

---

# Continuous vs. Discrete Optimization of Deep Neural Networks

---

**Omer Elkabetz**  
Tel Aviv University  
omer.elkabetz@cs.tau.ac.il

**Nadav Cohen**  
Tel Aviv University  
cohennadav@cs.tau.ac.il

## Abstract

Existing analyses of optimization in deep learning are either continuous, focusing on (variants of) gradient flow, or discrete, directly treating (variants of) gradient descent. Gradient flow is amenable to theoretical analysis, but is stylized and disregards computational efficiency. The extent to which it represents gradient descent is an open question in the theory of deep learning. The current paper studies this question. Viewing gradient descent as an approximate numerical solution to the initial value problem of gradient flow, we find that the degree of approximation depends on the curvature around the gradient flow trajectory. We then show that over deep neural networks with homogeneous activations, gradient flow trajectories enjoy favorable curvature, suggesting they are well approximated by gradient descent. This finding allows us to translate an analysis of gradient flow over deep linear neural networks into a guarantee that gradient descent efficiently converges to global minimum *almost surely* under random initialization. Experiments suggest that over simple deep neural networks, gradient descent with conventional step size is indeed close to gradient flow. We hypothesize that the theory of gradient flows will unravel mysteries behind deep learning.<sup>1</sup>

## 1 Introduction

The success of deep neural networks is fueled by the mysterious properties of gradient-based optimization, namely, the ability of (variants of) gradient descent to minimize non-convex training objectives while exhibiting tendency towards solutions that generalize well. Vast efforts are being directed at mathematically analyzing this phenomenon, with existing results typically falling into one of two categories: continuous or discrete. Continuous analyses usually focus on gradient flow (or variants thereof), which corresponds to gradient descent (or variants thereof) with infinitesimally small step size. Compared to their discrete (positive step size) counterparts, continuous settings are oftentimes far more amenable to theoretical analysis (*e.g.* they admit use of the theory of differential equations), but on the other hand are stylized, and disregard the critical aspect of computational efficiency (number of steps required for convergence). Works analyzing gradient flow over deep neural networks either accept the latter shortcomings (see for example [49, 4, 46]), or attempt to reproduce part of the results via completely separate analysis of gradient descent (*cf.* [30, 18, 5]). The extent to which gradient flow represents gradient descent is an open question in the theory of deep learning.

The current paper formally studies the foregoing question. Viewing gradient descent as a numerical method for approximately solving the initial value problem corresponding to gradient flow, we turn to the literature on numerical analysis, and invoke a fundamental theorem concerning the approximation error. The theorem implies that in general, the match between gradient descent and gradient flow is determined by the curvature around gradient flow’s trajectory. In particular, the “more convex” the trajectory, *i.e.* the larger the (possibly negative) minimal eigenvalue of the Hessian is around the trajectory, the bet-

---

<sup>1</sup>Due to lack of space, essential portions of this paper were deferred to supplementary material. We refer the reader to [21] for a self-contained version of the text.

ter the match is guaranteed to be.<sup>2</sup> We show that when applied to deep neural networks (fully connected as well as convolutional) with homogeneous activations (*e.g.* linear, rectified linear or leaky rectified linear), gradient flow emanating from near-zero initialization (as commonly employed in practice) follows trajectories that are “roughly convex,” in the sense that the minimal eigenvalue of the Hessian along them is far greater than in arbitrary points in space, particularly towards convergence. This implies that over deep neural networks, gradient descent with moderately small step size may in fact be close to its continuous limit, *i.e.* to gradient flow. We exemplify an application of this finding by translating an analysis of gradient flow over deep linear neural networks into a convergence guarantee for gradient descent. The guarantee we obtain is, to our knowledge, the first to ensure that a conventional gradient-based algorithm optimizing a deep (three or more layer) neural network of fixed (data-independent<sup>3</sup>) size efficiently converges<sup>4</sup> to global minimum *almost surely* under random (data-independent) near-zero initialization.

We corroborate our theoretical analysis through experiments with basic deep learning settings, which demonstrate that reducing the step size of gradient descent often leads to only slight changes in its trajectory. This confirms that, in basic settings, central aspects of deep neural network optimization may indeed be captured by gradient flow. Recent works (*e.g.* [8, 33, 53]) suggest that by appropriately modifying gradient flow it is possible to account for advanced settings as well, including ones with momentum, stochasticity and large step size. Encouraged by these developments, we hypothesize that the vast bodies of knowledge on continuous dynamical systems, and gradient flow in particular (see, *e.g.*, [23, 3]), will pave way to unraveling mysteries behind deep learning.

## 1.1 Contributions

The main contributions of this work are: (i) we conduct the first formal study for the discrepancy between continuous and discrete optimization of deep neural networks; (ii) we demonstrate the use of *generic* mathematical machinery for translating a continuous non-convex convergence result into a discrete one; (iii) to our knowledge, the discrete result we obtain forms the first guarantee of random (data-independent) near-zero initialization *almost surely* leading a conventional gradient-based algorithm optimizing a deep (three or more layer) neural network of fixed (data-independent) size to efficiently converge to global minimum; (iv) the fundamental theorem (from numerical analysis) we employ is seldom used in machine learning contexts and may be of independent interest; and (v) we provide empirical evidence suggesting that gradient descent over simple deep neural networks is often close to gradient flow.

## 2 Preliminaries: Numerical Solution of Initial Value Problems

Let  $d \in \mathbb{N}$ . Given a function  $\mathbf{g} : [0, \infty) \times \mathbb{R}^d \rightarrow \mathbb{R}^d$  (viewed as a time-dependent vector field) and a point  $\boldsymbol{\theta}_s \in \mathbb{R}^d$ , consider the *initial value problem*:

$$\boldsymbol{\theta}(0) = \boldsymbol{\theta}_s \quad , \quad \frac{d}{dt}\boldsymbol{\theta}(t) = \mathbf{g}(t, \boldsymbol{\theta}(t)) \quad \text{for } t \geq 0. \quad (1)$$

The following result — an extension of the well known Picard-Lindelöf Theorem — establishes that local Lipschitz continuity of  $\mathbf{g}(\cdot)$  suffices for ensuring existence and uniqueness of a solution  $\boldsymbol{\theta}(\cdot)$ .

**Theorem 1** (Existence-Uniqueness). *Consider the initial value problem in Equation (1), and suppose  $\mathbf{g}(\cdot)$  is locally Lipschitz continuous. Then, there exists a solution  $\boldsymbol{\theta} : [0, t_e) \rightarrow \mathbb{R}^d$ , where either: (i)  $t_e = \infty$ ; or (ii)  $t_e < \infty$  and  $\lim_{t \nearrow t_e} \|\boldsymbol{\theta}(t)\|_2 = \infty$ . Moreover, the solution is unique in the sense that any other solution  $\boldsymbol{\theta}' : [0, t'_e) \rightarrow \mathbb{R}^d$  must satisfy  $t'_e \leq t_e$  and  $\forall t \in [0, t'_e) : \boldsymbol{\theta}'(t) = \boldsymbol{\theta}(t)$ .*

*Proof.* The theorem is a direct consequence of the results in Section 1.5 of [25].<sup>5</sup> □

It is typically the case that the solution to Equation (1) cannot be expressed in closed form, and a numerical approximation is sought after. Various numerical methods for approximately solving initial value problems have been developed over the years (see Chapter 12 in [55] for an introduction). The most basic one, *Euler’s method*, is parameterized by a *step size*  $\eta > 0$ , and when applied to Equation (1) follows the recursive scheme:

<sup>2</sup>In addition to the minimal eigenvalue of the Hessian, local smoothness and Lipschitz constants also affect the guaranteed match between gradient descent and gradient flow. However, the impact of these constants is exponentially weaker than that of the Hessian’s minimal eigenvalue. For details see Theorem 3.

<sup>3</sup>By data-independence we mean that no assumptions on training data are made beyond it being subject to standard whitening and normalization procedures.

<sup>4</sup>We regard convergence as efficient if its computational complexity is polynomial in training set size and dimensions, as well as the desired level of accuracy.

<sup>5</sup>A minor subtlety is that in [25] the vector field  $\mathbf{g}(\cdot)$  is defined over an open domain. To account for this requirement, simply extend  $\mathbf{g}(\cdot)$  to the domain  $(-\infty, \infty) \times \mathbb{R}^d$  by setting  $\mathbf{g}(t, \mathbf{q}) = \mathbf{g}(0, \mathbf{q})$  for all  $t < 0$ ,  $\mathbf{q} \in \mathbb{R}^d$ .

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta \mathbf{g}(t_k, \boldsymbol{\theta}_k) \text{ for } k=0,1,2,\dots, \quad (2)$$

where  $t_k := k\eta$  and the initial point  $\boldsymbol{\theta}_0$  is typically set to  $\boldsymbol{\theta}_s$ . The motivation behind Euler's method is straightforward — a first order Taylor expansion of the exact solution  $\boldsymbol{\theta}(\cdot)$  around time  $t_k$  yields:  $\boldsymbol{\theta}(t_{k+1}) = \boldsymbol{\theta}(t_k + \eta) \approx \boldsymbol{\theta}(t_k) + \eta \frac{d}{dt} \boldsymbol{\theta}(t_k) = \boldsymbol{\theta}(t_k) + \eta \mathbf{g}(t_k, \boldsymbol{\theta}(t_k))$ , therefore if  $\boldsymbol{\theta}(t_k)$  is well approximated by  $\boldsymbol{\theta}_k$ , we may expect  $\boldsymbol{\theta}_{k+1}$  to resemble  $\boldsymbol{\theta}(t_{k+1})$ . The numerical solution produced by Euler's method may be viewed as a continuous polygonal curve:

$$\bar{\boldsymbol{\theta}}: [0, \infty) \rightarrow \mathbb{R}^d, \quad \bar{\boldsymbol{\theta}}(0) = \boldsymbol{\theta}_0, \quad \frac{d}{dt} \bar{\boldsymbol{\theta}}(t) = \mathbf{g}(t_k, \boldsymbol{\theta}_k) \text{ for } t \in (t_k, t_{k+1}), k=0,1,2,\dots \quad (3)$$

The quality of the numerical solution then boils down to the distance between this curve and the exact solution, *i.e.* between  $\bar{\boldsymbol{\theta}}(t)$  and  $\boldsymbol{\theta}(t)$  for  $t \geq 0$ . Many efforts have been made to derive tight bounds for this distance. We provide below a modern result known as ‘‘Fundamental Theorem.’’

**Theorem 2 (Fundamental Theorem).** *Consider the initial value problem in Equation (1), and suppose  $\mathbf{g}(\cdot)$  is continuously differentiable. Let  $\boldsymbol{\theta}: [0, t_e) \rightarrow \mathbb{R}^d$  be the solution to this problem (see Theorem 1), and let  $\bar{\boldsymbol{\theta}}: [0, \infty) \rightarrow \mathbb{R}^d$  be a continuous polygonal curve (Equation (3)) born from Euler's method (Equation (2)). For any  $t \in [0, t_e)$ ,  $\mathbf{q} \in \mathbb{R}^d$ , denote by  $J(t, \mathbf{q}) \in \mathbb{R}^{d,d}$  the Jacobian of  $\mathbf{g}(\cdot)$  with respect to its second argument at the point  $(t, \mathbf{q})$ , and by  $\lambda_{\max}(t, \mathbf{q})$  the maximal eigenvalue of  $\frac{1}{2}(J(t, \mathbf{q}) + J(t, \mathbf{q})^\top)$ .<sup>6</sup> Let  $m: [0, t_e) \rightarrow \mathbb{R}$  be an integrable function satisfying:  $\lambda_{\max}(t, \mathbf{q}) \leq m(t)$  for all  $t \in [0, t_e)$  and  $\mathbf{q} \in [\boldsymbol{\theta}(t), \bar{\boldsymbol{\theta}}(t)]$ , where  $[\boldsymbol{\theta}(t), \bar{\boldsymbol{\theta}}(t)]$  stands for the line segment (in  $\mathbb{R}^d$ ) between  $\boldsymbol{\theta}(t)$  and  $\bar{\boldsymbol{\theta}}(t)$ . Let  $\delta: [0, t_e) \rightarrow \mathbb{R}_{\geq 0}$  be an integrable function that meets:  $\|\frac{d}{dt} \bar{\boldsymbol{\theta}}(t^+) - \mathbf{g}(t, \bar{\boldsymbol{\theta}}(t))\|_2 \leq \delta(t)$  for all  $t \in [0, t_e)$ , where  $\frac{d}{dt} \bar{\boldsymbol{\theta}}(t^+)$  represents the right derivative of  $\bar{\boldsymbol{\theta}}(\cdot)$  at time  $t$ . Then, for all  $t \in [0, t_e)$ :*

$$\|\boldsymbol{\theta}(t) - \bar{\boldsymbol{\theta}}(t)\|_2 \leq e^{\mu(t)} \left( \|\boldsymbol{\theta}(0) - \bar{\boldsymbol{\theta}}(0)\|_2 + \int_0^t e^{-\mu(t')} \delta(t') dt' \right), \quad (4)$$

where  $\mu(t) := \int_0^t m(t') dt'$ .

*Proof.* The theorem is simply a restatement of Theorem 10.6 in [27].  $\square$

### 3 Continuous vs. Discrete Optimization: Match Determined by Convexity

Let  $f: \mathbb{R}^d \rightarrow \mathbb{R}$ , where  $d \in \mathbb{N}$ , be a twice continuously differentiable function which we would like to minimize. Consider continuous optimization via *gradient flow* initialized at  $\boldsymbol{\theta}_s \in \mathbb{R}^d$ :

$$\boldsymbol{\theta}(0) = \boldsymbol{\theta}_s, \quad \frac{d}{dt} \boldsymbol{\theta}(t) = -\nabla f(\boldsymbol{\theta}(t)) \text{ for } t \geq 0. \quad (5)$$

This is a special case of the initial value problem presented in Equation (1).<sup>7</sup> By Theorem 1, it admits a unique solution  $\boldsymbol{\theta}: [0, t_e) \rightarrow \mathbb{R}^d$ , where either: (i)  $t_e = \infty$ ; or (ii)  $t_e < \infty$  and  $\lim_{t \nearrow t_e} \|\boldsymbol{\theta}(t)\|_2 = \infty$ . Numerically approximating this solution via Euler's method (Equation (2)) yields a discrete optimization algorithm which is no other than *gradient descent*:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla f(\boldsymbol{\theta}_k) \text{ for } k=0,1,2,\dots, \quad (6)$$

where  $\eta > 0$  is the chosen step size. We may thus invoke the Fundamental Theorem (Theorem 2) and obtain a bound on the distance between the trajectories of gradient flow and gradient descent.

**Theorem 3.** *Consider the trajectory of gradient flow (solution to Equation (5))  $\boldsymbol{\theta}: [0, t_e) \rightarrow \mathbb{R}^d$ , and let  $\tilde{t} \in (0, t_e)$  and  $\epsilon > 0$ . Define  $\mathcal{D}_{\tilde{t}, \epsilon} := \bigcup_{t \in [0, \tilde{t}]} \mathcal{B}_\epsilon(\boldsymbol{\theta}(t))$ , where  $\mathcal{B}_\epsilon(\boldsymbol{\theta}(t)) \subset \mathbb{R}^d$  stands for the (closed) Euclidean ball of radius  $\epsilon$  centered at  $\boldsymbol{\theta}(t)$ . Let  $\beta_{\tilde{t}, \epsilon}, \gamma_{\tilde{t}, \epsilon} > 0$  be such that:  $\sup_{\mathbf{q} \in \mathcal{D}_{\tilde{t}, \epsilon}} \|\nabla^2 f(\mathbf{q})\|_{\text{spectral}} \leq \beta_{\tilde{t}, \epsilon}$  and  $\sup_{\mathbf{q} \in \mathcal{D}_{\tilde{t}, \epsilon}} \|\nabla f(\mathbf{q})\|_2 \leq \gamma_{\tilde{t}, \epsilon}$ . Let  $m: [0, \tilde{t}] \rightarrow \mathbb{R}$  be an integrable function satisfying:  $-\lambda_{\min}(\nabla^2 f(\mathbf{q})) \leq m(t)$  for all  $t \in [0, \tilde{t}]$  and  $\mathbf{q} \in \mathcal{B}_\epsilon(\boldsymbol{\theta}(t))$ , where  $\lambda_{\min}(\nabla^2 f(\mathbf{q}))$  stands for the minimal eigenvalue of  $\nabla^2 f(\mathbf{q})$ . Then, if the step size  $\eta > 0$  chosen for gradient descent (Equation (6)) satisfies:*

$$\eta < \inf_{t \in (0, \tilde{t}]} \frac{\epsilon - e^{\int_0^t m(t') dt'} \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}(0)\|_2}{\beta_{\tilde{t}, \epsilon} \gamma_{\tilde{t}, \epsilon} \int_0^t e^{\int_r^t m(t'') dt''} dt'}, \quad (7)$$

*the first  $\lfloor \tilde{t}/\eta \rfloor$  iterates of gradient descent will  $\epsilon$ -approximate the trajectory of gradient flow up to time  $\tilde{t}$ , *i.e.* we will have  $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}(k\eta)\|_2 \leq \epsilon$  for all  $k \in \{1, 2, \dots, \lfloor \tilde{t}/\eta \rfloor\}$ .*

*Proof sketch (for complete proof see Subappendix J.2).* The result follows from applying the Fundamental Theorem (Theorem 2) with  $\delta(\cdot)$  fixed at  $\beta_{\tilde{t}, \epsilon} \gamma_{\tilde{t}, \epsilon} \eta$ .  $\square$

<sup>6</sup>This maximal eigenvalue is known as the *logarithmic norm* of  $J(t, \mathbf{q})$  (cf. Section I.10 in [27]).

<sup>7</sup>The vector field in this case is time-independent (given by  $\mathbf{g}(t, \mathbf{q}) = -\nabla f(\mathbf{q})$  for all  $t \in [0, \infty)$ ,  $\mathbf{q} \in \mathbb{R}^d$ ). Initial value problems of this type are known as *autonomous*.

Theorem 3 gives a sufficient condition — upper bound on step size  $\eta$  (Equation (7)) — for gradient descent to follow gradient flow up to a given time  $\tilde{t}$ . The bound is inversely proportional to smoothness and Lipschitz constants ( $\beta_{\tilde{t},\epsilon}$  and  $\gamma_{\tilde{t},\epsilon}$  respectively), and more importantly, depends exponentially on the integral of  $m(\cdot)$  along the gradient flow trajectory, where  $m(\cdot)$  corresponds to minus the minimal eigenvalue of the Hessian. The smaller the integral of  $m(\cdot)$ , *i.e.* the larger (less negative or more positive) the minimal eigenvalue of the Hessian around the trajectory is, the more relaxed the bound will be. That is, *the “more convex” the objective function is around the gradient flow trajectory, the better the match between gradient flow and gradient descent is guaranteed to be.*

Corollary 1 below coarsely applies Theorem 3 by fixing  $m(\cdot)$  to minus the minimal eigenvalue of the Hessian *across the entire space*. If  $m(\cdot) \equiv m$  (now a constant) is negative, *i.e.* the objective function  $f(\cdot)$  is strongly convex, the upper bound on the step size  $\eta$  becomes constant, meaning it is independent of the time  $\tilde{t}$  until which gradient descent is required to follow gradient flow. If  $m$  is equal to zero, *i.e.*  $f(\cdot)$  is non-strongly convex, the upper bound on  $\eta$  mildly decreases with  $\tilde{t}$ , namely it scales as  $1/\tilde{t}$ . If on the other hand  $m$  is positive, meaning  $f(\cdot)$  is non-convex, the bound on  $\eta$  shrinks to zero (becoming prohibitively restrictive) exponentially fast as  $\tilde{t}$  grows. This suggests that as opposed to (strongly or non-strongly) convex objectives, over which gradient descent can easily be made to follow gradient flow, over non-convex objectives, in the worst case, gradient descent will immediately divert from gradient flow unless its step size is exponentially small. In Appendix B we present a simple example of such a worst case scenario. In this worst case, the minimal eigenvalue of the Hessian is bounded below and away from zero around the gradient flow trajectory. A question is then whether there are non-convex objectives in which the minimal eigenvalue of the Hessian around gradient flow trajectories is large enough for them to be followed by gradient descent. We will see that training losses of deep neural networks can meet this property.

**Corollary 1.** *Assume that the objective function  $f(\cdot)$  is non-negative and  $\beta$ -smooth with  $\beta > 0$ .<sup>8</sup> Denote  $m := -\inf_{\mathbf{q} \in \mathbb{R}^d} \lambda_{\min}(\nabla^2 f(\mathbf{q}))$ , where  $\lambda_{\min}(\nabla^2 f(\mathbf{q}))$  stands for the minimal eigenvalue of  $\nabla^2 f(\mathbf{q})$ . Consider the trajectory of gradient flow (solution to Equation (5))  $\boldsymbol{\theta} : [0, t_e) \rightarrow \mathbb{R}^d$ ,<sup>9</sup> and let  $\tilde{t} \in (0, t_e)$  and  $\epsilon > 0$ . Then, if the step size  $\eta > 0$  for gradient descent (Equation (6)) satisfies:*

$$\eta < \begin{cases} c(\epsilon - \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}(0)\|_2) |m| & , \text{if } m < 0 \quad (\text{strong convexity}) \\ c(\epsilon - \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}(0)\|_2) (1/\tilde{t}) & , \text{if } m = 0 \quad (\text{non-strong convexity}) \\ c(\epsilon - \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}(0)\|_2) e^{m\tilde{t}} (e^{m\tilde{t}} - 1)^{-1} m & , \text{if } m > 0 \quad (\text{non-convexity}) \end{cases} ,$$

where  $c := (\sqrt{2\beta^3 f(\boldsymbol{\theta}(0))} + \beta^2 \epsilon)^{-1}$ , we will have  $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}(k\eta)\|_2 \leq \epsilon$  for all  $k \in \{1, 2, \dots, \lfloor \tilde{t}/\eta \rfloor\}$ .

*Proof sketch (for complete proof see Subappendix J.3).* The result follows from applying Theorem 3 with  $\beta_{\tilde{t},\epsilon} = \beta$ ,  $\gamma_{\tilde{t},\epsilon} = \sqrt{2\beta f(\boldsymbol{\theta}(0))} + \beta\epsilon$  and  $m(\cdot) \equiv m$ .  $\square$

## 4 Optimization of Deep Neural Networks is Roughly Convex

Section 3 has shown that the extent to which gradient descent matches gradient flow depends on “how convex” the objective function is around the gradient flow trajectory. More precisely, the larger (less negative or more positive) the minimal eigenvalue of the Hessian is around this trajectory, the longer gradient descent (with given step size) is guaranteed to follow it.<sup>2</sup> In this section we establish that over training losses of deep neural networks (fully connected as well as convolutional) with homogeneous activations (*e.g.* linear, rectified linear or leaky rectified linear), when emanating from near-zero initialization (as commonly employed in practice), trajectories of gradient flow are “roughly convex,” in the sense that the minimal eigenvalue of the Hessian along them is far greater than in arbitrary points in space, particularly towards convergence. This finding suggests that when optimizing deep neural networks, gradient descent may closely resemble gradient flow. We demonstrate a formal application of the finding in Section 5, translating an analysis of gradient flow over deep linear neural networks into a guarantee of efficient convergence (to global minimum) for gradient descent, which applies *almost surely* with respect to a random near-zero initialization.

<sup>8</sup>Namely,  $\|\nabla^2 f(\mathbf{q})\|_{\text{spectral}} \leq \beta$  for all  $\mathbf{q} \in \mathbb{R}^d$ .

<sup>9</sup>Lemma 3 in Appendix A shows that in the current context ( $\beta$ -smoothness of the objective function  $f(\cdot)$ ), it necessarily holds that  $t_e = \infty$ , *i.e.* the trajectory of gradient flow is defined over  $[0, \infty)$ . For simplicity, the statement of the corollary does not rely on this fact.

#### 4.1 Fully Connected Architectures

Consider the mappings realized by a fully connected neural network with depth  $n \in \mathbb{N}_{\geq 2}$ , input dimension  $d_0 \in \mathbb{N}$ , hidden widths  $d_1, d_2, \dots, d_{n-1} \in \mathbb{N}$ , and output dimension  $d_n \in \mathbb{N}$ :

$$h_{\theta} : \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_n}, \quad h_{\theta}(\mathbf{x}) = W_n \sigma(W_{n-1} \sigma(W_{n-2} \cdots \sigma(W_1 \mathbf{x}) \cdots)), \quad (8)$$

where:  $W_j \in \mathbb{R}^{d_j, d_{j-1}}$ ,  $j = 1, 2, \dots, n$ , are learned weight matrices;  $\theta \in \mathbb{R}^d$ , with  $d := \sum_{j=1}^n d_j d_{j-1}$ , is their arrangement as a vector;<sup>10</sup> and  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  is a predetermined activation function that operates element-wise when applied to a vector.<sup>11</sup> We assume that  $\sigma(\cdot)$  is (positively) *homogeneous*, meaning  $\sigma(cz) = c\sigma(z)$  for all  $c \geq 0, z \in \mathbb{R}$ . This allows for linear ( $\sigma(z) = z$ ), as well as the commonly employed rectified linear ( $\sigma(z) = \max\{z, 0\}$ ) and leaky rectified linear ( $\sigma(z) = \max\{z, \bar{\alpha}z\}$  for some  $0 < \bar{\alpha} < 1$ ) activations.

Let  $\mathcal{Y}$  be a set of possible labels, and let  $\mathcal{S} = ((\mathbf{x}_i, y_i))_{i=1}^{|\mathcal{S}|}$ , with  $\mathbf{x}_i \in \mathbb{R}^{d_0}, y_i \in \mathcal{Y}$  for  $i = 1, 2, \dots, |\mathcal{S}|$ , be a sequence of labeled inputs. Given a loss function  $\ell : \mathbb{R}^{d_n} \times \mathcal{Y} \rightarrow \mathbb{R}$  convex and twice continuously differentiable in its first argument (common choices include square, logistic and exponential losses), we learn the weights of the neural network by minimizing its *training loss* — average loss over elements of  $\mathcal{S}$ :

$$f : \mathbb{R}^d \rightarrow \mathbb{R}, \quad f(\theta) = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell(h_{\theta}(\mathbf{x}_i), y_i). \quad (9)$$

Subsections 4.1.1 and 4.1.2 below show (for linear and non-linear activation functions, respectively) that although the minimal eigenvalue of  $\nabla^2 f(\theta)$  (Hessian of training loss) — denoted  $\lambda_{\min}(\nabla^2 f(\theta))$  — can in general be arbitrarily negative, along trajectories of gradient flow (which emanate from near-zero initialization) it is no less than moderately negative, approaching non-negativity towards convergence. In light of Section 3, this suggests that over fully connected deep neural networks, gradient flow may lend itself to approximation by gradient descent — a prospect we confirm (for a case with linear activation) in Section 5.

##### 4.1.1 Linear Activation

Assume that the activation function of the fully connected neural network (Equation (8)) is linear, i.e.  $\sigma(z) = z$ , and define the *end-to-end matrix*:

$$W_{n:1} := W_n W_{n-1} \cdots W_1 \in \mathbb{R}^{d_n, d_0}. \quad (10)$$

The mappings realized by the network can then be written as  $h_{\theta}(\mathbf{x}) = W_{n:1} \mathbf{x}$ , and the training loss as  $f(\theta) = \phi(W_{n:1})$ , where

$$\phi : \mathbb{R}^{d_n, d_0} \rightarrow \mathbb{R}, \quad \phi(W) = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell(W \mathbf{x}_i, y_i) \quad (11)$$

is convex and twice continuously differentiable. Lemma 1 below expresses  $\nabla^2 f(\theta)$  in this case.

**Lemma 1.** *For any  $\theta \in \mathbb{R}^d$ , regard  $\nabla^2 f(\theta)$  not only as a (symmetric) matrix in  $\mathbb{R}^{d,d}$ , but also as a quadratic form  $\nabla^2 f(\theta)[\cdot]$  that intakes a tuple  $(\Delta W_1, \Delta W_2, \dots, \Delta W_n) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$ , arranges it as a vector  $\Delta \theta \in \mathbb{R}^d$  (in correspondence with how weight matrices  $W_1, W_2, \dots, W_n$  are arranged to create  $\theta$ ), and returns  $\Delta \theta^\top \nabla^2 f(\theta) \Delta \theta \in \mathbb{R}$ . Similarly, for any  $W \in \mathbb{R}^{d_n, d_0}$ , regard  $\nabla^2 \phi(W)$  as a quadratic form  $\nabla^2 \phi(W)[\cdot]$  that intakes a matrix in  $\mathbb{R}^{d_n, d_0}$  and returns a scalar (non-negative since  $\phi(\cdot)$  is convex). Then,  $\nabla^2 f(\theta)$  is given by:*

$$\begin{aligned} \nabla^2 f(\theta)[\Delta W_1, \Delta W_2, \dots, \Delta W_n] &= \nabla^2 \phi(W_{n:1}) \left[ \sum_{j=1}^n W_{n:j+1} (\Delta W_j) W_{j-1:1} \right] \\ &\quad + 2 \text{Tr} \left( \nabla \phi(W_{n:1})^\top \sum_{1 \leq j < j' \leq n} W_{n:j'+1} (\Delta W_{j'}) W_{j'-1:j+1} (\Delta W_j) W_{j-1:1} \right), \end{aligned} \quad (12)$$

where  $W_{j':j}$ , for any  $j, j' \in \{1, 2, \dots, n\}$ , is defined as  $W_{j'} W_{j'-1} \cdots W_j$  if  $j \leq j'$ , and as an identity matrix (with size to be inferred by context) otherwise.

*Proof.* See Subappendix J.4. □

The following proposition makes use of Lemma 1 to show that (under mild conditions)  $\lambda_{\min}(\nabla^2 f(\theta))$  can be arbitrarily negative, i.e.  $\inf_{\theta \in \mathbb{R}^d} \lambda_{\min}(\nabla^2 f(\theta)) = -\infty$ .

**Proposition 1.** *Assume that the network is deep ( $n \geq 3$ ), and that the zero mapping is not a global minimizer of the training loss (meaning  $\nabla \phi(0) \neq 0$ ).<sup>12</sup> Then  $\inf_{\theta \in \mathbb{R}^d} \lambda_{\min}(\nabla^2 f(\theta)) = -\infty$ .*

<sup>10</sup>The exact order by which the entries of  $W_1, W_2, \dots, W_n$  are placed in  $\theta$  is insignificant for our purposes — all that matters is that the same order be used throughout.

<sup>11</sup>Our analysis can easily be extended to account for different activation functions at different hidden layers. We assume identical activation functions for simplicity of presentation.

<sup>12</sup>Both of these assumptions are necessary, in the sense that removing any of them (without imposing further assumptions) renders the proposition false — see Claim 1 in Appendix F.

*Proof.* See Subappendix J.5. □

Building on Lemma 1, Lemma 2 below provides a lower bound on  $\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta}))$ .

**Lemma 2.** For any  $\boldsymbol{\theta} \in \mathbb{R}^d$ :<sup>13</sup>

$$\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -(n-1)\sqrt{\min\{d_0, d_n\}} \|\nabla\phi(W_{n:1})\|_{Frobenius} \max_{\substack{\mathcal{J} \subseteq \{1, 2, \dots, n\} \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_{spectral}. \quad (13)$$

*Proof.* See Subappendix J.6. □

Assuming the training loss is non-constant and the network is deep ( $n \geq 3$ ), the infimum (over  $\boldsymbol{\theta} \in \mathbb{R}^d$ ) of the lower bound in Equation (13) is minus infinity. In particular, if  $\boldsymbol{\theta}$  is not a global minimizer ( $\nabla\phi(W_{n:1}) \neq 0$ ) and at least  $n-2$  of its weight matrices  $W_1, W_2, \dots, W_n$  are non-zero, then by rescaling the latter it is possible to take the lower bound to minus infinity while keeping the end-to-end matrix  $W_{n:1}$  (and thus the input-output mapping  $h_{\boldsymbol{\theta}}(\cdot)$  and the training loss value  $f(\boldsymbol{\theta})$ ) intact. However, gradient flow over fully connected neural networks (with homogeneous activations) initialized near zero is known to maintain balance between weight matrices — see [18] — and so along its trajectories the lower bound in Equation (13) takes a much tighter form. This is formalized in Proposition 2 below.

**Proposition 2.** If  $\boldsymbol{\theta} \in \mathbb{R}^d$  resides on a trajectory of gradient flow (over  $f(\cdot)$ ) emanating from some point  $\boldsymbol{\theta}_s \in \mathbb{R}^d$ , with  $\|\boldsymbol{\theta}_s\|_2 \leq \epsilon$  for some  $\epsilon \in (0, \frac{1}{2n}]$ , then:

$$\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -(n-1)\sqrt{\min\{d_0, d_n\}} \|\nabla\phi(W_{n:1})\|_{Frobenius} \|W_{n:1}\|_{spectral}^{1-2/n} - c\epsilon^{1-2/n}, \quad (14)$$

where  $c := \frac{4n(n-1)}{(4n)^{2/n}} \sqrt{\min\{d_0, d_n\}} \|\nabla\phi(W_{n:1})\|_{Frobenius} \max\{1, \max\{\|W_j\|_{spectral}\}_{j=1}^n\}^{2(n-2)}$ .

*Proof.* See Subappendix J.7. □

Assume the network is deep ( $n \geq 3$ ), and consider a trajectory of gradient flow (over  $f(\cdot)$ ) emanating from near-zero initialization. For every point on the trajectory, Proposition 2 may be applied with small  $\epsilon$ , leading the lower bound in Equation (14) to depend primarily on the sizes (norms) of the end-to-end matrix  $W_{n:1}$  and the gradient of the loss with respect to it, *i.e.*  $\nabla\phi(W_{n:1})$  (see Equations (10) and (11)). In the course of optimization,  $W_{n:1}$  is initially small, and (since the loss  $f(\boldsymbol{\theta}) = \phi(W_{n:1})$  is monotonically non-increasing) remains confined to sublevel sets of  $\phi(\cdot)$  (which is convex) thereafter.  $\nabla\phi(W_{n:1})$  on the other hand tends to zero upon convergence to global minimum. We conclude that the lower bound on  $\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta}))$  in Equation (14) starts off slightly negative, and approaches non-negativity (if and) as the trajectory converges to global minimum. In light of Section 3, this implies that the gradient flow trajectory may lend itself to approximation by gradient descent. Indeed, the results of the current Subsubsection are used in Section 5 to establish proximity between gradient flow and gradient descent, thereby translating an analysis of gradient flow into a guarantee of efficient convergence (to global minimum) for gradient descent.

#### 4.1.2 Non-Linear Activation

Due to lack of space, we defer our analysis for fully connected neural networks with non-linear activation to Appendix C. This analysis is similar in spirit to the one in Subsubsection 4.1.1 treating linear activation. In particular, it makes use of the fact that gradient flow initialized near zero maintains balance between weight matrices — *cf.* [18]. A key difference brought forth by non-linear activation is that the training loss  $f(\cdot)$  (Equation (9)) is no longer differentiable. We circumvent this challenge by excluding from the analysis points of non-differentiability, which form a negligible (closed and zero measure) set.

## 4.2 Convolutional Architectures

We account for convolutional neural networks by allowing for weight sharing and sparsity patterns to be imposed on the layers of the fully connected model analyzed in Subsection 4.1. Namely, we consider the exact same mappings as in Equation (8), but now, rather than being learned directly, the matrices  $W_j \in \mathbb{R}^{d_j, d_{j-1}}$ ,  $j = 1, 2, \dots, n$ , are determined by learned weight vectors  $\mathbf{w}_j \in \mathbb{R}^{d'_j}$ , with  $d'_j \in \mathbb{N}$ ,  $j = 1, 2, \dots, n$ , such that each entry of  $W_j$  is either fixed at zero or connected to a predetermined coordinate of  $\mathbf{w}_j$  (with no repetition of coordinates within the same row). The weight setting  $\boldsymbol{\theta} \in \mathbb{R}^d$  is then simply a concatenation of the weight vectors  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$ , and its dimension is accordingly  $d = \sum_{j=1}^n d'_j$ . Our analysis for this model (which includes convolutional neural networks as a special case) is essentially the same as that presented for fully connected neural networks with non-linear activation (Subsubsection 4.1.2). In particular, we use the fact that even with weight sharing and sparsity patterns imposed on the layers of a fully connected neural network (with homogeneous activation), when initialized near zero, gradient flow over the network maintains balance between weights of different layers — *cf.* [18]. For the complete analysis see Appendix D.

<sup>13</sup>Note that by convention, an empty product (*i.e.* a product over the elements of the empty set) is equal to one.

## 5 Continuous Proof of Discrete Convergence for Deep Linear Neural Networks

Section 3 invoked the Fundamental Theorem for numerical solution of initial value problems (Theorem 2) to show that, in general, the extent to which gradient descent provably matches gradient flow is determined by how large (less negative or more positive) the minimal eigenvalue of the Hessian is around the gradient flow trajectory.<sup>2</sup> Section 4 established that for training losses of deep neural networks, along trajectories of gradient flow emanating from near-zero initialization (as commonly employed in practice), the minimal eigenvalue of the Hessian is far greater than in arbitrary points in space, particularly towards convergence. In this section we combine the two findings, translating an analysis of gradient flow over deep linear neural networks into a convergence guarantee for gradient descent. The guarantee we obtain is, to our knowledge, the first to ensure that a conventional gradient-based algorithm optimizing a deep (three or more layer) neural network of fixed (data-independent<sup>3</sup>) size efficiently converges<sup>4</sup> to global minimum *almost surely* under random (data-independent) near-zero initialization.

Deep linear neural networks — fully connected neural networks with linear activation (see Subsection 4.1) — are perhaps the most common subject of theoretical study in the context of optimization in deep learning. Though trivial from an expressiveness point of view (realize only linear input-output mappings), they induce highly non-convex training losses, giving rise to highly non-trivial phenomena under gradient-based optimization. In recent years, various results concerning gradient flow over deep linear neural networks have been proven, most notably for the case of *balanced initialization* (see for example [49, 4, 34, 6, 46]). Under the notations of Subsection 4.1 (in particular with  $W_1, W_2, \dots, W_n$  standing for network weight matrices), balanced initialization means that when optimization commences:

$$W_{j+1}^\top W_{j+1} = W_j W_j^\top \text{ for } j = 1, 2, \dots, n-1. \quad (15)$$

The condition holds approximately with any near-zero initialization, and exactly when the following procedure (adaptation of Procedure 1 in [5]) is employed.

**Procedure 1** (random balanced initialization). *With a distribution  $\mathcal{P}$  over  $d_n$ -by- $d_0$  matrices of rank at most  $\min\{d_0, d_1, \dots, d_n\}$ , initialize  $W_j \in \mathbb{R}^{d_j, d_{j-1}}$ ,  $j = 1, 2, \dots, n$ , via following steps: (i) sample  $A \sim \mathcal{P}$ ; (ii) take singular value decomposition  $A = U \Sigma V^\top$ , where  $U \in \mathbb{R}^{d_n, \min\{d_0, d_n\}}$  and  $V \in \mathbb{R}^{d_0, \min\{d_0, d_n\}}$  have orthonormal columns, and  $\Sigma \in \mathbb{R}^{\min\{d_0, d_n\}, \min\{d_0, d_n\}}$  is diagonal and holds the singular values of  $A$ ; and (iii) set  $W_n \simeq U \Sigma^{1/n}$ ,  $W_{n-1} \simeq \Sigma^{1/n}$ ,  $W_{n-2} \simeq \Sigma^{1/n}$ ,  $\dots$ ,  $W_2 \simeq \Sigma^{1/n}$ ,  $W_1 \simeq \Sigma^{1/n} V^\top$ , where “ $\simeq$ ” stands for equality up to zero-valued padding.*

Compared to gradient flow, little is known about gradient descent when it comes to optimization of deep (three or more layer) linear neural networks. Indeed, there are relatively few results along this line (cf. [9, 30, 5]), and these are typically highly specific, built upon technical proofs that are difficult to generalize. Being able to obtain results via translation of gradient flow analyses is thus of prime interest.

We focus in this section on deep<sup>14</sup> linear neural networks trained for scalar regression per least-squares criterion. In the context of Subsection 4.1, this means that the activation function  $\sigma(\cdot)$  is linear ( $\sigma(z) = z$ ), the output dimension  $d_n$  is one, and the loss function  $\ell(\cdot)$  is the square loss (i.e.  $\mathcal{Y} = \mathbb{R}$  and  $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$ ). We assume that training inputs are whitened, i.e. have been transformed such that their empirical (uncentered) covariance matrix  $\Lambda_{xx} := \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \mathbf{x}_i \mathbf{x}_i^\top \in \mathbb{R}^{d_0, d_0}$  is equal to identity. A standard calculation (see Appendix G) shows that in this case the function  $\phi(\cdot)$  defined by Equation (11) becomes  $\phi(W) = \frac{1}{2} \|W - \Lambda_{yx}\|_{Frobenius}^2 + c$ , where  $\Lambda_{yx} := \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} y_i \mathbf{x}_i^\top \in \mathbb{R}^{1, d_0}$  is the empirical (uncentered) cross-covariance matrix between training labels and inputs, and  $c \in \mathbb{R}$  is a constant (independent of  $W$ ). We may thus write the training loss  $f(\cdot)$  (Equation (9)) as:

$$f(\boldsymbol{\theta}) = \frac{1}{2} \|W_{n:1} - \Lambda_{yx}\|_{Frobenius}^2 + c = \frac{1}{2} \|W_{n:1} - \Lambda_{yx}\|_{Frobenius}^2 + \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}), \quad (16)$$

where  $W_{n:1} \in \mathbb{R}^{1, d_0}$  is the network’s end-to-end matrix (Equation (10)). We disregard the degenerate case where  $\Lambda_{yx} = 0$ , i.e. where the zero mapping attains the global minimum, and assume that training labels are normalized (jointly scaled) such that  $\Lambda_{yx}$  has unit length ( $\|\Lambda_{yx}\|_{Frobenius} = 1$ ).

Proposition 3 below analyzes gradient flow over the training loss in Equation (16). Relying on a known characterization for the dynamics of the end-to-end matrix (cf. [4]), it establishes convergence to global minimum. Moreover, harnessing the results of Section 4, it derives a lower bound on (the integral of) the minimal eigenvalue of the Hessian around the gradient flow trajectory.

<sup>14</sup>Our results apply to shallow (two layer) networks as well. We highlight the deep (three or more layer) setting as it is far less understood (cf. [5]), and arguably more central to deep learning.

**Proposition 3.** Consider minimization of the training loss  $f(\cdot)$  in Equation (16) via gradient flow (Equation (5)) starting from initial point  $\theta_s \in \mathbb{R}^d$  that meets the balancedness condition (Equation (15)). Denote by  $W_{n:1,s}$  the initial value of the end-to-end matrix (Equation (10)), and suppose that  $\|W_{n:1,s}\|_{Frobenius} \in (0, 0.2]$  (initialization is small but non-zero). Assume that  $W_{n:1,s}$  is not antiparallel to  $\Lambda_{yx}$ , i.e.  $\nu := \text{Tr}(\Lambda_{yx}^\top W_{n:1,s}) / (\|\Lambda_{yx}\|_{Frobenius} \|W_{n:1,s}\|_{Frobenius}) \neq -1$ . Then, the trajectory of gradient flow is defined over infinite time, and with  $\theta: [0, \infty) \rightarrow \mathbb{R}^d$  representing this trajectory, for any  $\bar{\epsilon} > 0$ , the following time  $\bar{t}$  satisfies  $f(\theta(\bar{t})) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \leq \bar{\epsilon}$ :

$$\bar{t} = \frac{2n \left( \max\left\{1, \frac{3}{2} \cdot \frac{1-\nu}{1+\nu}\right\} \right)^n}{\|W_{n:1,s}\|_{Frobenius}} \ln \left( \frac{15n \max\left\{1, \frac{1-\nu}{1+\nu}\right\}}{\|W_{n:1,s}\|_{Frobenius} \min\{1, 2\bar{\epsilon}\}} \right). \quad (17)$$

Moreover, under the notations of Theorem 3, for any  $t > 0$  and  $\epsilon \in (0, \frac{1}{2n}]$  with corresponding  $\mathcal{D}_{t,\epsilon}$  ( $\epsilon$ -neighborhood of gradient flow trajectory up to time  $t$ ), we have the smoothness and Lipschitz constants  $\beta_{t,\epsilon} = 16n$  and  $\gamma_{t,\epsilon} = 6\sqrt{n}$  respectively, and the following (upper) bound on the integral of (minus) the minimal eigenvalue of the Hessian:

$$\int_0^t m(t') dt' \leq \frac{15n^3 \left( \max\left\{1, \frac{3}{2} \cdot \frac{1-\nu}{1+\nu}\right\} \right)^n t \epsilon}{\|W_{n:1,s}\|_{Frobenius}} + \ln \left( \frac{n^2 \left( e^2 \max\left\{1, \frac{1-\nu}{1+\nu}\right\} \right)^{5(n-1)/2}}{\|W_{n:1,s}\|_{Frobenius}^2} \right), \quad (18)$$

where the function  $m: [0, t] \rightarrow \mathbb{R}$  is non-negative.

*Proof.* See Subappendix J.8. □

Plugging the gradient flow results of Proposition 3 into the generic Theorem 3 translates them to the following convergence guarantee for gradient descent.

**Theorem 4.** Assume the same conditions as in Proposition 3, but with minimization via gradient descent (Equation (6)) instead of gradient flow.<sup>15</sup> Then, with  $\theta_0, \theta_1, \theta_2, \dots$  representing the iterates of gradient descent,  $W_{n:1,0}$  standing for the end-to-end matrix (Equation (10)) of the initial point  $\theta_0$ , and  $\nu := \text{Tr}(\Lambda_{yx}^\top W_{n:1,0}) / (\|\Lambda_{yx}\|_{Frobenius} \|W_{n:1,0}\|_{Frobenius})$ , for any  $\tilde{\epsilon} > 0$ , if the step size  $\eta$  meets:

$$\eta \leq \frac{\|W_{n:1,0}\|_{Frobenius}^5 \min\{1, \tilde{\epsilon}\}}{n^{17/2} e^{7n+6} \left( \max\left\{1, \frac{1-\nu}{1+\nu}\right\} \right)^{(11n-5)/2}} \left( \ln \left( \frac{15n \max\left\{1, \frac{1-\nu}{1+\nu}\right\}}{\|W_{n:1,0}\|_{Frobenius} \min\{1, \tilde{\epsilon}\}} \right) \right)^{-2} \in \tilde{\Omega} \left( \frac{\|W_{n:1,0}\|_{Frobenius}^5 \tilde{\epsilon}}{n^{17/2} \left( \text{poly}\left(\frac{1-\nu}{1+\nu}\right) \right)^n} \right), \quad (19)$$

it holds that  $f(\theta_k) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \leq \tilde{\epsilon}$ , where:

$$k = \left\lceil \frac{2n \left( \max\left\{1, \frac{3}{2} \cdot \frac{1-\nu}{1+\nu}\right\} \right)^n}{\|W_{n:1,0}\|_{Frobenius} \eta} \ln \left( \frac{15n \max\left\{1, \frac{1-\nu}{1+\nu}\right\}}{\|W_{n:1,0}\|_{Frobenius} \min\{1, \tilde{\epsilon}\}} \right) + 1 \right\rceil \in \tilde{\mathcal{O}} \left( \frac{n \left( \text{poly}\left(\frac{1-\nu}{1+\nu}\right) \right)^n \ln\left(\frac{1}{\tilde{\epsilon}}\right)}{\|W_{n:1,0}\|_{Frobenius} \eta} \right). \quad (20)$$

*Proof.* See Subappendix J.9. □

**Remark 1.** Theorem 3 — our generic tool for translating analyses between gradient flow and gradient descent — allows for the two to be initialized differently. Accordingly, the convergence guarantee of Theorem 4 may be extended to account for initialization which is not perfectly balanced, i.e. which satisfies Equation (15) only approximately. For details see Appendix H.

**Remark 2.** The convergence guarantee of Theorem 4 requires a number of iterates that scales exponentially with network depth ( $n$ ). [51] has proven that under mild conditions, for a deep linear neural network whose input, hidden and output dimensions are all equal to one (i.e., in our notations,  $d_0 = d_1 = \dots = d_n = 1$ ), such exponential dependence on depth is unavoidable. We defer to future work the question of whether this also holds in the context of Theorem 4.

Combining Theorem 4 with random balanced initialization (Procedure 1) yields what is, to our knowledge, the first guarantee of random (data-independent) near-zero initialization *almost surely* leading a conventional gradient-based algorithm optimizing a deep (three or more layer) neural network of fixed (data-independent) size to efficiently converge to global minimum.

**Corollary 2.** Consider minimization of the training loss  $f(\cdot)$  in Equation (16) via gradient descent (Equation (6)) emanating from a random balanced initialization (Procedure 1) whose underlying distribution  $\mathcal{P}$  is continuous and satisfies  $\Pr_{A \sim \mathcal{P}} [\|A\|_{Frobenius} \leq 0.2] = 1$ . Assume  $d_0$  (network input dimension) is greater than one, and let  $W_{n:1,0}$  and  $\nu$  be as defined in Theorem 4. Then, almost surely with respect to (i.e. with probability one over) initialization, for any  $\tilde{\epsilon} > 0$ , if the step size  $\eta$  meets Equation (19), the value of  $f(\cdot)$  after  $k$  iterates will be within  $\tilde{\epsilon}$  from global minimum, where  $k$  is given by Equation (20).

*Proof.* See Subappendix J.10. □

<sup>15</sup>The conditions on  $\theta_s$  in Proposition 3 are now satisfied by the initialization of gradient descent, i.e. by  $\theta_0$ .



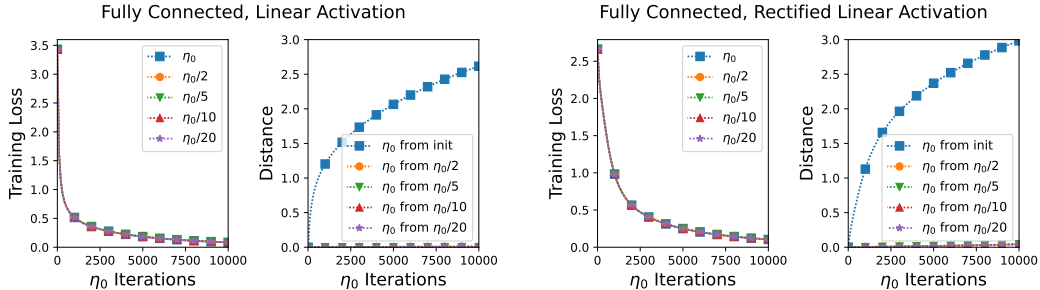


Figure 1: Over deep fully connected neural networks, trajectories of gradient descent with conventional step size barely change when step size is reduced, suggesting they are close to the continuous limit, *i.e.* to trajectories of gradient flow. Presented results were obtained on fully connected neural networks as analyzed in Subsection 4.1, trained to classify MNIST handwritten digits (28-by-28 grayscale images, each labeled as an integer between 0 and 9 — *cf.* [35]). Networks had depth  $n=3$ , input dimension  $d_0=784$  (corresponding to  $28 \cdot 28=784$  pixels), hidden widths  $d_1=d_2=50$  and output dimension  $d_3=10$  (corresponding to ten possible labels). Training was based on gradient descent applied to cross-entropy loss with no regularization, starting from a near-zero point drawn from Xavier distribution (*cf.* [24]). Separately on each network, we compared runs differing only in the step size  $\eta$ . Specifically, with  $\eta_0=0.001$  (standard choice of step size) and  $r$  ranging over  $\{2,5,10,20\}$ , we compared, in terms of training loss value and location in weight space, every iteration of a run using  $\eta=\eta_0$  to every  $r$ 'th iteration of a run in which  $\eta=\eta_0/r$ . Left pair of plots reports results obtained on a network with linear activation ( $\sigma(z)=z$ ), while right pair corresponds to a network with rectified linear activation ( $\sigma(z)=\max\{z,0\}$ ). In each pair, left plot displays training loss values, and right one shows (Euclidean) distances in weight space, namely, distance between initialization and run with  $\eta=\eta_0$ , alongside distances between run with  $\eta=\eta_0$  and runs having  $\eta=\eta_0/r$  for different values of  $r$ . Horizontal axes represent time in units of  $\eta=\eta_0$  iterations (meaning each time unit corresponds to  $r$  iterations of a run with  $\eta=\eta_0/r$ ). Notice that the drift between runs with different step sizes is minor compared to the distance traveled. For further implementation details, and results of similar experiments on convolutional neural networks, see Appendix I.

## 6 Experiments

In this section we corroborate our theory by presenting experiments suggesting that over simple deep neural networks, gradient descent with conventional step size is indeed close to the continuous limit, *i.e.* to gradient flow. Our experimental protocol is simple — on several deep neural networks classifying MNIST handwritten digits ([35]), we compare runs of gradient descent differing only in the step size  $\eta$ . Specifically, separately on each evaluated network, with  $\eta_0=0.001$  (standard choice of step size) and  $r$  ranging over  $\{2,5,10,20\}$ , we compare, in terms of training loss value and location in weight space, every iteration of a run using  $\eta=\eta_0$  to every  $r$ 'th iteration of a run in which  $\eta=\eta_0/r$ . Figure 1 reports the results obtained on fully connected neural networks (as analyzed in Subsection 4.1), with both linear and non-linear activation. As can be seen, reducing the step size  $\eta$  leads to only slight changes, suggesting that the trajectory of gradient descent with  $\eta=\eta_0$  is already close to the continuous limit. Similar results obtained on convolutional neural networks (see Subsection 4.2 for corresponding analysis) are reported by Figure 3 in Subappendix I.1.

Our experimental findings suggest that in practice, proximity between gradient descent and gradient flow may take place even when the step size of gradient descent is larger than permitted by current theory. Indeed, the theoretical machinery developed in this paper brings forth upper bounds on step size that guarantee proximity, and while such upper bounds can be asymptotically tight under worst case conditions (see Appendix B), they are by no means tight in every given scenario, and therefore larger step sizes may also admit proximity. For illustration, a step size of  $\eta_0$ , which in our experiments was seemingly sufficient for ensuring proximity, is many orders of magnitude greater than the upper bound on step size required by Theorem 4 (Equation (19)).

## 7 Related Work

Theoretical study of gradient-based optimization in deep learning is an extremely active area of research. While far too wide to fully cover here, we note that analyses in this area can broadly be categorized as continuous (see for example [49, 4, 34, 6, 1, 20, 57, 46, 31, 47, 60, 7, 62]) or discrete (*e.g.* [9, 26, 17, 2, 16, 66, 28]). There are works comprising analyses of both types (*cf.* [18, 30, 5, 61, 39, 19, 11, 12]), but with these developed separately, wherein continuous proofs typically serve as inspiration for discrete ones (which are often far more technical and brittle).

When relating continuous and discrete optimization, the algorithms at play are most commonly gradient flow and gradient descent. There are however works that draw analogies between other algorithms, replacing gradient flow on the continuous end and/or gradient descent on the discrete one (see, *e.g.*, [54, 58, 59, 45, 50, 37, 52, 63, 22, 43, 40, 8, 33, 53]). The literature includes works which, similarly to the current paper, provide formal results concerning the accumulated (non-local) discrepancy between continuous and discrete optimization (*cf.* [50, 43]). However, such works typically focus on simple objective functions (for example convex or quadratic), whereas we center on (non-convex and non-smooth) training losses of deep neural networks. Several recent works (*e.g.* [8, 33, 14]) also considered continuous vs. discrete optimization of deep neural networks, but they did not provide formal results concerning the accumulated discrepancy. We are not aware of any study (prior to the current) formally quantifying the accumulated discrepancy between continuous and discrete optimization of deep neural networks.

With regards to the convergence guarantee we obtain in Section 5 (via translation of gradient flow analysis to gradient descent) — Theorem 4 and Corollary 2 — relevant results are those that establish efficient convergence<sup>4</sup> to global minimum for a conventional (discrete) gradient-based algorithm optimizing a deep (three or more layer) neural network. Existing results meeting this criterion either: (i) apply to neural networks (linear or non-linear) whose size depends on the data (*i.e.* is not data-independent<sup>3</sup>), predominantly in an impractical fashion (*cf.* [65, 17, 2, 19, 64, 42]); or (ii) apply to linear neural networks of fixed (data-independent) size, similarly to our guarantee. Results of type (ii) often treat the residual setting, which boils down to (possibly scaled) identity initialization, perhaps with input and/or output layers initialized differently (see for example [9, 61, 66]). Exceptions include [5], [16] and [28]. [5] allows for random balanced initialization, as we do. Its results account for networks with multi-dimensional output, and require a number of iterates polynomial in network depth. Our guarantee on the other hand is limited to networks with one-dimensional output, and calls for a number of iterates scaling exponentially with network depth. However, while [5] demands that initialization be sufficiently close to global minimum, thereby excluding the possibility of saddle points being encountered, our guarantee holds *almost surely* (*i.e.* with probability one) under random (data-independent) near-zero initialization. The fact that we account for evasion of saddle points (in particular that at the origin, which is non-strict<sup>16</sup> when network depth is three or more) may be the source of the gap in number of iterates — see Remark 2. As for the results of [16] and [28], these also hold with high probability under random initialization, but they require network size to grow towards infinity in order for the probability to approach one.

## 8 Discussion

Our work puts forth a potential explanation to a puzzling phenomenon in deep learning, namely, the effect of weight decay ( $L_2$  regularization). While traditionally viewed as a regularizer, it is known (*cf.* [32]) that in deep learning, weight decay can assist in minimizing the training loss. In light of our findings, a possible reason for this is that weight decay translates to adding a positive constant to Hessian eigenvalues, thereby bringing gradient descent closer to gradient flow, which often enjoys favorable convergence properties. Theoretical and/or empirical investigation of this prospect is a potential avenue for future work.

Emerging evidence (*cf.* [38, 36, 29]) suggests that for (variants of) gradient descent optimizing deep neural networks, large step size is often beneficial in terms of generalization (*i.e.* in terms of test accuracy). While the large step size regime is not necessarily captured by standard (variants of) gradient flow (see [14]), recent works (*e.g.* [8, 33, 53]) argue that it is captured by a certain modified version of (variants of) gradient flow. Formally quantifying the discrepancy between gradient descent with large step size and such modified version of gradient flow is a promising direction for future research.

The demonstration we provided for translation of a gradient flow analysis to gradient descent (Section 5) culminated in a convergence guarantee, but in fact entails much more information. Namely, since the translated gradient flow analysis includes a careful trajectory characterization, not only do we know that gradient descent converges to global minimum (and how fast that happens), but we also have access to information about the trajectory it takes to get there. This allows, for example, shedding light on how saddle points (non-strict ones in particular<sup>16</sup>) are evaded. A nascent belief (*cf.* [5, 6]) is that understanding the trajectories of gradient descent is key to unraveling mysteries behind optimization and generalization (implicit regularization) in deep learning. The machinery developed in the current paper may contribute to this understanding, by translating results from the vast bodies of literature on continuous dynamical systems.

<sup>16</sup>A saddle point is said to be non-strict if its Hessian has no negative eigenvalues. Saddle points that are non-strict are generally regarded as more difficult to evade — *cf.* [5].

## Acknowledgments and Disclosure of Funding

We thank Sanjeev Arora, Noah Golowich, Wei Hu, Michael Lee, Zhiyuan Li, Kaifeng Lyu, Govind Menon and Zsolt Veraszto for helpful discussions. This work was supported by a Google Research Scholar Award, a Google Research Gift, the Yandex Initiative in Machine Learning, the Israel Science Foundation (grant 1780/21), Len Blavatnik and the Blavatnik Family Foundation, and Amnon and Anat Shashua.

## References

- [1] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.
- [2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via overparameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.
- [3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.
- [4] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pages 244–253, 2018.
- [5] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *International Conference on Learning Representations*, 2019.
- [6] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7413–7424, 2019.
- [7] Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. *arXiv preprint arXiv:2102.09769*, 2021.
- [8] David GT Barrett and Benoit Dherin. Implicit gradient regularization. *International Conference on Learning Representations*, 2021.
- [9] Peter Bartlett, Dave Helmbold, and Phil Long. Gradient descent with identity initialization efficiently learns positive definite linear transformations. In *International Conference on Machine Learning*, pages 520–529, 2018.
- [10] Jonathan Borwein and Adrian S Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010.
- [11] Lenaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.
- [12] Hung-Hsu Chou, Carsten Gieshoff, Johannes Maly, and Holger Rauhut. Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *arXiv preprint arXiv:2011.13772*, 2020.
- [13] Frank H Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.
- [14] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.
- [15] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.
- [16] Simon Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. In *International Conference on Machine Learning*, pages 1655–1664. PMLR, 2019.
- [17] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.
- [18] Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, pages 384–395, 2018.

- [19] Weinan E, Chao Ma, Qingcan Wang, and Lei Wu. Analysis of the gradient descent algorithm for a deep neural network model with skip-connections. *arXiv preprint arXiv:1904.05263*, 2019.
- [20] Armin Eftekhari. Training linear neural networks: Non-local convergence and complexity results. In *International Conference on Machine Learning*, pages 2836–2847. PMLR, 2020.
- [21] Omer Elkabetz and Nadav Cohen. Continuous vs. discrete optimization of deep neural networks. *arXiv preprint arXiv:2107.06608*, 2021.
- [22] Guilherme Franca, Daniel Robinson, and Rene Vidal. Admm and accelerated admm as continuous dynamical systems. In *International Conference on Machine Learning*, pages 1559–1567. PMLR, 2018.
- [23] Paul Glendinning. *Stability, instability and chaos: an introduction to the theory of nonlinear differential equations*. Cambridge university press, 1994.
- [24] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.
- [25] Christopher P Grant. Theory of ordinary differential equations. *Brigham Young University*, 2014.
- [26] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems*, 31:9461–9471, 2018.
- [27] Ernst Hairer, Syvert P Nørsett, and Gerhard Wanner. Solving ordinary differential equations i. nonstiff problems, volume 8 of, 1993.
- [28] Wei Hu, Lechao Xiao, and Jeffrey Pennington. Provable benefit of orthogonal initialization in optimizing deep linear networks. In *International Conference on Learning Representations*, 2020.
- [29] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. *International Conference on Learning Representations*, 2020.
- [30] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *International Conference on Learning Representations*, 2019.
- [31] Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 2020.
- [32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.
- [33] Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel LK Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. *International Conference on Learning Representations*, 2021.
- [34] Andrew K Lampinen and Surya Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. *International Conference on Learning Representations*, 2019.
- [35] Yann LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- [36] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.
- [37] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110. PMLR, 2017.
- [38] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, 2019.
- [39] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2019.
- [40] Daniil Merkulov and Ivan Oseledets. Stochastic gradient algorithms from ode splitting perspective. In *International Conference on Learning Representations, Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.
- [41] Leon Mirsky. A trace inequality of john von neumann. *Monatshefte für mathematik*, 79(4):303–306, 1975.

- [42] Asaf Noy, Yi Xu, Yonathan Aflalo, Lih Zelnik-Manor, and Rong Jin. A convergence theory towards practical over-parameterized deep neural networks. *arXiv preprint arXiv:2101.04243*, 2021.
- [43] Antonio Orvieto and Aurelien Lucchi. Shadowing properties of optimization algorithms. *Advances in Neural Information Processing Systems*, 32:12692–12703, 2019.
- [44] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.
- [45] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.
- [46] Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. In *Advances in Neural Information Processing Systems*, 2020.
- [47] Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in tensor factorization. In *International Conference on Machine Learning*. PMLR, 2021.
- [48] Halsey Lawrence Royden and Patrick Fitzpatrick. *Real analysis*, volume 32. Macmillan New York, 1988.
- [49] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *International Conference on Learning Representations*, 2014.
- [50] Damien Scieur, Vincent Roulet, Francis Bach, and Alexandre d’Aspremont. Integration methods and accelerated optimization algorithms. *arXiv preprint arXiv:1702.06751*, 2017.
- [51] Ohad Shamir. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. In *Conference on Learning Theory*, pages 2691–2713. PMLR, 2019.
- [52] Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*, 2018.
- [53] Samuel L Smith, Benoit Dherin, David GT Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. *International Conference on Learning Representations*, 2021.
- [54] Weijie Su, Stephen Boyd, and Emmanuel J Candès. A differential equation for modeling nesterov’s accelerated gradient method: theory and insights. In *Advances in neural information processing systems*, pages 2510–2518, 2014.
- [55] Endre Süli and David F Mayers. *An introduction to numerical analysis*. Cambridge university press, 2003.
- [56] Flemming Topsøe. Some bounds for the logarithmic function. *RGMIA Res. Rep. Collection*, 7(2):1–20, 2004.
- [57] Gal Vardi and Ohad Shamir. Implicit regularization in relu networks with the square loss. *arXiv preprint arXiv:2012.05156*, 2020.
- [58] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.
- [59] Ashia C Wilson, Benjamin Recht, and Michael I Jordan. A lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.
- [60] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.
- [61] Lei Wu, Qingcan Wang, and Chao Ma. Global convergence of gradient descent for deep linear residual networks. In *Advances in Neural Information Processing Systems*, volume 32, pages 13389–13398, 2019.
- [62] Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. *International Conference on Learning Representations*, 2021.
- [63] Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct runge-kutta discretization achieves acceleration. In *Advances in neural information processing systems*, pages 3904–3913, 2018.
- [64] Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. *Advances in neural information processing systems*, 2019.
- [65] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. arxiv e-prints. art. *arXiv preprint arXiv:1811.08888*, 2018.
- [66] Difan Zou, Philip M Long, and Quanquan Gu. On the global convergence of training deep linear resnets. In *International Conference on Learning Representations*, 2020.

## Checklist

1. For all authors...
  - (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
  - (b) Did you describe the limitations of your work? [Yes]
  - (c) Did you discuss any potential negative societal impacts of your work? [N/A]
  - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
  - (a) Did you state the full set of assumptions of all theoretical results? [Yes]
  - (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
  - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
  - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
  - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
  - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
  - (a) If your work uses existing assets, did you cite the creators? [Yes]
  - (b) Did you mention the license of the assets? [No] All assets used (MNIST data and PyTorch framework) are standard.
  - (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] There are no new assets other than our code (included in supplemental material).
  - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
  - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] All data used (MNIST) is standard.
5. If you used crowdsourcing or conducted research with human subjects...
  - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
  - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
  - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

## A Infinite Time for Gradient Flow Over Smooth Objective

By Theorem 1, gradient flow over a twice continuously differentiable objective function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  (Equation (5)) admits a unique solution  $\boldsymbol{\theta}: [0, t_e) \rightarrow \mathbb{R}^d$ , where either: (i)  $t_e = \infty$ ; or (ii)  $t_e < \infty$  and  $\lim_{t \nearrow t_e} \|\boldsymbol{\theta}(t)\|_2 = \infty$ . Lemma 3 below shows that if  $f(\cdot)$  is  $\beta$ -smooth then necessarily  $t_e = \infty$ .

**Lemma 3.** *Let  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  be twice continuously differentiable and  $\beta$ -smooth with  $\beta > 0$  (meaning  $\|\nabla^2 f(\mathbf{q})\|_{\text{spectral}} \leq \beta$  for all  $\mathbf{q} \in \mathbb{R}^d$ ). Then, for any  $\boldsymbol{\theta}_s \in \mathbb{R}^d$ , there exists a solution  $\boldsymbol{\theta}: [0, \infty) \rightarrow \mathbb{R}^d$  to gradient flow over  $f(\cdot)$  initialized at  $\boldsymbol{\theta}_s$  (Equation (5)).*

*Proof.* In light of Theorem 1, there exists a solution (to gradient flow over  $f(\cdot)$  initialized at  $\boldsymbol{\theta}_s$ )  $\boldsymbol{\theta}: [0, t_e) \rightarrow \mathbb{R}^d$ , where either: (i)  $t_e = \infty$ ; or (ii)  $t_e < \infty$  and  $\lim_{t \nearrow t_e} \|\boldsymbol{\theta}(t)\|_2 = \infty$ . It suffices to prove that condition (ii) is not satisfied. Assume by way of contradiction that it is. Then, there exists  $t_0 \in [0, t_e)$  such that for every  $t \in [t_0, t_e)$ ,  $\|\boldsymbol{\theta}(t)\|_2 \neq 0$  and we may write:

$$\begin{aligned} \frac{d}{dt} \|\boldsymbol{\theta}(t)\|_2 &= (\boldsymbol{\theta}(t) / \|\boldsymbol{\theta}(t)\|_2)^\top \frac{d}{dt} \boldsymbol{\theta}(t) \\ &= (\boldsymbol{\theta}(t) / \|\boldsymbol{\theta}(t)\|_2)^\top (-\nabla f(\boldsymbol{\theta}(t))) \\ &\leq \|\nabla f(\boldsymbol{\theta}(t))\|_2 \\ &= \|\nabla f(\mathbf{0}) + \nabla f(\boldsymbol{\theta}(t)) - \nabla f(\mathbf{0})\|_2 \\ &\leq \|\nabla f(\mathbf{0})\|_2 + \|\nabla f(\boldsymbol{\theta}(t)) - \nabla f(\mathbf{0})\|_2 \\ &\leq \|\nabla f(\mathbf{0})\|_2 + \beta \|\boldsymbol{\theta}(t)\|_2, \end{aligned}$$

where the first transition follows from the chain rule, the second holds since  $\boldsymbol{\theta}(\cdot)$  is a solution to gradient flow over  $f(\cdot)$ , the third is an application of the Cauchy-Schwartz inequality, the fourth is trivial, the fifth results from the triangle inequality, and the sixth is due to  $\beta$ -smoothness of  $f(\cdot)$ . Dividing by the right-hand side above and integrating between  $t_0$  and some  $t' \in [t_0, t_e)$ , we obtain:

$$\beta^{-1} \ln(\|\nabla f(\mathbf{0})\|_2 + \beta \|\boldsymbol{\theta}(t')\|_2) - \beta^{-1} \ln(\|\nabla f(\mathbf{0})\|_2 + \beta \|\boldsymbol{\theta}(t_0)\|_2) \leq t' - t_0,$$

which in turn implies:

$$\|\boldsymbol{\theta}(t')\|_2 \leq \beta^{-1} \left( (\|\nabla f(\mathbf{0})\|_2 + \beta \|\boldsymbol{\theta}(t_0)\|_2) \exp(\beta(t' - t_0)) - \|\nabla f(\mathbf{0})\|_2 \right).$$

We conclude that for any  $t' \in [t_0, t_e)$ , it holds that  $\|\boldsymbol{\theta}(t')\|_2 \leq c$ , where:

$$c := \beta^{-1} \left( (\|\nabla f(\mathbf{0})\|_2 + \beta \|\boldsymbol{\theta}(t_0)\|_2) \exp(\beta(t_e - t_0)) - \|\nabla f(\mathbf{0})\|_2 \right) < \infty.$$

This of course contradicts  $\lim_{t \nearrow t_e} \|\boldsymbol{\theta}(t)\|_2 = \infty$ , affirming that condition (ii) above is false.  $\square$

## B Worst Case Scenario

Theorem 3 in Section 3 established that if gradient descent (Equation (6)) is applied with step size  $\eta$  meeting a certain upper bound (Equation (7)), then its trajectory will  $\epsilon$ -approximate that of gradient flow (Equation (5)) up to a given time  $\tilde{t}$ . The upper bound on  $\eta$  decays exponentially with the integral of  $m(\cdot)$  along the gradient flow trajectory up to time  $\tilde{t}$ , where  $m(\cdot)$  corresponds to minus the minimal eigenvalue of the Hessian. Replacing  $m(\cdot)$  by a constant  $m$  equal to minus the minimal eigenvalue of the Hessian *across the entire space* results in a coarse bound, which for a non-convex objective ( $m > 0$ ) scales as  $e^{-m\tilde{t}}$  — see Corollary 1. The current appendix shows that in the worst case, such exponential scaling is necessary. That is, there exist objective functions and initializations with which the location of gradient flow at time  $\tilde{t}$  will not be  $\epsilon$ -approximated by the trajectory of gradient descent (at any iteration) unless the step size of gradient descent is  $\mathcal{O}(e^{-m\tilde{t}})$ . We prove this via an example, whose crux is that the gradient flow trajectories it entails traverse through regions where Hessian eigenvalues coincide with the minimal one across space.

Let  $a > 0$ ,  $b \geq 3$  and  $\epsilon \in (0, 1)$ . Define the ‘‘cut points’’  $z_c := be^{30} + 1$  and  $\bar{z}_c := b + 1$ , and the ‘‘transition width’’  $\bar{\rho} := \min\{e^{-12}/2, \epsilon/2b\}$ . Consider the functions  $\varphi, \bar{\varphi}: \mathbb{R} \rightarrow \mathbb{R}$  given by:

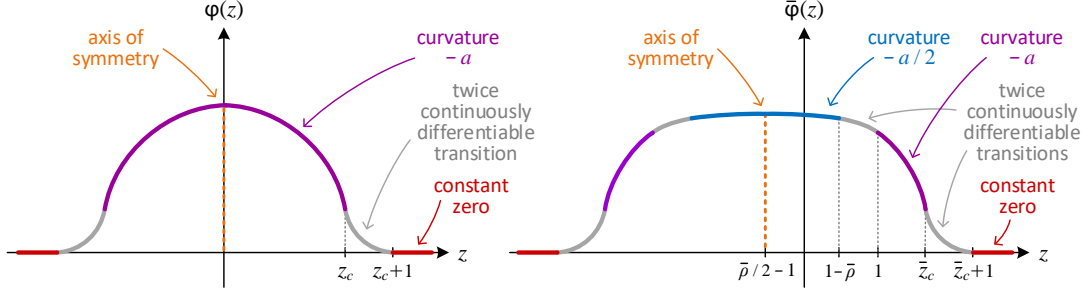


Figure 2: Illustrations of the functions  $\varphi(\cdot)$  and  $\bar{\varphi}(\cdot)$  defined in Equations (21) and (22) respectively.

$$\varphi(z) = \begin{cases} \frac{1}{2}a(z_c+1)^2 - \frac{5}{12}a - \frac{1}{2}az_c & , z=0 \\ \varphi(0) - \frac{1}{2}az^2 & , z \in (0, z_c) \\ \varphi(0) - \frac{1}{2}az^2 + a(\frac{2}{3} + z_c)(z - z_c)^3 - a(\frac{1}{4} + \frac{1}{2}z_c)(z - z_c)^4 & , z \in [z_c, z_c+1] \\ 0 & , z \in (z_c+1, \infty) \\ \varphi(|z|) & , z \in (-\infty, 0) \end{cases} \quad (21)$$

$$\bar{\varphi}(z) = \begin{cases} \frac{1}{2}a(\bar{z}_c+1)^2 + \frac{1}{12}a - \frac{1}{2}a\bar{z}_c - a(\frac{1}{2}\bar{\rho} - \frac{7}{48}\bar{\rho}^2) & , z = \frac{1}{2}\bar{\rho} - 1 \\ \bar{\varphi}(\frac{1}{2}\bar{\rho} - 1) - \frac{1}{4}a(z - (\frac{1}{2}\bar{\rho} - 1))^2 & , z \in (\frac{1}{2}\bar{\rho} - 1, 1 - \bar{\rho}) \\ \bar{\varphi}(\frac{1}{2}\bar{\rho} - 1) - \frac{1}{2}a + a(\frac{1}{2}\bar{\rho} - \frac{7}{48}\bar{\rho}^2) - \frac{1}{2}az^2 - \frac{1}{12}a\bar{\rho}^{-1}(z-1)^3 & , z \in [1 - \bar{\rho}, 1] \\ \bar{\varphi}(\frac{1}{2}\bar{\rho} - 1) - \frac{1}{2}a + a(\frac{1}{2}\bar{\rho} - \frac{7}{48}\bar{\rho}^2) - \frac{1}{2}az^2 & , z \in (1, \bar{z}_c) \\ \bar{\varphi}(\frac{1}{2}\bar{\rho} - 1) - \frac{1}{2}a + a(\frac{1}{2}\bar{\rho} - \frac{7}{48}\bar{\rho}^2) - \frac{1}{2}az^2 + a(\frac{2}{3} + \bar{z}_c)(z - \bar{z}_c)^3 - a(\frac{1}{4} + \frac{1}{2}\bar{z}_c)(z - \bar{z}_c)^4 & , z \in [\bar{z}_c, \bar{z}_c+1] \\ 0 & , z \in (\bar{z}_c+1, \infty) \\ \bar{\varphi}(|z - (\frac{1}{2}\bar{\rho} - 1)| + \frac{1}{2}\bar{\rho} - 1) & , z \in (-\infty, \frac{1}{2}\bar{\rho} - 1) \end{cases} \quad (22)$$

Both  $\varphi(\cdot)$  and  $\bar{\varphi}(\cdot)$  are twice continuously differentiable, non-negative and smooth,<sup>17</sup> with minimal curvature (second derivative) equal to  $-a$ .  $\varphi(\cdot)$  comprises three parts — (i) constant zero over  $(-\infty, -z_c - 1)$ ; (ii) quadratic with curvature  $-a$  over  $(-z_c, z_c)$ ; and (iii) constant zero over  $(z_c + 1, \infty)$  — with twice continuously differentiable transitions in-between.  $\bar{\varphi}(\cdot)$  consists of five parts — (i) constant zero over  $(-\infty, -\bar{z}_c - 3 + \bar{\rho})$ ; (ii) quadratic with curvature  $-a$  over  $(-\bar{z}_c - 2 + \bar{\rho}, -3 + \bar{\rho})$ ; (iii) quadratic with curvature  $-a/2$  over  $(-3 + 2\bar{\rho}, 1 - \bar{\rho})$ ; (iv) quadratic with curvature  $-a$  over  $(1, \bar{z}_c)$ ; and (v) constant zero over  $(\bar{z}_c + 1, \infty)$  — also joined by twice continuously differentiable transitions. Illustrations of  $\varphi(\cdot)$  and  $\bar{\varphi}(\cdot)$  are presented in Figure 2.

Let  $d \in \mathbb{N}_{\geq 3}$ , and consider the objective function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  defined by:

$$f(\mathbf{q}) = \varphi(q_1) + \bar{\varphi}(q_2) + 6aq_3^2, \quad (23)$$

where  $q_1, q_2$  and  $q_3$  stand for the first, second and third coordinates (respectively) of  $\mathbf{q} \in \mathbb{R}^d$ .  $f(\cdot)$  meets the conditions of Corollary 1 — it is twice continuously differentiable, non-negative and smooth.<sup>18</sup> The minimal eigenvalue of its Hessian across space (i.e.  $\inf_{\mathbf{q} \in \mathbb{R}^d} \lambda_{\min}(\nabla^2 f(\mathbf{q}))$ ), where  $\lambda_{\min}(\nabla^2 f(\mathbf{q}))$  represents the minimal eigenvalue of  $\nabla^2 f(\mathbf{q})$  is  $-a$ , meaning the constant  $m := -\inf_{\mathbf{q} \in \mathbb{R}^d} \lambda_{\min}(\nabla^2 f(\mathbf{q}))$  is equal to  $a$ . Building on the fact that in the region  $(0, z_c) \times (1, \bar{z}_c) \times \mathbb{R}^{d-2}$  the Hessian has eigenvalues coinciding with the minimum (i.e. equal to  $-a$ ), Proposition 4 below establishes the sought-after result — over  $f(\cdot)$ , there exist gradient flow trajectories whose  $\epsilon$ -approximation at a given time  $\bar{t}$  requires gradient descent to have step size  $\mathcal{O}(e^{-m\bar{t}})$ .

**Proposition 4.** *Let  $\theta_s = (\theta_{s,1}, \theta_{s,2}, \dots, \theta_{s,d}) \in \mathbb{R}^d$  be such that  $\theta_{s,1} \in (0.5, 1)$ ,  $\theta_{s,2} \in (e^{-12}/2 - 1, e^{-12} - 1)$  and  $\theta_{s,3} > 2$ . In the above context (in particular with the objective function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  defined*

<sup>17</sup>Their second derivatives are bounded.

<sup>18</sup>There exists  $\beta > 0$  such that  $\|\nabla^2 f(\mathbf{q})\|_{\text{spectral}} \leq \beta$  for all  $\mathbf{q} \in \mathbb{R}^d$ .



by Equation (23), for which  $m := -\inf_{\mathbf{q} \in \mathbb{R}^d} \lambda_{\min}(\nabla^2 f(\mathbf{q})) = a$ , denote by  $\boldsymbol{\theta}(\cdot)$  the trajectory of gradient flow initialized at  $\boldsymbol{\theta}_s$  (solution to Equation (5)), and by  $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots$  the iterates of gradient descent with step size  $\eta > 0$  (Equation (6)) emanating from the same point (i.e. with  $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_s$ ). Then, for any time  $\tilde{t} \in [\frac{2}{a} \ln(\frac{2-3\bar{\rho}/2}{\theta_{s,2} - (\bar{\rho}/2-1)}) + \frac{1}{a} \ln(\frac{2}{1-\bar{\rho}}), \frac{2}{a} \ln(\frac{2-3\bar{\rho}/2}{\theta_{s,2} - (\bar{\rho}/2-1)}) + \frac{1}{a} \ln(\frac{1+\bar{\rho}/4}{1-3\bar{\rho}/4}) + \frac{1}{a} \ln(b)]$ ,<sup>19</sup> if  $\eta \geq \frac{10^{14}}{a} e^{-a\tilde{t}} \epsilon$ , it holds that  $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}(\tilde{t})\|_2 > \epsilon$  for all  $k \in \mathbb{N} \cup \{0\}$ .<sup>20</sup>

*Proof sketch (for complete proof see Subappendix J.11).* Since  $f(\cdot)$  is additively separable (can be expressed as a sum of terms, each depending on a single input variable), the dynamics in  $\mathbb{R}^d$  induced by gradient flow and gradient descent can be analyzed separately for different coordinates. Restricting our attention to the first two coordinates, we observe that gradient flow and gradient descent initially traverse through an ‘‘anisotropic’’ region, where curvature is  $-a$  in the first coordinate and  $-a/2$  in the second, and from there move to an ‘‘isotropic’’ region, where curvature is  $-a$  in both the first and second coordinates. In the isotropic region, if gradient descent is placed along a gradient flow trajectory it will continue down the same path, but otherwise, if there is any discrepancy between gradient descent and gradient flow, this discrepancy will grow exponentially with time, namely will scale as  $e^{at}$ . Carefully characterizing the dynamics along the anisotropic region reveals that upon entrance to the isotropic one, there is indeed a discrepancy between gradient descent and gradient flow, the magnitude of which is proportional to  $\eta$  (step size of gradient descent). Since this magnitude scales as  $e^{at}$  thereafter, it will exceed  $\epsilon$  at time  $\tilde{t}$  if  $\eta \notin \mathcal{O}(e^{-a\tilde{t}} \epsilon)$ , which is what we set out to prove. The above analysis assumes  $\eta$  is no greater than a certain constant. However, larger values for  $\eta$  lead to divergence in the third coordinate (due to the term  $6aq_3^2$  in the definition of  $f(\cdot)$  — Equation (23)), thus these are accounted for as well (they preclude the possibility of gradient descent  $\epsilon$ -approximating gradient flow at time  $\tilde{t}$ ).  $\square$

## C Analysis for Fully Connected Architectures with Non-Linear Activation

In this appendix we provide our analysis for fully connected architectures with non-linear activation, outlined in Subsubsection 4.1.2.

When the (homogeneous) activation function of the fully connected neural network defined in Equation (8) (and surrounding text) is non-linear, i.e.  $\sigma(z) = \alpha \max\{z, 0\} - \bar{\alpha} \max\{-z, 0\}$  for some  $\alpha, \bar{\alpha} \in \mathbb{R}$ ,  $\alpha \neq \bar{\alpha}$ , the training loss  $f(\cdot)$  (Equation (9)) is (typically) not everywhere differentiable. It is however locally Lipschitz thus differentiable almost everywhere (see Theorem 9.1.2 in [10]). Moreover, as established by Proposition 9 in Appendix E, for almost every  $\boldsymbol{\theta}' \in \mathbb{R}^d$  there exist diagonal matrices  $D'_{i,j} \in \mathbb{R}^{d_j \times d_j}$ ,  $i = 1, 2, \dots, |\mathcal{S}|$ ,  $j = 1, 2, \dots, n-1$ , with diagonal elements in  $\{\alpha, \bar{\alpha}\}$ , such that  $f(\cdot)$  coincides with the function

$$\boldsymbol{\theta} \mapsto \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell(W_n D'_{i,n-1} W_{n-1} D'_{i,n-2} W_{n-2} \cdots D'_{i,1} W_1 \mathbf{x}_i, y_i) \quad (24)$$

on an open region  $\mathcal{D}_{\boldsymbol{\theta}'} \subseteq \mathbb{R}^d$  containing  $\boldsymbol{\theta}'$ , that is closed under positive rescaling of weight matrices (i.e. under  $(W_1, W_2, \dots, W_n) \mapsto (c_1 W_1, c_2 W_2, \dots, c_n W_n)$  with  $c_1, c_2, \dots, c_n > 0$ ). The notion of gradient flow over a non-differentiable locally Lipschitz objective function is typically formalized via differential inclusion and Clarke subdifferentials (cf. [15, 18]). To our knowledge there exists no analogue of the Fundamental Theorem (Theorem 2) that applies to this formalization, thus we focus on (open) regions of the form  $\mathcal{D}_{\boldsymbol{\theta}'}$ , where  $f(\cdot)$  is given by Equation (24), and in particular is twice continuously differentiable. On such regions the analysis of Section 3 applies, and since they constitute the entire weight space but a negligible (closed and zero measure) set, they can facilitate a ‘‘piecewise characterization’’ of the discrepancy between gradient flow and gradient descent.

Lemma 4 below expresses  $\nabla^2 f(\boldsymbol{\theta})$  for  $\boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\theta}'}$ .

**Lemma 4.** *Let  $\boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\theta}'}$ . For any  $i \in \{1, 2, \dots, |\mathcal{S}|\}$  and  $j, j' \in \{1, 2, \dots, n\}$  define  $(D'_{i,*} W_*)_{j':j}$  to be the matrix  $D'_{i,j'} W_{j'} D'_{i,j'-1} W_{j'-1} \cdots D'_{i,j} W_j$  (where by convention  $D'_{i,n} \in \mathbb{R}^{d_n \times d_n}$  stands for identity) if  $j \leq j'$ , and an identity matrix (with size to be inferred by context) otherwise. For  $i \in \{1, 2, \dots, |\mathcal{S}|\}$  let  $\nabla \ell_i \in \mathbb{R}^{d_n}$  and  $\nabla^2 \ell_i \in \mathbb{R}^{d_n \times d_n}$  be the gradient and Hessian (respectively) of the loss  $\ell(\cdot)$  at the point*

<sup>19</sup>Note that the upper bound on  $\tilde{t}$  can be made arbitrarily large via suitable (sufficiently large) choice of  $b$ .

<sup>20</sup>Since  $f(\cdot)$  is twice continuously differentiable and smooth,  $\boldsymbol{\theta}(\tilde{t})$  necessarily exists (see Lemma 3 in Appendix A).

$((D'_{i,*} W_*)_{n:1} \mathbf{x}_i, y_i)$  with respect to its first argument. Then, regarding Hessians as quadratic forms (see examples in Lemma 1), it holds that:

$$\begin{aligned} \nabla^2 f(\boldsymbol{\theta})[\Delta W_1, \Delta W_2, \dots, \Delta W_n] &= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i \left[ \sum_{j=1}^n (D'_{i,*} W_*)_{n:j+1} D'_{i,j}(\Delta W_j) (D'_{i,*} W_*)_{j-1:1} \mathbf{x}_i \right] \\ &+ \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i^\top \sum_{1 \leq j < j' \leq n} (D'_{i,*} W_*)_{n:j'+1} D'_{i,j'}(\Delta W_{j'}) (D'_{i,*} W_*)_{j'-1:j+1} D'_{i,j}(\Delta W_j) (D'_{i,*} W_*)_{j-1:1} \mathbf{x}_i. \end{aligned} \quad (25)$$

*Proof sketch (for complete proof see Subappendix J.12).* The proof is similar to that of Lemma 1. Namely, it expands the function in Equation (24) and then extracts second order terms.  $\square$

The following proposition employs Lemma 4 to show that (under mild conditions) there exists  $\boldsymbol{\theta} \in \mathbb{R}^d$  for which  $\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta}))$  is arbitrarily negative.

**Proposition 5.** *Assume that: (i) the network is deep ( $n \geq 3$ ); and (ii) the loss function  $\ell(\cdot)$  and training set  $\mathcal{S}$  are non-degenerate, in the sense that there exists a weight setting  $\boldsymbol{\theta} \in \mathbb{R}^d$  for which  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) \neq 0$ , where  $\nabla \ell(\cdot)$  stands for the gradient of  $\ell(\cdot)$  with respect to its first argument, and  $h_{\boldsymbol{\theta}}(\cdot)$  is the input-output mapping realized by the network (Equation (8)).<sup>21</sup> Then, it holds that  $\inf_{\boldsymbol{\theta} \in \mathbb{R}^d \text{ s.t. } \nabla^2 f(\boldsymbol{\theta}) \text{ exists}} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) = -\infty$ .*

*Proof sketch (for complete proof see Subappendix J.13).* Let  $\boldsymbol{\theta} \in \mathbb{R}^d$  be a weight setting realizing the non-degeneracy condition, i.e. for which  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) \neq 0$ . Without loss of generality, we may assume that  $\boldsymbol{\theta}$  satisfies the condition  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) < 0$  (if this is not the case then simply flip the signs of the entries in  $\boldsymbol{\theta}$  corresponding to the last weight matrix  $W_n$ ). From continuity, there exists a neighborhood of  $\boldsymbol{\theta}$  consisting of weight settings that all meet the latter condition. There must exist a region of the form  $\mathcal{D}_{\boldsymbol{\theta}'}$  intersecting this neighborhood (since these regions constitute all of  $\mathbb{R}^d$  but a zero measure set), so we may assume, without loss of generality, that  $\boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\theta}'}$ . Lemma 4 then applies. Moreover, since  $\mathcal{D}_{\boldsymbol{\theta}'}$  is closed under positive rescaling of weight matrices (i.e. of  $W_1, W_2, \dots, W_n$ ), the lemma remains applicable even when  $\boldsymbol{\theta}$  is subject to such rescaling. The proof proceeds by fixing  $\Delta W_1, \Delta W_2, \dots, \Delta W_n$  to certain values, and positively rescaling  $W_1, W_2, \dots, W_n$  in a certain way, such that the expression for  $\nabla^2 f(\boldsymbol{\theta})[\Delta W_1, \Delta W_2, \dots, \Delta W_n]$  provided in Lemma 4 becomes arbitrarily negative.  $\square$

Relying on Lemma 4, Lemma 5 below provides a lower bound on  $\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta}))$  for  $\boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\theta}'}$ .

**Lemma 5.** *With the notations of Lemma 4, for any  $\boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\theta}'}$ :<sup>13</sup>*

$$\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -\max\{|\alpha|, |\bar{\alpha}|\}^{n-1} \frac{n-1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \max_{\substack{\mathcal{J} \subseteq \{1, 2, \dots, n\} \\ |\mathcal{J}| = n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_{\text{Frobenius}}. \quad (26)$$

*Proof sketch (for complete proof see Subappendix J.14).* The proof is analogous to that of Lemma 2. Namely, it appeals to Lemma 4, and lower bounds the right-hand side of Equation (25). Convexity of  $\ell(\cdot)$  (with respect to its first argument) implies that the first summand is non-negative. For the second summand, we use known matrix inequalities (as well as the fact that  $\|D'_{i,j}\|_{\text{spectral}}$  is no greater than  $\max\{|\alpha|, |\bar{\alpha}|\}$  for  $j=1, 2, \dots, n-1$ , and equal to one for  $j=n$ ) to establish a lower bound of  $c \sum_{j=1}^n \|\Delta W_j\|_{\text{Frobenius}}^2$ , with  $c$  being the expression on the right-hand side of Equation (26).  $\square$

The lower bound in Equation (26) is highly sensitive to the scales of the individual weight matrices. Specifically, assuming the network is deep ( $n \geq 3$ ), if  $\boldsymbol{\theta}$  does not perfectly fit all non-zero training inputs (meaning there exists  $i \in \{1, 2, \dots, |\mathcal{S}|\}$  for which  $\nabla \ell_i \neq \mathbf{0}$  and  $\mathbf{x}_i \neq \mathbf{0}$ ), and if at least  $n-2$  of its weight matrices  $W_1, W_2, \dots, W_n$  are non-zero, then it is possible to rescale each  $W_j$  by  $c_j > 0$ , with  $\prod_{j=1}^n c_j = 1$ , such that the lower bound in Equation (26) becomes arbitrarily negative<sup>22</sup> despite the

<sup>21</sup> Assumptions (i) and (ii) are both necessary, in the sense that removing any of them (without imposing further assumptions) renders the proposition false — see Claim 2 in Appendix F. Assumption (ii) in particular is extremely mild, e.g. if  $\ell(\cdot)$  is the square loss (i.e.  $\mathcal{Y} = \mathbb{R}^{d_n}$  and  $\ell(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{2} \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2$ ), the slightest change in a single label ( $\mathbf{y}_i$ ) corresponding to a non-zero prediction ( $h_{\boldsymbol{\theta}}(\mathbf{x}_i) \neq \mathbf{0}$ ) can ensure the inequality.

<sup>22</sup> The bound remains applicable since  $\mathcal{D}_{\boldsymbol{\theta}'}$  is closed under positive rescaling of weight matrices.

input-output mapping  $h_{\theta}(\cdot)$  (and thus the training loss value  $f(\theta)$ ) remaining unchanged. Nevertheless, similarly to the case of linear activation (Subsubsection 4.1.1), we may employ the fact that gradient flow over fully connected neural networks (with homogeneous activations) initialized near zero maintains balance between weight matrices — cf. [18] — to show that along its trajectories, the lower bound in Equation (26) assumes a tighter form. This is done in Proposition 6 below.

**Proposition 6.** *If  $\theta \in \mathcal{D}_{\theta'}$  resides on a trajectory of gradient flow (over  $f(\cdot)$ )<sup>23</sup> initialized at some point  $\theta_s \in \mathbb{R}^d$ , with  $\|\theta_s\|_2 \leq \epsilon$  for some  $\epsilon > 0$ , then, using the notations of Lemma 4:*

$$\lambda_{\min}(\nabla^2 f(\theta)) \geq -\max\{|\alpha|, |\bar{\alpha}|\}^{n-1} \frac{n-1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \left( \min_{j \in \{1, 2, \dots, n\}} \|W_j\|_{\text{Frobenius}} + \epsilon \right)^{n-2}. \quad (27)$$

*Proof sketch (for complete proof see Subappendix J.15).* By the analysis of [18], for any  $j, j' \in \{1, 2, \dots, n\}$ , the quantity  $\|W_{j'}\|_{\text{Frobenius}}^2 - \|W_j\|_{\text{Frobenius}}^2$  is invariant (constant) along a gradient flow trajectory. This implies that along a trajectory emanating from a point with (Euclidean) norm  $\mathcal{O}(\epsilon)$ , it holds that  $\|W_{j'}\|_{\text{Frobenius}}^2 - \|W_j\|_{\text{Frobenius}}^2 \in \mathcal{O}(\epsilon^2)$  for all  $j, j' \in \{1, 2, \dots, n\}$ , which in turn implies  $\|W_{j'}\|_{\text{Frobenius}} \leq \min_{j \in \{1, 2, \dots, n\}} \|W_j\|_{\text{Frobenius}} + \mathcal{O}(\epsilon)$  for all  $j' \in \{1, 2, \dots, n\}$ . Plugging this into Equation (26) yields the desired result (Equation (27)).  $\square$

Assume the network is deep ( $n \geq 3$ ), and consider a trajectory of gradient flow (over  $f(\cdot)$ ) emanating from near-zero initialization. For every point on the trajectory, Proposition 6 may be applied with small  $\epsilon$ , leading the lower bound in Equation (27) to depend primarily on the *minimal* size (Frobenius norm) of a weight matrix  $W_j$ , and on  $\nabla \ell_1, \nabla \ell_2, \dots, \nabla \ell_{|\mathcal{S}|}$  — gradients of the loss function with respect to the predictions over the training set. In the course of optimization,  $W_1, W_2, \dots, W_n$  are initially small, and if a perfect fit of the training set is ultimately achieved,  $\nabla \ell_1, \nabla \ell_2, \dots, \nabla \ell_{|\mathcal{S}|}$  will converge to zero. Therefore, if not *all* weight matrices  $W_1, W_2, \dots, W_n$  become large during optimization, the lower bound on  $\lambda_{\min}(\nabla^2 f(\theta))$  in Equation (27) will only be moderately negative before approaching non-negativity (if and) as the trajectory converges to a perfect fit. In light of Section 3, this suggests that the gradient flow trajectory may lend itself to approximation by gradient descent. For a case with linear activation (Subsubsection 4.1.1) such prospect is theoretically verified in Section 5. For non-linear activation we provide empirical corroboration in Section 6, deferring to future work a complete theoretical affirmation.

## D Analysis for Convolutional Architectures

In this appendix we provide our analysis for convolutional architectures, outlined in Subsection 4.2.

Suppose we modify the fully connected neural network defined in Equation (8) (and surrounding text) by converting each learned weight matrix  $W_j \in \mathbb{R}^{d_j, d_{j-1}}$ ,  $j = 1, 2, \dots, n$ , into a function  $W_j: \mathbb{R}^{d_j} \rightarrow \mathbb{R}^{d_j, d_{j-1}}$ , with  $d_j' \in \mathbb{N}$ , that intakes a learned weight vector  $\mathbf{w}_j \in \mathbb{R}^{d_j'}$ , and returns a matrix where each element is either fixed at zero or connected to a predetermined coordinate of  $\mathbf{w}_j$ , with no repetition of coordinates within the same row (that is, each row of  $W_j(\cdot)$  realizes a function of the form  $\mathbf{w}_j \mapsto P\mathbf{w}_j$ , where  $P \in \mathbb{R}^{d_{j-1}, d_j'}$  is a matrix in which no row or column includes more than a single non-zero element, and all non-zero elements are equal to one). This allows imposing various weight sharing and sparsity patterns on the layers of the model, in particular ones giving rise to convolutional neural networks. The resulting input-output mapping has the form:

$$h_{\theta}: \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_n}, h_{\theta}(\mathbf{x}) = W_n(\mathbf{w}_n) \sigma(W_{n-1}(\mathbf{w}_{n-1}) \sigma(W_{n-2}(\mathbf{w}_{n-2}) \cdots \sigma(W_1(\mathbf{w}_1)\mathbf{x}) \cdots)), \quad (28)$$

where  $\theta \in \mathbb{R}^d$ , with  $d := \sum_{j=1}^n d_j'$ , is the concatenation of the weight vectors  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$ ,<sup>24</sup> and as before,  $\sigma: \mathbb{R} \rightarrow \mathbb{R}$  is a predetermined activation function (operating element-wise when

<sup>23</sup>Recall that in the current context, the optimized objective function  $f(\cdot)$  is locally Lipschitz but (typically) non-differentiable. Following a conventional formalization in such settings (cf. [15, 18]), we regard a curve in  $\mathbb{R}^d$  as a trajectory of gradient flow if it satisfies the differential inclusion  $\frac{d}{dt}\theta(t) \in -\partial f(\theta(t))$  for almost every time  $t$ , where  $\partial f(\theta(t)) \subseteq \mathbb{R}^d$  stands for the Clarke subdifferential (see [13]) of  $f(\cdot)$  at  $\theta(t)$ .

<sup>24</sup>The exact order by which  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$  are concatenated is insignificant for our purposes — all that matters is that the same order be used throughout.

applied to a vector) that is (positively) homogeneous, meaning there exist  $\alpha, \bar{\alpha} \in \mathbb{R}$  such that  $\sigma(z) = \alpha \max\{z, 0\} - \bar{\alpha} \max\{-z, 0\}$  for all  $z \in \mathbb{R}$ .<sup>25</sup>

Let  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  be the training loss defined by applying Equation (9) (and surrounding text) to the above neural network (*i.e.* with  $h_\theta(\cdot)$  given by Equation (28)). In line with our analysis of fully connected architectures (Subsection 4.1), we will show that although the minimal eigenvalue of  $\nabla^2 f(\theta)$  (Hessian of training loss) — denoted  $\lambda_{\min}(\nabla^2 f(\theta))$  — can in general be arbitrarily negative, along trajectories of gradient flow (which emanate from near-zero initialization) it is no less than moderately negative, approaching non-negativity towards convergence. In light of Section 3, this suggests that over deep convolutional neural networks, gradient flow may lend itself to approximation by gradient descent — a prospect we empirically corroborate in Subappendix I.1.

Proposition 10 in Appendix E establishes that for almost every  $\theta' \in \mathbb{R}^d$  there exist diagonal matrices  $D'_{i,j} \in \mathbb{R}^{d_j, d_j}$ ,  $i = 1, 2, \dots, |\mathcal{S}|$ ,  $j = 1, 2, \dots, n-1$ , with diagonal elements in  $\{\alpha, \bar{\alpha}\}$ , such that  $f(\cdot)$  coincides with the function

$$\theta \mapsto \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell(W_n(\mathbf{w}_n) D'_{i,n-1} W_{n-1}(\mathbf{w}_{n-1}) D'_{i,n-2} W_{n-2}(\mathbf{w}_{n-2}) \cdots D'_{i,1} W_1(\mathbf{w}_1) \mathbf{x}_i, y_i) \quad (29)$$

on an open region  $\mathcal{D}_{\theta'} \subseteq \mathbb{R}^d$  containing  $\theta'$ , that is closed under positive rescaling of weight vectors (*i.e.* under  $(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n) \mapsto (c_1 \mathbf{w}_1, c_2 \mathbf{w}_2, \dots, c_n \mathbf{w}_n)$  with  $c_1, c_2, \dots, c_n > 0$ ). Analogously to the case of fully connected architectures with non-linear activation (*cf.* Appendix C), we will focus on (open) regions of the form  $\mathcal{D}_{\theta'}$ , where  $f(\cdot)$  is given by Equation (29), and in particular is twice continuously differentiable. On such regions the analysis of Section 3 applies, and since they constitute the entire weight space but a negligible (closed and zero measure) set, they can facilitate a ‘‘piecewise characterization’’ of the discrepancy between gradient flow and gradient descent.<sup>26</sup>

Lemma 6 below expresses  $\nabla^2 f(\theta)$  for  $\theta \in \mathcal{D}_{\theta'}$ .

**Lemma 6.** *Let  $\theta \in \mathcal{D}_{\theta'}$ . For any  $i \in \{1, 2, \dots, |\mathcal{S}|\}$  and  $j, j' \in \{1, 2, \dots, n\}$  define  $(D'_{i,*} W_*(\mathbf{w}_*))_{j':j}$  to be the matrix  $D'_{i,j'} W_{j'}(\mathbf{w}_{j'}) D'_{i,j'-1} W_{j'-1}(\mathbf{w}_{j'-1}) \cdots D'_{i,j} W_j(\mathbf{w}_j)$  (where by convention  $D'_{i,n} \in \mathbb{R}^{d_n, d_n}$  stands for identity) if  $j \leq j'$ , and an identity matrix (with size to be inferred by context) otherwise. For  $i \in \{1, 2, \dots, |\mathcal{S}|\}$  let  $\nabla \ell_i \in \mathbb{R}^{d_n}$  and  $\nabla^2 \ell_i \in \mathbb{R}^{d_n, d_n}$  be the gradient and Hessian (respectively) of the loss  $\ell(\cdot)$  at the point  $((D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i, y_i)$  with respect to its first argument. Then, regarding Hessians as quadratic forms (see examples in Lemma 1), it holds that:*

$$\begin{aligned} \nabla^2 f(\theta)[\Delta \mathbf{w}_1, \Delta \mathbf{w}_2, \dots, \Delta \mathbf{w}_n] = & \quad (30) \\ & \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i \left[ \sum_{j=1}^n (D'_{i,*} W_*(\mathbf{w}_*))_{n:j+1} D'_{i,j} W_j(\Delta \mathbf{w}_j) (D'_{i,*} W_*(\mathbf{w}_*))_{j-1:1} \mathbf{x}_i \right] + \\ & \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i^\top \sum_{1 \leq j < j' \leq n} (D'_{i,*} W_*(\mathbf{w}_*))_{n:j'+1} D'_{i,j'} W_{j'}(\Delta \mathbf{w}_{j'}) (D'_{i,*} W_*(\mathbf{w}_*))_{j'-1:j+1} \cdot \\ & \quad D'_{i,j} W_j(\Delta \mathbf{w}_j) (D'_{i,*} W_*(\mathbf{w}_*))_{j-1:1} \mathbf{x}_i. \end{aligned}$$

*Proof sketch (for complete proof see Subappendix J.16).* The proof is similar to those of Lemmas 1 and 4. Namely, it expands the function in Equation (29) and then extracts second order terms.  $\square$

The following proposition employs Lemma 6 to show that (under mild conditions) there exists  $\theta \in \mathbb{R}^d$  for which  $\lambda_{\min}(\nabla^2 f(\theta))$  is arbitrarily negative.

**Proposition 7.** *Assume that: (i) the network is deep ( $n \geq 3$ ); and (ii) the network, loss function  $\ell(\cdot)$  and training set  $\mathcal{S}$  are non-degenerate, in the sense that there exists a weight setting  $\theta \in \mathbb{R}^d$  for which  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\theta, y_i)^\top h_\theta(\mathbf{x}_i) \neq 0$ , where  $\nabla \ell(\cdot)$  stands for the gradient of  $\ell(\cdot)$  with respect to its first*

<sup>25</sup>Similarly to our analysis of fully connected architectures (Subsection 4.1), that of convolutional architectures (current appendix) readily extends to the case of different (homogeneous) activation functions at different hidden layers.

<sup>26</sup>Such ‘‘piecewise characterization’’ is holistic when the activation function  $\sigma(\cdot)$  is linear, *i.e.*  $\sigma(z) = z$  (or more generally,  $\alpha = \bar{\alpha}$ ). Indeed, in this case  $f(\cdot)$  is twice continuously differentiable throughout, and we may take  $\mathcal{D}_{\theta'} = \mathbb{R}^d$ .

argument, and  $h_{\theta}(\cdot)$  is the input-output mapping realized by the network (Equation (28)).<sup>27</sup> Then, it holds that  $\inf_{\theta \in \mathbb{R}^d \text{ s.t. } \nabla^2 f(\theta) \text{ exists}} \lambda_{\min}(\nabla^2 f(\theta)) = -\infty$ .

*Proof sketch (for complete proof see Subappendix J.17).* The proof is analogous to that of Proposition 5. Specifically, it establishes that there exists  $\theta \in \mathcal{D}_{\theta'}$  for which  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_{\theta}(\mathbf{x}_i) < 0$ , and then makes use of Lemma 6 to show that fixing  $\Delta \mathbf{w}_1, \Delta \mathbf{w}_2, \dots, \Delta \mathbf{w}_n$  to certain values, and positively rescaling  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$  in a certain way, leads  $\nabla^2 f(\theta) [\Delta \mathbf{w}_1, \Delta \mathbf{w}_2, \dots, \Delta \mathbf{w}_n]$  to become arbitrarily negative.  $\square$

Relying on Lemma 6, Lemma 7 below provides a lower bound on  $\lambda_{\min}(\nabla^2 f(\theta))$  for  $\theta \in \mathcal{D}_{\theta'}$ .

**Lemma 7.** *With the notations of Lemma 6, for any  $\theta \in \mathcal{D}_{\theta'}$ :*<sup>13</sup>

$$\lambda_{\min}(\nabla^2 f(\theta)) \geq -\max\{|\alpha|, |\bar{\alpha}|\}^{n-1} \frac{n-1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \cdot \prod_{j=1}^n \|W_j(\cdot)\|_{op} \max_{\substack{\mathcal{J} \subseteq \{1, 2, \dots, n\} \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|\mathbf{w}_j\|_2, \quad (31)$$

where  $\|W_j(\cdot)\|_{op}$ ,  $j=1, 2, \dots, n$ , denotes the operator norm of  $W_j(\cdot)$  induced by the Frobenius norm.<sup>28</sup>

*Proof sketch (for complete proof see Subappendix J.18).* The proof mirrors those of Lemmas 2 and 5 — it establishes that the right-hand side of Equation (30) in Lemma 6 is lower bounded by  $c \sum_{j=1}^n \|\Delta \mathbf{w}_j\|_2^2$ , with  $c$  being the expression on the right-hand side of Equation (31).  $\square$

The lower bound in Equation (31) is highly sensitive to the scales of the individual weight vectors. Specifically, assuming the network is deep ( $n \geq 3$ ) and is non-degenerate, in the sense that all of its layers can realize non-zero mappings (that is, the activation function  $\sigma(\cdot)$  is not identically zero, i.e.  $\alpha$  and  $\bar{\alpha}$  are not both equal to zero, and for all  $j \in \{1, 2, \dots, n\}$ ,  $W_j(\cdot)$  is not the zero mapping, i.e.  $\|W_j(\cdot)\|_{op} > 0$ ), if  $\theta$  does not perfectly fit all non-zero training inputs (meaning there exists  $i \in \{1, 2, \dots, |\mathcal{S}|\}$  for which  $\nabla \ell_i \neq \mathbf{0}$  and  $\mathbf{x}_i \neq \mathbf{0}$ ), and if at least  $n-2$  of its weight vectors  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$  are non-zero, then it is possible to rescale each  $\mathbf{w}_j$  by  $c_j > 0$ , with  $\prod_{j=1}^n c_j = 1$ , such that the lower bound in Equation (31) becomes arbitrarily negative<sup>29</sup> despite the input-output mapping  $h_{\theta}(\cdot)$  (and thus the training loss value  $f(\theta)$ ) remaining unchanged. Nevertheless, as with fully connected architectures (see Subsection 4.1), gradient flow over convolutional architectures (i.e. over neural networks as defined in Equation (28) and surrounding text) initialized near zero maintains balance between weight vectors — cf. [18] — and so along its trajectories the lower bound in Equation (31) assumes a tighter form. This is formalized in Proposition 8 below.

**Proposition 8.** *If  $\theta \in \mathcal{D}_{\theta'}$  resides on a trajectory of gradient flow (over  $f(\cdot)$ )<sup>23</sup> initialized at some point  $\theta_s \in \mathbb{R}^d$ , with  $\|\theta_s\|_2 \leq \epsilon$  for some  $\epsilon > 0$ , then, using the notations of Lemmas 6 and 7:*

$$\lambda_{\min}(\nabla^2 f(\theta)) \geq -\max\{|\alpha|, |\bar{\alpha}|\}^{n-1} \frac{n-1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \cdot \prod_{j=1}^n \|W_j(\cdot)\|_{op} \left( \min_{j \in \{1, 2, \dots, n\}} \|\mathbf{w}_j\|_2 + \epsilon \right)^{n-2}. \quad (32)$$

*Proof sketch (for complete proof see Subappendix J.19).* By the analysis of [18], for any  $j, j' \in \{1, 2, \dots, n\}$ , the quantity  $\|\mathbf{w}_{j'}\|_2^2 - \|\mathbf{w}_j\|_2^2$  is invariant (constant) along a gradient flow trajectory. This implies that along a trajectory emanating from a point with (Euclidean) norm  $\mathcal{O}(\epsilon)$ , it holds that  $\|\mathbf{w}_{j'}\|_2^2 - \|\mathbf{w}_j\|_2^2 \in \mathcal{O}(\epsilon^2)$  for all  $j, j' \in \{1, 2, \dots, n\}$ , which in turn implies

<sup>27</sup> Assumptions (i) and (ii) are both necessary, in the sense that removing any of them (without imposing further assumptions) renders the proposition false — see Claim 3 in Appendix F. Assumption (ii) in particular is extremely mild, e.g. if  $\ell(\cdot)$  is the square loss (i.e.  $\mathcal{Y} = \mathbb{R}^{d_n}$  and  $\ell(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{2} \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2$ ), the slightest change in a single label ( $\mathbf{y}_i$ ) corresponding to a non-zero prediction ( $h_{\theta}(\mathbf{x}_i) \neq \mathbf{0}$ ) can ensure the inequality.

<sup>28</sup> From the structure of  $W_j(\cdot)$  (see beginning of this appendix) it follows that  $\|W_j(\cdot)\|_{op}$  is equal to square root of the maximal number of elements in  $W_j(\mathbf{w}_j)$  connected to the same coordinate of  $\mathbf{w}_j$ .

<sup>29</sup> The bound remains applicable since  $\mathcal{D}_{\theta'}$  is closed under positive rescaling of weight vectors.

$\|\mathbf{w}_{j'}\|_2 \leq \min_{j \in \{1, 2, \dots, n\}} \|\mathbf{w}_j\|_2 + \mathcal{O}(\epsilon)$  for all  $j' \in \{1, 2, \dots, n\}$ . Plugging this into Equation (31) yields the desired result (Equation (32)).  $\square$

Assume the network is deep ( $n \geq 3$ ) and non-degenerate ( $\alpha$  and  $\bar{\alpha}$  are not both equal to zero, and  $\|W_j(\cdot)\|_{op} > 0$  for all  $j \in \{1, 2, \dots, n\}$ ), and consider a trajectory of gradient flow (over  $f(\cdot)$ ) emanating from near-zero initialization. For every point on the trajectory, Proposition 8 may be applied with small  $\epsilon$ , leading the lower bound in Equation (32) to depend primarily on the *minimal* size (Euclidean norm) of a weight vector  $\mathbf{w}_j$ , and on  $\nabla \ell_1, \nabla \ell_2, \dots, \nabla \ell_{|\mathcal{S}|}$  — gradients of the loss function with respect to the predictions over the training set. In the course of optimization,  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$  are initially small, and if a perfect fit of the training set is ultimately achieved,  $\nabla \ell_1, \nabla \ell_2, \dots, \nabla \ell_{|\mathcal{S}|}$  will converge to zero. Therefore, if not *all* weight vectors  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$  become large during optimization, the lower bound on  $\lambda_{min}(\nabla^2 f(\boldsymbol{\theta}))$  in Equation (32) will only be moderately negative before approaching non-negativity (if and) as the trajectory converges to a perfect fit. In light of Section 3, this suggests that the gradient flow trajectory may lend itself to approximation by gradient descent. For a case of fully connected neural networks with linear activation (analyzed in Subsubsection 4.1.1), such prospect is theoretically verified in Section 5. For convolutional architectures (subject of the current appendix) we provide empirical corroboration in Subappendix I.1, deferring to future work a complete theoretical affirmation.

## E Regions of Differentiability

In this appendix we prove that for fully connected and convolutional architectures with non-linear activation, there exist regions of differentiability  $\mathcal{D}_{\boldsymbol{\theta}'}$  as described in Appendixes C and D respectively.

**Proposition 9** (regions of differentiability for fully connected architectures). *Consider a fully connected neural network as defined in Equation (8) (and surrounding text), and assume that its (homogeneous) activation function is non-linear, i.e.  $\sigma(z) = \alpha \max\{z, 0\} - \bar{\alpha} \max\{-z, 0\}$  for some  $\alpha, \bar{\alpha} \in \mathbb{R}$ ,  $\alpha \neq \bar{\alpha}$ . Then, for almost every (in the sense of Lebesgue measure)  $\boldsymbol{\theta}' \in \mathbb{R}^d$ , there exist diagonal matrices  $D'_{i,j} \in \mathbb{R}^{d_j \times d_j}$ ,  $i = 1, 2, \dots, |\mathcal{S}|$ ,  $j = 1, 2, \dots, n-1$ , with diagonal elements in  $\{\alpha, \bar{\alpha}\}$ , such that the training loss  $f(\cdot)$  (Equation (9)) coincides with the function defined in Equation (24) on an open region  $\mathcal{D}_{\boldsymbol{\theta}'} \subseteq \mathbb{R}^d$  containing  $\boldsymbol{\theta}'$ , that is closed under positive rescaling of weight matrices (i.e. under  $(W_1, W_2, \dots, W_n) \mapsto (c_1 W_1, c_2 W_2, \dots, c_n W_n)$  with  $c_1, c_2, \dots, c_n > 0$ ).*

*Proof.* If for  $\boldsymbol{\theta}' \in \mathbb{R}^d$  there exist diagonal matrices  $(D'_{i,j})_{i,j}$  and an open region  $\mathcal{D}_{\boldsymbol{\theta}'}$  as above, then we refer to  $\boldsymbol{\theta}'$  as an *admissible* weight setting, to  $(D'_{i,j})_{i,j}$  as its *activation matrices*, and to  $\mathcal{D}_{\boldsymbol{\theta}'}$  as its *differentiability region*.<sup>30</sup>

Without loss of generality, we may assume  $|\mathcal{S}| = 1$ , i.e. that the training set comprises a single labeled input  $(\mathbf{x}, y) \in \mathbb{R}^{d_0} \times \mathcal{Y}$ , meaning the training loss takes the form  $f(\boldsymbol{\theta}) = \ell(h_{\boldsymbol{\theta}}(\mathbf{x}), y)$ . To see this, assume the sought-after result holds for a single labeled input, and suppose  $|\mathcal{S}| > 1$ . We may then apply the result separately for each labeled input  $(\mathbf{x}_i, y_i)$ ,  $i = 1, 2, \dots, |\mathcal{S}|$ , and obtain, for every admissible  $\boldsymbol{\theta}' \in \mathbb{R}^d$ , activation matrices  $(D'_j{}^{(\mathbf{x}_i, y_i)})_{j=1}^{n-1}$  and a differentiability region  $\mathcal{D}_{\boldsymbol{\theta}'}{}^{(\mathbf{x}_i, y_i)}$ . Since the weight settings not admissible for a certain labeled input  $(\mathbf{x}_i, y_i)$  form a set of zero (Lebesgue) measure, those not admissible for any of the  $|\mathcal{S}|$  labeled inputs also constitute a zero measure set. That is, almost every  $\boldsymbol{\theta}' \in \mathbb{R}^d$  is jointly admissible for all  $((\mathbf{x}_i, y_i))_{i=1}^{|\mathcal{S}|}$ . Given such  $\boldsymbol{\theta}'$ , consider the activation matrices and differentiability regions obtained for the different labeled inputs —  $(D'_j{}^{(\mathbf{x}_i, y_i)})_{j=1}^{n-1}$  and  $\mathcal{D}_{\boldsymbol{\theta}'}{}^{(\mathbf{x}_i, y_i)}$ ,  $i = 1, 2, \dots, |\mathcal{S}|$ . Defining  $D'_{i,j} := D'_j{}^{(\mathbf{x}_i, y_i)}$ ,  $i = 1, 2, \dots, |\mathcal{S}|$ ,  $j = 1, 2, \dots, n-1$ , and  $\mathcal{D}_{\boldsymbol{\theta}'} := \bigcap_{i=1}^{|\mathcal{S}|} \mathcal{D}_{\boldsymbol{\theta}'}{}^{(\mathbf{x}_i, y_i)}$ , we have that  $\boldsymbol{\theta}'$  is admissible for  $\mathcal{S}$ , with activation matrices  $(D'_{i,j})_{i,j}$  and differentiability region  $\mathcal{D}_{\boldsymbol{\theta}'}$ . The sought-after result thus holds for  $\mathcal{S}$ .

In light of the above, we assume hereafter that  $\mathcal{S} = ((\mathbf{x}, y))$ . Recursively define the functions  $\mathbf{f}^{(j)} : \mathbb{R}^d \rightarrow \mathbb{R}^{d_j}$ ,  $j = 0, 1, \dots, n-1$ :

$$\mathbf{f}^{(0)}(\boldsymbol{\theta}) \equiv \mathbf{x} \quad , \quad \mathbf{f}^{(j)}(\boldsymbol{\theta}) = \sigma(W_j \mathbf{f}^{(j-1)}(\boldsymbol{\theta})) \quad \text{for } j = 1, 2, \dots, n-1 .$$

<sup>30</sup>Note that given an admissible weight setting, activation matrices and differentiability region are not necessarily determined uniquely.

We will prove by induction that given  $j' \in \{0, 1, \dots, n-1\}$ , for almost every  $\theta' \in \mathbb{R}^d$ , there exist diagonal matrices  $D'_j \in \mathbb{R}^{d_j, d_j}$ ,  $j = 1, 2, \dots, j'$ , with diagonal elements in  $\{\alpha, \bar{\alpha}\}$ , such that  $\mathbf{f}^{(j')}(\cdot)$  meets the following conditions on an open region  $\mathcal{D}_{\theta'} \subseteq \mathbb{R}^d$  containing  $\theta'$ , that is closed under positive rescaling of weight matrices:

- (i)  $\mathbf{f}^{(j')}(\cdot)$  coincides with the function  $\theta \mapsto D'_{j'} W_{j'} D'_{j'-1} W_{j'-1} \dots D'_1 W_1 \mathbf{x}$ ; and
- (ii) each entry of  $\mathbf{f}^{(j')}(\cdot)$  is either nowhere zero or identically zero.

Continuing the terminology defined earlier, in the context of  $\mathbf{f}^{(j')}(\cdot)$ ,  $j' = 0, 1, \dots, n-1$ , we refer to  $\theta'$ ,  $(D'_j)_j$  and  $\mathcal{D}_{\theta'}$  satisfying the above as *admissible*, *activation matrices* and *differentiability region*, respectively. Note that the training loss  $f(\cdot)$  can be expressed as  $f(\theta) = \ell(W_n \mathbf{f}^{(n-1)}(\theta), y)$ , and therefore proving the inductive hypothesis for  $j' = n-1$  yields the desired result. The base case for the induction ( $j' = 0$ ) is trivial, so all that remains is to establish the induction step.

Given  $j' \in \{1, 2, \dots, n-1\}$ , assume that the inductive hypothesis holds for  $j' - 1$ , and in the context of  $\mathbf{f}^{(j'-1)}(\cdot)$ , let  $\theta'$  be an admissible weight setting, with corresponding activation matrices  $(D'_j)_{j=1}^{j'-1}$  and differentiability region  $\mathcal{D}_{\theta'}$ . We refer to  $\theta'$  as *nullifying* if  $\mathbf{f}^{(j'-1)}(\theta') = \mathbf{0}$ , which implies  $\mathbf{f}^{(j'-1)}(\theta) = \mathbf{0}$  for all  $\theta \in \mathcal{D}_{\theta'}$ . In this case  $\theta'$  is clearly admissible in the context of  $\mathbf{f}^{(j')}(\cdot)$  (as activation matrices we may take  $(D'_j)_{j=1}^{j'-1}$  along with any diagonal matrix  $D'_{j'} \in \mathbb{R}^{d_{j'}, d_{j'}}$  whose diagonal elements are in  $\{\alpha, \bar{\alpha}\}$ , and as differentiability region we can simply use  $\mathcal{D}_{\theta'}$ ). Consider now the case where  $\theta'$  is non-nullifying, i.e. where  $\mathbf{f}^{(j'-1)}(\theta') \neq \mathbf{0}$ . We refer to  $\theta'$  as *regular* if all entries of  $W'_{j'} \mathbf{f}^{(j'-1)}(\theta')$  are non-zero, with  $W'_{j'} \in \mathbb{R}^{d_{j'}, d_{j'-1}}$  denoting the value of weight matrix  $j'$  held in  $\theta'$ . If  $\theta'$  is regular then it is admissible in the context of  $\mathbf{f}^{(j')}(\cdot)$ . To see this, note that a valid choice of activation matrices is  $(D'_j)_{j=1}^{j'-1}$  along with the diagonal matrix  $D'_{j'} \in \mathbb{R}^{d_{j'}, d_{j'}}$  whose diagonal elements corresponding to positive entries of  $W'_{j'} \mathbf{f}^{(j'-1)}(\theta')$  hold  $\alpha$ , and those corresponding to negative entries hold  $\bar{\alpha}$ . From continuity, and homogeneity with slopes  $\alpha$  and  $\bar{\alpha}$  of the activation function  $\sigma(\cdot)$ , there exists an open neighborhood of  $\theta'$  (subset of  $\mathcal{D}_{\theta'}$ ) on which conditions (i) and (ii) hold. Extending this neighborhood to include, for each of its weight settings  $\theta$ , all positive rescalings of weight matrices  $W_1, W_2, \dots, W_n$ , yields a valid differentiability region for  $\theta'$  in the context of  $\mathbf{f}^{(j')}(\cdot)$ , thereby confirming admissibility.

We conclude the proof by showing that almost every  $\theta' \in \mathbb{R}^d$  is admissible in the context of  $\mathbf{f}^{(j')}(\cdot)$ . Per the above, if  $\theta' \in \mathbb{R}^d$  does not meet this condition then it must either be inadmissible in the context of  $\mathbf{f}^{(j'-1)}(\cdot)$ , or be non-nullifying and irregular. By our inductive hypothesis, weight settings inadmissible in the context of  $\mathbf{f}^{(j'-1)}(\cdot)$  form a set of measure zero, so it suffices to show that the collection of non-nullifying and irregular weight settings, denoted  $\mathcal{C}$ , is also of measure zero. Note that whether a weight setting  $\theta$  is nullifying (i.e.  $\mathbf{f}^{(j'-1)}(\theta) = \mathbf{0}$ ) or not depends only on the weight matrices  $W_1, W_2, \dots, W_{j'-1}$ , and given these matrices, whether it is regular (i.e. all entries of  $W'_{j'} \mathbf{f}^{(j'-1)}(\theta')$  are non-zero) or not depends only on  $W_{j'}$ . We may thus apply Fubini's Theorem (cf. [48]), and compute the measure of  $\mathcal{C}$  by integrating over non-nullifying configurations of  $W_1, W_2, \dots, W_{j'-1}$ , where for each, the measure of values for  $W_{j'}, W_{j'+1}, \dots, W_n$  leading to irregularity is integrated. The latter measure is zero, since for any  $\mathbf{0} \neq \mathbf{q} \in \mathbb{R}^{d_{j'-1}}$ , the set  $\{W \in \mathbb{R}^{d_{j'}, d_{j'-1}} : \text{there exists a coordinate of } W\mathbf{q} \text{ equal to zero}\}$  has measure zero, thus its Cartesian product with  $\mathbb{R}^{d_{j'+1}, d_{j'}} \times \mathbb{R}^{d_{j'+2}, d_{j'+1}} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$  is also of measure zero. This implies that  $\mathcal{C}$  has measure zero, thereby completing the proof.  $\square$

**Proposition 10** (regions of differentiability for convolutional architectures). *Consider a neural network with weight sharing and sparsity as defined in Equation (28) (and surrounding text), and assume that its (homogeneous) activation function is non-linear, i.e.  $\sigma(z) = \alpha \max\{z, 0\} - \bar{\alpha} \max\{-z, 0\}$  for some  $\alpha, \bar{\alpha} \in \mathbb{R}$ ,  $\alpha \neq \bar{\alpha}$ . Then, for almost every (in the sense of Lebesgue measure)  $\theta' \in \mathbb{R}^d$ , there exist diagonal matrices  $D'_{i,j} \in \mathbb{R}^{d_j, d_j}$ ,  $i = 1, 2, \dots, |\mathcal{S}|$ ,  $j = 1, 2, \dots, n-1$ , with diagonal elements in  $\{\alpha, \bar{\alpha}\}$ , such that the training loss  $f(\cdot)$  (Equation (9)) coincides with the function defined in Equation (29) on an open region  $\mathcal{D}_{\theta'} \subseteq \mathbb{R}^d$  containing  $\theta'$ , that is closed under positive rescaling of weight vectors (i.e. under  $(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n) \mapsto (c_1 \mathbf{w}_1, c_2 \mathbf{w}_2, \dots, c_n \mathbf{w}_n)$  with  $c_1, c_2, \dots, c_n > 0$ ).*

*Proof.* The proof begins similarly to that of Proposition 9, and then takes a slightly different (more involved) route. We provide a self-contained presentation, repeating details from the proof of Proposition 9 as needed.

If for  $\theta' \in \mathbb{R}^d$  there exist diagonal matrices  $(D'_{i,j})_{i,j}$  and an open region  $\mathcal{D}_{\theta'}$  as in proposition statement, then we refer to  $\theta'$  as an *admissible* weight setting, to  $(D'_{i,j})_{i,j}$  as its *activation matrices*, and to  $\mathcal{D}_{\theta'}$  as its *differentiability region*.<sup>30</sup>

Without loss of generality, we may assume  $|\mathcal{S}|=1$ , i.e. that the training set comprises a single labeled input  $(\mathbf{x}, y) \in \mathbb{R}^{d_0} \times \mathcal{Y}$ , meaning the training loss takes the form  $f(\theta) = \ell(h_{\theta}(\mathbf{x}), y)$ . To see this, assume the sought-after result holds for a single labeled input, and suppose  $|\mathcal{S}|>1$ . We may then apply the result separately for each labeled input  $(\mathbf{x}_i, y_i)$ ,  $i = 1, 2, \dots, |\mathcal{S}|$ , and obtain, for every admissible  $\theta' \in \mathbb{R}^d$ , activation matrices  $(D'_j{}^{(\mathbf{x}_i, y_i)})_{j=1}^{n-1}$  and a differentiability region  $\mathcal{D}_{\theta'}{}^{(\mathbf{x}_i, y_i)}$ . Since the weight settings not admissible for a certain labeled input  $(\mathbf{x}_i, y_i)$  form a set of zero (Lebesgue) measure, those not admissible for any of the  $|\mathcal{S}|$  labeled inputs also constitute a zero measure set. That is, almost every  $\theta' \in \mathbb{R}^d$  is jointly admissible for all  $((\mathbf{x}_i, y_i))_{i=1}^{|\mathcal{S}|}$ . Given such  $\theta'$ , consider the activation matrices and differentiability regions obtained for the different labeled inputs —  $(D'_j{}^{(\mathbf{x}_i, y_i)})_{j=1}^{n-1}$  and  $\mathcal{D}_{\theta'}{}^{(\mathbf{x}_i, y_i)}$ ,  $i = 1, 2, \dots, |\mathcal{S}|$ . Defining  $D'_{i,j} := D'_j{}^{(\mathbf{x}_i, y_i)}$ ,  $i = 1, 2, \dots, |\mathcal{S}|$ ,  $j = 1, 2, \dots, n-1$ , and  $\mathcal{D}_{\theta'} := \cap_{i=1}^{|\mathcal{S}|} \mathcal{D}_{\theta'}{}^{(\mathbf{x}_i, y_i)}$ , we have that  $\theta'$  is admissible for  $\mathcal{S}$ , with activation matrices  $(D'_{i,j})_{i,j}$  and differentiability region  $\mathcal{D}_{\theta'}$ . The sought-after result thus holds for  $\mathcal{S}$ .

In light of the above, we assume hereafter that  $\mathcal{S} = ((\mathbf{x}, y))$ . Recursively define the functions  $\mathbf{f}^{(j)} : \mathbb{R}^d \rightarrow \mathbb{R}^{d_j}$ ,  $j = 0, 1, \dots, n-1$ :

$$\mathbf{f}^{(0)}(\theta) \equiv \mathbf{x} \quad , \quad \mathbf{f}^{(j)}(\theta) = \sigma(W_j(\mathbf{w}_j)\mathbf{f}^{(j-1)}(\theta)) \quad \text{for } j = 1, 2, \dots, n-1.$$

We will prove by induction that given  $j' \in \{0, 1, \dots, n-1\}$ , for almost every  $\theta' \in \mathbb{R}^d$ , there exist diagonal matrices  $D'_j \in \mathbb{R}^{d_j, d_j}$ ,  $j = 1, 2, \dots, j'$ , with diagonal elements in  $\{\alpha, \bar{\alpha}\}$ , such that  $\mathbf{f}^{(j')}(\cdot)$  meets the following conditions on an open region  $\mathcal{D}_{\theta'} \subseteq \mathbb{R}^d$  containing  $\theta'$ , that is closed under positive rescaling of weight vectors:

- (i)  $\mathbf{f}^{(j')}(\cdot)$  coincides with the function  $\theta \mapsto D'_j W_{j'}(\mathbf{w}_{j'}) D'_{j'-1} W_{j'-1}(\mathbf{w}_{j'-1}) \cdots D'_1 W_1(\mathbf{w}_1) \mathbf{x}$ ; and
- (ii) each entry of  $\mathbf{f}^{(j')}(\cdot)$  is either nowhere zero or identically zero.

Continuing the terminology defined earlier, in the context of  $\mathbf{f}^{(j')}(\cdot)$ ,  $j'=0, 1, \dots, n-1$ , we refer to  $\theta'$ ,  $(D'_j)_j$  and  $\mathcal{D}_{\theta'}$  satisfying the above as *admissible*, *activation matrices* and *differentiability region*, respectively. Note that the training loss  $f(\cdot)$  can be expressed as  $f(\theta) = \ell(W_n(\mathbf{w}_n)\mathbf{f}^{(n-1)}(\theta), y)$ , and therefore proving the inductive hypothesis for  $j' = n-1$  yields the desired result. The base case for the induction ( $j' = 0$ ) is trivial, so all that remains is to establish the induction step.

Given  $j' \in \{1, 2, \dots, n-1\}$ , assume that the inductive hypothesis holds for  $j'-1$ , and in the context of  $\mathbf{f}^{(j'-1)}(\cdot)$ , let  $\theta'$  be an admissible weight setting, with corresponding activation matrices  $(D'_j)_{j=1}^{j'-1}$  and differentiability region  $\mathcal{D}_{\theta'}$ . We define the *nullity pattern* of  $\theta'$  to be the vector  $\mathbf{e} \in \mathbb{R}^{d_{j'-1}}$  holding zero in the coordinates where  $\mathbf{f}^{(j'-1)}(\theta')$  holds zero, and one elsewhere (that is,  $\mathbf{e}$  is the vector obtained by setting to one all non-zero entries of  $\mathbf{f}^{(j'-1)}(\theta')$ ). With  $\mathbf{1} \in \mathbb{R}^{d_{j'}}$  standing for an all-ones vector, we refer to the coordinates of  $\mathbb{R}^{d_{j'}}$  where  $W_{j'}(\mathbf{1})\mathbf{e}$  holds zero as *infeasible*, and to the rest as *feasible*. Note that a coordinate of  $\mathbb{R}^{d_{j'}}$  is infeasible if and only if  $W_{j'}(\mathbf{q})\mathbf{f}^{(j'-1)}(\theta')$  holds zero in that coordinate for all  $\mathbf{q} \in \mathbb{R}^{d_{j'-1}}$ . We shall say that  $\theta'$  is *regular* if  $W_{j'}(\mathbf{w}'_{j'})\mathbf{f}^{(j'-1)}(\theta')$  is non-zero in all feasible coordinates, where  $\mathbf{w}'_{j'} \in \mathbb{R}^{d_{j'-1}}$  denotes the value of weight vector  $j'$  in  $\theta'$ . Hereafter we show that regularity of  $\theta'$  implies that it is admissible in the context of  $\mathbf{f}^{(j')}(\cdot)$ . By admissibility in the context of  $\mathbf{f}^{(j'-1)}(\cdot)$  we have that across  $\mathcal{D}_{\theta'}$ , each entry of  $\mathbf{f}^{(j'-1)}(\cdot)$  is either nowhere zero or identically zero. This implies the nullity pattern is constant across  $\mathcal{D}_{\theta'}$ , which in turn means the same for the set of infeasible coordinates. The coordinates where  $W_{j'}(\mathbf{w}'_{j'})\mathbf{f}^{(j'-1)}(\theta')$  holds zero thus vanish in  $W_{j'}(\mathbf{w}_{j'})\mathbf{f}^{(j'-1)}(\theta)$  for all  $\theta \in \mathcal{D}_{\theta'}$ . From continuity, and the fact that around any  $z \neq 0$ , the activation function  $\sigma(\cdot)$  is either nowhere zero or identically zero,<sup>31</sup> it follows that there exists an open neighborhood  $\mathcal{N} \subseteq \mathcal{D}_{\theta'}$  of  $\theta'$  on which condition (ii) holds. Let  $D'_{j'} \in \mathbb{R}^{d_{j'}, d_{j'}}$  be a diagonal matrix whose diagonal elements corresponding to positive entries in  $W_{j'}(\mathbf{w}'_{j'})\mathbf{f}^{(j'-1)}(\theta')$

<sup>31</sup>The latter is possible only if  $\alpha = 0$  or  $\bar{\alpha} = 0$ .



hold  $\alpha$ , those corresponding to negative entries hold  $\bar{\alpha}$ , and the rest hold either  $\alpha$  or  $\bar{\alpha}$ . Since  $\mathbf{f}^{(j'-1)}(\cdot)$  coincides with the function  $\boldsymbol{\theta} \mapsto D'_{j'-1} W_{j'-1}(\mathbf{w}_{j'-1}) D'_{j'-2} W_{j'-2}(\mathbf{w}_{j'-2}) \cdots D'_1 W_1(\mathbf{w}_1) \mathbf{x}$  on  $\mathcal{D}_{\boldsymbol{\theta}'}$ , and since  $\sigma(\cdot)$  is homogeneous with slopes  $\alpha$  and  $\bar{\alpha}$ , condition (i) holds across  $\mathcal{N}$ . Consider the extension of  $\mathcal{N}$  comprising, for each of its weight settings, all positive rescalings of weight vectors. Along with  $(D'_j)_{j=1}^{j'}$  as activation matrices, this extension serves as a valid differentiability region for  $\boldsymbol{\theta}'$  in the context of  $\mathbf{f}^{(j')}(\cdot)$ . The sought-after admissibility is thus established.

We conclude the proof by showing that almost every  $\boldsymbol{\theta}' \in \mathbb{R}^d$  is admissible in the context of  $\mathbf{f}^{(j')}(\cdot)$ . Per the above, if  $\boldsymbol{\theta}' \in \mathbb{R}^d$  does not meet this condition then either it is inadmissible in the context of  $\mathbf{f}^{(j'-1)}(\cdot)$ , or it is irregular. By our inductive hypothesis, weight settings inadmissible in the context of  $\mathbf{f}^{(j'-1)}(\cdot)$  form a set of measure zero, so it suffices to show that the collection of irregular weight settings, denoted  $\mathcal{C}$ , is also of measure zero. We first establish that  $\mathcal{C}$  is measurable. Let  $\mathbf{e} \in \mathbb{R}^{d_{j'-1}}$  be an arbitrary nullity pattern (vector with entries in  $\{0,1\}$ ), and consider the feasible coordinates it induces. The following two sets are measurable: weight settings with nullity pattern  $\mathbf{e}$ ; and weight settings  $\boldsymbol{\theta}$  for which  $W_{j'}(\mathbf{w}_{j'}) \mathbf{f}^{(j'-1)}(\boldsymbol{\theta})$  holds zero in at least one of the feasible coordinates induced by  $\mathbf{e}$ . The collection of irregular weight settings with nullity pattern  $\mathbf{e}$ , denoted  $\mathcal{C}_{\mathbf{e}}$ , is equal to the intersection of these two sets, and therefore is measurable. Taking union of  $\mathcal{C}_{\mathbf{e}}$  with  $\mathbf{e}$  ranging over all (finitely many) possible nullity patterns yields  $\mathcal{C}$ , from which it follows that the latter is indeed measurable. Given weight vectors  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{j'-1}$ , whether or not a weight setting  $\boldsymbol{\theta}$  is regular depends only on  $\mathbf{w}_{j'}$ . We may thus apply Fubini's Theorem (cf. [48]), and compute the measure of  $\mathcal{C}$  by integrating over configurations of  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{j'-1}$ , where for each, the measure of values for  $\mathbf{w}_{j'}, \mathbf{w}_{j'+1}, \dots, \mathbf{w}_n$  leading to irregularity is integrated. We now establish that the latter measure is zero, which in turn implies that  $\mathcal{C}$  has measure zero (thereby completing the proof). Since the Cartesian product of a zero measure subset of  $\mathbb{R}^{d_{j'}}$  with  $\mathbb{R}^{d_{j'+1}} \times \mathbb{R}^{d_{j'+2}} \times \dots \times \mathbb{R}^{d_n}$  has zero measure, it suffices to show that given any configuration of  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{j'-1}$ , the measure of values for  $\mathbf{w}_{j'}$  leading to irregularity is zero.  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{j'-1}$  fully determine  $\mathbf{f}^{(j'-1)}(\boldsymbol{\theta})$ , and as a consequence, the nullity pattern of  $\boldsymbol{\theta}$ . Consider the feasible coordinates induced by this nullity pattern. On each of these, the linear function  $\mathbf{w}_{j'} \mapsto W_{j'}(\mathbf{w}_{j'}) \mathbf{f}^{(j'-1)}(\boldsymbol{\theta})$  is not identically zero. The measure of values for  $\mathbf{w}_{j'}$  leading  $W_{j'}(\mathbf{w}_{j'}) \mathbf{f}^{(j'-1)}(\boldsymbol{\theta})$  to vanish in a feasible coordinate, i.e. leading  $\boldsymbol{\theta}$  to be irregular, is thus zero. This completes the proof.  $\square$

## F Necessity of Assumptions in Propositions 1, 5 and 7

In this appendix we prove that the assumptions in Propositions 1, 5 and 7 are necessary, in the sense that each of the latter becomes false if any of its assumptions are removed (and no further assumptions are imposed).

**Claim 1** (necessity of assumptions in Proposition 1). *In the context of Proposition 1, if the network is shallow ( $n=2$ ) or the zero mapping is a global minimizer of the training loss (meaning  $\nabla \phi(0) = 0$ ), then the stated result may not hold, i.e. it may be that  $\inf_{\boldsymbol{\theta} \in \mathbb{R}^d} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) > -\infty$ .*

*Proof.* Suppose the network is shallow ( $n=2$ ). With the notations of Lemma 1, for any  $\boldsymbol{\theta} \in \mathbb{R}^d$ ,  $(\Delta W_1, \Delta W_2) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1}$ :

$$\begin{aligned} \nabla^2 f(\boldsymbol{\theta})[\Delta W_1, \Delta W_2] &= \nabla^2 \phi(W_{2:1})[W_2(\Delta W_1) + (\Delta W_2)W_1] + 2\text{Tr}(\nabla \phi(W_{2:1})^\top (\Delta W_2)(\Delta W_1)) \\ &\geq 2\text{Tr}(\nabla \phi(W_{2:1})^\top (\Delta W_2)(\Delta W_1)) \\ &\geq -2\|\nabla \phi(W_{2:1})\|_{\text{Frobenius}} \|(\Delta W_2)(\Delta W_1)\|_{\text{Frobenius}} \\ &\geq -2\|\nabla \phi(W_{2:1})\|_{\text{Frobenius}} \|\Delta W_2\|_{\text{Frobenius}} \|\Delta W_1\|_{\text{Frobenius}} \\ &\geq -\|\nabla \phi(W_{2:1})\|_{\text{Frobenius}} (\|\Delta W_2\|_{\text{Frobenius}}^2 + \|\Delta W_1\|_{\text{Frobenius}}^2) \\ &= -\|\nabla \phi(W_{2:1})\|_{\text{Frobenius}} \|(\Delta W_1, \Delta W_2)\|_{\text{Frobenius}}^2, \end{aligned}$$

where the first transition follows from Lemma 1, the second holds since  $\phi(\cdot)$  is convex, the third is an application of the Cauchy-Schwarz inequality, the fourth follows from submultiplicativity of the Frobenius norm, and the latter two are based on simple arithmetics. It follows from the above that  $\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -\|\nabla \phi(W_{2:1})\|_{\text{Frobenius}}$ . Therefore if  $\nabla \phi(\cdot)$  is bounded (e.g. if  $\ell(\cdot)$  is the logistic loss — see Equation (11)) we will have  $\inf_{\boldsymbol{\theta} \in \mathbb{R}^d} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) > -\infty$ , as required.

It remains to show that if the zero mapping is a global minimizer of the training loss (meaning  $\nabla \phi(0) = 0$ ), then, regardless of network depth (i.e. with either  $n \geq 3$  or  $n = 2$ ), it may be that

$\inf_{\boldsymbol{\theta} \in \mathbb{R}^d} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) > -\infty$ . This is trivial — simply consider the case where the training set  $\mathcal{S}$  is such that  $\mathbf{x}_i = \mathbf{0}$  for all  $i = 1, 2, \dots, |\mathcal{S}|$ . The training loss in this case is constant (see Equations (8) and (9)), implying  $\inf_{\boldsymbol{\theta} \in \mathbb{R}^d} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) = 0$ .  $\square$

**Claim 2** (necessity of assumptions in Proposition 5). *In the context of Proposition 5, if assumptions (i) or (ii) are not satisfied, then the stated result may not hold, i.e. it may be that  $\inf_{\boldsymbol{\theta} \in \mathbb{R}^d \text{ s.t. } \nabla^2 f(\boldsymbol{\theta}) \text{ exists}} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) > -\infty$ .*

*Proof.* Suppose that assumption (i) is not satisfied, i.e. that the network is shallow ( $n = 2$ ). With the notations of Lemma 4, for any  $\boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\theta}'}$ ,  $(\Delta W_1, \Delta W_2) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1}$ :

$$\begin{aligned}
\nabla^2 f(\boldsymbol{\theta})[\Delta W_1, \Delta W_2] &= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i [W_2 D'_{i,1}(\Delta W_1) \mathbf{x}_i + (\Delta W_2) D'_{i,1} W_1 \mathbf{x}_i] \\
&\quad + \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i^\top (\Delta W_2) D'_{i,1} (\Delta W_1) \mathbf{x}_i \\
&\geq \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i^\top (\Delta W_2) D'_{i,1} (\Delta W_1) \mathbf{x}_i \\
&\geq -\frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|(\Delta W_2) D'_{i,1} (\Delta W_1) \mathbf{x}_i\|_2 \\
&\geq -\frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \|(\Delta W_2) D'_{i,1} (\Delta W_1)\|_{\text{spectral}} \\
&\geq -\frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \|\Delta W_2\|_{\text{spectral}} \|D'_{i,1}\|_{\text{spectral}} \|\Delta W_1\|_{\text{spectral}} \\
&\geq -\max\{|\alpha|, |\bar{\alpha}|\} \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \|\Delta W_2\|_{\text{spectral}} \|\Delta W_1\|_{\text{spectral}} \\
&\geq -\max\{|\alpha|, |\bar{\alpha}|\} \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \|\Delta W_2\|_{\text{Frobenius}} \|\Delta W_1\|_{\text{Frobenius}} \\
&\geq -\max\{|\alpha|, |\bar{\alpha}|\} \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 (\|\Delta W_2\|_{\text{Frobenius}}^2 + \|\Delta W_1\|_{\text{Frobenius}}^2) \\
&= -\max\{|\alpha|, |\bar{\alpha}|\} \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \|(\Delta W_1, \Delta W_2)\|_{\text{Frobenius}}^2,
\end{aligned}$$

where the first transition follows from Lemma 4, the second holds since  $\ell(\cdot)$  is convex with respect to its first argument (recall from Lemma 4 that  $\nabla^2 \ell_i$  is defined to be the Hessian of  $\ell(\cdot)$  at the point  $(W_2 D'_{i,1} W_1 \mathbf{x}_i, y_i)$  with respect to its first argument), the third is an application of the Cauchy-Schwarz inequality, the fourth follows from the spectral norm being the operator norm induced by the Euclidean norm, the fifth is due to submultiplicativity of the spectral norm, the sixth results from  $D'_{i,1}$  being diagonal with diagonal elements in  $\{\alpha, \bar{\alpha}\}$ , the seventh holds since spectral norm is upper bounded by Frobenius norm, and the latter two are based on simple arithmetics. It follows from the above that  $\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -\max\{|\alpha|, |\bar{\alpha}|\} \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2$ .

Consider the case where the gradient of  $\ell(\cdot)$  with respect to its first argument has Euclidean norm bounded by some constant  $c > 0$  (this holds, for example, if  $\ell(\cdot)$  is the logistic loss). Recalling (from Lemma 4) that  $\nabla \ell_i$  stands for this gradient at the point  $(W_2 D'_{i,1} W_1 \mathbf{x}_i, y_i)$ , we obtain  $\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -c \max\{|\alpha|, |\bar{\alpha}|\} \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\mathbf{x}_i\|_2$ . The latter holds for any  $\boldsymbol{\theta}$  belonging to any region of the form  $\mathcal{D}_{\boldsymbol{\theta}'}$ . Since these regions constitute the entire weight space but a zero measure set, and since by definition existence of  $\nabla^2 f(\boldsymbol{\theta})$  for some  $\boldsymbol{\theta} \in \mathbb{R}^d$  implies that  $f(\cdot)$  is twice continuously differentiable (and therefore  $\lambda_{\min}(\nabla^2 f(\cdot))$  is continuous) on a neighborhood of  $\boldsymbol{\theta}$ , it necessarily holds that  $\inf_{\boldsymbol{\theta} \in \mathbb{R}^d \text{ s.t. } \nabla^2 f(\boldsymbol{\theta}) \text{ exists}} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -c \max\{|\alpha|, |\bar{\alpha}|\} \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\mathbf{x}_i\|_2 > -\infty$ . This establishes necessity of assumption (i).

It remains to show that if assumption (ii) is not satisfied, i.e. if  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) = 0$  for all  $\boldsymbol{\theta} \in \mathbb{R}^d$ , then, regardless of whether or not assumption (i) holds (i.e. of whether  $n \geq 3$  or  $n = 2$ ), it may be that  $\inf_{\boldsymbol{\theta} \in \mathbb{R}^d \text{ s.t. } \nabla^2 f(\boldsymbol{\theta}) \text{ exists}} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) > -\infty$ . This is trivial — simply consider the case where the training set  $\mathcal{S}$  is such that  $\mathbf{x}_i = \mathbf{0}$  for all  $i = 1, 2, \dots, |\mathcal{S}|$ . The training loss in this case is constant (see Equations (8) and (9)), implying  $\inf_{\boldsymbol{\theta} \in \mathbb{R}^d \text{ s.t. } \nabla^2 f(\boldsymbol{\theta}) \text{ exists}} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) = 0$ .  $\square$

**Claim 3** (necessity of assumptions in Proposition 7). *In the context of Proposition 7, if assumptions (i) or (ii) are not satisfied, then the stated result may not hold, i.e. it may be that  $\inf_{\boldsymbol{\theta} \in \mathbb{R}^d \text{ s.t. } \nabla^2 f(\boldsymbol{\theta}) \text{ exists}} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) > -\infty$ .*

*Proof.* Suppose that assumption (i) is not satisfied, *i.e.* that the network is shallow ( $n=2$ ). With the notations of Lemmas 6 and 7, for any  $\boldsymbol{\theta} \in \mathcal{D}_{\theta'}$ ,  $(\Delta \mathbf{w}_1, \Delta \mathbf{w}_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ :

$$\begin{aligned}
\nabla^2 f(\boldsymbol{\theta})[\Delta \mathbf{w}_1, \Delta \mathbf{w}_2] &= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i [W_2(\mathbf{w}_2) D'_{i,1} W_1(\Delta \mathbf{w}_1) \mathbf{x}_i + W_2(\Delta \mathbf{w}_2) D'_{i,1} W_1(\mathbf{w}_1) \mathbf{x}_i] \\
&\quad + \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i^\top W_2(\Delta \mathbf{w}_2) D'_{i,1} W_1(\Delta \mathbf{w}_1) \mathbf{x}_i \\
&\geq \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i^\top W_2(\Delta \mathbf{w}_2) D'_{i,1} W_1(\Delta \mathbf{w}_1) \mathbf{x}_i \\
&\geq -\frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|W_2(\Delta \mathbf{w}_2) D'_{i,1} W_1(\Delta \mathbf{w}_1) \mathbf{x}_i\|_2 \\
&\geq -\frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \|W_2(\Delta \mathbf{w}_2) D'_{i,1} W_1(\Delta \mathbf{w}_1)\|_{\text{spectral}} \\
&\geq -\frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \|W_2(\Delta \mathbf{w}_2)\|_{\text{spectral}} \|D'_{i,1}\|_{\text{spectral}} \|W_1(\Delta \mathbf{w}_1)\|_{\text{spectral}} \\
&\geq -\max\{|\alpha|, |\bar{\alpha}|\} \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \|W_2(\Delta \mathbf{w}_2)\|_{\text{spectral}} \|W_1(\Delta \mathbf{w}_1)\|_{\text{spectral}} \\
&\geq -\max\{|\alpha|, |\bar{\alpha}|\} \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \|W_2(\Delta \mathbf{w}_2)\|_{\text{Frobenius}} \|W_1(\Delta \mathbf{w}_1)\|_{\text{Frobenius}} \\
&\geq -\max\{|\alpha|, |\bar{\alpha}|\} \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \|W_2(\cdot)\|_{\text{op}} \|\Delta \mathbf{w}_2\|_2 \|W_1(\cdot)\|_{\text{op}} \|\Delta \mathbf{w}_1\|_2 \\
&\geq -\max\{|\alpha|, |\bar{\alpha}|\} \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \|W_2(\cdot)\|_{\text{op}} \|W_1(\cdot)\|_{\text{op}} (\|\Delta \mathbf{w}_2\|_2^2 + \|\Delta \mathbf{w}_1\|_2^2) \\
&= -\max\{|\alpha|, |\bar{\alpha}|\} \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \prod_{j=1}^2 \|W_j(\cdot)\|_{\text{op}} \|(\Delta \mathbf{w}_1, \Delta \mathbf{w}_2)\|_{\text{Frobenius}}^2,
\end{aligned}$$

where the first transition follows from Lemma 6, the second holds since  $\ell(\cdot)$  is convex with respect to its first argument (recall from Lemma 6 that  $\nabla^2 \ell_i$  is defined to be the Hessian of  $\ell(\cdot)$  at the point  $(W_2(\mathbf{w}_1) D'_{i,1} W_1(\mathbf{w}_1) \mathbf{x}_i, y_i)$  with respect to its first argument), the third is an application of the Cauchy-Schwarz inequality, the fourth follows from the spectral norm being the operator norm induced by the Euclidean norm, the fifth is due to submultiplicativity of the spectral norm, the sixth results from  $D'_{i,1}$  being diagonal with diagonal elements in  $\{\alpha, \bar{\alpha}\}$ , the seventh holds since spectral norm is upper bounded by Frobenius norm, the eighth is due to the definition of  $\|W_j(\cdot)\|_{\text{op}}$  (operator norm of  $W_j(\cdot)$  induced by the Frobenius norm), and the latter two are based on simple arithmetics. The above implies that  $\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -\max\{|\alpha|, |\bar{\alpha}|\} \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \prod_{j=1}^2 \|W_j(\cdot)\|_{\text{op}}$ . Consider the case where the gradient of  $\ell(\cdot)$  with respect to its first argument has Euclidean norm bounded by some constant  $c > 0$  (this holds, for example, if  $\ell(\cdot)$  is the logistic loss). Recalling (from Lemma 6) that  $\nabla \ell_i$  stands for this gradient at the point  $(W_2(\mathbf{w}_2) D'_{i,1} W_1(\mathbf{w}_1) \mathbf{x}_i, y_i)$ , we obtain  $\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -c \max\{|\alpha|, |\bar{\alpha}|\} \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\mathbf{x}_i\|_2 \prod_{j=1}^2 \|W_j(\cdot)\|_{\text{op}}$ . The latter holds for any  $\boldsymbol{\theta}$  belonging to any region of the form  $\mathcal{D}_{\theta'}$ . Since these regions constitute the entire weight space but a zero measure set, and since by definition existence of  $\nabla^2 f(\boldsymbol{\theta})$  for some  $\boldsymbol{\theta} \in \mathbb{R}^d$  implies that  $f(\cdot)$  is twice continuously differentiable (and therefore  $\lambda_{\min}(\nabla^2 f(\cdot))$  is continuous) on a neighborhood of  $\boldsymbol{\theta}$ , it necessarily holds that:

$$\inf_{\boldsymbol{\theta} \in \mathbb{R}^d \text{ s.t. } \nabla^2 f(\boldsymbol{\theta}) \text{ exists}} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -c \max\{|\alpha|, |\bar{\alpha}|\} \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\mathbf{x}_i\|_2 \prod_{j=1}^2 \|W_j(\cdot)\|_{\text{op}} > -\infty.$$

This establishes necessity of assumption (i).

It remains to show that if assumption (ii) is not satisfied, *i.e.* if  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) = 0$  for all  $\boldsymbol{\theta} \in \mathbb{R}^d$ , then, regardless of whether or not assumption (i) holds (*i.e.* of whether  $n \geq 3$  or  $n=2$ ), it may be that  $\inf_{\boldsymbol{\theta} \in \mathbb{R}^d \text{ s.t. } \nabla^2 f(\boldsymbol{\theta}) \text{ exists}} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) > -\infty$ . This is trivial — simply consider the case where the training set  $\mathcal{S}$  is such that  $\mathbf{x}_i = \mathbf{0}$  for all  $i=1, 2, \dots, |\mathcal{S}|$ . The training loss in this case is constant (see Equations (28) and (9)), implying  $\inf_{\boldsymbol{\theta} \in \mathbb{R}^d \text{ s.t. } \nabla^2 f(\boldsymbol{\theta}) \text{ exists}} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) = 0$ .  $\square$

## G Training Loss for Least-Squares Linear Regression on Whitened Data

In this appendix we derive a simplified expression for the training loss corresponding to scalar linear regression on whitened data per least-squares criterion. Concretely, we simplify the function

$\phi: \mathbb{R}^{d_n, d_0} \rightarrow \mathbb{R}$  defined by Equation (11) in the special case where:  $d_n=1$ ; the empirical (uncentered) covariance matrix of the training inputs —  $\Lambda_{xx} := \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \mathbf{x}_i \mathbf{x}_i^\top \in \mathbb{R}^{d_0, d_0}$  — is equal to identity; and the loss function  $\ell: \mathbb{R}^{d_n} \times \mathcal{Y} \rightarrow \mathbb{R}$  is the square loss, *i.e.*  $\mathcal{Y} = \mathbb{R}$  and  $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$ .

Let  $X \in \mathbb{R}^{d_0, |\mathcal{S}|}$  and  $Y \in \mathbb{R}^{1, |\mathcal{S}|}$  be the matrices whose  $i$ 'th columns hold, respectively, the training input  $\mathbf{x}_i$  and its label  $y_i$ ,  $i = 1, 2, \dots, |\mathcal{S}|$ . Denote by  $\Lambda_{yx}$  the empirical (uncentered) cross-covariance matrix between training labels and inputs, *i.e.*  $\Lambda_{yx} := \frac{1}{|\mathcal{S}|} Y X^\top \in \mathbb{R}^{1, d_0}$ . In the special case under consideration, for any  $W \in \mathbb{R}^{1, d_0}$ :

$$\begin{aligned} \phi(W) &= \frac{1}{2|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} (W \mathbf{x}_i - y_i)^2 \\ &= \frac{1}{2|\mathcal{S}|} \|WX - Y\|_{Frobenius}^2 \\ &= \frac{1}{2|\mathcal{S}|} \text{Tr}((WX - Y)(WX - Y)^\top) \\ &= \frac{1}{2|\mathcal{S}|} \text{Tr}(WXX^\top W^\top) - \frac{1}{|\mathcal{S}|} \text{Tr}(YX^\top W^\top) + \frac{1}{2|\mathcal{S}|} \text{Tr}(YY^\top) \\ &= \frac{1}{2} \text{Tr}(W\Lambda_{xx}W^\top) - \text{Tr}(\Lambda_{yx}W^\top) + \frac{1}{2|\mathcal{S}|} \text{Tr}(YY^\top). \end{aligned}$$

Since  $\Lambda_{xx}$  is equal to identity, we have:

$$\begin{aligned} \phi(W) &= \frac{1}{2} \text{Tr}(WW^\top) - \text{Tr}(\Lambda_{yx}W^\top) + \frac{1}{2|\mathcal{S}|} \text{Tr}(YY^\top) \\ &= \frac{1}{2} \text{Tr}((W - \Lambda_{yx})(W - \Lambda_{yx})^\top) - \frac{1}{2} \text{Tr}(\Lambda_{yx}\Lambda_{yx}^\top) + \frac{1}{2|\mathcal{S}|} \text{Tr}(YY^\top) \\ &= \frac{1}{2} \|W - \Lambda_{yx}\|_{Frobenius}^2 - \frac{1}{2} \text{Tr}(\Lambda_{yx}\Lambda_{yx}^\top) + \frac{1}{2|\mathcal{S}|} \text{Tr}(YY^\top). \end{aligned}$$

$c := -\frac{1}{2} \text{Tr}(\Lambda_{yx}\Lambda_{yx}^\top) + \frac{1}{2|\mathcal{S}|} \text{Tr}(YY^\top)$  does not depend on  $W$ , so we arrive at the simplified form:

$$\phi(W) = \frac{1}{2} \|W - \Lambda_{yx}\|_{Frobenius}^2 + c.$$

## H Convergence with Unbalanced Initialization

In Section 5 we translated an analysis of gradient flow over deep linear neural networks — Proposition 3 — into a convergence guarantee for gradient descent — Theorem 4. In order to leverage known results concerning gradient flow over deep linear neural networks, Proposition 3 assumed that initialization is balanced (*i.e.* meets Equation (15)), which in turn led Theorem 4 to assume the same. We noted (Remark 1), however, that the generic tool used for the translation — Theorem 3 — allows for gradient flow and gradient descent to be initialized differently, thus it is possible to extend Theorem 4 so that it accounts for unbalanced initialization (*i.e.* for initialization which satisfies Equation (15) only approximately). The current appendix presents such an extension.

Consider the setting of Section 5 — depth  $n$  fully connected neural network as defined in Equation (8) (and surrounding text), with linear activation ( $\sigma(z) = z$ ) and output dimension  $d_n = 1$ , learned via minimization of square loss over whitened and normalized data, *i.e.* via minimization of the training loss  $f(\cdot)$  presented in Equation (16) (and surrounding text). For simplicity, we assume that the network's hidden widths are all equal to its input dimension, *i.e.*  $d_0 = d_1 = \dots = d_{n-1}$ .<sup>32</sup> Deviation from balancedness (Equation (15)) will be quantified per the following definition.

**Definition 1.** The *unbalancedness magnitude* of a weight setting  $\theta \in \mathbb{R}^d$  is defined to be:

$$\max_{j \in \{1, 2, \dots, n-1\}} \|W_{j+1}^\top W_{j+1} - W_j W_j^\top\|_{nuclear}, \quad (33)$$

where  $W_1, W_2, \dots, W_n$  denote the weight matrices constituting  $\theta$ .

By Lemma 8 below, small unbalancedness magnitude implies proximity to perfect balancedness.

**Lemma 8.** *For any weight setting  $\theta \in \mathbb{R}^d$  with unbalancedness magnitude (Definition 1) equal to  $\hat{\epsilon} \geq 0$ , there exists a weight setting  $\hat{\theta} \in \mathbb{R}^d$  which is balanced (has unbalancedness magnitude zero) and meets  $\|\theta - \hat{\theta}\|_2 \leq n^{1.5} \sqrt{\hat{\epsilon}}$ .*

<sup>32</sup>Lemma 8 is the only part of the analysis henceforth which relies on this assumption — generalizing the lemma to account for arbitrary hidden widths will accordingly generalize the entire analysis.

*Proof sketch (for complete proof see Subappendix J.20).* By Lemma 1 in [46], an analogous result holds in the case where all weight matrices are square (i.e.  $d_0 = d_1 = \dots = d_n$ ). The proof is based on a reduction to this case, attained by replacing  $W_n$  with  $\sqrt{W_n^\top W_n}$ .  $\square$

Including Lemma 8 in the translation of Proposition 3 via Theorem 3 yields Theorem 5 below — an extension of Theorem 4 that allows for unbalanced initialization.

**Theorem 5.** *Consider minimization of the training loss  $f(\cdot)$  in Equation (16) via gradient descent (Equation (6)). Denote by  $\theta_0, \theta_1, \theta_2, \dots$  the iterates of gradient descent, and by  $W_{n:1,0}$  the end-to-end matrix (Equation (10)) of the initial point  $\theta_0$ . Assume that  $\|W_{n:1,0}\|_{\text{Frobenius}} \in (0, 0.1]$  (initialization is small but non-zero), and that  $W_{n:1,0}$  is not antiparallel to  $\Lambda_{yx}$ , meaning:*

$$\nu := \text{Tr}(\Lambda_{yx}^\top W_{n:1,0}) / (\|\Lambda_{yx}\|_{\text{Frobenius}} \|W_{n:1,0}\|_{\text{Frobenius}}) \neq -1.$$

Let  $\tilde{\epsilon} > 0$ . Then, if the unbalancedness magnitude (Definition 1) of  $\theta_0$  is no greater than:

$$\hat{\epsilon} := \frac{\|W_{n:1,0}\|_{\text{Frobenius}}^8 \min\{1, \tilde{\epsilon}^2\}}{n^{15} e^{12n+6} \left(\max\left\{3, \frac{3-\nu}{1+\nu}\right\}\right)^{9n-5}} \left( \ln \left( \frac{23n \max\left\{3, \frac{3-\nu}{1+\nu}\right\}}{\|W_{n:1,0}\|_{\text{Frobenius}} \min\{1, \tilde{\epsilon}\}} \right) \right)^{-2} \in \tilde{\Omega} \left( \frac{\|W_{n:1,0}\|_{\text{Frobenius}}^8 \tilde{\epsilon}^2}{n^{15} \left(\text{poly}\left(\frac{3-\nu}{1+\nu}\right)\right)^n} \right), \quad (34)$$

and if the step size  $\eta$  meets:

$$\eta \leq \frac{\|W_{n:1,0}\|_{\text{Frobenius}}^5 \min\{1, \tilde{\epsilon}\}}{n^{17/2} e^{7n+10} \left(\max\left\{3, \frac{3-\nu}{1+\nu}\right\}\right)^{(11n-5)/2}} \left( \ln \left( \frac{23n \max\left\{3, \frac{3-\nu}{1+\nu}\right\}}{\|W_{n:1,0}\|_{\text{Frobenius}} \min\{1, \tilde{\epsilon}\}} \right) \right)^{-2} \in \tilde{\Omega} \left( \frac{\|W_{n:1,0}\|_{\text{Frobenius}}^5 \tilde{\epsilon}}{n^{17/2} \left(\text{poly}\left(\frac{3-\nu}{1+\nu}\right)\right)^n} \right), \quad (35)$$

it holds that  $f(\theta_k) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \leq \tilde{\epsilon}$  for some  $k \in \mathbb{N}$  satisfying.<sup>33</sup>

$$k \leq \frac{3n \left(\frac{3}{2} \max\left\{3, \frac{3-\nu}{1+\nu}\right\}\right)^n}{\|W_{n:1,0}\|_{\text{Frobenius}} \eta} \ln \left( \frac{23n \max\left\{3, \frac{3-\nu}{1+\nu}\right\}}{\|W_{n:1,0}\|_{\text{Frobenius}} \min\{1, \tilde{\epsilon}\}} \right) + 1 \in \tilde{\mathcal{O}} \left( \frac{n \left(\text{poly}\left(\frac{3-\nu}{1+\nu}\right)\right)^n \ln\left(\frac{1}{\tilde{\epsilon}}\right)}{\|W_{n:1,0}\|_{\text{Frobenius}} \eta} \right). \quad (36)$$

*Proof sketch (for complete proof see Subappendix J.21).* The proof begins by invoking Lemma 8 for obtaining a weight setting  $\hat{\theta}_0$  which is balanced and meets  $\|\theta_0 - \hat{\theta}_0\|_2 \leq n^{1.5} \sqrt{\tilde{\epsilon}}$ . It is then shown that as an initial point for gradient flow,  $\hat{\theta}_0$  satisfies the conditions of Proposition 3 (namely, in addition to being balanced, its end-to-end matrix has Frobenius norm in  $(0, 0.2]$  and is not antiparallel to  $\Lambda_{yx}$ ). From this point on, the proof is similar to that of Theorem 4 — it confirms that  $f(\theta_k) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \leq \tilde{\epsilon}$  by invoking Theorem 3 to establish that gradient descent approximates gradient flow sufficiently well until gradient flow is sufficiently close to global minimum. Throughout this process, the only deviation from the proof of Theorem 4 is that gradient descent and gradient flow are initialized differently — the former starts at  $\theta_0$ , whereas the latter sets off from the nearby point  $\hat{\theta}_0$ . Such discrepancy between initializations is permitted by Theorem 3.  $\square$

## I Further Experiments and Implementation Details

### I.1 Further Experiments

Figure 3 supplements Figure 1 from Section 6 by reporting results obtained on convolutional neural networks.

### I.2 Implementation Details

Below are implementation details omitted from our experimental reports (Section 6 and Subappendix I.1). Source code for reproducing the results, based on the PyTorch framework ([44]), can be found in [https://github.com/elkabzo/cont\\_disc\\_opt\\_dnn](https://github.com/elkabzo/cont_disc_opt_dnn).

<sup>33</sup>In addition to an upper bound (Equation (36)), the theorem’s proof (Subappendix J.21) also establishes an exact expression for  $k$  (Equation (92)). This expression includes terms that depend on  $\hat{\theta}_0$  — balanced weight setting near  $\theta_0$  whose existence is guaranteed by Lemma 8. Means for computing  $\hat{\theta}_0$  based on  $\theta_0$  are not provided by the lemma’s statement, but are brought forth by its proof (Subappendix J.20) — a constructive reduction to Lemma 1 in [46], which itself is proven constructively.

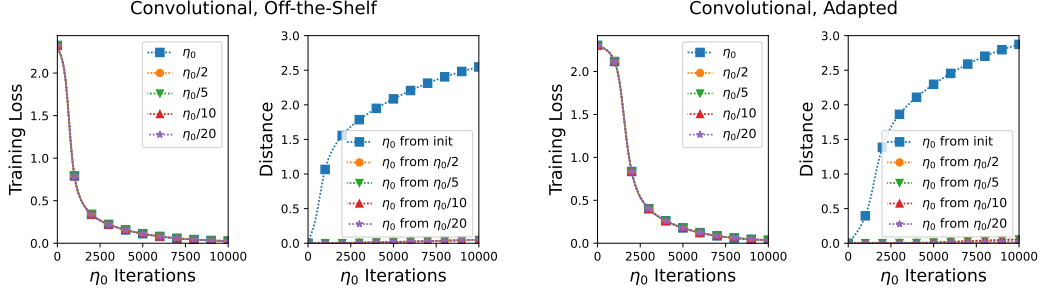


Figure 3: Over deep convolutional neural networks, trajectories of gradient descent with conventional step size barely change when step size is reduced, suggesting they are close to the continuous limit, *i.e.* to trajectories of gradient flow. This figure is identical to Figure 1, except that the results it reports were obtained on convolutional (rather than fully connected) neural networks. Specifically, left pair of plots reports results obtained on a network taken from the online tutorial “Deep Learning with PyTorch: A 60 Minute Blitz” (it comprises two convolutional layers followed by three linear layers, with rectified linear activation in each hidden layer, and max pooling in each convolutional layer),<sup>34</sup> while right pair corresponds to the same network slightly adapted (namely, with no biases in convolutional and linear layers, and with max pooling replaced by regular subsampling, *i.e.* by summarizing each pooling window through its top-left entry) so that it is captured by our theory (*cf.* Subsection 4.2). For further details see caption of Figure 1, as well as Subappendix I.2.

As customary, MNIST images were normalized before being used — we computed mean and standard deviation across all pixels in the dataset, and used those to shift and scale each pixel so as to ensure zero mean and unit standard deviation. To reduce run-time, rather than applying gradient descent to the full MNIST training set (60,000 labeled images), a subset of 1,000 labeled images (chosen once, uniformly at random) was used (altering the size of this subset did not yield a noticeable change in terms of final results). The Xavier distribution employed for initializing neural network weights was of type “uniform” (implemented by calling PyTorch `torch.nn.init.xavier_uniform()` method with default parameters). Experiments ran on an internal Intel Xeon server with eight NVIDIA GeForce RTX 2080 Ti graphical processing units.

## J Deferred Proofs

### J.1 Notations

We introduce notations to be used throughout the appendix. Beginning with matrix norms, we use  $\|\cdot\|_F$  for Frobenius norm,  $\|\cdot\|_n$  for nuclear norm and  $\|\cdot\|_s$  for spectral norm. We extend the notation established in Lemma 1 by regarding Hessians not only as matrices and quadratic forms, but also as bilinear forms. Namely, for any  $\theta \in \mathbb{R}^d$ , we regard  $\nabla^2 f(\theta)$  not only as a (symmetric) matrix in  $\mathbb{R}^{d,d}$  and a quadratic form  $\nabla^2 f(\theta)[\cdot, \cdot] : \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}} \rightarrow \mathbb{R}$ , but also as a bilinear form  $\nabla^2 f(\theta)[\cdot, \cdot]$  that intakes two tuples  $(\Delta W_1, \Delta W_2, \dots, \Delta W_n), (\Delta W'_1, \Delta W'_2, \dots, \Delta W'_n) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$  as its first and second arguments (respectively), arranges them as (respective) vectors  $\Delta\theta, \Delta\theta' \in \mathbb{R}^d$  (in correspondence with how weight matrices  $W_1, W_2, \dots, W_n$  are arranged to create  $\theta$ ), and returns  $\Delta\theta^\top \nabla^2 f(\theta) \Delta\theta' \in \mathbb{R}$ . Additionally, for any  $W \in \mathbb{R}^{d_n, d_0}$ , we extend the view of  $\nabla^2 \phi(W)$  as a quadratic form, and also see it as a bilinear form  $\nabla^2 \phi(W)[\cdot, \cdot]$  that intakes two matrices in  $\mathbb{R}^{d_n, d_0}$  and returns a scalar. We similarly extend the notation of Lemma 4, regarding the matrix  $\nabla^2 \ell_i \in \mathbb{R}^{d_n, d_n}$ , for any  $i \in \{1, 2, \dots, |\mathcal{S}|\}$ , as a bilinear form (in addition to its view as a quadratic form)  $\nabla^2 \ell_i[\cdot, \cdot] : \mathbb{R}^{d_n} \times \mathbb{R}^{d_n} \rightarrow \mathbb{R}$  defined by  $\nabla^2 \ell_i[\mathbf{v}, \mathbf{u}] = \mathbf{v}^\top \nabla^2 \ell_i \mathbf{u}$ . Finally, for any  $j \in \mathbb{N}$  we denote  $[j] := \{1, 2, \dots, j\}$ .

### J.2 Proof of Theorem 3

Let  $\bar{\theta}(\cdot)$  be the continuous polygonal curve corresponding to the iterates of gradient descent:

$$\bar{\theta} : [0, \infty) \rightarrow \mathbb{R}^d, \quad \bar{\theta}(0) = \theta_0, \quad \frac{d}{dt} \bar{\theta}(t) = -\nabla f(\theta_k) \text{ for } t \in (k\eta, (k+1)\eta), \quad k=0, 1, 2, \dots$$

<sup>34</sup>For exact specification of network see [https://pytorch.org/tutorials/beginner/blitz/neural\\_networks\\_tutorial.html#sphx-glr-beginner-blitz-neural-networks-tutorial-py](https://pytorch.org/tutorials/beginner/blitz/neural_networks_tutorial.html#sphx-glr-beginner-blitz-neural-networks-tutorial-py). Note that zero padding (two pixels wide, on each side) was applied to MNIST images for compliance with specified input size (32-by-32).

If  $\|\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t)\|_2 \leq \epsilon$  for all  $t \in [0, \tilde{t}]$  then we are done. Assume by contradiction that this is not the case, and define  $t_\epsilon := \inf\{t \in [0, \tilde{t}] : \|\bar{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t)\|_2 > \epsilon\}$ . It necessarily holds that  $\|\bar{\boldsymbol{\theta}}(0) - \boldsymbol{\theta}(0)\|_2 < \epsilon$  (otherwise the expression on the right-hand side of Equation (7) becomes negative as  $t \searrow 0$ , in contradiction to it being greater than  $\eta > 0$  for all  $t \in (0, \tilde{t}]$ ). By continuity, this implies  $t_\epsilon > 0$  and  $\|\bar{\boldsymbol{\theta}}(t_\epsilon) - \boldsymbol{\theta}(t_\epsilon)\|_2 = \epsilon$ . The trajectory of  $\bar{\boldsymbol{\theta}}(\cdot)$  between times 0 and  $t_\epsilon$ , *i.e.*  $\bar{\boldsymbol{\theta}}([0, t_\epsilon]) := \{\bar{\boldsymbol{\theta}}(t) : t \in [0, t_\epsilon]\}$ , is contained in  $\mathcal{D}_{\tilde{t}, \epsilon}$ . For any  $t \in [0, t_\epsilon]$ , the line segment (in  $\mathbb{R}^d$ ) between  $\bar{\boldsymbol{\theta}}(\lfloor t/\eta \rfloor \eta)$  and  $\bar{\boldsymbol{\theta}}(t)$  is a subset of  $\bar{\boldsymbol{\theta}}([0, t_\epsilon])$ , thus is contained in  $\mathcal{D}_{\tilde{t}, \epsilon}$  as well. We therefore have, for any  $t \in [0, t_\epsilon]$ :

$$\begin{aligned} \left\| \frac{d}{dt} \bar{\boldsymbol{\theta}}(t^+) - (-\nabla f(\bar{\boldsymbol{\theta}}(t))) \right\|_2 &= \left\| -\nabla f(\bar{\boldsymbol{\theta}}(\lfloor t/\eta \rfloor \eta)) - (-\nabla f(\bar{\boldsymbol{\theta}}(t))) \right\|_2 \\ &\leq \beta_{\tilde{t}, \epsilon} \|\bar{\boldsymbol{\theta}}(t) - \bar{\boldsymbol{\theta}}(\lfloor t/\eta \rfloor \eta)\|_2 \\ &= \beta_{\tilde{t}, \epsilon} \|\nabla f(\bar{\boldsymbol{\theta}}(\lfloor t/\eta \rfloor \eta))\|_2 (t - \lfloor t/\eta \rfloor \eta) \\ &\leq \beta_{\tilde{t}, \epsilon} \gamma_{\tilde{t}, \epsilon} \eta, \end{aligned}$$

where  $\frac{d}{dt} \bar{\boldsymbol{\theta}}(t^+)$  represents the right derivative of  $\bar{\boldsymbol{\theta}}(\cdot)$  at time  $t$ . The Fundamental Theorem (Theorem 2) may thus be applied with  $\delta(t) = \beta_{\tilde{t}, \epsilon} \gamma_{\tilde{t}, \epsilon} \eta$  for all  $t \in [0, t_\epsilon]$ , yielding:

$$\|\boldsymbol{\theta}(t_\epsilon) - \bar{\boldsymbol{\theta}}(t_\epsilon)\|_2 \leq e^{\int_0^{t_\epsilon} m(t') dt'} \|\boldsymbol{\theta}(0) - \bar{\boldsymbol{\theta}}(0)\|_2 + \beta_{\tilde{t}, \epsilon} \gamma_{\tilde{t}, \epsilon} \eta \int_0^{t_\epsilon} e^{\int_{t'}^{t_\epsilon} m(t'') dt''} dt'.$$

By our assumption on the step size (Equation (7)):

$$\eta < \frac{\epsilon - e^{\int_0^{t_\epsilon} m(t') dt'} \|\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}(0)\|_2}{\beta_{\tilde{t}, \epsilon} \gamma_{\tilde{t}, \epsilon} \int_0^{t_\epsilon} e^{\int_{t'}^{t_\epsilon} m(t'') dt''} dt'}.$$

Combining the latter two inequalities, we obtain  $\|\boldsymbol{\theta}(t_\epsilon) - \bar{\boldsymbol{\theta}}(t_\epsilon)\|_2 < \epsilon$ . Since it was previously noted that  $\|\bar{\boldsymbol{\theta}}(t_\epsilon) - \boldsymbol{\theta}(t_\epsilon)\|_2 = \epsilon$ , our proof by contradiction is complete.  $\square$

### J.3 Proof of Corollary 1

Non-negativity and  $\beta$ -smoothness of  $f(\cdot)$  imply  $\|\nabla f(\mathbf{q})\|_2 \leq \sqrt{2\beta f(\mathbf{q})}$  for all  $\mathbf{q} \in \mathbb{R}^d$ . Using this inequality, along with the fact that  $f(\cdot)$  is non-increasing during gradient flow, we have:

$$\sup_{t \in [0, t_\epsilon]} \|\nabla f(\boldsymbol{\theta}(t))\|_2 \leq \sup_{t \in [0, t_\epsilon]} \sqrt{2\beta f(\boldsymbol{\theta}(t))} \leq \sqrt{2\beta f(\boldsymbol{\theta}(0))}.$$

If  $\mathbf{q} \in \mathbb{R}^d$  lies no more than  $\epsilon$ -away from  $\boldsymbol{\theta}(\cdot)$ , *i.e.*  $\exists t \in [0, t_\epsilon] : \|\mathbf{q} - \boldsymbol{\theta}(t)\|_2 \leq \epsilon$ , then  $\beta$ -smoothness implies  $\|\nabla f(\mathbf{q})\|_2 \leq \|\nabla f(\boldsymbol{\theta}(t))\|_2 + \beta\epsilon$ , which in turn means  $\|\nabla f(\mathbf{q})\|_2 \leq \sqrt{2\beta f(\boldsymbol{\theta}(0))} + \beta\epsilon$ . We may therefore call Theorem 3 with  $\gamma_{\tilde{t}, \epsilon} = \sqrt{2\beta f(\boldsymbol{\theta}(0))} + \beta\epsilon$ , alongside  $\beta_{\tilde{t}, \epsilon} = \beta$  and  $m(\cdot) \equiv m$ . Simplifying the resulting bound on the step size (Equation (7)) then completes the proof.  $\square$

### J.4 Proof of Lemma 1

#### J.4.1 Sketch

With  $\Delta\boldsymbol{\theta}$  an arbitrary vector in  $\mathbb{R}^d$ , and  $(\Delta W_1, \Delta W_2, \dots, \Delta W_n)$  its corresponding matrix tuple, we expand:

$$f(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) = \phi((W_n + \Delta W_n)(W_{n-1} + \Delta W_{n-1}) \cdots (W_1 + \Delta W_1)),$$

and extract  $\nabla^2 f(\boldsymbol{\theta})$  from the second order terms.

#### J.4.2 Complete Proof

Recall that  $\boldsymbol{\theta} \in \mathbb{R}^d$  is an arrangement of  $(W_1, W_2, \dots, W_n) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$  as a vector. Let  $(\Delta W_1, \Delta W_2, \dots, \Delta W_n) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$ , and denote by  $\Delta\boldsymbol{\theta} \in \mathbb{R}^d$  its arrangement as a vector in corresponding order. Denote:

$$\begin{aligned} \Delta^{(1)} &:= \sum_{j=1}^n W_{n:j+1} (\Delta W_j) W_{j-1:1}, \\ \Delta^{(2)} &:= \sum_{1 \leq j < j' \leq n} W_{n:j'+1} (\Delta W_{j'}) W_{j'-1:j+1} (\Delta W_j) W_{j-1:1}, \\ \Delta^{(3:n)} &:= (W_n + \Delta W_n) \cdots (W_1 + \Delta W_1) - W_{n:1} - \Delta^{(1)} - \Delta^{(2)}. \end{aligned} \tag{37}$$

We now develop a second-order Taylor expansion of  $f(\boldsymbol{\theta})$ . Since the matrix tuple corresponding to  $(\boldsymbol{\theta} + \Delta\boldsymbol{\theta})$  is  $((W_1 + \Delta W_1), \dots, (W_n + \Delta W_n))$ , and  $f(\boldsymbol{\theta}) = \phi(W_{n:1})$  (see beginning of Subsubsection 4.1.1) on an open region containing  $\boldsymbol{\theta}$ , for sufficiently small  $\Delta\boldsymbol{\theta}$  we obtain:

$$f(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) = \phi\left((W_n + \Delta W_n) \dots (W_1 + \Delta W_1)\right) = \phi\left(W_{n:1} + \Delta^{(1)} + \Delta^{(2)} + \Delta^{(3:n)}\right). \quad (38)$$

Let  $\Delta W \in \mathbb{R}^{d_n, d_0}$ , the second-order Taylor expansion of the twice continuously differentiable  $\phi(\cdot)$  at the point  $W_{n:1}$  is given by:

$$\phi(W_{n:1} + \Delta W) = \phi(W_{n:1}) + \langle \nabla \phi(W_{n:1}), \Delta W \rangle + \frac{1}{2} \nabla^2 \phi(W_{n:1}) [\Delta W] + \mathcal{O}(\|\Delta W\|_F^2), \quad (39)$$

where the  $\mathcal{O}(\cdot)$  notation refers to some expression satisfying  $\lim_{a \rightarrow 0} (\mathcal{O}(a)/a) = 0$ . We continue to develop Equation (38) using Equation (39):

$$\begin{aligned} f(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) &= \phi\left(W_{n:1} + (\Delta^{(1)} + \Delta^{(2)} + \Delta^{(3:n)})\right) \\ &= \phi(W_{n:1}) + \langle \nabla \phi(W_{n:1}), \Delta^{(1)} + \Delta^{(2)} + \Delta^{(3:n)} \rangle + \\ &\quad \frac{1}{2} \nabla^2 \phi(W_{n:1}) \left[ \Delta^{(1)} + \Delta^{(2)} + \Delta^{(3:n)} \right] + \mathcal{O}(\|\Delta^{(1)} + \Delta^{(2)} + \Delta^{(3:n)}\|_F^2) \\ &= \phi(W_{n:1}) + \langle \nabla \phi(W_{n:1}), \Delta^{(1)} \rangle + \langle \nabla \phi(W_{n:1}), \Delta^{(2)} \rangle + \langle \nabla \phi(W_{n:1}), \Delta^{(3:n)} \rangle + \\ &\quad \frac{1}{2} \nabla^2 \phi(W_{n:1}) \left[ \Delta^{(1)} \right] + \frac{1}{2} \nabla^2 \phi(W_{n:1}) \left[ \Delta^{(2)} + \Delta^{(3:n)} \right] + \\ &\quad 2 \cdot \frac{1}{2} \nabla^2 \phi(W_{n:1}) \left[ \Delta^{(1)}, \Delta^{(2)} + \Delta^{(3:n)} \right] + \mathcal{O}(\|\Delta^{(1)} + \Delta^{(2)} + \Delta^{(3:n)}\|_F^2), \end{aligned}$$

where in the last transition we view  $\nabla^2 \phi$  as both a quadratic and a bilinear form (see Subappendix J.1). Notice that the following terms  $\langle \nabla \phi(W_{n:1}), \Delta^{(3:n)} \rangle$ ,  $\nabla^2 \phi(W_{n:1}) [\Delta^{(2)} + \Delta^{(3:n)}]$ ,  $\nabla^2 \phi(W_{n:1}) [\Delta^{(1)}, \Delta^{(2)} + \Delta^{(3:n)}]$  and  $\mathcal{O}(\|\Delta^{(1)} + \Delta^{(2)} + \Delta^{(3:n)}\|_F^2)$  are all  $\mathcal{O}(\|\Delta\boldsymbol{\theta}\|_F^2)$ , thus:

$$\begin{aligned} f(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) &= \phi(W_{n:1}) + \langle \nabla \phi(W_{n:1}), \Delta^{(1)} \rangle + \langle \nabla \phi(W_{n:1}), \Delta^{(2)} \rangle + \frac{1}{2} \nabla^2 \phi(W_{n:1}) \left[ \Delta^{(1)} \right] + \mathcal{O}(\|\Delta\boldsymbol{\theta}\|_F^2). \end{aligned}$$

This is a Taylor expansion of  $f(\cdot)$  at  $\boldsymbol{\theta}$  with a constant term  $\phi(W_{n:1})$ , a linear term  $\langle \nabla \phi(W_{n:1}), \Delta^{(1)} \rangle$ , a quadratic term  $\langle \nabla \phi(W_{n:1}), \Delta^{(2)} \rangle + \frac{1}{2} \nabla^2 \phi(W_{n:1}) [\Delta^{(1)}]$ , and a remainder term of  $\mathcal{O}(\|\Delta\boldsymbol{\theta}\|_F^2)$ . From uniqueness of the Taylor expansion it follows that the quadratic term is equal to  $\frac{1}{2} \nabla^2 f(\boldsymbol{\theta}) [\Delta W_1, \dots, \Delta W_n]$ . This implies:

$$\begin{aligned} \nabla^2 f(\boldsymbol{\theta}) [\Delta W_1, \dots, \Delta W_n] &= \nabla^2 \phi(W_{n:1}) \left[ \Delta^{(1)} \right] + 2 \langle \nabla \phi(W_{n:1}), \Delta^{(2)} \rangle \\ &= \nabla^2 \phi(W_{n:1}) \left[ \sum_{j=1}^n W_{n:j+1} (\Delta W_j) W_{j-1:1} \right] + \\ &\quad 2 \text{Tr} \left( \nabla \phi(W_{n:1})^\top \sum_{1 \leq j < j' \leq n} W_{n:j'+1} (\Delta W_{j'}) W_{j'-1:j+1} (\Delta W_j) W_{j-1:1} \right), \end{aligned}$$

where the last transition follows from plugging in the definitions of  $\Delta^{(1)}$  and  $\Delta^{(2)}$  (see Equation (37)).  $\square$

## J.5 Proof of Proposition 1

### J.5.1 Sketch

The proof is constructive — with  $c > 0$  arbitrary, we define a point  $\boldsymbol{\theta} \in \mathbb{R}^d$ , and a non-zero translation vector  $\Delta\boldsymbol{\theta} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ , such that  $\Delta\boldsymbol{\theta}^\top \nabla^2 f(\boldsymbol{\theta}) \Delta\boldsymbol{\theta} = -c \|\Delta\boldsymbol{\theta}\|_2^2$ .

### J.5.2 Complete Proof

Since  $\nabla \phi(0) \neq 0$ , there exists  $(\Delta W'_1, \Delta W'_2) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1}$  and  $(W'_3, \dots, W'_n) \in \mathbb{R}^{d_3, d_2} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$  such that  $\langle \nabla \phi(0), W'_n \dots W'_3 \Delta W'_2 \Delta W'_1 \rangle > 0$ . Notice that none of the following



matrices  $\Delta W'_1, \Delta W'_2, W'_3, \dots, W'_n$  are equal to zero. Define (while recalling the assumption of  $n \geq 3$ ):

$$\begin{aligned}\Delta W_1 &:= \Delta W'_1 \in \mathbb{R}^{d_1, d_0}, \\ \Delta W_2 &:= \Delta W'_2 \in \mathbb{R}^{d_2, d_1}, \\ \Delta W_3 &:= 0 \in \mathbb{R}^{d_3, d_2}, \\ \Delta W_j &:= 0 \in \mathbb{R}^{d_j, d_{j-1}} \text{ for } j \in \{1, 2, \dots, n\} \setminus \{1, 2, 3\}.\end{aligned}$$

For some arbitrary  $c > 0$ , we define:

$$\begin{aligned}W_1 &:= 0 \in \mathbb{R}^{d_1, d_0}, \\ W_2 &:= 0 \in \mathbb{R}^{d_2, d_1}, \\ W_3 &:= W'_3 \frac{-c \cdot \sum_{1 \leq j \leq n} \|\Delta W_j\|_F^2}{2 \langle \nabla \phi(0), W'_n \dots W'_3 \Delta W'_2 \Delta W'_1 \rangle} \in \mathbb{R}^{d_3, d_2}, \\ W_j &:= W'_j \in \mathbb{R}^{d_j, d_{j-1}} \text{ for } j \in \{1, 2, \dots, n\} \setminus \{1, 2, 3\}.\end{aligned}$$

Recall that we denote by  $\theta \in \mathbb{R}^d$  the arrangement of  $(W_1, W_2, \dots, W_n)$  as a vector. As shown in Lemma 1:

$$\begin{aligned}\nabla^2 f(\theta)[\Delta W_1, \dots, \Delta W_n] &= \nabla^2 \phi(W_{n:1}) \left[ \sum_{j=1}^n W_{n:j+1} (\Delta W_j) W_{j-1:1} \right] \\ &\quad + 2\text{Tr} \left( \nabla \phi(W_{n:1})^\top \sum_{1 \leq j < j' \leq n} W_{n:j'+1} (\Delta W_{j'}) W_{j'-1:j+1} (\Delta W_j) W_{j-1:1} \right).\end{aligned}\tag{40}$$

Notice the first summand in the right-hand side of Equation (40) is equal to zero:

$$\nabla^2 \phi(W_{n:1}) \left[ \sum_{j=1}^n W_{n:j+1} (\Delta W_j) W_{j-1:1} \right] = \nabla^2 \phi(W_{n:1}) \left[ \sum_{j=1}^n 0 \right] = 0.\tag{41}$$

We develop the expression of the second summand in the right-hand side of Equation (40):

$$\begin{aligned}2\text{Tr} \left( \nabla \phi(W_{n:1})^\top \sum_{1 \leq j < j' \leq n} W_{n:j'+1} (\Delta W_{j'}) W_{j'-1:j+1} (\Delta W_j) W_{j-1:1} \right) \\ = 2 \langle \nabla \phi(W_{n:1}), \sum_{1 \leq j < j' \leq n} W_{n:j'+1} (\Delta W_{j'}) W_{j'-1:j+1} (\Delta W_j) W_{j-1:1} \rangle \\ = 2 \langle \nabla \phi(0), W_n \dots W_3 \Delta W_2 \Delta W_1 \rangle \\ = -c \cdot \sum_{1 \leq j \leq n} \|\Delta W_j\|_F^2,\end{aligned}\tag{42}$$

where the last transition follows by plugging in the definitions of  $\Delta W_1$ ,  $\Delta W_2$  and  $W_j$  for  $j \in [n] \setminus \{1, 2\}$ . Plugging in Equations (41) and (42) in Equation (40), we obtain:

$$\nabla^2 f(\theta)[\Delta W_1, \dots, \Delta W_n] = -c \cdot \sum_{1 \leq j \leq n} \|\Delta W_j\|_F^2.\tag{43}$$

Noticing that  $\sum_{1 \leq j \leq n} \|\Delta W_j\|_F^2 \neq 0$ , Equation (43) implies  $\lambda_{\min}(\nabla^2 f(\theta)) \leq -c$ . This bound holds for every  $c > 0$ , thus yielding the desired result (*i.e.*  $\inf_{\theta \in \mathbb{R}^d} \lambda_{\min}(\nabla^2 f(\theta)) = -\infty$ ).  $\square$

## J.6 Proof of Lemma 2

### J.6.1 Sketch

Appealing to Lemma 1, we lower bound the right-hand side of Equation (12). Convexity of  $\phi(\cdot)$  implies that the first summand is non-negative. For the second summand, we use known matrix inequalities to establish a lower bound of  $c \sum_{j=1}^n \|\Delta W_j\|_{Frobenius}^2$ , with  $c$  being the expression on the right-hand side of Equation (13).

### J.6.2 Complete Proof

Recall that  $\theta \in \mathbb{R}^d$  is an arrangement of  $(W_1, W_2, \dots, W_n) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$  as a vector. Let  $(\Delta W_1, \Delta W_2, \dots, \Delta W_n) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$ , and denote by  $\Delta \theta \in \mathbb{R}^d$  its arrangement as a vector in corresponding order. As shown in Lemma 1:

$$\begin{aligned}\nabla^2 f(\theta)[\Delta W_1, \dots, \Delta W_n] &= \nabla^2 \phi(W_{n:1}) \left[ \sum_{j=1}^n W_{n:j+1} (\Delta W_j) W_{j-1:1} \right] \\ &\quad + 2\text{Tr} \left( \nabla \phi(W_{n:1})^\top \sum_{1 \leq j < j' \leq n} W_{n:j'+1} (\Delta W_{j'}) W_{j'-1:j+1} (\Delta W_j) W_{j-1:1} \right).\end{aligned}$$

Convexity of  $\phi(\cdot)$  implies that  $\nabla^2\phi(W_{n:1})$  is positive semi-definite, thus:

$$\nabla^2 f(\boldsymbol{\theta})[\Delta W_1, \dots, \Delta W_n] \geq 2\text{Tr}\left(\nabla\phi(W_{n:1})^\top \sum_{1 \leq j < j' \leq n} W_{n:j'+1}(\Delta W_{j'})W_{j'-1:j+1}(\Delta W_j)W_{j-1:1}\right).$$

Using a simple corollary of Von-Neumann's trace inequality (see [41]):

$$\begin{aligned} & \nabla^2 f(\boldsymbol{\theta})[\Delta W_1, \dots, \Delta W_n] \\ & \geq -2\|\nabla\phi(W_{n:1})\|_n \cdot \left\| \sum_{1 \leq j < j' \leq n} W_{n:j'+1}(\Delta W_{j'})W_{j'-1:j+1}(\Delta W_j)W_{j-1:1} \right\|_s. \end{aligned} \quad (44)$$

Upper bound the nuclear norm:

$$\|\nabla\phi(W_{n:1})\|_n \leq \sqrt{\min\{d_0, d_n\}} \|\nabla\phi(W_{n:1})\|_F. \quad (45)$$

The following bound holds:

$$\begin{aligned} & \left\| \sum_{1 \leq j < j' \leq n} W_{n:j'+1}(\Delta W_{j'})W_{j'-1:j+1}(\Delta W_j)W_{j-1:1} \right\|_s \\ & \leq \sum_{1 \leq j < j' \leq n} \|W_{n:j'+1}(\Delta W_{j'})W_{j'-1:j+1}(\Delta W_j)W_{j-1:1}\|_s \\ & \leq \sum_{1 \leq j < j' \leq n} \|\Delta W_{j'}\|_s \|\Delta W_j\|_s \cdot \prod_{k \in [n] \setminus \{j, j'\}} \|W_k\|_s \\ & \leq \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_s \cdot \sum_{1 \leq j < j' \leq n} \|\Delta W_{j'}\|_s \|\Delta W_j\|_s, \end{aligned} \quad (46)$$

where the first transition follows from triangle inequalities, the second inequality follows from sub-multiplicativity of the spectral norm, and the last inequality follows from maximizing the term  $\prod_{k \in [n] \setminus \{j, j'\}} \|W_k\|_s$  over  $j, j'$ . Plugging Equations (45) and (46) into Equation (44), we have:

$$\begin{aligned} & \nabla^2 f(\boldsymbol{\theta})[\Delta W_1, \dots, \Delta W_n] \\ & \geq -2\sqrt{\min\{d_0, d_n\}} \|\nabla\phi(W_{n:1})\|_F \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_s \cdot \sum_{1 \leq j < j' \leq n} \|\Delta W_{j'}\|_s \|\Delta W_j\|_s. \end{aligned}$$

It holds that:

$$\begin{aligned} & \sum_{1 \leq j < j' \leq n} \|\Delta W_{j'}\|_s \|\Delta W_j\|_s \\ & \leq \sum_{1 \leq j < j' \leq n} \|\Delta W_{j'}\|_F \|\Delta W_j\|_F \\ & = \frac{1}{2} \left( \sum_{j=1}^n \|\Delta W_j\|_F \right)^2 - \frac{1}{2} \sum_{j=1}^n \|\Delta W_j\|_F^2 \\ & \leq \frac{n}{2} \sum_{j=1}^n \|\Delta W_j\|_F^2 - \frac{1}{2} \sum_{j=1}^n \|\Delta W_j\|_F^2 \\ & = \frac{n-1}{2} \sum_{j=1}^n \|\Delta W_j\|_F^2, \end{aligned}$$

where the last inequality follows from the fact that the one-norm of a vector in  $\mathbb{R}^n$  is never greater than  $\sqrt{n}$  times its euclidean-norm. This leads us to:

$$\begin{aligned} & \nabla^2 f(\boldsymbol{\theta})[\Delta W_1, \dots, \Delta W_n] \\ & \geq -(n-1)\sqrt{\min\{d_0, d_n\}} \|\nabla\phi(W_{n:1})\|_F \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_s \cdot \sum_{j=1}^n \|\Delta W_j\|_F^2. \end{aligned}$$

The desired result readily follows:

$$\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -(n-1)\sqrt{\min\{d_0, d_n\}} \|\nabla\phi(W_{n:1})\|_F \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_s.$$

□

## J.7 Proof of Proposition 2

### J.7.1 Sketch

By the analysis of [18], the quantities  $W_{j+1}^\top W_{j+1} - W_j W_j^\top$ ,  $j = 1, 2, \dots, n-1$ , are invariant (constant) along a gradient flow trajectory, and therefore small if initialization is such. This implies that along a trajectory emanating from near-zero initialization, for every  $j = 1, 2, \dots, n-1$ , the singular values of  $W_j$  are similar to those of  $W_{j+1}$ , and the left singular vectors of  $W_j$  match the right ones of  $W_{j+1}$ . Products of adjacent weight matrices thus simplify, and we obtain  $\|W_j\|_{\text{spectral}} \approx \|W_{n:1}\|_{\text{spectral}}^{1/n}$  for  $j = 1, 2, \dots, n$ . Plugging this into Equation (13) yields the desired result (Equation (14)).

### J.7.2 Complete Proof

Denote by  $\boldsymbol{\theta}(t)$  the time dependent gradient flow trajectory starting at  $\boldsymbol{\theta}_s$  (i.e.  $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_s$ ) and by  $W_1(t), \dots, W_n(t)$  the corresponding time dependent curves of weight matrices induced by the flow. From the assumption  $\|\boldsymbol{\theta}_s\|_2 \leq \epsilon$  we can infer  $\|W_j(0)\|_F \leq \epsilon$  for all  $j \in \{1, 2, \dots, n\}$ . For  $j \in \{1, 2, \dots, n-1\}$ :

$$\begin{aligned} & \|W_{j+1}^\top(0)W_{j+1}(0) - W_j(0)W_j^\top(0)\|_s \\ & \leq \|W_{j+1}^\top(0)W_{j+1}(0)\|_s + \|W_j(0)W_j^\top(0)\|_s \\ & = \|W_{j+1}(0)\|_s^2 + \|W_j(0)\|_s^2 \\ & \leq \|W_{j+1}(0)\|_F^2 + \|W_j(0)\|_F^2 \leq 2\epsilon^2 \leq (2\epsilon)^2. \end{aligned}$$

Theorem 2.2 from [18] states that  $\frac{\partial}{\partial t}(W_j(t)W_j^\top(t) - W_{j+1}^\top(t)W_{j+1}(t)) = 0$  for all  $j \in \{1, 2, \dots, n-1\}$  and  $t \geq 0$ , thus:

$$\|W_{j+1}^\top(t)W_{j+1}(t) - W_j(t)W_j^\top(t)\|_s = \|W_{j+1}^\top(0)W_{j+1}(0) - W_j(0)W_j^\top(0)\|_s \leq (2\epsilon)^2.$$

We can rely on this condition in order to apply Lemma 9 below and get that for all  $t \geq 0$ :

$$\max_{j \in \{1, \dots, n\}} \|W_j(t)\|_s^n \leq \|W_{n:1}(t)\|_s + 4n\epsilon \cdot \max(1, \{\|W_j(t)\|_s\}_{j \in [n]})^{2n}.$$

Combining the latter inequality together with the result of Lemma 2 (Equation (13)), we get:

$$\begin{aligned} & \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta}(t))) \\ & \geq -(n-1)\sqrt{\min\{d_0, d_n\}} \|\nabla \phi(W_{n:1}(t))\|_F \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_j(t)\|_s \\ & \geq -(n-1)\sqrt{\min\{d_0, d_n\}} \|\nabla \phi(W_{n:1}(t))\|_F \max_{j \in [n]} \|W_j(t)\|_s^{n-2} \\ & = -(n-1)\sqrt{\min\{d_0, d_n\}} \|\nabla \phi(W_{n:1}(t))\|_F (\max_{j \in [n]} \|W_j(t)\|_s^n)^{\frac{n-2}{n}} \\ & \geq -(n-1)\sqrt{\min\{d_0, d_n\}} \|\nabla \phi(W_{n:1}(t))\|_F (\|W_{n:1}(t)\|_s + 4n\epsilon \max(1, \{\|W_j(t)\|_s\}_{j \in [n]})^{2n})^{\frac{n-2}{n}} \\ & \geq -(n-1)\sqrt{\min\{d_0, d_n\}} \|\nabla \phi(W_{n:1}(t))\|_F \|W_{n:1}(t)\|_s^{\frac{n-2}{2}} \\ & \quad - (n-1)\sqrt{\min\{d_0, d_n\}} \|\nabla \phi(W_{n:1}(t))\|_F (4n\epsilon \max(1, \{\|W_j(t)\|_s\}_{j \in [n]})^{2n})^{\frac{n-2}{n}}, \end{aligned}$$

where the last inequality follows from sub-additivity of any power between zero and one. Rewriting the inequality such that we remove the time notation as to be consistent with the proposition statement, we obtain:

$$\begin{aligned} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) & \geq -(n-1)\sqrt{\min\{d_0, d_n\}} \|\nabla \phi(W_{n:1})\|_F \|W_{n:1}\|_s^{1-2/n} \\ & \quad - (n-1)\sqrt{\min\{d_0, d_n\}} \|\nabla \phi(W_{n:1})\|_F (4n)^{\frac{n-2}{n}} \max(1, \{\|W_j(t)\|_s\}_{j \in [n]})^{2(n-2)} \epsilon^{\frac{n-2}{n}}. \end{aligned}$$

□

**Lemma 9.** Let  $A_i \in \mathbb{R}^{d_i, d_{i-1}}$  for  $i \in [n]$ . Denote  $\Delta_i := A_{i+1}^\top A_{i+1} - A_i A_i^\top$  for  $i \in [n-1]$ . Assume that  $\|\Delta_i\|_s \leq \frac{1}{2n}$  for  $i \in [n-1]$ . It holds that:

$$\max_{i \in [n]} \|A_i\|_s^n \leq \|A_{n:1}\|_s + 2n \sqrt{\max_{i \in [n-1]} \|\Delta_i\|_s} \cdot \max_{A \in \{I, A_1, \dots, A_n\}} \|A\|_s^{2n},$$

where we denote  $A_{j:i}$  as  $A_j \cdots A_{i+1} A_i$  for  $1 \leq i < j \leq n$  and as an identity matrix (with size to be inferred by context) otherwise.

*Proof.* Define  $A_{\max} := \max_{A \in \{I, A_1, \dots, A_n\}} \|A\|_s$  and  $\Delta_{\max} := \max_{i \in [n-1]} \|\Delta_i\|_s$ . Let  $\mathbf{v} \in \mathbb{R}^{d_0}$  such that  $\mathbf{v} \in \arg\max_{\|\mathbf{u}\|=1} \|A_1 \mathbf{u}\|_2$ . Define  $\mathbf{a}_i := \mathbf{v}^\top A_{n-i:1}^\top (A_{n-(i-1)}^\top A_{n-(i-1)})^i A_{n-i:1} \mathbf{v}$  for  $i \in [n]$ . For

$i \in [n-1]$  we have:

$$\begin{aligned}
& a_i - a_{i+1} \\
&= \mathbf{v}^\top A_{n-i:1}^\top (A_{n-(i-1)}^\top A_{n-(i-1)})^i A_{n-i:1} \mathbf{v} - \mathbf{v}^\top A_{n-(i+1):1}^\top (A_{n-i}^\top A_{n-i})^{i+1} A_{n-(i+1):1} \mathbf{v} \\
&= \mathbf{v}^\top A_{n-i:1}^\top (A_{n-(i-1)}^\top A_{n-(i-1)})^i A_{n-i:1} \mathbf{v} - \mathbf{v}^\top A_{n-(i+1):1}^\top A_{n-i}^\top (A_{n-i} A_{n-i}^\top)^i A_{n-i} A_{n-(i+1):1} \mathbf{v} \\
&= \mathbf{v}^\top A_{n-i:1}^\top (A_{n-(i-1)}^\top A_{n-(i-1)})^i A_{n-i:1} \mathbf{v} - \mathbf{v}^\top A_{n-i:1}^\top (A_{n-i} A_{n-i}^\top)^i A_{n-i:1} \mathbf{v} \\
&= \mathbf{v}^\top A_{n-i:1}^\top (A_{n-i} A_{n-i}^\top + \Delta_{n-i})^i A_{n-i:1} \mathbf{v} - \mathbf{v}^\top A_{n-i:1}^\top (A_{n-i} A_{n-i}^\top)^i A_{n-i:1} \mathbf{v} \\
&= \mathbf{v}^\top A_{n-i:1}^\top \left( (A_{n-i} A_{n-i}^\top + \Delta_{n-i})^i - (A_{n-i} A_{n-i}^\top)^i \right) A_{n-i:1} \mathbf{v} \\
&= \mathbf{v}^\top A_{n-i:1}^\top \left( \sum_{(b_1, \dots, b_i) \in \{0,1\}^i} \prod_{b \in \{b_1, \dots, b_i\}} (b A_{n-i} A_{n-i}^\top + (1-b) \Delta_{n-i}) - (A_{n-i} A_{n-i}^\top)^i \right) A_{n-i:1} \mathbf{v} \\
&= \mathbf{v}^\top A_{n-i:1}^\top \left( \sum_{(b_1, \dots, b_i) \in \{0,1\}^i \setminus (1, \dots, 1)} \prod_{b \in \{b_1, \dots, b_i\}} (b A_{n-i} A_{n-i}^\top + (1-b) \Delta_{n-i}) \right) A_{n-i:1} \mathbf{v},
\end{aligned}$$

where the fourth transition follows from the definition of  $\Delta_{n-i}$  and the second to last transition follows from unrolling  $(A_{n-i} A_{n-i}^\top + \Delta_{n-i})^i$ . Taking absolute value on  $a_i - a_{i+1}$  we obtain:

$$\begin{aligned}
& |a_i - a_{i+1}| \\
&= \left| \mathbf{v}^\top A_{n-i:1}^\top \left( \sum_{(b_1, \dots, b_i) \in \{0,1\}^i \setminus (1, \dots, 1)} \prod_{b \in \{b_1, \dots, b_i\}} (b A_{n-i} A_{n-i}^\top + (1-b) \Delta_{n-i}) \right) A_{n-i:1} \mathbf{v} \right| \\
&\leq \sum_{(b_1, \dots, b_i) \in \{0,1\}^i \setminus (1, \dots, 1)} \left| \mathbf{v}^\top A_{n-i:1}^\top \left( \prod_{b \in \{b_1, \dots, b_i\}} (b A_{n-i} A_{n-i}^\top + (1-b) \Delta_{n-i}) \right) A_{n-i:1} \mathbf{v} \right| \\
&\leq \sum_{(b_1, \dots, b_i) \in \{0,1\}^i \setminus (1, \dots, 1)} \|A_{n-i:1} \mathbf{v}\|_2 \left\| \prod_{b \in \{b_1, \dots, b_i\}} (b A_{n-i} A_{n-i}^\top + (1-b) \Delta_{n-i}) \right\|_s \|A_{n-i:1} \mathbf{v}\|_2 \\
&\leq \sum_{(b_1, \dots, b_i) \in \{0,1\}^i \setminus (1, \dots, 1)} \|A_{n-i:1}\|_s \left( \prod_{b \in \{b_1, \dots, b_i\}} \|b A_{n-i} A_{n-i}^\top + (1-b) \Delta_{n-i}\|_s \right) \|A_{n-i:1}\|_s \\
&\leq \sum_{(b_1, \dots, b_i) \in \{0,1\}^i \setminus (1, \dots, 1)} \|A_{n-i:1}\|_s^2 \prod_{b \in \{b_1, \dots, b_i\}} \left( b \|A_{n-i} A_{n-i}^\top\|_s + (1-b) \|\Delta_{n-i}\|_s \right) \\
&\leq \sum_{(b_1, \dots, b_i) \in \{0,1\}^i \setminus (1, \dots, 1)} A_{\max}^{2n} \prod_{b \in \{b_1, \dots, b_i\}} (b A_{\max}^2 + (1-b) \Delta_{\max}) A_{\max}^n \\
&= A_{\max}^{2n} \cdot \left( (A_{\max}^2 + \Delta_{\max})^i - A_{\max}^{2i} \right),
\end{aligned}$$

where the second transition follows from the triangle inequality, the third from Cauchy–Schwarz and the definition of the spectral norm, the fourth from sub-multiplicativity of the spectral norm, the fifth from sub-additivity of the spectral norm and the sixth from the definitions of  $A_{\max}$  and  $\Delta_{\max}$ . We continue by unrolling  $(A_{\max}^2 + \Delta_{\max})^i$ :

$$\begin{aligned}
& |a_i - a_{i+1}| \\
&\leq A_{\max}^{2n} \cdot \left( \sum_{k=0}^i \binom{i}{k} A_{\max}^{2(i-k)} \Delta_{\max}^k - A_{\max}^{2i} \right) \\
&= A_{\max}^{2n} \cdot \left( \sum_{k=1}^i \binom{i}{k} A_{\max}^{2(i-k)} \Delta_{\max}^k \right) \\
&\leq A_{\max}^{2n} \cdot \left( \sum_{k=1}^i n^k A_{\max}^{2n} \Delta_{\max}^k \right) \\
&= A_{\max}^{4n} \cdot \left( \sum_{k=1}^i (n \Delta_{\max})^k \right) \\
&\leq A_{\max}^{4n} \cdot \left( \sum_{k=1}^{\infty} (n \Delta_{\max})^k \right) \\
&= A_{\max}^{4n} \cdot \frac{n \Delta_{\max}}{1 - n \Delta_{\max}} \\
&\leq A_{\max}^{4n} \cdot 2n \Delta_{\max},
\end{aligned}$$

where the two last transitions follow from geometric series formula and the assumption  $\Delta_{\max} \leq \frac{1}{2n}$ . Overall we have that for  $i \in [n-1]$ :

$$|a_i - a_{i+1}| \leq 2n A_{\max}^{4n} \cdot \Delta_{\max}. \quad (47)$$

The following bound holds:

$$\begin{aligned}
\|A_{n:1}\|_s^2 &\geq \|A_{n:1}\mathbf{v}\|_2^2 \\
&= \mathbf{v}^\top A_{n:1}^\top A_{n:1} \mathbf{v} \\
&= \mathbf{v}^\top A_{n-1:1}^\top (A_n^\top A_n)^1 A_{n-1:1} \mathbf{v} \\
&= a_1 \\
&\geq a_2 - |a_2 - a_1| \\
&\geq a_3 - |a_3 - a_2| - |a_2 - a_1| \\
&\vdots \\
&\geq a_n - \sum_{i=1}^{n-1} |a_{i+1} - a_i| \\
&\geq a_n - \sum_{i=1}^{n-1} 2n A_{\max}^{4n} \cdot \Delta_{\max} \\
&\geq a_n - 2n^2 A_{\max}^{4n} \cdot \Delta_{\max} \\
&= \mathbf{v}^\top (A_1^\top A_1)^n \mathbf{v} - 2n^2 A_{\max}^{4n} \cdot \Delta_{\max} \\
&= \|A_1\|_s^{2n} - 2n^2 A_{\max}^{4n} \cdot \Delta_{\max},
\end{aligned}$$

where the second to last inequality follows from Equation (47). Overall we have:

$$\|A_1\|_s^{2n} \leq \|A_{n:1}\|_s^2 + 2n^2 A_{\max}^{4n} \cdot \Delta_{\max}. \quad (48)$$

For all  $i \in [n-1]$ :

$$\begin{aligned}
\|A_i\|_s^2 &= \|A_i A_i^\top\|_s \\
&= \|A_{i+1}^\top A_{i+1} - \Delta_i\|_s \\
&\geq \|A_{i+1}^\top A_{i+1}\|_s - \|\Delta_i\|_s \\
&\geq \|A_{i+1}\|_s^2 - \Delta_{\max}.
\end{aligned}$$

It follows that for  $i \in [n-1]$ :

$$\begin{aligned}
\|A_{i+1}\|_s^{2n} &\leq (\|A_i\|_s^2 + \Delta_{\max})^n \\
&= \sum_{k=0}^n \binom{n}{k} \|A_i\|_s^{2(n-k)} \Delta_{\max}^k \\
&= \|A_i\|_s^{2n} + \sum_{k=1}^n \binom{n}{k} \|A_i\|_s^{2(n-k)} \Delta_{\max}^k \\
&\leq \|A_i\|_s^{2n} + A_{\max}^{2n} \sum_{k=1}^{\infty} (n \Delta_{\max})^k \\
&= \|A_i\|_s^{2n} + A_{\max}^{2n} \cdot \frac{n \Delta_{\max}}{1 - n \Delta_{\max}} \\
&\leq \|A_i\|_s^{2n} + 2n A_{\max}^{2n} \cdot \Delta_{\max},
\end{aligned}$$

where the two last transitions follow from geometric series formula and the assumption  $\Delta_{\max} \leq \frac{1}{2n}$ . Using the above result repeatedly, we get that for  $i \in [n-1]$ :

$$\begin{aligned}
\|A_{i+1}\|_s^{2n} &\leq \|A_i\|_s^{2n} + 2n A_{\max}^{2n} \cdot \Delta_{\max} \\
&\vdots \\
&\leq \|A_1\|_s^{2n} + i \cdot 2n A_{\max}^{2n} \cdot \Delta_{\max} \\
&\leq \|A_1\|_s^{2n} + 2n^2 A_{\max}^{2n} \cdot \Delta_{\max}.
\end{aligned}$$

Overall we have that for  $i \in [n]$ :

$$\|A_i\|_s^{2n} \leq \|A_1\|_s^{2n} + 2n^2 A_{\max}^{2n} \cdot \Delta_{\max}. \quad (49)$$

Combining Equations (49) and (48) we get for  $i \in [n]$ :

$$\|A_i\|_s^{2n} \leq \|A_1\|_s^{2n} + 2n^2 A_{\max}^{2n} \cdot \Delta_{\max} \leq \|A_{n:1}\|_s^2 + 4n^2 A_{\max}^{4n} \cdot \Delta_{\max}.$$

This leads us to:

$$\begin{aligned}
\max_{i \in [n]} \|A_i\|_s^n &\leq \sqrt{\|A_{n:1}\|_s^2 + 4n^2 A_{\max}^{4n} \cdot \Delta_{\max}} \\
&\leq \sqrt{\|A_{n:1}\|_s^2} + \sqrt{4n^2 A_{\max}^{4n} \cdot \Delta_{\max}} \\
&= \|A_{n:1}\|_s + 2n \sqrt{\max_{i \in [n-1]} \|\Delta_i\|_s \cdot \max_{A \in \{I, A_1, \dots, A_n\}} \|A\|_s^{2n}},
\end{aligned}$$

where the second transition follows from sub-additivity of square root and the last transition follows from the definitions of  $A_{\max}$  and  $\Delta_{\max}$ .  $\square$

## J.8 Proof of Proposition 3

Subsubappendix J.8.1 below provides a brief proof sketch. The complete proof is delivered by the subsequent subsubappendices, organized as follows. Subsubappendix J.8.2 establishes preliminaries. Subsubappendix J.8.3 proves that the trajectory of gradient flow is defined over infinite time. Subsubappendix J.8.4 defines a reparameterization of the gradient flow trajectory, to be used as a technical tool. Subsubappendix J.8.5 lower bounds the minimal distance of the reparameterized trajectory from the origin. Subsubappendix J.8.6 confirms that the reparameterized trajectory escapes the origin. Subsubappendix J.8.7 establishes subsequent convergence, during which the reparameterized trajectory approaches global minimum exponentially fast. Subsubappendix J.8.8 shows that at time  $\bar{t}$  (defined in Equation (17)) the (original) gradient flow trajectory reaches  $\bar{\epsilon}$ -optimality. Subsubappendix J.8.9 analyzes the geometry of the optimization landscape around the gradient flow trajectory, namely, it confirms validity of the smoothness and Lipschitz constants  $\beta_{t,\epsilon}$  and  $\gamma_{t,\epsilon}$  (given in statement of Proposition 3) respectively, and bounds the integral of the minimal eigenvalue of the Hessian in accordance with Equation (18). Finally, Subsubappendix J.8.10 concludes.

### J.8.1 Sketch

By result of [4], gradient flow induces on the end-to-end matrix the following dynamics:

$$\frac{d}{dt} W_{n:1}(t) = -\nabla\phi(W_{n:1}(t)) \left( \|W_{n:1}(t)\|_{Frobenius}^{2-2/n} I_{d_0} + (n-1) [W_{n:1}^\top(t) W_{n:1}(t)]^{1-1/n} \right),$$

where  $I_{d_0} \in \mathbb{R}^{d_0, d_0}$  represents identity, and  $[\cdot]^c$ ,  $c \geq 0$ , stands for a power operator defined over positive semi-definite matrices (with  $c = 0$  yielding identity by definition). Carefully analyzing these dynamics, we characterize  $W_{n:1}(\cdot)$  — trajectory of end-to-end matrix — and show that, with  $\bar{t}$  given by Equation (17),  $\frac{1}{2} \|W_{n:1}(\bar{t}) - \Lambda_{yx}\|_{Frobenius}^2 \leq \bar{\epsilon}$  as required. For establishing Equation (18), we use the characterization of  $W_{n:1}(\cdot)$ , along with a lower bound on the minimal eigenvalue of the Hessian provided in Subsubsection 4.1.1. The expressions for  $\beta_{t,\epsilon}$  and  $\gamma_{t,\epsilon}$  are also derived using the characterization of  $W_{n:1}(\cdot)$  and geometric bounds (bounds on Hessian eigenvalues and gradient norm, respectively), but they involve much coarser computations.

### J.8.2 Preliminaries

We assume  $\bar{\epsilon} \leq \frac{1}{2}$  without loss of generality (a proof that is valid for  $\bar{\epsilon} = \frac{1}{2}$  automatically accounts for  $\bar{\epsilon} > \frac{1}{2}$  as well). Throughout the proof we identify matrices in  $\mathbb{R}^{1, d_0}$  with vectors in  $\mathbb{R}^{d_0}$ . For example, we identify the end-to-end matrix  $W_{n:1} \in \mathbb{R}^{1, d_0}$  (Equation (10)) with the vector  $\mathbf{w}_{n:1} \in \mathbb{R}^{d_0}$ , and the empirical (uncentered) cross-covariance matrix between training labels and inputs,  $\Lambda_{yx} \in \mathbb{R}^{1, d_0}$ , with the vector  $\boldsymbol{\lambda}_{yx} \in \mathbb{R}^{d_0}$ . Accordingly, we overload notation by regarding the function  $\phi(\cdot)$  (defined in Equation (11)) not only as a mapping from  $\mathbb{R}^{1, d_0}$  to  $\mathbb{R}$ , but also as one from  $\mathbb{R}^{d_0}$  to  $\mathbb{R}$ . Under the latter view,  $\phi(\cdot)$  is defined by  $\phi(\mathbf{w}) = \frac{1}{2} \|\mathbf{w} - \boldsymbol{\lambda}_{yx}\|_2^2 + \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q})$ . For  $t \geq 0$ , we denote by  $W_1(t) \in \mathbb{R}^{d_1, d_0}$ ,  $W_2(t) \in \mathbb{R}^{d_2, d_1}$ ,  $\dots$ ,  $W_{n-1}(t) \in \mathbb{R}^{d_{n-1}, d_{n-2}}$ ,  $W_n(t) \in \mathbb{R}^{1, d_{n-1}}$  the weight matrices constituting  $\boldsymbol{\theta}(t) \in \mathbb{R}^d$  (gradient flow trajectory at time  $t$ ), and by  $W_{n:1}(t) \in \mathbb{R}^{1, d_0}$  (or  $\mathbf{w}_{n:1}(t) \in \mathbb{R}^{d_0}$ ) the corresponding end-to-end matrix (*i.e.*  $W_{n:1}(t) := W_n(t) W_{n-1}(t) \cdots W_1(t)$ ).

**Definition 2.** Define  $\mathbf{h}: \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_0}$  by:

$$\mathbf{h}(\mathbf{w}) := \left( \|\mathbf{w}\|_2^{2-\frac{2}{n}} I_{d_0} + (n-1) [\mathbf{w} \mathbf{w}^\top]^{1-\frac{1}{n}} \right) \nabla\phi(\mathbf{w}),$$

where  $I_{d_0} \in \mathbb{R}^{d_0, d_0}$  represents identity, and  $[\cdot]^c$ ,  $c \geq 0$ , stands for a power operator defined over positive semi-definite matrices (with  $c = 0$  yielding identity by definition).

The importance of the vector field  $\mathbf{h}(\cdot)$  lies in the fact that it characterizes the dynamics of the end-to-end matrix — a result proven in [4], stated hereafter for completeness.

**Lemma 10.**  $\mathbf{w}_{n:1}(t)$  is a solution to the following initial value problem:

$$\mathbf{w}_{n:1}(0) = \mathbf{w}_{n:1,s} \quad , \quad \frac{d}{dt} \mathbf{w}_{n:1}(t) = -\mathbf{h}(\mathbf{w}_{n:1}(t)).$$

*Proof.* The lemma follows directly from Theorem 1 in [4].  $\square$

The following lemma will be used throughout the proof.

**Lemma 11.** *Let  $t \in [0, \infty) \cup \{\infty\}$ . Let  $q, \bar{q} : [0, t) \rightarrow \mathbb{R}$  be differentiable functions, and let  $g : [0, t) \times \mathbb{R} \rightarrow \mathbb{R}$  be some locally Lipschitz function. Assume that:*

- (i)  $q(0) \leq \bar{q}(0)$ ;
- (ii)  $\frac{d}{dt}q(t') \leq g(t', q(t'))$  for all  $t' \in [0, t)$ ; and
- (iii)  $\frac{d}{dt}\bar{q}(t') \geq g(t', \bar{q}(t'))$  for all  $t' \in [0, t)$ .

Then  $q(t') \leq \bar{q}(t')$  for all  $t' \in [0, t)$ .

*Proof.* The lemma is a direct consequence of Theorem 10.3 in [27].  $\square$

### J.8.3 Infinite Time

One of the assertions of Proposition 3 is that the gradient flow trajectory is defined over infinite time. This is confirmed by the following lemma.

**Lemma 12.** *The trajectory of gradient flow is defined over infinite time.*

*Proof.* by Theorem 1 we may denote the gradient flow trajectory by  $\theta : [0, t_e) \rightarrow \mathbb{R}^d$ , where either: (i)  $t_e = \infty$ ; or (ii)  $t_e < \infty$  and  $\lim_{t \nearrow t_e} \|\theta(t)\|_2 = \infty$ . Our objective is to show that  $t_e = \infty$ , thus it suffices to establish that  $\theta(\cdot)$  is bounded, i.e. there exists a constant larger than  $\|\theta(t)\|$  for all  $t \in [0, t_e)$ . Recall that  $\theta_s$  meets the balancedness condition (Equation (15)). Theorem 2.2 in [18] implies that the balancedness condition is preserved along the gradient flow trajectory, i.e. for any  $j \in [n-1]$  and  $t \in [0, t_e)$ , it holds that:

$$W_{j+1}^\top(t)W_{j+1}(t) = W_j(t)W_j^\top(t). \quad (50)$$

Using this relation repeatedly, we obtain:

$$\begin{aligned} \|\mathbf{w}_{n:1}(t)\|_2^2 &= \mathbf{w}_{n:1}^\top(t)\mathbf{w}_{n:1}(t) \\ &= W_{n:1}(t)W_{n:1}^\top(t) \\ &= W_{n:2}(t)W_1(t)W_1^\top(t)W_{n:2}^\top(t) \\ &= W_{n:2}(t)W_2^\top(t)W_2(t)W_{n:2}^\top(t) \\ &= W_{n:3}(t)W_2(t)W_2^\top(t)W_2(t)W_2^\top(t)W_{n:3}^\top(t) \\ &= W_{n:3}(t)W_3^\top(t)W_3(t)W_3^\top(t)W_3(t)W_{n:3}^\top(t) \\ &\vdots \\ &= (W_n(t)W_n^\top(t))^n \\ &= \|W_n(t)\|_F^{2n}. \end{aligned}$$

Since the balancedness condition implies that  $\|W_j(t)\|_F = \|W_{j+1}(t)\|_F$  for any  $t \in [0, t_e)$  and  $j \in [n-1]$  (to see this, simply apply trace to both sides of Equation (50)), we may conclude  $\|W_j(t)\|_F^2 = \|\mathbf{w}_{n:1}(t)\|_2^{2/n}$  for any  $j \in [n]$ . Gradient flow monotonically non-increases the objective it optimizes, i.e.  $f(\theta(t))$  is non-increasing. In particular it holds that  $f(\theta(t)) \leq f(\theta(0))$  for all  $t \in [0, t_e)$ . Relying on Equation (16), we obtain  $\|\mathbf{w}_{n:1}(t) - \boldsymbol{\lambda}_{yx}\|_2 \leq \|\mathbf{w}_{n:1}(0) - \boldsymbol{\lambda}_{yx}\|_2$  for all  $t \in [0, t_e)$ . By the triangle inequality we have that  $\|\mathbf{w}_{n:1}(t)\|_2 \leq \|\mathbf{w}_{n:1}(0)\|_2 + 2\|\boldsymbol{\lambda}_{yx}\|_2$ . Thus, for all  $t \in [0, t_e)$ :

$$\|\theta(t)\|_2^2 = \sum_{j=1}^n \|W_j(t)\|_F^2 = n\|\mathbf{w}_{n:1}(t)\|_2^{2/n} \leq n(\|\mathbf{w}_{n:1}(0)\|_2 + \|\boldsymbol{\lambda}_{yx}\|_2)^{2/n}.$$

This completes the proof.  $\square$

### J.8.4 Reparameterization

Consider the initial value problem:

$$\mathbf{u}(0) = \mathbf{w}_{n:1}(0) \quad , \quad \frac{d}{dt} \mathbf{u}(t) = -\|\mathbf{u}(t)\|_2 (n\mathbf{u}(t) - \boldsymbol{\lambda}_{yx}) + (n-1)\|\mathbf{u}(t)\|_2^{-1} \mathbf{u}(t) \mathbf{u}(t)^\top \boldsymbol{\lambda}_{yx}. \quad (51)$$

Lemma 13 below establishes existence of a unique solution to this problem.

**Lemma 13.** *The initial value problem in Equation (51) admits a solution  $\mathbf{u} : [0, t_e) \rightarrow \mathbb{R}^{d_0} \setminus \{\mathbf{0}\}$ , where either: (i)  $t_e = \infty$ ; or (ii)  $t_e < \infty$  and  $\lim_{t \nearrow t_e} \|\mathbf{u}(t)\|_2 \in \{0, \infty\}$ . Moreover, the solution is unique in the sense that any other solution  $\mathbf{u}' : [0, t'_e) \rightarrow \mathbb{R}^{d_0} \setminus \{\mathbf{0}\}$  must satisfy  $t'_e \leq t_e$  and  $\forall t \in [0, t'_e) : \mathbf{u}'(t) = \mathbf{u}(t)$ .*

*Proof.* Define  $\mathbf{g} : [0, \infty) \times \mathbb{R}^{d_0} \setminus \{\mathbf{0}\} \rightarrow \mathbb{R}^{d_0}$  by:

$$\mathbf{g}(t, \mathbf{w}) := -\|\mathbf{w}\|_2 (n\mathbf{w} - \boldsymbol{\lambda}_{yx}) + (n-1)\|\mathbf{w}\|_2^{-1} \mathbf{w} \mathbf{w}^\top \boldsymbol{\lambda}_{yx}.$$

The dynamics in Equation (51) can be written as  $\frac{d}{dt} \mathbf{u}(t) = \mathbf{g}(t, \mathbf{u}(t))$ . Since  $\mathbf{g}(\cdot)$  is locally Lipschitz continuous, the lemma follows directly from the results in Section 1.5 of [25].  $\square$

Hereafter, we denote by  $\mathbf{u} : [0, t_e) \rightarrow \mathbb{R}^{d_0} \setminus \{\mathbf{0}\}$  the (unique) solution to Equation (51). In the remainder of the current subsubappendix we will show that  $\mathbf{u}(\cdot)$  is a reparameterization of the gradient flow trajectory, or more precisely, of  $\mathbf{w}_{n:1}(\cdot)$ .

The following definition overloads notation by extending the scalar  $\nu$  (defined in the statement of Proposition 3) to a function.

**Definition 3.** Define  $\nu : [0, t_e) \rightarrow [-1, 1]$  by  $\nu(t) = \frac{\boldsymbol{\lambda}_{yx}^\top \mathbf{u}(t)}{\|\boldsymbol{\lambda}_{yx}\|_2 \|\mathbf{u}(t)\|_2}$ .

Notice that  $\nu(0)$  coincides with the original (scalar) definition of  $\nu$ . Lemma 14 below makes use of  $\nu(\cdot)$  for characterizing the dynamics of the norm of  $\mathbf{u}(\cdot)$ .

**Lemma 14.** *For all  $t \in [0, t_e)$ :*

$$\frac{d}{dt} \|\mathbf{u}(t)\|_2 = n \|\mathbf{u}(t)\|_2 \left( \nu(t) - \|\mathbf{u}(t)\|_2 \right).$$

*Proof.* Recall that  $\|\boldsymbol{\lambda}_{yx}\| = 1$ . For all  $t \in [0, t_e)$ , it holds that:

$$\begin{aligned} \frac{d}{dt} \|\mathbf{u}(t)\|_2 &= \frac{\mathbf{u}(t)^\top}{\|\mathbf{u}(t)\|_2} \frac{d}{dt} \mathbf{u}(t) \\ &= \frac{\mathbf{u}(t)^\top}{\|\mathbf{u}(t)\|_2} \left( -\|\mathbf{u}(t)\|_2 (n\mathbf{u}(t) - \boldsymbol{\lambda}_{yx}) + (n-1)\|\mathbf{u}(t)\|_2^{-1} \mathbf{u}(t) \mathbf{u}(t)^\top \boldsymbol{\lambda}_{yx} \right) \\ &= -n \|\mathbf{u}(t)\|_2^2 + \|\mathbf{u}(t)\|_2 \nu(t) + (n-1) \|\mathbf{u}(t)\|_2 \nu(t) \\ &= n \|\mathbf{u}(t)\|_2 \left( \nu(t) - \|\mathbf{u}(t)\|_2 \right), \end{aligned}$$

where the first transition follows from the chain rule and derivative of a (non-zero) vector norm, and the second transition follows from  $\mathbf{u}(\cdot)$  being a solution to Equation (51).  $\square$

Relying on Lemma 14, Lemma 15 below derives upper and lower bounds for the norm of  $\mathbf{u}(\cdot)$ .

**Lemma 15.** *For all  $t \in [0, t_e)$ :*

$$\|\mathbf{u}(0)\|_2 e^{-2nt} \leq \|\mathbf{u}(t)\|_2 \leq \|\mathbf{u}(0)\|_2 e^{nt}.$$

*Proof.* We start by proving the upper bound. Recall that by Lemma 14 we have that  $\frac{d}{dt} \|\mathbf{u}(t)\|_2 = n \|\mathbf{u}(t)\|_2 (\nu(t) - \|\mathbf{u}(t)\|_2)$ . It holds that  $\frac{d}{dt} \|\mathbf{u}(t)\|_2 \leq n \|\mathbf{u}(t)\|_2$ , as  $\nu(t) \leq 1$  (by Definition 3). Integrating over time:

$$\ln(\|\mathbf{u}(t)\|_2) - \ln(\|\mathbf{u}(0)\|_2) = \int_0^t \frac{1}{\|\mathbf{u}(t')\|_2} \frac{d}{dt} \|\mathbf{u}(t')\|_2 dt' \leq \int_0^t n dt' = nt.$$

It follows that  $\|\mathbf{u}(t)\|_2 \leq \|\mathbf{u}(0)\|_2 e^{nt}$ .



Moving on to the lower bound, define  $g: [0, t_e) \times \mathbb{R} \rightarrow \mathbb{R}$  by:

$$g(t, z) := \begin{cases} -nz(1+z) & z \geq 1 \\ -2nz & z < 1 \end{cases}.$$

Note that  $g(\cdot)$  is locally Lipschitz continuous. For all  $t \in [0, t_e)$ , it holds that:

$$\|\mathbf{u}(0)\|_2 e^{-2nt} \leq \|\mathbf{u}(0)\|_2 = \|\mathbf{w}_{n:1}(0)\|_2 < 1,$$

where the equality follows from  $u(\cdot)$  being a solution to Equation (51), and the last inequality follows from an assumption made in Proposition 3. Using this fact, the following holds for all  $t \in [0, t_e)$ :

$$\frac{d}{dt} (\|\mathbf{u}(0)\|_2 e^{-2nt}) = -2n\|\mathbf{u}(0)\|_2 e^{-2nt} = g(t, \|\mathbf{u}(0)\|_2 e^{-2nt}).$$

On the other hand, recalling that  $\nu(t) \geq -1$  (by Definition 3) for all  $t \in [0, t_e)$ , it holds (for both cases  $\|\mathbf{u}(t)\|_2 < 1$  and  $\|\mathbf{u}(t)\|_2 \geq 1$ ) that:

$$\frac{d}{dt} \|\mathbf{u}(t)\|_2 = n\|\mathbf{u}(t)\|_2 (\nu(t) - \|\mathbf{u}(t)\|_2) \geq g(t, \|\mathbf{u}(t)\|_2).$$

We may now use Lemma 11 to conclude  $\|\mathbf{u}(0)\|_2 e^{-2nt} \leq \|\mathbf{u}(t)\|_2$  for all  $t \in [0, t_e)$ .  $\square$

Taken together, Lemmas 13 and 15 imply that  $\mathbf{u}(\cdot)$  is defined over infinite time. We formalize this in Lemma 16 below.

**Lemma 16.** *It holds that  $t_e = \infty$ , i.e. we may write  $\mathbf{u}: [0, \infty) \rightarrow \mathbb{R}^{d_0} \setminus \{\mathbf{0}\}$ .*

*Proof.* Assume by contradiction that  $t_e < \infty$ . Lemma 13 implies  $\lim_{t \nearrow t_e} \|\mathbf{u}(t)\|_2 \in \{0, \infty\}$ . On the other hand, by Lemma 15 we have that  $\liminf_{t \nearrow t_e} \|\mathbf{u}(0)\|_2 e^{-2nt_e} \geq \|\mathbf{u}(0)\|_2 e^{-2nt_e}$  and  $\limsup_{t \nearrow t_e} \|\mathbf{u}(0)\|_2 e^{-2nt_e} \leq \|\mathbf{u}(0)\|_2 e^{-2nt_e}$ , which is a contradiction. Hence it must be that  $t_e = \infty$ .  $\square$

Finally, we are in a position to prove that  $\mathbf{u}(\cdot)$  is indeed a (monotonic) reparameterization of  $\mathbf{w}_{n:1}(\cdot)$ .

**Lemma 17.** *For all  $t \geq 0$ :*

$$\mathbf{w}_{n:1}(\xi(t)) = \mathbf{u}(t),$$

where  $\xi: [0, \infty) \rightarrow \mathbb{R}_{\geq 0}$  is defined by  $\xi(t) := \int_0^t \|\mathbf{u}(t')\|_2^{-(1-2/n)} dt'$ .

*Proof.* Define  $\mathbf{g}: [0, \infty) \times \mathbb{R}^{d_0} / \{\mathbf{0}\} \rightarrow \mathbb{R}^{d_0}$  by  $\mathbf{g}(t, \mathbf{w}) := -\mathbf{h}(\mathbf{w}) / \|\mathbf{u}(t)\|_2^{1-2/n}$ . Note that  $\mathbf{g}(\cdot)$  is locally Lipschitz continuous. Define the following initial value problem:

$$\mathbf{q}(0) = \mathbf{w}_{n:1}(0) \quad , \quad \frac{d}{dt} \mathbf{q}(t) = \mathbf{g}(t, \mathbf{q}(t)). \quad (52)$$

We will show both  $\mathbf{u}(\cdot)$  and  $\mathbf{w}_{n:1}(\xi(\cdot))$  are solutions to Equation (52), which by uniqueness implies  $\mathbf{u}(\cdot) = \mathbf{w}_{n:1}(\xi(\cdot))$  for all  $t \geq 0$ , as required. By the definition of  $\mathbf{u}(\cdot)$  (solution to Equation (51)) it holds that  $\mathbf{u}(0) = \mathbf{w}_{n:1}(0) = \mathbf{w}_{n:1}(\xi(0))$ . With the help of Lemma 10 we establish the following for  $t \geq 0$ :

$$\frac{d}{dt} (\mathbf{w}_{n:1}(\xi(t))) = \frac{d}{dt} \mathbf{w}_{n:1}(\xi(t)) \cdot \frac{d\xi}{dt}(t) = -\mathbf{h}(\mathbf{w}_{n:1}(\xi(t))) / \|\mathbf{u}(t)\|_2^{1-2/n} = \mathbf{g}(t, \mathbf{w}_{n:1}(\xi(t))).$$

Recall that  $\mathbf{u}(\cdot)$  is a solution to Equation (51). For all  $t \geq 0$ , it holds that:

$$\begin{aligned} \frac{d}{dt} \mathbf{u}(t) &= -\|\mathbf{u}(t)\|_2 (n\mathbf{u}(t) - \boldsymbol{\lambda}_{yx}) + (n-1)\|\mathbf{u}(t)\|_2^{-1} \mathbf{u}(t) \mathbf{u}(t)^\top \boldsymbol{\lambda}_{yx} \\ &= -\|\mathbf{u}(t)\|_2 (\mathbf{u}(t) - \boldsymbol{\lambda}_{yx}) - (n-1)\|\mathbf{u}(t)\|_2 \mathbf{u}(t) + (n-1)\|\mathbf{u}(t)\|_2^{-1} \mathbf{u}(t) \mathbf{u}(t)^\top \boldsymbol{\lambda}_{yx} \\ &= -\|\mathbf{u}(t)\|_2 (\mathbf{u}(t) - \boldsymbol{\lambda}_{yx}) - (n-1) [\mathbf{u}(t) \mathbf{u}(t)^\top]^\frac{1}{2} \mathbf{u}(t) + (n-1) [\mathbf{u}(t) \mathbf{u}(t)^\top]^\frac{1}{2} \boldsymbol{\lambda}_{yx} \\ &= -\|\mathbf{u}(t)\|_2 (\mathbf{u}(t) - \boldsymbol{\lambda}_{yx}) - (n-1) [\mathbf{u}(t) \mathbf{u}(t)^\top]^\frac{1}{2} (\mathbf{u}(t) - \boldsymbol{\lambda}_{yx}) \\ &= -\|\mathbf{u}(t)\|_2 (\mathbf{u}(t) - \boldsymbol{\lambda}_{yx}) - (n-1) [\mathbf{u}(t) \mathbf{u}(t)^\top]^\frac{1}{2} \nabla \phi(\mathbf{u}(t)) \\ &= -\left( \|\mathbf{u}(t)\|_2^{2-2/n} (\mathbf{u}(t) - \boldsymbol{\lambda}_{yx}) + (n-1) [\mathbf{u}(t) \mathbf{u}(t)^\top]^{1-1/n} \nabla \phi(\mathbf{u}(t)) \right) / \|\mathbf{u}(t)\|_2^{1-2/n} \\ &= -\mathbf{h}(\mathbf{u}(t)) / \|\mathbf{u}(t)\|_2^{1-2/n} \\ &= \mathbf{g}(t, \mathbf{u}(t)). \end{aligned}$$

The above confirms that  $\mathbf{u}(\cdot)$  and  $\mathbf{w}_{n:1}(\xi(\cdot))$  are both solutions to Equation (52), thereby completing the proof.  $\square$

### J.8.5 Minimal Distance From Origin

In this subsubappendix we derive a lower bound on the minimal distance of  $\mathbf{u}(\cdot)$  — solution to Equation (51), which by Lemma 17 is a reparameterization of  $\mathbf{w}_{n:1}(\cdot)$  — from the origin. We denote this minimal distance by  $u_{\min}$ , *i.e.* we let  $u_{\min} := \inf_{t \geq 0} \|\mathbf{u}(t)\|_2$ .

Recall the function  $\nu(\cdot)$  from Definition 3. Lemma 18 below establishes several properties of this function.

**Lemma 18.** *For all  $t \geq 0$ , the following hold:*

- (i)  $\nu(t) \in (-1, 1]$ ;
- (ii)  $\frac{d}{dt}\nu(t) = 1 - \nu(t)^2$ ;
- (iii)  $\nu(t) = 1 - 2 \cdot \frac{1 - \nu(0)}{1 + \nu(0)} / \left( \frac{1 - \nu(0)}{1 + \nu(0)} + e^{2t} \right)$ ; and
- (iv)  $\lim_{t \rightarrow \infty} \nu(t) = 1$ .

*Proof.* Recall  $\|\boldsymbol{\lambda}_{yx}\|_2 = 1$ . It holds that:

$$\begin{aligned} \frac{d}{dt}\nu(t) &= \boldsymbol{\lambda}_{yx}^\top \frac{d}{dt} \left( \frac{\mathbf{u}(t)}{\|\mathbf{u}(t)\|_2} \right) \\ &= \boldsymbol{\lambda}_{yx}^\top \frac{\frac{d}{dt}\mathbf{u}(t)\|\mathbf{u}(t)\|_2 - \mathbf{u}(t)\frac{d}{dt}\|\mathbf{u}(t)\|_2}{\|\mathbf{u}(t)\|_2^2} \\ &= \frac{\boldsymbol{\lambda}_{yx}^\top \frac{d}{dt}\mathbf{u}(t)\|\mathbf{u}(t)\|_2 - \mathbf{u}(t)\frac{d}{dt}\|\mathbf{u}(t)\|_2}{\|\mathbf{u}(t)\|_2}. \end{aligned}$$

Plugging in the expression for  $\frac{d}{dt}\mathbf{u}(t)$  from Equation (51) and the one of  $\frac{d}{dt}\|\mathbf{u}(t)\|_2$  from Lemma 14 (while dividing by  $\|\mathbf{u}(t)\|_2$ ) affirms property (ii):

$$\begin{aligned} \frac{d}{dt}\nu(t) &= \frac{\boldsymbol{\lambda}_{yx}^\top}{\|\mathbf{u}(t)\|_2} \left( \left( (n-1) \frac{\mathbf{u}(t)\mathbf{u}(t)^\top}{\|\mathbf{u}(t)\|_2} \boldsymbol{\lambda}_{yx} - \|\mathbf{u}(t)\|_2 (n\mathbf{u}(t) - \boldsymbol{\lambda}_{yx}) \right) - \left( \mathbf{u}(t)n(\nu(t) - \|\mathbf{u}(t)\|_2) \right) \right) \\ &= \left( (n-1)\nu(t)^2 - \|\mathbf{u}(t)\|_2 (n\nu(t) - 1/\|\mathbf{u}(t)\|_2) \right) - \left( \nu(t)n(\nu(t) - \|\mathbf{u}(t)\|_2) \right) \\ &= (n-1)\nu(t)^2 - n\nu(t)\|\mathbf{u}(t)\|_2 + 1 - n\nu(t)^2 + n\nu(t)\|\mathbf{u}(t)\|_2 \\ &= 1 - \nu(t)^2. \end{aligned}$$

By Theorem 1, the initial value problem which  $\nu(t)$  solves (*i.e.*  $\nu(0) = 0$  and  $\frac{d}{dt}\nu(t) = 1 - \nu(t)^2$ ) admits a unique solution. Since  $t \mapsto 1 - 2 \cdot \frac{1 - \nu(0)}{1 + \nu(0)} / \left( \frac{1 - \nu(0)}{1 + \nu(0)} + e^{2t} \right)$  is a solution to this problem, it must be that  $\nu(t) = 1 - 2 \cdot \frac{1 - \nu(0)}{1 + \nu(0)} / \left( \frac{1 - \nu(0)}{1 + \nu(0)} + e^{2t} \right)$ . This confirms property (iii). Properties (i) and (iv) immediately follow.  $\square$

Below we define a point in time that will turn out to be one at which the distance of  $\mathbf{u}(\cdot)$  from the origin is minimal (*i.e.* is equal to  $u_{\min}$ ).

**Definition 4.** Let  $t_m := \inf\{t \geq 0 : \nu(t) \geq \|\mathbf{u}(t)\|_2\}$ , where by convention  $t_m = \infty$  if  $\nu(t) < \|\mathbf{u}(t)\|_2$  for all  $t \geq 0$ .

Lemma 19 below establishes that  $t_m$  is finite, that the norm of  $\mathbf{u}(\cdot)$  is monotonically decreasing until  $t_m$  and monotonically non-decreasing thereafter, and that this norm remains smaller than one.

**Lemma 19.** *It holds that:*

- (i)  $t_m < \infty$ ;
- (ii)  $\frac{d}{dt}\|\mathbf{u}(t)\|_2 < 0$  for all  $t \in [0, t_m)$ ;
- (iii)  $\frac{d}{dt}\|\mathbf{u}(t)\|_2 \geq 0$  for all  $t \in [t_m, \infty)$ ; and
- (iv)  $\|\mathbf{u}(t)\|_2 < 1$  for all  $t \geq 0$ .

*Proof.* We start by treating the special case where  $\nu(0) = 1$ . Recall that by assumption  $\|\mathbf{w}_{n:1}(0)\| < 1$ . Together with Equation (51) this implies  $\|\mathbf{u}(0)\|_2 = \|\mathbf{w}_{n:1}(0)\|_2 < 1 = \nu(0)$ . Thus, by definition  $t_m = 0$ . We trivially obtain properties (i) and (ii). By Lemma 14 together with property (iii) from Lemma 18 it holds that  $\frac{d}{dt}\|\mathbf{u}(t)\|_2 = n\|\mathbf{u}(t)\|_2(1 - \|\mathbf{u}(t)\|_2)$ . These dynamics, along with the initial value  $\|\mathbf{u}(0)\|_2$ , induce an initial value problem whose unique solution is  $\|\mathbf{u}(t)\|_2 = e^{nt} / (e^{nt} + \|\mathbf{u}(0)\|^{-1} - 1)$ . This confirms properties (iii) and (iv). From this point onward we assume  $\nu(0) \neq 1$ .

By definition of  $t_m$ , it holds that  $\|\mathbf{u}(t)\|_2 > \nu(t)$  for  $t \in [0, t_m)$ . Together with Lemma 14 this implies property (ii).

Relying on property (ii), we have  $\|\mathbf{u}(t)\|_2 \leq \|\mathbf{u}(0)\|_2 < 1$  for  $t \in [0, t_m)$ , where we used the fact that  $\|\mathbf{u}(0)\|_2 = \|\mathbf{w}_{n:1}(0)\|_2 < 1$ . Assume by contradiction that property (i) does not hold, i.e.  $t_m = \infty$ . This means that  $\nu(t) < \|\mathbf{u}(t)\|_2 \leq \|\mathbf{u}(0)\|_2 < 1$  for all  $t \geq 0$ . On the other hand, by Lemma 18,  $\lim_{t \rightarrow \infty} \nu(t) = 1$  — a contradiction. Thus, property (i) must hold.

From the definition of  $t_m$ , together with continuity of  $\nu(t)$  and  $\|\mathbf{u}(t)\|_2$ , it must be that  $\nu(t_m) \geq \|\mathbf{u}(t_m)\|_2$ . Define  $\bar{t}_m := \inf\{t \geq t_m : \nu(t) < \|\mathbf{u}(t)\|_2\}$ , where by convention the infimum of the empty set is equal to infinity. Property (iii) of Lemma 18 together with  $\nu(0) < 1$  imply that  $\nu(t) < 1$  for all  $t \geq 0$  (recall that we are treating the case  $\nu(0) \neq 1$ ). We have previously shown (in the proof of this lemma) that  $\|\mathbf{u}(t)\|_2 < 1$  for  $t \in [0, t_m)$ . By definition of  $\bar{t}_m$ , it holds that  $\|\mathbf{u}(t)\|_2 \leq \nu(t) < 1$  for all  $t \in [t_m, \bar{t}_m)$ . Relying on this inequality together with Lemma 14, we have that  $\frac{d}{dt}\|\mathbf{u}(t)\|_2 \geq 0$  for  $t \in [t_m, \bar{t}_m)$ . If  $\bar{t}_m = \infty$ , then we obtain properties (iii) and (iv), thereby finishing the proof. Assume by contradiction that this is not the case, i.e.  $\bar{t}_m < \infty$ . From continuity  $\|\mathbf{u}(\bar{t}_m)\|_2 = \nu(\bar{t}_m)$ . By Lemmas 14 and 18:

$$\frac{d}{dt}\nu(t)|_{t=\bar{t}_m} = 1 - \nu(\bar{t}_m)^2 > 0 = n\|\mathbf{u}(\bar{t}_m)\|_2 \left( \nu(\bar{t}_m) - \|\mathbf{u}(\bar{t}_m)\|_2 \right) = \frac{d}{dt}\|\mathbf{u}(t)\|_2|_{t=\bar{t}_m},$$

implying existence of a right neighborhood of  $\bar{t}_m$  which contradicts its definition.  $\square$

As a direct consequence of Lemma 19, we obtain that the distance of  $\mathbf{u}(\cdot)$  from the origin is indeed minimal at time  $t_m$ . This is formalized in Lemma 20 below.

**Lemma 20.** *It holds that  $\|\mathbf{u}(t_m)\|_2 = u_{\min}$ . Moreover, if  $\nu(0) \leq \|\mathbf{u}(0)\|_2$  then  $\nu(t_m) = u_{\min}$ .*

*Proof.*  $\|\mathbf{u}(t_m)\|_2 = u_{\min}$  directly follows from properties (ii) and (iii) of Lemma 19. In the case where  $\nu(0) \leq \|\mathbf{u}(0)\|_2$ , from continuity of  $\nu(t)$  and  $\|\mathbf{u}(t)\|_2$ , and from the definition of  $t_m$ , it must be that  $\|\mathbf{u}(t_m)\|_2 = \nu(t_m)$ . Together with  $\|\mathbf{u}(t_m)\|_2 = u_{\min}$ , this concludes the proof.  $\square$

Finally, we are ready to establish a lower bound for  $u_{\min}$ .

**Lemma 21.** *It holds that:*

$$u_{\min} \geq \|\mathbf{w}_{n:1}(0)\|_2 \min \left\{ 1, \left( \frac{2}{3} \frac{1+\nu(0)}{1-\nu(0)} \right)^n \right\},$$

where in the case  $\nu(0) = 1$  the fraction  $(1+\nu(0))/(1-\nu(0))$  is to be interpreted as equal to infinity, leading to  $u_{\min} \geq \|\mathbf{w}_{n:1}(0)\|_2$ .

*Proof.* We split the proof into two possible cases: (i)  $\|\mathbf{u}(0)\|_2 \leq \nu(0)$ ; and (ii)  $\|\mathbf{u}(0)\|_2 > \nu(0)$ .

In case (i), by definition  $t_m = 0$ . By taking Lemma 20 together with Equation (51), we obtain  $u_{\min} = \|\mathbf{u}(t_m)\|_2 = \|\mathbf{u}(0)\|_2 = \|\mathbf{w}_{n:1}(0)\|_2$ .

Moving on to case (ii), recall that by assumption  $\|\mathbf{w}_{n:1}(0)\|_2 \leq 0.2$  and  $\nu(0) \neq -1$ . By Equation (51), we have that  $\|\mathbf{u}(0)\|_2 = \|\mathbf{w}_{n:1}(0)\|_2$ . Define  $t_b := \frac{1}{2} \ln \left( \frac{1+\|\mathbf{u}(0)\|_2}{1-\|\mathbf{u}(0)\|_2} \cdot \frac{1-\nu(0)}{1+\nu(0)} \right)$ , which we will show upper bounds  $t_m$ . Plugging  $t = t_b$  into the explicit expression for  $\nu(t)$  given in property (iii) of Lemma 18, we have that  $\nu(t_b) = \|\mathbf{u}(0)\|_2$ . Taking this together with Lemma 20, we obtain  $\nu(t_b) = \|\mathbf{u}(0)\|_2 \geq u_{\min} = \nu(t_m)$ . Note that  $\nu(0) < \|\mathbf{u}(0)\|_2 < 1$ , and property (iii) of Lemma 18, together imply that  $\nu(t)$  is (strictly) monotonically increasing. Thus, it must be that  $t_b \geq t_m$ . Combining this observation with Lemma 15 yields:

$$\begin{aligned} u_{\min} &= \|\mathbf{u}(t_m)\|_2 \\ &\geq \|\mathbf{u}(0)\|_2 \exp(-2nt_m) \\ &\geq \|\mathbf{u}(0)\|_2 \exp(-2nt_b) \\ &= \|\mathbf{u}(0)\|_2 \left( \frac{1+\|\mathbf{u}(0)\|_2}{1-\|\mathbf{u}(0)\|_2} \cdot \frac{1-\nu(0)}{1+\nu(0)} \right)^{-n}. \end{aligned}$$

Recalling that  $\|\mathbf{u}(0)\|_2 = \|\mathbf{w}_{n:1}(0)\|_2 \leq 0.2$  enables us to conclude the proof for this case.  $\square$

### J.8.6 Escape From Origin

Recall that  $\mathbf{u}(\cdot)$  is the (unique) solution to Equation (51), which by Lemma 17 is a reparameterization of  $\mathbf{w}_{n:1}(\cdot)$ . Recall also the function  $\nu(\cdot)$  from Definition 3, which quantifies the alignment between  $\mathbf{u}(\cdot)$  and  $\lambda_{yx}$ . The current subsubappendix defines a certain point in time (Definition 5), establishes that after this point  $\mathbf{u}(\cdot)$  and  $\lambda_{yx}$  are highly aligned (Lemma 22), and shows that this alignment is accompanied by an escape of  $\mathbf{u}(\cdot)$  from the origin (Lemma 23).

**Definition 5.** Define  $t_a := \frac{1}{2} \ln(\max\{5 \cdot \frac{1-\nu(0)}{1+\nu(0)}, 1\})$ .

**Lemma 22.** For all  $t \geq t_a$ :  $\nu(t) \geq \frac{2}{3}$ .

*Proof.* We split the proof into two possible cases: (i)  $\nu(0) \geq \frac{2}{3}$ ; and (ii)  $\nu(0) < \frac{2}{3}$ .

For case (i), it holds that  $t_a = 0$ . We conclude the proof for this case by relying on Lemma 18, which implies that  $\nu(t)$  is monotonically non-decreasing.

Moving on to case (ii), it holds that  $t_a = \frac{1}{2} \ln(5 \cdot \frac{1-\nu(0)}{1+\nu(0)})$ . Plugging  $t = t_a$  into the explicit expression for  $\nu(t)$  given in Lemma 18, we obtain  $\nu(t_a) = \frac{2}{3}$ . We conclude the proof for this case by once again relying on the fact that  $\nu(t)$  is monotonically non-decreasing.  $\square$

**Lemma 23.** For all  $t \geq 0$ :

$$\|\mathbf{u}(t_a + t)\|_2 \geq \frac{2}{3} \cdot \frac{\exp(\frac{2}{3}nt)}{\exp(\frac{2}{3}nt) + \frac{2}{3}u_{\min}^{-1} - 1},$$

where (as defined in Subsubappendix J.8.5)  $u_{\min} := \inf_{t \geq 0} \|\mathbf{u}(t)\|_2$ .

*Proof.* Define  $g: [0, \infty) \rightarrow \mathbb{R}$  by  $g(z) := nz(\frac{2}{3} - z)$ . Notice that  $g(\cdot)$  is locally Lipschitz continuous. Define  $\bar{u}: [0, \infty) \rightarrow \mathbb{R}$  by  $\bar{u}(t) := \frac{2}{3} \exp(\frac{2}{3}nt) / (\exp(\frac{2}{3}nt) + \frac{2}{3}u_{\min}^{-1} - 1)$ . It holds that  $\|\mathbf{u}(t_a)\|_2 \geq u_{\min} = \bar{u}(0)$ . By Lemmas 14 and 22,  $\frac{d}{dt} \|\mathbf{u}(t)\|_2 \geq g(\|\mathbf{u}(t)\|_2)$  for all  $t \geq t_a$ . Furthermore, notice that  $\frac{d}{dt} \bar{u}(t) = g(\bar{u}(t))$ . We may now use Lemma 11 to obtain  $\|\mathbf{u}(t_a + t)\|_2 \geq \bar{u}(t)$  for all  $t \geq 0$ .  $\square$

### J.8.7 Convergence

Recall that  $\mathbf{u}(\cdot)$  is the (unique) solution to Equation (51), and (by Lemma 17) a reparameterization of  $\mathbf{w}_{n:1}(\cdot)$ . Recall also the (alignment guaranteeing) time  $t_a$  from Definition 5, the notation  $u_{\min} := \inf_{t \geq 0} \|\mathbf{u}(t)\|_2$ , and the fact that (by Lemma 21)  $u_{\min} > 0$ . In this subsubappendix we define a certain time duration (Definition 6), and show that after it elapses from  $t_a$ : (i) the norm of  $\mathbf{u}(\cdot)$  is on the order of one, which is the norm of the target solution  $\lambda_{yx}$  (Lemma 24); and (ii)  $\mathbf{u}(\cdot)$  converges to  $\lambda_{yx}$  exponentially fast (Lemma 25).

**Definition 6.** Define  $t_c := \frac{3}{2n} \ln(\frac{2n}{3u_{\min}})$ .<sup>35</sup>

**Lemma 24.** For all  $t \geq 0$ :  $\|\mathbf{u}(t_a + t_c + t)\|_2 \geq \frac{2n}{3(n+1)}$ .

*Proof.* Let  $t \geq 0$ . From Lemma 23:

$$\|\mathbf{u}(t_a + t_c + t)\|_2 \geq \frac{2}{3} \cdot \frac{\exp(\frac{2}{3}n(t_c + t))}{\exp(\frac{2}{3}n(t_c + t)) + \frac{2}{3}u_{\min}^{-1} - 1} \geq \frac{2}{3} \cdot \frac{\exp(\frac{2}{3}nt_c)}{\exp(\frac{2}{3}nt_c) + \frac{2}{3}u_{\min}^{-1}} = \frac{2}{3} \cdot \frac{n(\frac{2}{3}u_{\min}^{-1})}{n(\frac{2}{3}u_{\min}^{-1}) + (\frac{2}{3}u_{\min}^{-1})} = \frac{2n}{3(n+1)}.$$

$\square$

**Lemma 25.** For all  $t \geq 0$ :  $\|\lambda_{yx} - \mathbf{u}(t_a + t_c + t)\|_2 \leq \frac{6}{5} \exp(-\frac{2n}{3(n+1)}t)$ .

<sup>35</sup>Note that, since  $n \geq 2$  and  $u_{\min} \leq \|\mathbf{u}(0)\|_2 = \|\mathbf{w}_{n:1}(0)\|_2 \leq 0.2$ , the time duration  $t_c$  is necessarily positive.

*Proof.* Property (iv) from Lemma 19 together with  $\|\lambda_{yx}\|_2 = 1$  imply that  $\|\lambda_{yx} - \mathbf{u}(t)\|_2 \neq 0$  for all  $t \geq 0$ . Relying on Equation (51), while recalling the function  $\nu(\cdot)$  from Definition 3, we obtain:

$$\begin{aligned}
\frac{d}{dt} \|\lambda_{yx} - \mathbf{u}(t)\|_2 &= \frac{(\lambda_{yx} - \mathbf{u}(t))^\top}{\|\lambda_{yx} - \mathbf{u}(t)\|_2} \frac{d}{dt} (\lambda_{yx} - \mathbf{u}(t)) \\
&= \frac{(\lambda_{yx} - \mathbf{u}(t))^\top}{\|\lambda_{yx} - \mathbf{u}(t)\|_2} \left( \|\mathbf{u}(t)\|_2 (n\mathbf{u}(t) - \lambda_{yx}) - (n-1) \|\mathbf{u}(t)\|_2^{-1} \mathbf{u}(t) \mathbf{u}(t)^\top \lambda_{yx} \right) \\
&= \frac{(\lambda_{yx} - \mathbf{u}(t))^\top}{\|\lambda_{yx} - \mathbf{u}(t)\|_2} \left( \|\mathbf{u}(t)\|_2 (\mathbf{u}(t) - \lambda_{yx}) + (n-1) \|\mathbf{u}(t)\|_2 \mathbf{u}(t) - (n-1) \nu(t) \mathbf{u}(t) \right) \\
&= \frac{(\lambda_{yx} - \mathbf{u}(t))^\top}{\|\lambda_{yx} - \mathbf{u}(t)\|_2} \left( \|\mathbf{u}(t)\|_2 (\mathbf{u}(t) - \lambda_{yx}) + (n-1) (\|\mathbf{u}(t)\|_2 - \nu(t)) \mathbf{u}(t) \right) \\
&= -\|\mathbf{u}(t)\|_2 \|\lambda_{yx} - \mathbf{u}(t)\|_2 + (n-1) (\|\mathbf{u}(t)\|_2 - \nu(t)) \frac{\nu(t) \|\mathbf{u}(t)\|_2 - \|\mathbf{u}(t)\|_2^2}{\|\lambda_{yx} - \mathbf{u}(t)\|_2} \\
&= -\|\mathbf{u}(t)\|_2 \|\lambda_{yx} - \mathbf{u}(t)\|_2 - \frac{(n-1) \|\mathbf{u}(t)\|_2}{\|\lambda_{yx} - \mathbf{u}(t)\|_2} (\|\mathbf{u}(t)\|_2 - \nu(t))^2,
\end{aligned}$$

for all  $t \geq 0$ . By Lemma 24, we may bound  $\frac{d}{dt} \|\lambda_{yx} - \mathbf{u}(t)\|_2 \leq -\frac{2n}{3(n+1)} \|\lambda_{yx} - \mathbf{u}(t)\|_2$  for all  $t \geq t_a + t_c$ . Let  $t' \geq 0$ . We integrate  $\frac{d}{dt} \|\lambda_{yx} - \mathbf{u}(t)\|_2 / \|\lambda_{yx} - \mathbf{u}(t)\|_2$  from  $t = t_a + t_c$  to  $t = t_a + t_c + t'$  in order to obtain  $\|\lambda_{yx} - \mathbf{u}(t_a + t_c + t')\|_2 \leq \|\lambda_{yx} - \mathbf{u}(t_a + t_c)\|_2 \exp(-\frac{2n}{3(n+1)} t')$ . Recall that by assumption  $\|\mathbf{w}_{n:1}(0)\| \leq 0.2$ . By Equation (51) we have that  $\|\mathbf{u}(0)\|_2 = \|\mathbf{w}_{n:1}(0)\|_2$ . We conclude the proof by noting that  $\|\lambda_{yx} - \mathbf{u}(0)\|_2 \leq \|\lambda_{yx}\|_2 + \|\mathbf{u}(0)\|_2 \leq \frac{6}{5}$ .  $\square$

### J.8.8 Time to Convergence

Recall that  $\mathbf{u}(\cdot)$  is the (unique) solution to Equation (51), and that Lemma 17 presents a monotonically increasing function  $\xi(\cdot)$  satisfying  $\mathbf{w}_{n:1}(\xi(t)) = \mathbf{u}(t)$  for all  $t \geq 0$ . Recall also the function  $\nu(\cdot)$  from Definition 3, quantifying the alignment between  $\mathbf{u}(\cdot)$  and  $\lambda_{yx}$ . Finally, recall the times  $t_a$  and  $t_c$  from Definitions 5 and 6, which guarantee alignment and initiation of exponential convergence, respectively. The current subsubappendix makes use of the above to establish that at the time  $\bar{t}$  defined in Equation (17),  $\mathbf{w}_{n:1}(\cdot)$  is  $\bar{\epsilon}$ -optimal, i.e.  $\frac{1}{2} \|\mathbf{w}_{n:1}(\bar{t}) - \lambda_{yx}\|_2^2 \leq \bar{\epsilon}$ .

We begin by defining a certain time duration (Definition 7), and showing that it elapsing from  $t_a + t_c$  ensures that  $\mathbf{u}(\cdot)$  is  $\bar{\epsilon}$ -optimal (Lemma 26).

**Definition 7.** Define  $t_{\bar{\epsilon}} := \frac{3(n+1)}{2n} \ln\left(\frac{6}{5\sqrt{2\bar{\epsilon}}}\right)$ .<sup>36</sup>

**Lemma 26.** It holds that  $\frac{1}{2} \|\mathbf{u}(t_a + t_c + t_{\bar{\epsilon}}) - \lambda_{yx}\|_2^2 \leq \bar{\epsilon}$ .

*Proof.* The proof follows from plugging  $t = t_{\bar{\epsilon}}$  into the result of Lemma 25.  $\square$

Moving from the reparameterized to the original gradient flow trajectory, i.e. from  $\mathbf{u}(\cdot)$  to  $\mathbf{w}_{n:1}(\cdot)$ , we immediately obtain  $\bar{\epsilon}$ -optimality of  $\mathbf{w}_{n:1}(\cdot)$  at time  $\xi(t_a + t_c + t_{\bar{\epsilon}})$ .

**Lemma 27.** It holds that  $\frac{1}{2} \|\mathbf{w}_{n:1}(\xi(t_a + t_c + t_{\bar{\epsilon}})) - \lambda_{yx}\|_2^2 \leq \bar{\epsilon}$ .

*Proof.* The proof immediately follows from Lemmas 17 and 26.  $\square$

Lemma 28 below shows that the time  $\bar{t}$  defined in Equation (17) (recall that the scalar  $\nu$  there, defined in the preceding text, coincides with the value taken by the function  $\nu(\cdot)$  at zero) is greater than or equal to  $\xi(t_a + t_c + t_{\bar{\epsilon}})$ .

**Lemma 28.** It holds that  $\bar{t} \geq \xi(t_a + t_c + t_{\bar{\epsilon}})$ .

*Proof.* By Lemma 19 we have that  $\|\mathbf{u}(t)\|_2 < 1$  for all  $t \geq 0$ . Recall the notation  $u_{\min} := \inf_{t \geq 0} \|\mathbf{u}(t)\|_2$ , and the fact that (by Lemma 21)  $u_{\min} \geq \|\mathbf{w}_{n:1}(0)\|_2 \min\left\{1, \left(\frac{2}{3} \frac{1 + \nu(0)}{1 - \nu(0)}\right)^n\right\}$ . For all  $t \geq 0$ :

$$\xi(t) = \int_0^t \|\mathbf{u}(t')\|_2^{-2(1-2/n)} dt' \leq \int_0^t u_{\min}^{-2} dt' = t u_{\min}^{-2}. \quad (53)$$

<sup>36</sup>Note that  $t_{\bar{\epsilon}} > 0$ , since we assume  $\bar{\epsilon} \leq 1/2$  without loss of generality (cf. Subsubappendix J.8.2).

Recall from Subsubappendix J.8.2 that we assume (without loss of generality)  $\bar{\epsilon} \leq \frac{1}{2}$ . The following holds:

$$\begin{aligned}
t_a + t_c + t_{\bar{\epsilon}} &= \frac{1}{2} \ln \left( \max \left\{ 5 \cdot \frac{1-\nu(0)}{1+\nu(0)}, 1 \right\} \right) + \frac{3}{2n} \ln \left( \frac{2n}{3u_{\min}} \right) + \frac{3(n+1)}{2n} \ln \left( \frac{6}{5\sqrt{2\bar{\epsilon}}} \right) \\
&\leq \frac{1}{2} \ln \left( 5 \max \left\{ \frac{1-\nu(0)}{1+\nu(0)}, 1 \right\} \right) + \ln \left( \frac{2n}{3u_{\min}} \right) + 4 \ln \left( \frac{1}{\sqrt{\bar{\epsilon}}} \right) \\
&\leq \ln \left( \sqrt{5} \max \left\{ \frac{1-\nu(0)}{1+\nu(0)}, 1 \right\} \cdot \frac{2n}{3u_{\min}} \cdot (1/\bar{\epsilon})^2 \right) \\
&\leq \ln \left( 5n \max \left\{ \frac{1-\nu(0)}{1+\nu(0)}, 1 \right\} \cdot (1/\bar{\epsilon})^2 \cdot u_{\min}^{-1} \right).
\end{aligned} \tag{54}$$

Using Equations (53) and (54), we conclude the proof:

$$\begin{aligned}
&\xi(t_a + t_c + t_{\bar{\epsilon}}) \\
&\leq (t_a + t_c + t_{\bar{\epsilon}}) u_{\min}^{-1} \\
&\leq \ln \left( 5n \max \left\{ \frac{1-\nu(0)}{1+\nu(0)}, 1 \right\} \cdot (1/\bar{\epsilon})^2 \cdot u_{\min}^{-1} \right) \cdot u_{\min}^{-1} \\
&\leq \ln \left( 5n (1/\bar{\epsilon})^2 \|\mathbf{w}_{n:1}(0)\|_2^{-1} \max \left\{ 1, \left( \frac{3}{2} \cdot \frac{1-\nu(0)}{1+\nu(0)} \right)^{n+1} \right\} \right) \cdot \|\mathbf{w}_{n:1}(0)\|_2^{-1} \max \left\{ 1, \left( \frac{3}{2} \cdot \frac{1-\nu(0)}{1+\nu(0)} \right)^n \right\} \\
&\leq \ln \left( 15n (1/2\bar{\epsilon}) \|\mathbf{w}_{n:1}(0)\|_2^{-1} \max \left\{ 1, \frac{1-\nu(0)}{1+\nu(0)} \right\} \right) \cdot 2n \|\mathbf{w}_{n:1}(0)\|_2^{-1} \max \left\{ 1, \left( \frac{3}{2} \cdot \frac{1-\nu(0)}{1+\nu(0)} \right)^n \right\} \\
&= \bar{t}.
\end{aligned}$$

□

Combining Lemmas 27 and 28 with the fact that, in general, gradient flow monotonically non-increases the objective it optimizes, we obtain the result which the current subsubappendix set out to prove —  $\bar{\epsilon}$ -optimality of  $\mathbf{w}_{n:1}(\cdot)$  at time  $\bar{t}$ .

**Lemma 29.** *It holds that  $f(\boldsymbol{\theta}(\bar{t})) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) = \frac{1}{2} \|\mathbf{w}_{n:1}(\bar{t}) - \boldsymbol{\lambda}_{yx}\|_2^2 \leq \bar{\epsilon}$ .*

*Proof.* The proof follows directly from Equation (16), Lemmas 27 and 28, and the fact that  $f(\boldsymbol{\theta}(\cdot))$  is monotonically non-increasing. □

### J.8.9 Geometric Analysis

The current subsubappendix analyzes the geometry of the optimization landscape around the gradient flow trajectory. Namely, under the notations of Theorem 3, for  $t > 0$ ,  $\epsilon \in (0, \frac{1}{2n}]$  and corresponding  $\mathcal{D}_{t,\epsilon}$  ( $\epsilon$ -neighborhood of gradient flow trajectory up to time  $t$ ), it establishes a smoothness constant  $\beta_{t,\epsilon} = 16n$ , a Lipschitz constant  $\gamma_{t,\epsilon} = 6\sqrt{n}$ , and the (upper) bound on the integral of (minus) the minimal eigenvalue of the Hessian given in Equation (18) (with the function  $m(\cdot)$  there being non-negative).

Recall (from Lemma 19) that the (Euclidean) norm of  $\mathbf{u}(\cdot)$  — the (unique) solution to Equation (51) — is upper bounded by one. Since (by Lemma 17)  $\mathbf{u}(\cdot)$  is a reparameterization of  $\mathbf{w}_{n:1}(\cdot)$ , the norm of  $\mathbf{w}_{n:1}(\cdot)$  is upper bounded by one as well. This allows proving the following result.

**Lemma 30.** *For all  $t' \geq 0$ :*

$$(\|\mathbf{w}_{n:1}(t')\|_2^{1/n} + \epsilon)^n \leq \|\mathbf{w}_{n:1}(t')\|_2 + 2n\epsilon.$$

*Proof.* For all  $t' \geq 0$ :

$$\begin{aligned}
(\|\mathbf{w}_{n:1}(t')\|_2^{1/n} + \epsilon)^n &= \sum_{j=0}^n \binom{n}{j} \|\mathbf{w}_{n:1}(t')\|_2^{(n-j)/n} \epsilon^j \\
&\leq \sum_{j=0}^n n^j \|\mathbf{w}_{n:1}(t')\|_2^{(n-j)/n} \epsilon^j \\
&= \|\mathbf{w}_{n:1}(t')\|_2 + \sum_{j=1}^n n^j \|\mathbf{w}_{n:1}(t')\|_2^{(n-j)/n} \epsilon^j.
\end{aligned}$$

By Lemma 17 we have that  $\mathbf{w}_{n:1}(\xi(t')) = \mathbf{u}(t')$ , where  $\xi(t') := \int_0^{t'} \|\mathbf{u}(t'')\|_2^{-(1-2/n)} dt''$ .  $\xi(\cdot)$  is unbounded since  $\|\mathbf{u}(\cdot)\|_2$  is bounded by property (iv) of Lemma 19. It follows that:

$$\begin{aligned}
(\|\mathbf{w}_{n:1}(t')\|_2^{1/n} + \epsilon)^n &\leq \|\mathbf{w}_{n:1}(t')\|_2 + \sum_{j=1}^{\infty} (n\epsilon)^j \\
&= \|\mathbf{w}_{n:1}(t')\|_2 + \frac{n\epsilon}{1-n\epsilon} \\
&\leq \|\mathbf{w}_{n:1}(t')\|_2 + 2n\epsilon,
\end{aligned}$$

where the second transition follows from the formula for geometric sum (notice that  $n\epsilon < 1$  since by assumption  $\epsilon \leq 1/2n$ ), and the last transition follows from the assumption  $\epsilon \leq 1/2n$ .  $\square$

Building on Lemma 30, and the fact that (by assumption) the gradient flow trajectory  $\boldsymbol{\theta}(\cdot)$  emanates from a balanced initialization (*i.e.* an initialization whose weight matrices satisfy the condition in Equation (15)) — which by [18] implies that  $\boldsymbol{\theta}(t')$  is balanced for any  $t' \geq 0$  — the lemma below establishes different properties for weight settings lying  $\epsilon$ -away from the trajectory.

**Lemma 31.** *Let  $t' \geq 0$  and let  $\boldsymbol{\theta}_\epsilon \in \mathbb{R}^d$  be a weight setting satisfying  $\|\boldsymbol{\theta}_\epsilon - \boldsymbol{\theta}(t')\|_2 \leq \epsilon$ . Denote by  $W_{1,\epsilon} \in \mathbb{R}^{d_1, d_0}$ ,  $W_{2,\epsilon} \in \mathbb{R}^{d_2, d_1}$ ,  $\dots$ ,  $W_{n-1,\epsilon} \in \mathbb{R}^{d_{n-1}, d_{n-2}}$ ,  $W_{n,\epsilon} \in \mathbb{R}^{1, d_{n-1}}$  the weight matrices constituting  $\boldsymbol{\theta}_\epsilon$ , and by  $\mathbf{w}_{n:1,\epsilon} \in \mathbb{R}^{d_0}$  the corresponding end-to-end matrix  $W_{n,\epsilon} W_{n-1,\epsilon} \cdots W_{1,\epsilon}$  (in vectorized form). Then, the following hold:*

- (i)  $\|\mathbf{w}_{n:1,\epsilon} - \mathbf{w}_{n:1}(t')\|_2 \leq (\|\mathbf{w}_{n:1}(t')\|_2^{1/n} + \epsilon)^n - \|\mathbf{w}_{n:1}(t')\|_2$ ;
- (ii)  $\|\nabla \phi(\mathbf{w}_{n:1,\epsilon})\|_2 \leq \|\nabla \phi(\mathbf{w}_{n:1}(t'))\|_2 + 2n\epsilon$ ; and
- (iii) for any  $\mathcal{J} \subseteq [n] \setminus \emptyset$ ,  $\prod_{j \in \mathcal{J}} \|W_{j,\epsilon}\|_F \leq (\|\mathbf{w}_{n:1}(t')\|_2 + 2n\epsilon)^{\frac{|\mathcal{J}|}{n}}$ .

*Proof.* For brevity, throughout this proof we omit the time  $t'$  from our notation, *i.e.* we denote  $\boldsymbol{\theta}(t')$ ,  $\mathbf{w}_{n:1}(t')$ ,  $W_{n:1}(t')$  and  $W_1(t'), \dots, W_n(t')$  by  $\boldsymbol{\theta}$ ,  $\mathbf{w}_{n:1}$ ,  $W_{n:1}$  and  $W_1, \dots, W_n$  respectively.

Starting with property (i), we have that:

$$\begin{aligned}
& \|\mathbf{w}_{n:1,\epsilon} - \mathbf{w}_{n:1}\|_2 \\
&= \|W_{n:1,\epsilon} - W_{n:1}\|_F \\
&= \|(W_n + W_{n,\epsilon} - W_n) - (W_1 + W_{1,\epsilon} - W_1) - W_{n:1}\|_F \\
&= \left\| \sum_{(b_1, \dots, b_n) \in \{0,1\}^n} (b_n W_n + (1-b_n)(W_{n,\epsilon} - W_n)) - (b_1 W_1 + (1-b_1)(W_{1,\epsilon} - W_1)) - W_{n:1} \right\|_F \\
&= \left\| \sum_{(b_1, \dots, b_n) \in \{0,1\}^n \setminus \{1\}^n} (b_n W_n + (1-b_n)(W_{n,\epsilon} - W_n)) - (b_1 W_1 + (1-b_1)(W_{1,\epsilon} - W_1)) \right\|_F \\
&\leq \sum_{(b_1, \dots, b_n) \in \{0,1\}^n \setminus \{1\}^n} \|(b_n W_n + (1-b_n)(W_{n,\epsilon} - W_n)) - (b_1 W_1 + (1-b_1)(W_{1,\epsilon} - W_1))\|_F \\
&\leq \sum_{(b_1, \dots, b_n) \in \{0,1\}^n \setminus \{1\}^n} \|b_n W_n + (1-b_n)(W_{n,\epsilon} - W_n)\|_F + \|b_1 W_1 + (1-b_1)(W_{1,\epsilon} - W_1)\|_F \\
&\leq \sum_{(b_1, \dots, b_n) \in \{0,1\}^n \setminus \{1\}^n} (b_n \|W_n\|_F + (1-b_n) \|W_{n,\epsilon} - W_n\|_F) + (b_1 \|W_1\|_F + (1-b_1) \|W_{1,\epsilon} - W_1\|_F) \\
&\leq \sum_{(b_1, \dots, b_n) \in \{0,1\}^n \setminus \{1\}^n} (b_n \|W_n\|_F + (1-b_n)\epsilon) + (b_1 \|W_1\|_F + (1-b_1)\epsilon),
\end{aligned} \tag{55}$$

where the inequalities follow from sub-multiplicativity and sub-additivity of Frobenius norm, as well as the assumption  $\|\boldsymbol{\theta}_\epsilon - \boldsymbol{\theta}\|_2 \leq \epsilon$ . Recall that  $\boldsymbol{\theta}_s$  meets the balancedness condition (Equation (15)). Theorem 2.2 from [18] implies that the balancedness condition holds along the gradient flow trajectory. Therefore,  $\boldsymbol{\theta}$  is balanced, *i.e.* for any  $j \in [n-1]$  it holds that  $W_{j+1}^\top W_{j+1} = W_j W_j^\top$ . Using this relation repeatedly, we obtain:

$$\begin{aligned}
\|\mathbf{w}_{n:1}\|_2^2 &= \mathbf{w}_{n:1}^\top \mathbf{w}_{n:1} \\
&= W_{n:1} W_{n:1}^\top \\
&= W_{n:2} W_1 W_1^\top W_{n:2}^\top \\
&= W_{n:2} W_2^\top W_2 W_{n:2}^\top \\
&= W_{n:3} W_2 W_2^\top W_2 W_2^\top W_{n:3}^\top \\
&= W_{n:3} W_3^\top W_3 W_3^\top W_{n:3}^\top \\
&\vdots \\
&= (W_n W_n^\top)^n \\
&= \|W_n\|_F^{2n}.
\end{aligned} \tag{56}$$

Since the balancedness condition implies that  $\|W_j\|_F = \|W_{j+1}\|_F$  for any  $j \in [n-1]$ , we may conclude  $\|W_j\|_F = \|\mathbf{w}_{n:1}\|_2^{1/n}$  for any  $j \in [n]$ . This, along with Equation (55), establishes property (i):

$$\begin{aligned} & \|\mathbf{w}_{n:1,\epsilon} - \mathbf{w}_{n:1}\|_2 \\ & \leq \sum_{(b_1, \dots, b_n) \in \{0,1\}^n \setminus \{1\}^n} (b_n \|\mathbf{w}_{n:1}\|_2^{1/n} + (1-b_n)\epsilon) - (b_1 \|\mathbf{w}_{n:1}\|_2^{1/n} + (1-b_1)\epsilon) \\ & = \sum_{(b_1, \dots, b_n) \in \{0,1\}^n} (b_n \|\mathbf{w}_{n:1}\|_2^{1/n} + (1-b_n)\epsilon) - (b_1 \|\mathbf{w}_{n:1}\|_2^{1/n} + (1-b_1)\epsilon) - \|\mathbf{w}_{n:1}\|_2 \\ & = (\|\mathbf{w}_{n:1}\|_2^{1/n} + \epsilon)^n - \|\mathbf{w}_{n:1}\|_2. \end{aligned}$$

Moving to property (ii), we have that:

$$\begin{aligned} \|\nabla\phi(\mathbf{w}_{n:1,\epsilon})\|_2 &= \|\mathbf{w}_{n:1,\epsilon} - \boldsymbol{\lambda}_{yx}\|_2 \\ &= \|\mathbf{w}_{n:1} - \boldsymbol{\lambda}_{yx} + \mathbf{w}_{n:1,\epsilon} - \mathbf{w}_{n:1}\|_2 \\ &\leq \|\mathbf{w}_{n:1} - \boldsymbol{\lambda}_{yx}\|_2 + \|\mathbf{w}_{n:1,\epsilon} - \mathbf{w}_{n:1}\|_2 \\ &= \|\nabla\phi(\mathbf{w}_{n:1})\|_2 + \|\mathbf{w}_{n:1,\epsilon} - \mathbf{w}_{n:1}\|_2. \end{aligned}$$

applying property (i), together with Lemma 30, we obtain property (ii):

$$\|\nabla\phi(\mathbf{w}_{n:1,\epsilon})\|_2 \leq \|\nabla\phi(\mathbf{w}_{n:1})\|_2 + (\|\mathbf{w}_{n:1}\|_2^{1/n} + \epsilon)^n - \|\mathbf{w}_{n:1}\|_2 \leq \|\nabla\phi(\mathbf{w}_{n:1})\|_2 + 2n\epsilon.$$

Regarding property (iii), for any  $\mathcal{J} \subseteq [n]$  we have that:

$$\begin{aligned} \prod_{j \in \mathcal{J}} \|W_{j,\epsilon}\|_F &= \prod_{j \in \mathcal{J}} \|W_j + W_{j,\epsilon} - W_j\|_F \\ &\leq \prod_{j \in \mathcal{J}} (\|W_j\|_F + \|W_{j,\epsilon} - W_j\|_F) \\ &\leq \prod_{j \in \mathcal{J}} (\|W_j\|_F + \epsilon) \\ &= \prod_{j \in \mathcal{J}} (\|\mathbf{w}_{n:1}\|_2^{1/n} + \epsilon) \\ &= (\|\mathbf{w}_{n:1}\|_2^{1/n} + \epsilon)^{|\mathcal{J}|} \\ &= \left( (\|\mathbf{w}_{n:1}\|_2^{1/n} + \epsilon)^n \right)^{\frac{|\mathcal{J}|}{n}}, \end{aligned}$$

where the third transition follows from the assumption  $\|\boldsymbol{\theta}_\epsilon - \boldsymbol{\theta}\|_2 \leq \epsilon$ , and the fourth from Equation (56). Applying Lemma 30 concludes the proof of property (iii), and the entire lemma.  $\square$

The following lemma analyzes the Hessian and gradient of the training loss  $f(\cdot)$ , bounding their spectral and Euclidean norms respectively.

**Lemma 32.** *For any weight setting  $\boldsymbol{\theta} \in \mathbb{R}^d$  with corresponding weight matrices  $W_1 \in \mathbb{R}^{d_1, d_0}, W_2 \in \mathbb{R}^{d_2, d_1}, \dots, W_{n-1} \in \mathbb{R}^{d_{n-1}, d_{n-2}}, W_n \in \mathbb{R}^{1, d_{n-1}}$ , the following hold:<sup>13</sup>*

$$\begin{aligned} (i) \quad & \|\nabla^2 f(\boldsymbol{\theta})\|_s \leq n \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-1}} \prod_{j \in \mathcal{J}} \|W_j\|_F^2 + 2n \|\nabla\phi(\mathbf{w}_{n:1})\|_2 \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_F; \text{ and} \\ (ii) \quad & \|\nabla f(\boldsymbol{\theta})\|_2 \leq \sqrt{n} \|\nabla\phi(\mathbf{w}_{n:1})\|_2 \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-1}} \prod_{j \in \mathcal{J}} \|W_j\|_F. \end{aligned}$$

*Proof.* Let  $\Delta W_1 \in \mathbb{R}^{d_1, d_0}, \Delta W_2 \in \mathbb{R}^{d_2, d_1}, \dots, \Delta W_{n-1} \in \mathbb{R}^{d_{n-1}, d_{n-2}}, \Delta W_n \in \mathbb{R}^{1, d_{n-1}}$ .

We begin with property (i). By Lemma 1 we have that:

$$\begin{aligned} \nabla^2 f(\boldsymbol{\theta})[\Delta W_1, \dots, \Delta W_n] &= \nabla^2 \phi(W_{n:1}) \left[ \sum_{j=1}^n W_{n:j+1} (\Delta W_j) W_{j-1:1} \right] \\ &+ 2\text{Tr} \left( \nabla\phi(W_{n:1})^\top \sum_{1 \leq j < j' \leq n} W_{n:j'+1} (\Delta W_{j'}) W_{j'-1:j+1} (\Delta W_j) W_{j-1:1} \right), \end{aligned} \tag{57}$$



where  $W_{j':j}$ , for any  $j, j' \in \{1, 2, \dots, n\}$ , is defined as  $W_{j'} W_{j'-1} \dots W_j$  if  $j \leq j'$ , and as an identity matrix (with size to be inferred by context) otherwise. We will upper bound each of the two summands on the right-hand side of Equation (57). We bound the first summand as follows:

$$\begin{aligned}
& \nabla^2 \phi(W_{n:1}) \left[ \sum_{j=1}^n W_{n:j+1} (\Delta W_j) W_{j-1:1} \right] \\
&= \left\| \sum_{j=1}^n W_{n:j+1} (\Delta W_j) W_{j-1:1} \right\|_F^2 \\
&\leq \left( \sum_{j=1}^n \|W_{n:j+1} (\Delta W_j) W_{j-1:1}\|_F \right)^2 \\
&\leq n \sum_{j=1}^n \|W_{n:j+1} (\Delta W_j) W_{j-1:1}\|_F^2 \\
&\leq n \sum_{j=1}^n \|W_n\|_F^2 \dots \|W_{j+1}\|_F^2 \|\Delta W_j\|_F^2 \|W_{j-1}\|_F^2 \dots \|W_1\|_F^2 \\
&\leq n \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-1}} \prod_{j \in \mathcal{J}} \|W_j\|_F^2 \cdot \sum_{j=1}^n \|\Delta W_j\|_F^2,
\end{aligned} \tag{58}$$

where the first transition follows from the fact that the Hessian of  $\phi(\cdot)$  is an identity (since  $\phi(W) = \frac{1}{2} \|W - \Lambda_{yx}\|_F^2 + c$ ), the second transition follows from the triangle inequality, the third transition follows from the one-norm of a vector in  $\mathbb{R}^n$  being no greater than  $\sqrt{n}$  times its Euclidean norm, and the fourth transition follows from sub-multiplicativity of Frobenius norm. Moving on to bounding the second summand on the right-hand side of Equation (57):

$$\begin{aligned}
& 2\text{Tr} \left( \nabla \phi(W_{n:1})^\top \sum_{1 \leq j < j' \leq n} W_{n:j'+1} (\Delta W_{j'}) W_{j'-1:j+1} (\Delta W_j) W_{j-1:1} \right) \\
&\leq 2 \|\nabla \phi(W_{n:1})\|_F \left\| \sum_{1 \leq j < j' \leq n} W_{n:j'+1} (\Delta W_{j'}) W_{j'-1:j+1} (\Delta W_j) W_{j-1:1} \right\|_F \\
&\leq 2 \|\nabla \phi(W_{n:1})\|_F \sum_{1 \leq j < j' \leq n} \|W_{n:j'+1} (\Delta W_{j'}) W_{j'-1:j+1} (\Delta W_j) W_{j-1:1}\|_F \\
&\leq 2 \|\nabla \phi(W_{n:1})\|_F \sum_{1 \leq j < j' \leq n} \|\Delta W_{j'}\|_F \|\Delta W_j\|_F \cdot \prod_{j'' \in [n] \setminus \{j, j'\}} \|W_{j''}\|_F \\
&\leq 2 \|\nabla \phi(W_{n:1})\|_F \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_F \cdot \sum_{1 \leq j < j' \leq n} \|\Delta W_{j'}\|_F \|\Delta W_j\|_F,
\end{aligned}$$

where the first transition follows from Cauchy-Schwartz inequality, the second and third from sub-additivity and sub-multiplicativity of Frobenius norm respectively. It holds that:

$$\sum_{1 \leq j < j' \leq n} \|\Delta W_{j'}\|_F \|\Delta W_j\|_F \leq \left( \sum_{j=1}^n \|\Delta W_j\|_F \right)^2 \leq n \sum_{j=1}^n \|\Delta W_j\|_F^2,$$

where the last transition follows from the fact that the one-norm of a vector in  $\mathbb{R}^n$  is never greater than  $\sqrt{n}$  times its Euclidean norm. This leads us to:

$$\begin{aligned}
& 2\text{Tr} \left( \nabla \phi(W_{n:1})^\top \sum_{1 \leq j < j' \leq n} W_{n:j'+1} (\Delta W_{j'}) W_{j'-1:j+1} (\Delta W_j) W_{j-1:1} \right) \\
&\leq 2n \|\nabla \phi(W_{n:1})\|_F \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_F \cdot \sum_{j=1}^n \|\Delta W_j\|_F^2.
\end{aligned} \tag{59}$$

Plugging Equations (58) and (59) into Equation (57), we obtain:

$$\begin{aligned}
& \nabla^2 f(\boldsymbol{\theta}) [\Delta W_1, \dots, \Delta W_n] \leq \\
& \left( n \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-1}} \prod_{j \in \mathcal{J}} \|W_j\|_F^2 + 2n \|\nabla \phi(W_{n:1})\|_F \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_F \right) \sum_{j=1}^n \|\Delta W_j\|_F^2.
\end{aligned}$$

This proves property (i).

Moving on to property (ii), we overload notation by allowing the function  $f(\cdot)$  to intake the tuple  $(W_1, W_2, \dots, W_n)$  (in which case  $W_1, \dots, W_n$  are arranged as  $\boldsymbol{\theta}$ , and the value  $f(\boldsymbol{\theta})$  is returned). In Appendix A of [4] it is shown that:

$$\begin{aligned}
& \nabla f(W_1, \dots, W_n) = \\
& \left( (W_{n:2})^\top \nabla \phi(W_{n:1}), \dots, (W_{n:j+1})^\top \nabla \phi(W_{n:1}) (W_{j-1:1})^\top, \dots, \nabla \phi(W_{n:1}) (W_{n-1:1})^\top \right).
\end{aligned}$$

It follows that:

$$\begin{aligned}
\|\nabla f(\boldsymbol{\theta})\|_2^2 &= \|\nabla f(W_1, \dots, W_n)\|_{Frobenius}^2 \\
&= \sum_{j=1}^n \|(W_{n:j+1})^\top \nabla \phi(W_{n:1})(W_{j-1:1})^\top\|_F^2 \\
&\leq \sum_{j=1}^n \|\nabla \phi(W_{n:1})\|_F^2 \prod_{i \in [n] \setminus \{j\}} \|W_i\|_F^2 \\
&\leq n \|\nabla \phi(W_{n:1})\|_F^2 \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-1}} \prod_{j \in \mathcal{J}} \|W_j\|_F^2,
\end{aligned}$$

where the second transition follows from sub-multiplicativity of Frobenius norm. Taking square root of both sides of the inequality concludes the proof of property (ii), and the entire lemma.  $\square$

Combining Lemmas 31 and 32, Lemma 33 below establishes the smoothness and Lipschitz constants  $\beta_{t,\epsilon} = 16n$  and  $\gamma_{t,\epsilon} = 6\sqrt{n}$  respectively.

**Lemma 33.** *It holds that  $\sup_{\mathbf{q} \in \mathcal{D}_{t,\epsilon}} \|\nabla^2 f(\mathbf{q})\|_s \leq 16n$  and  $\sup_{\mathbf{q} \in \mathcal{D}_{t,\epsilon}} \|\nabla f(\mathbf{q})\|_2 \leq 6\sqrt{n}$ .*

*Proof.* Under the conditions and notations of Lemma 31, for any  $\mathcal{J} \subseteq [n]$ :

$$\prod_{j \in \mathcal{J}} \|W_{j,\epsilon}\|_F \leq (\|\mathbf{w}_{n:1}(t')\|_2 + 2n\epsilon)^{\frac{|\mathcal{J}|}{n}}. \quad (60)$$

By Lemma 17 we have that  $\mathbf{w}_{n:1}(\xi(t')) = \mathbf{u}(t')$ , where  $\xi(t') := \int_0^{t'} \|\mathbf{u}(t'')\|_2^{-(1-2/n)} dt''$ .  $\xi(\cdot)$  is unbounded since  $\|\mathbf{u}(\cdot)\|_2 < 1$  by property (iv) of Lemma 19. This implies  $\|\mathbf{w}_{n:1}(t')\|_2 < 1$ , which together with the fact that by definition  $\epsilon \leq 1/2n$ , means:

$$\prod_{j \in \mathcal{J}} \|W_{j,\epsilon}\|_F \leq (1+1)^{\frac{|\mathcal{J}|}{n}} \leq 2. \quad (61)$$

It holds that:

$$\begin{aligned}
&\|\nabla \phi(\mathbf{w}_{n:1,\epsilon})\|_2 \leq \\
\|\nabla \phi(\mathbf{w}_{n:1}(t'))\|_2 + 2n\epsilon &= \|\mathbf{w}_{n:1}(t') - \boldsymbol{\lambda}_{yx}\|_2 + 2n\epsilon \leq \|\mathbf{w}_{n:1}(t')\|_2 + \|\boldsymbol{\lambda}_{yx}\|_2 + 2n\epsilon \leq 3,
\end{aligned} \quad (62)$$

where the first transition follows from Lemma 31, and the last from  $\|\mathbf{w}_{n:1}(t')\|_2 < 1$ ,  $\|\boldsymbol{\lambda}_{yx}\|_2 = 1$  and  $\epsilon \leq 1/2n$ . We conclude the proof by plugging Equations (61) and (62) into the results of Lemma 32, while noticing that arbitrary  $t' \geq 0$  and  $\boldsymbol{\theta}_\epsilon$  account for all  $\mathbf{q} \in \mathcal{D}_{t,\epsilon}$ .  $\square$

Lemma 34 below employs Lemma 2 from our analysis in Section 4, along with Lemma 31 above, for deriving a lower bound on the minimal eigenvalue of the Hessian (of the training loss  $f(\cdot)$ ) in the vicinity of a point along the gradient flow trajectory.

**Lemma 34.** *For all  $t' \geq 0$ :*

$$\inf_{\substack{\mathbf{q} \in \mathbb{R}^d \\ \|\mathbf{q} - \boldsymbol{\theta}(t')\|_2 \leq \epsilon}} \lambda_{\min}(\nabla^2 f(\mathbf{q})) \geq -(n-1) (\|\nabla \phi(\mathbf{w}_{n:1,\epsilon})\|_2 + 2n\epsilon) (\|\mathbf{w}_{n:1}(t')\|_2 + 2n\epsilon)^{1-\frac{2}{n}},$$

where  $\lambda_{\min}(\nabla^2 f(\mathbf{q}))$  stands for the minimal eigenvalue of  $\nabla^2 f(\mathbf{q})$ .

*Proof.* Let  $\boldsymbol{\theta}_\epsilon \in \mathbb{R}^d$  be a weight setting satisfying  $\|\boldsymbol{\theta}_\epsilon - \boldsymbol{\theta}(t')\|_2 \leq \epsilon$ . Denote by  $W_{1,\epsilon} \in \mathbb{R}^{d_1, d_0}, W_{2,\epsilon} \in \mathbb{R}^{d_2, d_1}, \dots, W_{n-1,\epsilon} \in \mathbb{R}^{d_{n-1}, d_{n-2}}, W_{n,\epsilon} \in \mathbb{R}^{1, d_{n-1}}$  the weight matrices constituting  $\boldsymbol{\theta}_\epsilon$ , and by  $\mathbf{w}_{n:1,\epsilon} \in \mathbb{R}^{d_0}$  the corresponding end-to-end matrix  $W_{n,\epsilon} W_{n-1,\epsilon} \cdots W_{1,\epsilon}$  (in vectorized form). Lemma 2 ensures:

$$\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta}_\epsilon)) \geq -(n-1) \|\nabla \phi(\mathbf{w}_{n:1,\epsilon})\|_2 \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_{j,\epsilon}\|_s.$$

We conclude the proof by bounding spectral norms with Frobenius norms, and applying properties (ii) and (iii) from Lemma 31.  $\square$

Lemma 34 implies that, under the notations of Theorem 3, we may choose the function  $m(\cdot)$  to be as follows:

$$m: [0, t] \rightarrow \mathbb{R}, \quad m(t') = (n-1)(\|\nabla\phi(\mathbf{w}_{n:1}(t'))\|_2 + 2n\epsilon)(\|\mathbf{w}_{n:1}(t')\|_2 + 2n\epsilon)^{1-\frac{2}{n}}. \quad (63)$$

Lemma 35 below bounds the integral of this choice of  $m(\cdot)$  in accordance with Equation (18) (recall that the scalar  $\nu$  there, defined in the preceding text, coincides with the value taken by the function  $\nu(\cdot)$  from Definition 3 at zero). For doing so, it makes use of the reparameterized trajectory  $\mathbf{u}(\cdot)$ , and splits the reparameterized integral into two parts corresponding to two time intervals: before exponentially fast convergence is guaranteed to have commenced (*i.e.* until time  $t_a + t_c$  — see Subsubappendix J.8.7), and afterwards.

**Lemma 35.** *With the function  $m(\cdot)$  defined by Equation (63), Equation (18) is satisfied.*

*Proof.* We apply a change of variable using the (continuously differentiable and strictly increasing) function  $\xi(\cdot)$  defined in Lemma 17:

$$\int_0^t m(t') dt' = \int_{\xi^{-1}(0)}^{\xi^{-1}(t)} m(\xi(t')) \frac{d}{dt'} \xi(t') dt'.$$

Notice that  $\xi(0) = 0$  and  $\frac{d}{dt'} \xi(t') = \|\mathbf{u}(t')\|^{-(1-2/n)}$ . Plugging this and the definition of  $m(\cdot)$  (Equation (63)) into the above leads to:

$$\int_0^t m(t') dt' = \int_0^{\xi^{-1}(t)} (n-1) \left( \|\nabla\phi(\mathbf{w}_{n:1}(\xi(t')))\|_2 + 2n\epsilon \right) \left( \|\mathbf{w}_{n:1}(\xi(t'))\|_2 + 2n\epsilon \right)^{1-\frac{2}{n}} \|\mathbf{u}(t')\|^{\frac{2}{n}-1} dt'.$$

Since (by Lemma 17)  $\mathbf{w}_{n:1}(\xi(t)) = \mathbf{u}(t)$ , we have that:

$$\int_0^t m(t') dt' = \int_0^{\xi^{-1}(t)} (n-1) \left( \|\nabla\phi(\mathbf{u}(t'))\|_2 + 2n\epsilon \right) \left( \|\mathbf{u}(t')\|_2 + 2n\epsilon \right)^{1-\frac{2}{n}} \|\mathbf{u}(t')\|^{\frac{2}{n}-1} dt'.$$

Recall the notation  $u_{\min} := \inf_{t \geq 0} \|\mathbf{u}(t)\|_2$  and that, by Lemma 21,  $u_{\min} > 0$ . It holds that:

$$\begin{aligned} \int_0^t m(t') dt' &= \int_0^{\xi^{-1}(t)} (n-1) \left( \|\nabla\phi(\mathbf{u}(t'))\|_2 + 2n\epsilon \right) \left( 1 + 2n\epsilon \|\mathbf{u}(t')\|^{-1} \right)^{1-\frac{2}{n}} dt' \\ &\leq \int_0^{\xi^{-1}(t)} (n-1) \left( \|\nabla\phi(\mathbf{u}(t'))\|_2 + 2n\epsilon \right) \left( 1 + 2n\epsilon \|\mathbf{u}(t')\|^{-1} \right) dt' \\ &\leq \int_0^{\xi^{-1}(t)} (n-1) \left( \|\nabla\phi(\mathbf{u}(t'))\|_2 + 2n\epsilon \right) \left( 1 + 2n\epsilon u_{\min}^{-1} \right) dt' \\ &= (n-1) \left( 1 + 2n\epsilon u_{\min}^{-1} \right) \left( \int_0^{\xi^{-1}(t)} \|\nabla\phi(\mathbf{u}(t'))\|_2 dt' + 2n\epsilon \xi^{-1}(t) \right). \end{aligned}$$

Per Lemma 19 we know that  $\|\mathbf{u}(t')\|_2 < 1$  for all  $t' \geq 0$ . Thus, by the definition of  $\xi(\cdot)$ , for all  $t' \geq 0$  it holds that  $\xi(t') \geq t'$ , which (since  $\xi(\cdot)$  is strictly increasing) implies  $\xi^{-1}(t) \leq t$ . This leads to:

$$\begin{aligned} &\int_0^t m(t') dt' \\ &\leq (n-1) \left( 1 + 2n\epsilon u_{\min}^{-1} \right) \left( \int_0^t \|\nabla\phi(\mathbf{u}(t'))\|_2 dt' + 2n\epsilon t \right) \\ &= (n-1) \int_0^t \|\nabla\phi(\mathbf{u}(t'))\|_2 dt' + (n-1) 2n\epsilon t + (n-1) 2n\epsilon u_{\min}^{-1} \left( \int_0^t \|\nabla\phi(\mathbf{u}(t'))\|_2 dt' + 2n\epsilon t \right) \\ &\leq (n-1) \int_0^t \|\nabla\phi(\mathbf{u}(t'))\|_2 dt' + 2n^2 \epsilon t + 4n^3 \epsilon^2 u_{\min}^{-1} t + 2n^2 \epsilon u_{\min}^{-1} \int_0^t \|\nabla\phi(\mathbf{u}(t'))\|_2 dt'. \end{aligned}$$

It holds that  $\|\nabla\phi(\mathbf{u}(t'))\|_2 = \|\mathbf{u}(t') - \boldsymbol{\lambda}_{yx}\|_2 \leq \|\mathbf{u}(t')\|_2 + \|\boldsymbol{\lambda}_{yx}\|_2 \leq 2$  for all  $t' \geq 0$  (recall that  $\|\boldsymbol{\lambda}_{yx}\|_2 = 1$  by assumption). Thus:

$$\begin{aligned} \int_0^t m(t') dt' &\leq (n-1) \int_0^t \|\nabla\phi(\mathbf{u}(t'))\|_2 dt' + 2n^2 \epsilon t + 4n^3 \epsilon^2 u_{\min}^{-1} t + 2n^2 \epsilon u_{\min}^{-1} \cdot 2t \\ &\leq (n-1) \int_0^t \|\nabla\phi(\mathbf{u}(t'))\|_2 dt' + 3 \cdot \max \left\{ 2n^2 \epsilon t, 4n^3 \epsilon^2 u_{\min}^{-1} t, 4n^2 \epsilon u_{\min}^{-1} t \right\} \\ &\leq (n-1) \int_0^t \|\nabla\phi(\mathbf{u}(t'))\|_2 dt' + 3 \cdot 4n^3 \epsilon u_{\min}^{-1} t. \end{aligned} \quad (64)$$

We may bound the latter integral as follows:

$$\begin{aligned}
\int_0^t \|\nabla\phi(\mathbf{u}(t'))\|_2 dt' &\leq \int_0^\infty \|\nabla\phi(\mathbf{u}(t'))\|_2 dt' \\
&= \int_0^{t_a+t_c} \|\nabla\phi(\mathbf{u}(t'))\|_2 dt' + \int_{t_a+t_c}^\infty \|\nabla\phi(\mathbf{u}(t'))\|_2 dt' \\
&= \int_0^{t_a+t_c} \|\mathbf{u}(t') - \boldsymbol{\lambda}_{yx}\|_2 dt' + \int_0^\infty \|\mathbf{u}(t_a+t_c+t') - \boldsymbol{\lambda}_{yx}\|_2 dt',
\end{aligned}$$

where  $t_a$  and  $t_c$  are given by Definitions 5 and 6 respectively. Notice that  $\|\mathbf{u}(t') - \boldsymbol{\lambda}_{yx}\|_2$  is monotonically non-increasing (since  $\mathbf{u}(\cdot)$  is a monotonic reparameterization of  $\mathbf{w}_{n:1}(\cdot)$ , and gradient flow monotonically non-increases the objective it optimizes). Applying this fact, as well as Lemma 25, we obtain:

$$\begin{aligned}
\int_0^t \|\nabla\phi(\mathbf{u}(t'))\|_2 dt' &\leq \int_0^{t_a+t_c} \|\mathbf{u}(0) - \boldsymbol{\lambda}_{yx}\|_2 dt' + \frac{6}{5} \int_0^\infty \exp\left(-\frac{2n}{3(n+1)}t'\right) dt' \\
&= \|\mathbf{u}(0) - \boldsymbol{\lambda}_{yx}\|_2 (t_a+t_c) + \frac{6}{5} \cdot \frac{3(n+1)}{2n} \\
&\leq \frac{6}{5}(t_a+t_c) + 3,
\end{aligned}$$

where the last transition follows from the assumptions  $\|\mathbf{w}_{n:1}(0)\|_2 \leq 0.2$  and  $\|\boldsymbol{\lambda}_{yx}\|_2 = 1$ . Plug in the definitions of  $t_a$  and  $t_c$  (Definitions 5 and 6 respectively):

$$\begin{aligned}
&\int_0^t \|\nabla\phi(\mathbf{u}(t'))\|_2 dt' \\
&\leq \frac{6}{5} \left( \frac{1}{2} \ln(\max\{5 \cdot \frac{1-\nu(0)}{1+\nu(0)}, 1\}) + \frac{3}{2n} \ln\left(\frac{2n}{3u_{\min}}\right) \right) + 3 \\
&= \frac{3}{5} \ln(\max\{5 \cdot \frac{1-\nu(0)}{1+\nu(0)}, 1\}) + \frac{9}{5n} \ln\left(\frac{2n}{3u_{\min}}\right) + 3 \\
&= \frac{3}{5n} \ln\left(\max\left\{\left\{5 \cdot \frac{1-\nu(0)}{1+\nu(0)}, 1\right\}\right\}^n \cdot \left(\frac{2n}{3u_{\min}}\right)^3 \cdot e^{5n}\right) \\
&\leq \frac{3}{5n} \ln\left(5^n \max\left\{\left\{\frac{1-\nu(0)}{1+\nu(0)}, 1\right\}\right\}^n \cdot \left(\frac{2n}{3}\right)^3 \max\left\{\left\{\frac{3}{2} \cdot \frac{1-\nu(0)}{1+\nu(0)}, 1\right\}\right\}^{3n} \|\mathbf{w}_{n:1}(0)\|_2^{-3} \cdot e^{5n}\right) \\
&\leq \frac{3}{5n} \ln\left(n^3 \|\mathbf{w}_{n:1}(0)\|_2^{-3} e^{8n} \max\left\{\left\{\frac{1-\nu(0)}{1+\nu(0)}, 1\right\}\right\}^{4n}\right),
\end{aligned} \tag{65}$$

where the fourth transition follows from Lemma 21. Plug Equation (65) into Equation (64):

$$\begin{aligned}
&\int_0^t m(t') dt' \\
&\leq \frac{3(n-1)}{5n} \ln\left(n^3 \|\mathbf{w}_{n:1}(0)\|_2^{-3} e^{8n} \max\left\{\left\{\frac{1-\nu(0)}{1+\nu(0)}, 1\right\}\right\}^{4n}\right) + 12n^3 \epsilon u_{\min}^{-1} t \\
&\leq \ln\left(n^2 \|\mathbf{w}_{n:1}(0)\|_2^{-2} e^{5(n-1)} \max\left\{\left\{\frac{1-\nu(0)}{1+\nu(0)}, 1\right\}\right\}^{\frac{5}{2}(n-1)}\right) + 15n^3 u_{\min}^{-1} \epsilon t.
\end{aligned}$$

We conclude the proof with the help of Lemma 21:

$$\int_0^t m(t') dt' \leq \frac{15n^3 \max\left\{\left\{\frac{3}{2} \cdot \frac{1-\nu(0)}{1+\nu(0)}, 1\right\}\right\}^n t \epsilon}{\|\mathbf{w}_{n:1}(0)\|_2} + \ln\left(\frac{n^2 e^{5(n-1)} \max\left\{\left\{\frac{1-\nu(0)}{1+\nu(0)}, 1\right\}\right\}^{\frac{5}{2}(n-1)}}{\|\mathbf{w}_{n:1}(0)\|_2^2}\right).$$

□

### J.8.10 Conclusion

Lemmas 12, 29, 33 and 35, along with the fact that by the definition of  $m(\cdot)$  (Equation (63)) it is non-negative, together form a complete proof for Proposition 3. □

## J.9 Proof of Theorem 4

### J.9.1 Sketch

*Proof sketch (for complete proof see Subappendix J.9).* The proof calls Proposition 3 with  $\bar{\epsilon}$  and  $\epsilon$  small enough such that for any  $t > 0$  and  $\mathbf{q}' \in \mathbb{R}^d$ , if gradient flow at time  $t$  is  $\bar{\epsilon}$ -optimal (meaning  $f(\boldsymbol{\theta}(t)) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \leq \bar{\epsilon}$ ) and is  $\epsilon$ -approximated by  $\mathbf{q}'$  (i.e.  $\|\mathbf{q}' - \boldsymbol{\theta}(t)\|_2 \leq \epsilon$ ), then  $\mathbf{q}'$  is  $\bar{\epsilon}$ -optimal ( $f(\mathbf{q}') - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \leq \bar{\epsilon}$ ). The proposition implies that gradient flow is  $\bar{\epsilon}$ -optimal at the time  $t$  given in Equation (17). Since gradient flow monotonically non-increases  $f(\cdot)$ , it is  $\bar{\epsilon}$ -optimal at any time after  $\bar{t}$

as well. With  $\eta$  and  $k$  adhering to Equations (19) and (20) respectively, we have  $k\eta \geq \bar{t}$ , so it suffices to show that when its step size is  $\eta$ , the first  $k$  iterates of gradient descent  $\epsilon$ -approximate the trajectory of gradient flow up to time  $k\eta$ . This follows directly from delivering to Theorem 3 the geometric results of Proposition 3 (bound on integral of minimal eigenvalue of the Hessian, as well as smoothness and Lipschitz constants) corresponding to  $\mathcal{D}_{k\eta, \epsilon}$  —  $\epsilon$ -neighborhood of gradient flow trajectory up to time  $k\eta$ .  $\square$

### J.9.2 Complete Proof

Let  $\tilde{\epsilon} > 0$ , and consider  $\eta > 0$  and  $k \in \mathbb{N}$  adhering to Equations (19) and (20) respectively. We would like to show that with step size  $\eta$ , iterate  $k$  of gradient descent is  $\tilde{\epsilon}$ -optimal, *i.e.*  $f(\boldsymbol{\theta}_k) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \leq \tilde{\epsilon}$ . Without loss of generality, we may assume  $\tilde{\epsilon} \leq 1$  (a proof that is valid for  $\tilde{\epsilon} = 1$  automatically accounts for  $\tilde{\epsilon} > 1$  as well). Define:

$$\bar{\epsilon} := \tilde{\epsilon}/2, \quad \epsilon := \frac{\|W_{n:1,0}\|_F \bar{\epsilon}}{15n^3 \left(\max\left\{1, \frac{3}{2}, \frac{1-\nu}{1+\nu}\right\}\right)^n k\eta}. \quad (66)$$

Invoking Proposition 3 with initial point  $\boldsymbol{\theta}_s = \boldsymbol{\theta}_0$ , time  $t = k\eta$  and  $\bar{\epsilon}, \epsilon$  as defined above (note that  $\epsilon \in (0, 1/(2n))$ ), we obtain that the gradient flow trajectory emanating from  $\boldsymbol{\theta}_0$  is defined over infinite time, and with  $\boldsymbol{\theta} : [0, \infty) \rightarrow \mathbb{R}^d$  representing this trajectory, the following time  $\bar{t}$  satisfies  $f(\boldsymbol{\theta}(\bar{t})) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \leq \bar{\epsilon}$ :

$$\bar{t} = \frac{2n \left(\max\left\{1, \frac{3}{2}, \frac{1-\nu}{1+\nu}\right\}\right)^n}{\|W_{n:1,0}\|_F} \ln \left( \frac{15n \max\left\{1, \frac{1-\nu}{1+\nu}\right\}}{\|W_{n:1,0}\|_F \min\{1, 2\bar{\epsilon}\}} \right). \quad (67)$$

Moreover, we obtain that under the notations of Theorem 3, in correspondence with  $\mathcal{D}_{k\eta, \epsilon}$  ( $\epsilon$ -neighborhood of gradient flow trajectory up to time  $k\eta$ ) are the smoothness and Lipschitz constants  $\beta_{k\eta, \epsilon} = 16n$  and  $\gamma_{k\eta, \epsilon} = 6\sqrt{n}$  respectively, and the following (upper) bound on the integral of (minus) the minimal eigenvalue of the Hessian:

$$\int_0^{k\eta} m(t') dt' \leq \frac{15n^3 \left(\max\left\{1, \frac{3}{2}, \frac{1-\nu}{1+\nu}\right\}\right)^n k\eta \epsilon}{\|W_{n:1,0}\|_F} + \ln \left( \frac{n^2 \left(e^2 \max\left\{1, \frac{1-\nu}{1+\nu}\right\}\right)^{5(n-1)/2}}{\|W_{n:1,0}\|_F^2} \right), \quad (68)$$

where the function  $m : [0, k\eta] \rightarrow \mathbb{R}$  is non-negative.

Notice that  $k = \lceil \bar{t}/\eta + 1 \rceil$  and therefore  $k\eta \geq \bar{t}$ . Combining this with the fact that the gradient flow trajectory is  $\bar{\epsilon}$ -optimal at time  $\bar{t}$ , and that in general gradient flow monotonically non-increases the objective it optimizes, we infer  $\bar{\epsilon}$ -optimality of the gradient flow trajectory at time  $k\eta$ , *i.e.*  $f(\boldsymbol{\theta}(k\eta)) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \leq \bar{\epsilon}$ . We will invoke Theorem 3 for showing that, in addition to being  $\bar{\epsilon}$ -optimal, the gradient flow trajectory at time  $k\eta$  is also  $\epsilon$ -approximated by iterate  $k$  of gradient descent, *i.e.*  $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}(k\eta)\|_2 \leq \epsilon$ . This, along with  $f(\cdot)$  being  $6\sqrt{n}$ -Lipschitz across  $\mathcal{D}_{k\eta, \epsilon}$  ( $\epsilon$ -neighborhood of gradient flow trajectory up to time  $k\eta$ ), yields the desired result —  $\tilde{\epsilon}$ -optimality for iterate  $k$  of gradient descent:

$$\begin{aligned} & f(\boldsymbol{\theta}_k) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \\ &= \left( f(\boldsymbol{\theta}_k) - f(\boldsymbol{\theta}(k\eta)) \right) + \left( f(\boldsymbol{\theta}(k\eta)) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \right) \\ &\leq \left( 6\sqrt{n} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}(k\eta)\|_2 \right) + \left( f(\boldsymbol{\theta}(k\eta)) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \right) \\ &\leq 6\sqrt{n} \cdot \epsilon + \bar{\epsilon} \\ &\leq \tilde{\epsilon}, \end{aligned}$$

where the last transition follows from the definitions of  $\epsilon$  and  $\bar{\epsilon}$  (Equation (66)).

We conclude the proof by showing that indeed  $\|\theta_k - \theta(k\eta)\|_2 \leq \epsilon$ . Equation (68), the definition of  $\epsilon$  (Equation (66)) and the condition  $\tilde{\epsilon} \leq 1$  together imply:

$$\begin{aligned} \int_0^{k\eta} m(t') dt' &\leq \frac{15n^3 \left(\max\left\{1, \frac{3}{2} \cdot \frac{1-\nu}{1+\nu}\right\}\right)^n k\eta\epsilon}{\|W_{n:1,0}\|_F} + \ln\left(\frac{n^2 \left(e^2 \max\left\{1, \frac{1-\nu}{1+\nu}\right\}\right)^{5(n-1)/2}}{\|W_{n:1,0}\|_F^2}\right) \\ &= \tilde{\epsilon} + \ln\left(\frac{n^2 \left(e^2 \max\left\{1, \frac{1-\nu}{1+\nu}\right\}\right)^{5(n-1)/2}}{\|W_{n:1,0}\|_F^2}\right) \\ &\leq 1 + \ln\left(\frac{n^2 \left(e^2 \max\left\{1, \frac{1-\nu}{1+\nu}\right\}\right)^{5(n-1)/2}}{\|W_{n:1,0}\|_F^2}\right) \\ &< \ln\left(\frac{3n^2 \left(e^2 \max\left\{1, \frac{1-\nu}{1+\nu}\right\}\right)^{5(n-1)/2}}{\|W_{n:1,0}\|_F^2}\right). \end{aligned} \quad (69)$$

Recalling the fact that  $k = \lfloor \bar{t}/\eta + 1 \rfloor$ , the expression for  $\bar{t}$  (Equation (67)), and the definition of  $\bar{\epsilon}$  (Equation (66)), we have:

$$\begin{aligned} k\eta = \lfloor \bar{t}/\eta + 1 \rfloor \eta &\leq \bar{t} + \eta = \frac{2n \left(\max\left\{1, \frac{3}{2} \cdot \frac{1-\nu}{1+\nu}\right\}\right)^n}{\|W_{n:1,0}\|_F} \ln\left(\frac{15n \max\left\{1, \frac{1-\nu}{1+\nu}\right\}}{\|W_{n:1,0}\|_F \bar{\epsilon}}\right) + \eta \\ &< \frac{3n \left(\max\left\{1, \frac{3}{2} \cdot \frac{1-\nu}{1+\nu}\right\}\right)^n}{\|W_{n:1,0}\|_F} \ln\left(\frac{15n \max\left\{1, \frac{1-\nu}{1+\nu}\right\}}{\|W_{n:1,0}\|_F \bar{\epsilon}}\right), \end{aligned} \quad (70)$$

where the last transition makes use of the upper bound on  $\eta$  given in Equation (19). It holds that:

$$\begin{aligned} &\epsilon^{-1} \beta_{k\eta, \epsilon} \gamma_{k\eta, \epsilon} k\eta e^{\int_0^{k\eta} m(t') dt'} \\ &< \frac{15n^3 \left(\max\left\{1, \frac{3}{2} \cdot \frac{1-\nu}{1+\nu}\right\}\right)^n k\eta}{\|W_{n:1,0}\|_F \bar{\epsilon}} \cdot 16n \cdot 6\sqrt{n} \cdot k\eta \cdot \frac{3n^2 \left(e^2 \max\left\{1, \frac{1-\nu}{1+\nu}\right\}\right)^{5(n-1)/2}}{\|W_{n:1,0}\|_F^2} \\ &< \frac{4500n^{13/2} e^{6n-5} \left(\max\left\{1, \frac{1-\nu}{1+\nu}\right\}\right)^{(7n-5)/2}}{\|W_{n:1,0}\|_F^3 \bar{\epsilon}} (k\eta)^2 \\ &< \frac{4500n^{13/2} e^{6n-5} \left(\max\left\{1, \frac{1-\nu}{1+\nu}\right\}\right)^{(7n-5)/2}}{\|W_{n:1,0}\|_F^3 \bar{\epsilon}} \cdot \frac{9n^2 \left(\max\left\{1, \frac{3}{2} \cdot \frac{1-\nu}{1+\nu}\right\}\right)^{2n}}{\|W_{n:1,0}\|_F^2} \left(\ln\left(\frac{15n \max\left\{1, \frac{1-\nu}{1+\nu}\right\}}{\|W_{n:1,0}\|_F \bar{\epsilon}}\right)\right)^2 \\ &< \frac{n^{17/2} e^{7n+6} \left(\max\left\{1, \frac{1-\nu}{1+\nu}\right\}\right)^{(11n-5)/2}}{\|W_{n:1,0}\|_F^5 \bar{\epsilon}} \left(\ln\left(\frac{15n \max\left\{1, \frac{1-\nu}{1+\nu}\right\}}{\|W_{n:1,0}\|_F \bar{\epsilon}}\right)\right)^2 \\ &\leq 1/\eta, \end{aligned}$$

where the first transition follows from Equation (69) and the definition of  $\epsilon$  (Equation (66)); the third makes use of Equation (70); and the last is due to the upper bound on  $\eta$  given in Equation (19). Rearrange the derived inequality:

$$\eta < \frac{\epsilon}{\beta_{k\eta, \epsilon} \gamma_{k\eta, \epsilon} k\eta e^{\int_0^{k\eta} m(t') dt'}}.$$

Since  $m(\cdot)$  is non-negative, it holds that:

$$\frac{\epsilon}{\beta_{k\eta, \epsilon} \gamma_{k\eta, \epsilon} k\eta e^{\int_0^{k\eta} m(t') dt'}} \leq \inf_{t \in (0, k\eta)} \frac{\epsilon}{\beta_{k\eta, \epsilon} \gamma_{k\eta, \epsilon} \int_0^t e^{\int_{t'}^t m(t'') dt''} dt'},$$

and therefore:

$$\eta < \inf_{t \in (0, k\eta)} \frac{\epsilon}{\beta_{k\eta, \epsilon} \gamma_{k\eta, \epsilon} \int_0^t e^{\int_{t'}^t m(t'') dt''} dt'}. \quad (71)$$

We now invoke Theorem 3 with  $\epsilon$  as we have defined (Equation (66)), time  $\tilde{t} = k\eta$ , and  $\beta_{k\eta, \epsilon}, \gamma_{k\eta, \epsilon}$  and  $m(\cdot)$  as produced by Proposition 3. Recalling that in our context gradient flow and gradient descent are initialized identically, *i.e.*  $\theta(0) = \theta_0$ , we conclude from Equation (71) that the first  $\lfloor k\eta/\eta \rfloor = k$  iterates of gradient descent  $\epsilon$ -approximate the gradient flow trajectory up to time  $k\eta$ , *i.e.*  $\|\theta_{k'} - \theta(k'\eta)\|_2 \leq \epsilon$  for all  $k' \in \{1, 2, \dots, k\}$ . In particular  $\|\theta_k - \theta(k\eta)\|_2 \leq \epsilon$ , as required.  $\square$

## J.10 Proof of Corollary 2

It suffices to show that the conditions of Theorem 4 are almost surely satisfied. Initialization is balanced by construction, and since  $W_{n:1,0}$  (initial end-to-end matrix) follows the distribution  $\mathcal{P}$ , it almost surely has Frobenius norm no greater than 0.2. Moreover, since  $\mathcal{P}$  is continuous, and the line in  $\mathbb{R}^{1,d_0}$  passing through the origin and  $\Lambda_{yx}$  has (Lebesgue) measure zero,  $W_{n:1,0}$  is almost surely not equal to zero and not antiparallel to  $\Lambda_{yx}$ . This completes the proof.  $\square$

## J.11 Proof of Proposition 4

The proof is organized as follows. Subsubappendix J.11.1 establishes preliminaries. Subsubappendixes J.11.2, J.11.3 and J.11.4 respectively analyze the trajectories of gradient flow and gradient descent in three different regions of the objective function: (i) “anisotropic” region where curvatures in first and second coordinates differ; (ii) transition region between the previous and the next; and (iii) “isotropic” region where curvatures in first and second coordinates are identical. Subsubappendix J.11.5 shows that the location of gradient flow at time  $\tilde{t}$  is not  $\epsilon$ -approximated by different portions of the gradient descent trajectory. Finally, Subsubappendix J.11.6 concludes.

### J.11.1 Preliminaries

Consider an arbitrary time

$$\tilde{t} \in \left[ \frac{2}{a} \ln \left( \frac{2-3\bar{\rho}/2}{\bar{\theta}_{s,2} - (\bar{\rho}/2 - 1)} \right) + \frac{1}{a} \ln \left( \frac{2}{1-\bar{\rho}} \right), \frac{2}{a} \ln \left( \frac{2-3\bar{\rho}/2}{\bar{\theta}_{s,2} - (\bar{\rho}/2 - 1)} \right) + \frac{1}{a} \ln \left( \frac{1+\bar{\rho}/4}{1-3\bar{\rho}/4} \right) + \frac{1}{a} \ln(b) \right],$$

and suppose the step size  $\eta$  is greater than or equal to  $\frac{10^{14}}{a} e^{-a\tilde{t}} \epsilon$ . We aim to prove  $\|\theta_k - \theta(\tilde{t})\|_2 > \epsilon$  for all  $k \in \mathbb{N} \cup \{0\}$ .

Since the objective function  $f(\cdot)$  (defined in Equation (23)) is additively separable (can be expressed as a sum of terms, each depending on a single input variable), the dynamics in  $\mathbb{R}^d$  induced by gradient flow and gradient descent can be analyzed separately for different coordinates. Lemma 36 below analyzes the dynamics in the third coordinate, establishing the sought after result for the case where  $\eta$  is greater than  $\frac{1}{6a}$ .

**Lemma 36.** *Assume  $\eta > \frac{1}{6a}$ . Then  $\|\theta_k - \theta(\tilde{t})\|_2 > \epsilon$  for all  $k \in \mathbb{N} \cup \{0\}$ .*

*Proof.* Denote by  $\hat{\theta}(\cdot)$  the third coordinate of the gradient flow trajectory  $\theta(\cdot)$ . Similarly, for any  $k \in \mathbb{N} \cup \{0\}$ , denote by  $\hat{\theta}_k$  the third coordinate of the gradient descent iterate  $\theta_k$ . For any  $k \in \mathbb{N} \cup \{0\}$ , it holds that:

$$|\hat{\theta}_{k+1}| = |\hat{\theta}_k - \eta \frac{\partial f}{\partial q_3}(\theta_k)| = |\hat{\theta}_k - 12a\eta \hat{\theta}_k| = |\hat{\theta}_k| \cdot |1 - 12a\eta| > |\hat{\theta}_k|.$$

Thus, we may conclude  $|\hat{\theta}_k| > |\hat{\theta}_0|$  for any  $k \in \mathbb{N}$ . The solution to the gradient flow equation of the third coordinate (i.e.  $\frac{d}{dt} \hat{\theta}(t) = -12a\hat{\theta}(t)$ ) is  $\hat{\theta}(t) = \hat{\theta}(0)e^{-12at}$ . Recall that  $\hat{\theta}(0) = \hat{\theta}_0 > 2$  and notice that  $\tilde{t} \geq \frac{\ln(2)}{12a}$ . For any  $k \in \mathbb{N} \cup \{0\}$  we have:

$$\|\theta(\tilde{t}) - \theta_k\|_2 \geq |\hat{\theta}(\tilde{t}) - \hat{\theta}_k| \geq |\hat{\theta}_k| - |\hat{\theta}(\tilde{t})| \geq |\hat{\theta}_0| - \left| \hat{\theta} \left( \frac{\ln(2)}{12a} \right) \right| = |\hat{\theta}_0| - \frac{1}{2} |\hat{\theta}_0| > 1 > \epsilon.$$

$\square$

It remains to treat the case where  $\eta$  is no greater than  $\frac{1}{6a}$ . In the remainder of the proof we restrict our attention to this case, i.e. we assume  $\eta \in \left[ \frac{10^{14}}{a} e^{-a\tilde{t}} \epsilon, \frac{1}{6a} \right]$ . Special focus will be devoted to the dynamics in the first two coordinates. Denote by  $\theta(\cdot)$  and  $\bar{\theta}(\cdot)$  the first and second coordinates, respectively, of the gradient flow trajectory  $\theta(\cdot)$ . Similarly, for  $k \in \mathbb{N} \cup \{0\}$ , denote by  $\theta_k$  and  $\bar{\theta}_k$  the first and second coordinates, respectively, of the gradient descent iterate  $\theta_k$ . The following lemma shows that in the first two coordinates, the trajectories of gradient flow and gradient descent are monotonically non-decreasing.

**Lemma 37.** *The functions  $\theta(\cdot)$  and  $\bar{\theta}(\cdot)$ , and the series  $(\theta_k)_{k=0}^\infty$  and  $(\bar{\theta}_k)_{k=0}^\infty$ , are all monotonically non-decreasing.*

*Proof.* The results follows from the fact that the derivative of  $\varphi(\cdot)$  over  $[0, \infty)$ , and that of  $\bar{\varphi}(\cdot)$  over  $[\frac{\bar{\rho}}{2} - 1, \infty)$ , are both non-positive.  $\square$

With Lemma 37 at hand, we consider three regions (in  $\mathbb{R}^d$ ) which may be traversed by the trajectories of gradient flow and gradient descent: (i) “anisotropic” region  $[0, z_c) \times [\bar{\rho}/2 - 1, 1 - \bar{\rho}) \times \mathbb{R}^{d-2}$ , where the curvatures of  $f(\cdot)$  in the first and second coordinates differ (namely, they equal  $-a$  and  $-a/2$  respectively); (ii) transition region  $[0, z_c) \times [1 - \bar{\rho}, 1) \times \mathbb{R}^{d-2}$ ; and (iii) “isotropic” region  $[0, z_c) \times [1, \bar{z}_c) \times \mathbb{R}^{d-2}$ , where the curvatures of  $f(\cdot)$  in the first and second coordinates are identical (namely, they both equal  $-a$ ). As we now show, throughout the above regions, the trajectories of gradient flow and gradient descent admit simple characterizations for their first coordinate.

**Lemma 38.** *It holds that  $\theta(t) = \theta(0)e^{at}$  for all  $t \in [0, a^{-1} \ln(z_c/\theta(0))]$ , and  $\theta_k = \theta_0(1+a\eta)^k$  for all  $k \in \{0, 1, \dots, \lceil \ln(z_c/\theta_0)/\ln(1+a\eta) \rceil\}$ .*

*Proof.* Notice that  $\theta(0) \in (0, z_c)$ . For any  $t \in [0, \infty)$  such that  $\theta(t) \in (0, z_c)$ , we obtain:

$$\frac{d\theta}{dt}(t) = -\frac{d\varphi}{dz}(\theta(t)) = a\theta(t).$$

The function  $t \mapsto \theta(0)e^{at}$  is a solution to this initial value problem valid through  $t \in (0, \ln(\frac{z_c}{\theta_0})/a)$ , and from uniqueness of the solution together with continuity of  $\theta(\cdot)$ , we conclude that  $\theta(t) = \theta(0)e^{at}$  for  $t \in [0, \ln(\frac{z_c}{\theta_0})/a]$ .

Moving on to gradient descent. Notice that  $\theta_0 \in (0, z_c)$ , and for any  $k \in \mathbb{N}$  such that  $\theta_{k-1} \in (0, z_c)$  we have:

$$\theta_k = \theta_{k-1} - \eta \frac{d\varphi}{dz}(\theta_{k-1}) = \theta_{k-1} + a\eta\theta_{k-1} = \theta_{k-1}(1+a\eta).$$

It follows that  $\theta_k = \theta_0(1+a\eta)^k$  for any  $k \in \{0, 1, \dots, \lceil \ln(\frac{z_c}{\theta_0})/\ln(1+a\eta) \rceil\}$ , where by plugging in  $k = \lceil \ln(\frac{z_c}{\theta_0})/\ln(1+a\eta) \rceil$  we obtain  $\theta_{k-1} < z_c$ .  $\square$

Compared to the first coordinate, in the second coordinate the trajectories of gradient flow and gradient descent are more involved — analyses for the anisotropic, transition and isotropic regions are conducted in Subsubappendixes J.11.2, J.11.3 and J.11.4 respectively.

## J.11.2 Anisotropic Region

The current subsubappendix analyzes the second coordinate of the gradient flow and gradient descent trajectories throughout the anisotropic region, or more specifically, when the second coordinate is in the range  $[\bar{\rho}/2 - 1, 1 - \bar{\rho})$ . Beginning with gradient flow, we recall that (by Lemma 37) the second coordinate of the trajectory is monotonically non-decreasing, and consider the time at which it exits the range  $[\bar{\rho}/2 - 1, 1 - \bar{\rho})$ .

**Definition 8.** Define  $t_{1-\bar{\rho}} := \inf\{t \geq 0 : \bar{\theta}(t) \geq 1 - \bar{\rho}\}$ .<sup>37</sup>

Lemma 39 below provides an explicit expression for  $t_{1-\bar{\rho}}$ , and for the second coordinate of the gradient flow trajectory until this time.

**Lemma 39.** *The following hold:*

- (i)  $t_{1-\bar{\rho}} = \frac{2}{a} \ln((4-3\bar{\rho})/(2\bar{\theta}(0)+2-\bar{\rho}))$ ; and
- (ii)  $\bar{\theta}(t) = (\bar{\theta}(0) - (\bar{\rho}/2 - 1))e^{at/2} + (\bar{\rho}/2 - 1)$  for all  $t \in [0, t_{1-\bar{\rho}}]$ .

*Proof.* Notice that  $\bar{\theta}(0) \in (\frac{\bar{\rho}}{2} - 1, 1 - \bar{\rho})$ . For any  $t \in [0, \infty)$  such that  $\bar{\theta}(t) \in (\frac{\bar{\rho}}{2} - 1, 1 - \bar{\rho})$ , we obtain:

$$\frac{d\bar{\theta}}{dt}(t) = -\frac{d\bar{\varphi}}{dz}(\bar{\theta}(t)) = \frac{a}{2}(\bar{\theta}(t) - (\frac{\bar{\rho}}{2} - 1)).$$

The function  $t \mapsto (\bar{\theta}(0) - (\frac{1}{2}\bar{\rho} - 1))e^{at/2} + (\frac{1}{2}\bar{\rho} - 1)$  is a solution to this initial value problem valid through  $t \in (0, \frac{2}{a} \ln((4-3\bar{\rho})/(2\bar{\theta}(0)+2-\bar{\rho})))$ , and from uniqueness of the solution together with continuity of  $\bar{\theta}(\cdot)$ , we conclude that  $\bar{\theta}(t) = (\bar{\theta}(0) - (\frac{1}{2}\bar{\rho} - 1))e^{at/2} + (\frac{1}{2}\bar{\rho} - 1)$  for  $t \in [0, \frac{2}{a} \ln((4-3\bar{\rho})/(2\bar{\theta}(0)+2-\bar{\rho}))]$ . This, along with the definition of  $t_{1-\bar{\rho}}$ , implies that  $t_{1-\bar{\rho}} = \frac{2}{a} \ln((4-3\bar{\rho})/(2\bar{\theta}(0)+2-\bar{\rho}))$ .  $\square$

<sup>37</sup>Note that by convention, the infimum of the empty set is equal to infinity.



Moving on to gradient descent, we provide a treatment analogous to that of gradient flow. Namely, we recall that (by Lemma 37) the second coordinate of the trajectory is monotonically non-decreasing, consider the iteration at which it exits the range  $[\bar{\rho}/2 - 1, 1 - \bar{\rho})$ , and present an explicit expression for the index of this iteration as well as the second coordinate of the gradient descent trajectory until the iteration is reached.

**Definition 9.** Define  $k_{1-\bar{\rho}} := \inf \{k \in \mathbb{N} \cup \{0\} : \bar{\theta}_k \geq 1 - \bar{\rho}\}$ .<sup>37</sup>

**Lemma 40.** *The following hold:*

- (i)  $k_{1-\bar{\rho}} = \lceil (\ln(2 - 3\bar{\rho}/2) - \ln(\bar{\theta}_0 + 1 - \bar{\rho}/2)) / \ln(1 + a\eta/2) \rceil$ ; and
- (ii)  $\bar{\theta}_k = (\bar{\theta}_0 - (\bar{\rho}/2 - 1))(1 + a\eta/2)^k + (\bar{\rho}/2 - 1)$  for all  $k \in \{0, 1, \dots, k_{1-\bar{\rho}}\}$ .

*Proof.* Notice that  $\bar{\theta}_0 \in (\frac{\bar{\rho}}{2} - 1, 1 - \bar{\rho})$ . For any  $k \in \mathbb{N}$  such that  $\bar{\theta}_{k-1} \in (\frac{\bar{\rho}}{2} - 1, 1 - \bar{\rho})$ , we obtain:

$$\bar{\theta}_k = \bar{\theta}_{k-1} - \eta \frac{d\bar{\varphi}}{dz}(\bar{\theta}_{k-1}) = \bar{\theta}_{k-1} + \frac{a}{2}\eta(\bar{\theta}_{k-1} - (\frac{\bar{\rho}}{2} - 1)).$$

Subtract  $(\frac{\bar{\rho}}{2} - 1)$  from both sides of the equation:

$$(\bar{\theta}_k - (\frac{\bar{\rho}}{2} - 1)) = (\bar{\theta}_{k-1} - (\frac{\bar{\rho}}{2} - 1)) + \frac{a}{2}\eta(\bar{\theta}_{k-1} - (\frac{\bar{\rho}}{2} - 1)).$$

This leads us to:

$$(\bar{\theta}_k - (\frac{\bar{\rho}}{2} - 1)) = (\bar{\theta}_{k-1} - (\frac{\bar{\rho}}{2} - 1))(1 + \frac{a}{2}\eta).$$

It follows that  $(\bar{\theta}_k - (\frac{\bar{\rho}}{2} - 1)) = (\bar{\theta}_0 - (\frac{\bar{\rho}}{2} - 1))(1 + \frac{a}{2}\eta)^k$  for any  $k \in \{0, 1, \dots, k_{1-\bar{\rho}}\}$ . From the definition of  $k_{1-\bar{\rho}}$  it must be equal to  $\lceil (\ln(2 - 3\bar{\rho}/2) - \ln(\bar{\theta}_0 + 1 - \bar{\rho}/2)) / \ln(1 + a\eta/2) \rceil$ , as if it is smaller we get  $\bar{\theta}_{k_{1-\bar{\rho}}} < 1 - \bar{\rho}$ , and if it is larger we get  $\bar{\theta}_{k_{1-\bar{\rho}}-1} \geq 1 - \bar{\rho}$ , both contradicting the definition of  $k_{1-\bar{\rho}}$ .  $\square$

We conclude this subsubappendix by combining its results with Lemma 38, thereby showing that the gradient flow trajectory between initialization and time  $t_{1-\bar{\rho}}$ , and the gradient descent trajectory between initialization and iteration  $k_{1-\bar{\rho}}$ , both lie in the anisotropic region.

**Lemma 41.** *It holds that  $(\theta(t), \bar{\theta}(t)) \in [0, z_c) \times [\bar{\rho}/2 - 1, 1 - \bar{\rho})$  for all  $t \in [0, t_{1-\bar{\rho}})$ , and  $(\theta_k, \bar{\theta}_k) \in [0, z_c) \times [\bar{\rho}/2 - 1, 1 - \bar{\rho})$  for all  $k \in \{0, 1, \dots, k_{1-\bar{\rho}} - 1\}$ .*

*Proof.* We start by proving the result for gradient flow. By assumption it holds that  $\bar{\theta}(0) \in [\bar{\rho}/2 - 1, 1 - \bar{\rho})$ . From monotonicity of  $\bar{\theta}(\cdot)$  (Lemma 37) together with the definition of  $t_{1-\bar{\rho}}$ , we have that  $\bar{\theta}(t) \in [\bar{\rho}/2 - 1, 1 - \bar{\rho})$  for all  $t \in [0, t_{1-\bar{\rho}})$ . Recall that we assume  $\theta(0) \in (0.5, 1)$ . By Lemma 38 we have that  $\theta(t) \in [\theta(0), z_c)$  for all  $t \in [0, \frac{1}{a} \ln(z_c/\theta(0))]$ . By Lemma 39  $t_{1-\bar{\rho}} = \frac{2}{a} \ln((4 - 3\bar{\rho}) / (2\bar{\theta}(0) + 2 - \bar{\rho}))$ . Since  $\frac{2}{a} \ln((4 - 3\bar{\rho}) / (2\bar{\theta}(0) + 2 - \bar{\rho})) \leq \frac{1}{a} \ln(z_c/\theta(0))$  (can be verified by recalling the assumptions on  $a, z_c, \bar{\rho}, \theta(0)$  and  $\bar{\theta}(0)$ ), it follows that  $\theta(t) \in [0, z_c)$  for all  $t \in [0, t_{1-\bar{\rho}})$ . In conclusion, we have shown that  $(\theta(t), \bar{\theta}(t)) \in [0, z_c) \times [\bar{\rho}/2 - 1, 1 - \bar{\rho})$  for all  $t \in [0, t_{1-\bar{\rho}})$ .

Moving on to gradient descent, by assumption it holds that  $\bar{\theta}_0 \in [\bar{\rho}/2 - 1, 1 - \bar{\rho})$ . From monotonicity of  $(\bar{\theta}_k)_{k=0}^\infty$  (Lemma 37) together with the definition of  $k_{1-\bar{\rho}}$ , we have that  $\bar{\theta}_k \in [\bar{\rho}/2 - 1, 1 - \bar{\rho})$  for all  $k \in \{0, 1, \dots, k_{1-\bar{\rho}} - 1\}$ . Recall that we assume  $\theta_0 \in (0.5, 1)$ . By Lemma 38 we have that  $\theta_k \in [\theta(0), z_c)$  for all  $k \in \{0, 1, \dots, \lceil \ln(z_c/\theta_0) / \ln(1 + a\eta) \rceil - 1\}$ . By Lemma 40  $k_{1-\bar{\rho}} = \lceil (\ln(2 - 3\bar{\rho}/2) - \ln(\bar{\theta}_0 + 1 - \bar{\rho}/2)) / \ln(1 + a\eta/2) \rceil$ . Since  $k_{1-\bar{\rho}} \leq \lceil \ln(z_c/\theta_0) / \ln(1 + a\eta) \rceil$  (can be verified by recalling the assumptions on  $a, z_c, \bar{\rho}, \theta_0$  and  $\bar{\theta}_0$ ), it follows that  $\theta_k \in [0, z_c)$  for all  $k \in \{0, 1, \dots, k_{1-\bar{\rho}} - 1\}$ . In conclusion, we have shown that  $(\theta_k, \bar{\theta}_k) \in [0, z_c) \times [\bar{\rho}/2 - 1, 1 - \bar{\rho})$  for all  $k \in \{0, 1, \dots, k_{1-\bar{\rho}} - 1\}$ .  $\square$

### J.11.3 Transition Region

The current subsubappendix analyzes the second coordinate of the gradient flow and gradient descent trajectories throughout the transition region, or more specifically, when the second coordinate is in the range  $[1 - \bar{\rho}, 1)$ . Beginning with gradient flow, we recall that (by Lemma 37) the second coordinate of the trajectory is monotonically non-decreasing, and consider the time at which it exits the range  $[1 - \bar{\rho}, 1)$ .

**Definition 10.** Define  $t_1 := \inf \{t \geq 0 : \bar{\theta}(t) \geq 1\}$ .<sup>37</sup>

Using  $t_{1-\bar{\rho}}$  from Definition 8, Lemma 42 below provides lower and upper bounds for  $t_1$ , and an upper bound for the ratio between the second coordinate of the gradient flow trajectory at time  $t_1$ , and its first coordinate at the same time.

**Lemma 42.** *The following hold:*

- (i)  $t_{1-\bar{\rho}} + a^{-1} \ln((4+\bar{\rho})/(4-3\bar{\rho})) \leq t_1 \leq t_{1-\bar{\rho}} + a^{-1} \ln(1/(1-\bar{\rho}))$ ; and
- (ii)  $\bar{\theta}(t_1)/\theta(t_1) \leq ((\bar{\theta}(0)+1-\bar{\rho}/2)/(2-3\bar{\rho}/2))^2/\theta(0)$ .

*Proof.* We start by proving property (i). Lemma 39 implies  $\bar{\theta}(t_{1-\bar{\rho}}) = 1 - \bar{\rho}$ . Recall that by Lemma 37,  $\bar{\theta}(\cdot)$  is monotonically non-decreasing. For any  $t \in [0, \infty)$  such that  $\bar{\theta}(t) \in [1 - \bar{\rho}, 1]$ , we have:

$$\frac{d\bar{\theta}}{dt}(t) = -\frac{d\bar{\varphi}}{dz}(\bar{\theta}(t)) = a\bar{\theta}(t) + \frac{a}{4\bar{\rho}}(\bar{\theta}(t) - 1)^2. \quad (72)$$

By lower bounding Equation (72) we get  $\frac{d\bar{\theta}}{dt}(t) \geq a\bar{\theta}(t)$  (which implies  $t_1 < \infty$ ). Dividing both sides of this inequality by  $\bar{\theta}(t)$  and integrating over time from  $t_{1-\bar{\rho}}$  until  $t_1$  we have that  $\bar{\theta}(t_1) \geq (1-\bar{\rho})e^{a(t_1-t_{1-\bar{\rho}})}$ . From continuity of  $\bar{\theta}(\cdot)$ , the definition of  $t_1$  and the fact that  $t_1 < \infty$ , we have that  $\theta(t_1) = 1$ . This implies  $(1-\bar{\rho})e^{a(t_1-t_{1-\bar{\rho}})} \leq 1$ . We may conclude  $t_1 \leq t_{1-\bar{\rho}} + a^{-1} \ln(1/(1-\bar{\rho}))$ . We now turn to upper bound Equation (72). For any  $t \in [0, \infty)$  such that  $\bar{\theta}(t) \in [1 - \bar{\rho}, 1]$ :

$$\frac{d\bar{\theta}}{dt}(t) \leq a\bar{\theta}(t) + \frac{a}{4\bar{\rho}}(\bar{\rho})^2 = a\bar{\theta}(t) + \frac{a\bar{\rho}}{4}.$$

Dividing both sides of this inequality by  $\bar{\theta}(t)$  and integrating over time from  $t_{1-\bar{\rho}}$  until  $t_1$ , we have that  $\bar{\theta}(t_1) \leq (1 - \frac{3}{4}\bar{\rho})e^{a(t_1-t_{1-\bar{\rho}}) - \frac{\bar{\rho}}{4}}$ . Since  $\theta(t_1) = 1$ , this implies  $(1 - \frac{3}{4}\bar{\rho})e^{a(t_1-t_{1-\bar{\rho}}) - \frac{\bar{\rho}}{4}} \geq 1$ . We may conclude  $t_1 \geq t_{1-\bar{\rho}} + a^{-1} \ln((4+\bar{\rho})/(4-3\bar{\rho}))$ .

Moving on to property (ii), by property (i) we know that  $t_1 < \infty$ . Recall that  $\bar{\theta}(t_1) = 1$ . Lemma 39 showed that  $t_{1-\bar{\rho}} = \frac{2}{a} \ln((4-3\bar{\rho})/(2\bar{\theta}(0) + 2 - \bar{\rho}))$ . Lemma 38 ensures  $\theta(t) = \theta_0 e^{at}$  for  $t \in [0, \ln(\frac{z_c}{\bar{\theta}(0)})/a]$ . Notice that  $t_1 \leq t_{1-\bar{\rho}} + a^{-1} \ln(1/(1-\bar{\rho})) \leq \ln(\frac{z_c}{\bar{\theta}(0)})/a$ , where the first inequality follows from property (i). It holds that:

$$\bar{\theta}(t_1)/\theta(t_1) = 1/(\theta(0)e^{at_1}) \leq 1/(\theta(0)e^{at_{1-\bar{\rho}}}) \leq \left(\frac{2\bar{\theta}(0)+2-\bar{\rho}}{4-3\bar{\rho}}\right)^2/\theta(0). \quad \square$$

Moving on to gradient descent, we recall that here too the second coordinate of the trajectory is monotonically non-decreasing (see Lemma 37), and consider the iteration at which this second coordinate exits the range  $[1 - \bar{\rho}, 1]$ .

**Definition 11.** Define  $k_1 := \inf\{k \in \mathbb{N} \cup \{0\} : \bar{\theta}_k \geq 1\}$ .<sup>37</sup>

Using  $k_{1-\bar{\rho}}$  from Definition 9, Lemma 43 below provides an upper bound for  $k_1$ , and a lower bound for the ratio between the second coordinate of the gradient descent trajectory at iteration  $k_1$ , and its first coordinate at the same iteration.

**Lemma 43.** *The following hold:*

- (i)  $k_1 \leq k_{1-\bar{\rho}} + \lceil \max\{0, -\ln(\bar{\theta}_{k_{1-\bar{\rho}}})/\ln(1+a\eta)\} \rceil$ ; and
- (ii)  $\bar{\theta}_{k_1}/\theta_{k_1} \geq ((\bar{\theta}_0+1-\bar{\rho}/2)/(2-3\bar{\rho}/2))^2/(\theta_0(1-a\eta/10))$ .

*Proof.* We start by proving property (i). By the definition of  $\bar{\theta}_{k_{1-\bar{\rho}}}$  (and from the fact that it is finite from Lemma 40), we know that  $\bar{\theta}_{k_{1-\bar{\rho}}} \geq 1 - \bar{\rho}$ . Recall that by Lemma 37  $(\bar{\theta}_k)_{k=0}^\infty$  is monotonically non-decreasing. Notice that it is possible for  $k_1$  to be equal to  $k_{1-\bar{\rho}}$ ; this will be the case if  $\bar{\theta}_{k_{1-\bar{\rho}}} \geq 1$ . For any  $k \in \mathbb{N}$  such that  $\bar{\theta}_{k-1} \in [1 - \bar{\rho}, 1]$ , we obtain:

$$\bar{\theta}_k = \bar{\theta}_{k-1} - \eta \frac{d\bar{\varphi}}{dz}(\bar{\theta}_{k-1}) = \bar{\theta}_{k-1} + \eta \left( a\bar{\theta}_{k-1} + \frac{a}{4\bar{\rho}}(\bar{\theta}_{k-1} - 1)^2 \right) \geq \bar{\theta}_{k-1} + a\eta\bar{\theta}_{k-1} = \bar{\theta}_{k-1}(1+a\eta).$$

It follows that  $\bar{\theta}_k \geq \bar{\theta}_{k_{1-\bar{\rho}}}(1+a\eta)^{k-k_{1-\bar{\rho}}}$  for any  $k \in \{k_{1-\bar{\rho}}, k_{1-\bar{\rho}} + 1, \dots, k_1\}$ . Plugging  $k = k_{1-\bar{\rho}} + \lceil \max\{0, -\ln(\bar{\theta}_{k_{1-\bar{\rho}}})/\ln(1+a\eta)\} \rceil$  yields  $\bar{\theta}_{k_1} \geq \bar{\theta}_{k_{1-\bar{\rho}}}(1+a\eta)^{k-k_{1-\bar{\rho}}} \geq 1$ . From monotonicity

of  $(\bar{\theta}_k)_{k=0}^\infty$ , we may conclude that  $k_1 \leq k_{1-\bar{\rho}} + \lceil \max\{0, -\ln(\bar{\theta}_{k_{1-\bar{\rho}}})/\ln(1+a\eta)\} \rceil$ , thereby finishing the proof of property (i).

Moving on to property (ii), with the help of Lemma 38, Lemma 40 and property (i), we obtain:

$$\begin{aligned}
\frac{\theta_{k_1}}{\bar{\theta}_{k_1}} &\leq \theta_{k_1} \\
&= \theta_0(1+a\eta)^{k_1} \\
&= \theta_0 \exp\left(\ln(1+a\eta)k_1\right) \\
&\leq \theta_0 \exp\left(\ln(1+a\eta)\left(\left\lceil \frac{\ln(2-3\bar{\rho}/2) - \ln(\bar{\theta}_0 + 1 - \bar{\rho}/2)}{\ln(1+a\eta/2)} \right\rceil + \left\lceil \frac{-\ln(\bar{\theta}_{k_{1-\bar{\rho}}})}{\ln(1+a\eta)} \right\rceil\right)\right) \\
&\leq \theta_0 \exp\left(\ln(1+a\eta)\left(2 + \frac{\ln(2-3\bar{\rho}/2) - \ln(\bar{\theta}_0 + 1 - \bar{\rho}/2)}{\ln(1+a\eta/2)} - \frac{\ln(\bar{\theta}_{k_{1-\bar{\rho}}})}{\ln(1+a\eta)}\right)\right) \\
&= \theta_0 \left(\frac{2-3\bar{\rho}/2}{\bar{\theta}_0 + 1 - \bar{\rho}/2}\right)^{\frac{\ln(1+a\eta)}{\ln(1+a\eta/2)}} \frac{1}{\bar{\theta}_{k_{1-\bar{\rho}}}} (1+a\eta)^2 \\
&\leq \frac{\theta_0}{1-\bar{\rho}} \left(\frac{2-3\bar{\rho}/2}{\bar{\theta}_0 + 1 - \bar{\rho}/2}\right)^{\frac{\ln(1+a\eta)}{\ln(1+a\eta/2)}} (1+a\eta)^2.
\end{aligned}$$

Equation (3) in [56] states that  $\frac{2z}{2+z} \leq \ln(1+z) \leq \frac{z}{2} \cdot \frac{2+z}{1+z}$  for all  $z \geq 0$ . This, along with the fact that  $\eta \leq \frac{1}{6a}$  (see Subsubappendix J.11.1), leads us to  $\frac{\ln(1+a\eta)}{\ln(1+a\eta/2)} \leq \left(\frac{a\eta}{2} \cdot \frac{2+a\eta}{1+a\eta}\right) / \left(\frac{a\eta}{2+a\eta/2}\right) = \frac{4+3a\eta+(a\eta)^2/2}{2+2a\eta} = 2 + \frac{(a\eta)^2/2 - a\eta}{2+2a\eta} \leq 2 - a\eta/3$ . Thus:

$$\frac{\theta_{k_1}}{\bar{\theta}_{k_1}} \leq \frac{\theta_0}{1-\bar{\rho}} \left(\frac{2-3\bar{\rho}/2}{\bar{\theta}_0 + 1 - \bar{\rho}/2}\right)^{2-a\eta/3} (1+a\eta)^2 = \frac{\theta_0}{1-\bar{\rho}} \left(\frac{2-3\bar{\rho}/2}{\bar{\theta}_0 + 1 - \bar{\rho}/2}\right)^2 \left(\frac{\bar{\theta}_0 + 1 - \bar{\rho}/2}{2-3\bar{\rho}/2}\right)^{a\eta/3} (1+a\eta)^2.$$

By assumption on  $\bar{\theta}_0$  and  $\bar{\rho}$ , namely  $\bar{\theta}_0 \leq e^{-12} - 1$  and  $\bar{\rho} \in [0, e^{-12}/2]$ , we may bound as follows:

$$\frac{\theta_{k_1}}{\bar{\theta}_{k_1}} \leq \frac{\theta_0}{1-\bar{\rho}} \left(\frac{2-3\bar{\rho}/2}{\bar{\theta}_0 + 1 - \bar{\rho}/2}\right)^2 (e^{-12})^{a\eta/3} (1+a\eta)^2 = \frac{\theta_0}{1-\bar{\rho}} \left(\frac{2-3\bar{\rho}/2}{\bar{\theta}_0 + 1 - \bar{\rho}/2}\right)^2 e^{-4a\eta} (1+a\eta)^2.$$

Since  $1+z \leq e^z$  for all  $z \geq 0$ , we have that:

$$\frac{\theta_{k_1}}{\bar{\theta}_{k_1}} \leq \frac{\theta_0}{1-\bar{\rho}} \left(\frac{2-3\bar{\rho}/2}{\bar{\theta}_0 + 1 - \bar{\rho}/2}\right)^2 \frac{1}{1+4a\eta} (1+a\eta)^2 = \frac{\theta_0}{1-\bar{\rho}} \left(\frac{2-3\bar{\rho}/2}{\bar{\theta}_0 + 1 - \bar{\rho}/2}\right)^2 \left(1 - \frac{4}{1+4a\eta} a\eta\right) (1+2a\eta + (a\eta)^2).$$

Once again relying on the fact that  $\eta \leq \frac{1}{6a}$ , we obtain:

$$\frac{\theta_{k_1}}{\bar{\theta}_{k_1}} \leq \frac{\theta_0}{1-\bar{\rho}} \left(\frac{2-3\bar{\rho}/2}{\bar{\theta}_0 + 1 - \bar{\rho}/2}\right)^2 \left(1 - \frac{12}{5} a\eta\right) \left(1 + \frac{11}{5} a\eta\right) \leq \frac{\theta_0}{1-\bar{\rho}} \left(\frac{2-3\bar{\rho}/2}{\bar{\theta}_0 + 1 - \bar{\rho}/2}\right)^2 (1 - a\eta/5).$$

We will show that  $(1 - a\eta/5)/(1 - \bar{\rho}) \leq 1 - a\eta/10$ , thereby finishing the proof, as this leads to  $\bar{\theta}_{k_1}/\theta_{k_1} \geq ((\bar{\theta}_0 + 1 - \bar{\rho}/2)/(2 - 3\bar{\rho}/2))^2 / (\theta_0(1 - a\eta/10))$ . It holds that:

$$a\tilde{t} \leq 2\ln\left(\frac{2-3\bar{\rho}/2}{\bar{\theta}_0 - (\bar{\rho}/2 - 1)}\right) + \ln\left(\frac{1+\bar{\rho}/4}{1-3\bar{\rho}/4}\right) + \ln(b) \leq 2\ln\left(\frac{2}{e^{-12}/4}\right) + \ln(e) + \ln(b) \leq 30 + \ln(b), \quad (73)$$

where the first transition follows from the upper bound for  $\tilde{t}$ ; and the second follows from the definition  $\bar{\rho} := \min\{e^{-12}/2, \epsilon/2b\}$  together with the assumption  $\bar{\theta}_0 \in (e^{-12}/2 - 1, e^{-12} - 1)$ . The following holds:

$$\bar{\rho} \leq \frac{\epsilon}{2b} = 10^{13} \epsilon \cdot \frac{1}{10^{13} \cdot 2b} \leq 10^{13} \epsilon \cdot \frac{1}{e^{30b}} = 10^{13} \epsilon \cdot e^{-(30+\ln(b))} \leq 10^{13} \epsilon \cdot e^{-a\tilde{t}} \leq a\eta/10,$$

where the first transition follows from the definition of  $\bar{\rho}$ ; the fifth from Equation (73); and the last from the assumption  $\eta \geq e^{-a\tilde{t}} \cdot 10^{14} \epsilon/a$ . It follows that:

$$\frac{1-a\eta/5}{1-\bar{\rho}} \leq \frac{1-a\eta/5}{1-a\eta/10} = 1 - \frac{a\eta/10}{1-a\eta/10} \leq 1 - a\eta/10. \quad \square$$

We conclude this subsubappendix by combining its results with Lemma 38, thereby showing that the gradient flow trajectory between times  $t_{1-\bar{\rho}}$  and  $t_1$ , and the gradient descent trajectory between iterations  $k_{1-\bar{\rho}}$  and  $k_1$ , both lie in the transition region.

**Lemma 44.** *It holds that  $(\theta(t), \bar{\theta}(t)) \in [0, z_c) \times [1 - \bar{\rho}, 1)$  for all  $t \in [t_{1-\bar{\rho}}, t_1)$ , and  $(\theta_k, \bar{\theta}_k) \in [0, z_c) \times [1 - \bar{\rho}, 1)$  for all  $k \in \{k_{1-\bar{\rho}}, k_{1-\bar{\rho}} + 1, \dots, k_1 - 1\}$ .*

*Proof.* We start by proving the result for gradient flow. Lemma 39 implies  $t_{1-\bar{\rho}} < \infty$ , and similarly Lemma 42 implies  $t_1 < \infty$ . From continuity of  $\bar{\theta}(\cdot)$  together with the definitions of  $t_{1-\bar{\rho}}$  and  $t_1$ , we have that  $\bar{\theta}(t_{1-\bar{\rho}}) = 1 - \bar{\rho}$  and  $\bar{\theta}(t_1) = 1$ . By monotonicity of  $\bar{\theta}(\cdot)$  (Lemma 37) we conclude  $\theta(t) \in [1 - \bar{\rho}, 1)$  for all  $t \in [t_{1-\bar{\rho}}, t_1)$ . Recall that we assume  $\theta(0) \in (0.5, 1)$ . Lemma 38 implies  $\theta(t) \in [\theta(0), z_c)$  for all  $t \in [0, a^{-1} \ln(z_c/\theta(0))]$ . By Lemma 42  $t_1 \leq t_{1-\bar{\rho}} + a^{-1} \ln(1/(1 - \bar{\rho}))$ . By recalling the explicit expression for  $t_{1-\bar{\rho}}$  from Lemma 39, and all the assumptions on  $a, z_c, \bar{\rho}, \theta(0)$  and  $\bar{\theta}(0)$ , it can be seen that  $t_{1-\bar{\rho}} + a^{-1} \ln(1/(1 - \bar{\rho})) \leq a^{-1} \ln(z_c/\theta(0))$ , which implies  $t_1 \leq a^{-1} \ln(z_c/\theta(0))$ . It follows that  $\theta(t) \in [0, z_c)$  for all  $t \in [t_{1-\bar{\rho}}, t_1)$ . Overall we proved that  $(\theta(t), \bar{\theta}(t)) \in [0, z_c) \times [1 - \bar{\rho}, 1)$  for all  $t \in [t_{1-\bar{\rho}}, t_1)$ , as required.

Moving on to gradient descent, Lemma 40 implies  $k_{1-\bar{\rho}} < \infty$ , and similarly Lemma 43 implies  $k_1 < \infty$ . By definition of  $k_{1-\bar{\rho}}$  we have that  $\bar{\theta}_{k_{1-\bar{\rho}}} \geq 1 - \bar{\rho}$ , and by definition of  $k_1$  it holds that  $\bar{\theta}_{k_1 - 1} < 1$ . From monotonicity of  $(\bar{\theta}_k)_{k=0}^\infty$  (Lemma 37) and the definitions of  $k_{1-\bar{\rho}}$  and  $k_1$  (from Definitions 9 and 11 respectively), we know that  $\bar{\theta}_k \in [1 - \bar{\rho}, 1)$  for all  $k \in \{k_{1-\bar{\rho}}, k_{1-\bar{\rho}} + 1, \dots, k_1 - 1\}$ . Recall that we assume  $\theta_0 \in (0.5, 1)$ . Lemma 38 implies  $\theta_k \in [\theta_0, z_c)$  for all  $k \in \{0, 1, \dots, \lceil \ln(z_c/\theta_0) / \ln(1 + a\eta) \rceil - 1\}$ . By Lemma 43,  $k_1 \leq k_{1-\bar{\rho}} + \lceil \max\{0, -\ln(\bar{\theta}_{k_{1-\bar{\rho}}}) / \ln(1 + a\eta)\} \rceil$ . By recalling the definition of  $k_{1-\bar{\rho}}$  (Definition 9) and also its explicit expression from Lemma 40, while also recalling the assumptions on  $a, z_c, \bar{\rho}, \theta_0$  and  $\bar{\theta}_0$ , it can be seen that  $k_{1-\bar{\rho}} + \lceil \max\{0, -\ln(\bar{\theta}_{k_{1-\bar{\rho}}}) / \ln(1 + a\eta)\} \rceil \leq \lceil \ln(z_c/\theta_0) / \ln(1 + a\eta) \rceil$ , which implies  $k_1 \leq \lceil \ln(z_c/\theta_0) / \ln(1 + a\eta) \rceil$ . It follows that  $\theta_k \in [0, z_c)$  for all  $k \in \{k_{1-\bar{\rho}}, k_{1-\bar{\rho}} + 1, \dots, k_1 - 1\}$ . Overall we proved that  $(\theta_k, \bar{\theta}_k) \in [0, z_c) \times [1 - \bar{\rho}, 1)$  for all  $k \in \{k_{1-\bar{\rho}}, k_{1-\bar{\rho}} + 1, \dots, k_1 - 1\}$ .  $\square$

#### J.11.4 Isotropic Region

The current subsubappendix analyzes the second coordinate of the gradient flow and gradient descent trajectories throughout the isotropic region, or more specifically, when the second coordinate is in the range  $[1, \bar{z}_c)$ . Beginning with gradient flow, we recall that (by Lemma 37) the second coordinate of the trajectory is monotonically non-decreasing, and consider the time at which it exits the range  $[1, \bar{z}_c)$ .

**Definition 12.** Define  $t_{\bar{z}_c} := \inf\{t \geq 0 : \bar{\theta}(t) \geq \bar{z}_c\}$ .<sup>37</sup>

Using  $t_1$  from Definition 10, Lemma 45 below provides an expression for  $t_{\bar{z}_c}$ , and for the second coordinate of the gradient flow trajectory between times  $t_1$  and  $t_{\bar{z}_c}$ , *i.e.* between the time it enters the range  $[1, \bar{z}_c)$  and that at which it exits.

**Lemma 45.** *The following hold:*

- (i)  $t_{\bar{z}_c} = t_1 + a^{-1} \ln(\bar{z}_c)$ ; and
- (ii)  $\bar{\theta}(t) = e^{a(t-t_1)}$  for all  $t \in [t_1, t_{\bar{z}_c}]$ .

*Proof.* For any  $t \in [0, \infty)$  such that  $\bar{\theta}(t) \in [1, \bar{z}_c)$ , we obtain:

$$\frac{d\bar{\theta}}{dt}(t) = -\frac{d\bar{\varphi}}{dz}(\bar{\theta}(t)) = a\bar{\theta}(t).$$

Lemma 42, the definition of  $t_1$  and continuity of  $\bar{\theta}(\cdot)$  together imply that  $\bar{\theta}(t_1) = 1$ . The function  $t \mapsto e^{a(t-t_1)}$  is a solution to this initial value problem (starting at  $t_1$ ), valid through  $t \in [t_1, t_1 + a^{-1} \ln(\bar{z}_c))$ , and from uniqueness of the solution together with continuity of  $\bar{\theta}(\cdot)$ , we conclude that  $\bar{\theta}(t) = e^{a(t-t_1)}$  for  $t \in [t_1, t_1 + a^{-1} \ln(\bar{z}_c)]$ . This, along with the definition of  $t_{\bar{z}_c}$ , implies that  $t_{\bar{z}_c} = t_1 + a^{-1} \ln(\bar{z}_c)$ .  $\square$

Moving on to gradient descent, we recall that here too the second coordinate of the trajectory is monotonically non-decreasing (see Lemma 37), and consider the iteration at which this second coordinate exits the range  $[1, \bar{z}_c)$ .

**Definition 13.** Define  $k_{\bar{z}_c} := \inf\{k \in \mathbb{N} \cup \{0\} : \bar{\theta}_k \geq \bar{z}_c\}$ .<sup>37</sup>

Using  $k_1$  from Definition 11, Lemma 46 below provides an expression for  $k_{\bar{z}_c}$ , and for the second coordinate of the gradient descent trajectory between iterations  $k_1$  and  $k_{\bar{z}_c}$ , *i.e.* between the iteration where it enters the range  $[1, \bar{z}_c)$  and that at which it exits.

**Lemma 46.** *It holds that:*

- (i)  $k_{\bar{z}_c} = k_1 + \lceil \ln(\bar{z}_c/\bar{\theta}_{k_1})/\ln(1+a\eta) \rceil$ ; and
- (ii)  $\bar{\theta}_k = \bar{\theta}_{k_1}(1+a\eta)^{k-k_1}$  for all  $k \in \{k_1, k_1+1, \dots, k_{\bar{z}_c}\}$ .

*Proof.* Lemma 43 and the definition of  $k_1$  imply that  $\bar{\theta}_{k_1} \geq 1$ . Recall that by Lemma 37  $(\bar{\theta}_k)_{k=0}^\infty$  is monotonically non-decreasing. For any  $k \in \mathbb{N}$  such that  $\bar{\theta}_{k-1} \in [1, \bar{z}_c)$ , we obtain:

$$\bar{\theta}_k = \bar{\theta}_{k-1} - \eta \frac{d\bar{\varphi}}{dz}(\bar{\theta}_{k-1}) = \bar{\theta}_{k-1} + a\eta\bar{\theta}_{k-1}.$$

The solution of this recursive equation is  $\bar{\theta}_k = \bar{\theta}_{k_1}(1+a\eta)^{k-k_1}$  for all  $k \geq k_1$  such that  $\bar{\theta}_{k-1} \in [1, \bar{z}_c)$  *i.e.* for all  $k \in \{k_1, k_1+1, \dots, k_1 + \lceil \ln(\bar{z}_c/\bar{\theta}_{k_1})/\ln(1+a\eta) \rceil\}$ . From the definition of  $k_{\bar{z}_c}$  it must be equal to  $k_1 + \lceil \ln(\bar{z}_c/\bar{\theta}_{k_1})/\ln(1+a\eta) \rceil$ , as if it is smaller we get  $\bar{\theta}_{k_{\bar{z}_c}} < \bar{z}_c$ , and if it is larger we get  $\bar{\theta}_{k_{\bar{z}_c-1}} \geq \bar{z}_c$ , both contradicting the definition of  $k_{\bar{z}_c}$ .  $\square$

We conclude this subsubappendix by combining its results with Lemma 38, thereby showing that the gradient flow trajectory between times  $t_1$  and  $t_{\bar{z}_c}$ , and the gradient descent trajectory between iterations  $k_1$  and  $k_{\bar{z}_c}$ , both lie in the isotropic region.

**Lemma 47.** *It holds that  $(\theta(t), \bar{\theta}(t)) \in [0, z_c) \times [1, \bar{z}_c)$  for all  $t \in [t_1, t_{\bar{z}_c})$ , and  $(\theta_k, \bar{\theta}_k) \in [0, z_c) \times [1, \bar{z}_c)$  for all  $k \in \{k_1, k_1+1, \dots, k_{\bar{z}_c}-1\}$ .*

*Proof.* We start by proving the result for gradient flow. Lemma 42 implies  $t_1 < \infty$ , and similarly Lemma 45 implies  $t_{\bar{z}_c} < \infty$ . From continuity of  $\bar{\theta}(\cdot)$  and the definition of  $t_1$  we have that  $\bar{\theta}(t_1) = 1$ , and similarly by definition of  $t_{\bar{z}_c}$  we have  $\bar{\theta}(t_{\bar{z}_c}) = \bar{z}_c$ . By monotonicity of  $\bar{\theta}(\cdot)$  (Lemma 37) we conclude  $\bar{\theta}(t) \in [1, \bar{z}_c)$  for all  $t \in [t_1, t_{\bar{z}_c})$ . Recall that we assume  $\theta(0) \in (0.5, 1)$ . By Lemma 38 we have that  $\theta(t) \in [\theta(0), z_c)$  for all  $t \in [0, a^{-1}\ln(z_c/\theta(0))]$ . Recall the explicit expression for  $t_{1-\bar{\rho}}$  from Lemma 39. By Lemma 42  $t_1 \leq t_{1-\bar{\rho}} + a^{-1}\ln(1/(1-\bar{\rho}))$ , and by Lemma 45 we have that  $t_{\bar{z}_c} = t_1 + a^{-1}\ln(\bar{z}_c)$ . Overall, we obtain  $t_{\bar{z}_c} \leq \frac{2}{a}\ln((4-3\bar{\rho})/(2\bar{\theta}(0)+2-\bar{\rho})) + a^{-1}\ln(1/(1-\bar{\rho})) + a^{-1}\ln(\bar{z}_c)$ . By recalling the assumptions on  $a, z_c, \bar{\rho}, \theta(0)$  and  $\bar{\theta}(0)$ , it can be shown that  $t_{\bar{z}_c} \leq a^{-1}\ln(z_c/\theta(0))$ . It follows that  $\theta(t) \in [\theta(0), z_c)$  for all  $t \in [t_1, t_{\bar{z}_c})$ . Overall we proved that  $(\theta(t), \bar{\theta}(t)) \in [0, z_c) \times [1, \bar{z}_c)$  for all  $t \in [t_1, t_{\bar{z}_c})$ , as required.

Moving on to gradient descent, Lemma 43 implies  $k_1 < \infty$ , and similarly Lemma 46 implies  $k_{\bar{z}_c} < \infty$ . By definition of  $k_1$  we have that  $\bar{\theta}_{k_1} \geq 1$ , and by definition of  $k_{\bar{z}_c}$  we have  $\bar{\theta}_{k_{\bar{z}_c}-1} < \bar{z}_c$ . From monotonicity of  $(\bar{\theta}_k)_{k=0}^\infty$  (Lemma 37) we conclude  $\bar{\theta}_k \in [1, \bar{z}_c)$  for all  $k \in \{k_1, k_1+1, \dots, k_{\bar{z}_c}-1\}$ . Recall that we assume  $\theta_0 \in (0.5, 1)$ . By Lemma 38  $\theta_k \in [\theta_0, z_c)$  for all  $k \in \{0, 1, \dots, \lceil \ln(z_c/\theta_0)/\ln(1+a\eta) \rceil - 1\}$ . Recall the definition of  $k_{1-\bar{\rho}}$  and also its explicit expression from Lemma 40. By Lemma 43 we have that  $k_1 \leq k_{1-\bar{\rho}} + \lceil \max\{0, -\ln(\bar{\theta}_{k_{1-\bar{\rho}}})/\ln(1+a\eta)\} \rceil$ . By Lemma 46  $k_{\bar{z}_c} = k_1 + \lceil \ln(\bar{z}_c/\bar{\theta}_{k_1})/\ln(1+a\eta) \rceil$ . Overall, we obtain  $k_{\bar{z}_c} \leq \lceil (\ln(2-3\bar{\rho}/2) - \ln(\bar{\theta}_0 + 1 - \bar{\rho}/2))/\ln(1+a\eta/2) \rceil + \lceil \max\{0, -\ln(\bar{\theta}_{k_{1-\bar{\rho}}})/\ln(1+a\eta)\} \rceil + \lceil \ln(\bar{z}_c/\bar{\theta}_{k_1})/\ln(1+a\eta) \rceil$ . By recalling the assumptions on  $a, z_c, \bar{\rho}, \theta_0$  and  $\bar{\theta}_0$ , it can be shown that  $k_{\bar{z}_c} \leq \lceil \ln(z_c/\theta_0)/\ln(1+a\eta) \rceil$ . It follows that  $\theta_k \in [\theta_0, z_c)$  for all  $k \in \{k_1, k_1+1, \dots, k_{\bar{z}_c}-1\}$ . Overall we proved that  $(\theta_k, \bar{\theta}_k) \in [0, z_c) \times [1, \bar{z}_c)$  for all  $k \in \{k_1, k_1+1, \dots, k_{\bar{z}_c}-1\}$ , as required.  $\square$

### J.11.5 Inapproximation

The current subsubappendix shows that the location of gradient flow at time  $\tilde{t}$  is not  $\epsilon$ -approximated by different portions of the gradient descent trajectory. Key to the derived results is the following lemma, which establishes that in the isotropic region, for both gradient flow and gradient descent trajectories, the first two coordinates proceed in a straight line at an exponential pace.

**Lemma 48.** *It holds that  $(\theta(t), \bar{\theta}(t)) = (\theta(t_1), \bar{\theta}(t_1)) \cdot e^{\alpha(t-t_1)}$  for all  $t \in [t_1, t_{\bar{z}_c})$ , and  $(\theta_k, \bar{\theta}_k) = (\theta_{k_1}, \bar{\theta}_{k_1}) \cdot (1+a\eta)^{k-k_1}$  for all  $k \in \{k_1, k_1+1, \dots, k_{\bar{z}_c}-1\}$ , where  $t_1, t_{\bar{z}_c}, k_1$  and  $k_{\bar{z}_c}$  are given by Definitions 10, 12, 11 and 13 respectively.*

*Proof.* We start by proving the result for gradient flow. Lemma 45 implies  $\bar{\theta}(t) = \bar{\theta}(t_1)e^{a(t-t_1)}$  for all  $t \in [t_1, t_{\bar{z}_c}]$ . By Lemma 38 we have that  $\theta(t) = \theta(0)e^{at}$  for all  $t \in [0, a^{-1}\ln(z_c/\theta(0))]$ . Recall the explicit expression for  $t_{1-\bar{\rho}}$  from Lemma 39. By Lemma 42  $t_1 \leq t_{1-\bar{\rho}} + a^{-1}\ln(1/(1-\bar{\rho}))$ , and by Lemma 45 we have that  $t_{\bar{z}_c} = t_1 + a^{-1}\ln(\bar{z}_c)$ . Overall, we obtain  $t_{\bar{z}_c} \leq \frac{2}{a}\ln((4-3\bar{\rho})/(2\bar{\theta}(0)+2-\bar{\rho})) + a^{-1}\ln(1/(1-\bar{\rho})) + a^{-1}\ln(\bar{z}_c)$ . By recalling the assumptions on  $a, z_c, \bar{\rho}, \theta(0)$  and  $\bar{\theta}(0)$ , it can be shown that  $t_{\bar{z}_c} \leq a^{-1}\ln(z_c/\theta(0))$ . It follows that  $\theta(t) = \theta(0)e^{at} = \theta(t_1)e^{a(t-t_1)}$  for all  $t \in [t_1, t_{\bar{z}_c}]$ .

Moving on to gradient descent, Lemma 46 implies  $\bar{\theta}_k = \bar{\theta}_{k_1}(1+a\eta)^{k-k_1}$  for all  $k \in \{k_1, k_1+1, \dots, k_{\bar{z}_c}-1\}$ . By Lemma 38 we have that  $\theta_k = \theta_0(1+a\eta)^k$  for all  $k \in \{0, 1, \dots, \lceil \ln(z_c/\theta_0)/\ln(1+a\eta) \rceil\}$ . Recall the definition of  $k_{1-\bar{\rho}}$  and its explicit expression from Lemma 40. Lemma 43 ensures  $k_1 \leq k_{1-\bar{\rho}} + \lceil \max\{0, -\ln(\bar{\theta}_{k_{1-\bar{\rho}}})/\ln(1+a\eta)\} \rceil$ . By Lemma 46  $k_{\bar{z}_c} = k_1 + \lceil \ln(\bar{z}_c/\bar{\theta}_{k_1})/\ln(1+a\eta) \rceil$ . Putting it all together, we obtain  $k_{\bar{z}_c} \leq \lceil (\ln(2-3\bar{\rho}/2) - \ln(\bar{\theta}_0 + 1 - \bar{\rho}/2))/\ln(1+a\eta/2) \rceil + \lceil \max\{0, -\ln(\bar{\theta}_{k_{1-\bar{\rho}}})/\ln(1+a\eta)\} \rceil + \lceil \ln(\bar{z}_c/\bar{\theta}_{k_1})/\ln(1+a\eta) \rceil$ . By recalling the assumptions on  $a, z_c, \bar{\rho}, \theta_0$  and  $\bar{\theta}_0$ , it can be shown that  $k_{\bar{z}_c} \leq \lceil \ln(z_c/\theta_0)/\ln(1+a\eta) \rceil$ . It follows that  $\theta_k = \theta_0(1+a\eta)^k = \theta_{k_1}(1+a\eta)^{k-k_1}$  for all  $k \in \{k_1, k_1+1, \dots, k_{\bar{z}_c}-1\}$ .  $\square$

**Lemma 49.** *It holds that  $\|\theta_k - \theta(\tilde{t})\| > \epsilon$  for all  $k \in \{0, 1, \dots, k_1-1\}$ , where  $k_1$  is given by Definition 11.*

*Proof.* Recall that  $\tilde{t} \in [\frac{2}{a}\ln(\frac{2-3\bar{\rho}/2}{\bar{\theta}(0)-(\bar{\rho}/2-1)}) + \frac{1}{a}\ln(\frac{2}{1-\bar{\rho}}), \frac{2}{a}\ln(\frac{2-3\bar{\rho}/2}{\bar{\theta}(0)-(\bar{\rho}/2-1)}) + \frac{1}{a}\ln(\frac{1+\bar{\rho}/4}{1-3\bar{\rho}/4}) + \frac{1}{a}\ln(b)]$  and  $\epsilon < 1$ . Lemmas 39 and 42 imply  $\tilde{t} \in [t_1 + \frac{1}{a}\ln(2), t_1 + \frac{1}{a}\ln(b)]$ . Notice that  $t_1 \leq t_1 + \frac{1}{a}\ln(2) \leq \tilde{t} \leq t_1 + \frac{1}{a}\ln(b) \leq t_1 + \frac{1}{a}\ln(b+1) = t_{\bar{z}_c}$ . Thus, by Lemma 47, gradient flow is in the isotropic region at time  $\tilde{t}$ . We may use Lemma 45 together with monotonicity of  $\bar{\theta}(\cdot)$  (Lemma 37) to obtain  $\bar{\theta}(\tilde{t}) \geq \bar{\theta}(t_1 + \frac{1}{a}\ln(2)) = e^{a\ln(2)/a} = 2$ . From the definition of  $k_1$  (and from the fact that it is finite from Lemma 43) we know that  $|\bar{\theta}_{k_1-1}| \leq 1$ . For all  $k \in \{0, 1, \dots, k_1-1\}$ , using monotonicity of  $(\bar{\theta}_k)_{k=0}^\infty$  (Lemma 37), we may conclude:

$$\|\theta_k - \theta(\tilde{t})\|_2 \geq |\bar{\theta}_k - \bar{\theta}(\tilde{t})| \geq |\bar{\theta}(\tilde{t})| - |\bar{\theta}_k| \geq |\bar{\theta}(\tilde{t})| - |\bar{\theta}_{k_1-1}| \geq 2 - 1 = 1 > \epsilon.$$

$\square$

**Lemma 50.** *It holds that  $\|\theta_k - \theta(\tilde{t})\| > \epsilon$  for all  $k \in \{k_1, k_1+1, \dots, k_{\bar{z}_c}-1\}$ , where  $k_1$  and  $k_{\bar{z}_c}$  are given by Definitions 11 and 13 respectively.*

*Proof.* Recall that  $\tilde{t} \in [\frac{2}{a}\ln(\frac{2-3\bar{\rho}/2}{\bar{\theta}(0)-(\bar{\rho}/2-1)}) + \frac{1}{a}\ln(\frac{2}{1-\bar{\rho}}), \frac{2}{a}\ln(\frac{2-3\bar{\rho}/2}{\bar{\theta}(0)-(\bar{\rho}/2-1)}) + \frac{1}{a}\ln(\frac{1+\bar{\rho}/4}{1-3\bar{\rho}/4}) + \frac{1}{a}\ln(b)]$  and  $\epsilon < 1$ . Lemmas 39 and 42 imply  $\tilde{t} \in [t_1 + \frac{1}{a}\ln(2), t_1 + \frac{1}{a}\ln(b)]$ . Notice that  $t_1 \leq t_1 + \frac{1}{a}\ln(2) \leq \tilde{t} \leq t_1 + \frac{1}{a}\ln(b) \leq t_1 + \frac{1}{a}\ln(b+1) = t_{\bar{z}_c}$ . Thus, by Lemma 47, gradient flow is in the isotropic region at time  $\tilde{t}$ . By Lemma 43 we know that  $k_1 < \infty$ , and by Lemma 47 it holds that  $\bar{\theta}_{k_1} \geq 1$ . By monotonicity of  $(\theta_k)_{k=0}^\infty$  (Lemma 37) and since  $\theta_0 > 0$ , we know that  $\theta_k > 0$  for all  $k \in \mathbb{N}$ . For all  $k \in \{k_1, k_1+1, \dots, k_{\bar{z}_c}-1\}$ , Lemma 48 ensures  $\theta_k/\bar{\theta}_k = \theta_{k_1}/\bar{\theta}_{k_1}$ , thus  $(\theta_k, \bar{\theta}_k) \in \{(q, \bar{q}) : q, \bar{q} \in (0, \infty) \text{ s.t. } q/\bar{q} = \theta_{k_1}/\bar{\theta}_{k_1}\} = \{c(\theta_{k_1}, \bar{\theta}_{k_1}) : c > 0\}$ . This leads us to:

$$\|\theta_k - \theta(\tilde{t})\|_2 \geq \|(\theta_k, \bar{\theta}_k) - (\theta(\tilde{t}), \bar{\theta}(\tilde{t}))\|_2 \geq \inf_{c>0} \|c(\theta_{k_1}, \bar{\theta}_{k_1}) - (\theta(\tilde{t}), \bar{\theta}(\tilde{t}))\|_2.$$

Minimizing over  $c > 0$  yields  $c_{\min} = \langle (\theta(\tilde{t}), \bar{\theta}(\tilde{t})), (\theta_{k_1}, \bar{\theta}_{k_1}) \rangle / \|(\theta_{k_1}, \bar{\theta}_{k_1})\|_2^2$ . Note that since gradient flow is in the isotropic region at time  $\tilde{t}$ , then  $\|(\theta(\tilde{t}), \bar{\theta}(\tilde{t}))\|_2 \neq 0$ . We obtain:

$$\begin{aligned} \|\theta_k - \theta(\tilde{t})\|_2 &\geq \left\| (\theta_{k_1}, \bar{\theta}_{k_1}) \langle (\theta(\tilde{t}), \bar{\theta}(\tilde{t})), (\theta_{k_1}, \bar{\theta}_{k_1}) \rangle / \|(\theta_{k_1}, \bar{\theta}_{k_1})\|_2^2 - (\theta(\tilde{t}), \bar{\theta}(\tilde{t})) \right\|_2 \\ &= \sqrt{\|(\theta(\tilde{t}), \bar{\theta}(\tilde{t}))\|_2^2 - \left\langle (\theta(\tilde{t}), \bar{\theta}(\tilde{t})), \frac{(\theta_{k_1}, \bar{\theta}_{k_1})}{\|(\theta_{k_1}, \bar{\theta}_{k_1})\|_2} \right\rangle^2} \\ &= \|(\theta(\tilde{t}), \bar{\theta}(\tilde{t}))\|_2 \sqrt{1 - \left\langle \frac{(\theta(\tilde{t}), \bar{\theta}(\tilde{t}))}{\|(\theta(\tilde{t}), \bar{\theta}(\tilde{t}))\|_2}, \frac{(\theta_{k_1}, \bar{\theta}_{k_1})}{\|(\theta_{k_1}, \bar{\theta}_{k_1})\|_2} \right\rangle^2} \\ &\geq |\theta(\tilde{t})| \sqrt{1 - \left\langle \frac{(\theta(\tilde{t}), \bar{\theta}(\tilde{t}))}{\|(\theta(\tilde{t}), \bar{\theta}(\tilde{t}))\|_2}, \frac{(\theta_{k_1}, \bar{\theta}_{k_1})}{\|(\theta_{k_1}, \bar{\theta}_{k_1})\|_2} \right\rangle^2}. \end{aligned}$$

By Lemma 38 we have that  $\theta(t) = \theta(0)e^{at}$  for all  $t \in [0, a^{-1} \ln(z_c/\theta(0))]$ . Recall the explicit expression for  $t_{1-\bar{\rho}}$  from Lemma 39. By Lemma 42  $t_1 \leq t_{1-\bar{\rho}} + a^{-1} \ln(1/(1-\bar{\rho}))$ , and by Lemma 45 we have that  $t_{\bar{z}_c} = t_1 + a^{-1} \ln(\bar{z}_c)$ . Since  $\tilde{t} \leq t_{\bar{z}_c}$ , overall we obtain  $\tilde{t} \leq \frac{2}{a} \ln((4-3\bar{\rho})/(2\bar{\theta}(0)+2-\bar{\rho})) + a^{-1} \ln(1/(1-\bar{\rho})) + a^{-1} \ln(\bar{z}_c)$ . By recalling the assumptions on  $a, z_c, \bar{\rho}, \theta(0)$  and  $\bar{\theta}(0)$ , it can be shown that  $\tilde{t} \leq a^{-1} \ln(z_c/\theta(0))$ . It follows that:

$$\|\theta_k - \theta(\tilde{t})\|_2 \geq \theta(0)e^{a\tilde{t}} \sqrt{1 - \left\langle \frac{(\theta(\tilde{t}), \bar{\theta}(\tilde{t}))}{\|(\theta(\tilde{t}), \bar{\theta}(\tilde{t}))\|_2}, \frac{(\theta_{k_1}, \bar{\theta}_{k_1})}{\|(\theta_{k_1}, \bar{\theta}_{k_1})\|_2} \right\rangle^2}.$$

Recall that  $\tilde{t} \in [t_1, t_{\bar{z}_c}]$ . Lemma 47 implies that both  $\bar{\theta}(\tilde{t})$  and  $\bar{\theta}(t_1)$  are greater or equal to one. Since Lemma 48 implies  $\bar{\theta}(\tilde{t})/\bar{\theta}(\tilde{t}) = \theta(t_1)/\bar{\theta}(t_1)$ , it follows that:

$$\begin{aligned} \|\theta_k - \theta(\tilde{t})\|_2 &\geq \theta(0)e^{a\tilde{t}} \sqrt{1 - \left\langle \frac{(\theta(t_1), \bar{\theta}(t_1))}{\|(\theta(t_1), \bar{\theta}(t_1))\|_2}, \frac{(\theta_{k_1}, \bar{\theta}_{k_1})}{\|(\theta_{k_1}, \bar{\theta}_{k_1})\|_2} \right\rangle^2} \\ &= \theta(0)e^{a\tilde{t}} \sqrt{1 - \left\langle \frac{(\theta(t_1)/\bar{\theta}(t_1), 1)}{\|(\theta(t_1)/\bar{\theta}(t_1), 1)\|_2}, \frac{(\theta_{k_1}/\bar{\theta}_{k_1}, 1)}{\|(\theta_{k_1}/\bar{\theta}_{k_1}, 1)\|_2} \right\rangle^2} \end{aligned}$$

Note that the latter inner product is between positively correlated unit vectors, and that it is squared. We use Lemmas 42 and 43, ensuring that  $\bar{\theta}(t_1)/\theta(t_1) \leq ((\bar{\theta}(0) + 1 - \bar{\rho}/2)/(2 - 3\bar{\rho}/2))^2/\theta(0)$ , and  $\bar{\theta}_{k_1}/\theta_{k_1} \geq ((\bar{\theta}_0 + 1 - \bar{\rho}/2)/(2 - 3\bar{\rho}/2))^2/(\theta_0(1 - a\eta/10))$  respectively. For brevity, denote  $\alpha := \theta(0)((4 - 3\bar{\rho})/(2\bar{\theta}(0) + 2 - \bar{\rho}))^2$  and  $\beta := (1 - a\eta/10)$ . Notice that  $\beta \in (0, 1)$ . Recall that by definition  $\theta(0) = \theta_0$  and  $\bar{\theta}(0) = \bar{\theta}_0$ . Thus,  $\theta_{t_1}/\bar{\theta}_{t_1} \leq \alpha\beta < \alpha \leq \theta(t_1)/\bar{\theta}(t_1)$ . Replacing  $\theta(t_1)/\bar{\theta}(t_1)$  with  $\alpha$ , and  $\theta_{k_1}/\bar{\theta}_{k_1}$  with  $\alpha\beta$ , decreases the angle between the unit vectors, thereby increasing their inner product. We thus have that:

$$\begin{aligned} \|\theta_k - \theta(\tilde{t})\|_2 &\geq \theta(0)e^{a\tilde{t}} \sqrt{1 - \left\langle \frac{(\alpha, 1)}{\|(\alpha, 1)\|_2}, \frac{(\alpha\beta, 1)}{\|(\alpha\beta, 1)\|_2} \right\rangle^2} \\ &= \theta(0)e^{a\tilde{t}} \sqrt{1 - \left( \frac{\alpha^2\beta + 1}{\sqrt{\alpha^2 + 1} \cdot \sqrt{\alpha^2\beta^2 + 1}} \right)^2} \\ &= \theta(0)e^{a\tilde{t}} \sqrt{\frac{\alpha^4\beta^2 + \alpha^2\beta^2 + \alpha^2 + 1}{\alpha^4\beta^2 + \alpha^2\beta^2 + \alpha^2 + 1} - \left( \frac{\alpha^4\beta^2 + 2\alpha^2\beta + 1}{\alpha^4\beta^2 + \alpha^2\beta^2 + \alpha^2 + 1} \right)} \\ &= \theta(0)e^{a\tilde{t}} \sqrt{\frac{\alpha^2(1-\beta)^2}{\alpha^4\beta^2 + \alpha^2\beta^2 + \alpha^2 + 1}} \end{aligned}$$

Since  $\beta \in (0, 1)$  and  $\alpha \geq 1$  (can be verified by recalling the assumptions on  $a, \eta, \bar{\rho}, \theta(0)$  and  $\bar{\theta}(0)$ ), we obtain:

$$\|\theta_k - \theta(\tilde{t})\|_2 \geq \theta(0)e^{a\tilde{t}} \left( \frac{1-\beta}{2\alpha} \right).$$

Plugging in  $\alpha$  and  $\beta$ , we obtain:

$$\|\theta_k - \theta(\tilde{t})\|_2 \geq \theta(0)e^{a\tilde{t}} \frac{a\eta}{10} \cdot \frac{1}{2\bar{\theta}(0)} \left( \frac{2\bar{\theta}(0) + 2 - \bar{\rho}}{4 - 3\bar{\rho}} \right)^2 = \frac{1}{20} \left( \frac{2\bar{\theta}(0) + 2 - \bar{\rho}}{4 - 3\bar{\rho}} \right)^2 a\eta e^{a\tilde{t}}.$$

The definition of  $\bar{\rho}$ , and the assumptions on  $\theta(0)$  and  $\bar{\theta}(0)$  lead us to:

$$\|\theta_k - \theta(\tilde{t})\|_2 \geq \frac{1}{20} \left( \frac{0.5e^{-12}}{4} \right)^2 a\eta e^{a\tilde{t}} > 10^{-14} a\eta e^{a\tilde{t}}.$$

Since  $\eta \geq \epsilon 10^{14} e^{-a\tilde{t}}/a$ , we have  $\|\theta_k - \theta(\tilde{t})\|_2 > \epsilon$ , and this holds for all  $k \in \{k_1, k_1 + 1, \dots, k_{\bar{z}_c} - 1\}$ .  $\square$

**Lemma 51.** *It holds that  $\|\theta_k - \theta(\tilde{t})\| > \epsilon$  for all  $k \in \{k_{\bar{z}_c}, k_{\bar{z}_c} + 1, k_{\bar{z}_c} + 2, \dots\}$ , where  $k_{\bar{z}_c}$  is given by Definition 13.*

*Proof.* Recall that  $\tilde{t} \in [\frac{2}{a} \ln(\frac{2-3\bar{\rho}/2}{\bar{\theta}(0) - (\bar{\rho}/2 - 1)}) + \frac{1}{a} \ln(\frac{2}{1-\bar{\rho}}), \frac{2}{a} \ln(\frac{2-3\bar{\rho}/2}{\bar{\theta}(0) - (\bar{\rho}/2 - 1)}) + \frac{1}{a} \ln(\frac{1+\bar{\rho}/4}{1-3\bar{\rho}/4}) + \frac{1}{a} \ln(b)]$  and  $\epsilon < 1$ . Lemmas 39 and 42 imply  $\tilde{t} \in [t_1 + \frac{1}{a} \ln(2), t_1 + \frac{1}{a} \ln(b)]$ . Notice that  $t_1 \leq t_1 + \frac{1}{a} \ln(2) \leq \tilde{t} \leq t_1 + \frac{1}{a} \ln(b) \leq t_1 + \frac{1}{a} \ln(b+1) = t_{\bar{z}_c}$ . Thus, by Lemma 47, gradient flow is in the isotropic region at time  $\tilde{t}$ . We may use Lemma 45 together with monotonicity of  $\bar{\theta}(\cdot)$

(Lemma 37) to obtain  $\bar{\theta}(\tilde{t}) \leq \bar{\theta}(t_1 + \frac{1}{a} \ln(b)) = e^{a \ln(b)/a} = b$ . Lemma 37 further ensures monotonicity of  $(\bar{\theta}_k)_{k=0}^\infty$ . By the definition of  $\bar{\theta}_{k_{\bar{z}_c}}$  (and from the fact that it is finite from Lemma 46) we know that  $\bar{\theta}_{k_{\bar{z}_c}} \geq \bar{z}_c = b + 1$ . This implies, for all  $k \in \{k_{\bar{z}_c}, k_{\bar{z}_c} + 1, k_{\bar{z}_c} + 2, \dots\}$ :

$$\|\boldsymbol{\theta}_k - \boldsymbol{\theta}(\tilde{t})\|_2 \geq |\bar{\theta}_k - \bar{\theta}(\tilde{t})| \geq |\bar{\theta}_k| - |\bar{\theta}(\tilde{t})| \geq |\bar{\theta}_{k_{\bar{z}_c}}| - |\bar{\theta}(\tilde{t})| \geq (b+1) - b = 1 > \epsilon.$$

□

### J.11.6 Conclusion

Taken together, Lemmas 49, 50 and 51 form a proof for Proposition 4 in the case where the step size  $\eta$  is no greater than  $\frac{1}{6a}$ . The complementary case  $\eta > \frac{1}{6a}$  is accounted for by Lemma 36.

### J.12 Proof of Lemma 4

This proof is very similar to that of Lemma 1 (see Subappendix J.4). We repeat all details for completeness. Recall that  $\boldsymbol{\theta} \in \mathbb{R}^d$  is an arrangement of  $(W_1, W_2, \dots, W_n) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$  as a vector. Let  $(\Delta W_1, \Delta W_2, \dots, \Delta W_n) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$ , and denote by  $\Delta \boldsymbol{\theta} \in \mathbb{R}^d$  its arrangement as a vector in corresponding order. Denote the following for  $i \in \{1, \dots, |\mathcal{S}|\}$ :

$$\Delta_i^{(1)} := \sum_{j=1}^n (D'_{i,*} W_*)_{n:j+1} D'_{i,j} (\Delta W_j) (D'_{i,*} W_*)_{j-1:1}, \quad (74)$$

$$\Delta_i^{(2)} := \sum_{1 \leq j < j' \leq n} (D'_{i,*} W_*)_{n:j'+1} D'_{i,j'} (\Delta W_{j'}) (D'_{i,*} W_*)_{j'-1:j+1} D'_{i,j} (\Delta W_j) (D'_{i,*} W_*)_{j-1:1}, \quad (75)$$

$$\Delta_i^{(3:n)} := D'_{i,n} (W_n + \Delta W_n) \dots D'_{i,1} (W_1 + \Delta W_1) - (D'_{i,*} W_*)_{n:1} - \Delta_i^{(1)} - \Delta_i^{(2)}. \quad (76)$$

We now develop a second-order Taylor expansion of  $f(\boldsymbol{\theta})$ . Since the matrix tuple corresponding to  $(\boldsymbol{\theta} + \Delta \boldsymbol{\theta})$  is  $((W_1 + \Delta W_1), \dots, (W_n + \Delta W_n))$ , and the function  $f(\cdot)$  coincides with the function given in Equation (24) on an open region containing  $\boldsymbol{\theta}$ , for sufficiently small  $\Delta \boldsymbol{\theta}$  we obtain:

$$\begin{aligned} & f(\boldsymbol{\theta} + \Delta \boldsymbol{\theta}) \\ &= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell(D'_{i,n} (W_n + \Delta W_n) \dots D'_{i,1} (W_1 + \Delta W_1) \mathbf{x}_i, y_i) \\ &= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell\left(\left((D'_{i,*} W_*)_{n:1} + \Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3:n)}\right) \mathbf{x}_i, y_i\right) \\ &= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell\left((D'_{i,*} W_*)_{n:1} \mathbf{x}_i + (\Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i, y_i\right), \end{aligned} \quad (77)$$

where the second transition follows from the definition of  $\Delta_i^{(3:n)}$  (Equation (76)). Let  $\Delta \mathbf{v} \in \mathbb{R}^{d_n}$ . For every  $i \in \{1, \dots, |\mathcal{S}|\}$ , the second-order Taylor expansion of  $\ell(\cdot)$  with respect to its first argument at the point  $((D'_{i,*} W_*)_{n:1} \mathbf{x}_i, y_i)$  is given by:

$$\ell\left((D'_{i,*} W_*)_{n:1} \mathbf{x}_i + \Delta \mathbf{v}, y_i\right) = \ell\left((D'_{i,*} W_*)_{n:1} \mathbf{x}_i, y_i\right) + \langle \nabla \ell_i, \Delta \mathbf{v} \rangle + \frac{1}{2} \nabla^2 \ell_i [\Delta \mathbf{v}] + \mathcal{O}(\|\Delta \mathbf{v}\|_2^2), \quad (78)$$



where the  $\mathcal{O}(\cdot)$  notation refers to some expression satisfying  $\lim_{a \rightarrow 0} (\mathcal{O}(a)/a) = 0$ . We continue to develop Equation (77) using Equation (78):

$$\begin{aligned}
& f(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) \\
&= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \left( \ell((D'_{i,*} W_*)_{n:1} \mathbf{x}_i, y_i) + \langle \nabla \ell_i, (\Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i \rangle + \right. \\
&\quad \left. \frac{1}{2} \nabla^2 \ell_i [(\Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i] + \mathcal{O}(\|(\Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i\|_2^2) \right) \\
&= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell((D'_{i,*} W_*)_{n:1} \mathbf{x}_i, y_i) + \\
&\quad \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \langle \nabla \ell_i, \Delta_i^{(1)} \mathbf{x}_i \rangle + \langle \nabla \ell_i, \Delta_i^{(2)} \mathbf{x}_i \rangle + \langle \nabla \ell_i, \Delta_i^{(3:n)} \mathbf{x}_i \rangle + \\
&\quad \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \frac{1}{2} \nabla^2 \ell_i [\Delta_i^{(1)} \mathbf{x}_i] + \frac{1}{2} \nabla^2 \ell_i [(\Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i] + 2 \cdot \frac{1}{2} \nabla^2 \ell_i [\Delta_i^{(1)} \mathbf{x}_i, (\Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i] + \\
&\quad \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \mathcal{O}(\|(\Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i\|_2^2),
\end{aligned}$$

where in the last transition we view  $\nabla^2 \ell_i$  as both a quadratic and a bilinear form (see Subappendix J.1). Notice that  $\langle \nabla \ell_i, \Delta_i^{(3:n)} \mathbf{x}_i \rangle$ ,  $\frac{1}{2} \nabla^2 \ell_i [(\Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i]$ ,  $\nabla^2 \ell_i [\Delta_i^{(1)} \mathbf{x}_i, (\Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i]$  and  $\mathcal{O}(\|(\Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i\|_2^2)$  are all  $\mathcal{O}(\|\Delta\boldsymbol{\theta}\|_2^2)$ , thus:

$$\begin{aligned}
& f(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) \\
&= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell((D'_{i,*} W_*)_{n:1} \mathbf{x}_i, y_i) + \langle \nabla \ell_i, \Delta_i^{(1)} \mathbf{x}_i \rangle + \langle \nabla \ell_i, \Delta_i^{(2)} \mathbf{x}_i \rangle + \frac{1}{2} \nabla^2 \ell_i [\Delta_i^{(1)} \mathbf{x}_i] + \mathcal{O}(\|\Delta\boldsymbol{\theta}\|_2^2).
\end{aligned}$$

This is a Taylor expansion of  $f(\cdot)$  evaluated at  $\boldsymbol{\theta}$  with a constant term  $\frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell((D'_{i,*} W_*)_{n:1} \mathbf{x}_i, y_i)$ , a linear term  $\frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \langle \nabla \ell_i, \Delta_i^{(1)} \mathbf{x}_i \rangle$ , a quadratic term of two summands  $\frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \langle \nabla \ell_i, \Delta_i^{(2)} \mathbf{x}_i \rangle + \frac{1}{2} \nabla^2 \ell_i [\Delta_i^{(1)} \mathbf{x}_i]$ , and a remainder term of  $\mathcal{O}(\|\Delta\boldsymbol{\theta}\|_2^2)$ . From uniqueness of the Taylor expansion, the quadratic term must be equal to  $\frac{1}{2} \nabla^2 f(\boldsymbol{\theta}) [\Delta W_1, \dots, \Delta W_n]$ . This implies:

$$\begin{aligned}
& \nabla^2 f(\boldsymbol{\theta}) [\Delta W_1, \dots, \Delta W_n] \\
&= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \left( \nabla^2 \ell_i [\Delta_i^{(1)} \mathbf{x}_i] + 2 \langle \nabla \ell_i, \Delta_i^{(2)} \mathbf{x}_i \rangle \right) \\
&= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \left( \nabla^2 \ell_i \left[ \sum_{j=1}^n (D'_{i,*} W_*)_{n:j+1} D'_{i,j} (\Delta W_j) (D'_{i,*} W_*)_{j-1:1} \mathbf{x}_i \right] + \right. \\
&\quad \left. 2 \langle \nabla \ell_i, \sum_{1 \leq j < j' \leq n} (D'_{i,*} W_*)_{n:j'+1} D'_{i,j'} (\Delta W_{j'}) (D'_{i,*} W_*)_{j'-1:j+1} D'_{i,j} (\Delta W_j) (D'_{i,*} W_*)_{j-1:1} \mathbf{x}_i \rangle \right),
\end{aligned}$$

where the last transition follows from plugging in the definitions of  $\Delta^{(1)}$  and  $\Delta^{(2)}$  (see Equations (74) and (75)).  $\square$

### J.13 Proof of Proposition 5

From assumption (ii) there exists some  $\boldsymbol{\theta} \in \mathbb{R}^d$  such that  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) \neq 0$ . Define  $(W_1, W_2, \dots, W_n) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$  to be the weight matrices constituting  $\boldsymbol{\theta}$ . We may assume  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) < 0$  without loss of generality, as we can negate the vectors  $h_{\boldsymbol{\theta}}(\mathbf{x}_i) \in \mathbb{R}^{d_n}$  for all  $i \in \{1, 2, \dots, |\mathcal{S}|\}$  by flipping the signs of the entries in  $\boldsymbol{\theta}$  corresponding to the last

weight matrix  $W_n$  (see Equation (8)). From continuity, there exists a neighborhood  $\mathcal{N}$  of  $\theta$  such that for all  $\tilde{\theta} \in \mathcal{N}$  it holds that  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_{\tilde{\theta}}(\mathbf{x}_i) < 0$ . Moreover, as discussed in Appendix C, for almost all  $\theta' \in \mathbb{R}^d$  there exists an open region  $\mathcal{D}_{\theta'} \subseteq \mathbb{R}^d$  containing  $\theta'$ , which is closed under positive rescaling of weight matrices and across which  $f(\cdot)$  coincides with a function as given in Equation (24). There must exist some  $\theta'$  in the neighborhood  $\mathcal{N}$  for which a region of the type  $\mathcal{D}_{\theta'}$  exists. We may assume, without loss of generality, that  $\theta \in \mathcal{D}_{\theta'}$ . Notice that none of the matrices  $W_1, W_2, \dots, W_n$  are equal to zero (as that would lead to  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_{\theta}(\mathbf{x}_i) = 0$ ). Define the following weight matrices parameterized by  $a > 0$  (while recalling that  $n \geq 3$  by assumption (i)):

$$\begin{aligned} W_1(a) &:= W_1 \cdot a^{-2} \in \mathbb{R}^{d_1, d_0}, \\ W_2(a) &:= W_2 \cdot a^{-2} \in \mathbb{R}^{d_2, d_1}, \\ W_3(a) &:= W_3 \cdot a \in \mathbb{R}^{d_3, d_2}, \\ W_j(a) &:= W_j \in \mathbb{R}^{d_j, d_{j-1}} \text{ for } j \in \{1, 2, \dots, n\} \setminus \{1, 2, 3\}, \end{aligned}$$

and denote by  $\theta(a) \in \mathbb{R}^d$  their corresponding weight setting. Since  $\mathcal{D}_{\theta'}$  is closed under positive rescaling of weight matrices, it holds that  $\{\theta(a) : a > 0\} \subseteq \mathcal{D}_{\theta'}$ . Define:

$$\begin{aligned} \Delta W_1 &:= W_1 \in \mathbb{R}^{d_1, d_0}, \\ \Delta W_2 &:= W_2 \in \mathbb{R}^{d_2, d_1}, \\ \Delta W_j &:= 0 \in \mathbb{R}^{d_j, d_{j-1}} \text{ for } j \in [n] \setminus \{1, 2\}. \end{aligned}$$

For  $a > 0$ ,  $i \in \{1, 2, \dots, |\mathcal{S}|\}$  and  $j, j' \in \{1, 2, \dots, n\}$ , define  $(D'_{i,*} W_*(a))_{j':j}$  to be the matrix  $D'_{i,j'} W_{j'}(a) D'_{i,j'-1} W_{j'-1}(a) \dots D'_{i,j} W_j(a)$  (where by convention  $D'_{i,n} \in \mathbb{R}^{d_n, d_n}$  stands for identity) if  $j \leq j'$ , and an identity matrix (with size to be inferred by context) otherwise. For  $i \in \{1, 2, \dots, |\mathcal{S}|\}$  and  $a > 0$  let  $\nabla \ell_i(a) \in \mathbb{R}^{d_n}$  and  $\nabla^2 \ell_i(a) \in \mathbb{R}^{d_n, d_n}$  be the gradient and Hessian (respectively) of the loss  $\ell(\cdot)$  at the point  $((D'_{i,*} W_*(a))_{n:1} \mathbf{x}_i, y_i)$  with respect to its first argument. For every  $a > 0$ , since  $\theta(a) \in \mathcal{D}_{\theta'}$  we may apply Lemma 4, obtaining:

$$\begin{aligned} \nabla^2 f(\theta(a))[\Delta W_1, \dots, \Delta W_n] &= \\ \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i(a) \left[ \sum_{j=1}^n (D'_{i,*} W_*(a))_{n:j+1} D'_{i,j} (\Delta W_j) (D'_{i,*} W_*(a))_{j-1:1} \mathbf{x}_i \right] &+ \\ \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i(a)^\top \sum_{1 \leq j < j' \leq n} (D'_{i,*} W_*(a))_{n:j'+1} D'_{i,j'} (\Delta W_{j'}) (D'_{i,*} W_*(a))_{j'-1:j+1} D'_{i,j} (\Delta W_j) (D'_{i,*} W_*(a))_{j-1:1} \mathbf{x}_i, & \end{aligned} \quad (79)$$

where we regard Hessians as quadratic forms (see Subappendix J.1). Plugging in the definitions of  $W_j(a)$  and  $\Delta W_j$  for  $j \in [n]$  we have:

$$\begin{aligned} \nabla^2 f(\theta(a))[\Delta W_1, \dots, \Delta W_n] &= \\ &= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i(a) \left[ 2a^{-1} (D'_{i,*} W_*)_{n:1} \mathbf{x}_i \right] + \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i(a)^\top a (D'_{i,*} W_*)_{n:1} \mathbf{x}_i \\ &= \frac{4}{a^2} \cdot \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i(a) \left[ (D'_{i,*} W_*)_{n:1} \mathbf{x}_i \right] + a \cdot \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i(a)^\top h_{\theta}(\mathbf{x}_i), \end{aligned} \quad (80)$$

where the second transition follows from pulling  $2/a$  out of the quadratic operator and the fact that  $h_{\theta}(\mathbf{x}_i) = (D'_{i,*} W_*)_{n:1} \mathbf{x}_i$ . Note that  $\lim_{a \rightarrow \infty} (D'_{i,*} W_*)_{n:1} = 0$ . Since  $\ell(\cdot)$  is twice continuously differentiable in its first argument, it holds that  $\lim_{a \rightarrow \infty} \nabla^2 \ell_i(a) = \lim_{a \rightarrow \infty} \nabla^2 \ell((D'_{i,*} W_*(a))_{n:1} \mathbf{x}_i, y_i) = \nabla^2 \ell(\mathbf{0}, y_i)$ , and similarly  $\lim_{a \rightarrow \infty} \nabla \ell_i(a) = \lim_{a \rightarrow \infty} \nabla \ell((D'_{i,*} W_*(a))_{n:1} \mathbf{x}_i, y_i) = \nabla \ell(\mathbf{0}, y_i)$ . Therefore,

in the limit  $a \rightarrow \infty$ , Equation (80) becomes:

$$\begin{aligned}
& \lim_{a \rightarrow \infty} \left( \nabla^2 f(\boldsymbol{\theta}(a)) [\Delta W_1, \dots, \Delta W_n] \right) \\
&= \lim_{a \rightarrow \infty} \left( \frac{4}{a^2} \cdot \frac{4}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i(a) \left[ (D'_{i,*} W_*)_{n:1} \mathbf{x}_i \right] \right) + \lim_{a \rightarrow \infty} \left( a \cdot \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i(a)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) \right) \\
&= \lim_{a \rightarrow \infty} \left( \frac{4}{a^2} \right) \cdot \lim_{a \rightarrow \infty} \left( \frac{4}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i(a) \left[ (D'_{i,*} W_*)_{n:1} \mathbf{x}_i \right] \right) + \lim_{a \rightarrow \infty} \left( a \cdot \lim_{a \rightarrow \infty} \left( \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i(a)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) \right) \right) \\
&= 0 \cdot \left( \frac{4}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell(\mathbf{0}, y_i) \left[ (D'_{i,*} W_*)_{n:1} \mathbf{x}_i \right] \right) + \lim_{a \rightarrow \infty} \left( a \cdot \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i(\mathbf{0}, y_i)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) \right) \\
&= -\infty,
\end{aligned}$$

where the second transition is valid since the multiplied limits are finite and the limit inside a limit is non-zero, and the last transition follows from  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) < 0$ . Notice that the matrices  $\Delta W_1, \Delta W_2, \dots, \Delta W_n$  are independent of  $a$ , thus it must hold that  $\lim_{a \rightarrow \infty} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta}(a))) = -\infty$ . This in particular implies the desired result:

$$\inf_{\boldsymbol{\theta} \in \mathbb{R}^d \text{ s.t. } \nabla^2 f(\boldsymbol{\theta}) \text{ exists}} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) = -\infty.$$

□

#### J.14 Proof of Lemma 5

This proof is very similar to that of Lemma 2 (see Subappendix J.6). We repeat all details for completeness. Recall that  $\boldsymbol{\theta} \in \mathbb{R}^d$  is an arrangement of  $(W_1, W_2, \dots, W_n) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$  as a vector. Let  $(\Delta W_1, \Delta W_2, \dots, \Delta W_n) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$ , and denote by  $\Delta \boldsymbol{\theta} \in \mathbb{R}^d$  its arrangement as a vector in corresponding order. As shown in Lemma 4:

$$\begin{aligned}
& \nabla^2 f(\boldsymbol{\theta}) [\Delta W_1, \dots, \Delta W_n] = \\
& \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i \left[ \sum_{j=1}^n (D'_{i,*} W_*)_{n:j+1} D'_{i,j} (\Delta W_j) (D'_{i,*} W_*)_{j-1:1} \mathbf{x}_i \right] + \\
& \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i^\top \sum_{1 \leq j < j' \leq n} (D'_{i,*} W_*)_{n:j'+1} D'_{i,j'} (\Delta W_{j'}) (D'_{i,*} W_*)_{j'-1:j+1} D'_{i,j} (\Delta W_j) (D'_{i,*} W_*)_{j-1:1} \mathbf{x}_i,
\end{aligned}$$

where we regard Hessians as quadratic forms (see Subappendix J.1). Convexity of  $\ell(\cdot)$  in its first argument implies that for  $i \in \{1, 2, \dots, |\mathcal{S}|\}$ ,  $\nabla^2 \ell_i$  is positive semi-definite, thus:

$$\begin{aligned}
& \nabla^2 f(\boldsymbol{\theta}) [\Delta W_1, \dots, \Delta W_n] \geq \\
& \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i^\top \sum_{1 \leq j < j' \leq n} (D'_{i,*} W_*)_{n:j'+1} D'_{i,j'} (\Delta W_{j'}) (D'_{i,*} W_*)_{j'-1:j+1} D'_{i,j} (\Delta W_j) (D'_{i,*} W_*)_{j-1:1} \mathbf{x}_i.
\end{aligned}$$

Applying Cauchy-Schwarz and triangle inequalities, we get:

$$\begin{aligned}
& \nabla^2 f(\boldsymbol{\theta}) [\Delta W_1, \dots, \Delta W_n] \\
& \geq - \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \sum_{1 \leq j < j' \leq n} \left\| (D'_{i,*} W_*)_{n:j'+1} D'_{i,j'} (\Delta W_{j'}) (D'_{i,*} W_*)_{j'-1:j+1} D'_{i,j} (\Delta W_j) (D'_{i,*} W_*)_{j-1:1} \mathbf{x}_i \right\|_2 \\
& \geq - \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \sum_{1 \leq j < j' \leq n} \|\Delta W_j\|_s \|\Delta W_{j'}\|_s \prod_{k \in [n] \setminus \{j, j'\}} \|W_k\|_s \prod_{k=1}^n \|D'_{i,k}\|_s \cdot \|\mathbf{x}_i\|_2 \\
& \geq - \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \cdot \left( \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_s \right) \max\{|\alpha|, |\bar{\alpha}|\}^{n-1} \|\mathbf{x}_i\|_2 \sum_{1 \leq j < j' \leq n} \|\Delta W_j\|_s \|\Delta W_{j'}\|_s,
\end{aligned}$$

where the second transition follows from the definition and sub-multiplicativity of spectral norm, and the last transition follows from maximizing  $\prod_{k \in [n] \setminus \{j, j'\}} \|W_k\|_s$  over  $j, j'$ , upper bounding  $\|D'_{i,j}\|_s \leq \max\{|\alpha|, |\bar{\alpha}|\}$  for  $j \in [n-1]$  and recalling that  $D'_{i,n}$  is an identity matrix, meaning  $\|D'_{i,n}\|_s = 1$ . It holds that:

$$\begin{aligned} & \sum_{1 \leq j < j' \leq n} \|\Delta W_{j'}\|_s \|\Delta W_j\|_s \\ & \leq \sum_{1 \leq j < j' \leq n} \|\Delta W_{j'}\|_F \|\Delta W_j\|_F \\ & = \frac{1}{2} \left( \sum_{j=1}^n \|\Delta W_j\|_F \right)^2 - \frac{1}{2} \sum_{j=1}^n \|\Delta W_j\|_F^2 \\ & \leq \frac{n}{2} \sum_{j=1}^n \|\Delta W_j\|_F^2 - \frac{1}{2} \sum_{j=1}^n \|\Delta W_j\|_F^2 \\ & = \frac{n-1}{2} \sum_{j=1}^n \|\Delta W_j\|_F^2, \end{aligned}$$

where the last inequality follows from the fact that the one-norm of a vector in  $\mathbb{R}^n$  is never greater than  $\sqrt{n}$  times its euclidean-norm. This leads us to the following bound:

$$\begin{aligned} & \nabla^2 f(\boldsymbol{\theta})[\Delta W_1, \dots, \Delta W_n] \\ & \geq -\max\{|\alpha|, |\bar{\alpha}|\}^{n-1} \left( \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_s \right) \frac{n-1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \sum_{j=1}^n \|\Delta W_j\|_F^2. \end{aligned}$$

The desired result readily follows:

$$\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -\max\{|\alpha|, |\bar{\alpha}|\}^{n-1} \left( \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_F \right) \frac{n-1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2.$$

□

### J.15 Proof of Proposition 6

Recall that  $(W_1, W_2, \dots, W_n) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$  are the weight matrices constituting  $\boldsymbol{\theta} \in \mathbb{R}^d$ , and denote by  $(W_{1,s}, W_{2,s}, \dots, W_{n,s}) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$  those that constitute  $\boldsymbol{\theta}_s$ . For  $j, j' \in [n]$ :

$$\left| \|W_{j,s}\|_F^2 - \|W_{j',s}\|_F^2 \right| \leq \max\{ \|W_{j,s}\|_F^2, \|W_{j',s}\|_F^2 \} \leq \max_{j \in [n]} \|W_{j,s}\|_F^2 \leq \|\boldsymbol{\theta}_s\|_2^2 \leq \epsilon^2.$$

Corollary 2.1 from [18] implies that throughout a gradient flow trajectory differences between squared Frobenius norms of weight matrices are constant. Therefore, for  $j, j' \in [n]$ :

$$\left| \|W_j\|_F^2 - \|W_{j'}\|_F^2 \right| = \left| \|W_{j,s}\|_F^2 - \|W_{j',s}\|_F^2 \right| \leq \epsilon^2. \quad (81)$$

If the network is shallow (*i.e.*  $n=2$ ), then Equation (27) coincides with Equation (26), thus the desired result follows trivially from Lemma 5. Hereafter we assume that the network is deep (*i.e.*  $n \geq 3$ ). It holds that:

$$\begin{aligned} & \max_{\substack{\mathcal{J} \subseteq [n], |\mathcal{J}|=n-2}} \prod_{j \in \mathcal{J}} \|W_j\|_F \leq \max_{j \in [n]} \|W_j\|_F^{n-2} \\ & = \left( \min_{j \in [n]} \|W_j\|_F^2 + \max_{j \in [n]} \|W_j\|_F^2 - \min_{j \in [n]} \|W_j\|_F^2 \right)^{\frac{n-2}{2}} \\ & \leq \left( \min_{j \in [n]} \|W_j\|_F^2 + \epsilon^2 \right)^{\frac{n-2}{2}} \\ & = \left( \sqrt{\min_{j \in [n]} \|W_j\|_F^2 + \epsilon^2} \right)^{n-2} \\ & \leq \left( \min_{j \in [n]} \|W_j\|_F + \epsilon \right)^{n-2}, \end{aligned}$$

where the third transition follows from Equation (81) and the last transition follows from subadditivity of square root. Combining the latter inequality together with the result of Lemma 5 (Equation (26)), we obtain the desired result:

$$\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -\max\{|\alpha|, |\bar{\alpha}|\}^{n-1} \frac{n-1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \left( \min_{j \in [n]} \|W_j\|_F + \epsilon \right)^{n-2}.$$

□

## J.16 Proof of Lemma 6

This proof is very similar to that of Lemmas 1 and 4 (see Subappendixes J.4 and J.12 respectively). We repeat all details for completeness. Recall that  $\boldsymbol{\theta} \in \mathbb{R}^d$  is a concatenation of  $(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \dots \times \mathbb{R}^{d_n}$  as a vector. Let  $(\Delta \mathbf{w}_1, \Delta \mathbf{w}_2, \dots, \Delta \mathbf{w}_n) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \dots \times \mathbb{R}^{d_n}$ , and denote by  $\Delta \boldsymbol{\theta} \in \mathbb{R}^d$  its concatenation as a vector in corresponding order. Denote the following for  $i \in \{1, \dots, |\mathcal{S}|\}$ :

$$\begin{aligned} \Delta_i^{(1)} &:= \sum_{j=1}^n (D'_{i,*} W_*(\mathbf{w}_*))_{n:j+1} D'_{i,j}(W_j(\Delta \mathbf{w}_j)) (D'_{i,*} W_*(\mathbf{w}_*))_{j-1:1}, \\ \Delta_i^{(2)} &:= \sum_{1 \leq j < j' \leq n} (D'_{i,*} W_*(\mathbf{w}_*))_{n:j'+1} D'_{i,j'}(W_{j'}(\Delta \mathbf{w}_{j'})) (D'_{i,*} W_*(\mathbf{w}_*))_{j'-1:j+1} \\ &\quad D'_{i,j}(W_j(\Delta \mathbf{w}_j)) (D'_{i,*} W_*(\mathbf{w}_*))_{j-1:1}, \\ \Delta_i^{(3:n)} &:= D'_{i,n}(W_n(\mathbf{w}_n) + W_n(\Delta \mathbf{w}_n)) \cdots D'_{i,1}(W_1(\mathbf{w}_1) + W_1(\Delta \mathbf{w}_1)) \\ &\quad - (D'_{i,*} W_*(\mathbf{w}_*))_{n:1} - \Delta_i^{(1)} - \Delta_i^{(2)}. \end{aligned} \tag{82}$$

We now develop a second-order Taylor expansion of  $f(\boldsymbol{\theta})$ . Since the vector tuple corresponding to  $(\boldsymbol{\theta} + \Delta \boldsymbol{\theta})$  is  $((\mathbf{w}_1 + \Delta \mathbf{w}_1), \dots, (\mathbf{w}_n + \Delta \mathbf{w}_n))$ , and the function  $f(\cdot)$  coincides with the function given in Equation (29) on an open region containing  $\boldsymbol{\theta}$ , for sufficiently small  $\Delta \boldsymbol{\theta}$  we obtain:

$$\begin{aligned} f(\boldsymbol{\theta} + \Delta \boldsymbol{\theta}) &= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell \left( D'_{i,n}(W_n(\mathbf{w}_n + \Delta \mathbf{w}_n)) \cdots D'_{i,1}(W_1(\mathbf{w}_1 + \Delta \mathbf{w}_1)) \mathbf{x}_i, y_i \right) \\ &= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell \left( D'_{i,n}(W_n(\mathbf{w}_n) + W_n(\Delta \mathbf{w}_n)) \cdots D'_{i,1}(W_1(\mathbf{w}_1) + W_1(\Delta \mathbf{w}_1)) \mathbf{x}_i, y_i \right) \\ &= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell \left( ((D'_{i,*} W_*(\mathbf{w}_*))_{n:1} + \Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i, y_i \right) \\ &= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell \left( (D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i + (\Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i, y_i \right), \end{aligned} \tag{83}$$

where the second transition follows from linearity of  $W_j(\cdot)$  for  $j \in \{1, \dots, n\}$  and the third transition follows from the definition of  $\Delta_i^{(3:n)}$  (Equation (82)). Let  $\Delta \mathbf{v} \in \mathbb{R}^{d_n}$ . For every  $i \in \{1, \dots, |\mathcal{S}|\}$ , the second-order Taylor expansion of  $\ell(\cdot)$  with respect to its first argument at  $((D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i, y_i)$  is given by:

$$\begin{aligned} \ell((D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i + \Delta \mathbf{v}, y_i) &= \\ \ell((D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i, y_i) &+ \langle \nabla \ell_i, \Delta \mathbf{v} \rangle + \frac{1}{2} \nabla^2 \ell_i[\Delta \mathbf{v}] + \mathcal{O}(\|\Delta \mathbf{v}\|_2^2), \end{aligned} \tag{84}$$

where the  $\mathcal{O}(\cdot)$  notation refers to some expression satisfying  $\lim_{a \rightarrow 0} (\mathcal{O}(a)/a) = 0$ . We continue to develop Equation (83) using Equation (84):

$$\begin{aligned}
& f(\boldsymbol{\theta} + \Delta \boldsymbol{\theta}) \\
&= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \left( \ell((D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i, y_i) + \langle \nabla \ell_i, (\Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i \rangle + \right. \\
&\quad \left. \frac{1}{2} \nabla^2 \ell_i [(\Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i] + \mathcal{O}(\|(\Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i\|_2^2) \right) \\
&= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell((D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i, y_i) + \\
&\quad \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \langle \nabla \ell_i, \Delta_i^{(1)} \mathbf{x}_i \rangle + \langle \nabla \ell_i, \Delta_i^{(2)} \mathbf{x}_i \rangle + \langle \nabla \ell_i, \Delta_i^{(3:n)} \mathbf{x}_i \rangle + \\
&\quad \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \frac{1}{2} \nabla^2 \ell_i [\Delta_i^{(1)} \mathbf{x}_i] + \frac{1}{2} \nabla^2 \ell_i [(\Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i] + 2 \cdot \frac{1}{2} \nabla^2 \ell_i [\Delta_i^{(1)} \mathbf{x}_i, (\Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i] + \\
&\quad \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \mathcal{O}(\|(\Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i\|_2^2),
\end{aligned}$$

where in the last transition we view  $\nabla^2 \ell_i$  as both a quadratic and a bilinear form (see Subappendix J.1). Notice that  $\langle \nabla \ell_i, \Delta_i^{(3:n)} \mathbf{x}_i \rangle$ ,  $\frac{1}{2} \nabla^2 \ell_i [(\Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i]$ ,  $\nabla^2 \ell_i [\Delta_i^{(1)} \mathbf{x}_i, (\Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i]$  and  $\mathcal{O}(\|(\Delta_i^{(1)} + \Delta_i^{(2)} + \Delta_i^{(3:n)}) \mathbf{x}_i\|_2^2)$  are all  $\mathcal{O}(\|\Delta \boldsymbol{\theta}\|_2^2)$ , thus:

$$\begin{aligned}
& f(\boldsymbol{\theta} + \Delta \boldsymbol{\theta}) = \\
& \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell((D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i, y_i) + \langle \nabla \ell_i, \Delta_i^{(1)} \mathbf{x}_i \rangle + \langle \nabla \ell_i, \Delta_i^{(2)} \mathbf{x}_i \rangle + \frac{1}{2} \nabla^2 \ell_i [\Delta_i^{(1)} \mathbf{x}_i] + \mathcal{O}(\|\Delta \boldsymbol{\theta}\|_2^2).
\end{aligned}$$

This is in fact a Taylor expansion of the function  $f(\cdot)$  evaluated at the point  $\boldsymbol{\theta}$  with a constant term  $\frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \ell((D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i, y_i)$ , a linear term  $\frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \langle \nabla \ell_i, \Delta_i^{(1)} \mathbf{x}_i \rangle$ , a quadratic term  $\frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \langle \nabla \ell_i, \Delta_i^{(2)} \mathbf{x}_i \rangle + \frac{1}{2} \nabla^2 \ell_i [\Delta_i^{(1)} \mathbf{x}_i]$ , and a remainder term of  $\mathcal{O}(\|\Delta \boldsymbol{\theta}\|_2^2)$ . From uniqueness of the Taylor expansion, the quadratic term must be equal to  $\frac{1}{2} \nabla^2 f(\boldsymbol{\theta})[\Delta \mathbf{w}_1, \dots, \Delta \mathbf{w}_n]$ . This implies:

$$\begin{aligned}
& \nabla^2 f(\boldsymbol{\theta})[\Delta \mathbf{w}_1, \dots, \Delta \mathbf{w}_n] \\
&= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \left( \nabla^2 \ell_i [\Delta_i^{(1)} \mathbf{x}_i] + 2 \langle \nabla \ell_i, \Delta_i^{(2)} \mathbf{x}_i \rangle \right) \\
&= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i \left[ \sum_{j=1}^n (D'_{i,*} W_*(\mathbf{w}_*))_{n:j+1} D'_{i,j} (W_j(\Delta \mathbf{w}_j)) (D'_{i,*} W_*(\mathbf{w}_*))_{j-1:1} \mathbf{x}_i \right] + \\
&\quad \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i^\top \sum_{1 \leq j < j' \leq n} (D'_{i,*} W_*(\mathbf{w}_*))_{n:j'+1} D'_{i,j'} (W_{j'}(\Delta \mathbf{w}_{j'})) (D'_{i,*} W_*(\mathbf{w}_*))_{j'-1:1} \mathbf{x}_i,
\end{aligned}$$

where the last transition follows from plugging in the definitions of  $\Delta^{(1)}$  and  $\Delta^{(2)}$  (see Equation (82)).  $\square$

### J.17 Proof of Proposition 7

This proof is very similar to that of Proposition 5 (see Subappendix J.13). We repeat all details for completeness. From assumption (ii) there exists some  $\boldsymbol{\theta} \in \mathbb{R}^d$  such that  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_\theta(\mathbf{x}_i) \neq 0$ .

Define  $(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \dots \times \mathbb{R}^{d_n}$  to be the weight vectors constituting  $\theta$ . We may assume  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_\theta(\mathbf{x}_i) < 0$  without loss of generality, as we can negate the vectors  $h_\theta(\mathbf{x}_i) \in \mathbb{R}^{d_n}$  for all  $i \in \{1, 2, \dots, |\mathcal{S}|\}$  by flipping the signs of the entries in  $\theta$  corresponding to the last vector  $\mathbf{w}_n$  (see Equation (28)). From continuity, there exists a neighborhood  $\mathcal{N}$  of  $\theta$  such that for all  $\tilde{\theta} \in \mathcal{N}$  it holds that  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_{\tilde{\theta}}(\mathbf{x}_i) < 0$ . Moreover, as discussed in Appendix D, for almost all  $\theta' \in \mathbb{R}^d$  there exists an open region  $\mathcal{D}_{\theta'} \subseteq \mathbb{R}^d$  containing  $\theta'$ , which is closed under positive rescaling of weight matrices and across which  $f(\cdot)$  coincides with a function as given in Equation (29). There must exist some  $\theta'$  in the neighborhood  $\mathcal{N}$  for which a region of the type  $\mathcal{D}_{\theta'}$  exists. We may assume, without loss of generality, that  $\theta \in \mathcal{D}_{\theta'}$ . Notice that none of the weight vectors  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$  are equal to zero (as that would lead to  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_\theta(\mathbf{x}_i) = 0$ ). Define the following weight vectors parameterized by  $a > 0$  (while recalling that  $n \geq 3$  by assumption (i)):

$$\begin{aligned} \mathbf{w}_1(a) &:= \mathbf{w}_1 \cdot a^{-2} \in \mathbb{R}^{d_1}, \\ \mathbf{w}_2(a) &:= \mathbf{w}_2 \cdot a^{-2} \in \mathbb{R}^{d_2}, \\ \mathbf{w}_3(a) &:= \mathbf{w}_3 \cdot a \in \mathbb{R}^{d_3}, \\ \mathbf{w}_j(a) &:= \mathbf{w}_j \in \mathbb{R}^{d_j} \text{ for } j \in \{1, 2, \dots, n\} / \{1, 2, 3\}, \end{aligned}$$

and denote by  $\theta(a) \in \mathbb{R}^d$  their corresponding weight setting. Since  $\mathcal{D}_{\theta'}$  is closed under positive rescaling of weight vectors, it holds that  $\{\theta(a) : a > 0\} \subseteq \mathcal{D}_{\theta'}$ . Define:

$$\begin{aligned} \Delta \mathbf{w}_1 &:= \mathbf{w}_1 \in \mathbb{R}^{d_1}, \\ \Delta \mathbf{w}_2 &:= \mathbf{w}_2 \in \mathbb{R}^{d_2}, \\ \Delta \mathbf{w}_j &:= \mathbf{0} \in \mathbb{R}^{d_j} \text{ for } j \in [n] / \{1, 2\}. \end{aligned}$$

For  $a > 0$ ,  $i \in \{1, 2, \dots, |\mathcal{S}|\}$  and  $j, j' \in \{1, 2, \dots, n\}$ , define  $(D'_{i,*} W_*(\mathbf{w}_*(a)))_{j':j}$  to be the matrix  $D'_{i,j'} W_{j'}(\mathbf{w}_{j'}(a)) D'_{i,j'-1} W_{j'-1}(\mathbf{w}_{j'-1}(a)) \dots D'_{i,j} W_j(\mathbf{w}_j(a))$  (where by convention  $D'_{i,n} \in \mathbb{R}^{d_n \times d_n}$  stands for identity) if  $j \leq j'$ , and an identity matrix (with size to be inferred by context) otherwise. For  $i \in \{1, 2, \dots, |\mathcal{S}|\}$  and  $a > 0$  let  $\nabla \ell_i(a) \in \mathbb{R}^{d_n}$  and  $\nabla^2 \ell_i(a) \in \mathbb{R}^{d_n \times d_n}$  be the gradient and Hessian (respectively) of the loss  $\ell(\cdot)$  at the point  $((D'_{i,*} W_*(\mathbf{w}_*(a)))_{n:1} \mathbf{x}_i, y_i)$  with respect to its first argument. For every  $a > 0$ , since  $\theta(a) \in \mathcal{D}_{\theta'}$ , we may apply Lemma 6, obtaining:

$$\begin{aligned} \nabla^2 f(\theta(a))[\Delta \mathbf{w}_1, \dots, \Delta \mathbf{w}_n] &= \\ \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i(a) \left[ \sum_{j=1}^n (D'_{i,*} W_*(\mathbf{w}_*(a)))_{n:j+1} D'_{i,j} (W_j(\Delta \mathbf{w}_j)) (D'_{i,*} W_*(\mathbf{w}_*(a)))_{j-1:1} \mathbf{x}_i \right] &+ \\ \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i(a)^\top \sum_{\substack{j'=j+1 \\ 1 \leq j < j' \leq n}} (D'_{i,*} W_*(\mathbf{w}_*(a)))_{n:j'+1} D'_{i,j'} (W_{j'}(\Delta \mathbf{w}_{j'})) (D'_{i,*} W_*(\mathbf{w}_*(a)))_{j'-1:j+1} & \\ & D'_{i,j} (W_j(\Delta \mathbf{w}_j)) (D'_{i,*} W_*(\mathbf{w}_*(a)))_{j-1:1} \mathbf{x}_i, \end{aligned} \tag{85}$$

where we regard Hessians as quadratic forms (see Subappendix J.1). Plugging in the definitions of  $\mathbf{w}_j(a)$  and  $\Delta \mathbf{w}_j$  for  $j \in [n]$  and relying on linearity of  $W_j(\cdot)$  for  $j \in [n]$ , we have:

$$\begin{aligned} \nabla^2 f(\theta(a))[\Delta \mathbf{w}_1, \dots, \Delta \mathbf{w}_n] &= \\ &= \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i(a) \left[ 2a^{-1} (D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i \right] + \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i(a)^\top a (D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i \\ &= \frac{4}{a^2} \cdot \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i(a) \left[ (D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i \right] + a \cdot \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i(a)^\top h_\theta(\mathbf{x}_i), \end{aligned} \tag{86}$$

where the second transition follows from pulling  $2/a$  out of the quadratic operator and the fact that  $h_\theta(\mathbf{x}_i) = (D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i$ . Note that  $\lim_{a \rightarrow \infty} (D'_{i,*} W_*(\mathbf{w}_*(a)))_{n:1} = 0$ . Since the function  $\ell(\cdot)$  is twice continuously differentiable in its first argument, it holds that  $\lim_{a \rightarrow \infty} \nabla^2 \ell_i(a) = \lim_{a \rightarrow \infty} \nabla^2 \ell((D'_{i,*} W_*(\mathbf{w}_*(a)))_{n:1} \mathbf{x}_i, y_i) = \nabla^2 \ell(\mathbf{0}, y_i)$ , and similarly we

have that  $\lim_{a \rightarrow \infty} \nabla \ell_i(a) = \lim_{a \rightarrow \infty} \nabla \ell((D'_{i,*} W_*(\mathbf{w}_*(a)))_{n:1} \mathbf{x}_i, y_i) = \nabla \ell(\mathbf{0}, y_i)$ . Therefore, in the limit  $a \rightarrow \infty$ , Equation (86) becomes:

$$\begin{aligned}
& \lim_{a \rightarrow \infty} \left( \nabla^2 f(\boldsymbol{\theta}(a)) [\Delta \mathbf{w}_1, \dots, \Delta \mathbf{w}_n] \right) \\
&= \lim_{a \rightarrow \infty} \left( \frac{4}{a^2} \cdot \frac{4}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i(a) \left[ (D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i \right] \right) + \lim_{a \rightarrow \infty} \left( a \cdot \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i(a)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) \right) \\
&= \lim_{a \rightarrow \infty} \left( \frac{4}{a^2} \right) \lim_{a \rightarrow \infty} \left( \frac{4}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i(a) \left[ (D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i \right] \right) + \lim_{a \rightarrow \infty} \left( a \lim_{a \rightarrow \infty} \left( \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i(a)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) \right) \right) \\
&= 0 \cdot \left( \frac{4}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell(\mathbf{0}, y_i) \left[ (D'_{i,*} W_*(\mathbf{w}_*))_{n:1} \mathbf{x}_i \right] \right) + \lim_{a \rightarrow \infty} \left( a \cdot \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i(\mathbf{0}, y_i)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) \right) \\
&= -\infty,
\end{aligned}$$

where the second transition is valid since the multiplied limits are finite and the limit inside a limit is non-zero, and the last transition follows from  $\sum_{i=1}^{|\mathcal{S}|} \nabla \ell(\mathbf{0}, y_i)^\top h_{\boldsymbol{\theta}}(\mathbf{x}_i) < 0$ . Notice that the vectors  $\Delta \mathbf{w}_1, \Delta \mathbf{w}_2, \dots, \Delta \mathbf{w}_n$  are independent of  $a$ , thus it must hold that  $\lim_{a \rightarrow \infty} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta}(a))) = -\infty$ . This in particular implies the desired result:

$$\inf_{\boldsymbol{\theta} \in \mathbb{R}^d \text{ s.t. } \nabla^2 f(\boldsymbol{\theta}) \text{ exists}} \lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) = -\infty.$$

□

### J.18 Proof of Lemma 7

This proof is very similar to that of Lemmas 2 and 5 (see Subappendixes J.6 and J.14 respectively). We repeat all details for completeness. Recall that  $\boldsymbol{\theta} \in \mathbb{R}^d$  is a concatenation of  $(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \dots \times \mathbb{R}^{d_n}$  as a vector. Let  $(\Delta \mathbf{w}_1, \Delta \mathbf{w}_2, \dots, \Delta \mathbf{w}_n) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \dots \times \mathbb{R}^{d_n}$ , and denote by  $\Delta \boldsymbol{\theta} \in \mathbb{R}^d$  its concatenation as a vector in corresponding order. As shown in Lemma 6:

$$\begin{aligned}
& \nabla^2 f(\boldsymbol{\theta}) [\Delta \mathbf{w}_1, \dots, \Delta \mathbf{w}_n] = \\
& \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla^2 \ell_i \left[ \sum_{j=1}^n (D'_{i,*} W_*(\mathbf{w}_*))_{n:j+1} D'_{i,j} (W_j(\Delta \mathbf{w}_j)) (D'_{i,*} W_*(\mathbf{w}_*))_{j-1:1} \mathbf{x}_i \right] + \\
& \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i^\top \sum_{1 \leq j < j' \leq n} (D'_{i,*} W_*(\mathbf{w}_*))_{n:j'+1} D'_{i,j'} (W_{j'}(\Delta \mathbf{w}_{j'})) (D'_{i,*} W_*(\mathbf{w}_*))_{j'-1:j+1} \cdot \\
& \qquad \qquad \qquad D'_{i,j} (W_j(\Delta \mathbf{w}_j)) (D'_{i,*} W_*(\mathbf{w}_*))_{j-1:1} \mathbf{x}_i,
\end{aligned}$$

where we regard Hessians as quadratic forms (see Subappendix J.1). Convexity of  $\ell(\cdot)$  in its first argument implies that for  $i \in \{1, 2, \dots, |\mathcal{S}|\}$ ,  $\nabla^2 \ell_i$  is positive semi-definite, thus:

$$\begin{aligned}
& \nabla^2 f(\boldsymbol{\theta}) [\Delta \mathbf{w}_1, \dots, \Delta \mathbf{w}_n] \geq \\
& \frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \nabla \ell_i^\top \sum_{1 \leq j < j' \leq n} (D'_{i,*} W_*(\mathbf{w}_*))_{n:j'+1} D'_{i,j'} (W_{j'}(\Delta \mathbf{w}_{j'})) (D'_{i,*} W_*(\mathbf{w}_*))_{j'-1:j+1} \cdot \\
& \qquad \qquad \qquad D'_{i,j} (W_j(\Delta \mathbf{w}_j)) (D'_{i,*} W_*(\mathbf{w}_*))_{j-1:1} \mathbf{x}_i.
\end{aligned}$$



Applying Cauchy-Schwarz and triangle inequalities, we get:

$$\begin{aligned}
& \nabla^2 f(\boldsymbol{\theta})[\Delta \mathbf{w}_1, \dots, \Delta \mathbf{w}_n] \\
& \geq -\frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \cdot \sum_{1 \leq j < j' \leq n} \|(D'_{i,*} W_*(\mathbf{w}_*))_{n:j'+1} D'_{i,j'}(W_{j'}(\Delta \mathbf{w}_{j'}))(D'_{i,*} W_*(\mathbf{w}_*))_{j'-1:j+1} \cdot \\
& \qquad \qquad \qquad D'_{i,j}(W_j(\Delta \mathbf{w}_j))(D'_{i,*} W_*(\mathbf{w}_*))_{j-1:1} \mathbf{x}_i\|_2 \\
& \geq -\frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \cdot \sum_{1 \leq j < j' \leq n} \|W_j(\Delta \mathbf{w}_j)\|_s \|W_{j'}(\Delta \mathbf{w}_{j'})\|_s \prod_{k \in [n]/\{j,j'\}} \|W_k(\mathbf{w}_k)\|_s \prod_{k \in [n]} \|D'_{i,k}\|_s \cdot \|\mathbf{x}_i\|_2 \\
& \geq -\frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \cdot \prod_{j=1}^n \|W_j(\cdot)\|_{op} \sum_{1 \leq j < j' \leq n} \|\Delta \mathbf{w}_j\|_2 \|\Delta \mathbf{w}_{j'}\|_2 \prod_{k \in [n]/\{j,j'\}} \|\mathbf{w}_k\|_2 \prod_{k=1}^n \|D'_{i,k}\|_s \cdot \|\mathbf{x}_i\|_2 \\
& \geq -\frac{2}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \prod_{j=1}^n \|W_j(\cdot)\|_{op} \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2^j \in \mathcal{J}}} \prod \|\mathbf{w}_j\|_2 \max\{|\alpha|, |\bar{\alpha}|\}^{n-1} \|\mathbf{x}_i\|_2 \sum_{1 \leq j < j' \leq n} \|\Delta \mathbf{w}_j\|_2 \|\Delta \mathbf{w}_{j'}\|_2,
\end{aligned}$$

where the second transition follows from the definition and sub-multiplicativity of spectral norm, the third transition follows from bounding spectral norms with Frobenius norms and the definition of  $\|W_j(\cdot)\|_{op}$ , and the last transition follows from maximizing  $\prod_{k \in [n]/\{j,j'\}} \|\mathbf{w}_k\|_2$  over  $j, j'$ , upper bounding  $\|D'_{i,j}\|_s \leq \max\{|\alpha|, |\bar{\alpha}|\}$  for  $j \in [n-1]$  and recalling that  $D'_{i,n}$  is an identity matrix, meaning  $\|D'_{i,n}\|_s = 1$ . It holds that:

$$\begin{aligned}
& \sum_{1 \leq j < j' \leq n} \|\Delta \mathbf{w}_{j'}\|_2 \|\Delta \mathbf{w}_j\|_2 \\
& = \frac{1}{2} \left( \sum_{j=1}^n \|\Delta \mathbf{w}_j\|_2 \right)^2 - \frac{1}{2} \sum_{j=1}^n \|\Delta \mathbf{w}_j\|_2^2 \\
& \leq \frac{n}{2} \sum_{j=1}^n \|\Delta \mathbf{w}_j\|_2^2 - \frac{1}{2} \sum_{j=1}^n \|\Delta \mathbf{w}_j\|_2^2 \\
& = \frac{n-1}{2} \sum_{j=1}^n \|\Delta \mathbf{w}_j\|_2^2,
\end{aligned}$$

where the last inequality follows from the fact that the one-norm of a vector in  $\mathbb{R}^n$  is never greater than  $\sqrt{n}$  times its euclidean-norm. This leads us to the following bound:

$$\begin{aligned}
& \nabla^2 f(\boldsymbol{\theta})[\Delta \mathbf{w}_1, \dots, \Delta \mathbf{w}_n] \geq \\
& -\max\{|\alpha|, |\bar{\alpha}|\}^{n-1} \prod_{j=1}^n \|W_j(\cdot)\|_{op} \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2^j \in \mathcal{J}}} \prod \|\mathbf{w}_j\|_2 \frac{n-1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \sum_{j=1}^n \|\Delta \mathbf{w}_j\|_2^2.
\end{aligned}$$

The desired result readily follows:

$$\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -\max\{|\alpha|, |\bar{\alpha}|\}^{n-1} \frac{n-1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \prod_{j=1}^n \|W_j(\cdot)\|_{op} \max_{\substack{\mathcal{J} \subseteq [n] \\ |\mathcal{J}|=n-2^j \in \mathcal{J}}} \prod \|\mathbf{w}_j\|_2.$$

□

## J.19 Proof of Proposition 8

This proof is very similar to that of Proposition 6 (see Subappendix J.15). Recall that  $(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n) \in \mathbb{R}^{d'_1} \times \mathbb{R}^{d'_2} \times \dots \times \mathbb{R}^{d'_n}$  are the weight vectors constituting  $\boldsymbol{\theta} \in \mathbb{R}^d$ , and denote by  $(\mathbf{w}_{1,s}, \mathbf{w}_{2,s}, \dots, \mathbf{w}_{n,s}) \in \mathbb{R}^{d'_1} \times \mathbb{R}^{d'_2} \times \dots \times \mathbb{R}^{d'_n}$  those that constitute  $\boldsymbol{\theta}_s$ . For  $j, j' \in [n]$ :

$$\left| \|\mathbf{w}_{j,s}\|_2^2 - \|\mathbf{w}_{j',s}\|_2^2 \right| \leq \max\{\|\mathbf{w}_{j,s}\|_2^2, \|\mathbf{w}_{j',s}\|_2^2\} \leq \max_{j \in [n]} \|\mathbf{w}_{j,s}\|_2^2 \leq \|\boldsymbol{\theta}_s\|_2^2 \leq \epsilon^2.$$

Theorem 2.3 from [18] implies that throughout a gradient flow trajectory differences between squared Euclidean norms of weight vectors are constant. Therefore, for  $j, j' \in [n]$ :

$$\left| \|\mathbf{w}_j\|_2^2 - \|\mathbf{w}_{j'}\|_2^2 \right| = \left| \|\mathbf{w}_{j,s}\|_2^2 - \|\mathbf{w}_{j',s}\|_2^2 \right| \leq \epsilon^2. \tag{87}$$

If the network is shallow (*i.e.*  $n = 2$ ), then Equation (32) coincides with Equation (31), thus the desired result follows trivially from Lemma 7. Hereafter we assume that the network is deep (*i.e.*  $n \geq 3$ ). It holds that:

$$\begin{aligned}
\max_{\mathcal{J} \subseteq [n], |\mathcal{J}|=n-2} \prod_{j \in \mathcal{J}} \|\mathbf{w}_j\|_2 &\leq \max_{j \in [n]} \|\mathbf{w}_j\|_2^{n-2} \\
&= \left( \min_{j \in [n]} \|\mathbf{w}_j\|_2^2 + \max_{j \in [n]} \|\mathbf{w}_j\|_2^2 - \min_{j \in [n]} \|\mathbf{w}_j\|_2^2 \right)^{\frac{n-2}{2}} \\
&\leq \left( \min_{j \in [n]} \|\mathbf{w}_j\|_2^2 + \epsilon^2 \right)^{\frac{n-2}{2}} \\
&= \left( \sqrt{\min_{j \in [n]} \|\mathbf{w}_j\|_2^2 + \epsilon^2} \right)^{n-2} \\
&\leq \left( \min_{j \in [n]} \|\mathbf{w}_j\|_2 + \epsilon \right)^{n-2},
\end{aligned}$$

where the third transition follows from Equation (87) and the last transition follows from subadditivity of square root. Combining the latter inequality together with the result of Lemma 7 (Equation (31)), we obtain the desired result:

$$\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta})) \geq -\max\{|\alpha|, |\bar{\alpha}|\}^{n-1} \frac{n-1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \|\nabla \ell_i\|_2 \|\mathbf{x}_i\|_2 \prod_{j \in [n]} \|W_j(\cdot)\|_{op} \left( \min_{j \in [n]} \|\mathbf{w}_j\|_2 + \epsilon \right)^{n-2}.$$

□

## J.20 Proof of Lemma 8

In this proof we overload the definition of unbalancedness magnitude (Definition 1) to account for arbitrary matrix dimensions, namely, for any matrices  $A_1, \dots, A_n$  such that the product  $A_n \cdots A_1$  is defined, we refer to  $\max_{j \in [n-1]} \|A_{j+1}^\top A_{j+1} - A_j A_j^\top\|_n$  as their unbalancedness magnitude. Recall that  $\boldsymbol{\theta} \in \mathbb{R}^d$  is the arrangement of  $W_1, W_2, \dots, W_{n-1} \in \mathbb{R}^{d_0, d_0}$  and  $W_n \in \mathbb{R}^{d_n, d_0}$  as a vector. Define the matrices  $B_1, B_2, \dots, B_n \in \mathbb{R}^{d_0, d_0}$  as follows:  $B_j := W_j$  for  $j \in [n-1]$  and  $B_n := \sqrt{W_n^\top W_n}$ . Notice that:

$$B_n^\top B_n = \sqrt{W_n^\top W_n} \sqrt{W_n^\top W_n} = W_n^\top W_n,$$

thus the unbalancedness magnitude of  $B_1, \dots, B_n$  is equal to that of  $W_1, \dots, W_n$ , *i.e.* to  $\hat{\epsilon}$ . Define the matrices  $C_1, C_2, \dots, C_n \in \mathbb{R}^{d_0, d_0}$  by transposing and reversing the order of  $B_1, \dots, B_n$ , formally:  $C_j := B_{n-j+1}^\top$  for  $j \in [n]$ . Notice that transposition and order reversal do not change the unbalancedness magnitude. Namely, since for  $j \in [n-1]$  we have that  $\|C_{j+1}^\top C_{j+1} - C_j C_j^\top\|_n = \|B_{n-j} B_{n-j}^\top - B_{n-j+1}^\top B_{n-j+1}\|_n$ , the unbalancedness magnitude of  $C_1, \dots, C_n$  is equal to that of  $B_1, \dots, B_n$ , *i.e.* to  $\hat{\epsilon}$ . Applying Lemma 1 from [46] to  $C_1, \dots, C_n$ , we conclude that there exists  $\hat{C}_1, \dots, \hat{C}_n \in \mathbb{R}^{d_0, d_0}$  which are balanced (*i.e.* have unbalancedness magnitude zero), such that  $\|C_j - \hat{C}_j\|_F \leq (j-1)\sqrt{\hat{\epsilon}}$  for  $j \in [n]$ . Pay special notice to the fact that  $\hat{C}_1 = C_1$  (as the Frobenius norm of the discrepancy is zero). Define the matrices  $\hat{B}_1, \hat{B}_2, \dots, \hat{B}_n \in \mathbb{R}^{d_0, d_0}$  by transposing and reversing the order of  $\hat{C}_1, \dots, \hat{C}_n$ , formally:  $\hat{B}_j := \hat{C}_{n-j+1}^\top$  for  $j \in [n]$ . Relying again on the fact that transposition and order reversal do not change unbalancedness magnitude, we have that  $\hat{B}_1, \dots, \hat{B}_n$ , similarly to  $\hat{C}_1, \dots, \hat{C}_n$ , are balanced. Define the matrices  $\hat{W}_1, \hat{W}_2, \dots, \hat{W}_{n-1} \in \mathbb{R}^{d_0, d_0}$  and  $\hat{W}_n \in \mathbb{R}^{d_n, d_0}$  as follows:  $\hat{W}_j := \hat{B}_j$  for  $j \in [n-1]$  and  $\hat{W}_n := W_n$ . Notice that the dimensions of  $\hat{W}_1, \hat{W}_2, \dots, \hat{W}_n$  correspond to those of  $W_1, W_2, \dots, W_n$ , and in particular that these are valid weight matrices. We denote their corresponding weight setting by  $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ . Notice that:

$$\hat{W}_n^\top \hat{W}_n = W_n^\top W_n = \sqrt{W_n^\top W_n} \sqrt{W_n^\top W_n} = B_n B_n = C_1 C_1 = \hat{C}_1 \hat{C}_1 = \hat{B}_n \hat{B}_n,$$

which means that  $\hat{W}_1, \hat{W}_2, \dots, \hat{W}_n$  are balanced, as they have the same unbalancedness magnitude of  $\hat{B}_1, \dots, \hat{B}_n$ , *i.e.* zero. Furthermore we have that:

$$\begin{aligned}
\|\hat{\theta} - \theta\|_2 &= \|(\hat{W}_n, \hat{W}_{n-1}, \dots, \hat{W}_1) - (W_n, W_{n-1}, \dots, W_1)\|_F \\
&= \|(W_n, \hat{B}_{n-1}, \dots, \hat{B}_1) - (W_n, B_{n-1}, \dots, B_1)\|_F \\
&= \sqrt{\|W_n - W_n\|_F^2 + \|\hat{B}_{n-1} - B_{n-1}\|_F^2 + \dots + \|\hat{B}_1 - B_1\|_F^2} \\
&= \sqrt{0 + \|\hat{C}_2^\top - C_2^\top\|_F^2 + \dots + \|\hat{C}_n^\top - C_n^\top\|_F^2} \\
&\leq \sqrt{(n-1) \cdot (n-1)^2 \tilde{\epsilon}} \\
&\leq n^{1.5} \sqrt{\tilde{\epsilon}},
\end{aligned}$$

where the second transition follows from the definitions of  $\hat{W}_1, \dots, \hat{W}_{n-1}$  and  $B_1, \dots, B_{n-1}$ , the third from the definition of Frobenius norm, the fourth from the definitions of  $\hat{B}_1, \dots, \hat{B}_{n-1}$  and  $C_1, \dots, C_{n-1}$  and the fifth transition follows from the conclusion of Lemma 1 from [46] applied to  $C_1, \dots, C_n$ .  $\square$

### J.21 Proof of Theorem 5

Without loss of generality, we may assume  $\tilde{\epsilon} \leq 1$  (a proof that is valid for  $\tilde{\epsilon} = 1$  automatically accounts for  $\tilde{\epsilon} > 1$  as well). Given that the unbalancedness magnitude (Definition 1) of  $\theta_0$  is no greater than  $\hat{\epsilon}$  (defined in Equation (34)), by Lemma 8, there exists a weight setting  $\hat{\theta}_0 \in \mathbb{R}^d$  which is balanced (has unbalancedness magnitude zero) and meets  $\|\theta_0 - \hat{\theta}_0\|_2 \leq n^{1.5} \sqrt{\tilde{\epsilon}}$ . Denote by  $(\hat{W}_{1,0}, \hat{W}_{2,0}, \dots, \hat{W}_{n,0}) \in \mathbb{R}^{d_1, d_0} \times \mathbb{R}^{d_2, d_1} \times \dots \times \mathbb{R}^{d_n, d_{n-1}}$  the weight matrices corresponding to  $\hat{\theta}_0$ , and by  $\hat{W}_{n:1,0} \in \mathbb{R}^{d_n, d_0}$  its end-to-end matrix (*i.e.*  $\hat{W}_{n:1,0} := \hat{W}_{n,0} \hat{W}_{n-1,0} \dots \hat{W}_{1,0}$ ). Define  $\hat{\nu}$  as  $\text{Tr}(\Lambda_{yx}^\top \hat{W}_{n:1,0}) / (\|\Lambda_{yx}\|_F \|\hat{W}_{n:1,0}\|_F)$  if  $\|\hat{W}_{n:1,0}\|_F \neq 0$ , and as 0 otherwise. The following lemma establishes several bounds relating  $\hat{W}_{n:1,0}$  and  $\hat{\nu}$  to  $W_{n:1,0}$  and  $\nu$  respectively.

**Lemma 52.** *The following hold:*

$$\|W_{n:1,0} - \hat{W}_{n:1,0}\|_F \leq \frac{1}{3} \|W_{n:1,0}\|_F; \quad (88)$$

$$\hat{\nu} \geq \min\left\{-\frac{1}{2}, \text{sign}(\nu) \frac{|\nu|+1}{2}\right\}; \quad (89)$$

$$\|\hat{W}_{n:1,0}\|_F^{-1} \leq \frac{3}{2} \|W_{n:1,0}\|_F^{-1}; \text{ and} \quad (90)$$

$$\max\left\{1, \frac{1-\hat{\nu}}{1+\hat{\nu}}\right\} \leq \max\left\{3, \frac{3-\nu}{1+\nu}\right\}. \quad (91)$$

Proof for Lemma 52 is provided in Subsubappendix J.21.1.

Given  $\eta > 0$  adhering to Equation (35), define:

$$k := \left\lceil \frac{2n \left(\max\left\{1, \frac{3}{2} \cdot \frac{1-\hat{\nu}}{1+\hat{\nu}}\right\}\right)^n \ln\left(\frac{15n \max\left\{1, \frac{1-\hat{\nu}}{1+\hat{\nu}}\right\}}{\|\hat{W}_{n:1,0}\|_F \tilde{\epsilon}}\right)}{\|\hat{W}_{n:1,0}\|_F \eta} + 1 \right\rceil. \quad (92)$$

Taken together, Equations (90) and (91) imply that  $k$  adheres to the upper bound in Equation (36). It thus suffices to show that with step size  $\eta$ , iterate  $k$  of gradient descent is  $\tilde{\epsilon}$ -optimal, *i.e.*  $f(\theta_k) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \leq \tilde{\epsilon}$ .

Equations (88) and (89) respectively imply that  $\|\hat{W}_{n:1,0}\|_F \leq 0.2$  and  $\hat{\nu} \neq -1$ . Therefore, as an initial point for gradient flow, the (balanced) weight setting  $\hat{\theta}_0$  satisfies the conditions of Proposition 3. Define:

$$\bar{\epsilon} := \tilde{\epsilon}/2, \quad \epsilon := \frac{\|\hat{W}_{n:1,0}\|_F \bar{\epsilon}}{15n^3 \left(\max\left\{1, \frac{3}{2} \cdot \frac{1-\hat{\nu}}{1+\hat{\nu}}\right\}\right)^n k \eta}, \quad (93)$$

and invoke Proposition 3 with initial point  $\theta_s = \hat{\theta}_0$ , time  $t = k\eta$  and  $\bar{\epsilon}, \epsilon$  as above (note that  $\epsilon \in (0, 1/(2n))$ ). From the proposition we obtain that the gradient flow trajectory emanating from  $\hat{\theta}_0$

is defined over infinite time, and with  $\hat{\boldsymbol{\theta}}: [0, \infty) \rightarrow \mathbb{R}^d$  representing this trajectory, the following time  $\bar{t}$  satisfies  $f(\hat{\boldsymbol{\theta}}(\bar{t})) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \leq \bar{\epsilon}$ :

$$\bar{t} = \frac{2n \left( \max \left\{ 1, \frac{3}{2} \cdot \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^n}{\|\hat{W}_{n:1,0}\|_F} \ln \left( \frac{15n \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\}}{\|\hat{W}_{n:1,0}\|_F \min \{1, 2\bar{\epsilon}\}} \right). \quad (94)$$

Moreover, we obtain that under the notations of Theorem 3, in correspondence with  $\mathcal{D}_{k\eta, \epsilon}$  ( $\epsilon$ -neighborhood of gradient flow trajectory up to time  $k\eta$ ) are the smoothness and Lipschitz constants  $\beta_{k\eta, \epsilon} = 16n$  and  $\gamma_{k\eta, \epsilon} = 6\sqrt{n}$  respectively, and the following (upper) bound on the integral of (minus) the minimal eigenvalue of the Hessian:

$$\int_0^{k\eta} m(t') dt' \leq \frac{15n^3 \left( \max \left\{ 1, \frac{3}{2} \cdot \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^n k\eta \epsilon}{\|\hat{W}_{n:1,0}\|_F} + \ln \left( \frac{n^2 \left( e^2 \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{5(n-1)/2}}{\|\hat{W}_{n:1,0}\|_F^2} \right), \quad (95)$$

where the function  $m: [0, k\eta] \rightarrow \mathbb{R}$  is non-negative.

Notice that  $k = \lfloor \bar{t}/\eta + 1 \rfloor$  and therefore  $k\eta \geq \bar{t}$ . Combining this with the fact that the gradient flow trajectory  $\hat{\boldsymbol{\theta}}(\cdot)$  is  $\bar{\epsilon}$ -optimal at time  $\bar{t}$ , and that in general gradient flow monotonically non-increases the objective it optimizes, we infer  $\bar{\epsilon}$ -optimality of the gradient flow trajectory at time  $k\eta$ , i.e.  $\hat{\boldsymbol{\theta}}(k\eta) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \leq \bar{\epsilon}$ . We will invoke Theorem 3 for showing that, in addition to being  $\bar{\epsilon}$ -optimal, the gradient flow trajectory at time  $k\eta$  is also  $\epsilon$ -approximated by iterate  $k$  of gradient descent, i.e.  $\|\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}(k\eta)\|_2 \leq \epsilon$ . This, along with  $f(\cdot)$  being  $6\sqrt{n}$ -Lipschitz across  $\mathcal{D}_{k\eta, \epsilon}$  ( $\epsilon$ -neighborhood of gradient flow trajectory up to time  $k\eta$ ), yields the desired result —  $\bar{\epsilon}$ -optimality for iterate  $k$  of gradient descent:

$$\begin{aligned} & f(\boldsymbol{\theta}_k) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \\ &= \left( f(\boldsymbol{\theta}_k) - f(\hat{\boldsymbol{\theta}}(k\eta)) \right) + \left( f(\hat{\boldsymbol{\theta}}(k\eta)) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \right) \\ &\leq \left( 6\sqrt{n} \|\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}(k\eta)\|_2 \right) + \left( f(\hat{\boldsymbol{\theta}}(k\eta)) - \min_{\mathbf{q} \in \mathbb{R}^d} f(\mathbf{q}) \right) \\ &\leq 6\sqrt{n} \cdot \epsilon + \bar{\epsilon} \\ &\leq \tilde{\epsilon}, \end{aligned}$$

where the last transition follows from the definitions of  $\epsilon$  and  $\bar{\epsilon}$  (Equation (93)).

We conclude the proof by showing that indeed  $\|\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}(k\eta)\|_2 \leq \epsilon$ . Equation (95), the definition of  $\epsilon$  (Equation (93)) and the condition  $\tilde{\epsilon} \leq 1$  together imply:

$$\begin{aligned} \int_0^{k\eta} m(t') dt' &\leq \frac{15n^3 \left( \max \left\{ 1, \frac{3}{2} \cdot \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^n k\eta \epsilon}{\|\hat{W}_{n:1,0}\|_F} + \ln \left( \frac{n^2 \left( e^2 \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{5(n-1)/2}}{\|\hat{W}_{n:1,0}\|_F^2} \right) \quad (96) \\ &= \tilde{\epsilon} + \ln \left( \frac{n^2 \left( e^2 \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{5(n-1)/2}}{\|\hat{W}_{n:1,0}\|_F^2} \right) \\ &\leq 1 + \ln \left( \frac{n^2 \left( e^2 \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{5(n-1)/2}}{\|\hat{W}_{n:1,0}\|_F^2} \right) \\ &< \ln \left( \frac{3n^2 \left( e^2 \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{5(n-1)/2}}{\|\hat{W}_{n:1,0}\|_F^2} \right). \end{aligned}$$

Recalling the expressions for  $k$  and  $\bar{t}$  (Equations (92) and (94) respectively), and the definition of  $\bar{\epsilon}$  (Equation (93)), we have:

$$\begin{aligned} k\eta = \lfloor \bar{t}/\eta + 1 \rfloor \eta &\leq \bar{t} + \eta = \frac{2n \left( \max \left\{ 1, \frac{3}{2} \cdot \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^n}{\|\hat{W}_{n:1,0}\|_F} \ln \left( \frac{15n \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\}}{\|\hat{W}_{n:1,0}\|_F \bar{\epsilon}} \right) + \eta \quad (97) \\ &< \frac{3n \left( \max \left\{ 1, \frac{3}{2} \cdot \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^n}{\|\hat{W}_{n:1,0}\|_F} \ln \left( \frac{15n \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\}}{\|\hat{W}_{n:1,0}\|_F \bar{\epsilon}} \right), \end{aligned}$$

where the last transition makes use of the upper bound on  $\eta$  given in Equation (35). It holds that:

$$\begin{aligned}
& 4n^3 \epsilon^{-2} e^{2 \int_0^{k\eta} m(t') dt'} \\
& < 4n^3 \frac{225n^6 \left( \max \left\{ 1, \frac{3}{2}, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{2n} (k\eta)^2}{\|\hat{W}_{n:1}(0)\|_F^2 \tilde{\epsilon}^2} \cdot \frac{9n^4 \left( e^2 \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{5(n-1)}}{\|\hat{W}_{n:1}(0)\|_F^4} \\
& < \frac{8100n^{13} e^{11n-10} \left( \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{7n-5}}{\|\hat{W}_{n:1}(0)\|_F^6 \tilde{\epsilon}^2} (k\eta)^2 \\
& < \frac{8100n^{13} e^{11n-10} \left( \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{7n-5}}{\|\hat{W}_{n:1}(0)\|_F^6 \tilde{\epsilon}^2} \cdot \frac{9n^2 \left( \max \left\{ 1, \frac{3}{2}, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{2n}}{\|\hat{W}_{n:1}(0)\|_F^2} \left( \ln \left( \frac{15n \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\}}{\|\hat{W}_{n:1}(0)\|_F \tilde{\epsilon}} \right) \right)^2 \\
& < \frac{n^{15} e^{12n+2} \left( \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{9n-5}}{\|\hat{W}_{n:1}(0)\|_F^8 \tilde{\epsilon}^2} \left( \ln \left( \frac{15n \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\}}{\|\hat{W}_{n:1}(0)\|_F \tilde{\epsilon}} \right) \right)^2 \\
& \leq \frac{n^{15} e^{12n+2} \left( \max \left\{ 3, \frac{3-\nu}{1+\nu} \right\} \right)^{9n-5}}{\left(\frac{2}{3}\right)^8 \|\hat{W}_{n:1}(0)\|_F^8 \tilde{\epsilon}^2} \left( \ln \left( \frac{15n \max \left\{ 3, \frac{3-\nu}{1+\nu} \right\}}{\frac{2}{3} \|\hat{W}_{n:1}(0)\|_F \tilde{\epsilon}} \right) \right)^2 \\
& \leq \frac{n^{15} e^{12n+6} \left( \max \left\{ 3, \frac{3-\nu}{1+\nu} \right\} \right)^{9n-5}}{\|\hat{W}_{n:1}(0)\|_F^8 \tilde{\epsilon}^2} \left( \ln \left( \frac{23n \max \left\{ 3, \frac{3-\nu}{1+\nu} \right\}}{\|\hat{W}_{n:1}(0)\|_F \tilde{\epsilon}} \right) \right)^2 \\
& = 1/\hat{\epsilon},
\end{aligned}$$

where the first transition follows from Equation (96) and the definition of  $\epsilon$  (Equation (93)); the third makes use of Equation (97); the fifth relies on Equations (90) and (91); and the last is based on the definition of  $\hat{\epsilon}$  (Equation (34)) and the condition  $\tilde{\epsilon} \leq 1$ . Rearranging the derived inequality gives  $\sqrt{\hat{\epsilon}} < \frac{1}{2} n^{-1.5} \epsilon e^{-\int_0^{k\eta} m(t') dt'}$ . Combining this with the fact that  $\|\theta_0 - \hat{\theta}(0)\|_2 \leq n^{1.5} \sqrt{\hat{\epsilon}}$ , we obtain:

$$\epsilon - e^{\int_0^{k\eta} m(t') dt'} \|\theta_0 - \hat{\theta}(0)\|_2 \geq \epsilon - e^{\int_0^{k\eta} m(t') dt'} n^{1.5} \sqrt{\hat{\epsilon}} > \epsilon - \frac{1}{2} \epsilon = \frac{1}{2} \epsilon. \quad (98)$$

We now have:

$$\begin{aligned}
& \beta_{k\eta, \epsilon} \gamma_{k\eta, \epsilon} k\eta e^{\int_0^{k\eta} m(t') dt'} / \left( \epsilon - e^{\int_0^{k\eta} m(t') dt'} \|\theta_0 - \hat{\theta}(0)\|_2 \right) \\
& < \beta_{k\eta, \epsilon} \gamma_{k\eta, \epsilon} k\eta e^{\int_0^{k\eta} m(t') dt'} 2\epsilon^{-1} \\
& < (16n)(6\sqrt{n})k\eta \cdot \frac{3n^2 \left( e^2 \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{5(n-1)/2}}{\|\hat{W}_{n:1,0}\|_F^2} \cdot 2 \frac{15n^3 \left( \max \left\{ 1, \frac{3}{2}, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^n k\eta}{\|\hat{W}_{n:1,0}\|_F \tilde{\epsilon}} \\
& < \frac{9000n^{13/2} e^{6n-5} \left( \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{(7n-5)/2}}{\|\hat{W}_{n:1,0}\|_F^3 \tilde{\epsilon}} (k\eta)^2 \\
& < \frac{9000n^{13/2} e^{6n-5} \left( \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{(7n-5)/2}}{\|\hat{W}_{n:1,0}\|_F^3 \tilde{\epsilon}} \cdot \frac{9n^2 \left( \max \left\{ 1, \frac{3}{2}, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{2n}}{\|\hat{W}_{n:1,0}\|_F^2} \left( \ln \left( \frac{15n \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\}}{\|\hat{W}_{n:1,0}\|_F \tilde{\epsilon}} \right) \right)^2 \\
& < \frac{n^{17/2} e^{7n+7} \left( \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\} \right)^{(11n-5)/2}}{\|\hat{W}_{n:1,0}\|_F^5 \tilde{\epsilon}} \left( \ln \left( \frac{15n \max \left\{ 1, \frac{1-\hat{\nu}}{1+\hat{\nu}} \right\}}{\|\hat{W}_{n:1,0}\|_F \tilde{\epsilon}} \right) \right)^2 \\
& \leq \frac{n^{17/2} e^{7n+7} \left( \max \left\{ 3, \frac{3-\nu}{1+\nu} \right\} \right)^{(11n-5)/2}}{\left(\frac{2}{3}\right)^5 \|\hat{W}_{n:1,0}\|_F^5 \tilde{\epsilon}} \left( \ln \left( \frac{15n \max \left\{ 3, \frac{3-\nu}{1+\nu} \right\}}{\frac{2}{3} \|\hat{W}_{n:1,0}\|_F \tilde{\epsilon}} \right) \right)^2 \\
& \leq \frac{n^{17/2} e^{7n+10} \left( \max \left\{ 3, \frac{3-\nu}{1+\nu} \right\} \right)^{(11n-5)/2}}{\|\hat{W}_{n:1,0}\|_F^5 \tilde{\epsilon}} \left( \ln \left( \frac{23n \max \left\{ 3, \frac{3-\nu}{1+\nu} \right\}}{\|\hat{W}_{n:1,0}\|_F \tilde{\epsilon}} \right) \right)^2 \\
& \leq 1/\eta,
\end{aligned}$$

where the first transition is due to Equation (98); the second makes use of  $\beta_{k\eta, \epsilon} = 16n$ ,  $\gamma_{k\eta, \epsilon} = 6\sqrt{n}$ , Equation (96) and the definition of  $\epsilon$  (Equation (93)); the fourth relies on Equation (97); the sixth

is an outcome of Equations (90) and (91); and the last follows from the upper bound on  $\eta$  given in Equation (35), as well as the condition  $\tilde{\epsilon} \leq 1$ . Rearrange the inequality above:

$$\eta < \frac{\epsilon - e^{\int_0^{k\eta} m(t') dt'} \|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}(0)\|_2}{\beta_{k\eta, \epsilon} \gamma_{k\eta, \epsilon} k\eta e^{\int_0^{k\eta} m(t') dt'}}.$$

Since  $m(\cdot)$  is non-negative, it holds that:

$$\frac{\epsilon - e^{\int_0^{k\eta} m(t') dt'} \|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}(0)\|_2}{\beta_{k\eta, \epsilon} \gamma_{k\eta, \epsilon} k\eta e^{\int_0^{k\eta} m(t') dt'}} \leq \inf_{t \in (0, k\eta]} \frac{\epsilon - e^{\int_0^t m(t') dt'} \|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}(0)\|_2}{\beta_{k\eta, \epsilon} \gamma_{k\eta, \epsilon} \int_0^t e^{\int_{t'}^t m(t'') dt''} dt'},$$

and therefore:

$$\eta < \inf_{t \in (0, k\eta]} \frac{\epsilon - e^{\int_0^t m(t') dt'} \|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}(0)\|_2}{\beta_{k\eta, \epsilon} \gamma_{k\eta, \epsilon} \int_0^t e^{\int_{t'}^t m(t'') dt''} dt'}. \quad (99)$$

We now invoke Theorem 3 with  $\epsilon$  as we have defined (Equation (93)), time  $\tilde{t} = k\eta$ , and  $\beta_{k\eta, \epsilon}$ ,  $\gamma_{k\eta, \epsilon}$  and  $m(\cdot)$  as produced by Proposition 3. The theorem implies that, by Equation (99), the first  $\lceil k\eta/\eta \rceil = k$  iterates of gradient descent  $\epsilon$ -approximate the gradient flow trajectory up to time  $k\eta$ , i.e.  $\|\boldsymbol{\theta}_{k'} - \hat{\boldsymbol{\theta}}(k'\eta)\|_2 \leq \epsilon$  for all  $k' \in \{1, 2, \dots, k\}$ . In particular  $\|\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}(k\eta)\|_2 \leq \epsilon$ , as required.  $\square$

### J.21.1 Proof of Lemma 52

For conciseness, in the current proof we omit a second subscript “0” from our notation. Namely, we use  $W_{n:1}$  and  $\hat{W}_{n:1}$  as shorthand for  $W_{n:1,0}$  and  $\hat{W}_{n:1,0}$  respectively, and for any  $j \in [n]$ ,  $W_j$  and  $\hat{W}_j$  serve as shorthand for  $W_{j,0}$  and  $\hat{W}_{j,0}$  respectively.

We start by proving Equation (88). The following matrix  $\hat{W}_{j':j}$ , for any  $j, j' \in [n]$ , is defined as  $\hat{W}_{j'} \hat{W}_{j'-1} \dots \hat{W}_j$  if  $j \leq j'$ , and as an identity matrix (with size to be inferred by context) otherwise. Recall that  $\hat{\boldsymbol{\theta}}_0$  meets the balancedness condition, i.e.  $\hat{W}_{j+1}^\top \hat{W}_j = \hat{W}_j \hat{W}_{j+1}^\top$  for all  $j \in [n-1]$ . Using this relation repeatedly (while recalling that  $d_n = 1$ ), we have:

$$\begin{aligned} \|\hat{W}_{n:1}\|_F^2 &= \hat{W}_{n:1} \hat{W}_{n:1}^\top \\ &= \hat{W}_{n:2} \hat{W}_1 \hat{W}_1^\top \hat{W}_{n:2}^\top \\ &= \hat{W}_{n:2} \hat{W}_2^\top \hat{W}_2 \hat{W}_{n:2}^\top \\ &= \hat{W}_{n:3} \hat{W}_2 \hat{W}_2^\top \hat{W}_2 \hat{W}_2^\top \hat{W}_{n:3}^\top \\ &= \hat{W}_{n:3} \hat{W}_3^\top \hat{W}_3 \hat{W}_3^\top \hat{W}_{n:3}^\top \\ &\vdots \\ &= (\hat{W}_n \hat{W}_n^\top)^n \\ &= \|\hat{W}_n\|_F^{2n}. \end{aligned}$$

Since the balancedness condition implies that  $\|\hat{W}_j\|_F = \|\hat{W}_{j+1}\|_F$  for any  $j \in [n-1]$ , we may conclude  $\|\hat{W}_j\|_F = \|\hat{W}_{n:1}\|_F^{1/n}$  for any  $j \in [n]$ . It holds that:

$$\begin{aligned} &\|W_{n:1} - \hat{W}_{n:1}\|_F \\ &= \|(\hat{W}_n + W_n - \hat{W}_n) - (\hat{W}_1 + W_1 - \hat{W}_1) - \hat{W}_{n:1}\|_F \\ &= \left\| \sum_{(b_1, \dots, b_n) \in \{0,1\}^n} (b_n \hat{W}_n + (1-b_n)(W_n - \hat{W}_n)) - (b_1 \hat{W}_1 + (1-b_1)(W_1 - \hat{W}_1)) - \hat{W}_{n:1} \right\|_F \\ &= \left\| \sum_{(b_1, \dots, b_n) \in \{0,1\}^n \setminus \{1\}^n} (b_n \hat{W}_n + (1-b_n)(W_n - \hat{W}_n)) - (b_1 \hat{W}_1 + (1-b_1)(W_1 - \hat{W}_1)) \right\|_F \\ &\leq \sum_{(b_1, \dots, b_n) \in \{0,1\}^n \setminus \{1\}^n} \left\| (b_n \hat{W}_n + (1-b_n)(W_n - \hat{W}_n)) - (b_1 \hat{W}_1 + (1-b_1)(W_1 - \hat{W}_1)) \right\|_F \\ &\leq \sum_{(b_1, \dots, b_n) \in \{0,1\}^n \setminus \{1\}^n} \left\| b_n \hat{W}_n + (1-b_n)(W_n - \hat{W}_n) \right\|_F - \left\| b_1 \hat{W}_1 + (1-b_1)(W_1 - \hat{W}_1) \right\|_F \\ &\leq \sum_{(b_1, \dots, b_n) \in \{0,1\}^n \setminus \{1\}^n} (b_n \|\hat{W}_n\|_F + (1-b_n) \|W_n - \hat{W}_n\|_F) - (b_1 \|\hat{W}_1\|_F + (1-b_1) \|W_1 - \hat{W}_1\|_F), \end{aligned}$$

where the inequalities follow from sub-multiplicativity and sub-additivity of Frobenius norm. Since  $\|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_0\|_2 \leq \sqrt{n^3 \hat{\epsilon}}$  and  $\|\hat{W}_j\|_F = \|\hat{W}_{n:1}\|_F^{1/n}$  for any  $j \in [n]$ , we obtain:

$$\begin{aligned}
& \|W_{n:1} - \hat{W}_{n:1}\|_F \\
& \leq \sum_{(b_1, \dots, b_n) \in \{0,1\}^n \setminus \{1\}^n} (b_n \| \hat{W}_{n:1} \|_F^{1/n} + (1-b_n) \sqrt{n^3 \hat{\epsilon}}) - (b_1 \| \hat{W}_{n:1} \|_F^{1/n} + (1-b_1) \sqrt{n^3 \hat{\epsilon}}) \\
& = (\| \hat{W}_{n:1} \|_F^{1/n} + \sqrt{n^3 \hat{\epsilon}})^n - \| \hat{W}_{n:1} \|_F \\
& = \sum_{j=0}^n \binom{n}{j} \| \hat{W}_{n:1} \|_F^{(n-j)/n} (n^3 \hat{\epsilon})^{j/2} - \| \hat{W}_{n:1} \|_F \\
& = \sum_{j=1}^n \binom{n}{j} \| \hat{W}_{n:1} \|_F^{(n-j)/n} (n^3 \hat{\epsilon})^{j/2} \\
& \leq \sum_{j=1}^n n^j \max\{1, \| \hat{W}_{n:1} \|_F\} (n^3 \hat{\epsilon})^{j/2} \\
& = \max\{1, \| \hat{W}_{n:1} \|_F\} \sum_{j=1}^n (n^5 \hat{\epsilon})^{j/2} \\
& \leq \max\{1, \| \hat{W}_{n:1} \|_F\} \sum_{j=1}^{\infty} (n^5 \hat{\epsilon})^{j/2}.
\end{aligned}$$

Since  $\sqrt{n^5 \hat{\epsilon}} < 1$  (relying on the definition of  $\hat{\epsilon}$  in Equation (34)), we obtain:

$$\|W_{n:1} - \hat{W}_{n:1}\|_F \leq \max\{1, \| \hat{W}_{n:1} \|_F\} \frac{\sqrt{n^5 \hat{\epsilon}}}{1 - \sqrt{n^5 \hat{\epsilon}}}. \quad (100)$$

By the definition of  $\hat{\epsilon}$  (Equation (34)), it follows that  $\sqrt{n^5 \hat{\epsilon}} / (1 - \sqrt{n^5 \hat{\epsilon}}) \leq \frac{1}{3} \|W_{n:1}\|_F$ , thus:

$$\|W_{n:1} - \hat{W}_{n:1}\|_F \leq \frac{1}{3} \max\{1, \| \hat{W}_{n:1} \|_F\} \|W_{n:1}\|_F.$$

We conclude the proof of Equation (88) by showing that  $\| \hat{W}_{n:1} \|_F \leq 1$ . Indeed, assuming that this is not the case, *i.e.*  $\| \hat{W}_{n:1} \|_F > 1$ , while recalling that  $\|W_{n:1}\|_F \leq 0.1$ , leads us to a contradiction:

$$\| \hat{W}_{n:1} \|_F \leq \|W_{n:1}\|_F + \| \hat{W}_{n:1} - W_{n:1} \|_F \leq 0.1 \| \hat{W}_{n:1} \|_F + \frac{1}{3} \|W_{n:1}\|_F \| \hat{W}_{n:1} \|_F < \| \hat{W}_{n:1} \|_F.$$

Note that in addition to Equation (88), from Equation (100) and the fact that  $\| \hat{W}_{n:1} \|_F \leq 1$ , we may also establish the following:

$$\|W_{n:1} - \hat{W}_{n:1}\|_F \leq \| \hat{W}_{n:1} \|_F \frac{\sqrt{n^5 \hat{\epsilon}}}{1 - \sqrt{n^5 \hat{\epsilon}}}. \quad (101)$$

Moving on to the proof of Equation (89), we split the analysis into the following two cases: (i)  $\nu \in [0, 1]$ ; and (ii)  $\nu \in (-1, 0)$ . We start by analyzing case (i). Note that Equation (88) together with the fact that  $\|W_{n:1}\|_F \neq 0$  imply  $\| \hat{W}_{n:1} \|_F \neq 0$ . It holds that:

$$\begin{aligned}
\hat{\nu} &= \frac{\langle \Lambda_{yx}, \hat{W}_{n:1} \rangle}{\| \Lambda_{yx} \|_F \| \hat{W}_{n:1} \|_F} \\
&= \frac{\langle \Lambda_{yx}, W_{n:1} + \hat{W}_{n:1} - W_{n:1} \rangle}{\| \Lambda_{yx} \|_F \| \hat{W}_{n:1} \|_F} \\
&= \frac{\langle \Lambda_{yx}, W_{n:1} \rangle}{\| \Lambda_{yx} \|_F \| \hat{W}_{n:1} \|_F} + \frac{\langle \Lambda_{yx}, \hat{W}_{n:1} - W_{n:1} \rangle}{\| \Lambda_{yx} \|_F \| \hat{W}_{n:1} \|_F} \\
&= \nu \cdot \frac{\|W_{n:1}\|_F}{\| \hat{W}_{n:1} \|_F} + \frac{\langle \Lambda_{yx}, \hat{W}_{n:1} - W_{n:1} \rangle}{\| \Lambda_{yx} \|_F \| \hat{W}_{n:1} \|_F} \\
&\geq 0 + \frac{\langle \Lambda_{yx}, \hat{W}_{n:1} - W_{n:1} \rangle}{\| \Lambda_{yx} \|_F \| \hat{W}_{n:1} \|_F}.
\end{aligned}$$

Recall that  $\| \Lambda_{yx} \|_F = 1$ . We may finish the proof for case (i) by using Cauchy-Schwartz and triangle inequalities together with Equation (88):

$$\begin{aligned}
\hat{\nu} &\geq -\frac{1 \cdot \| \hat{W}_{n:1} - W_{n:1} \|_F}{1 \cdot \| \hat{W}_{n:1} \|_F} \\
&= -\frac{\| \hat{W}_{n:1} - W_{n:1} \|_F}{\|W_{n:1} + \hat{W}_{n:1} - W_{n:1}\|_F} \\
&\geq -\frac{\| \hat{W}_{n:1} - W_{n:1} \|_F}{\|W_{n:1}\|_F - \| \hat{W}_{n:1} - W_{n:1} \|_F} \\
&\geq -\frac{\|W_{n:1}\|_F / 3}{2 \|W_{n:1}\|_F / 3} \\
&= -\frac{1}{2}.
\end{aligned}$$

Regarding case (ii) (i.e.  $\nu \in (-1, 0)$ ), we have that:

$$\begin{aligned}
|\hat{\nu}| &= \frac{|\langle \Lambda_{yx}, \hat{W}_{n:1} \rangle|}{\|\Lambda_{yx}\|_F \|\hat{W}_{n:1}\|_F} \\
&= \frac{|\langle \Lambda_{yx}, W_{n:1} + \hat{W}_{n:1} - W_{n:1} \rangle|}{1 \cdot \|W_{n:1} + \hat{W}_{n:1} - W_{n:1}\|_F} \\
&= \frac{|\langle \Lambda_{yx}, W_{n:1} \rangle + \langle \Lambda_{yx}, \hat{W}_{n:1} - W_{n:1} \rangle|}{\|W_{n:1} + \hat{W}_{n:1} - W_{n:1}\|_F} \\
&\leq \frac{|\langle \Lambda_{yx}, W_{n:1} \rangle| + |\langle \Lambda_{yx}, \hat{W}_{n:1} - W_{n:1} \rangle|}{(\|W_{n:1}\|_F - \|\hat{W}_{n:1} - W_{n:1}\|_F)} \\
&\leq \frac{|\langle \Lambda_{yx}, W_{n:1} \rangle| + 1 \cdot \|\hat{W}_{n:1} - W_{n:1}\|_F}{\|W_{n:1}\|_F - \|\hat{W}_{n:1} - W_{n:1}\|_F} \\
&= |\nu| - |\nu| + \frac{|\langle \Lambda_{yx}, W_{n:1} \rangle| + \|\hat{W}_{n:1} - W_{n:1}\|_F}{\|W_{n:1}\|_F - \|\hat{W}_{n:1} - W_{n:1}\|_F} \\
&= |\nu| - \frac{|\langle \Lambda_{yx}, W_{n:1} \rangle|}{\|W_{n:1}\|_F} + \frac{|\langle \Lambda_{yx}, W_{n:1} \rangle| + \|\hat{W}_{n:1} - W_{n:1}\|_F}{\|W_{n:1}\|_F - \|\hat{W}_{n:1} - W_{n:1}\|_F} \\
&= |\nu| - \frac{|\langle \Lambda_{yx}, W_{n:1} \rangle|}{1 \cdot \|W_{n:1}\|_F} + \frac{|\langle \Lambda_{yx}, W_{n:1} \rangle| + \|\hat{W}_{n:1} - W_{n:1}\|_F}{\|W_{n:1}\|_F - \|\hat{W}_{n:1} - W_{n:1}\|_F} \\
&= |\nu| + \|\hat{W}_{n:1} - W_{n:1}\|_F \cdot \frac{\|W_{n:1}\|_F^{-1} |\langle \Lambda_{yx}, W_{n:1} \rangle| + 1}{\|W_{n:1}\|_F - \|\hat{W}_{n:1} - W_{n:1}\|_F} \\
&= |\nu| + \|\hat{W}_{n:1} - W_{n:1}\|_F \cdot \frac{|\nu| + 1}{\|W_{n:1}\|_F - \|\hat{W}_{n:1} - W_{n:1}\|_F},
\end{aligned}$$

where the first transition relies on  $\|\hat{W}_{n:1}\|_F \neq 0$ ; the second uses  $\|\Lambda_{yx}\|_F = 1$ ; the fourth uses triangle inequality, and relies on Equation (88) ensuring positive denominator; the fifth uses Cauchy-Schwartz and  $\|\Lambda_{yx}\|_F = 1$ ; and both the eighth and the last follow from  $\|\Lambda_{yx}\|_F = 1$ . It holds that:

$$\begin{aligned}
|\hat{\nu}| &\leq |\nu| + \|\hat{W}_{n:1} - W_{n:1}\|_F \cdot \frac{3}{2} \frac{|\nu| + 1}{\|W_{n:1}\|_F} \\
&\leq |\nu| + \|\hat{W}_{n:1} - W_{n:1}\|_F \cdot \frac{3}{\|W_{n:1}\|_F} \\
&\leq |\nu| + 3 \frac{\sqrt{n^5 \hat{\epsilon}}}{1 - \sqrt{n^5 \hat{\epsilon}}} \cdot \frac{\|\hat{W}_{n:1}\|_F}{\|W_{n:1}\|_F} \\
&\leq |\nu| + 3 \frac{\sqrt{n^5 \hat{\epsilon}}}{1 - \sqrt{n^5 \hat{\epsilon}}} \cdot \frac{\|W_{n:1}\|_F + \|\hat{W}_{n:1} - W_{n:1}\|_F}{\|W_{n:1}\|_F} \\
&\leq |\nu| + 4 \frac{\sqrt{n^5 \hat{\epsilon}}}{1 - \sqrt{n^5 \hat{\epsilon}}} \\
&\leq |\nu| + 4 \frac{(1+\nu)/16}{1-1/2} \\
&= |\nu| + \frac{1-|\nu|}{2} \\
&= \frac{|\nu| + 1}{2},
\end{aligned}$$

where the first transition uses Equation (88); the second relies on  $|\nu| \leq 1$ ; the third uses Equation (101); the fourth follows from triangle inequality; the fifth uses Equation (88); the sixth follows from the definition of  $\hat{\epsilon}$  (Equation (34)), namely that  $\sqrt{n^5 \hat{\epsilon}} \leq (1+\nu)/16$  and  $\sqrt{n^5 \hat{\epsilon}} \leq 1/2$ ; and the seventh relies on the assumption of case (ii) (i.e.  $\nu < 0$ ). After proving both cases (i) and (ii), we may conclude Equation (89).

Equation (90) follows from triangle inequality and Equation (88):

$$\|\hat{W}_{n:1}\|_F \geq \|W_{n:1}\|_F - \|\hat{W}_{n:1} - W_{n:1}\|_F \geq \|W_{n:1}\|_F - \frac{1}{3} \|W_{n:1}\|_F = \frac{2}{3} \|W_{n:1}\|_F.$$

To prove Equation (91), it suffices to show that  $\frac{1-\hat{\nu}}{1+\hat{\nu}} \leq 3$  or  $\frac{1-\hat{\nu}}{1+\hat{\nu}} \leq \frac{3-\nu}{1+\nu}$ . We prove this separately for the following two cases:  $-\frac{1}{2} \leq \text{sign}(\nu) \frac{|\nu|+1}{2}$  and  $-\frac{1}{2} > \text{sign}(\nu) \frac{|\nu|+1}{2}$ . In the case of  $-\frac{1}{2} \leq \text{sign}(\nu) \frac{|\nu|+1}{2}$ , Equation (89) implies  $\hat{\nu} \geq -\frac{1}{2}$ . Thus, we have that  $\frac{1-\hat{\nu}}{1+\hat{\nu}} \leq \frac{1-(-0.5)}{1-0.5} = 3$ , thereby proving that Equation (91) holds for this case. For the other case (i.e.  $-0.5 > \text{sign}(\nu) \frac{|\nu|+1}{2}$ ), we have that  $\nu < 0$ , and Equation (89) implies  $\hat{\nu} \geq -\frac{1-\nu}{2}$ . Thus, we have that  $\frac{1-\hat{\nu}}{1+\hat{\nu}} \leq \frac{1-(-(1-\nu)/2)}{1-(1-\nu)/2} = \frac{1.5-\nu/2}{0.5+\nu/2} = \frac{3-\nu}{1+\nu}$ , thereby proving that Equation (91) holds for the second (and last) case.  $\square$