Implicit Regularization in ReLU Networks with the Square Loss

Gal Vardi GAL.VARDI@WEIZMANN.AC.IL and **Ohad Shamir** OHAD.SHAMIR@WEIZMANN.AC.IL
Weizmann Institute of Science

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Understanding the implicit regularization (or implicit bias) of gradient descent has recently been a very active research area. However, the implicit regularization in nonlinear neural networks is still poorly understood, especially for regression losses such as the square loss. Perhaps surprisingly, we prove that even for a single ReLU neuron, it is impossible to characterize the implicit regularization with the square loss by any explicit function of the model parameters (although on the positive side, we show it can be characterized approximately). For one hidden-layer networks, we prove a similar result, where in general it is impossible to characterize implicit regularization properties in this manner, except for the "balancedness" property identified in Du et al. (2018). Our results suggest that a more general framework than the one considered so far may be needed to understand implicit regularization for nonlinear predictors, and provides some clues on what this framework should be.

1. Introduction

A major open question in the theory of deep learning is how neural networks generalize even when trained without any explicit regularization, and when there are far more learnable parameters than training examples. In such an underdetermined optimization problem, there are many global minima with zero training loss, and gradient descent seems to prefer solutions that generalize well (see Zhang et al. (2016)). Hence, it is believed that gradient descent induces an *implicit regularization* (or implicit bias) (Neyshabur et al., 2014, 2017), and characterizing this regularization/bias has been a subject for an extensive research in recent years.

The focus in the existing research is on finding a regularization function $\mathcal{R}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the parameters of the model, such that if we apply gradient descent on the average loss, then it converges in some sense to a global optimum that minimizes $\mathcal{R}$. Thus, the function $\mathcal{R}$ determines to which global optimum gradient descent converges. For example, it is known that for linear regression, gradient descent on the average square loss converges to the zero-loss solution with minimal ℓ_2 norm (Zhang et al., 2016). For linear classification and exponentially-tailed losses (such as the logistic or exponential losses), gradient descent on linearly separable data converges in direction to the minimal ℓ_2 norm predictor which attains a fixed positive margin over the data, also known as the max-margin solution (cf. Soudry et al. (2018)). Thus, in both cases $\mathcal{R}$ is essentially the ℓ_2 norm.

Once we move beyond simple linear classification and regression, the situation gets more complicated. This is especially so when considering the square loss (a setting which we focus on in this paper), where existing works still focus on predictors which express linear functions[1]. For example, there has been much effort to characterize the implicit regularization of gradient descent in matrix factorization. This problem corresponds to training a depth-2 linear neural network, and is

---

1. See the related work section below for discussion of results on classification losses, which are relatively easier to tackle.

considered a well-studied test-bed for studying implicit regularization in deep learning. Gunasekar et al. (2018c) conjectured that the implicit regularization in matrix factorization is the nuclear norm, and proved it for some restricted cases. This conjecture was further studied in a string of works (e.g., Belabbas (2020); Arora et al. (2019); Razin and Cohen (2020)), and was formally refuted by Li et al. (2020). Focusing on a more specific setting, Woodworth et al. (2020) recently studied the implicit regularization in *diagonal linear neural networks*, namely, linear networks where the weight matrices have a diagonal structure. They find the regularization function and show how it interpolates between the ℓ_1 and ℓ_2 norms depending on the initialization scale. This result was further generalized in Yun et al. (2020).

In this paper, we study implicit regularization with the square loss in nonlinear neural networks, when trained using gradient descent with infinitesimal step sizes (a.k.a. gradient flow). Perhaps surprisingly, we show that already for very simple such networks (involving a single ReLU neuron, or one thin hidden layer), the implicit regularization *cannot* be expressed by any explicit function of the model parameters. In other words, the model used to capture the implicit regularization in previous works cannot be used to capture the implicit regularization of nonlinear models in general (at least with respect to the square loss). However, on the positive side, we show that this implicit regularization can sometimes be captured by an explicit $\mathcal{R}$, but only up to some constant approximation factor. In a bit more detail, our contributions are as follows:

- We start with single-neuron networks, namely, $\mathbf{x} \mapsto \sigma(\langle \mathbf{x}, \mathbf{w} \rangle)$. If the activation function σ is strictly monotonic, then it is not hard to show that the implicit regularization is the ℓ_2 norm. However, if σ is the ReLU function, then we show that the implicit regularization is not expressible by any nontrivial function of $\mathbf{w}$. A bit more precisely, suppose that $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ is such that if gradient flow converges to a zero-loss solution $\mathbf{w}^*$, then $\mathbf{w}^*$ is a zero-loss solution that minimizes $\mathcal{R}$. We show that such a function $\mathcal{R}$ must be constant in $\mathbb{R}^d \setminus \{\mathbf{0}\}$. Hence, perhaps surprisingly, the approach of precisely specifying the implicit bias of gradient flow via a regularization function is not feasible in single-neuron networks.

- On the positive side, we show that while the implicit bias of a ReLU neuron cannot be specified exactly by a regularization function, it can be expressed approximately, within a factor of 2, by the ℓ_2 norm. That is, let $\mathbf{w}^*$ be a zero-loss solution with a minimal ℓ_2 norm, and assume that gradient flow converges to some point $\mathbf{w}(\infty)$, then $\|\mathbf{w}(\infty)\|_2 \leq 2 \|\mathbf{w}^*\|_2$. Assuming $\|\mathbf{w}^*\|$ is not too large, such a bound on the ℓ_2 norm can be used to derive good statistical generalization guarantees, via standard techniques (cf. Shalev-Shwartz and Ben-David (2014)).

- We extend our study to depth-2 ReLU networks. For such networks, an important implicit bias property shown in Du et al. (2018) is that gradient flow enforces the differences between square norms across different layers to remain invariant. Starting from a point close to zero, it implies that the magnitudes of all layers are automatically balanced. Thus, gradient flow induces a bias toward "balanced" layers. However, this bias is rather weak and is not related to properties that may allow us to obtain generalization guarantees, such as small norms, sparsity, or low rank. We consider networks with one hidden ReLU neuron, i.e., $\mathbf{x} \mapsto v \cdot \sigma(\langle \mathbf{x}, \mathbf{w} \rangle)$ (where $v, \mathbf{w}$ are the trained parameters), which is the simplest case of a depth-2 ReLU network. We show that the *only* bias which can be specified by a function $\mathcal{R}$ of the model parameters is the balancedness property: Namely, $\mathcal{R}$ is constant in the set of parameters that satisfy this property. Thus, no other implicit bias property can be expressed by such a

function for single-hidden-neuron networks. This also has implications for general depth-2 ReLU networks: If $\mathcal{R}$ is a function that specifies the implicit regularization of depth-2 networks, then in the special case of single-hidden-neuron networks, it does not induce bias except for enforcing the balancedness property. Hence, the implicit regularization function of depth-2 ReLU networks cannot be generally related to properties such as small norms or sparsity, which should apply also to single-hidden-neuron networks.

We again emphasize that since our results show the impossibility of characterizing the implicit regularization already for very simple networks, similar difficulties will be encountered in more general cases (such as ReLU networks of any depth and width).

Overall, our results suggest that the implicit regularization for ReLU networks with the square loss may not be expressible by an explicit function $\mathcal{R}$ of the model parameters. On the positive side, they also suggest how this can be overcome by changing the model: Instead of looking for such an $\mathcal{R}$ which captures the bias exactly, we might try to capture it *approximately*, as we did in the single ReLU neuron case. Another possible direction is to find a regularization function $\mathcal{R}$ that is *data-dependent* in some simple way[2], and does not depend just on the model parameters. We believe these are both interesting directions for further research.

Related Work

As we already discussed, implicit regularization in matrix factorization and linear neural networks was extensively studied, as a first step toward understanding implicit regularization in more complex models (see, e.g., Gunasekar et al. (2018c); Razin and Cohen (2020); Arora et al. (2019); Belabbas (2020); Eftekhari and Zygalakis (2020); Li et al. (2018); Ma et al. (2018); Woodworth et al. (2020); Gidel et al. (2019); Li et al. (2020); Yun et al. (2020)). In Razin and Cohen (2020) it was shown that gradient flow in matrix factorization may approach a global minimum at infinity rather than converging to a global minimum with a finite norm. This result suggests that the implicit regularization in matrix factorization may not be expressible by norms. The main conceptual differences from our paper are that we rule out all regularization functions, and that our result holds already if we assume that gradient flow converges to some global minimum with a finite norm.

Oymak and Soltanolkotabi (2019) studied the implicit bias of gradient descent in certain nonlinear models. They showed that under some assumptions, gradient descent is guarranteed to converge to a zero-loss solution with a bounded ℓ_2 norm. Their results have implications on the implicit regularization of single-neuron networks with a strictly monotonic activation function, which we discuss in Remark 2. Williams et al. (2019) and Jin and Montúfar (2020) studied the dynamics and implicit bias of gradient descent in wide depth-2 ReLU networks with input dimension 1.

The conceptually closest works to ours are Dauber et al. (2020); Suggala et al. (2018), which show that variants of gradient descent on some (specially crafted) convex optimization problems do not converge to a closest Euclidean solution, or cannot be explained by any reasonable implicit regularization function. These works, as well as ours, consider the limitations of implicit regularization, but our results apply to a standard learning problem, namely, learning neural networks. Thus, while the results of Dauber et al. (2020); Suggala et al. (2018) demonstrate that there are learning problems that cannot be explained by any reasonable regularization function, we show that this phenomenon occurs already for simple ReLU networks.

---

2. Otherwise, a trivial solution is to define $\mathcal{R}(\boldsymbol{\theta}, \mathcal{X})$ (where $\mathcal{X}$ is the dataset) to be 0 if $\boldsymbol{\theta}$ are the parameters returned by gradient flow given $\mathcal{X}$, and 1 otherwise.

The implicit bias of gradient descent in classification tasks is also widely studied. Soudry et al. (2018) showed that gradient descent on linearly-separable binary classification problem with an exponentially-tailed loss (e.g., the exponential loss and the logistic loss), converges to the maximum margin direction. This analysis was extended to other loss functions, tighter convergence rates, nonseparable data, and variants of gradient-based optimization algorithms (Nacson et al., 2019; Ji and Telgarsky, 2018b; Ji et al., 2020; Gunasekar et al., 2018a; Shamir, 2020). The problem was also studied for more complex models, such as linear neural networks (Gunasekar et al., 2018b; Ji and Telgarsky, 2018a; Moroshko et al., 2020; Yun et al., 2020), and neural networks with homogeneous activation functions (Lyu and Li, 2019; Chizat and Bach, 2020; Xu et al., 2018).

2. Preliminaries

Notations. We use bold-faced letters to denote vectors, e.g., $\mathbf{x} = (x_1, \ldots, x_d)$. For $\mathbf{x} \in \mathbb{R}^d$ we denote by $\|\mathbf{x}\|$ the Euclidean norm.

Single-neuron networks. Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be a training dataset, where for every i we have $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We consider empirical-loss minimization of a single-neuron network, with respect to the square loss. Thus, the objective is given by

$$\mathcal{L}(\mathbf{w}) := \frac{1}{2} \sum_{i=1}^n \left(\sigma(\langle \mathbf{x}_i, \mathbf{w} \rangle) - y_i \right)^2 , \tag{1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonlinear activation function. Let $X \in \mathbb{R}^{n \times d}$ denote the data matrix, i.e., the rows of X are $\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top$, and let $\mathbf{y} = (y_1, \ldots, y_n)$. We have $\mathcal{L}(\mathbf{w}) = \frac{1}{2} \|\sigma(X\mathbf{w}) - \mathbf{y}\|^2$, where σ is applied component-wise. We assume that the data is realizable, that is, $\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = 0$. Moreover, we focus on settings where the network is *overparameterized*, in the sense that $\mathcal{L}$ has multiple (or even infinitely many) global minima.

We analyze the implicit regularization of gradient flow on the objective given by Eq. 1. This setting captures the behaviour of gradient descent with infinitesimally small step size. Let $\mathbf{w}(t)$ be the trajectory of gradient flow, where $\mathbf{w}(0)$ is the initial point. The dynamics of $\mathbf{w}(t)$ is given by the differential equation $\dot{\mathbf{w}}(t) := \frac{d\mathbf{w}(t)}{dt} = -\nabla \mathcal{L}(\mathbf{w}(t))$. If $\lim_{t \to \infty} \mathbf{w}(t)$ exists then we denote it by $\mathbf{w}(\infty)$. We note that gradient flow is not guaranteed to converge to a global minimum (cf. Yehudai and Shamir (2020)). In the overparameterized setting there can be infinitely many global minima, and we study to which one gradient flow converges, assuming that it converges to a global minimum.

3. Warm up – strictly monotonic activation functions

We first analyze the implicit regularization of gradient flow on the objective given by Eq. 1, where the activation function σ is strictly monotonic. In this case, we show that the implicit bias is simply the ℓ_2 norm: Namely, the point $\mathbf{w}$ closest to $\mathbf{w}(0)$ on which $\mathcal{L}(\mathbf{w}) = 0$.

Theorem 1 *Consider gradient flow on the objective given by Eq. 1, where $\sigma : \mathbb{R} \to \mathbb{R}$ is a strictly monotonic activation function. If $\mathbf{w}(\infty)$ exists and $\mathcal{L}(\mathbf{w}(\infty)) = 0$, then we have $\mathbf{w}(\infty) \in \operatorname{argmin}_{\mathbf{w}} \|\mathbf{w} - \mathbf{w}(0)\|$ s.t. $\mathcal{L}(\mathbf{w}) = 0$.*

Proof Since σ is strictly monotonic, then for every $1 \leq i \leq n$ we have $\sigma(\langle \mathbf{x}_i, \mathbf{w} \rangle) = y_i$ iff $\langle \mathbf{x}_i, \mathbf{w} \rangle = \sigma^{-1}(y_i)$. The KKT optimality conditions for the problem $\min_{\mathbf{w}} \|\mathbf{w} - \mathbf{w}(0)\|^2$ s.t. $\langle \mathbf{x}_i, \mathbf{w} \rangle = \sigma^{-1}(y_i), \; i = 1, \ldots, n$, are

$$\exists \lambda_1, \ldots, \lambda_n \text{ s.t. } \mathbf{w} = \mathbf{w}(0) + \sum_{i=1}^{n} \lambda_i \mathbf{x}_i$$

$$\forall 1 \leq i \leq n, \; \langle \mathbf{x}_i, \mathbf{w} \rangle = \sigma^{-1}(y_i) \, .$$

Note that since the constraints are affine functions, then the conditions are both necessary and sufficient. The second condition holds in $\mathbf{w}(\infty)$ since $\mathcal{L}(\mathbf{w}(\infty)) = 0$. Moreover, we have $\dot{\mathbf{w}}(t) = -\nabla \mathcal{L}(\mathbf{w}(t)) = -\sum_{i=1}^{n} \left(\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) - y_i \right) \sigma'(\mathbf{x}_i^\top \mathbf{w}(t)) \mathbf{x}_i$. Thus, $\dot{\mathbf{w}}(t)$ is always spanned by the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and hence the first condition also holds. ∎

Remark 2 *By Oymak and Soltanolkotabi (2019), if σ is continuously differentiable and there are positive constants γ, Γ such that $0 < \gamma \leq \sigma'(z) \leq \Gamma$ for every $z \in \mathbb{R}$, then $\mathbf{w}(t)$ is guaranteed to converge to a global optimum $\mathbf{w}(\infty)$ that is closest to $\mathbf{w}(0)$. In Theorem 1, our assumption on σ is weaker, but we assume that $\mathbf{w}(\infty)$ exists and $\mathcal{L}(\mathbf{w}(\infty)) = 0$ (which is a common assumption when considering implicit regularization).*

4. The implicit regularization of a ReLU neuron is not expressible by w

In this section we study the implicit regularization of gradient flow on the objective given by Eq. 1, where $\mathbf{w}(0) = \mathbf{0}$ and $\sigma : \mathbb{R} \to \mathbb{R}$ is the ReLU function. We impose the convention that even though the ReLU function is not differentiable at 0, we take $\sigma'(0) > 0$. This convention is required for analyzing gradient flow with $\mathbf{w}(0) = \mathbf{0}$, since if $\sigma'(0) = 0$ then $\mathbf{w}(t)$ would stay at $\mathbf{0}$ indefinitely. For simplicty, in our proofs we assume that $\sigma'(0) = 1$, but our results hold for every positive value.

Assume that the implicit regularization can be expressed as a function $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$, namely, if gradient-flow converges to a global minimum $\mathbf{w}^*$, then we have $\mathbf{w}^* \in \operatorname{argmin}_{\mathbf{w}} \mathcal{R}(\mathbf{w})$ s.t. $\mathcal{L}(\mathbf{w}) = 0$. We will first show that unlike the case of strictly monotonic activations, the implicit bias of $\mathcal{R}$ is *not* the ℓ_2 norm, and in fact, data-dependent in a nontrivial way. Then, we will extend this result, and show that $\mathcal{R}$ is trivial, namely, constant in $\mathbb{R}^d \setminus \{\mathbf{0}\}$. Thus, the behavior of $\mathbf{w}(\infty)$ has a complicated data-dependent behavior, and cannot be written down as minimizing some function dependent only on $\mathbf{w}$, among all zero-loss solutions.

4.1. $\mathcal{R}$ is not the ℓ_2 norm

We show that the implicit regularization of a single ReLU neuron is not the ℓ_2 norm. Thus, the implicit regularization of ReLU is different than that of strictly monotonic activation functions considered in Section 3. Specifically, in the following theorem we analyze the behavior of gradient flow for some family of inputs. The theorem considers a datasets of the form $(X_\gamma, \mathbf{y}) \in \mathbb{R}^{3 \times 3} \times \mathbb{R}^3$, where γ is a parameter. We show that gradient flow starting from $\mathbf{w}(0) = \mathbf{0}$ with the input $(X_\gamma, \mathbf{y})$, converges to a zero-loss solution which is not a minimal-norm zero-loss solution, and that the parameter γ controls the ratio between $\|\mathbf{w}(\infty)\|$ and the minimal-norm of a zero-loss solution. Hence, the implicit regularization is not the ℓ_2 norm. We will also use this theorem later in order to show additional results.

Theorem 3 *Let $\gamma \geq 0$, and let*

$$X_\gamma = \begin{bmatrix} 3 & -1 & 0 \\ 4 & 2 & 0 \\ 0 & \gamma & \gamma \end{bmatrix} .$$

Let $\alpha > 0$ and let $\mathbf{y}_\alpha = \alpha \cdot (16, 18, 0)^\top$. Consider gradient flow with input $(X_\gamma, \mathbf{y}_\alpha)$, starting from $\mathbf{w}(0) = \mathbf{0}$, on the objective given by Eq. 1, where $\sigma : \mathbb{R} \to \mathbb{R}$ is the ReLU function. Let $\mathcal{W}_\alpha = \{(5\alpha, -\alpha, s)^\top : s \leq 0\} \subseteq \mathbb{R}^3$. Note that for every $\mathbf{w} \in \mathcal{W}_\alpha$ we have $\sigma(X_\gamma \mathbf{w}) = \mathbf{y}_\alpha$, and thus $\mathcal{L}(\mathbf{w}) = 0$. Moreover, note that the global optimum with the minimal ℓ_2 norm is $\alpha \cdot (5, -1, 0)^\top$. We have:

- *If $\gamma = 0$ then $\mathbf{w}(\infty) = \alpha \cdot (5, -1, 0)^\top$.*

- *If $\gamma = 1$ then $\mathbf{w}(\infty) = \alpha \cdot (5, -1, s)^\top$ for some $s \in (-0.045, -0.035)$ independent of α.*

- *If $\gamma = 2$ then $\mathbf{w}(\infty) = \alpha \cdot (5, -1, s)^\top$ for some $s \in (-0.12, -0.1)$ independent of α.*

- *If $\gamma = 5$ then $\mathbf{w}(\infty) = \alpha \cdot (5, -1, s)^\top$ for some $s \in (-0.22, -0.2)$ independent of α.*

Proof [Proof sketch (for complete proof see Appendix A)] The trajectory $\mathbf{w}(t)$ is the solution of the initial value problem defined by the differential equation

$$\dot{\mathbf{w}}(t) = -\nabla \mathcal{L}(\mathbf{w}(t)) = -\sum_{i=1}^{3} \left(\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) - y_i \right) \sigma'(\mathbf{x}_i^\top \mathbf{w}(t)) \mathbf{x}_i , \tag{2}$$

and the initial condition $\mathbf{w}(0) = \mathbf{0}$. Here, $\mathbf{x}_i^\top$ are the rows of the matrix X_γ, and y_i are the components of $\mathbf{y}_\alpha$.

In Figure 1, we show the trajectories $\mathbf{w}(t)$ of gradient descent (an empirical simulation), for $\alpha = 1$ and $\gamma \in \{0, 1, 2, 5, 20\}$. The grey line denotes the set $\mathcal{W}_\alpha$, and the hyperplane corresponds to $\mathbf{x}_3^\top \mathbf{w} = 0$ (where $\gamma > 0$). Note that the trajectories for $\gamma > 0$ consist of two parts. In the first part $\mathbf{x}_3^\top \mathbf{w}(t) \geq 0$, and in the second part $\mathbf{x}_3^\top \mathbf{w}(t) \leq 0$. We note that in the second part, the component $w_3(t)$ remains constant. Indeed, by Eq. 2, the change in $\mathbf{w}(t)$ is spanned by $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$, and if $\mathbf{x}_3^\top \mathbf{w}(t) \leq 0$ then it is spanned only by $\mathbf{x}_1, \mathbf{x}_2$ (note that if $\mathbf{x}_3^\top \mathbf{w}(t) = 0$ then $\sigma(\mathbf{x}_3^\top \mathbf{w}(t)) - y_3 = 0$). Since $\mathbf{x}_1, \mathbf{x}_2$ have 0 as their third component, then $w_3(t)$ does not change. Intuitively, the parameter γ controls the extent to which $w_3(t)$ decreases during the first part, before reaching the second part in which it stays constant. In the proof we analyze these trajectories formally.

For $\gamma = 0$, we show that the solution to Eq. 2 is $\mathbf{w}(t) = (w_1(t), w_2(t), 0)$, where $\tilde{\mathbf{w}}(t) = (w_1(t), w_2(t))$ is given by $\tilde{\mathbf{w}}(t) = \tilde{X}^{-1} \tilde{\mathbf{y}} - \exp\left(-t \tilde{X}^\top \tilde{X} \right) \tilde{X}^{-1} \tilde{\mathbf{y}}$. Here, $\tilde{X} \in \mathbb{R}^{2 \times 2}$ is obtained from X_0 by omitting the third column and third row, $\tilde{\mathbf{y}} = (y_1, y_2)^\top$, and $\exp(\cdot)$ denotes the matrix exponential. Note that $\tilde{\mathbf{w}}(\infty) = \tilde{X}^{-1} \tilde{\mathbf{y}} = \alpha \cdot (5, -1)^\top$. Moreover, note that in order to show that this trajectory satisfies Eq. 2, it suffices to show that $\mathbf{x}_i^\top \mathbf{w}(t) \geq 0$ for every $i \in \{1, 2, 3\}$ and $t \geq 0$, and that $\mathbf{w}(t)$ satisfies the linear differential equation obtained form Eq. 2 by plugging in $\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) = \mathbf{x}_i^\top \mathbf{w}(t)$ and $\sigma'(\mathbf{x}_i^\top \mathbf{w}(t)) = 1$.

For $\gamma > 0$, we show that the solution to Eq. 2 consists of two parts: The first part is where $t \in [0, t_1]$ for some $t_1 > 0$, and second part is where $t \in [t_1, \infty)$. In the first part, we have $\mathbf{w}(t) = \alpha(5, -1, 1)^\top - \exp\left(-t X_\gamma^\top X_\gamma \right) \alpha(5, -1, 1)^\top$, and $\mathbf{x}_i^\top \mathbf{w}(t) \geq 0$ for every $i \in \{1, 2, 3\}$ and $t \in [0, t_1]$. Then, in the second part, $w_3(t)$ remains constant, and $\tilde{\mathbf{w}}(t) = (w_1(t), w_2(t))$ is given by
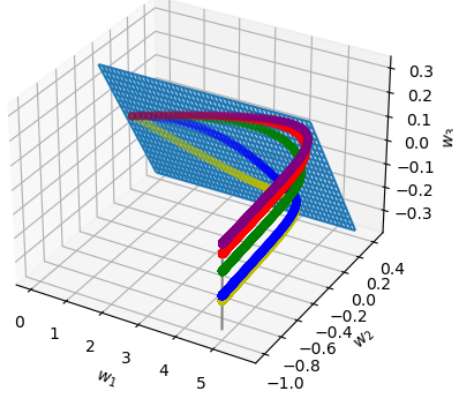
Figure 1: The trajectories for $\gamma = 0$ (purple), $\gamma = 1$ (red), $\gamma = 2$ (green), $\gamma = 5$ (blue), and $\gamma = 20$ (yellow).

$\tilde{\mathbf{w}}(t) = \alpha(5, -1)^\top + \exp\left(-(t - t_1)\tilde{X}_\gamma^\top \tilde{X}_\gamma\right)\left(\tilde{\mathbf{w}}(t_1) - \alpha(5, -1)^\top\right)$. Here, $\tilde{X}_\gamma \in \mathbb{R}^{2 \times 2}$ is obtained from X_γ by omitting the third column and third row. Note that $\tilde{\mathbf{w}}(\infty) = \alpha \cdot (5, -1)^\top$. In this part, we have $\mathbf{x}_i^\top \mathbf{w}(t) \geq 0$ for $i \in \{1, 2\}$ and $\mathbf{x}_3^\top \mathbf{w}(t) \leq 0$.

In the proof, we show for both parts, that the trajectories satisfy Eq. 2. Moreover, we investigate the value of $w_3(\infty) = w_3(t_1)$ in order to obtain the required bounds for each $\gamma \in \{1, 2, 5\}$. ∎

Remark 4 *Theorem 3 shows that by modifying the parameter γ we can control the ratio $\frac{\|\mathbf{w}(\infty)\|}{\|w^*\|}$, where $\mathbf{w}^*$ is a minimal-norm zero-loss solution. In Section 5 we will show that this ratio is bounded.*

4.2. $\mathcal{R}$ is trivial

We now prove the following theorem, which implies that $\mathcal{R}$ is trivial and does not induce bias.

Theorem 5 *Consider gradient flow starting from $\mathbf{w}(0) = \mathbf{0}$, on the objective given by Eq. 1, where $\sigma : \mathbb{R} \to \mathbb{R}$ is the ReLU function. Let $\mathcal{R} : \mathbb{R}^3 \to \mathbb{R}$, such that for every input $(X, \mathbf{y}) \in \mathbb{R}^{3 \times 3} \times \mathbb{R}^3$ where gradient flow converges to $\mathbf{w}(\infty)$ with $\mathcal{L}(\mathbf{w}(\infty)) = 0$, we have $\mathbf{w}(\infty) \in \operatorname{argmin}_{\mathbf{w}} \mathcal{R}(\mathbf{w})$ s.t. $\mathcal{L}(\mathbf{w}) = 0$. Then, $\mathcal{R}$ is constant in $\mathbb{R}^3 \setminus \{\mathbf{0}\}$.*

Let $X \in \mathbb{R}^{3 \times 3}$ and $\mathbf{y} \in \mathbb{R}^3$. Let $\mathbf{x}_1^\top, \mathbf{x}_2^\top, \mathbf{x}_3^\top$ be the rows of X. Suppose that given input $(X, \mathbf{y})$, gradient flow with $\mathbf{w}(0) = \mathbf{0}$ converges to $\mathbf{w}(\infty)$. Note that by rotating $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$, the trajectory $\mathbf{w}(t)$ is rotated accordingly Thus, for every rotation matrix $M \in SO(3)$, given input $(XM^\top, \mathbf{y})$ gradient flow converges to $M\mathbf{w}(\infty)$.

Let $\alpha > 0$. By Theorem 3, all points in $\mathcal{W}_\alpha$ are global minima for all inputs $\{(X_\gamma, \mathbf{y}_\alpha) : \gamma \geq 0\}$, but gradient-flow converges to different points $\mathbf{w}^\gamma \in \mathcal{W}_\alpha$, depending on γ. Hence, we

have $\mathcal{R}(\mathbf{w}^\gamma) = \min_{\mathbf{w} \in \mathcal{W}_\alpha} \mathcal{R}(\mathbf{w})$. Thus, Theorem 3 implies that $\mathcal{R}(\mathbf{w}^0) = \mathcal{R}(\mathbf{w}^1) = \mathcal{R}(\mathbf{w}^2) = \mathcal{R}(\mathbf{w}^5) = \min_{\mathbf{w} \in \mathcal{W}_\alpha} \mathcal{R}(\mathbf{w})$. We use this property in order to prove Theorem 5.

Let $M \in SO(3)$. We denote $M\mathcal{W}_\alpha = \{M\mathbf{w} : \mathbf{w} \in \mathcal{W}_\alpha\}$. By changing the input X_γ in Theorem 3 to $X_\gamma M^\top$, we deduce that $\mathcal{R}(M\mathbf{w}^0) = \mathcal{R}(M\mathbf{w}^1) = \min_{\mathbf{w} \in M\mathcal{W}_\alpha} \mathcal{R}(\mathbf{w})$. Indeed, all points in $M\mathcal{W}_\alpha$ are global minima for all inputs $\{(X_\gamma M^\top, \mathbf{y}_\alpha) : \gamma \geq 0\}$, but gradient-flow converges to different points $M\mathbf{w}^\gamma \in M\mathcal{W}_\alpha$, depending on γ. Let $c > 0$ be such that $\mathbf{w}^1 = \alpha \cdot (5, -1, -c)$. For every $\mathbf{w} \in \mathbb{R}^3$ with $\|\mathbf{w}\| = \|\mathbf{w}^0\| = \|\alpha \cdot (5, -1, 0)\| = \alpha\sqrt{26}$, and $\mathbf{v} \in \mathbb{R}^3$ such that $\langle \mathbf{w}, \mathbf{v} \rangle = 0$ and $\|\mathbf{v}\| = \alpha \cdot c$, by choosing an appropriate rotation M we obtain $\mathcal{R}(\mathbf{w}) = \mathcal{R}(\mathbf{w} + \mathbf{v}) = \min_{\beta \geq 0} \mathcal{R}(\mathbf{w} + \beta\mathbf{v})$. Finally, as we show in the following lemma, this property implies that $\mathcal{R}$ is constant over $\mathbb{R}^3 \setminus \{\mathbf{0}\}$, and thus we complete the proof of Theorem 5.

Lemma 6 *Let $q, c > 0$ be constants. Let $d \geq 3$, and let $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ be a function such that for every $\alpha > 0$, vector $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\| = \alpha q$, and vector $\mathbf{v} \in \mathbb{R}^d$ such that $\|\mathbf{v}\| = \alpha c$ and $\langle \mathbf{w}, \mathbf{v} \rangle = 0$, we have $\mathcal{R}(\mathbf{w}) = \mathcal{R}(\mathbf{w} + \mathbf{v}) = \min_{\beta \geq 0} \mathcal{R}(\mathbf{w} + \beta\mathbf{v})$. Then, $\mathcal{R}(\mathbf{w}) = \mathcal{R}(\mathbf{w}')$ for every $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d \setminus \{\mathbf{0}\}$.*

Proof [Proof sketch (for complete proof see Appendix B)] First, we show that $\mathcal{R}$ is radial, namely, there exists a function $f : \mathbb{R}_+ \to \mathbb{R}$ such that $\mathcal{R}(\mathbf{w}) = f(\|\mathbf{w}\|)$ for every $\mathbf{w} \in \mathbb{R}^d$. Assume that $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ are such that $\|\mathbf{w}\| = \|\mathbf{w}'\| = \alpha q$ for some $\alpha > 0$, and $\|\mathbf{w} - \mathbf{w}'\|$ is sufficiently small. We show that there exists $\mathbf{u}^* \in \mathbb{R}^d$ such that $\mathbf{u}^* = \mathbf{w} + \mathbf{v} = \mathbf{w}' + \mathbf{v}'$ for some $\mathbf{v}, \mathbf{v}'$ with $\|\mathbf{v}\| = \|\mathbf{v}'\| = \alpha c$ and $\langle \mathbf{w}, \mathbf{v} \rangle = \langle \mathbf{w}', \mathbf{v}' \rangle = 0$. By our assumption on $\mathcal{R}$, it implies that $\mathcal{R}(\mathbf{w}) = \mathcal{R}(\mathbf{u}^*) = \mathcal{R}(\mathbf{w}')$, and hence $\mathcal{R}$ is radial.

Thus, it suffices to show that $f(r) = f(r')$ for every $0 < r < r'$. Using our assumption on $\mathcal{R}$, we show that for every $m \in \mathbb{N}$ we have $f(r) = f\left(r\left(1 + \frac{c^2}{q^2}\right)^{m/2}\right)$. The assumption on $\mathcal{R}$ also implies that $f(r) = \min_{r'' \geq r} f(r'') \leq f(r')$, and that $f(r') \leq f\left(r\left(1 + \frac{c^2}{q^2}\right)^{m/2}\right)$ for a sufficiently large m. Therefore, we have $f(r) \leq f(r') \leq f\left(r\left(1 + \frac{c^2}{q^2}\right)^{m/2}\right) = f(r)$, and hence $f(r) = f(r')$ as required. ∎

Remark 7 ($\mathcal{R}$ **is trivial also for** $d > 3$) *Theorem 5 implies that the implicit regularization function $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ is trivial for $d = 3$. By padding the inputs $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ from Theorem 3 with zeros, we can obtain a d-dimensional version of Theorem 3 for $d > 3$. Then, using the same arguments as above, it implies that Theorem 5 holds for every $d \geq 3$.*

5. The implicit regularization of a ReLU neuron is approximately the ℓ_2 norm

While the implicit regularization of ReLU neuron cannot be expressed as a function $\mathcal{R}(\mathbf{w})$, in this section we show that it can be expressed approximately, within a factor of 2, by the ℓ_2 norm. This implies that even without early stopping, if the data can be labeled by a ReLU neuron with small ℓ_2 norm, then gradient flow will converge to a ReLU neuron whose ℓ_2 norm is not much larger. Since a ReLU neuron is just a linear function composed with a fixed nonlinearity, this can be used to derive good statistical generalization guarantees, via standard techniques (cf. Shalev-Shwartz and Ben-David (2014)).

Theorem 8 *Consider gradient flow on the objective given by Eq. 1, where $\sigma : \mathbb{R} \to \mathbb{R}$ is a monotonically non-decreasing activation function (e.g., ReLU). Assume that $\mathbf{w}(\infty)$ exists and $\mathcal{L}(\mathbf{w}(\infty)) = 0$. Let $\mathbf{w}^* \in \arg\min_{\mathbf{w}} \|\mathbf{w} - \mathbf{w}(0)\|$ s.t. $\mathcal{L}(\mathbf{w}) = 0$. Then, $\|\mathbf{w}(\infty) - \mathbf{w}(0)\| \le 2 \cdot \|\mathbf{w}^* - \mathbf{w}(0)\|$.*

Proof First, note that

$$\langle \nabla \mathcal{L}(\mathbf{w}(t)), \mathbf{w}(t) - \mathbf{w}^* \rangle = \left\langle \sum_{i=1}^{n} \left(\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) - \sigma(\mathbf{x}_i^\top \mathbf{w}^*) \right) \sigma'(\mathbf{x}_i^\top \mathbf{w}(t)) \mathbf{x}_i, \mathbf{w}(t) - \mathbf{w}^* \right\rangle$$

$$= \sum_{i=1}^{n} \left(\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) - \sigma(\mathbf{x}_i^\top \mathbf{w}^*) \right) \sigma'(\mathbf{x}_i^\top \mathbf{w}(t))(\mathbf{x}_i^\top \mathbf{w}(t) - \mathbf{x}_i^\top \mathbf{w}^*) .$$

Since σ is monotonically non-decreasing then $\left(\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) - \sigma(\mathbf{x}_i^\top \mathbf{w}^*) \right) (\mathbf{x}_i^\top \mathbf{w}(t) - \mathbf{x}_i^\top \mathbf{w}^*) \ge 0$, and $\sigma'(\mathbf{x}_i^\top \mathbf{w}(t)) \ge 0$. Hence, $\langle \nabla \mathcal{L}(\mathbf{w}(t)), \mathbf{w}(t) - \mathbf{w}^* \rangle \ge 0$. Therefore, we have

$$\frac{d}{dt} \left(\frac{1}{2} \|\mathbf{w}(t) - \mathbf{w}^*\|^2 \right) = \langle \mathbf{w}(t) - \mathbf{w}^*, \dot{\mathbf{w}}(t) \rangle = \langle \mathbf{w}(t) - \mathbf{w}^*, -\nabla \mathcal{L}(\mathbf{w}(t)) \rangle \le 0 .$$

Thus, $\|\mathbf{w}(\infty) - \mathbf{w}^*\| \le \|\mathbf{w}(0) - \mathbf{w}^*\|$. Hence,

$$\|\mathbf{w}(\infty) - \mathbf{w}(0)\| = \|\mathbf{w}(\infty) - \mathbf{w}^* + \mathbf{w}^* - \mathbf{w}(0)\| \le \|\mathbf{w}(\infty) - \mathbf{w}^*\| + \|\mathbf{w}^* - \mathbf{w}(0)\|$$

$$= 2 \cdot \|\mathbf{w}^* - \mathbf{w}(0)\| .$$

∎

Remark 9 *Note that Theorem 8 implies that if $\mathbf{w}(0) = \mathbf{0}$ then $\|\mathbf{w}(\infty)\| \le 2 \cdot \|\mathbf{w}^*\|$. By Theorem 5, the implicit regularization cannot be expressed as a function of $\mathbf{w}$. Hence, while the implicit regularization of a ReLU neuron cannot be expressed exactly as a function of $\mathbf{w}$, it can be expressed approximately, within a factor of 2, by the ℓ_2 norm.*

6. Depth-2 ReLU networks

In Section 4, we showed that the implicit regularization of a single ReLU neuron is not expressible by any nontrivial function of $\mathbf{w}$. In this section we will show an analogous result for networks with one hidden ReLU neuron (namely, $\mathbf{x} \mapsto v \cdot \sigma(\langle \mathbf{x}, \mathbf{w} \rangle)$), which is the simplest case of a depth-2 ReLU network. Since our focus here is on impossibility results, then our results for this simple case imply impossibility results also for more general cases.

By Du et al. (2018), in feed-forward ReLU networks, gradient flow enforces the differences between square norms across different layers to remain invariant. Thus, if gradient flow starts from a point close to $\mathbf{0}$, then the magnitudes of all layers are automatically balanced. In the case of single-hidden-neuron networks, it implies that $\|\mathbf{w}\|$ (the norm of the weights vector of the first layer) is roughly equal to $|v|$ (the weight in the second layer). Hence, gradient flow induce a bias toward balanced layers. However, this bias is considered weak and does not allow us to derive generalization guarantees. Hence, there has been much effort to characterize the implicit regularization and to understand whether it is related to properties such as small norms, sparsity, or low ranks (cf. Gunasekar et al. (2018c); Razin and Cohen (2020); Arora et al. (2019); Woodworth et al. (2020);

Belabbas (2020); Li et al. (2020)). We show that in single-hidden-neuron networks, the *only* bias which can be specified by a regularization function $\mathcal{R}(\boldsymbol{\theta})$ (where $\boldsymbol{\theta}$ are the network parameters) is the balancedness property described above. Namely, $\mathcal{R}(\boldsymbol{\theta})$ is constant in the set of parameters that satisfy the balacedness property.

Recall that in our study of single-neuron networks in Section 4, we first analyzed the behavior of gradient flow for some specific inputs (Theorem 3), and then used this result in order to show that the implicit regularization function $\mathcal{R}$ is trivial (Theorem 5). Here, we proceed in a similar manner, but the first part is based on empirical results rather than on a theoretical proof. Thus, we first run gradient descent for some specific inputs, and observe empirically that it converges to points that clearly satisfy a certain technical property. Then, making the explicit assumption that the technical property holds, we prove that the regularization function $\mathcal{R}(\boldsymbol{\theta})$ is constant in the set of parameters that satisfy the balacedness property.

We now proceed to the formal results. Let $\boldsymbol{\theta} = (\mathbf{w}, v) \in \mathbb{R}^d \times \mathbb{R}$, and consider the neural network $N_{\boldsymbol{\theta}}(\mathbf{x}) = v \cdot \sigma(\langle \mathbf{x}, \mathbf{w} \rangle)$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is the ReLU function. Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be a training dataset, and let $X \in \mathbb{R}^{n \times d}$ be the corresponding data matrix. We analyze the implicit regularization of gradient flow on the objective

$$\mathcal{L}_{X,\mathbf{y}}(\boldsymbol{\theta}) := \frac{1}{2} \sum_{i=1}^n \left(N_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i \right)^2 . \tag{3}$$

We assume that the data is realizable, i.e., $\min_{\boldsymbol{\theta}} \mathcal{L}_{X,\mathbf{y}}(\boldsymbol{\theta}) = 0$. The run of gradient flow starts from $\boldsymbol{\theta}(0) = (\mathbf{w}(0), v(0))$, where $\mathbf{w}(0) = \mathbf{0}$, and $v(0)$ is a small number[3]. Let $\boldsymbol{\theta}^\epsilon(t)$ be the trajectory of gradient flow where $\boldsymbol{\theta}^\epsilon(0) = (\mathbf{0}, \epsilon)$ for some $\epsilon > 0$, and let $\boldsymbol{\theta}^\epsilon(\infty) = (\mathbf{w}^\epsilon(\infty), v^\epsilon(\infty)) := \lim_{t \to \infty} \boldsymbol{\theta}^\epsilon(t)$ (assuming that the limit exists). In addition, we define the *limit point* of gradient flow (on inputs $(X, \mathbf{y})$) as $\boldsymbol{\theta}^* = (\mathbf{w}^*, v^*) := \lim_{\epsilon \to 0^+} \boldsymbol{\theta}^\epsilon(\infty)$, assuming the limit exists. This follows a standard practice in analyzing the implicit regularization of gradient flow, where we consider the flow's limit assuming it starts infinitesimally close to $\mathbf{0}$ (see, e.g., Gunasekar et al. (2018c); Razin and Cohen (2020); Arora et al. (2019); Woodworth et al. (2020); Li et al. (2020)).

We are interested in understanding the properties of any possible implicit regularization function $\mathcal{R} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$: Namely, a function that for every input $(X, \mathbf{y}) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ to gradient flow, if $\boldsymbol{\theta}^*$ exists and $\mathcal{L}_{X,\mathbf{y}}(\boldsymbol{\theta}^*) = 0$, then we have $\boldsymbol{\theta}^* \in \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{R}(\boldsymbol{\theta})$ s.t. $\mathcal{L}_{X,\mathbf{y}}(\boldsymbol{\theta}) = 0$. Whereas for a single ReLU neuron, we showed that $\mathcal{R}$ is necessarily constant, the situation for a single hidden-neuron is more complex, because gradient flow is already known to induce the following nontrivial balacedness property[4]:

Lemma 10 (Du et al. (2018)) *Let $\boldsymbol{\theta}(t) = (\mathbf{w}(t), v(t))$ be the trajectory of gradient flow on Eq. 3, starting from some $\boldsymbol{\theta}(0) \in \mathbb{R}^d \times \mathbb{R}$. Then, $\frac{d}{dt} \left(\|\mathbf{w}\|^2 - v^2 \right) = 0$.*

As a consequence, we have the following corollary.

Corollary 11 *Let $\epsilon > 0$ and let $\boldsymbol{\theta}^\epsilon(t) = (\mathbf{w}(t), v(t))$. For every $t \geq 0$ we have $(v(t))^2 = \|\mathbf{w}(t)\|^2 + \epsilon^2 > 0$.*

---

3. Note that if we start from $(\mathbf{0}, 0)$ then gradient flow stays at this point indefinitely, since the gradient there is $\mathbf{0}$. This is avoided by making either the scalar $v(0)$ or the vector $\mathbf{w}(0)$ non-zero, and we focus on the former as it makes the analysis cleaner.

4. Du et al. (2018) showed the lemma for the more general case of fully-connected networks with homogeneous activation functions. See Appendix C for a simpler proof for our setting.

By Corollary 11, we have $v^\epsilon(\infty) = \sqrt{\|\mathbf{w}^\epsilon(\infty)\|^2 + \epsilon^2}$, and hence $v^* = \|\mathbf{w}^*\|$. As a result, not all parameters $(\mathbf{w}^*, v^*)$ can be reached by gradient flow – only those which satisfy the "balancedness" property above. However, this property in itself is rather weak, and does not necessarily induce bias towards properties which may aid in generalization, such as small norms or sparsity. Is there any stronger implicit regularization at play here? To answer that, it is natural to consider the behavior of the implicit regularization function $\mathcal{R}$ on the set $\mathcal{S} := \{(\mathbf{w}, v) \in \mathbb{R}^d \times \mathbb{R} : \|\mathbf{w}\| = v\}$ of "balanced" parameters. In what follows, we will argue that $\mathcal{R}$ is constant on $\mathcal{S} \setminus \{\mathbf{0}\}$. Thus, $\mathcal{R}$ may induce a bias toward $\mathcal{S}$, namely, enforce the "balancedness" of $\mathbf{w}^*$ and v^*, but it does not induce any additional bias within $\mathcal{S}$.

In order to show that $\mathcal{R}$ is constant in $\mathcal{S} \setminus \{\mathbf{0}\}$, we start with an empirical study of gradient descent for some specific inputs. Then, under a mild assumption based on the empirical results, we prove that $\mathcal{R}$ is constant.

6.1. Empirical results

In Theorem 3 we showed that the implicit regularization of a single ReLU neuron is not the ℓ_2 norm. We now demonstrate an analogous behavior in single-hidden-neuron networks.

Let $X_\gamma \in \mathbb{R}^{3 \times 3}$ be the matrix from Theorem 3. Let $\mathbf{y} \in \mathbb{R}^3$ be the vector from Theorem 3 with $\alpha = 1$, i.e., $\mathbf{y} = (16, 18, 0)^\top$. Let $\mathcal{Q} = \{(5, -1, q)^\top : q \le 0\}$. Note that for every $\mathbf{u} \in \mathcal{Q}$ and $\gamma \ge 0$ we have $\sigma(X_\gamma \mathbf{u}) = \mathbf{y}$. By Theorem 3, for different values of $\gamma \ge 0$, gradient flow on a single neuron with the input $(X_\gamma, \mathbf{y})$ converges to different points $\mathbf{w}^\gamma$ in a set of global minima. We now show empirically that a similar behavior appears also in single-hidden-neuron networks: For different values of $\gamma \ge 0$, gradient descent with the input $(X_\gamma, \mathbf{y})$, starting from $(\mathbf{0}, \epsilon)$ for a fixed $\epsilon > 0$, converges to different points $\boldsymbol{\theta}^\gamma = (\mathbf{w}^\gamma, v^\gamma)$ with $v^\gamma \mathbf{w}^\gamma \in \mathcal{Q}$. We denote $\mathbf{u}^\gamma = v^\gamma \mathbf{w}^\gamma$.
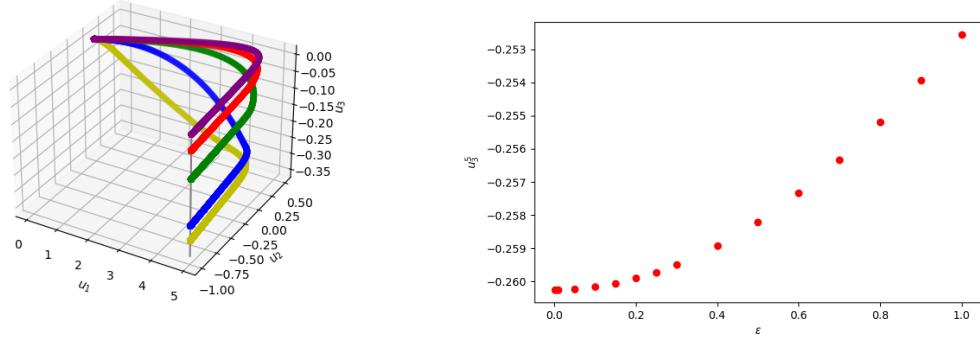
In Figure 2 (left), we show the trajectories $\mathbf{u}(t) = v(t) \cdot \mathbf{w}(t)$ of gradient descent starting from $\boldsymbol{\theta}(0) = (\mathbf{0}, 0.001)$, for $\gamma \in \{0, 1, 2, 5, 20\}$. The grey line denotes the set $\mathcal{Q}$. We note that while the trajectories in Figure 2 (left) have similar shapes to the trajectories shown in Figure 1 for the case of a single-neuron network, they are not identical. Indeed, the dynamics of the problems are different. This experiment demonstrates that for $\epsilon = 0.001$ and different values of $\gamma \ge 0$, gradient descent converges to different points $\mathbf{u}^\gamma \in \mathcal{Q}$.

Since we are interested in gradient flow starting from $(\mathbf{0}, \epsilon)$ for $\epsilon \to 0$, we need to consider $\mathbf{u}^\gamma$ where $\epsilon \to 0$. Consider $\gamma \in \{0, 5\}$, and let $\mathbf{u}^0 = (u_1^0, u_2^0, u_3^0)$ and $\mathbf{u}^5 = (u_1^5, u_2^5, u_3^5)$. We now show empirical evidence that strongly supports the following assumption, which will be the key used to prove that the implicit regularization function is constant as discussed earlier:

Assumption 12 *For every sufficiently small $\epsilon > 0$, gradient flow starting from $(\mathbf{0}, \epsilon)$ on the inputs $(X_0, \mathbf{y})$ and $(X_5, \mathbf{y})$, converges to zero loss. Moreover, $\lim_{\epsilon \to 0^+} u_3^5$ exists and is negative.*

We turn to present this empirical evidence. Since we are interested in simulating gradient flow, we used gradient descent with a small learning rate of 10^{-5}. We ran gradient descent starting from $(\mathbf{0}, \epsilon)$, for $\gamma \in \{0, 5\}$ and $\epsilon \in \{10^{-i} : i = 0, \ldots, 5\}$. In all cases we reached loss smaller than 10^{-15}, which suggests that the first part of the assumption holds. We now turn to the second part of the assumption. In Figure 2 (right) we show for different values of ϵ, the value of u_3^5 in the last iteration of gradient descent starting from $(\mathbf{0}, \epsilon)$ and reaching loss smaller than 10^{-15}. This experiment suggests that $\lim_{\epsilon \to 0^+} u_3^5$ exists and is negative.

Finally, we show that Assumption 12 easily implies the following corollary:

11

(a) The trajectories $\mathbf{u}(t)$ for the input $(X_\gamma, \mathbf{y})$ starting from $\boldsymbol{\theta}(0) = (\mathbf{0}, 0.001)$, for $\gamma = 0$ (purple), $\gamma = 1$ (red), $\gamma = 2$ (green), $\gamma = 5$ (blue), and $\gamma = 20$ (yellow).

(b) The value of u_3^5 in the last iteration of gradient descent starting from $(\mathbf{0}, \epsilon)$ and reaching loss smaller than 10^{-15}.

Figure 2: Experiments on single-hidden-neuron networks.

Corollary 13 *Under Assumption 12, for the inputs $(X_0, \mathbf{y})$ and $(X_5, \mathbf{y})$, the limit points of gradient flow are $\boldsymbol{\theta}^* = (\mathbf{w}^*, v^*)$ and $\tilde{\boldsymbol{\theta}}^* = (\tilde{\mathbf{w}}^*, \tilde{v}^*)$ (respectively), such that $v^*\mathbf{w}^* = (5, -1, 0)^\top$ and $\tilde{v}^*\tilde{\mathbf{w}}^* = (5, -1, s)^\top$ for some $s < 0$.*

Proof First, since by Assumption 12 for every sufficiently small $\epsilon > 0$ the loss converges to zero, then for every sufficiently small $\epsilon > 0$ we have $(u_1^0, u_2^0) = (u_1^5, u_2^5) = (5, -1)$, and hence we also have $\lim_{\epsilon \to 0^+}(u_1^0, u_2^0) = \lim_{\epsilon \to 0^+}(u_1^5, u_2^5) = (5, -1)$. Clearly, for every $\epsilon > 0$ we have $u_3^0 = 0$, since if $\gamma = 0$ then the third component of $\mathbf{u}(t)$ is 0 for every $t \geq 0$. Therefore, $\lim_{\epsilon \to 0^+} u_3^0 = 0$. Finally, by Assumption 12, $\lim_{\epsilon \to 0^+} u_3^5$ exists and is negative. Overall, we have $\lim_{\epsilon \to 0^+} \mathbf{u}^5 = (5, -1, s)^\top$ for some $s < 0$, and $\lim_{\epsilon \to 0^+} \mathbf{u}^0 = (5, -1, 0)^\top$. ∎

6.2. $\mathcal{R}$ is constant in the set of balanced parameters

We now show that the regularization function $\mathcal{R}$ is constant in the set $\mathcal{S}$ of balanced parameters. Since the limit point $\boldsymbol{\theta}^* = (\mathbf{w}^*, v^*) \in \mathbb{R}^d \times \mathbb{R}$ of gradient flow is in $\mathcal{S}$, then it is uniquely determined by the point $v^*\mathbf{w}^* \in \mathbb{R}^d$, and hence, intuitively, we can think about $\boldsymbol{\theta}^*$ as a point in $\mathbb{R}^d$. More formally, let $\Psi : \mathbb{R}^d \to \mathcal{S}$ be such that $\Psi(\mathbf{u}) = \left(\frac{\mathbf{u}}{\sqrt{\|\mathbf{u}\|}}, \sqrt{\|\mathbf{u}\|} \right)$ for $\mathbf{u} \neq \mathbf{0}$, and $\Psi(\mathbf{0}) = (\mathbf{0}, 0)$. Note that for every $(\mathbf{w}, v) \in \mathcal{S}$ we have $\Psi(v\mathbf{w}) = (\mathbf{w}, v)$, and that Ψ is bijective. Let $\mathcal{R}' : \mathbb{R}^d \to \mathbb{R}$ be such that $\mathcal{R}'(\mathbf{u}) = \mathcal{R}(\Psi(\mathbf{u}))$. We state our main theorem on single-hidden-neuron networks:

Theorem 14 *Under Assumption 12, the function $\mathcal{R}' : \mathbb{R}^3 \to \mathbb{R}$ is constant in $\mathbb{R}^3 \setminus \{\mathbf{0}\}$. Therefore, $\mathcal{R} : \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}$ is constant in $\mathcal{S} \setminus \{\mathbf{0}\}$.*

We prove the theorem in Appendix D. Intuitively, the theorem follows from the following argument. By Corollary 13, for the input $(X_0, \mathbf{y})$, gradient flow converges to $\boldsymbol{\theta}^*$ such that $\mathbf{u}^* :=$

$\Psi^{-1}(\boldsymbol{\theta}^*) = (5, -1, 0)^\top$, even though every $\mathbf{u} \in \mathcal{Q} = \{(5, -1, q)^\top : q \leq 0\}$ corresponds to a zero-loss solution. Hence, gradient flow prefers $\mathbf{u}^*$ over all other points in $\mathcal{Q}$. Similarly, for the input $(X_5, \mathbf{y})$ gradient flow converges to $\tilde{\boldsymbol{\theta}}^*$ such that $\tilde{\mathbf{u}}^* := \Psi^{-1}(\tilde{\boldsymbol{\theta}}^*) = (5, -1, s)^\top$ for some $s < 0$, even though every $\mathbf{u} \in \mathcal{Q}$ corresponds to a zero-loss solution. Hence, it prefers $\tilde{\mathbf{u}}^*$ over all other points in $\mathcal{Q}$. It implies that $\mathcal{R}'(\mathbf{u}^*) = \mathcal{R}'(\tilde{\mathbf{u}}^*) = \min_{\mathbf{u} \in \mathcal{Q}} \mathcal{R}'(\mathbf{u})$. Then, by rotating and scaling the inputs to gradient flow, and by using Lemma 6, we show that $\mathcal{R}'$ is constant in $\mathbb{R}^3 \setminus \{\mathbf{0}\}$.

Remark 15 ($\mathcal{R}$ is constant in $\mathcal{S} \setminus \{\mathbf{0}\}$ also for $d > 3$) *Theorem 14 implies that $\mathcal{R}' : \mathbb{R}^d \to \mathbb{R}$ is trivial for $d = 3$. By padding the inputs from Assumption 12 with zeros, we can obtain a d-dimensional version of Assumption 12 for $d > 3$. Then, using the same arguments as above, it implies that Theorem 14 holds for every $d \geq 3$.*

Acknowledgments

References

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.

Mohamed Ali Belabbas. On implicit regularization: Morse functions and applications to matrix factorization. *arXiv preprint arXiv:2001.04264*, 2020.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

Assaf Dauber, Meir Feder, Tomer Koren, and Roi Livni. Can implicit bias explain generalization? stochastic convex optimization as a case study. *arXiv preprint arXiv:2003.06152*, 2020.

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.

Armin Eftekhari and Konstantinos Zygalakis. Implicit regularization in matrix sensing: A geometric view leads to stronger results. *arXiv preprint arXiv:2008.12091*, 2020.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, pages 3202–3211, 2019.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018a.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018b.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018c.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018a.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018b.

Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136. PMLR, 2020.

Hui Jin and Guido Montúfar. Implicit bias of gradient descent for mean squared error regression with wide neural networks. 2020.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *arXiv preprint arXiv:2012.09839*, 2020.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

Edward Moroshko, Blake E Woodworth, Suriya Gunasekar, Jason D Lee, Nati Srebro, and Daniel Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. *Advances in Neural Information Processing Systems*, 33, 2020.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960, 2019.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *arXiv preprint arXiv:2005.06398*, 2020.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Ohad Shamir. Gradient methods never overfit on separable data. *arXiv preprint arXiv:2007.00028*, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. *Advances in Neural Information Processing Systems*, 31:10608–10619, 2018.

Francis Williams, Matthew Trager, Daniele Panozzo, Claudio Silva, Denis Zorin, and Joan Bruna. Gradient dynamics of shallow univariate relu networks. In *Advances in Neural Information Processing Systems*, pages 8378–8387, 2019.

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. *arXiv preprint arXiv:2002.09277*, 2020.

Tengyu Xu, Yi Zhou, Kaiyi Ji, and Yingbin Liang. When will gradient methods converge to max-margin classifier under relu models? *arXiv preprint arXiv:1806.04339*, 2018.

Gilad Yehudai and Ohad Shamir. Learning a single neuron with gradient methods. *arXiv preprint arXiv:2001.05205*, 2020.

Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. *arXiv preprint arXiv:2010.02501*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Appendix A. Proof of Theorem 3

Lemma 16 *Let $A \in \mathbb{R}^{d \times d}$ be an invertible matrix, and let $\mathbf{y} \in \mathbb{R}^d$. Consider the dynamics of $\mathbf{w}(t)$ given by the differential equation*

$$\dot{\mathbf{w}}(t) = -A^\top (A\mathbf{w}(t) - \mathbf{y}) .$$

Then, we have

$$\mathbf{w}(t) = \exp(-tA^\top A) \left(\mathbf{w}(0) - A^{-1}\mathbf{y} \right) + A^{-1}\mathbf{y} , \tag{4}$$

where $\exp(\cdot)$ denotes the matrix exponential.

The proof of Lemma 16 follows easily by differentiating the expression in Eq. 4.

A.1. Proof for $\gamma = 0$

The trajectory $\mathbf{w}(t)$ obeys the dynamics

$$\dot{\mathbf{w}}(t) = -\nabla\mathcal{L}(\mathbf{w}(t)) = -\sum_{i=1}^{3} \left(\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) - y_i\right)\sigma'(\mathbf{x}_i^\top \mathbf{w}(t))\mathbf{x}_i$$

$$= -\sum_{i=1}^{2} \left(\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) - y_i\right)\sigma'(\mathbf{x}_i^\top \mathbf{w}(t))\mathbf{x}_i \ ,$$

where the last equality is since $\mathbf{x}_3 = \mathbf{0}$. Hence, since we also have $\mathbf{w}(0) = \mathbf{0}$, then $\mathbf{w}(t)$ is spanned by $\mathbf{x}_1, \mathbf{x}_2$. Note that the third components in $\mathbf{x}_1, \mathbf{x}_2$ are 0, and therefore the third component in $\mathbf{w}(t)$ is 0 for all $t \geq 0$. For a vector $\mathbf{v} \in \{\mathbf{w}(t), \mathbf{x}_1, \mathbf{x}_2, \mathbf{y}\}$, we denote $\tilde{\mathbf{v}} = (v_1, v_2) \in \mathbb{R}^2$. We also denote by $\tilde{X} \in \mathbb{R}^{2\times 2}$ the matrix whose rows are $\tilde{\mathbf{x}}_1^\top$ and $\tilde{\mathbf{x}}_2^\top$. Note that for $i \in \{1, 2\}$ we have $\mathbf{x}_i^\top \mathbf{w}(t) = \tilde{\mathbf{x}}_i^\top \tilde{\mathbf{w}}(t)$. Thus, we have

$$\dot{\tilde{\mathbf{w}}}(t) = -\sum_{i=1}^{2} \left(\sigma(\tilde{\mathbf{x}}_i^\top \tilde{\mathbf{w}}(t)) - y_i\right)\sigma'(\tilde{\mathbf{x}}_i^\top \tilde{\mathbf{w}}(t))\tilde{\mathbf{x}}_i \ . \tag{5}$$

If $\tilde{\mathbf{x}}_i^\top \tilde{\mathbf{w}}(t) \geq 0$ for $i \in \{1, 2\}$ then the above equals

$$-\sum_{i=1}^{2} (\tilde{\mathbf{x}}_i^\top \tilde{\mathbf{w}}(t) - y_i)\tilde{\mathbf{x}}_i = -\tilde{X}^\top (\tilde{X}\tilde{\mathbf{w}}(t) - \tilde{\mathbf{y}}) \ .$$

By Lemma 16, the solution to the above equation is

$$\tilde{\mathbf{w}}(t) = \exp\left(-t\tilde{X}^\top \tilde{X}\right)\left(\tilde{\mathbf{w}}(0) - \tilde{X}^{-1}\tilde{\mathbf{y}}\right) + \tilde{X}^{-1}\tilde{\mathbf{y}} \ . \tag{6}$$

Let $\tilde{\mathcal{B}} = \{\tilde{\mathbf{w}} \in \mathbb{R}^2 : \tilde{\mathbf{x}}_1^\top \tilde{\mathbf{w}} \geq 0, \tilde{\mathbf{x}}_2^\top \tilde{\mathbf{w}} \geq 0\}$. We will show that the trajectory from Eq. 6 satisfies $\tilde{\mathbf{w}}(t) \in \tilde{\mathcal{B}}$ for every $t \geq 0$. Therefore, it satisfies Eq. 5 for every $t \geq 0$. It implies that $\tilde{\mathbf{w}}(\infty) = \tilde{X}^{-1}\tilde{\mathbf{y}} = \alpha \cdot (5, -1)^\top$, and therefore $\mathbf{w}(\infty) = \alpha \cdot (5, -1, 0)^\top$ as required.

Lemma 17 *The trajectory from Eq. 6 satisfies $\tilde{\mathbf{w}}(t) \in \tilde{\mathcal{B}}$ for every $t \geq 0$.*

Proof The eigenvalues of $\tilde{X}^\top \tilde{X}$ are $\tilde{\lambda}_1 = 15 + 5\sqrt{5}$ and $\tilde{\lambda}_2 = 15 - 5\sqrt{5}$. Let $\tilde{D} = \mathrm{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2)$, and let $\tilde{U} \in \mathbb{R}^{2\times 2}$ such that $\tilde{X}^\top \tilde{X} = \tilde{U}\tilde{D}\tilde{U}^\top$. By Eq. 6 and since $\tilde{\mathbf{w}}(0) = \mathbf{0}$, we have

$$\tilde{\mathbf{w}}(t) = \tilde{U}\exp\left(-t\tilde{D}\right)\tilde{U}^\top\left(\mathbf{0} - \alpha(5, -1)^\top\right) + \alpha(5, -1)^\top \ ,$$

where $\exp\left(-t\tilde{D}\right) = \mathrm{diag}(e^{-\tilde{\lambda}_1 t}, e^{-\tilde{\lambda}_2 t})$.

By straightforward calculations we obtain

$$\frac{1}{\alpha} \cdot \tilde{\mathbf{x}}_1^\top \tilde{\mathbf{w}}(t) = -(3, -1)\tilde{U}\exp\left(-t\tilde{D}\right)\tilde{U}^\top (5, -1)^\top + (3, -1)(5, -1)^\top$$

$$= -\left(8 + 2\sqrt{5}\right)e^{-\tilde{\lambda}_1 t} - \left(8 - 2\sqrt{5}\right)e^{-\tilde{\lambda}_2 t} + 16$$

$$\geq -\left(8 + 2\sqrt{5}\right) - \left(8 - 2\sqrt{5}\right) + 16 = 0 \ .$$

Moreover, we have

$$\frac{1}{\alpha} \cdot \tilde{\mathbf{x}}_2^\top \tilde{\mathbf{w}}(t) = -(4,2)\tilde{U} \exp\left(-t\tilde{D}\right) \tilde{U}^\top (5,-1)^\top + (4,2)(5,-1)^\top$$

$$= -\left(5\sqrt{5}+9\right) e^{-\tilde{\lambda}_1 t} + \left(5\sqrt{5}-9\right) e^{-\tilde{\lambda}_2 t} + 18 .$$

Since $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq 0$, then the above is at least

$$-\left(5\sqrt{5}+9\right) e^{-\tilde{\lambda}_1 t} + \left(5\sqrt{5}-9\right) e^{-\tilde{\lambda}_1 t} + 18 = -18e^{-\tilde{\lambda}_1 t} + 18 \geq 0 .$$

∎

A.2. Proof for $\gamma > 0$

We show that the trajectory $\mathbf{w}(t)$ consists of two parts. In the first part, where $t \in [0, t_1]$ for some $t_1 > 0$, we have $\mathbf{x}_i^\top \mathbf{w}(t) \geq 0$ for every $1 \leq i \leq 3$. Then, for $t \in [t_1, \infty)$, we have $\mathbf{x}_1^\top \mathbf{w}(t) \geq 0$, $\mathbf{x}_2^\top \mathbf{w}(t) \geq 0$, and $\mathbf{x}_3^\top \mathbf{w}(t) \leq 0$.

A.2.1. PART 1

Recall that the trajectory $\mathbf{w}(t)$ obeys the dynamics

$$\dot{\mathbf{w}}(t) = -\nabla \mathcal{L}(\mathbf{w}(t)) = -\sum_{i=1}^{3} \left(\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) - y_i\right) \sigma'(\mathbf{x}_i^\top \mathbf{w}(t))\mathbf{x}_i . \tag{7}$$

If $\mathbf{x}_i^\top \mathbf{w}(t) \geq 0$ for every $1 \leq i \leq 3$ then the above equals

$$-\sum_{i=1}^{3}(\mathbf{x}_i^\top \mathbf{w}(t) - y_i)\mathbf{x}_i = -X^\top (X\mathbf{w}(t) - \mathbf{y}) .$$

By Lemma 16, the solution to the above equation is

$$\mathbf{w}(t) = \exp\left(-tX^\top X\right) \left(\mathbf{w}(0) - X^{-1}\mathbf{y}\right) + X^{-1}\mathbf{y} . \tag{8}$$

Let $\mathcal{B} = \{\mathbf{w} \in \mathbb{R}^3 : \mathbf{x}_i^\top \mathbf{w} \geq 0, \ i = 1\ldots 3\}$. We will show that the trajectory from Eq. 8 satisfies $\mathbf{w}(t) \in \mathcal{B}$ for every $t \in [0, t_1]$ for some $t_1 > 0$. Therefore, it satisfies Eq. 7 for $t \in [0, t_1]$. Hence, gradient flow follows the trajectory from Eq. 8 where $t \in [0, t_1]$. We will also investigate the values of the components of $\mathbf{w}(t_1)$, since it is the starting point of the second part of the trajectory.

Note that for every γ we have $X\alpha(5,-1,1)^\top = \alpha(16,18,0)^\top = \mathbf{y}$, and therefore $X^{-1}\mathbf{y} = \alpha(5,-1,1)^\top$. Since we also have $\mathbf{w}(0) = \mathbf{0}$, then Eq. 8 implies

$$\mathbf{w}(t) = \alpha(5,-1,1)^\top - \exp\left(-tX^\top X\right) \alpha(5,-1,1)^\top .$$

Let $\lambda_1, \lambda_2, \lambda_3$ be the eigenvalues of $X^\top X$, let $D = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$, and let $U \in \mathbb{R}^{3\times 3}$ such that $X^\top X = UDU^\top$. Thus, we have

$$\mathbf{w}(t) = \alpha(5,-1,1)^\top - U \exp(-tD)U^\top \alpha(5,-1,1)^\top , \tag{9}$$

where $\exp(-tD) = \text{diag}\left(e^{-\lambda_1 t}, e^{-\lambda_2 t}, e^{-\lambda_3 t}\right)$.

Lemma 18 *Let $\mathbf{w}(t)$ be the trajectory from Eq. 9. For every $\gamma \in \{1, 2, 5\}$, there exists $t_1 > 0$, such that $\mathbf{w}(t) \in \mathcal{B}$ for every $t \in [0, t_1]$, $\mathbf{x}_3^\top \mathbf{w}(t_1) = 0$, and $\mathbf{w}(t_1) = \alpha \cdot (a, b, -b)$ for some a, b (independent of α), where $a \in (4, 5)$ and b satisfies the following:*

- *If $\gamma = 1$ then $b \in (0.035, 0.045)$.*

- *If $\gamma = 2$ then $b \in (0.1, 0.12)$.*

- *If $\gamma = 5$ then $b \in (0.2, 0.22)$.*

We defer the proof of Lemma 18 to later sections (the proof for $\gamma = 1$ is given in Section A.2.3, for $\gamma = 2$ in Section A.2.4, and for $\gamma = 5$ in Section A.2.5).

A.2.2. PART 2

In Lemma 18 we established the trajectory for $t \in [0, t_1]$. We need to find the trajectory that satisfies the dynamics in Eq. 7 for $t \in [t_1, \infty)$, with the initial value $\mathbf{w}(t_1)$ obtained in Lemma 18 .

Let $\mathcal{B}' = \{\mathbf{w} \in \mathbb{R}^3 : \mathbf{x}_1^\top \mathbf{w} \geq 0, \mathbf{x}_2^\top \mathbf{w} \geq 0, \mathbf{x}_3^\top \mathbf{w} \leq 0\}$. Consider the dynamics from Eq. 7, where $\mathbf{w}(t) \in \mathcal{B}'$. Since $\sigma(\mathbf{x}_3^\top \mathbf{w}(t)) = 0$ and $y_3 = 0$, and since $\mathbf{x}_i^\top \mathbf{w}(t) \geq 0$ for $i \in \{1, 2\}$, then we have

$$
\begin{aligned}
\dot{\mathbf{w}}(t) &= -\sum_{i=1}^{3} \left(\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) - y_i\right) \sigma'(\mathbf{x}_i^\top \mathbf{w}(t))\mathbf{x}_i \\
&= -\sum_{i=1}^{2} \left(\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) - y_i\right) \sigma'(\mathbf{x}_i^\top \mathbf{w}(t))\mathbf{x}_i \\
&= -\sum_{i=1}^{2} (\mathbf{x}_i^\top \mathbf{w}(t) - y_i)\mathbf{x}_i \;.
\end{aligned}
$$

We will find the trajectory $\mathbf{w}(t)$ that satisfies

$$
\dot{\mathbf{w}}(t) = -\sum_{i=1}^{2} (\mathbf{x}_i^\top \mathbf{w}(t) - y_i)\mathbf{x}_i \;, \tag{10}
$$

with the initial value $\mathbf{w}(t_1)$, and show that $\mathbf{w}(t) \in \mathcal{B}'$ for every $t \in [t_1, \infty)$. Such a trajectory obeys the dynamics in Eq. 7 for $t \in [t_1, \infty)$, with the initial value $\mathbf{w}(t_1)$, as required.

For a vector $\mathbf{v} \in \{\mathbf{w}(t), \mathbf{x}_1, \mathbf{x}_2, \mathbf{y}\}$, we denote $\tilde{\mathbf{v}} = (v_1, v_2) \in \mathbb{R}^2$. We also denote by $\tilde{X} \in \mathbb{R}^{2 \times 2}$ the matrix whose rows are $\tilde{\mathbf{x}}_1^\top$ and $\tilde{\mathbf{x}}_2^\top$. Since both $\mathbf{x}_1$ and $\mathbf{x}_2$ have 0 as their third components, then $\mathbf{w}(t)$ satisfies Eq. 10 iff the third component $w_3(t)$ satisfies

$$
\dot{w}_3(t) = 0 \;, \tag{11}
$$

and the first two components $\tilde{\mathbf{w}}(t)$ satisfy

$$
\dot{\tilde{\mathbf{w}}}(t) = -\sum_{i=1}^{2} (\tilde{\mathbf{x}}_i^\top \tilde{\mathbf{w}}(t) - y_i)\tilde{\mathbf{x}}_i = -\tilde{X}^\top (\tilde{X}\tilde{\mathbf{w}}(t) - \tilde{\mathbf{y}}) \;. \tag{12}
$$

By Lemma 16, the trajectory that follows from Eq. 12 is

$$\tilde{\mathbf{w}}(t) = \exp\left(-(t - t_1)\tilde{X}^\top \tilde{X}\right)\left(\tilde{\mathbf{w}}(t_1) - \tilde{X}^{-1}\tilde{\mathbf{y}}\right) + \tilde{X}^{-1}\tilde{\mathbf{y}}$$

$$= \exp\left(-(t - t_1)\tilde{X}^\top \tilde{X}\right)\left(\tilde{\mathbf{w}}(t_1) - \alpha(5, -1)^\top\right) + \alpha(5, -1)^\top . \qquad (13)$$

Consider the trajectory $\mathbf{w}(t)$ for $t \geq t_1$, such that the first two components follow the trajectory $\tilde{\mathbf{w}}(t)$ from the above equation, and the third component is constant, i.e., $w_3(t) = w_3(t_1)$ for all $t \geq t_1$. Note that this trajectory satisfies Eq. 11 and Eq. 12, and hence satisfies Eq. 10. It also satisfies the initial condition $\mathbf{w}(t_1)$. Thus, it remains to show that $\mathbf{w}(t) \in \mathcal{B}'$ for every $t \in [t_1, \infty)$. Since by Eq.13 we have $\tilde{\mathbf{w}}(\infty) = \alpha(5, -1)^\top$, then we have $\mathbf{w}(\infty) = (5\alpha, -\alpha, w_3(t_1))^\top$. By Lemma 18, it follows that $w_3(t_1)$ is in the required interval, and thus it completes the proof of the theorem.

The eigenvalues of $\tilde{X}^\top \tilde{X}$ are $\tilde{\lambda}_1 = 15 + 5\sqrt{5}$ and $\tilde{\lambda}_2 = 15 - 5\sqrt{5}$. Let $\tilde{D} = \mathrm{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2)$, and let $\tilde{U} \in \mathbb{R}^{2 \times 2}$ such that $\tilde{X}^\top \tilde{X} = \tilde{U}\tilde{D}\tilde{U}^\top$. By Eq. 13, we have

$$\tilde{\mathbf{w}}(t) = \tilde{U} \exp\left(-(t - t_1)\tilde{D}\right)\tilde{U}^\top \left(\tilde{\mathbf{w}}(t_1) - \alpha(5, -1)^\top\right) + \alpha(5, -1)^\top , \qquad (14)$$

where $\exp\left(-(t - t_1)\tilde{D}\right) = \mathrm{diag}\left(e^{-\tilde{\lambda}_1(t-t_1)}, e^{-\tilde{\lambda}_2(t-t_1)}\right)$.

From the following two lemmas it follows that $\mathbf{w}(t) \in \mathcal{B}'$ for all $t \geq t_1$.

Lemma 19 *For every $t \geq t_1$ and $i \in \{1, 2\}$, we have $\mathbf{x}_i^\top \mathbf{w}(t) \geq 0$.*

Proof Since $\mathbf{x}_1$ and $\mathbf{x}_2$ have 0 in their third components, then it suffices to show that $\tilde{\mathbf{x}}_i^\top \tilde{\mathbf{w}}(t) \geq 0$ for every $i \in \{1, 2\}$ and $t \geq t_1$. By Lemma 18, we have $\mathbf{w}(t_1) = \alpha \cdot (a, b, -b)^\top$, where $a \in (4, 5)$ and $b \in (0, 1)$.

For every $t \geq t_1$, by straightforward calculations we obtain

$$\frac{1}{\alpha} \cdot \tilde{\mathbf{x}}_1^\top \tilde{\mathbf{w}}(t) = (3, -1)\tilde{U} \exp\left(-(t - t_1)\tilde{D}\right)\tilde{U}^\top (a - 5, b + 1)^\top + (3, -1)(5, -1)^\top$$

$$= \left(\frac{(3 + \sqrt{5})a}{2} + \frac{(\sqrt{5} - 1)b}{2} - 8 - 2\sqrt{5}\right)e^{-\tilde{\lambda}_1(t-t_1)}$$

$$+ \left(\frac{(3 - \sqrt{5})a}{2} - \frac{(\sqrt{5} + 1)b}{2} - 8 + 2\sqrt{5}\right)e^{-\tilde{\lambda}_2(t-t_1)} + 16 .$$

Since $a \in (4, 5)$ and $b \in (0, 1)$, the above is at least

$$\left(\frac{(3 + \sqrt{5}) \cdot 4}{2} + \frac{(\sqrt{5} - 1) \cdot 0}{2} - 8 - 2\sqrt{5}\right)e^{-\tilde{\lambda}_1(t-t_1)}$$

$$+ \left(\frac{(3 - \sqrt{5}) \cdot 4}{2} - \frac{(\sqrt{5} + 1) \cdot 1}{2} - 8 + 2\sqrt{5}\right)e^{-\tilde{\lambda}_2(t-t_1)} + 16$$

$$= -2e^{-\tilde{\lambda}_1(t-t_1)} - \left(2 + \frac{\sqrt{5} + 1}{2}\right)e^{-\tilde{\lambda}_2(t-t_1)} + 16$$

$$\geq -2 - 2 - \frac{\sqrt{5} + 1}{2} + 16 \geq 0 .$$

Next, we have

$$\frac{1}{\alpha} \cdot \tilde{\mathbf{x}}_2^\top \tilde{\mathbf{w}}(t) = (4,2)\tilde{U}\exp\left(-(t-t_1)\tilde{D}\right)\tilde{U}^\top(a-5, b+1)^\top + (4,2)(5,-1)^\top$$

$$= \left((\sqrt{5}+2)a + b - 9 - 5\sqrt{5}\right)e^{-\tilde{\lambda}_1(t-t_1)} + \left(-(\sqrt{5}-2)a + b - 9 + 5\sqrt{5}\right)e^{-\tilde{\lambda}_2(t-t_1)} + 18$$

$$\geq \left((\sqrt{5}+2)\cdot 4 + 0 - 9 - 5\sqrt{5}\right)e^{-\tilde{\lambda}_1(t-t_1)} + \left(-(\sqrt{5}-2)\cdot 5 + 0 - 9 + 5\sqrt{5}\right)e^{-\tilde{\lambda}_2(t-t_1)} + 18$$

$$= -(\sqrt{5}+1)e^{-\tilde{\lambda}_1(t-t_1)} + e^{-\tilde{\lambda}_2(t-t_1)} + 18 \geq -(\sqrt{5}+1) + 18 \geq 0 \,.$$

$\blacksquare$

Lemma 20 *For every $t \geq t_1$, we have $\mathbf{x}_3^\top \mathbf{w}(t) \leq 0$.*

Proof By Lemma 18, we have $\mathbf{w}(t_1) = \alpha \cdot (a, b, -b)^\top$, where $a \in (4,5)$ and $b \in (0,1)$. The first two components in $\mathbf{w}(t)$ are $\tilde{\mathbf{w}}(t)$, and the third component is $-\alpha b$. Hence, in order to show that

$$\mathbf{x}_3^\top \mathbf{w}(t) = (0, \gamma, \gamma)\mathbf{w}(t) = (0, \gamma)\tilde{\mathbf{w}}(t) - \alpha b\gamma \leq 0 \,,$$

it suffices to show that $\frac{1}{\alpha} \cdot (0,1)\tilde{\mathbf{w}}(t) \leq b$.

We have

$$\frac{1}{\alpha} \cdot (0,1)\tilde{\mathbf{w}}(t) = (0,1)\tilde{U}\exp\left(-(t-t_1)\tilde{D}\right)\tilde{U}^\top(a-5, b+1)^\top + (0,1)(5,-1)^\top$$

$$= \left(\frac{\sqrt{5}a}{10} + \left(\frac{1}{2} - \frac{\sqrt{5}}{5}\right)b + \frac{1}{2} - \frac{7\sqrt{5}}{10}\right)e^{-\tilde{\lambda}_1(t-t_1)}$$

$$+ \left(-\frac{\sqrt{5}a}{10} + \left(\frac{1}{2} + \frac{\sqrt{5}}{5}\right)b + \frac{1}{2} + \frac{7\sqrt{5}}{10}\right)e^{-\tilde{\lambda}_2(t-t_1)} - 1 \,. \qquad (15)$$

Note that the above equals to

$$\frac{\sqrt{5}a}{10}\left(e^{-\tilde{\lambda}_1(t-t_1)} - e^{-\tilde{\lambda}_2(t-t_1)}\right) + \left(\left(\frac{1}{2} - \frac{\sqrt{5}}{5}\right)b + \frac{1}{2} - \frac{7\sqrt{5}}{10}\right)e^{-\tilde{\lambda}_1(t-t_1)}$$

$$+ \left(\left(\frac{1}{2} + \frac{\sqrt{5}}{5}\right)b + \frac{1}{2} + \frac{7\sqrt{5}}{10}\right)e^{-\tilde{\lambda}_2(t-t_1)} - 1 \,.$$

Since $\tilde{\lambda}_1 > \tilde{\lambda}_2 > 0$, then the above expression is monotonically descreasing w.r.t. a. Also, since $a \in (4,5)$, then by plugging in $a = 4$ to Eq. 15 we obtain

$$\frac{1}{\alpha} \cdot (0,1)\tilde{\mathbf{w}}(t) \leq \left(-\frac{3\sqrt{5}}{10} + \left(\frac{1}{2} - \frac{\sqrt{5}}{5}\right)b + \frac{1}{2}\right)e^{-\tilde{\lambda}_1(t-t_1)}$$

$$+ \left(\frac{3\sqrt{5}}{10} + \left(\frac{1}{2} + \frac{\sqrt{5}}{5}\right)b + \frac{1}{2}\right)e^{-\tilde{\lambda}_2(t-t_1)} - 1 \,.$$

Let $g(t)$ denote the r.h.s. of the above inequality. Note that $g(t_1) = b$. Hence, in order to show that $g(t) \leq b$ for every $t \geq t_1$, it suffices to show that $g'(t) \leq 0$ for all $t \geq t_1$.

We denote

$$C_1 = -\frac{3\sqrt{5}}{10} + \left(\frac{1}{2} - \frac{\sqrt{5}}{5}\right) b + \frac{1}{2} \,,$$

$$C_2 = \frac{3\sqrt{5}}{10} + \left(\frac{1}{2} + \frac{\sqrt{5}}{5}\right) b + \frac{1}{2} \,.$$

Thus, $g(t) = C_1 e^{-\tilde{\lambda}_1(t-t_1)} + C_2 e^{-\tilde{\lambda}_2(t-t_1)} - 1$. Since $b \in (0,1)$, it is easy to verify that $C_1 < 0$ and $C_2 > 0$. Thus,

$$g'(t) = -\tilde{\lambda}_1 C_1 e^{-\tilde{\lambda}_1(t-t_1)} - \tilde{\lambda}_2 C_2 e^{-\tilde{\lambda}_2(t-t_1)}$$
$$\leq -\tilde{\lambda}_1 C_1 e^{-\tilde{\lambda}_1(t-t_1)} - \tilde{\lambda}_2 C_2 e^{-\tilde{\lambda}_1(t-t_1)}$$
$$= \left[-\tilde{\lambda}_1 C_1 - \tilde{\lambda}_2 C_2\right] e^{-\tilde{\lambda}_1(t-t_1)} \,.$$

We have

$$-\tilde{\lambda}_1 C_1 - \tilde{\lambda}_2 C_2 = -(15 + 5\sqrt{5}) \left(-\frac{3\sqrt{5}}{10} + \left(\frac{1}{2} - \frac{\sqrt{5}}{5}\right) b + \frac{1}{2}\right)$$
$$- (15 - 5\sqrt{5}) \left(\frac{3\sqrt{5}}{10} + \left(\frac{1}{2} + \frac{\sqrt{5}}{5}\right) b + \frac{1}{2}\right) \,.$$

Since $b \geq 0$ then the above is at most

$$-(15 + 5\sqrt{5}) \left(-\frac{3\sqrt{5}}{10} + \frac{1}{2}\right) - (15 - 5\sqrt{5}) \left(\frac{3\sqrt{5}}{10} + \frac{1}{2}\right) = 0 \,.$$

∎

A.2.3. PROOF OF LEMMA 18 FOR $\gamma = 1$

Here, we have $\lambda_1 = \frac{27}{2} + \frac{\sqrt{649}}{2}$, $\lambda_2 = 5$, and $\lambda_3 = \frac{27}{2} - \frac{\sqrt{649}}{2}$. The proof of Lemma 18 for $\gamma = 1$ follows from the following lemmas.

Lemma 21 *Let $\mathbf{w}(t)$ be the trajectory from Eq. 9. For $i \in \{1, 2\}$, we have $\mathbf{x}_i^\top \mathbf{w}(t) \geq 0$ for every $t \geq 0$.*

Proof By straightforward calculations, we obtain

$$\frac{1}{\alpha} \cdot \mathbf{x}_1^\top \mathbf{w}(t) = (3, -1, 0)(5, -1, 1)^\top - (3, -1, 0)U \exp(-tD)U^\top(5, -1, 1)^\top$$
$$= 16 - \left(\frac{58}{9} + \frac{1354\sqrt{649}}{5841}\right) e^{-\lambda_1 t} - \frac{28}{9} e^{-\lambda_2 t} - \left(\frac{58}{9} - \frac{1354\sqrt{649}}{5841}\right) e^{-\lambda_3 t}$$
$$\geq 16 - \left(\frac{58}{9} + \frac{1354\sqrt{649}}{5841}\right) - \frac{28}{9} - \left(\frac{58}{9} - \frac{1354\sqrt{649}}{5841}\right) = 0 \,.$$

Moreover, we have

$$\frac{1}{\alpha} \cdot \mathbf{x}_2^\top \mathbf{w}(t) = (4,2,0)(5,-1,1)^\top - (4,2,0)U \exp(-tD)U^\top(5,-1,1)^\top$$

$$= 18 - \left(\frac{89}{9} + \frac{2357\sqrt{649}}{5841}\right)e^{-\lambda_1 t} + \frac{16}{9}e^{-\lambda_2 t} + \left(\frac{2357\sqrt{649}}{5841} - \frac{89}{9}\right)e^{-\lambda_3 t}\ .$$

Since $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$, then the above is at least

$$18 - \left(\frac{89}{9} + \frac{2357\sqrt{649}}{5841}\right)e^{-\lambda_1 t} + \frac{16}{9}e^{-\lambda_1 t} + \left(\frac{2357\sqrt{649}}{5841} - \frac{89}{9}\right)e^{-\lambda_1 t}$$

$$= 18 - 18e^{-\lambda_1 t} \geq 0\ .$$

■

Lemma 22 *Let $\mathbf{w}(t)$ be the trajectory from Eq. 9. There exists $t_1 \in (0.169, 0.17)$ such that $\mathbf{x}_3^\top \mathbf{w}(t_1) = 0$ and for every $t \in [0, t_1]$ we have $\mathbf{x}_3^\top \mathbf{w}(t) \geq 0$.*

Proof We have

$$\frac{1}{\alpha} \cdot \mathbf{x}_3^\top \mathbf{w}(t) = (0,1,1)(5,-1,1)^\top - (0,1,1)U \exp(-tD)U^\top(5,-1,1)^\top$$

$$= -\left(\frac{10}{9} + \frac{10\sqrt{649}}{5841}\right)e^{-\lambda_1 t} + \frac{20}{9}e^{-\lambda_2 t} - \left(\frac{10}{9} - \frac{10\sqrt{649}}{5841}\right)e^{-\lambda_3 t}\ .$$

Let

$$g(t) = -\left(\frac{10}{9} + \frac{10\sqrt{649}}{5841}\right) + \frac{20}{9}e^{(\lambda_1-\lambda_2)t} - \left(\frac{10}{9} - \frac{10\sqrt{649}}{5841}\right)e^{(\lambda_1-\lambda_3)t}$$

Note that $\frac{1}{\alpha} \cdot \mathbf{x}_3^\top \mathbf{w}(t) = e^{-\lambda_1 t}g(t)$. Thus, $\mathbf{x}_3^\top \mathbf{w}(t) = 0$ iff $g(t) = 0$, and $\mathbf{x}_3^\top \mathbf{w}(t) \geq 0$ iff $g(t) \geq 0$. We will show that there exists $t_1 \in (0.169, 0.17)$ such that $g(t_1) = 0$ and for every $t \in [0, t_1]$ we have $g(t) \geq 0$.

Consider the derivative of $g(t)$

$$g'(t) = \frac{20}{9}(\lambda_1 - \lambda_2)e^{(\lambda_1-\lambda_2)t} - \left(\frac{10}{9} - \frac{10\sqrt{649}}{5841}\right)(\lambda_1 - \lambda_3)e^{(\lambda_1-\lambda_3)t}\ .$$

We have $g'(t) = 0$ iff

$$\ln\left(\frac{20}{9}(\lambda_1 - \lambda_2)\right) + (\lambda_1 - \lambda_2)t = \ln\left(\left(\frac{10}{9} - \frac{10\sqrt{649}}{5841}\right)(\lambda_1 - \lambda_3)\right) + (\lambda_1 - \lambda_3)t\ .$$

Note that this equation has a single solution. Since $g(0) = 0$, it follows that g has at most one additional root. It is easy to verify that $g(0.169) > 0$ and $g(0.17) < 0$. Hence, $g(t) = 0$ iff $t \in \{0, t_1\}$ for some $t_1 \in (0.169, 0.17)$, and $g(t) > 0$ for every $t \in (0, t_1)$. ■

Lemma 23

$$(1,0,0)\mathbf{w}(t_1) \in \alpha \cdot (4,5) \, .$$

Proof We have

$$\frac{1}{\alpha} \cdot (1,0,0)\mathbf{w}(t_1) = (1,0,0)(5,-1,1)^\top - (1,0,0)U \exp(-t_1 D)U^\top (5,-1,1)^\top$$

$$= 5 - \left(\frac{1013\sqrt{649}}{11682} + \frac{41}{18} \right) e^{-\lambda_1 t_1} - \frac{4}{9}e^{-\lambda_2 t_1} - \left(\frac{41}{18} - \frac{1013\sqrt{649}}{11682} \right) e^{-\lambda_3 t_1} \, .$$

By Lemma 26 we have $t_1 \in (0.169, 0.17)$, and hence

$$\frac{1}{\alpha} \cdot (1,0,0)\mathbf{w}(t_1)$$

$$\geq 5 - \left(\frac{1013\sqrt{649}}{11682} + \frac{41}{18} \right) e^{-\lambda_1 \cdot 0.169} - \frac{4}{9}e^{-\lambda_2 \cdot 0.169} - \left(\frac{41}{18} - \frac{1013\sqrt{649}}{11682} \right) e^{-\lambda_3 \cdot 0.169} \, .$$

It is easy to verify that the above expression is greater than 4.

We also have

$$\frac{1}{\alpha} \cdot (1,0,0)\mathbf{w}(t_1)$$

$$\leq 5 - \left(\frac{1013\sqrt{649}}{11682} + \frac{41}{18} \right) e^{-\lambda_1 \cdot 1.7} - \frac{4}{9}e^{-\lambda_2 \cdot 1.7} - \left(\frac{41}{18} - \frac{1013\sqrt{649}}{11682} \right) e^{-\lambda_3 \cdot 1.7} < 5 \, .$$

$$\blacksquare$$

Lemma 24

$$(0,1,0)\mathbf{w}(t_1) = -(0,0,1)\mathbf{w}(t_1) \in \alpha \cdot (0.035, 0.045) \, .$$

Proof We have

$$\frac{1}{\alpha} \cdot (0,0,1)\mathbf{w}(t_1) = (0,0,1)(5,-1,1)^\top - (0,0,1)U \exp(-t_1 D)U^\top (5,-1,1)^\top$$

$$= 1 - \left(\frac{13}{18} - \frac{311\sqrt{649}}{11682} \right) e^{-\lambda_1 t_1} + \frac{4}{9}e^{-\lambda_2 t_1} - \left(\frac{13}{18} + \frac{311\sqrt{649}}{11682} \right) e^{-\lambda_3 t_1} \, .$$

By Lemma 26 we have $t_1 \in (0.169, 0.17)$, and hence

$$\frac{1}{\alpha} \cdot (0,0,1)\mathbf{w}(t_1)$$

$$\leq 1 - \left(\frac{13}{18} - \frac{311\sqrt{649}}{11682} \right) e^{-\lambda_1 \cdot 0.17} + \frac{4}{9}e^{-\lambda_2 \cdot 0.169} - \left(\frac{13}{18} + \frac{311\sqrt{649}}{11682} \right) e^{-\lambda_3 \cdot 0.17} \, .$$

It is easy to verify that the above expression is smaller than -0.035.

We also have

$$\frac{1}{\alpha} \cdot (0,0,1)\mathbf{w}(t_1)$$

$$\geq 1 - \left(\frac{13}{18} - \frac{311\sqrt{649}}{11682}\right) e^{-\lambda_1 \cdot 0.169} + \frac{4}{9} e^{-\lambda_2 \cdot 0.17} - \left(\frac{13}{18} + \frac{311\sqrt{649}}{11682}\right) e^{-\lambda_3 \cdot 0.169} .$$

It is easy to verify that the above expression is greater than -0.045.

Finally, by Lemma 22, we have

$$0 = \mathbf{x}_3^\top \mathbf{w}(t_1) = (0,1,1)\mathbf{w}(t_1) = (0,1,0)\mathbf{w}(t_1) + (0,0,1)\mathbf{w}(t_1) ,$$

and hence $(0,1,0)\mathbf{w}(t_1) = -(0,0,1)\mathbf{w}(t_1)$. ∎

A.2.4. PROOF OF LEMMA 18 FOR $\gamma = 2$

Here, we have $\lambda_1 = 14 + 2\sqrt{39}$, $\lambda_2 = 10$, and $\lambda_3 = 14 - 2\sqrt{39}$. The proof of Lemma 18 for $\gamma = 2$ follows from the following lemmas.

Lemma 25 *Let $\mathbf{w}(t)$ be the trajectory from Eq. 9. For $i \in \{1,2\}$, we have $\mathbf{x}_i^\top \mathbf{w}(t) \geq 0$ for every $t \geq 0$.*

Proof By straightforward calculations, we obtain

$$\frac{1}{\alpha} \cdot \mathbf{x}_1^\top \mathbf{w}(t) = (3,-1,0)(5,-1,1)^\top - (3,-1,0)U \exp(-tD)U^\top(5,-1,1)^\top$$

$$= 16 - \left(\frac{47}{7} + \frac{17\sqrt{39}}{21}\right) e^{-\lambda_1 t} - \frac{18}{7} e^{-\lambda_2 t} - \left(\frac{47}{7} - \frac{17\sqrt{39}}{21}\right) e^{-\lambda_3 t}$$

$$\geq 16 - \left(\frac{47}{7} + \frac{17\sqrt{39}}{21}\right) - \frac{18}{7} - \left(\frac{47}{7} - \frac{17\sqrt{39}}{21}\right) = 0 .$$

Moreover, we have

$$\frac{1}{\alpha} \cdot \mathbf{x}_2^\top \mathbf{w}(t) = (4,2,0)(5,-1,1)^\top - (4,2,0)U \exp(-tD)U^\top(5,-1,1)^\top$$

$$= 18 - \left(\frac{463\sqrt{39}}{273} + \frac{66}{7}\right) e^{-\lambda_1 t} + \frac{6}{7} e^{-\lambda_2 t} + \left(\frac{463\sqrt{39}}{273} - \frac{66}{7}\right) e^{-\lambda_3 t} .$$

Since $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$, then the above is at least

$$18 - \left(\frac{463\sqrt{39}}{273} + \frac{66}{7}\right) e^{-\lambda_1 t} + \frac{6}{7} e^{-\lambda_1 t} + \left(\frac{463\sqrt{39}}{273} - \frac{66}{7}\right) e^{-\lambda_1 t}$$

$$= 18 - 18 e^{-\lambda_1 t} \geq 0 .$$

∎

Lemma 26 *Let* $\mathbf{w}(t)$ *be the trajectory from Eq. 9. There exists* $t_1 \in (0.138, 0.139)$ *such that* $\mathbf{x}_3^\top \mathbf{w}(t_1) = 0$*, and for every* $t \in [0, t_1]$ *we have* $\mathbf{x}_3^\top \mathbf{w}(t) \geq 0$.

Proof We have

$$\frac{1}{\alpha} \cdot \mathbf{x}_3^\top \mathbf{w}(t) = (0, 2, 2)(5, -1, 1)^\top - (0, 2, 2)U \exp(-tD)U^\top(5, -1, 1)^\top$$

$$= -\left(\frac{15}{7} + \frac{40\sqrt{39}}{273}\right)e^{-\lambda_1 t} + \frac{30}{7}e^{-\lambda_2 t} - \left(\frac{15}{7} - \frac{40\sqrt{39}}{273}\right)e^{-\lambda_3 t}.$$

Let

$$g(t) = -\left(\frac{15}{7} + \frac{40\sqrt{39}}{273}\right) + \frac{30}{7}e^{(\lambda_1 - \lambda_2)t} - \left(\frac{15}{7} - \frac{40\sqrt{39}}{273}\right)e^{(\lambda_1 - \lambda_3)t}.$$

Note that $\frac{1}{\alpha} \cdot \mathbf{x}_3^\top \mathbf{w}(t) = e^{-\lambda_1 t}g(t)$. Thus, $\mathbf{x}_3^\top \mathbf{w}(t) = 0$ iff $g(t) = 0$, and $\mathbf{x}_3^\top \mathbf{w}(t) \geq 0$ iff $g(t) \geq 0$. We will show that there exists $t_1 \in (0.138, 0.139)$ such that $g(t_1) = 0$ and for every $t \in [0, t_1]$ we have $g(t) \geq 0$.

Consider the derivative of $g(t)$

$$g'(t) = \frac{30}{7}(\lambda_1 - \lambda_2)e^{(\lambda_1 - \lambda_2)t} - \left(\frac{15}{7} - \frac{40\sqrt{39}}{273}\right)(\lambda_1 - \lambda_3)e^{(\lambda_1 - \lambda_3)t}.$$

We have $g'(t) = 0$ iff

$$\ln\left(\frac{30}{7}(\lambda_1 - \lambda_2)\right) + (\lambda_1 - \lambda_2)t = \ln\left(\left(\frac{15}{7} - \frac{40\sqrt{39}}{273}\right)(\lambda_1 - \lambda_3)\right) + (\lambda_1 - \lambda_3)t.$$

Note that this equation has a single solution. Since $g(0) = 0$, it follows that g has at most one additional root. It is easy to verify that $g(0.138) > 0$ and $g(0.139) < 0$. Hence, $g(t) = 0$ iff $t \in \{0, t_1\}$ for some $t_1 \in (0.138, 0.139)$, and $g(t) > 0$ for every $t \in (0, t_1)$. ∎

Lemma 27

$$(1, 0, 0)\mathbf{w}(t_1) \in \alpha \cdot (4, 5).$$

Proof We have

$$\frac{1}{\alpha} \cdot (1, 0, 0)\mathbf{w}(t_1) = (1, 0, 0)(5, -1, 1)^\top - (1, 0, 0)U \exp(-t_1 D)U^\top(5, -1, 1)^\top$$

$$= 5 - \left(\frac{16}{7} + \frac{181\sqrt{39}}{546}\right)e^{-\lambda_1 t_1} - \frac{3}{7}e^{-\lambda_2 t_1} - \left(\frac{16}{7} - \frac{181\sqrt{39}}{546}\right)e^{-\lambda_3 t_1}.$$

By Lemma 26 we have $t_1 \in (0.138, 0.139)$, and hence

$$\frac{1}{\alpha} \cdot (1, 0, 0)\mathbf{w}(t_1)$$

$$\geq 5 - \left(\frac{16}{7} + \frac{181\sqrt{39}}{546}\right)e^{-\lambda_1 \cdot 0.138} - \frac{3}{7}e^{-\lambda_2 \cdot 0.138} - \left(\frac{16}{7} - \frac{181\sqrt{39}}{546}\right)e^{-\lambda_3 \cdot 0.138}.$$

It is easy to verify that the above expression is greater than 4.

We also have

$$\frac{1}{\alpha}\cdot(1,0,0)\mathbf{w}(t_1)$$
$$\leq 5 - \left(\frac{16}{7} + \frac{181\sqrt{39}}{546}\right)e^{-\lambda_1\cdot 0.139} - \frac{3}{7}e^{-\lambda_2\cdot 0.139} - \left(\frac{16}{7} - \frac{181\sqrt{39}}{546}\right)e^{-\lambda_3\cdot 0.139} < 5\,.$$

∎

Lemma 28

$$(0,1,0)\mathbf{w}(t_1) = -(0,0,1)\mathbf{w}(t_1) \in \alpha\cdot(0.1,0.12)\,.$$

Proof We have

$$\frac{1}{\alpha}\cdot(0,0,1)\mathbf{w}(t_1) = (0,0,1)(5,-1,1)^\top - (0,0,1)U\exp(-t_1 D)U^\top(5,-1,1)^\top$$
$$= 1 - \left(\frac{13}{14} - \frac{61\sqrt{39}}{546}\right)e^{-\lambda_1 t_1} + \frac{6}{7}e^{-\lambda_2 t_1} - \left(\frac{13}{14} + \frac{61\sqrt{39}}{546}\right)e^{-\lambda_3 t_1}\,.$$

By Lemma 26 we have $t_1 \in (0.138, 0.139)$, and hence

$$\frac{1}{\alpha}\cdot(0,0,1)\mathbf{w}(t_1)$$
$$\leq 1 - \left(\frac{13}{14} - \frac{61\sqrt{39}}{546}\right)e^{-\lambda_1\cdot 0.139} + \frac{6}{7}e^{-\lambda_2\cdot 0.138} - \left(\frac{13}{14} + \frac{61\sqrt{39}}{546}\right)e^{-\lambda_3\cdot 0.139}\,.$$

It is easy to verify that the above expression is smaller than -0.1.

We also have

$$\frac{1}{\alpha}\cdot(0,0,1)\mathbf{w}(t_1)$$
$$\geq 1 - \left(\frac{13}{14} - \frac{61\sqrt{39}}{546}\right)e^{-\lambda_1\cdot 0.138} + \frac{6}{7}e^{-\lambda_2\cdot 0.139} - \left(\frac{13}{14} + \frac{61\sqrt{39}}{546}\right)e^{-\lambda_3\cdot 0.138}\,.$$

It is easy to verify that the above expression is greater than -0.12.

Finally, by Lemma 26, we have

$$0 = \mathbf{x}_3^\top\mathbf{w}(t_1) = 2\cdot(0,1,1)\mathbf{w}(t_1) = 2\left((0,1,0)\mathbf{w}(t_1) + (0,0,1)\mathbf{w}(t_1)\right)\,,$$

and hence $(0,1,0)\mathbf{w}(t_1) = -(0,0,1)\mathbf{w}(t_1)$. ∎

A.2.5. PROOF OF LEMMA 18 FOR $\gamma = 5$

Here, we have $\lambda_1 = \frac{5\sqrt{105}}{2} + \frac{55}{2}$, $\lambda_2 = 25$, and $\lambda_3 = \frac{55}{2} - \frac{5\sqrt{105}}{2}$. The proof of Lemma 18 for $\gamma = 5$ follows from the following lemmas.

Lemma 29 *Let $\mathbf{w}(t)$ be the trajectory from Eq. 9. For $i \in \{1, 2\}$, we have $\mathbf{x}_i^\top \mathbf{w}(t) \geq 0$ for every $t \geq 0$.*

Proof By straightforward calculations, we obtain

$$\frac{1}{\alpha} \cdot \mathbf{x}_1^\top \mathbf{w}(t) = (3, -1, 0)(5, -1, 1)^\top - (3, -1, 0)U \exp(-tD)U^\top (5, -1, 1)^\top$$

$$= 16 + \left(\frac{34\sqrt{105}}{273} - \frac{14}{13} \right) e^{-\lambda_1 t} - \frac{180}{13} e^{-\lambda_2 t} - \left(\frac{34\sqrt{105}}{273} + \frac{14}{13} \right) e^{-\lambda_3 t}.$$

Since $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$, then the above is at least

$$16 + \left(\frac{34\sqrt{105}}{273} - \frac{14}{13} \right) e^{-\lambda_1 t} - \frac{180}{13} e^{-\lambda_3 t} - \left(\frac{34\sqrt{105}}{273} + \frac{14}{13} \right) e^{-\lambda_3 t}$$

$$= 16 + \left(\frac{34\sqrt{105}}{273} - \frac{14}{13} \right) e^{-\lambda_1 t} - \left(\frac{180}{13} + \frac{34\sqrt{105}}{273} + \frac{14}{13} \right) e^{-\lambda_3 t}.$$

Let $g(t)$ denote the above expression. It is easy to verify that $g(0) = 0$. Thus, it suffices to show that $g'(t) \geq 0$ for every $t \geq 0$. We have

$$g'(t) = -\lambda_1 \left(\frac{34\sqrt{105}}{273} - \frac{14}{13} \right) e^{-\lambda_1 t} + \lambda_3 \left(\frac{180}{13} + \frac{34\sqrt{105}}{273} + \frac{14}{13} \right) e^{-\lambda_3 t}$$

$$\geq -\lambda_1 \left(\frac{34\sqrt{105}}{273} - \frac{14}{13} \right) e^{-\lambda_1 t} + \lambda_3 \left(\frac{180}{13} + \frac{34\sqrt{105}}{273} + \frac{14}{13} \right) e^{-\lambda_1 t}$$

$$= \left[-\lambda_1 \left(\frac{34\sqrt{105}}{273} - \frac{14}{13} \right) + \lambda_3 \left(\frac{180}{13} + \frac{34\sqrt{105}}{273} + \frac{14}{13} \right) \right] e^{-\lambda_1 t}.$$

By plugging in λ_1 and λ_2, it is easy to verify that the above expression is positive.

Next, we have

$$\frac{1}{\alpha} \cdot \mathbf{x}_2^\top \mathbf{w}(t) = (4, 2, 0)(5, -1, 1)^\top - (4, 2, 0)U \exp(-tD)U^\top (5, -1, 1)^\top$$

$$= 18 - \left(\frac{37\sqrt{105}}{273} - \frac{3}{13} \right) e^{-\lambda_1 t} - \frac{240}{13} e^{-\lambda_2 t} + \left(\frac{37\sqrt{105}}{273} + \frac{3}{13} \right) e^{-\lambda_3 t}$$

$$\geq 18 - \left(\frac{37\sqrt{105}}{273} - \frac{3}{13} \right) e^{-\lambda_3 t} - \frac{240}{13} e^{-\lambda_3 t} + \left(\frac{37\sqrt{105}}{273} + \frac{3}{13} \right) e^{-\lambda_3 t}$$

$$= 18 - 18 e^{-\lambda_3 t} \geq 0.$$

∎

Lemma 30 *Let $\mathbf{w}(t)$ be the trajectory from Eq. 9. There exists $t_1 \in (0.086, 0.087)$ such that $\mathbf{x}_3^\top \mathbf{w}(t_1) = 0$, and for every $t \in [0, t_1]$ we have $\mathbf{x}_3^\top \mathbf{w}(t) \geq 0$.*

Proof We have

$$\frac{1}{\alpha} \cdot \mathbf{x}_3^\top \mathbf{w}(t) = (0, 5, 5)(5, -1, 1)^\top - (0, 5, 5)U \exp(-tD)U^\top (5, -1, 1)^\top$$

$$= -\left(\frac{46\sqrt{105}}{273} + \frac{30}{13}\right) e^{-\lambda_1 t} + \frac{60}{13} e^{-\lambda_2 t} - \left(\frac{30}{13} - \frac{46\sqrt{105}}{273}\right) e^{-\lambda_3 t} .$$

Let

$$g(t) = -\left(\frac{46\sqrt{105}}{273} + \frac{30}{13}\right) + \frac{60}{13} e^{(\lambda_1 - \lambda_2)t} - \left(\frac{30}{13} - \frac{46\sqrt{105}}{273}\right) e^{(\lambda_1 - \lambda_3)t} .$$

Note that $\frac{1}{\alpha} \cdot \mathbf{x}_3^\top \mathbf{w}(t) = e^{-\lambda_1 t} g(t)$. Thus, $\mathbf{x}_3^\top \mathbf{w}(t) = 0$ iff $g(t) = 0$, and $\mathbf{x}_3^\top \mathbf{w}(t) \geq 0$ iff $g(t) \geq 0$. We will show that there exists $t_1 \in (0.086, 0.087)$ such that $g(t_1) = 0$ and for every $t \in [0, t_1]$ we have $g(t) \geq 0$.

Consider the derivative of $g(t)$

$$g'(t) = \frac{60}{13}(\lambda_1 - \lambda_2)e^{(\lambda_1 - \lambda_2)t} - \left(\frac{30}{13} - \frac{46\sqrt{105}}{273}\right)(\lambda_1 - \lambda_3)e^{(\lambda_1 - \lambda_3)t} .$$

We have $g'(t) = 0$ iff

$$\ln\left(\frac{60}{13}(\lambda_1 - \lambda_2)\right) + (\lambda_1 - \lambda_2)t = \ln\left(\left(\frac{30}{13} - \frac{46\sqrt{105}}{273}\right)(\lambda_1 - \lambda_3)\right) + (\lambda_1 - \lambda_3)t .$$

Note that this equation has a single solution. Since $g(0) = 0$, it follows that g has at most one additional root. It is easy to verify that $g(0.086) > 0$ and $g(0.087) < 0$. Hence, $g(t) = 0$ iff $t \in \{0, t_1\}$ for some $t_1 \in (0.086, 0.087)$, and $g(t) > 0$ for every $t \in (0, t_1)$. ■

Lemma 31

$$(1, 0, 0)\mathbf{w}(t_1) \in \alpha \cdot (4, 5) .$$

Proof We have

$$\frac{1}{\alpha} \cdot (1, 0, 0)\mathbf{w}(t_1) = (1, 0, 0)(5, -1, 1)^\top - (1, 0, 0)U \exp(-t_1 D)U^\top (5, -1, 1)^\top$$

$$= 5 - \left(\frac{5}{26} - \frac{31\sqrt{105}}{2730}\right) e^{-\lambda_1 t_1} - \frac{60}{13} e^{-\lambda_2 t_1} - \left(\frac{5}{26} + \frac{31\sqrt{105}}{2730}\right) e^{-\lambda_3 t_1} .$$

By Lemma 30 we have $t_1 \in (0.086, 0.087)$, and hence

$$\frac{1}{\alpha} \cdot (1, 0, 0)\mathbf{w}(t_1)$$

$$\geq 5 - \left(\frac{5}{26} - \frac{31\sqrt{105}}{2730}\right) e^{-\lambda_1 \cdot 0.086} - \frac{60}{13} e^{-\lambda_2 \cdot 0.086} - \left(\frac{5}{26} + \frac{31\sqrt{105}}{2730}\right) e^{-\lambda_3 \cdot 0.086} .$$

It is easy to verify that the above expression is greater than 4.

We also have

$$\frac{1}{\alpha} \cdot (1, 0, 0)\mathbf{w}(t_1)$$
$$\geq 5 - \left(\frac{5}{26} - \frac{31\sqrt{105}}{2730}\right) e^{-\lambda_1 \cdot 0.087} - \frac{60}{13} e^{-\lambda_2 \cdot 0.087} - \left(\frac{5}{26} + \frac{31\sqrt{105}}{2730}\right) e^{-\lambda_3 \cdot 0.087} < 5 .$$

∎

Lemma 32

$$(0, 1, 0)\mathbf{w}(t_1) = -(0, 0, 1)\mathbf{w}(t_1) \in \alpha \cdot (0.2, 0.22) .$$

Proof We have

$$\frac{1}{\alpha} \cdot (0, 0, 1)\mathbf{w}(t_1) = (0, 0, 1)(5, -1, 1)^\top - (0, 0, 1)U \exp(-t_1 D)U^\top (5, -1, 1)^\top$$
$$= 1 - \left(\frac{25}{26} - \frac{31\sqrt{105}}{546}\right) e^{-\lambda_1 t_1} + \frac{12}{13} e^{-\lambda_2 t_1} - \left(\frac{25}{26} + \frac{31\sqrt{105}}{546}\right) e^{-\lambda_3 t_1} .$$

By Lemma 30 we have $t_1 \in (0.086, 0.087)$, and hence

$$\frac{1}{\alpha} \cdot (0, 0, 1)\mathbf{w}(t_1)$$
$$\leq 1 - \left(\frac{25}{26} - \frac{31\sqrt{105}}{546}\right) e^{-\lambda_1 \cdot 0.087} + \frac{12}{13} e^{-\lambda_2 \cdot 0.086} - \left(\frac{25}{26} + \frac{31\sqrt{105}}{546}\right) e^{-\lambda_3 \cdot 0.087} .$$

It is easy to verify that the above expression is smaller than -0.2.

We also have

$$\frac{1}{\alpha} \cdot (0, 0, 1)\mathbf{w}(t_1)$$
$$\geq 1 - \left(\frac{25}{26} - \frac{31\sqrt{105}}{546}\right) e^{-\lambda_1 \cdot 0.086} + \frac{12}{13} e^{-\lambda_2 \cdot 0.087} - \left(\frac{25}{26} + \frac{31\sqrt{105}}{546}\right) e^{-\lambda_3 \cdot 0.086} .$$

It is easy to verify that the above expression is greater than -0.22.

Finally, by Lemma 30, we have

$$0 = \mathbf{x}_3^\top \mathbf{w}(t_1) = 5 \cdot (0, 1, 1)\mathbf{w}(t_1) = 5\left((0, 1, 0)\mathbf{w}(t_1) + (0, 0, 1)\mathbf{w}(t_1)\right) ,$$

and hence $(0, 1, 0)\mathbf{w}(t_1) = -(0, 0, 1)\mathbf{w}(t_1)$.

∎

Appendix B. Proof of Lemma 6

Lemma 33 *The function $\mathcal{R}$ is radial. That is, there exists a function $f : \mathbb{R}_+ \to \mathbb{R}$ such that $\mathcal{R}(\mathbf{w}) = f(\|\mathbf{w}\|)$ for every $\mathbf{w} \in \mathbb{R}^d$.*

Proof Let $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ and $\alpha > 0$ such that $\|\mathbf{w}\| = \|\mathbf{w}'\| = \alpha q$. We denote $r = \|\mathbf{w}\|$. Let $c' = \frac{c}{q}$, and let $\epsilon = \|\mathbf{w} - \mathbf{w}'\|$. Assume w.l.o.g. that $\epsilon \leq \frac{2rc'}{\sqrt{1+(c')^2}}$. We will show that $\mathcal{R}(\mathbf{w}) = \mathcal{R}(\mathbf{w}')$.

Let $\mathbf{u} = \frac{\mathbf{w}+\mathbf{w}'}{2}\left(1 + \frac{\epsilon^2}{4r^2-\epsilon^2}\right)$, and let $\mathbf{e} \in \mathbb{R}^d$ be a vector orthogonal to $\mathbf{w}, \mathbf{w}'$, such that $\|\mathbf{e}\|^2 = r^2\left((c')^2 - \frac{\epsilon^2}{4r^2-\epsilon^2}\right)$. Note that

$$4r^2 - \epsilon^2 \geq 4r^2 - \frac{4r^2(c')^2}{1+(c')^2} = \frac{4r^2}{1+(c')^2} \geq \frac{\epsilon^2}{(c')^2} .$$

Thus, $(c')^2 \geq \frac{\epsilon^2}{4r^2-\epsilon^2}$. Hence $r^2\left((c')^2 - \frac{\epsilon^2}{4r^2-\epsilon^2}\right) \geq 0$, and therefore $\mathbf{e}$ is well-defined.

Let $\mathbf{u}^* = \mathbf{u} + \mathbf{e}$. We will show that there is a vector $\mathbf{v}$ such that $\mathbf{u}^* = \mathbf{w} + \mathbf{v}$, $\|\mathbf{v}\| = \alpha c$, and $\langle \mathbf{w}, \mathbf{v} \rangle = 0$. It implies that $\mathcal{R}(\mathbf{u}^*) = \mathcal{R}(\mathbf{w})$. Likewise, we will show that there is a vector $\mathbf{v}'$ such that $\mathbf{u}^* = \mathbf{w}' + \mathbf{v}'$, $\|\mathbf{v}'\| = \alpha c$ and $\langle \mathbf{w}', \mathbf{v}' \rangle = 0$. It implies that $\mathcal{R}(\mathbf{u}^*) = \mathcal{R}(\mathbf{w}')$, and hence $\mathcal{R}(\mathbf{w}) = \mathcal{R}(\mathbf{w}')$, as required.

We first show that $\|\mathbf{u}^* - \mathbf{w}\| = \alpha c$. We have

$$\|\mathbf{u}^* - \mathbf{w}\|^2 = \|\mathbf{u} + \mathbf{e} - \mathbf{w}\|^2 .$$

Since $\mathbf{e}$ is orthogonal to $\mathbf{w}, \mathbf{w}'$, it is also orthogonal to $\mathbf{u}$, and thus the above equals

$$\|\mathbf{e}\|^2 + \|\mathbf{u} - \mathbf{w}\|^2 = r^2\left((c')^2 - \frac{\epsilon^2}{4r^2-\epsilon^2}\right) + \|\mathbf{u} - \mathbf{w}\|^2 . \tag{16}$$

Moreover,

$$\|\mathbf{u} - \mathbf{w}\|^2 = \left\|\frac{\mathbf{w}+\mathbf{w}'}{2}\left(1 + \frac{\epsilon^2}{4r^2-\epsilon^2}\right) - \mathbf{w}\right\|^2 = \left\|\frac{\mathbf{w}'-\mathbf{w}}{2} + \frac{\mathbf{w}'+\mathbf{w}}{2}\cdot\frac{\epsilon^2}{4r^2-\epsilon^2}\right\|^2 .$$

Since $\left\langle \frac{\mathbf{w}'-\mathbf{w}}{2}, \frac{\mathbf{w}'+\mathbf{w}}{2} \right\rangle = 0$, the above equals

$$\left\|\frac{\mathbf{w}'-\mathbf{w}}{2}\right\|^2 + \left\|\frac{\mathbf{w}'+\mathbf{w}}{2}\cdot\frac{\epsilon^2}{4r^2-\epsilon^2}\right\|^2 = \frac{\epsilon^2}{4} + \left(\frac{\epsilon^2}{4r^2-\epsilon^2}\right)^2\left\|\frac{\mathbf{w}'+\mathbf{w}}{2}\right\|^2 . \tag{17}$$

Note that

$$\epsilon^2 = \|\mathbf{w} - \mathbf{w}'\|^2 = 2r^2 - 2\cdot\langle \mathbf{w}, \mathbf{w}' \rangle , \tag{18}$$

and hence

$$\|\mathbf{w} + \mathbf{w}'\|^2 = 2r^2 + 2\cdot\langle \mathbf{w}, \mathbf{w}' \rangle = 4r^2 - \epsilon^2 .$$

Combining the above with Eq. 17 we have

$$\|\mathbf{u} - \mathbf{w}\|^2 = \frac{\epsilon^2}{4} + \left(\frac{\epsilon^2}{4r^2-\epsilon^2}\right)^2\cdot\frac{1}{4}\cdot(4r^2 - \epsilon^2) = \frac{1}{4}\left(\epsilon^2 + \frac{\epsilon^4}{4r^2-\epsilon^2}\right)$$

$$= \frac{r^2\epsilon^2}{4r^2-\epsilon^2} .$$

Plugging the above into Eq. 16, we have

$$\|\mathbf{u}^* - \mathbf{w}\|^2 = r^2 \left((c')^2 - \frac{\epsilon^2}{4r^2 - \epsilon^2}\right) + \frac{r^2\epsilon^2}{4r^2 - \epsilon^2} = (rc')^2 = \left(\frac{rc}{q}\right)^2 = (\alpha c)^2 \ .$$

We now show that $\langle \mathbf{w}, \mathbf{u}^* - \mathbf{w}\rangle = 0$. We have

$$\langle \mathbf{w}, \mathbf{u}^* - \mathbf{w}\rangle = \langle \mathbf{w}, \mathbf{u} + \mathbf{e} - \mathbf{w}\rangle \ .$$

Since $\mathbf{e}$ is orthogonal to $\mathbf{w}$, the above equals

$$\langle \mathbf{w}, \mathbf{u} - \mathbf{w}\rangle = \langle \mathbf{w}, \mathbf{u}\rangle - r^2$$

$$= \frac{\langle \mathbf{w}, \mathbf{w} + \mathbf{w}'\rangle}{2}\left(1 + \frac{\epsilon^2}{4r^2 - \epsilon^2}\right) - r^2$$

$$= \left(r^2 + \langle \mathbf{w}, \mathbf{w}'\rangle\right) \cdot \frac{1}{2}\left(1 + \frac{\epsilon^2}{4r^2 - \epsilon^2}\right) - r^2 \ .$$

By Eq. 18, the above equals

$$\left(2r^2 - \frac{\epsilon^2}{2}\right) \cdot \frac{1}{2}\left(1 + \frac{\epsilon^2}{4r^2 - \epsilon^2}\right) - r^2 = \frac{4r^2 - \epsilon^2}{4} \cdot \frac{4r^2}{4r^2 - \epsilon^2} - r^2 = 0 \ .$$

Finally, the proof that $\|\mathbf{u}^* - \mathbf{w}'\| = \alpha c$ and that $\langle \mathbf{w}', \mathbf{u}^* - \mathbf{w}'\rangle = 0$ is similar. ∎

By Lemma 33, we have $\mathcal{R}(\mathbf{w}) = f(\|\mathbf{w}\|)$ for every $\mathbf{w} \in \mathbb{R}^d$, for some $f : \mathbb{R}_+ \to \mathbb{R}$. We will show that for every $0 < r < r'$ we have $f(r) = f(r')$. Let $\mathbf{w} \in \mathbb{R}^d$ such that $\|\mathbf{w}\| = r$, and let $\alpha = \frac{r}{q}$. By our assumption on $\mathcal{R}$, for a vector $\mathbf{v} \in \mathbb{R}^d$ such that $\|\mathbf{v}\| = \alpha c$ and $\langle \mathbf{w}, \mathbf{v}\rangle = 0$, we have $\mathcal{R}(\mathbf{w}) = \mathcal{R}(\mathbf{w} + \mathbf{v}) = \min_{\beta \geq 0} \mathcal{R}(\mathbf{w} + \beta\mathbf{v})$.

Since $\mathcal{R}(\mathbf{w}) = \mathcal{R}(\mathbf{w} + \mathbf{v})$ then we have $f(r) = f(\|\mathbf{w}\|) = f(\|\mathbf{w} + \mathbf{v}\|)$. Note that

$$\|\mathbf{w} + \mathbf{v}\| = \sqrt{\|\mathbf{w}\|^2 + \|\mathbf{v}\|^2} = \sqrt{r^2 + (\alpha c)^2} = r\sqrt{1 + \frac{c^2}{q^2}} \ .$$

Hence, we have $f(r) = f\left(r\sqrt{1 + \frac{c^2}{q^2}}\right)$. By repeating the above steps, we have for every $m \in \mathbb{N}$ that $f(r) = f\left(r\left(\sqrt{1 + \frac{c^2}{q^2}}\right)^m\right)$. Let m' be a sufficiently large integer, such that $r\left(\sqrt{1 + \frac{c^2}{q^2}}\right)^{m'} > r'$.

Since $\mathcal{R}(\mathbf{w}) = \min_{\beta \geq 0} \mathcal{R}(\mathbf{w} + \beta\mathbf{v})$ then we have $f(r) = \min_{r'' \geq r} f(r'') \leq f(r')$. By repeating this argument with $\mathbf{w}' \in \mathbb{R}^d$ such that $\|\mathbf{w}'\| = r'$ we also have $f(r') \leq f\left(r\left(\sqrt{1 + \frac{c^2}{q^2}}\right)^{m'}\right)$.

Overall, we have

$$f(r) \leq f(r') \leq f\left(r\left(\sqrt{1 + \frac{c^2}{q^2}}\right)^{m'}\right) = f(r) \ ,$$

and thus $f(r) = f(r')$ as required.

Appendix C. Proof of Lemma 10

We have

$$
\begin{aligned}
\frac{d}{dt}\left(\|\mathbf{w}(t)\|^2 - (v(t))^2\right) &= 2\mathbf{w}(t)^\top\dot{\mathbf{w}}(t) - 2v(t)\dot{v}(t)\\
&= -2\mathbf{w}(t)^\top\frac{\partial\mathcal{L}_{X,\mathbf{y}}(\boldsymbol{\theta}(t))}{\partial\mathbf{w}} + 2v(t)\frac{\partial\mathcal{L}_{X,\mathbf{y}}(\boldsymbol{\theta}(t))}{\partial v}\\
&= -2\mathbf{w}(t)^\top\sum_{i=1}^{n}\left(v(t)\sigma(\mathbf{x}_i^\top\mathbf{w}(t)) - y_i\right)v(t)\sigma'(\mathbf{x}_i^\top\mathbf{w}(t))\mathbf{x}_i\\
&\quad + 2v(t)\sum_{i=1}^{n}\left(v(t)\sigma(\mathbf{x}_i^\top\mathbf{w}(t)) - y_i\right)\sigma(\mathbf{x}_i^\top\mathbf{w}(t))\\
&= -2\sum_{i=1}^{n}\left(v(t)\sigma(\mathbf{x}_i^\top\mathbf{w}(t)) - y_i\right)v(t)\sigma'(\mathbf{x}_i^\top\mathbf{w}(t))\mathbf{w}(t)^\top\mathbf{x}_i\\
&\quad + 2\sum_{i=1}^{n}\left(v(t)\sigma(\mathbf{x}_i^\top\mathbf{w}(t)) - y_i\right)v(t)\sigma(\mathbf{x}_i^\top\mathbf{w}(t))\,.
\end{aligned}
$$

Since the ReLU function σ satisfies $\sigma'(\mathbf{x}_i^\top\mathbf{w}(t))(\mathbf{x}_i^\top\mathbf{w}(t)) = \sigma(\mathbf{x}_i^\top\mathbf{w}(t))$, then the above equals 0.

Appendix D. Proof of Theorem 14

We start with some additional notations. For an input $(X, \mathbf{y}) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$ to gradient flow, we denote $\mathcal{U}_{X,\mathbf{y}} = \{\mathbf{u} \in \mathbb{R}^d : \mathcal{L}_{X,\mathbf{y}}(\Psi(\mathbf{u})) = 0\}$, namely, the vectors in $\mathbb{R}^d$ that correspond to global minima of $\mathcal{L}_{X,\mathbf{y}}$. Let $\mathbf{u} \in \mathbb{R}^d$ and let $(\mathbf{w}, v) = \Psi(\mathbf{u})$. Note that $\mathbf{u} \in \mathcal{U}_{X,\mathbf{y}}$ iff $\mathbf{y} = v \cdot \sigma(X\mathbf{w})$ iff $\mathbf{y} = \sigma(vX\mathbf{w}) = \sigma(X\mathbf{u})$. Therefore, we have $\mathcal{U}_{X,\mathbf{y}} = \{\mathbf{u} \in \mathbb{R}^d : \sigma(X\mathbf{u}) = \mathbf{y}\}$. Let $\mathbf{u}^* = v^*\mathbf{w}^* = \Psi^{-1}(\boldsymbol{\theta}^*)$, namely, the vector that corresponds to the limit point. Since $\boldsymbol{\theta}^* \in \operatorname{argmin}_{\{\boldsymbol{\theta}\in\mathcal{S}:\mathcal{L}_{X,\mathbf{y}}(\boldsymbol{\theta})=0\}}\mathcal{R}(\boldsymbol{\theta})$, then $\mathbf{u}^* \in \operatorname{argmin}_{\mathbf{u}\in\mathcal{U}_{X,\mathbf{y}}}\mathcal{R}'(\mathbf{u})$. Indeed, we have

$$
\begin{aligned}
\mathcal{R}'(\mathbf{u}^*) = \mathcal{R}(\Psi(\mathbf{u}^*)) = \mathcal{R}(\boldsymbol{\theta}^*) &= \min_{\{\boldsymbol{\theta}\in\mathcal{S}:\mathcal{L}_{X,\mathbf{y}}(\boldsymbol{\theta})=0\}}\mathcal{R}(\boldsymbol{\theta})\\
&= \min_{\{\mathbf{u}\in\mathbb{R}^d:\mathcal{L}_{X,\mathbf{y}}(\Psi(\mathbf{u}))=0\}}\mathcal{R}(\Psi(\mathbf{u})) = \min_{\mathbf{u}\in\mathcal{U}_{X,\mathbf{y}}}\mathcal{R}'(\mathbf{u})\,,
\end{aligned}
$$

and $\mathbf{u}^* \in \mathcal{U}_{X,\mathbf{y}}$ since $\mathcal{L}_{X,\mathbf{y}}(\Psi(\mathbf{u}^*)) = \mathcal{L}_{X,\mathbf{y}}(\boldsymbol{\theta}^*) = 0$.

Note that by Corollary 13, for the inputs $(X_0, \mathbf{y})$ and $(X_5, \mathbf{y})$, the limit points of gradient flow are $\boldsymbol{\theta}^* = (\mathbf{w}^*, v^*)$ and $\tilde{\boldsymbol{\theta}}^* = (\tilde{\mathbf{w}}^*, \tilde{v}^*)$ (respectively), such that $\mathbf{u}^* = \Psi^{-1}(\boldsymbol{\theta}^*) = v^*\mathbf{w}^* = (5, -1, 0)^\top$ and $\tilde{\mathbf{u}}^* = \Psi^{-1}(\tilde{\boldsymbol{\theta}}^*) = \tilde{v}^*\tilde{\mathbf{w}}^* = (5, -1, s)^\top$, and therefore we have the following. Let $\mathbf{u}' = \tilde{\mathbf{u}}^* - \mathbf{u}^* = (0, 0, s)^\top$. We have $\mathbf{u}^*, \tilde{\mathbf{u}}^*, \mathbf{u}' \neq \mathbf{0}$, and $\langle\mathbf{u}^*, \mathbf{u}'\rangle = 0$. Moreover, since for every $\mathbf{u} \in \mathcal{Q} = \{(5, -1, q)^\top : q \leq 0\}$ we have $\sigma(X_0\mathbf{u}) = \sigma(X_5\mathbf{u}) = \mathbf{y}$, then $\mathcal{Q} \subseteq \mathcal{U}_{X_0,\mathbf{y}} \cap \mathcal{U}_{X_5,\mathbf{y}}$, and therefore for every $\beta \geq 0$ we have $\mathbf{u}^* + \beta\mathbf{u}' \in \mathcal{U}_{X_0,\mathbf{y}} \cap \mathcal{U}_{X_5,\mathbf{y}}$. Since $\mathbf{u}^* \in \operatorname{argmin}_{\mathbf{u}\in\mathcal{U}_{X_0,\mathbf{y}}}\mathcal{R}'(\mathbf{u})$ and $\tilde{\mathbf{u}}^* \in \operatorname{argmin}_{\mathbf{u}\in\mathcal{U}_{X_5,\mathbf{y}}}\mathcal{R}'(\mathbf{u})$, then we have $\mathcal{R}'(\mathbf{u}^*) = \mathcal{R}'(\tilde{\mathbf{u}}^*) = \min_{\beta\geq 0}\mathcal{R}'(\mathbf{u}^* + \beta\mathbf{u}')$.

We now utilize the following lemma, which considers the effect on the limit point of rotating and scaling the input to gradient flow. The lemma follows by analyzing the trajectories of gradient flow with the modified inputs. See Section D.1 for a proof.

Lemma 34 *Let $(X, \mathbf{y}) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ be an input to gradient flow, let $\boldsymbol{\theta}^* = (\mathbf{w}^*, v^*)$ be its limit point, and let $\mathbf{u}^* = v^* \mathbf{w}^* = \Psi^{-1}(\boldsymbol{\theta}^*)$.*

1. *Let $M \in SO(d)$ be a rotation matrix. The limit point of gradient flow with input $(XM^\top, \mathbf{y})$ is $\boldsymbol{\theta}^*_M = (M\mathbf{w}^*, v^*)$. Thus, $\Psi^{-1}(\boldsymbol{\theta}^*_M) = M\mathbf{u}^*$.*

2. *Let $\alpha > 0$. The limit point of gradient flow with input $(X, \alpha\mathbf{y})$ is $\boldsymbol{\theta}^*_\alpha = \sqrt{\alpha} \cdot (\mathbf{w}^*, v^*)$. Thus, $\Psi^{-1}(\boldsymbol{\theta}^*_\alpha) = \alpha \mathbf{u}^*$.*

By rotating and scaling the inputs from Corollary 13, and using Lemma 34, we obtain the following corollary (see Section D.2 for a formal proof).

Corollary 35 *Under Assumption 12, there is a vector $\mathbf{u}^* \in \mathbb{R}^3 \setminus \{\mathbf{0}\}$ and a constant $c > 0$, such that for every $\alpha > 0$ and $\mathbf{u} \in \mathbb{R}^3$ with $\|\mathbf{u}\| = \alpha \cdot \|\mathbf{u}^*\|$, and for every $\mathbf{u}_\perp \in \mathbb{R}^3$ such that $\langle \mathbf{u}_\perp, \mathbf{u} \rangle = 0$ and $\|\mathbf{u}_\perp\| = \alpha c$, we have $\mathcal{R}'(\mathbf{u}) = \mathcal{R}'(\mathbf{u} + \mathbf{u}_\perp) = \min_{\beta \geq 0} \mathcal{R}'(\mathbf{u} + \beta \mathbf{u}_\perp)$.*

Finally, Theorem 14 follows immediately from Corollary 35 and Lemma 6.

D.1. Proof of Lemma 34

For the input $(X, \mathbf{y})$, gradient flow obeys the following dynamics.

$$\dot{\mathbf{w}}(t) = -\frac{\partial \mathcal{L}_{X,\mathbf{y}}((\mathbf{w}(t), v(t)))}{\partial \mathbf{w}} = -\sum_{i=1}^{n} \left(v(t)\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) - y_i \right) v(t)\sigma'(\mathbf{x}_i^\top \mathbf{w}(t))\mathbf{x}_i \,, \quad (19)$$

and

$$\dot{v}(t) = -\frac{\partial \mathcal{L}_{X,\mathbf{y}}((\mathbf{w}(t), v(t)))}{\partial v} = -\sum_{i=1}^{n} \left(v(t)\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) - y_i \right) \sigma(\mathbf{x}_i^\top \mathbf{w}(t)) \,. \quad (20)$$

D.1.1. PROOF OF PART (1)

Let $\epsilon > 0$ be a constant, and let $\boldsymbol{\theta}^\epsilon(t) = (\mathbf{w}(t), v(t))$ be the trajectory of gradient flow with the input $(X, \mathbf{y})$ starting from $\boldsymbol{\theta}^\epsilon(0) = (\mathbf{0}, \epsilon)$. Let $\tilde{\mathbf{w}}(t) = M\mathbf{w}(t)$ and $\tilde{v}(t) = v(t)$. By Eq. 19 we have

$$\dot{\tilde{\mathbf{w}}}(t) = M\dot{\mathbf{w}}(t) = -M\sum_{i=1}^{n} \left(v(t)\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) - y_i \right) v(t)\sigma'(\mathbf{x}_i^\top \mathbf{w}(t))\mathbf{x}_i$$

$$= -M\sum_{i=1}^{n} \left(v(t)\sigma((M\mathbf{x}_i)^\top (M\mathbf{w}(t))) - y_i \right) v(t)\sigma'((M\mathbf{x}_i)^\top (M\mathbf{w}(t)))\mathbf{x}_i$$

$$= -\sum_{i=1}^{n} \left(\tilde{v}(t)\sigma((M\mathbf{x}_i)^\top \tilde{\mathbf{w}}(t)) - y_i \right) \tilde{v}(t)\sigma'((M\mathbf{x}_i)^\top \tilde{\mathbf{w}}(t))(M\mathbf{x}_i)$$

$$= -\frac{\partial \mathcal{L}_{XM^\top,\mathbf{y}}((\tilde{\mathbf{w}}(t), \tilde{v}(t)))}{\partial \mathbf{w}} \,.$$

Moreover, by Eq. 20 we have

$$\dot{\tilde{v}}(t) = \dot{v}(t) = -\sum_{i=1}^{n} \left(v(t)\sigma(\mathbf{x}_i^\top \mathbf{w}(t)) - y_i \right) \sigma(\mathbf{x}_i^\top \mathbf{w}(t))$$

$$= -\sum_{i=1}^{n} \left(\tilde{v}(t)\sigma((M\mathbf{x}_i)^\top \tilde{\mathbf{w}}(t)) - y_i \right) \sigma((M\mathbf{x}_i)^\top \tilde{\mathbf{w}}(t))$$

$$= -\frac{\partial \mathcal{L}_{XM^\top, \mathbf{y}}((\tilde{\mathbf{w}}(t), \tilde{v}(t)))}{\partial v} .$$

Hence, $\tilde{\boldsymbol{\theta}}^\epsilon(t) = (\tilde{\mathbf{w}}(t), \tilde{v}(t))$ is the trajectory of gradient flow with the input $(XM^\top, \mathbf{y})$ staring from $\tilde{\boldsymbol{\theta}}^\epsilon(0) = (\tilde{\mathbf{w}}(0), \tilde{v}(0)) = (\mathbf{0}, \epsilon)$. Therefore,

$$\lim_{\epsilon \to 0^+} \tilde{\boldsymbol{\theta}}^\epsilon(\infty) = \lim_{\epsilon \to 0^+} (M\mathbf{w}(\infty), v(\infty)) = (M\mathbf{w}^*, v^*) .$$

D.1.2. PROOF OF PART (2)

Let $\epsilon > 0$ be a constant, and let $\boldsymbol{\theta}^{\epsilon/\sqrt{\alpha}}(t) = (\mathbf{w}(t), v(t))$ be the trajectory of gradient flow with the input $(X, \mathbf{y})$ starting from $\boldsymbol{\theta}^{\epsilon/\sqrt{\alpha}}(0) = \left(\mathbf{0}, \frac{\epsilon}{\sqrt{\alpha}} \right)$. Let $\tilde{\mathbf{w}}(t) = \sqrt{\alpha}\mathbf{w}(\alpha t)$ and $\tilde{v}(t) = \sqrt{\alpha}v(\alpha t)$. By Eq. 19 we have

$$\dot{\tilde{\mathbf{w}}}(t) = \sqrt{\alpha} \cdot \alpha \cdot \dot{\mathbf{w}}(\alpha t)$$

$$= -\alpha\sqrt{\alpha} \sum_{i=1}^{n} \left(v(\alpha t)\sigma(\mathbf{x}_i^\top \mathbf{w}(\alpha t)) - y_i \right) v(\alpha t)\sigma'(\mathbf{x}_i^\top \mathbf{w}(\alpha t))\mathbf{x}_i$$

$$= -\sum_{i=1}^{n} \left(\sqrt{\alpha}v(\alpha t)\sigma(\mathbf{x}_i^\top \sqrt{\alpha}\mathbf{w}(\alpha t)) - (\alpha y_i) \right) (\sqrt{\alpha}v(\alpha t))\sigma'(\mathbf{x}_i^\top \mathbf{w}(\alpha t))\mathbf{x}_i .$$

Since $\sigma'(\mathbf{x}_i^\top \mathbf{w}(\alpha t)) = \sigma'(\mathbf{x}_i^\top \sqrt{\alpha}\mathbf{w}(\alpha t))$, the above equals

$$-\sum_{i=1}^{n} \left(\tilde{v}(t)\sigma(\mathbf{x}_i^\top \tilde{\mathbf{w}}(t)) - (\alpha y_i) \right) \tilde{v}(t)\sigma'(\mathbf{x}_i^\top \tilde{\mathbf{w}}(t))\mathbf{x}_i = -\frac{\partial \mathcal{L}_{X, \alpha\mathbf{y}}((\tilde{\mathbf{w}}(t), \tilde{v}(t)))}{\partial \mathbf{w}} .$$

Moreover, by Eq. 20 we have

$$\dot{\tilde{v}}(t) = \sqrt{\alpha} \cdot \alpha \cdot \dot{v}(\alpha t)$$

$$= -\alpha\sqrt{\alpha} \sum_{i=1}^{n} \left(v(\alpha t)\sigma(\mathbf{x}_i^\top \mathbf{w}(\alpha t)) - y_i \right) \sigma(\mathbf{x}_i^\top \mathbf{w}(\alpha t))$$

$$= -\sum_{i=1}^{n} \left(\sqrt{\alpha}v(\alpha t)\sigma(\mathbf{x}_i^\top \sqrt{\alpha}\mathbf{w}(\alpha t)) - (\alpha y_i) \right) \sigma(\mathbf{x}_i^\top \sqrt{\alpha}\mathbf{w}(\alpha t))$$

$$= -\sum_{i=1}^{n} \left(\tilde{v}(t)\sigma(\mathbf{x}_i^\top \tilde{\mathbf{w}}(t)) - (\alpha y_i) \right) \sigma(\mathbf{x}_i^\top \tilde{\mathbf{w}}(t))$$

$$= -\frac{\partial \mathcal{L}_{X, \alpha\mathbf{y}}((\tilde{\mathbf{w}}(t), \tilde{v}(t)))}{\partial v} .$$

Hence, $\tilde{\boldsymbol{\theta}}^{\epsilon}(t) = (\tilde{\mathbf{w}}(t), \tilde{v}(t))$ is the trajectory of gradient flow with the input $(X, \alpha \mathbf{y})$ staring from $\tilde{\boldsymbol{\theta}}^{\epsilon}(0) = (\tilde{\mathbf{w}}(0), \tilde{v}(0)) = \left(\mathbf{0}, \sqrt{\alpha} \cdot \frac{\epsilon}{\sqrt{\alpha}}\right) = (\mathbf{0}, \epsilon)$. Therefore,

$$\lim_{\epsilon \to 0^+} \tilde{\boldsymbol{\theta}}^{\epsilon}(\infty) = \lim_{\epsilon \to 0^+} (\sqrt{\alpha} \mathbf{w}(\infty), \sqrt{\alpha} v(\infty)) = \sqrt{\alpha} \cdot (\mathbf{w}^*, v^*) .$$

D.2. Proof of Corollary 35

Let $X_0, X_5 \in \mathbb{R}^{3 \times 3}$ and $\mathbf{y} \in \mathbb{R}^3$ from Corollary 13, such that for the corresponding limit points $\boldsymbol{\theta}^*, \tilde{\boldsymbol{\theta}}^*$, and vectors $\mathbf{u}^* = \Psi^{-1}(\boldsymbol{\theta}^*)$, $\tilde{\mathbf{u}}^* = \Psi^{-1}(\tilde{\boldsymbol{\theta}}^*)$, and $\mathbf{u}' = \tilde{\mathbf{u}}^* - \mathbf{u}^*$, we have $\|\mathbf{u}'\| > 0$ and $\langle \mathbf{u}^*, \mathbf{u}' \rangle = 0$. Furthermore, for every $\beta \geq 0$ we have $\mathbf{u}^* + \beta \mathbf{u}' \in \mathcal{U}_{X_0,\mathbf{y}} \cap \mathcal{U}_{X_5,\mathbf{y}}$. We denote $c = \|\mathbf{u}'\|$. Let $\mathbf{u} \in \mathbb{R}^3$ and $\alpha > 0$ such that $\|\mathbf{u}\| = \alpha \cdot \|\mathbf{u}^*\|$, and let $\mathbf{u}_\perp \in \mathbb{R}^3$ such that $\langle \mathbf{u}, \mathbf{u}_\perp \rangle = 0$ and $\|\mathbf{u}_\perp\| = \alpha c$.

Let $M \in SO(3)$ be a rotaion matrix such that $\alpha M \mathbf{u}^* = \mathbf{u}$ and $\alpha M \mathbf{u}' = \mathbf{u}_\perp$. Let $\boldsymbol{\theta}_{\alpha M}^*$ and $\tilde{\boldsymbol{\theta}}_{\alpha M}^*$ be the limit points of gradient flow with inputs $(X_0 M^\top, \alpha \mathbf{y})$ and $(X_5 M^\top, \alpha \mathbf{y})$, respectively. By Lemma 34, we have $\Psi^{-1}(\boldsymbol{\theta}_{\alpha M}^*) = \alpha M \mathbf{u}^* = \mathbf{u}$, and

$$\Psi^{-1}(\tilde{\boldsymbol{\theta}}_{\alpha M}^*) = \alpha M \tilde{\mathbf{u}}^* = \alpha M (\mathbf{u}^* + \mathbf{u}') = \alpha M \mathbf{u}^* + \alpha M \mathbf{u}' = \mathbf{u} + \mathbf{u}_\perp .$$

Let $\{\mathbf{x}_i^\top\}_{i=1}^3$ and $\{\tilde{\mathbf{x}}_i^\top\}_{i=1}^3$ be the rows of X_0 and X_5 (respectively). Since for every $\beta \geq 0$ we have $\mathbf{u}^* + \beta \mathbf{u}' \in \mathcal{U}_{X_0,\mathbf{y}} \cap \mathcal{U}_{X_5,\mathbf{y}}$, then for every $1 \leq i \leq 3$ and $\mathbf{x} \in \{\mathbf{x}_i, \tilde{\mathbf{x}}_i\}$, and every $\beta \geq 0$, we have

$$\alpha y_i = \alpha \sigma(\mathbf{x}^\top(\mathbf{u}^* + \beta \mathbf{u}')) = \alpha \sigma((M\mathbf{x})^\top(M\mathbf{u}^* + M\beta \mathbf{u}')) = \sigma((M\mathbf{x})^\top(\alpha M \mathbf{u}^* + \alpha M \beta \mathbf{u}'))$$
$$= \sigma((M\mathbf{x})^\top(\mathbf{u} + \beta \mathbf{u}_\perp)).$$

Hence, $\mathbf{u} + \beta \mathbf{u}_\perp \in \mathcal{U}_{X_0 M^\top, \alpha \mathbf{y}} \cap \mathcal{U}_{X_5 M^\top, \alpha \mathbf{y}}$. Since $\mathbf{u} \in \text{argmin}_{\bar{\mathbf{u}} \in \mathcal{U}_{X_0 M^\top, \alpha \mathbf{y}}} \mathcal{R}'(\bar{\mathbf{u}})$ and $\mathbf{u} + \mathbf{u}_\perp \in \text{argmin}_{\bar{\mathbf{u}} \in \mathcal{U}_{X_5 M^\top, \alpha \mathbf{y}}} \mathcal{R}'(\bar{\mathbf{u}})$, then we have $\mathcal{R}'(\mathbf{u}) = \mathcal{R}'(\mathbf{u} + \mathbf{u}_\perp) = \min_{\beta \geq 0} \mathcal{R}'(\mathbf{u} + \beta \mathbf{u}_\perp)$.