0 \in \mathbb{R}^d$ and positive constant R , we denote

$$C_f(z_0, R) = \sup_{z \in \mathcal{B}(z_0, R)} \|f(z)\|_2, \quad (\text{B.12})$$

$$C'_f(z_0, R) = \sup_{z \in \mathcal{B}(z_0, R)} \|\nabla f(z)\|_{op}, \quad (\text{B.13})$$

along with the corresponding quantities for the $\{g_i\}_{1 \leq i \leq nm}$ and h . We also let, for any $1 \leq i \leq m$,

$$C''_{h,i}(z_0, R) = \sup_{z \in \mathcal{B}(z_0, R)} \|\nabla^2 h_i(z)\|_{op}, \quad (\text{B.14})$$

and introduce the stopping times, for any positive constant a and any $\omega \in \Omega$

$$T_a = \inf\{0 \leq t \leq T_0 \wedge \bar{T}_0 \wedge T_{\mathcal{U}} : \sigma_{\min}(G(z_\varepsilon(t, \cdot))) \wedge \sigma_{\min}(G(z(t, \cdot))) \leq a\}, \quad (\text{B.15})$$

$$T_{z_0, R}(\omega) = \inf\{0 \leq t \leq T_0 \wedge \bar{T}_0 \wedge T_{\mathcal{U}} : z_\varepsilon(t, \omega) \notin \mathcal{B}(z_0(\omega), R) \text{ or } z(t, \omega) \notin \mathcal{B}(z_0(\omega), R)\}, \quad (\text{B.16})$$

and we will denote $T_{z_0, R} = \inf_{\omega \in \Omega} T_{z_0, R}(\omega)$. Using the above notations, for any $\delta > 0$, consider the following quantities

$$L_{z_\varepsilon} = \sum_{i=1}^m \frac{u_i(\varepsilon)}{v_i(\varepsilon)} C_{g_i}(z_0, R) + C_f(z_0, R), \quad (\text{B.17})$$

$$C_\alpha = \frac{C_f(z_0, R)}{a} \left(\sum_{i=1}^m C'_{h_i}(z_0, R)^2 \right)^{1/2}, \quad (\text{B.18})$$

$$L_z = C_\alpha \left(\sum_{i=1}^m C_{g_i}(z_0, R)^2 \right)^{1/2} + C_f(z_0, R), \quad (\text{B.19})$$

$$C'_\alpha = \left(\frac{1}{a} \left(\sum_{i=1}^m (C'_f C'_{h_i} + C_f C''_{h_i})^2 \right) + \frac{m}{a} \sup_{1 \leq i, j \leq n} (C_{g_i} C''_{h_j} + C'_{g_i} C'_{h_j}) C_\alpha \right)^{1/2}, \quad (\text{B.20})$$

$$K_{v_z} = \frac{\delta^2}{\sqrt{3}} L_{z_\varepsilon}(z_0, R) \left(\sum_{i=1}^m \frac{u_i(\varepsilon)}{v_i(\varepsilon)} C'_{g_i}(z_0, R) + C'_f(z_0, R) \right), \quad (\text{B.21})$$

$$K_{v_{z_\varepsilon}} = \frac{\delta^2}{\sqrt{3}} C'_f(z_0, R) L_z(z_0, R) + \frac{\delta^2 \sqrt{2m}}{\sqrt{3}} \left(C_\alpha^2 L_z^2 \left(\sup_{1 \leq i \leq m} C'_{g_i} \right)^2 + \left(\sup_{1 \leq i \leq m} C_{g_i} \right)^2 (C'_\alpha L_z)^2 \right)^{1/2}, \quad (\text{B.22})$$

as well as

$$K = \left(C'_f + C_\alpha \left(\sum_{i=1}^m (C'_{g_i})^2 \right)^{1/2} + C'_\alpha \left(\sum_{i=1}^m C_{g_i}^2 \right)^{1/2} \right), \quad (\text{B.23})$$

$$\ell(\delta, \varepsilon) = K_{v_z} + K_{v_{z_\varepsilon}} + \frac{(\sum_{i=1}^m C_{g_i}^2)^{1/2}}{a} \left(\left(\sum_{i=1}^m (2u_i(\varepsilon) + C''_{h_i} L_z^2 \delta^2)^2 \right)^{1/2} + K_{v_{z_\varepsilon}} \left(\sum_i^n (C'_{h_i}(z_0, R))^2 \right)^{1/2} \right). \quad (\text{B.24})$$

We then have the following result:

Proposition B.1. *Using the notations and under the assumptions outlined above, fix a and R such that $T_a \wedge T_{z_0, R} > 0$, and denote $S = T_S \wedge T_a \wedge T_{z_0, R} \wedge T_0 \wedge \bar{T}_0 \wedge T_{\mathcal{U}}$. Then, for any $\delta > 0$ and $n \in \mathbb{N}$ such that $n\delta \leq S$,*

$$\|z_\varepsilon(n\delta, \cdot) - z(n\delta, \cdot)\|_2 \leq (\|z_\varepsilon(0, \cdot) - z(0, \cdot)\|_2 + n\ell(\delta, \varepsilon)) \exp(Kn\delta). \quad (\text{B.25})$$

Proof. We begin by writing, for any $t < S$ and $\delta > 0$ such that $t + \delta \leq S$,

$$z_\varepsilon(t + \delta, \omega) = z_\varepsilon(t, \omega) + \int_0^\delta f(z_\varepsilon(t + s, \omega)) ds + \sum_{i=1}^m \int_0^\delta g_i(z_\varepsilon(t + s, \omega)) \int_\Omega \frac{h_i(z_\varepsilon(t + s, \nu))}{v_i(\varepsilon)} d\rho(\nu) ds, \quad (\text{B.26})$$

$$z(t + \delta, \omega) = z(t, \omega) + \int_0^\delta f(z(t + s, \omega)) ds + \sum_{i=1}^m \int_0^\delta \alpha_i(t + s) g_i(z(t + s, \omega)) ds. \quad (\text{B.27})$$

Let us introduce the vectors v_{z_ε}, v_z , defined as

$$v_{z_\varepsilon}(t, \omega) = z_\varepsilon(t + \delta, \omega) - z_\varepsilon(t, \omega) - \delta f(z_\varepsilon(t, \omega)) - \sum_{i=1}^m g_i(z_\varepsilon(t, \omega)) \int_0^\delta \int_\Omega \frac{h_i(z_\varepsilon(t + s, \nu))}{v_i(\varepsilon)} d\rho(\nu) ds, \quad (\text{B.28})$$

$$v_z(t, \omega) = z(t + \delta, \omega) - z(t, \omega) - \delta f(z(t, \omega)) - \delta \sum_{i=1}^m \alpha_i(t) g_i(z(t, \omega)), \quad (\text{B.29})$$

and turn to bounding their norms.

Bounding the norms of v_{z_ε}, v_z . Combining Eq.(B.26) with Eq.(B.28), and Eq.(B.27) with Eq.(B.29), we find

$$\begin{aligned} v_{z_\varepsilon}(t, \omega) &= \int_0^\delta f(z_\varepsilon(t + s, \omega)) ds - \delta f(z_\varepsilon(t, \omega)) \\ &+ \sum_{i=1}^m \left(\int_0^\delta g_i(z_\varepsilon(t + s, \omega)) \int_\Omega \frac{h_i(z_\varepsilon(t + s, \nu))}{v_i(\varepsilon)} d\rho(\nu) ds - g_i(z_\varepsilon(t, \omega)) \int_0^\delta \int_\Omega \frac{h_i(z_\varepsilon(t + s, \nu))}{v_i(\varepsilon)} d\rho(\nu) ds \right) \end{aligned} \quad (\text{B.30})$$

and

$$v_z(t, \omega) = \int_0^\delta f(z(t + s, \omega)) ds - \delta f(z(t, \omega)) + \sum_{i=1}^m \left(\int_0^\delta \alpha_i(t + s) g_i(z(t + s, \omega)) ds - \delta \alpha_i(t) g_i(z(t, \omega)) \right). \quad (\text{B.31})$$

Starting with v_{z_ε} , a first order Taylor expansion then gives

$$f(z_\varepsilon(t + s, \omega)) = f(z_\varepsilon(t, \omega)) + s Df[z(\tilde{t}, \omega)] \partial_t z(\tilde{t}, \omega), \quad (\text{B.32})$$

for some $\tilde{t} \in [t, t + s]$. But then, owing to Eq.(B.2) and the definition of S ,

$$\sup_{0 \leq t \leq S, \omega \in \Omega} \|\partial_t z_\varepsilon(t, \omega)\|_2 \leq \sum_{i=1}^m \frac{u_i(\varepsilon)}{v_i(\varepsilon)} C_{g_i}(z_0, R) + C_f(z_0, R), \quad (\text{B.33})$$

so that, letting $L_{z_\varepsilon}(z_0, R)$ denote the right-hand-side of the above inequality, we find

$$\left\| \int_0^\delta f(z_\varepsilon(t+s, \omega)) ds - \delta f(z_\varepsilon(t, \omega)) \right\|_2 \leq \sqrt{\delta} \left(\int_0^\delta s^2 \|Df[z(\tilde{t}(s), \omega)]\|_{op}^2 \|\partial_t z(\tilde{t}(s), \omega)\|_2^2 ds \right)^{1/2} \quad (\text{B.34})$$

$$\leq \frac{\delta^2}{\sqrt{3}} C'_f(z_0, R) L_{z_\varepsilon}(z_0, R). \quad (\text{B.35})$$

Letting, for any $1 \leq i \leq m$, $\beta_{\varepsilon,i}(t) = \int_\Omega \frac{h_i(z_\varepsilon(t, \omega))}{v_i(\varepsilon)} d\rho(\omega)$, we turn to bounding the second term in Eq.(B.30). Starting again with a first order Taylor expansion, we find

$$g_i(z_\varepsilon(t+s, \omega)) = g_i(z_\varepsilon(t, \omega)) + s Dg_i[z_\varepsilon(\tilde{t}, \omega)] \partial_t z_\varepsilon(\tilde{t}, \omega), \quad (\text{B.36})$$

for some $\tilde{t}(s) \in [t, t+\delta]$. Then,

$$\begin{aligned} & \left\| \int_0^\delta g_i(z_\varepsilon(t+s, \omega)) \beta_{\varepsilon,i}(t+s) ds - g_i(z_\varepsilon(t, \omega)) \int_0^\delta \beta_{\varepsilon,i}(t+s) ds \right\|_2, \\ & \leq \left\| \int_0^\delta s Dg_i[z_\varepsilon(\tilde{t}(s), \omega)] \partial_t z_\varepsilon(\tilde{t}(s), \omega) \beta_{\varepsilon,i}(t+s) ds \right\|_2, \\ & \leq \frac{\delta^2}{\sqrt{3}} C'_{g_i}(z_0, R) L_{z_\varepsilon}(z_0, R) \frac{u_i(\varepsilon)}{v_i(\varepsilon)}, \end{aligned} \quad (\text{B.37})$$

where we have used the upper bound (B.33) and the fact that $\sup_{0 \leq t \leq S} |\beta_{\varepsilon,i}(t)| \leq \frac{u_i(\varepsilon)}{v_i(\varepsilon)}$. Combining the bounds (B.35) and (B.37), along with the triangle inequality, we reach

$$\|v_{z_\varepsilon}(t, \omega)\|_2 \leq \frac{\delta^2}{\sqrt{3}} L_{z_\varepsilon}(z_0, R) \left(\sum_{i=1}^m \frac{u_i(\varepsilon)}{v_i(\varepsilon)} C'_{g_i}(z_0, R) + C'_f(z_0, R) \right). \quad (\text{B.38})$$

Turning to v_z , we begin by noting that, for any $t \leq S$,

$$\left\| \sum_{i=1}^m \alpha_i(t) g_i(z(t, \omega)) \right\|_2 \leq \sum_{i=1}^m |\alpha_i(t)| C_{g_i}(z_0, R) \quad (\text{B.39})$$

$$\leq \|\alpha(t)\|_2 \left(\sum_{i=1}^m C_{g_i}(z_0, R)^2 \right)^{1/2}, \quad (\text{B.40})$$

owing to the Cauchy-Schwarz inequality. But then, recalling that $\alpha(t) = -\mathbf{G}(t)^{-1} \mathbf{F}(t)$, we find

$$\|\alpha(t)\|_2 \leq \frac{\|\mathbf{F}(t)\|_2}{a}, \quad (\text{B.41})$$

where, recalling the definition of $\mathbf{F}(t) := \mathbf{F}(z(t, \cdot))$, we have the upper bound

$$\|\mathbf{F}(t)\|_2 \leq C_f(z_0, R) \left(\sum_{i=1}^m C'_{h_i}(z_0, R)^2 \right)^{1/2} \quad (\text{B.42})$$

Letting C_F denote the right-hand-side of the above inequality and $C_\alpha = \frac{C_F}{a}$, we reach

$$\sup_{0 \leq t \leq S, \omega \in \Omega} \|\partial_t z(t, \omega)\|_2 \leq C_\alpha \left(\sum_{i=1}^m C_{g_i}(z_0, R)^2 \right)^{1/2} + C_f(z_0, R). \quad (\text{B.43})$$

Letting $L_z(z_0, R)$ denote the right-hand-side of inequality (B.43), an argument similar to the one used to obtain the bound (B.35) gives

$$\left\| \int_0^\delta f(z(t+s, \omega)) ds - \delta f(z(t, \omega)) \right\|_2 \leq \frac{\delta^2}{\sqrt{3}} C'_f(z_0, R) L_z(z_0, R). \quad (\text{B.44})$$

Now, for any $0 < t, t' \leq S$, the triangle inequality gives

$$\|\alpha(t) - \alpha(t')\|_2 \leq \|\mathbf{G}^{-1}(t) - \mathbf{G}^{-1}(t')\|_{op} \|\mathbf{F}(t)\|_2 + \|\mathbf{G}^{-1}(t)\|_{op} \|\mathbf{F}(t) - \mathbf{F}(t')\|_2. \quad (\text{B.45})$$

Then, for any $1 \leq i \leq m$, the fact that $f, \nabla h_i$ are bounded Lipschitz (on $\mathcal{B}(z_0, R)$) gives,

$$|[\mathbf{F}(t) - \mathbf{F}(t')]_i| \leq \int_{\Omega} |(\langle f(z(t, \omega)), \nabla h_i(z(t, \omega)) \rangle - \langle f(z(t', \omega)), \nabla h_i(z(t', \omega)) \rangle)| d\rho(\omega) \quad (\text{B.46})$$

$$\leq (C'_f C'_{h_i} + C_f C''_{h_i}) L_z |t - t'|, \quad (\text{B.47})$$

leading to

$$\|\mathbf{F}(t) - \mathbf{F}(t')\|_2 \leq \left(\sum_{i=1}^m (C'_f C'_{h_i} + C_f C''_{h_i})^2 \right)^{1/2} L_z |t - t'|. \quad (\text{B.48})$$

Then, noting that $\mathbf{G}^{-1}(t) - \mathbf{G}^{-1}(t') = \mathbf{G}^{-1}(t') (\mathbf{G}(t') - \mathbf{G}(t)) \mathbf{G}^{-1}(t)$

$$\|\mathbf{G}^{-1}(t) - \mathbf{G}^{-1}(t')\|_{op} \leq \frac{1}{a^2} \|\mathbf{G}(t) - \mathbf{G}(t')\|_{op} \quad (\text{B.49})$$

$$\leq \frac{m}{a^2} \sup_{1 \leq i, j \leq m} |[\mathbf{G}(t) - \mathbf{G}(t')]_{ij}| \quad (\text{B.50})$$

$$\leq \frac{m}{a^2} \sup_{1 \leq i, j \leq m} (C_{g_i} C''_{h_j} + C'_{g_i} C'_{h_j}) L_z |t - t'|. \quad (\text{B.51})$$

Letting

$$C'_\alpha = \left(\frac{1}{a} \left(\sum_{i=1}^m (C'_f C'_{h_i} + C_f C''_{h_i})^2 \right)^{1/2} + \frac{m}{a^2} \sup_{1 \leq i, j \leq n} (C_{g_i} C''_{h_j} + C'_{g_i} C'_{h_j}) C_F \right), \quad (\text{B.52})$$

we reach

$$\|\alpha(t) - \alpha(t')\|_2 \leq C'_\alpha L_z |t - t'|. \quad (\text{B.53})$$

Then, for any $1 \leq i \leq m$, we write

$$\begin{aligned} \alpha_i(t+s)g_i(z(t+s), \omega) - \alpha_i(t)g_i(z(t, \omega)) &= \alpha_i(t+s)g_i(z(t+s), \omega) - \alpha_i(t+s)g_i(z(t), \omega), \\ &\quad + \alpha_i(t+s)g_i(z(t), \omega) - \alpha_i(t)g_i(z(t), \omega), \end{aligned}$$

so that

$$\|\alpha_i(t+s)g_i(z(t+s), \omega) - \alpha_i(t)g_i(z(t), \omega)\|_2 \leq |\alpha_i(t+s)| C'_{g_i} L_z s + C_{g_i} |\alpha_i(t+s) - \alpha_i(t)|, \quad (\text{B.54})$$

and

$$\begin{aligned} &\left\| \sum_{i=1}^m \left(\int_0^\delta \alpha_i(t+s)g_i(z(t+s), \omega) ds - \delta \alpha_i(t)g_i(z(t), \omega) \right) \right\|_2 \leq \\ &\sqrt{2\delta} \sum_{i=1}^m \left(\int_0^\delta (|\alpha_i(t+s)| C'_{g_i} L_z s)^2 ds + \int_0^\delta (C_{g_i} |\alpha_i(t+s) - \alpha_i(t)|)^2 ds \right)^{1/2} \\ &\leq \sqrt{2\delta m} \left(\frac{\delta^3}{3} L_z^2 \left(\sup_{1 \leq i \leq n} C'_{g_i} \right)^2 \sum_{i=1}^m \alpha_i(t+s)^2 + \left(\sup_{1 \leq i \leq n} C_{g_i} \right)^2 \int_0^\delta \|\alpha(t+s) - \alpha(t)\|_2^2 ds \right)^{1/2} \\ &\leq \frac{\sqrt{2m}}{\sqrt{3}} \delta^2 \left(C_\alpha^2 L_z^2 \left(\sup_{1 \leq i \leq n} C'_{g_i} \right)^2 + \left(\sup_{1 \leq i \leq n} C_{g_i} \right)^2 (C'_\alpha)^2 \right)^{1/2} \end{aligned}$$

We finally reach

$$\|v_z(t, \omega)\|_2 \leq \frac{\delta^2}{\sqrt{3}} C'_f(z_0, R) L_z(z_0, R) + \frac{\delta^2 \sqrt{2m}}{\sqrt{3}} \left(C_\alpha^2 L_z^2 \left(\sup_{1 \leq i \leq n} C'_{g_i} \right)^2 + \left(\sup_{1 \leq i \leq n} C_{g_i} \right)^2 (C'_\alpha L_z)^2 \right)^{1/2}. \quad (\text{B.55})$$

In what follows, we will respectively denote $K_{v_{z_\varepsilon}}$ and K_{v_z} the bounds obtained at Eq.(B.38) and Eq.(B.55). We now turn to showing that the vector $\alpha(t)$ is close to the vector $\alpha_\varepsilon(t)$ whose entries are given by

$$\alpha_{\varepsilon, i}(t) = \frac{1}{\delta} \int_0^\delta \int_{\Omega} \frac{h_i(z_\varepsilon(t+s), \omega)}{v_i(\varepsilon)} d\rho(\omega) ds, \quad \forall 1 \leq i \leq m. \quad (\text{B.56})$$

Approximating the $\alpha_{\varepsilon,i}(t)$ with the $\alpha_i(t)$. Owing to the definitions of v_{z_ε}, v_z , we may write

$$z_\varepsilon(t + \delta, \omega) = z_\varepsilon(t, \omega) + \delta f(z_\varepsilon(t, \omega)) + \delta \sum_{i=1}^m \alpha_{\varepsilon,i}(t) g_i(z_\varepsilon(t, \omega)) + v_{z_\varepsilon}(t, \omega), \quad (\text{B.57})$$

$$z(t + \delta, \omega) = z(t, \omega) + \delta f(z(t, \omega)) + \delta \sum_{i=1}^m \alpha_i(t) g_i(z(t, \omega)) + v_z(t, \omega). \quad (\text{B.58})$$

Then, for any $1 \leq i \leq m$, an order two Taylor expansion gives that

$$\begin{aligned} h_i(z_\varepsilon(t + \delta, \omega)) &= h_i(z_\varepsilon(t, \omega)) + \langle \nabla h_i(z_\varepsilon(t, \omega)), z_\varepsilon(t + \delta, \omega) - z_\varepsilon(t, \omega) \rangle, \\ &\quad + \frac{1}{2} \nabla^2 h_i(\tilde{z}_\varepsilon) [(z_\varepsilon(t + \delta, \omega) - z_\varepsilon(t, \omega))^{\otimes 2}], \end{aligned} \quad (\text{B.59})$$

for some $\tilde{z}_\varepsilon \in \mathcal{B}(z_\varepsilon(t, \omega), \|z_\varepsilon(t + \delta, \omega) - z_\varepsilon(t, \omega)\|_2) \subset \mathcal{B}(z_{\varepsilon,0}(\omega), 3R)$. Letting

$$s_i(t, \omega) = h_i(z_\varepsilon(t + \delta, \omega)) - h_i(z_\varepsilon(t, \omega)) - \frac{1}{2} \nabla^2 h_i(\tilde{z}_\varepsilon) [(z_\varepsilon(t + \delta, \omega) - z_\varepsilon(t, \omega))^{\otimes 2}] \quad (\text{B.60})$$

and replacing the increment $z_\varepsilon(t + \delta, \omega) - z_\varepsilon(t, \omega)$ using Eq.(B.57) in Eq.(B.59), and integrating with respect to ω , we find, for any $1 \leq j \leq m$,

$$\int_{\Omega} \langle \nabla h_j(z_\varepsilon(t, \omega)), \delta f(z_\varepsilon(t, \omega)) + \delta \sum_{i=1}^m \alpha_{\varepsilon,i}(t) g_i(z_\varepsilon(t, \omega)) + v_{z_\varepsilon}(t, \omega) \rangle d\rho(\omega) = \int_{\Omega} s_j(t, \omega) d\rho(\omega). \quad (\text{B.61})$$

Now, defining the vectors $\mathbf{s}(t)$ and $\mathbf{V}_{z_\varepsilon}(t)$, whose entries are respectively given by

$$[\mathbf{s}(t)]_{1 \leq i \leq m} = \int_{\Omega} s_i(t, \omega) d\rho(\omega), \quad (\text{B.62})$$

$$[\mathbf{V}_{z_\varepsilon}(t)]_{1 \leq i \leq m} = \int_{\Omega} \langle \nabla h_i(z_\varepsilon(t, \omega)), v_{z_\varepsilon}(t, \omega) \rangle d\rho(\omega), \quad (\text{B.63})$$

we may rewrite Eq.(B.61) as

$$\delta (\mathbf{G}(z_\varepsilon(t, \cdot)) \boldsymbol{\alpha}_\varepsilon(t) + \mathbf{F}(z_\varepsilon(t, \cdot))) = -\mathbf{V}_{z_\varepsilon}(t) + \mathbf{s}(t). \quad (\text{B.64})$$

Abbreviating $\mathbf{G}(z_\varepsilon(t, \cdot))$ with $\mathbf{G}^\varepsilon(t)$ and $\mathbf{F}(z_\varepsilon(t, \cdot))$ with $\mathbf{F}^\varepsilon(t)$, and letting $\tilde{\boldsymbol{\alpha}}_\varepsilon(t) = -\mathbf{G}^\varepsilon(t)^{-1} \mathbf{F}^\varepsilon(t)$, we may write

$$\begin{aligned} \|\boldsymbol{\alpha}(t) - \boldsymbol{\alpha}_\varepsilon(t)\|_2 &\leq \|\boldsymbol{\alpha}(t) - \tilde{\boldsymbol{\alpha}}_\varepsilon(t)\|_2 + \frac{1}{\delta} \|\mathbf{G}^\varepsilon(t)^{-1} (-\mathbf{V}_{z_\varepsilon}(t) + \mathbf{s}(t))\| \\ &\leq \|\boldsymbol{\alpha}(t) - \tilde{\boldsymbol{\alpha}}_\varepsilon(t)\|_2 + \frac{1}{\delta a} (\|\mathbf{V}_{z_\varepsilon}(t)\|_2 + \|\mathbf{s}(t)\|_2). \end{aligned} \quad (\text{B.65})$$

Now, using the definition of $s_i(t)$ at Eq.(B.60), we find

$$\sup_{0 \leq t \leq S} \left| \int_{\Omega} s_i(t, \omega) d\rho(\omega) \right| \leq 2u_i(\varepsilon) + C''_{h_i} L_z^2 \delta^2, \quad (\text{B.66})$$

and thus

$$\|\mathbf{s}(t)\|_2 \leq \left(\sum_{i=1}^m (2u_i(\varepsilon) + C''_{h_i} L_z^2 \delta^2)^2 \right)^{1/2}. \quad (\text{B.67})$$

Also,

$$\|\mathbf{V}_{z_\varepsilon}(t)\|_2 \leq K_{v_{z_\varepsilon}} \left(\sum_i^m (C'_{h_i}(z_0, R))^2 \right)^{1/2}. \quad (\text{B.68})$$

Furthermore, the argument employed to establish Eq.(B.53) also implies that

$$\|\boldsymbol{\alpha}(t) - \tilde{\boldsymbol{\alpha}}_\varepsilon(t)\|_2 \leq C'_\alpha \|z(t) - z_\varepsilon(t)\|_2. \quad (\text{B.69})$$

In summary, we reach

$$\begin{aligned} \|\boldsymbol{\alpha}(t) - \boldsymbol{\alpha}_\varepsilon(t)\|_2 &\leq \frac{1}{\delta a} \left(\left(\sum_{i=1}^m (2u_i(\varepsilon) + C''_{h_i} L_z^2 \delta^2)^2 \right)^{1/2} + K_{v_{z_\varepsilon}} \left(\sum_i^m (C'_{h_i}(z_0, R))^2 \right)^{1/2} \right) \\ &\quad + C'_\alpha \|z(t, \omega) - z_\varepsilon(t, \omega)\|_2. \end{aligned} \quad (\text{B.70})$$

Concluding the proof. To conclude the proof, we use Eq. (B.57), (B.58) to obtain that, for all $\omega \in \Omega$,

$$\begin{aligned} \|z_\varepsilon(t + \delta, \omega) - z(t + \delta, \omega)\|_2 &\leq (1 + \delta C'_f) \|z_\varepsilon(t, \omega) - z(t, \omega)\|_2 + K_{v_z} + K_{v_{z_\varepsilon}}, \\ &\quad + \delta \left\| \sum_{i=1}^m (\alpha_{\varepsilon,i}(t) g_i(z_\varepsilon(t, \omega)) - \alpha_i(t) g_i(z(t, \omega))) \right\|_2. \end{aligned} \quad (\text{B.71})$$

But then,

$$\begin{aligned} \left\| \sum_{i=1}^m (\alpha_{\varepsilon,i}(t) g_i(z_\varepsilon(t, \omega)) - \alpha_i(t) g_i(z(t, \omega))) \right\|_2 &\leq \sum_{i=1}^m \|\alpha_{\varepsilon,i}(t) - \alpha_i(t)\|_2 \|g_i(z_\varepsilon(t, \omega))\|_2 \\ &\quad + \sum_{i=1}^m \|\alpha_i(t)\|_2 \|g_i(z_\varepsilon(t, \omega)) - g_i(z(t, \omega))\|_2 \\ &\leq \left(\sum_{i=1}^m C_{g_i}^2 \right)^{1/2} \|\alpha(t) - \alpha_\varepsilon(t)\|_2 + C_\alpha \left(\sum_{i=1}^m (C'_{g_i})^2 \right)^{1/2} \|z(t, \omega) - z_\varepsilon(t, \omega)\|_2. \end{aligned}$$

Gathering the above bounds, we introduce the quantities

$$K = \left(C'_f + C_\alpha \left(\sum_{i=1}^m (C'_{g_i})^2 \right)^{1/2} + C'_\alpha \left(\sum_{i=1}^m C_{g_i}^2 \right)^{1/2} \right), \quad (\text{B.72})$$

$$\ell(\delta, \varepsilon) = K_{v_z} + K_{v_{z_\varepsilon}} + \frac{(\sum_{i=1}^m C_{g_i}^2)^{1/2}}{a} \left(\left(\sum_{i=1}^m (2u_i(\varepsilon) + C''_{h_i} L_z^2 \delta^2)^2 \right)^{1/2} + K_{v_{z_\varepsilon}} \left(\sum_i (C'_{h_i}(z_0, R))^2 \right)^{1/2} \right), \quad (\text{B.73})$$

so that we finally obtain the following control:

$$\|z_\varepsilon(t + \delta, \omega) - z(t + \delta, \omega)\|_2 \leq (1 + K\delta) \|z_\varepsilon(t, \omega) - z(t, \omega)\|_2 + \ell(\delta, \varepsilon). \quad (\text{B.74})$$

Applying the discrete Gronwall inequality then gives that, for any $n \in \mathbb{N}$ such that $n\delta \leq S$,

$$\|z_\varepsilon(n\delta, \omega) - z(n\delta, \omega)\|_2 \leq (\|z_\varepsilon(0, \omega) - z(0, \omega)\|_2 + n\ell(\delta, \varepsilon)) \exp(Kn\delta). \quad (\text{B.75})$$

□

We now specialize this result to a family of dynamical systems related to (2.8)-(2.9), using the same notations. For any integer D , consider the following evolution:

$$\partial_t a(t, \omega) = \sum_{k=0}^{D-1} \sigma_k s^k(t, \omega) \frac{(\varphi_k - \sigma_k \overline{as^k}(t))}{v_k(\varepsilon)} + \sigma_D \varphi_D s^D(t, \omega), \quad (\text{B.76})$$

$$\partial_t s(t, \omega) = \sum_{k=1}^{D-1} k \sigma_k a(t, \omega) s^{k-1}(t, \omega) \frac{(\varphi_k - \sigma_k \overline{as^k}(t))}{v_k(\varepsilon)} + D \sigma_D \varphi_D a(t, \omega) s^{D-1}(t, \omega), \quad (\text{B.77})$$

with initial condition $a(0, \cdot), s(0, \cdot) \in \mathcal{C}_M(\Omega, \mathbb{R})$. For any pair $z_1(t, \cdot), z_2(t, \cdot) \in \mathcal{C}(\Omega, \mathbb{R})$, define the p -dimensional vector $\alpha(z_1(t, \cdot), z_2(t, \cdot))$ as any solution to the linear system

$$\forall 0 \leq p \leq D-1 : \quad \sum_{k=0}^{D-1} \alpha_k(t) \left(\sigma_k \overline{z_2^{p+k}}(t) + p k \sigma_k \overline{z_1^2 z_2^{p+k-2}}(t) \right) = -\sigma_D \varphi_D \left(\overline{z_2^{p+D}}(t) + D p \overline{z_1^2 z_2^{p+D-2}}(t) \right). \quad (\text{B.78})$$

Assume that there exists a pair $z_1(0, \cdot), z_2(0, \cdot)$ such that the system (B.78) has a unique solution, and consider the dynamical system

$$\partial_t z_1(t, \omega) = \sum_{k=0}^{D-1} \sigma_k z_2^k(t, \omega) \alpha_k(t) + \sigma_D \varphi_D z_2^D(t, \omega), \quad (\text{B.79})$$

$$\partial_t z_2(t, \omega) = \sum_{k=1}^{D-1} k \sigma_k z_1(t, \omega) z_2^{k-1}(t, \omega) \alpha_k(t) + D \sigma_D \varphi_D z_1(t, \omega) z_2^{D-1}(t, \omega), \quad (\text{B.80})$$

initialized with $z_1(0, \cdot), z_2(0, \cdot)$. Finally, define $z_\varepsilon(t, \omega) = (a(t, \omega), s(t, \omega))$ and $z(t, \omega) = (z_1(t, \omega), z_2(t, \omega))$.

Corollary B.2. *With the same notations as in Proposition B.1, assume that there exists a constant C_0 (independent on ε) such that, for each $0 \leq k \leq D-1$:*

$$u_k(\varepsilon) \leq C_0 v_k(\varepsilon), \quad (\text{B.81})$$

and denote $u^*(\varepsilon) = \sup_k \{u_k(\varepsilon)\}_{k=0}^{D-1}$. Then, there exist constants C_1, C_2 depending only on K, R, C_0 and the coefficients $\{\sigma_k, \varphi_k\}_{k=0}^D$ such that, for any $0 \leq t \leq S$,

$$\|z_\varepsilon(t, \cdot) - z(t, \cdot)\| \leq \left(\|z_\varepsilon(0, \cdot) - z(0, \cdot)\| + C_1 \sqrt{u^*(\varepsilon)} \right) \exp(C_2 t). \quad (\text{B.82})$$

Proof. We apply Proposition B.1 to the evolution (B.76)-(B.77), the form of which implies that there exist constants C_1, C_2, C_3 depending only on $D, R, \{\sigma_k, \varphi_k\}_{k=0}^D$ such that $K \leq C_1$ and $l(\delta, \varepsilon) \leq C_2 \delta^2 + C_3 u^*(\varepsilon)$. Choosing $n = \lfloor \frac{t}{\delta} \rfloor$, we find, for any $\delta < 1$,

$$\|z_\varepsilon(t, \cdot) - z(t, \cdot)\| \leq \left(\|z_\varepsilon(0, \cdot) - z(0, \cdot)\| + \frac{t}{\delta} (C_1 \delta^2 + C_3 u^*(\varepsilon)) \right) \exp(C_2 t). \quad (\text{B.83})$$

Optimizing over δ then gives the desired result. $\square$

APPENDIX C. AUXILIARY LEMMAS

Lemma C.1. *Let $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ be a Lipschitz continuous function and let $F(t) = \sup_{r \leq t} |f(r)|$. Then F is Lipschitz continuous. Furthermore, if there exists a positive, increasing function g such that, for every $t \geq 0$, $|\partial_t f(t)| \leq g(t)$, then $\partial_t F(t) \leq g(t)$ for (Lebesgue) almost every t .*

Proof. For any $t \geq 0$ and since f is continuous, we may define $r_t = \arg \sup_{r \leq t} |f(r)|$. Furthermore, let

$$\tilde{r} = \begin{cases} r_{t+s} & \text{if } r_{t+s} \leq t \\ t & \text{otherwise} \end{cases} \quad (\text{C.1})$$

Then, $F(t) \geq |f(\tilde{r})|$ and therefore

$$F(t) \geq |f(r_{t+s})| + |f(\tilde{r})| - |f(r_{t+s})|. \quad (\text{C.2})$$

By definition $|f(r_{t+s})| = F(t+s)$ and $r_{t+s} - \tilde{r} \leq s$, and, by the Lipschitz assumption, $||f(\tilde{r})| - |f(r_{t+s})|| \leq Ls$ for some constant L . Upon noting that $F(t+s) \geq F(t)$, this completes the proof of the first part of the statement. Now, since F is Lipschitz continuous, its derivative exists (Lebesgue) almost everywhere and, we have

$$\partial_t F(t) = \begin{cases} 0 & \text{if } r_t < t \\ \partial_t f(t) & \text{if } r_t = t. \end{cases} \quad (\text{C.3})$$

$\square$

Lemma C.2. *For any positive real numbers α, β , consider the ODE*

$$\dot{x}(t) = -\alpha x(t) + \beta \sqrt{x(t)}, \quad (\text{C.4})$$

$$x(0) > 0. \quad (\text{C.5})$$

Then, for all $t \geq 0$,

$$x(t) = \left(\sqrt{x(0)} \exp\left(-\frac{\alpha}{2}t\right) + \frac{\beta}{\alpha} \left(1 - \exp\left(-\frac{\alpha}{2}t\right)\right) \right)^2. \quad (\text{C.6})$$

Proof. This is a Bernoulli equation, see e.g. [52], and, clearly, $x(t) > 0$ for any $t > 0$, so that existence and uniqueness of maximal solutions follows straightforwardly. Making the change of variable $y(t) = \sqrt{x(t)}$, we find

$$\dot{y}(t) = -\frac{\alpha}{2}y(t) + \frac{\beta}{2}, \quad (\text{C.7})$$

so that

$$y(t) = y(0) \exp\left(-\frac{\alpha}{2}t\right) + \frac{\beta}{\alpha} \left(1 - \exp\left(-\frac{\alpha}{2}t\right)\right). \quad (\text{C.8})$$

$\square$

Lemma C.3. *Let $z : \mathbb{R}_+ \rightarrow \mathbb{R}^d$ be a continuous function. Then, for any $t \geq 0$,*

$$\left\| \int_0^t z(s) ds \right\|_2 \leq \int_0^t \|z(s)\|_2 ds, \quad (\text{C.9})$$

$$\left\| \int_0^t z(s) ds \right\|_2 \leq \sqrt{t} \left(\int_0^t \|z(s)\|_2^2 ds \right)^{1/2}. \quad (\text{C.10})$$

Proof. The first inequality is a consequence of the convexity of the Euclidean norm and Jensen’s inequality. For the second inequality, we apply the Cauchy-Schwarz inequality to obtain

$$\left\| \int_0^t z(s) ds \right\|_2^2 \leq \sum_{i=1}^d t \int_0^t z_i^2(s) ds. \quad (\text{C.11})$$

□

REFERENCES

- [1] E. ABBE, E. B. ADSERA, AND T. MISIAKIEWICZ, *The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks*, in Conference on Learning Theory, PMLR, 2022, pp. 4782–4887.
- [2] ———, *Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics*, in The Thirty Sixth Annual Conference on Learning Theory, PMLR, 2023, pp. 2552–2623.
- [3] L. ARNABOLDI, L. STEPHAN, F. KRZAKALA, AND B. LOUREIRO, *From high-dimensional $\mathcal{E}$ mean-field dynamics to dimensionless odes: A unifying approach to sgd in two-layers networks*, in The Thirty Sixth Annual Conference on Learning Theory, PMLR, 2023, pp. 1199–1227.
- [4] D. ARPIT, S. JASTRZĘBSKI, N. BALLAS, D. KRUEGER, E. BENGIO, M. S. KANWAL, T. MAHARAJ, A. FISCHER, A. COURVILLE, Y. BENGIO, ET AL., *A closer look at memorization in deep networks*, in International conference on machine learning, PMLR, 2017, pp. 233–242.
- [5] J. BA, M. A. ERDOĞDU, T. SUZUKI, Z. WANG, D. WU, AND G. YANG, *High-dimensional asymptotics of feature learning: How one gradient step improves the representation*, Advances in Neural Information Processing Systems, 35 (2022), pp. 37932–37946.
- [6] F. BACH, *Breaking the curse of dimensionality with convex neural networks*, Journal of Machine Learning Research, 18 (2017), pp. 1–53.
- [7] G. BEN AROUS, R. GHEISSARI, J. HUANG, AND A. JAGANNATH, *High-dimensional sgd aligns with emerging outlier eigenspaces*, in International Conference on Learning Representations, vol. 2024, 2024, pp. 47732–47778.
- [8] G. BEN AROUS, R. GHEISSARI, AND A. JAGANNATH, *Algorithmic thresholds for tensor pca*, The Annals of Probability, 48 (2020), pp. 2052–2087.
- [9] ———, *Online stochastic gradient descent on non-convex losses from high-dimensional inference*, Journal of Machine Learning Research, 22 (2021), pp. 1–51.
- [10] ———, *High-dimensional limit theorems for sgd: Effective dynamics and critical scaling*, Communications on Pure and Applied Mathematics, 77 (2024), pp. 2030–2080.
- [11] Y. BENGIO, I. GOODFELLOW, A. COURVILLE, ET AL., *Deep learning*, vol. 1, MIT press Cambridge, MA, USA, 2017.
- [12] R. BERTHIER, *Incremental learning in diagonal linear networks*, Journal of Machine Learning Research, 24 (2023), pp. 1–26.
- [13] R. BERTHIER, A. MONTANARI, AND K. ZHOU, *Learning time-scales in two-layers neural networks*, Foundations of Computational Mathematics, (2024), pp. 1–84.
- [14] A. BIETTI, J. BRUNA, AND L. PILLAUD-VIVIEN, *On learning gaussian multi-index models with gradient flow part i: General properties and two-timescale learning*, Communications on Pure and Applied Mathematics, 78 (2025), pp. 2354–2435.
- [15] A. BIETTI, J. BRUNA, C. SANFORD, AND M. J. SONG, *Learning single-index models with shallow neural networks*, Advances in neural information processing systems, 35 (2022), pp. 9768–9783.
- [16] V. S. BORKAR, *Stochastic approximation with two time scales*, Systems & Control Letters, 29 (1997), pp. 291–294.
- [17] N. BOUMAL, *An introduction to optimization on smooth manifolds*, Cambridge University Press, 2023.
- [18] L. CHIZAT, *Sparse optimization on measures with over-parameterized gradient descent*, Mathematical Programming, 194 (2022), pp. 487–532.
- [19] L. CHIZAT AND F. BACH, *On the global convergence of gradient descent for over-parameterized models using optimal transport*, Advances in neural information processing systems, 31 (2018).
- [20] ———, *Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss*, in Conference on learning theory, PMLR, 2020, pp. 1305–1338.
- [21] L. CHIZAT, E. OYALLON, AND F. BACH, *On lazy training in differentiable programming*, Advances in neural information processing systems, 32 (2019).
- [22] G. DALAL, G. THOPPE, B. SZÖRÉNYI, AND S. MANNOR, *Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning*, in Conference On Learning Theory, PMLR, 2018, pp. 1199–1233.
- [23] A. DAMIAN, J. LEE, AND M. SOLTANOLKOTABI, *Neural networks can learn representations with gradient descent*, in Conference on Learning Theory, PMLR, 2022, pp. 5413–5452.

- [24] Y. DANDI, F. KRZAKALA, B. LOUREIRO, L. PESCE, AND L. STEPHAN, *How two-layer neural networks learn, one (giant) step at a time*, Journal of Machine Learning Research, 25 (2024), pp. 1–65.
- [25] A. DANIELY AND E. MALACH, *Learning parities with neural networks*, Advances in Neural Information Processing Systems, 33 (2020), pp. 20356–20365.
- [26] S. DU, J. LEE, H. LI, L. WANG, AND X. ZHAI, *Gradient descent finds global minima of deep neural networks*, in International conference on machine learning, PMLR, 2019, pp. 1675–1685.
- [27] R. DUDEJA AND D. HSU, *Learning single-index models in gaussian space*, in Conference On Learning Theory, PMLR, 2018, pp. 1887–1930.
- [28] M. GEIGER, S. SPIGLER, A. JACOT, AND M. WYART, *Disentangling feature and lazy training in deep neural networks*, Journal of Statistical Mechanics: Theory and Experiment, 2020 (2020), p. 113301.
- [29] B. GHORBANI, S. MEI, T. MISIAKIEWICZ, AND A. MONTANARI, *When do neural networks outperform kernel methods?*, Advances in Neural Information Processing Systems, 33 (2020), pp. 14820–14830.
- [30] M. GLASGOW, *Sgd finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the xor problem*, in International Conference on Learning Representations, vol. 2024, 2024, pp. 52419–52430.
- [31] M. GLASGOW AND J. BRUNA, *Uniform-in-time weak propagation-of-chaos in shallow neural networks*, arXiv preprint arXiv:2605.22010, (2026).
- [32] M. GLASGOW, D. WU, AND J. BRUNA, *Mean-field analysis of polynomial-width two-layer neural network beyond finite time horizon*, Proceedings of Machine Learning Research vol. 1 (2025), p. 79.
- [33] S. GOLDT, M. ADVANI, A. M. SAXE, F. KRZAKALA, AND L. ZDEBOROVÁ, *Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup*, Advances in neural information processing systems, 32 (2019).
- [34] K. HAJJAR AND L. CHIZAT, *On the symmetries in the dynamics of wide two-layer neural networks.*, Electronic Research Archive, 31 (2023).
- [35] M. H. HOLMES, *Introduction to perturbation methods*, Springer Science & Business Media, 2012.
- [36] M. HONG, H.-T. WAI, Z. WANG, AND Z. YANG, *A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic*, SIAM Journal on Optimization, 33 (2023), pp. 147–180.
- [37] A. JACOT, F. GABRIEL, AND C. HONGLER, *Neural tangent kernel: Convergence and generalization in neural networks*, Advances in neural information processing systems, 31 (2018).
- [38] A. JACOT, F. GED, B. ŞİMŞEK, C. HONGLER, AND F. GABRIEL, *Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity*, arXiv preprint arXiv:2106.15933, (2021).
- [39] V. R. KONDA AND J. N. TSITSIKLIS, *Convergence rate of linear two-time-scale stochastic approximation*, (2004).
- [40] J. D. LEE, K. OKO, T. SUZUKI, AND D. WU, *Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit*, Advances in Neural Information Processing Systems, 37 (2024), pp. 58716–58756.
- [41] A. MAHANKALI, H. ZHANG, K. DONG, M. GLASGOW, AND T. MA, *Beyond ntk with vanilla gradient descent: A mean-field analysis of neural networks with polynomial width, samples, and time*, Advances in Neural Information Processing Systems, 36 (2023), pp. 57367–57480.
- [42] P. MARION AND R. BERTHIER, *Leveraging the two-timescale regime to demonstrate convergence of neural networks*, Advances in Neural Information Processing Systems, 36 (2023), pp. 64996–65029.
- [43] S. MEI, T. MISIAKIEWICZ, AND A. MONTANARI, *Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit*, in Conference on learning theory, PMLR, 2019, pp. 2388–2464.
- [44] S. MEI, A. MONTANARI, AND P.-M. NGUYEN, *A mean field view of the landscape of two-layer neural networks*, Proceedings of the National Academy of Sciences, 115 (2018), pp. E7665–E7671.
- [45] S. PESME AND N. FLAMMARION, *Saddle-to-saddle dynamics in diagonal linear networks*, Advances in Neural Information Processing Systems, 36 (2023), pp. 7475–7505.
- [46] L. PILLAUD-VIVIEN AND A. SCHERTZER, *Joint learning in the gaussian single index model*, arXiv preprint arXiv:2505.21336, (2025).
- [47] M. REFINETTI, S. GOLDT, F. KRZAKALA, AND L. ZDEBOROVÁ, *Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed*, in International Conference on Machine Learning, PMLR, 2021, pp. 8936–8947.
- [48] G. ROTSKOFF AND E. VANDEN-EIJNDEN, *Trainability and accuracy of artificial neural networks: An interacting particle system approach*, Communications on Pure and Applied Mathematics, 75 (2022), pp. 1889–1935.
- [49] D. SAAD AND S. SOLLA, *Dynamics of on-line gradient descent learning for multilayer neural networks*, Advances in neural information processing systems, 8 (1995).
- [50] A. SAXE, J. MCCLELLAND, AND S. GANGULI, *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*, in Proceedings of the International Conference on Learning Representations 2014, International Conference on Learning Representations 2014, 2014.
- [51] Y. S. TAN AND R. VERSHYNIN, *Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval*, Journal of Machine Learning Research, 24 (2023), pp. 1–47.
- [52] G. TESCHL, *Ordinary differential equations and dynamical systems*, vol. 140, American Mathematical Soc., 2012.
- [53] R. VEIGA, L. STEPHAN, B. LOUREIRO, F. KRZAKALA, AND L. ZDEBOROVÁ, *Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks*, Advances in Neural Information Processing Systems, 35 (2022), pp. 23244–23255.
- [54] S. WOJTOWYTSCH, *On the convergence of gradient descent training for two-layer relu-networks in the mean field regime*, arXiv preprint arXiv:2005.13530, (2020).
- [55] B. WOODWORTH, S. GUNASEKAR, J. D. LEE, E. MOROSHO, P. SAVARESE, I. GOLAN, D. SOUDRY, AND N. SREBRO, *Kernel and rich regimes in overparametrized models*, in Conference on Learning Theory, PMLR, 2020, pp. 3635–3673.

- [56] G. YEHUDAI AND S. OHAD, *Learning a single neuron with gradient methods*, in Conference on Learning Theory, PMLR, 2020, pp. 3756–3786.
- [57] G. YEHUDAI AND O. SHAMIR, *On the power and limitations of random features for understanding neural networks*, Advances in neural information processing systems, 32 (2019).