



# Landscape Analysis for Shallow Neural Networks: Complete Classification of Critical Points for Affine Target Functions

Patrick Cheridito<sup>1</sup> · Arnulf Jentzen<sup>2,3</sup> · Florian Rossmannek<sup>1</sup>

Received: 21 June 2021 / Accepted: 8 June 2022 / Published online: 5 July 2022  
© The Author(s) 2022

## Abstract

In this paper, we analyze the landscape of the true loss of neural networks with one hidden layer and ReLU, leaky ReLU, or quadratic activation. In all three cases, we provide a complete classification of the critical points in the case where the target function is affine and one-dimensional. In particular, we show that there exist no local maxima and clarify the structure of saddle points. Moreover, we prove that non-global local minima can only be caused by ‘dead’ ReLU neurons. In particular, they do not appear in the case of leaky ReLU or quadratic activation. Our approach is of a combinatorial nature and builds on a careful analysis of the different types of hidden neurons that can occur.

**Keywords** Neural networks · Shallow networks · Landscape analysis · Loss surface · Nonconvex optimization

**Mathematics Subject Classification** 68T07

## 1 Introduction

An important aspect of neural network theory in machine learning is the dynamic behavior of gradient-based training algorithms. Although empirical evidence suggests that training is often successful, meaning that the algorithm reaches a point that is close

---

Communicated by Dejan Slepceev.

---

✉ Florian Rossmannek  
florian.rossmannek@math.ethz.ch

<sup>1</sup> Department of Mathematics, ETH Zurich, Zurich, Switzerland

<sup>2</sup> School of Data Science and Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China

<sup>3</sup> Applied Mathematics: Institute for Analysis and Numerics, Faculty of Mathematics and Computer Science, University of Münster, Münster, Germany

to a global minimum of the loss function measuring the error (see, e.g., LeCun et al. 2015), a full theoretical understanding of gradient-based methods in network models is still lacking. One branch of recent research has been investigating the effects of overparametrization, i.e. using an exceedingly large number of neurons in the network model, on the convergence behavior (we refer to Chizat et al. (2019), Allen-Zhu et al. (2019) and the references therein for more details on this), but here we focus on landscape analysis of the loss surface. This landscape analysis provides an indirect tool for studying the dynamics of gradient-based algorithms, as these dynamics are governed by the loss surface. One goal of landscape analysis is a better understanding of the occurrence and frequency of critical points of the loss function and obtaining information about their type, that is, whether they constitute extrema, local extrema, or saddle points. Using the hierarchical structure of networks, some partial results have been obtained (see Fukumizu and Amari 2000). However, the choice of activation function in the network model can have a significant impact on the landscape. For instance, it is known that the loss surface of a linear network, that is, a network with the identity function as activation, only has global minima and saddle points but no non-global local minima (see Baldi and Hornik 1989; Kawaguchi 2016). However, the picture becomes less clear if a nonlinearity is introduced (see Mannelli et al. 2020; Safran and Shamir 2018).

In the last decade, progress has been made in this more difficult nonlinear case. In Choromanska et al. (2015a), the loss surface has been studied by relating it to a model from statistical physics. This way, detailed results have been obtained about the frequency and quality of local minima. Although the findings of Choromanska et al. (2015a) are theoretically insightful, their theory is based on assumptions that are not met in practice (see Choromanska et al. 2015b). In Soudry and Hoffer (2017), similar results have been obtained for networks with one hidden layer with less unrealistic assumptions. We refer to Dauphin et al. (2014) for experimental findings, on which (Choromanska et al. 2015a; Soudry and Hoffer 2017) is based.

Besides work studying the effects of overparametrization on gradient-based methods directly, there have also been investigations of its impact on the loss landscape. For instance, it has been shown in Safran and Shamir (2016) that taking larger networks increases the likelihood to start from a good initialization with small loss or from which there exists a monotonically decreasing path to a global minimum. However, it is still not fully understood in which situations a gradient-based training algorithm follows such a path. If the quadratic activation function is used in a network with one hidden layer, then in the overparametrized regime only global minima and strict saddle points remain, but no non-global local minima; see (Du and Lee 2018; Venturi et al. 2019). Even for deeper architectures, all non-global local minima disappear with high probability for any activation function if the width of the last hidden layer is increased (see Soudry and Carmon 2016; Soltanolkotabi et al. 2019; Livni et al. 2014) and, under some regularity assumptions on the activation, this continues to hold if any of the hidden layers is sufficiently wide and the proceeding layers have a pyramidal structure (see Nguyen and Hein 2017). However, note that these results only apply in this level of generality if the loss is measured with respect to a finite set of data. In particular, these global minima are potentially prone to overfitting.

In contrast to the literature mentioned above, our results concern the landscape of the true loss instead of the empirical loss. The final goal in machine learning is to minimize not only the empirical loss, but the true loss, so it is of essence to understand its landscape. In this paper, we consider networks with a single hidden layer with (leaky) rectified linear unit (ReLU) or quadratic activation. As an alternative to the popular theme of overparametrization, we do not impose assumptions on the network model that are not met in practice, but instead focus on *special target functions*. In Cheridito et al. (2021), this strategy has been pursued with constant target functions. In this paper, we expand the scope from constant to affine functions. This represents a first step toward a better understanding of the true loss landscape corresponding to general target functions.

In this framework with affine target functions, we provide a complete classification of the critical points of the true loss. We do so by unfolding the combinatorics of the problem, governed by different types of hidden neurons appearing in a network. We find that ReLU networks admit non-global local minima regardless of the number of hidden neurons. At the same time, it turns out that these local minima are solely caused by ‘dead’ ReLU neurons. In particular, for leaky ReLU networks, which are often used to avoid the problem of dead neurons, there are only saddle points and global minima. This suggests that using leaky ReLU instead of ReLU not only makes sense to avoid issues with training itself, but also to work with a better behaved loss surface on which training takes place to begin with. Interestingly, also for the quadratic activation, non-global local minima do not appear, which is in line with the observations in Du and Lee (2018), Venturi et al. (2019) for the discretized loss but does not require overparametrization. In addition, for networks with quadratic activation, all saddle points have a constant realization function, whereas for (leaky) ReLU networks we show that there exist saddle points with a non-constant realization.

These complete classifications in the proposed approach to consider special target functions shed new light on important aspects of gradient-based methods in the training of networks. Knowledge of the loss surface can be transformed into results about convergence of such methods as done in, e.g., (Jentzen and Riekert 2021). In a smooth setting, a recent strand of work has shown that the domain of attraction of saddle points under gradient descent has zero Lebesgue measure as long as the Hessian at the saddle points has a strictly negative eigenvalue (see Lee et al. 2019, 2016; Panageas and Piliouras 2017). This indicates that it also becomes necessary to study the spectrum of the Hessian of the loss function as previously pursued in, e.g., (Pennington and Bahri 2017; Du and Lee 2018). Using the classification in this paper, we are able to derive results about the existence of strictly negative eigenvalues of the Hessian at most of the saddle points (understood in a suitable sense because we have to deal with differentiability issues arising from the (leaky) ReLU activation). Furthermore, the set of non-global local minima, being caused by dead ReLU neurons, consists of a single connected component in the parameter space. In particular, these extrema are not isolated. The behavior of (stochastic) gradient descent at not necessarily isolated local minima has been studied in, e.g., (Fehrman et al. 2020).

The remainder of this article is organized as follows. The first activation function we consider is the ReLU activation in Sect. 2. We begin by introducing the relevant notation and definitions, including a new description of the types of hidden neurons

that can appear in a ReLU network, in Sects. 2.1 and 2.2. The first main result, the classification for ReLU networks, is Theorem 2.4 in Sect. 2.3. The remainder of Sect. 2 is dedicated to proving the classification. More precisely, we discuss a few important ingredients for the proof in Sect. 2.4. Thereafter, Sect. 2.5 is devoted to differentiability and regularity properties of the loss function in view of the non-differentiability of the ReLU activation. The heart of the proof is contained in Sects. 2.6 and 2.7. Finally, we establish in Sect. 2.8 a special case of Theorem 2.4 and deduce it in full generality afterward in Sect. 2.9. Section 3 is concerned with extending the classification to leaky ReLU, stated as our second main result in Theorem 3.5, which heavily relies on understanding the ReLU case. To conclude, we also classify the critical points for networks with the quadratic activation in our third main result, Theorem 4.1 in Sect. 4.

## 2 Classification for ReLU Activation

### 2.1 Notation and Formal Problem Description

We consider shallow networks, by which we mean networks with a single hidden layer. For simplicity, we focus on networks with a single input and output neuron. The set of such networks with  $N \in \mathbb{N}$  hidden neurons can be parametrized by  $\mathbb{R}^{3N+1}$ . We begin by describing the problem for the ReLU activation function  $x \mapsto \max\{x, 0\}$ . We will always write an element  $\phi \in \mathbb{R}^{3N+1}$  as  $\phi = (w, b, v, c)$ , where  $w, b, v \in \mathbb{R}^N$  and  $c \in \mathbb{R}$ . The realization of the network  $\phi$  with ReLU activation is the function  $f_\phi \in C(\mathbb{R}, \mathbb{R})$  given by

$$f_\phi(x) = c + \sum_{j=1}^N v_j \max\{w_j x + b_j, 0\}. \quad (2.1)$$

We suppose that the objective is to approximate an affine function on an interval  $[T_0, T_1]$  in the  $L^2$ -norm. In other words, given  $\mathcal{A} = (\alpha, \beta) \in \mathbb{R}^2$  and  $T = (T_0, T_1) \in \mathbb{R}^2$ , one tries to minimize the loss function  $\mathcal{L}_{N,T,\mathcal{A}} \in C(\mathbb{R}^{3N+1}, \mathbb{R})$  given by

$$\mathcal{L}_{N,T,\mathcal{A}}(\phi) = \int_{T_0}^{T_1} (f_\phi(x) - \alpha x - \beta)^2 dx. \quad (2.2)$$

The purpose of the first half of this paper is to classify the critical points of the loss function  $\mathcal{L}_{N,T,\mathcal{A}}$ . Since the ReLU function is not differentiable at 0, we work with the generalized gradient  $\mathcal{G}_{N,T,\mathcal{A}}: \mathbb{R}^{3N+1} \rightarrow \mathbb{R}^{3N+1}$  of the loss obtained by taking right-hand partial derivatives;

$$(\mathcal{G}_{N,T,\mathcal{A}}(\phi))_k = \lim_{h \downarrow 0} \frac{\mathcal{L}_{N,T,\mathcal{A}}(\phi + h e_k) - \mathcal{L}_{N,T,\mathcal{A}}(\phi)}{h} \quad (2.3)$$

for all  $k \in \{1, \dots, 3N + 1\}$ , where  $e_k$  is the  $k$ th unit vector in  $\mathbb{R}^{3N+1}$ . The function  $\mathcal{G}_{N,T,\mathcal{A}}$  is defined on the entire parameter space  $\mathbb{R}^{3N+1}$  and agrees with the gradient

of  $\mathcal{L}_{N,T,\mathcal{A}}$  if the latter exists. We verify this and study regularity properties of  $\mathcal{L}_{N,T,\mathcal{A}}$  more thoroughly in Sect. 2.5.

**Definition 2.1** Let  $N \in \mathbb{N}$  and  $\mathcal{A}, T \in \mathbb{R}^2$ . Then, we call  $\phi \in \mathbb{R}^{3N+1}$  a critical point of  $\mathcal{L}_{N,T,\mathcal{A}}$  if  $\mathcal{G}_{N,T,\mathcal{A}}(\phi) = 0$  and a saddle point if it is a critical point but not a local extremum.<sup>1</sup>

It can be shown that if  $\phi$  is a critical point of  $\mathcal{L}_{N,T,\mathcal{A}}$ , then 0 belongs to the limiting sub-differential of  $\mathcal{L}_{N,T,\mathcal{A}}$ ; see<sup>2</sup> [Eberle et al. (2021), Prop. 2.12]. With Definition 2.1, it is not immediately clear whether all local extrema are critical points. However, we will show that this is the case by demonstrating that local extrema are points of differentiability of the loss function. In particular, Definition 2.1 is well suited for our purposes. The next notion relates the outer bias, i.e., the coordinate  $c$ , to the target function  $x \mapsto \alpha x + \beta$ .

**Definition 2.2** Let  $N \in \mathbb{N}$ ,  $\phi = (w, b, v, c) \in \mathbb{R}^{3N+1}$ ,  $\mathcal{A} = (\alpha, \beta) \in \mathbb{R}^2$ , and  $T = (T_0, T_1) \in \mathbb{R}^2$ . Then, we say that  $\phi$  is  $(T, \mathcal{A})$ -centered if  $c = \frac{\alpha}{2}(T_0 + T_1) + \beta$ .

To motivate this definition, note that  $\frac{\alpha}{2}(T_0 + T_1) + \beta$  is the best constant  $L^2$ -approximation of the function  $[T_0, T_1] \rightarrow \mathbb{R}$ ,  $x \mapsto \alpha x + \beta$ .

## 2.2 Different Types of Hidden Neurons

In this section, we introduce a few notions that describe how the different hidden neurons in a network are contributing to the realization function. In the definition below, we introduce sets  $I_j$ , which are defined such that  $[T_0, T_1] \setminus I_j$  is the interval on which the output of the  $j$ th hidden neuron is rendered zero by the ReLU activation.

**Definition 2.3** Let  $N \in \mathbb{N}$ ,  $\phi = (w, b, v, c) \in \mathbb{R}^{3N+1}$ ,  $j \in \{1, \dots, N\}$ , and  $T_0, T_1 \in \mathbb{R}$  such that  $T_0 < T_1$ . Then, we denote by  $I_j$  the set given by  $I_j = \{x \in [T_0, T_1] : w_j x + b_j \geq 0\}$ , we say that the  $j$ th hidden neuron of  $\phi$  is

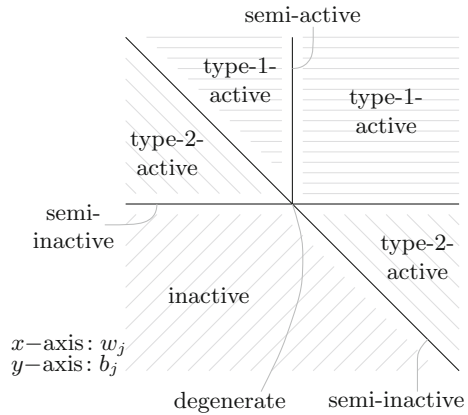
- *inactive* if  $I_j = \emptyset$ ,
- *semi-inactive* if  $\#I_j = 1$ ,
- *semi-active* if  $w_j = 0 < b_j$ ,
- *active* if  $w_j \neq 0 < b_j + \max_{k \in \{0,1\}} w_j T_k$ ,
- *type-1-active* if  $w_j \neq 0 \leq b_j + \min_{k \in \{0,1\}} w_j T_k$ ,
- *type-2-active* if  $\emptyset \neq I_j \cap (T_0, T_1) \neq (T_0, T_1)$ ,
- *degenerate* if  $|w_j| + |b_j| = 0$ ,
- *non-degenerate* if  $|w_j| + |b_j| > 0$ ,
- *flat* if  $v_j = 0$ ,
- *non-flat* if  $v_j \neq 0$ ,

and we say that  $t \in \mathbb{R}$  is the breakpoint of the  $j$ th hidden neuron of  $\phi$  if  $w_j \neq 0 = w_j t + b_j$ .

<sup>1</sup> We consider non-strict local extrema, i.e.  $\phi$  is a local minimum (maximum) of  $\mathcal{L}_{N,T,\mathcal{A}}$  if  $\mathcal{L}_{N,T,\mathcal{A}}(\phi) \leq \mathcal{L}_{N,T,\mathcal{A}}(\psi)$  ( $\geq$ ) for all  $\psi$  in an open neighborhood of  $\phi$ , allowing equality  $\mathcal{L}_{N,T,\mathcal{A}}(\phi) = \mathcal{L}_{N,T,\mathcal{A}}(\psi)$ .

<sup>2</sup> In Eberle et al. (2021), the authors use a different generalization of the gradient, which can be obtained by taking left-hand partial derivatives. However, if  $\mathcal{G}_{N,T,\mathcal{A}}$  is zero at some  $\phi$ , then its left-hand analog is also zero at  $\phi$ , so [Eberle et al. (2021), Prop. 2.12] is applicable.

**Fig. 1** Regions with different types of a hidden neuron as seen in the  $(w_j, b_j)$ -plane in the case  $T_0 = 0, T_1 = 1$ . The general case is obtained by a shear transformation.



Let us briefly motivate these notions. Every hidden neuron is exactly one of: inactive, semi-inactive, semi-active, active, or degenerate. Moreover, observe that  $I_j$  is always an interval.

For an inactive neuron, applying the ReLU activation function yields the constant zero function on  $[T_0, T_1]$ . The breakpoint  $t_j$  might not exist (if  $w_j = 0$  and  $b_j < 0$ ), or it might exist and lie outside of  $[T_0, T_1]$  with  $t_j < T_0$  if  $w_j < 0$  and  $t_j > T_1$  if  $w_j > 0$ . Note that inactivity is a stable condition in the sense that a small perturbation of an inactive neuron remains inactive.

Applying the ReLU activation to a semi-inactive neuron also yields the constant zero function on  $[T_0, T_1]$ . But in this case, a breakpoint must exist and be equal to one of the endpoints  $T_0, T_1$  (which one depends on the sign of  $w_j$  similarly to the inactive case). However, a perturbation of a semi-inactive neuron may yield a (semi-)inactive or a type-2-active neuron; see Fig. 1. In this sense, semi-inactive neurons are boundary cases.

The realization of a semi-active neuron is also constant, but not necessarily zero since the corresponding interval  $I_j$  is  $[T_0, T_1]$ . As can be seen from Fig. 1, perturbing a semi-active neuron always yields a semi- or type-1-active neuron.

Non-flat active neurons provide a non-constant contribution to the overall realization function. Note that a hidden neuron is active exactly if it is type-1- or type-2-active. These two types distinguish whether the breakpoint  $t_j$ , which exists in either case, lies outside or inside the interval  $(T_0, T_1)$  and, hence, whether the contribution of the neuron is affine (corresponding to  $I_j = [T_0, T_1]$ ) or piecewise affine (corresponding to  $I_j = [T_0, t_j]$  or  $I_j = [t_j, T_1]$ ). Type-1 and type-2-active neurons both form two connected components in the  $(w_j, b_j)$ -plane; see Fig. 1. A perturbation of an active neuron remains active.

The case  $w_j = 0 = b_j$  is called degenerate because it leads to problems with differentiability. Perturbing a degenerate neuron may yield any of the other types of neurons.

Lastly, a flat neuron also does not contribute to the overall realization, but the reason for this lies between the second and third layer and not between the first and second one, which is why this case deserves a separate notion.

### 2.3 Classification of the Critical Points of the Loss Function

Now, we are ready to provide a classification of the critical points of the loss function.

**Theorem 2.4** *Let  $N \in \mathbb{N}$ ,  $\phi = (w, b, v, c) \in \mathbb{R}^{3N+1}$ ,  $\mathcal{A} = (\alpha, \beta) \in \mathbb{R}^2$ , and  $T = (T_0, T_1) \in \mathbb{R}^2$  satisfy  $\alpha \neq 0$  and  $0 \leq T_0 < T_1$ . Then, the following hold:*

- (I)  $\phi$  is not a local maximum of  $\mathcal{L}_{N,T,\mathcal{A}}$ .
- (II) If  $\phi$  is a critical point or a local extremum of  $\mathcal{L}_{N,T,\mathcal{A}}$ , then  $\mathcal{L}_{N,T,\mathcal{A}}$  is differentiable at  $\phi$  with gradient  $\nabla \mathcal{L}_{N,T,\mathcal{A}}(\phi) = 0$ .
- (III)  $\phi$  is a non-global local minimum of  $\mathcal{L}_{N,T,\mathcal{A}}$  if and only if  $\phi$  is  $(T, \mathcal{A})$ -centered and, for all  $j \in \{1, \dots, N\}$ , the  $j$ th hidden neuron of  $\phi$  is
  - (a) inactive,
  - (b) semi-inactive with  $I_j = \{T_0\}$  and  $\alpha v_j > 0$ , or
  - (c) semi-inactive with  $I_j = \{T_1\}$  and  $\alpha v_j < 0$ .
- (IV)  $\phi$  is a saddle point of  $\mathcal{L}_{N,T,\mathcal{A}}$  if and only if  $\phi$  is  $(T, \mathcal{A})$ -centered,  $\phi$  does not have any type-1-active neurons,  $\phi$  does not have any non-flat semi-active neurons,  $\phi$  does not have any non-flat degenerate neurons, and exactly one of the following two items holds:
  - (a)  $\phi$  does not have any type-2-active neurons and there exists  $j \in \{1, \dots, N\}$  such that the  $j$ th hidden neuron of  $\phi$  is
    - (i) flat semi-active,
    - (ii) semi-inactive with  $I_j = \{T_0\}$  and  $\alpha v_j \leq 0$ ,
    - (iii) semi-inactive with  $I_j = \{T_1\}$  and  $\alpha v_j \geq 0$ , or
    - (iv) flat degenerate.
  - (b) There exists  $n \in \{2, 4, 6, \dots\}$  such that  $(\bigcup_{j \in \{1, \dots, N\}, w_j \neq 0} \{-\frac{b_j}{w_j}\}) \cap (T_0, T_1) = \bigcup_{i=1}^n \{T_0 + \frac{i(T_1-T_0)}{n+1}\}$  and, for all  $j \in \{1, \dots, N\}$ ,  $i \in \{1, \dots, n\}$  with  $w_j \neq 0 = b_j + w_j(T_0 + \frac{i(T_1-T_0)}{n+1})$ , it holds that  $\text{sign}(w_j) = (-1)^{i+1}$  and  $\sum_{k \in \{1, \dots, N\}, w_k \neq 0 = b_k + w_k(T_0 + \frac{i(T_1-T_0)}{n+1})} v_k w_k = \frac{2\alpha}{n+1}$ .
- (V) If  $\phi$  is a non-global local minimum of  $\mathcal{L}_{N,T,\mathcal{A}}$  or a saddle point of  $\mathcal{L}_{N,T,\mathcal{A}}$  without type-2-active neurons, then  $f_\phi(x) = \frac{\alpha}{2}(T_0+T_1) + \beta$  for all  $x \in [T_0, T_1]$ .
- (VI) If  $\phi$  is a saddle point of  $\mathcal{L}_{N,T,\mathcal{A}}$  with at least one type-2-active neuron, then there exists  $n \in \{2, 4, 6, \dots\}$  such that  $n \leq N$  and, for all  $i \in \{0, \dots, n\}$ ,  $x \in [T_0 + \frac{i(T_1-T_0)}{n+1}, T_0 + \frac{(i+1)(T_1-T_0)}{n+1}]$ , one has

$$f_\phi(x) = \alpha x + \beta - \frac{(-1)^i \alpha}{n+1} \left( x - T_0 - \frac{(i + \frac{1}{2})(T_1 - T_0)}{n+1} \right). \tag{2.4}$$

Theorem 2.4.(IV.b) says that the set of breakpoints of all type-2-active neurons agrees with the set of  $n$  equally spaced points  $T_0 < q_1 < \dots < q_n < T_1$ . Furthermore, for any type-2-active neuron with breakpoint  $q_i$ , the sign of the coordinate  $w$  is given by  $(-1)^{i+1}$ . Lastly, the sum of  $v_k w_k$ , where  $k$  ranges over all type-2-active neurons with breakpoint  $q_i$ , is equal to  $\frac{2\alpha}{n+1}$ . The term  $v_k w_k$  is the contribution of the  $k$ th hidden neuron to the slope of the realization.

**Remark 2.5** Note that, by Theorem 2.4.(II), all local extrema and all critical points of  $\mathcal{L}_{N,T,\mathcal{A}}$ , which we defined as zeros of  $\mathcal{G}_{N,T,\mathcal{A}}$ , are actually critical points of  $\mathcal{L}_{N,T,\mathcal{A}}$  in the classical sense, i.e. points of differentiability of  $\mathcal{L}_{N,T,\mathcal{A}}$  with vanishing gradient. In particular, the classification in Theorem 2.4 turns out to be a classification of the critical points in the classical sense as well.

**Remark 2.6** Gradient descent-type algorithms typically use generalized gradients to train ReLU networks. For instance, they might compute  $\mathcal{G}$ , its left-hand analog, the average of the two, or quantities obtained by artificially defining the derivative of the ReLU function at 0. For each of these versions, a similar classification of critical points could be derived.

Theorem 2.4.(V) shows that any non-global local minimum has the constant realization  $\frac{\alpha}{2}(T_0 + T_1) + \beta$ . In particular, there is only one value that the loss function can take at non-global local minima. Similarly, it follows from Theorem 2.4.(VI) that a saddle point can lead to exactly one of  $\lfloor N/2 \rfloor + 1$  possible loss values.

**Corollary 2.7** Let  $N \in \mathbb{N}$ ,  $\mathcal{A} = (\alpha, \beta) \in \mathbb{R}^2$ , and  $T = (T_0, T_1) \in \mathbb{R}^2$  satisfy  $0 \leq T_0 < T_1$ , and assume that  $\phi \in \mathbb{R}^{3N+1}$  is a critical point of  $\mathcal{L}_{N,T,\mathcal{A}}$ . Then, the following hold:

- (i) If  $\phi$  is a non-global local minimum of  $\mathcal{L}_{N,T,\mathcal{A}}$ , then  $\mathcal{L}_{N,T,\mathcal{A}}(\phi) = \frac{1}{12}\alpha^2(T_1 - T_0)^3$ .
- (ii) If  $\phi$  is a saddle point of  $\mathcal{L}_{N,T,\mathcal{A}}$ , then there exists  $n \in \{0, 2, 4, \dots\}$  such that  $n \leq N$  and  $\mathcal{L}_{N,T,\mathcal{A}}(\phi) = \frac{1}{12(n+1)^4}\alpha^2(T_1 - T_0)^3$ .

Formally, Corollary 2.7 only follows from Theorem 2.4 for  $\alpha \neq 0$ . But for  $\alpha = 0$ , it holds trivially since for constant target functions there exist no critical points other than global minima (see Cheridito et al. 2021).

### 2.4 Ingredients for the Proof of the Classification

As a first step, let us provide a simple argument to establish Theorem 2.4.(I).

**Lemma 2.8** Let  $N \in \mathbb{N}$ ,  $\mathcal{A} \in \mathbb{R}^2$ , and  $T = (T_0, T_1) \in \mathbb{R}^2$  satisfy  $T_0 < T_1$ . Then,  $\mathcal{L}_{N,T,\mathcal{A}}$  does not have any local maxima.

**Proof** Write  $\mathcal{A} = (\alpha, \beta)$ . The lemma directly follows from the simple fact that

$$\mathcal{L}_{N,T,\mathcal{A}}(w, b, v, c) = \int_{T_0}^{T_1} \left( c + \sum_{j=1}^N v_j \max\{w_j x + b_j, 0\} - \alpha x - \beta \right)^2 dx \tag{2.5}$$

is strictly convex in  $c$ . □

As a consequence of this lemma, whenever we want to show that a critical point  $\phi$  is a saddle point, it suffices to show that it is not a local minimum, that is, it suffices to show that, in every neighborhood of  $\phi$ ,  $\mathcal{L}$  attains a value that is below  $\mathcal{L}(\phi)$ .



**Remark 2.9** The previous proof only used linearity of the realization function in the  $c$ -coordinate and strict convexity of the square function. In particular, the same argument shows that the square loss never has local maxima regardless of the target function, the activation function, and the architecture of the network.

Let us now provide a sketch of the proofs to come. Instead of proving Theorem 2.4 directly, we first assume that the affine target function is the identity on the interval  $[0, 1]$ , corresponding to the special case  $T_0 = \beta = 0$  and  $T_1 = \alpha = 1$  in Theorem 2.4. Afterward, we will verify that the general case can always be reduced to this one. For convenience of notation, we assume the following convention to hold throughout the remainder of Sect. 2.

**Setting 2.10** Fix  $N \in \mathbb{N}$  and denote  $\mathcal{L} = \mathcal{L}_{N,(0,1),(1,0)}$  and  $\mathcal{G} = \mathcal{G}_{N,(0,1),(1,0)}$ . We say that a network  $\phi \in \mathbb{R}^{3N+1}$  is centered if it is  $((0, 1), (1, 0))$ -centered.

The generalized gradient  $\mathcal{G}$  was defined in terms of the right-hand partial derivatives of  $\mathcal{L}$ . These are given by

$$\begin{aligned} \frac{\partial^+}{\partial w_j} \mathcal{L}(\phi) &= 2v_j \int_{I_j} x(f_\phi(x) - x)dx, & \frac{\partial^+}{\partial b_j} \mathcal{L}(\phi) &= 2v_j \int_{I_j} (f_\phi(x) - x)dx, \\ \frac{\partial^+}{\partial v_j} \mathcal{L}(\phi) &= 2 \int_{I_j} (w_j x + b_j)(f_\phi(x) - x)dx, & \frac{\partial^+}{\partial c} \mathcal{L}(\phi) &= 2 \int_0^1 (f_\phi(x) - x)dx. \end{aligned} \tag{2.6}$$

Regularity properties of the loss function will be discussed in detail in the next section. We will see then that these right-hand partial derivatives are proper partial derivatives if the  $j$ th hidden neuron is flat or non-degenerate. If these partial derivatives are zero, then we encounter the system of equations

$$\begin{aligned} 0 &= 2v_j \int_{I_j} x(f_\phi(x) - x)dx, \\ 0 &= 2v_j \int_{I_j} (f_\phi(x) - x)dx, \\ 0 &= 2 \int_{I_j} (w_j x + b_j)(f_\phi(x) - x)dx, \\ 0 &= 2 \int_0^1 (f_\phi(x) - x)dx, \end{aligned} \tag{2.7}$$

from which we deduce that any non-flat non-degenerate neuron of a critical point or local extremum  $\phi$  satisfies

$$\int_{I_j} (f_\phi(x) - x)dx = 0 = \int_{I_j} x(f_\phi(x) - x)dx. \tag{2.8}$$

This simple observation will be used repeatedly in the proof of Theorem 2.4. Moreover, for a type-1-active neuron (for which  $I_j = [0, 1]$ ), (2.8) is even satisfied if the neuron

is flat as can be seen from the third and fourth line of (2.7). Here is an example of how (2.8) can be employed: note that any affine function  $f : [0, 1] \rightarrow \mathbb{R}$  satisfying

$$\int_0^1 (f(x) - x)dx = 0 = \int_0^1 x(f(x) - x)dx \tag{2.9}$$

necessarily equals the identity on  $[0, 1]$ . Thus, if  $\phi$  is a critical point or local extremum of  $\mathcal{L}$  for which  $f_\phi$  is affine and if  $\phi$  admits a type-1-active or non-flat semi-active neuron (so that  $I_j = [0, 1]$ ), then we obtain from (2.8) that  $\phi$  is a global minimum. If  $f_\phi$  is not affine, we will be able to develop similar arguments for each affine piece of  $f_\phi$ . In this case, we will obtain a system of equations from (2.7) that intricately describes the combinatorics of the realization function.

### 2.5 Differentiability of the Loss Function

Since the ReLU function is not differentiable at 0, the loss function is not everywhere differentiable. However, a simple argument establishes that  $\mathcal{L}$  is differentiable at any of its global minima as the following lemma shows.

**Lemma 2.11** *Let  $\phi \in \mathbb{R}^{3N+1}$ . If  $f_\phi(x) = x$  for all  $x \in [0, 1]$ , then  $\mathcal{L}$  is differentiable at  $\phi$ .*

**Proof** It is well known that the realization function  $\mathbb{R}^{3N+1} \rightarrow C([0, 1], \mathbb{R})$ ,  $\phi \mapsto f_\phi|_{[0,1]}$  is locally Lipschitz continuous if  $C([0, 1], \mathbb{R})$  is equipped with the supremum norm (see, e.g., Petersen et al. 2020). Thus, there is a constant  $L > 0$  depending only on  $N$  and  $\phi$  with  $|f_{\phi+\psi}(x) - f_\phi(x)| \leq L\|\psi\|$  uniformly on  $[0, 1]$  for all  $\psi$  sufficiently close to  $\phi$ . Then,

$$\frac{\mathcal{L}(\phi + \psi) - \mathcal{L}(\phi)}{\|\psi\|} = \frac{1}{\|\psi\|} \int_0^1 (f_{\phi+\psi}(x) - f_\phi(x))^2 dx \leq L^2 \|\psi\|, \tag{2.10}$$

which shows that  $\mathcal{L}$  is differentiable at  $\phi$ . □

The next result shows that there even are regions in the parameter space where  $\mathcal{L}$  is infinitely often differentiable in spite of the ReLU activation.

**Lemma 2.12** *The loss function  $\mathcal{L}$  is everywhere analytic in  $(v, c)$ . Moreover, if the  $j$ th hidden neuron of  $\phi \in \mathbb{R}^{3N+1}$  is inactive, semi-active, or type-1-active with breakpoint neither 0 nor 1 for some  $j \in \{1, \dots, N\}$ , then  $\mathcal{L}$  is also analytic in  $(w_j, b_j, v, c)$  in a neighborhood of  $\phi$ , and mixed partial derivatives of any order can be obtained by differentiating under the integral. In particular,*

$$\begin{aligned} \frac{\partial}{\partial w_j} \mathcal{L}(\phi) &= 2v_j \int_{I_j} x(f_\phi(x) - x)dx, & \frac{\partial}{\partial b_j} \mathcal{L}(\phi) &= 2v_j \int_{I_j} (f_\phi(x) - x)dx, \\ \frac{\partial}{\partial v_j} \mathcal{L}(\phi) &= 2 \int_{I_j} (w_j x + b_j)(f_\phi(x) - x)dx, & \frac{\partial}{\partial c} \mathcal{L}(\phi) &= 2 \int_0^1 (f_\phi(x) - x)dx. \end{aligned} \tag{2.11}$$

**Proof** For the first part, note that  $\mathcal{L}$  is a polynomial in the coordinates  $(v, c)$ . Secondly, assume that the  $j$ th hidden neuron of  $\phi^0 \in \mathbb{R}^{3N+1}$  is inactive. Then, for all  $\phi$  in a sufficiently small neighborhood of  $\phi^0$  and all  $x \in [0, 1]$  we have  $\max\{w_j x + b_j, 0\} = 0$ . Hence,  $\mathcal{L}$  is constant in the coordinates  $(w_j, b_j)$  near  $\phi^0$  and it is a polynomial in  $(w_j, b_j, v, c)$ . Thirdly, assume that the  $j$ th hidden neuron of  $\phi^0$  is semi-active or type-1-active with breakpoint neither 0 nor 1. Then, for all  $\phi$  in a sufficiently small neighborhood of  $\phi^0$  and all  $x \in [0, 1]$  we have  $\max\{w_j x + b_j, 0\} = w_j x + b_j$ . In particular,  $\mathcal{L}$  is a polynomial in the coordinates  $(w_j, b_j, v, c)$  near  $\phi^0$ . The statement about differentiating under the integral follows from dominated convergence.  $\square$

In regions of the parameter space not covered by Lemma 2.12, we cannot guarantee as much regularity of the loss function, but we can still hope for differentiability. Indeed, we already noted in the proof of Lemma 2.11 that the realization function  $\mathbb{R}^{3N+1} \rightarrow C([0, 1], \mathbb{R}), \phi \mapsto f_\phi|_{[0,1]}$  is locally Lipschitz continuous. So, it follows from Rademacher’s theorem that  $\mathcal{G}$  is, in fact, equal to the true gradient  $\nabla\mathcal{L}$  of  $\mathcal{L}$  almost everywhere. In the next result, we obtain insights about the measure-zero set on which  $\mathcal{G}$  may not be the true gradient.

**Lemma 2.13** *For all  $j \in \{1, \dots, N\}$ , the right-hand partial derivatives  $\partial^+\mathcal{L}(\phi)/\partial w_j$  and  $\partial^+\mathcal{L}(\phi)/\partial b_j$  exist everywhere and are given by*

$$\frac{\partial^+}{\partial w_j} \mathcal{L}(\phi) = 2v_j \int_{I_j} x(f_\phi(x) - x)dx \quad \text{and} \quad \frac{\partial^+}{\partial b_j} \mathcal{L}(\phi) = 2v_j \int_{I_j} (f_\phi(x) - x)dx. \tag{2.12}$$

*Moreover, if the  $j$ th hidden neuron is flat or non-degenerate, then  $\mathcal{L}$  is differentiable in  $(w_j, b_j, v, c)$  and, in particular, the right-hand partial derivatives  $\partial^+\mathcal{L}(\phi)/\partial w_j$  and  $\partial^+\mathcal{L}(\phi)/\partial b_j$  are proper partial derivatives.*

**Proof** Let  $\phi \in \mathbb{R}^{3N+1}$  be arbitrary and denote by  $\phi_h, h = (h^1, h^2) \in \mathbb{R}^2$ , the network with the same coordinates as  $\phi$  except in the  $j$ th hidden neuron, where  $\phi_h$  has coordinates  $w_j + h^1$  and  $b_j + h^2$ . We use the notation  $I_j^h$  for the interval  $I_j$  associated to  $\phi_h$  and denote

$$\varepsilon = \mathcal{L}(\phi_h) - \mathcal{L}(\phi) - 2v_j h^1 \int_{I_j} x(f_\phi(x) - x)dx - 2v_j h^2 \int_{I_j} (f_\phi(x) - x)dx. \tag{2.13}$$

The proof is complete if we can show that  $\varepsilon$  goes to zero faster than  $(h^1, h^2)$ . To do that, we estimate the two terms of the last line of

$$\begin{aligned} \varepsilon &= \int_0^1 (f_{\phi_h}(x) - f_\phi(x))^2 dx + 2 \int_0^1 (f_{\phi_h}(x) - f_\phi(x))(f_\phi(x) - x) dx \\ &\quad - 2v_j \int_{I_j} (h^1 x + h^2)(f_\phi(x) - x) dx = \int_0^1 (f_{\phi_h}(x) - f_\phi(x))^2 dx \end{aligned}$$

$$+2v_j \int_0^1 (w_j x + b_j + h^1 x + h^2)(f_\phi(x) - x)(\mathbb{1}_{I_j^h}(x) - \mathbb{1}_{I_j}(x))dx. \tag{2.14}$$

To control the first term, we use local Lipschitz continuity of the realization function, which yields a constant  $L > 0$  depending only on  $\phi$  so that  $|f_{\phi_h}(x) - f_\phi(x)| \leq L(|h^1| + |h^2|)$  uniformly on  $[0, 1]$  for all sufficiently small  $h$ . To estimate the second term, we note that the absolute value of  $\mathbb{1}_{I_j^h} - \mathbb{1}_{I_j}$  is the indicator function of the symmetric difference  $I_j \Delta I_j^h$ . By definition of these sets, we obtain the bound  $|w_j x + b_j| \leq |h^1 x + h^2|$  for any  $x \in I_j \Delta I_j^h$ . This yields

$$\frac{|\varepsilon|}{|h^1| + |h^2|} \leq L^2(|h^1| + |h^2|) + 4|v_j| \int_0^1 |f_\phi(x) - x| \mathbb{1}_{I_j \Delta I_j^h}(x) dx. \tag{2.15}$$

The term  $L^2(|h^1| + |h^2|)$  vanishes as  $h \rightarrow 0$ . We need to argue that the second term also vanishes as  $h \rightarrow 0$ . If the  $j$ th hidden neuron is flat, then the second term is trivially zero. On the other hand, if the  $j$ th hidden neuron is non-degenerate, then the Lebesgue measure of  $I_j \Delta I_j^h$  tends to zero as  $h \rightarrow 0$ . Thus, in this case, the integral also vanishes as  $h \rightarrow 0$ . If the  $j$ th hidden neuron is non-flat degenerate, then we consider the directional derivatives from the right, i.e. with  $h^1, h^2 \downarrow 0$ . But then  $I_j = [0, 1] = I_j^h$ , so  $\mathbb{1}_{I_j \Delta I_j^h}$  is constantly zero.  $\square$

It is well known that a multivariate function is continuously differentiable if it has continuous partial derivatives. The following result is a slight extension for the loss function  $\mathcal{L}$ .

**Lemma 2.14** *The loss function  $\mathcal{L}$  is continuously differentiable on the set of networks without degenerate neurons. In addition,  $\mathcal{L}$  is differentiable at networks without non-flat degenerate neurons.*

**Proof** The preceding two results established existence of all partial derivatives of first order at networks without degenerate neurons. Furthermore, these partial derivatives are continuous in the network parameters. This is clear for  $(v, c)$  and it also holds for  $(w, b)$  because the endpoints of  $I_j$  vary continuously in  $w_j$  and  $b_j$  as long as not both are zero. This concludes the first statement.

To prove that  $\mathcal{L}$  is still differentiable if flat degenerate neurons appear, assume without loss of generality that the first  $M \leq N$  hidden neurons of  $\phi \in \mathbb{R}^{3N+1}$  are flat degenerate and the remaining  $N - M$  hidden neurons are non-degenerate. Denote by  $\phi_1 \in \mathbb{R}^{3M+1}$  the network comprised of the first  $M$  hidden neurons of  $\phi$  (with zero outer bias) and by  $\phi_2 \in \mathbb{R}^{3(N-M)+1}$  the network comprised of the last  $N - M$  hidden neurons. We write  $\mathcal{L}_{N-M}$  for the loss defined on networks with  $N - M$  hidden neurons. Then, for any perturbation  $\phi_h = \phi + h \in \mathbb{R}^{3N+1}$  of  $\phi$  with the same decomposition into its first  $M$  and last  $N - M$  hidden neurons, we can write  $f_{\phi_h}(x) = f_{\phi_{1,h}}(x) + f_{\phi_{2,h}}(x)$  and, hence,

$$\mathcal{L}(\phi_h) = \int_0^1 f_{\phi_{1,h}}(x)^2 dx + 2 \int_0^1 f_{\phi_{1,h}}(x)(f_{\phi_{2,h}}(x) - x) dx + \mathcal{L}_{N-M}(\phi_{2,h}). \tag{2.16}$$

Since the first  $M$  hidden neurons of  $\phi$  are flat degenerate,  $f_{\phi_{1,h}}(x)$  is given by

$$f_{\phi_{1,h}}(x) = \sum_{j=1}^M h_{j+2N} \max\{h_j x + h_{j+N}, 0\}. \tag{2.17}$$

In particular,  $f_{\phi_{1,h}}(x)/\|h\| \rightarrow 0$  uniformly in  $x \in [0, 1]$  as  $h \rightarrow 0$ . Denote by  $\tilde{h}$  the last  $3(N - M)$  components of  $h$ . Since  $\phi_2$  has only non-degenerate neurons,  $\mathcal{L}_{N-M}$  is differentiable at  $\phi_2$  with some gradient  $A$ . Using that the first  $M$  hidden neurons of  $\phi$  do not contribute to its realization and, hence,  $\mathcal{L}(\phi) = \mathcal{L}_{N-M}(\phi_2)$ , we find

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{\mathcal{L}(\phi_h) - \mathcal{L}(\phi) - A\tilde{h}}{\|h\|} &= \lim_{h \rightarrow 0} \frac{\mathcal{L}_{N-M}(\phi_{2,h}) - \mathcal{L}_{N-M}(\phi_2) - A\tilde{h}}{\|\tilde{h}\|} \frac{\|\tilde{h}\|}{\|h\|} \\ &+ \lim_{h \rightarrow 0} \frac{1}{\|h\|} \left( \int_0^1 f_{\phi_{1,h}}(x)^2 dx + 2 \int_0^1 f_{\phi_{1,h}}(x)(f_{\phi_{2,h}}(x) - x) dx \right) = 0. \end{aligned} \tag{2.18}$$

This proves differentiability of  $\mathcal{L}$  at  $\phi$ . □

So far, we have seen that, in some regions of the parameter space, the loss is differentiable while in others it may not be. In the following, we show that, for type-2-active neurons, one even has twice continuous differentiability.

**Lemma 2.15** *Let  $i, j \in \{1, \dots, N\}$ . If the  $i$ th and  $j$ th hidden neuron of  $\phi \in \mathbb{R}^{3N+1}$  are type-2-active, then  $\mathcal{L}$  is twice continuously differentiable in  $(w_i, w_j, b_i, b_j, v, c)$  in a neighborhood of  $\phi$  in  $\mathbb{R}^{3N+1}$ .*

**Proof** Note that we established twice continuous differentiability of  $\mathcal{L}$  in  $(v, c)$  in Lemma 2.12. Suppose the  $i$ th and  $j$ th hidden neuron of  $\phi^0 = (w^0, b^0, v^0, c^0) \in \mathbb{R}^{3N+1}$  are type-2-active. Since a small perturbation of a type-2-active neuron remains type-2-active and since a type-2-active neuron is non-degenerate, it follows from Lemma 2.13 that  $\mathcal{L}$  is differentiable in  $(w_j, b_j)$  in a neighborhood  $U \subseteq \mathbb{R}^{3N+1}$  of  $\phi^0$  with partial derivatives

$$\frac{\partial}{\partial w_j} \mathcal{L}(\phi) = 2v_j \int_{I_j} x(f_\phi(x) - x) dx \quad \text{and} \quad \frac{\partial}{\partial b_j} \mathcal{L}(\phi) = 2v_j \int_{I_j} (f_\phi(x) - x) dx \tag{2.19}$$

for any  $\phi = (w, b, v, c) \in U$ . Because the  $j$ th hidden neuron is assumed to be type-2-active, the interval  $I_j^0$  is exactly  $[0, t_j^0]$  or  $[t_j^0, 1]$  for the breakpoint  $t_j^0 = -b_j^0/w_j^0 \in (0, 1)$ . Assume  $I_j^0 = [0, t_j^0]$  as the other case is dealt with analogously. By shrinking  $U$  if necessary, we therefore integrate over  $[0, -b_j/w_j]$  in the above partial derivatives for all  $\phi = (w, b, v, c) \in U$ . In particular, the integration boundaries vary smoothly in  $(w_j, b_j)$  in  $U$ . So, it follows from Leibniz' rule that these partial derivatives are continuously differentiable with respect to  $(w_j, b_j)$ . Furthermore, since  $t_j = -b_j/w_j$  does not depend on  $(w_i, b_i, v, c)$ , it follows from dominated convergence that  $\partial \mathcal{L}(\phi)/\partial w_j$

and  $\partial\mathcal{L}(\phi)/\partial b_j$  are also differentiable with respect to  $(w_i, b_i, v, c)$ . The mixed partial derivative with respect to  $w_i$  and  $w_j$  is given by

$$\frac{\partial}{\partial w_i} \frac{\partial}{\partial w_j} \mathcal{L}(\phi) = 2v_j \int_{I_j} x \frac{\partial}{\partial w_i} f_\phi(x) dx = 2v_i v_j \int_{I_i \cap I_j} x^2 dx. \tag{2.20}$$

That the  $i$ th and  $j$ th hidden neuron are type-2-active ensures that  $\int_{I_i \cap I_j} x^2 dx$  is continuous in  $(w_i, w_j, b_i, b_j)$  and, hence, that  $\partial^2\mathcal{L}(\phi)/(\partial w_i \partial w_j)$  is continuous in  $(w_i, w_j, b_i, b_j, v, c)$ . Analogous considerations show that all mixed partial derivatives with respect to  $w_i, w_j, b_i, b_j, v, c$  up to second order exist and are continuous. Thus,  $\mathcal{L}$  restricted to  $(w_i, w_j, b_i, b_j, v, c)$  is twice continuously differentiable in a neighborhood of  $\phi^0$ .  $\square$

**Remark 2.16** We mentioned in Remark 2.5 that all critical points and local extrema of  $\mathcal{L}$  are actually proper critical points and, hence, the classification actually does not deal with points of non-differentiability. Furthermore, by modifying the gradient descent algorithm and the initialization in an appropriate way, one can ensure that the trajectories of the algorithm avoid any points of non-differentiability; see (Wojtowycsh 2020) and also the appendix in Chizat and Bach (2020). Nonetheless, to formally prove the classification, including that all critical points are proper, an extensive regularity analysis of the loss function as done in this section is necessary.

**2.6 Critical Points of the Loss Function with Affine Realization**

In this and the next section, we develop the building blocks necessary for proving the main result. The first lemma establishes one direction of the equivalence in Theorem 2.4.(III).

**Lemma 2.17** *Suppose  $\phi \in \mathbb{R}^{3N+1}$  is centered and all of its hidden neurons satisfy one of the properties (III.a)–(III.c) in Theorem 2.4. Then,  $\phi$  is a local minimum of  $\mathcal{L}$ .*

**Proof** Denote by  $J_0 \subseteq \{1, \dots, N\}$  the set of those hidden neurons of  $\phi$  that satisfy Theorem 2.4.(III.b), and, likewise, denote by  $J_1 \subseteq \{1, \dots, N\}$  the set of those hidden neurons of  $\phi$  that satisfy Theorem 2.4.(III.c). Write  $\phi = (w^0, b^0, v^0, c^0)$  and consider  $\psi = (w, b, v, c) \in U$  in a small neighborhood  $U$  of  $\phi$ . Since a small perturbation of an inactive neuron remains inactive, we have for all  $\psi \in U$  and every  $x \in [0, 1]$  that

$$f_\psi(x) = c + \sum_{j \in J_0 \cup J_1} v_j \max\{w_j x + b_j, 0\} \tag{2.21}$$

if  $U$  is small enough. Moreover, for any  $j \in J_0$  and  $\psi \in U$ , note that  $\max\{w_j x + b_j, 0\} = 0$  for all  $x \in [1/4, 1]$ . Similarly,  $\max\{w_j x + b_j, 0\} = 0$  for all  $x \in [0, 3/4]$  if  $j \in J_1$ . Since we also know  $v_j^0 > 0$  for all  $j \in J_0$  and  $v_j^0 < 0$  for all  $j \in J_1$ , we

find that the realization of  $\psi \in U$  satisfies

$$f_\psi(x) = \begin{cases} c + \sum_{j \in J_0} v_j \max\{w_j x + b_j, 0\} \geq c & \text{if } x \in [0, 1/4] \\ c & \text{if } x \in [1/4, 3/4] \\ c + \sum_{j \in J_1} v_j \max\{w_j x + b_j, 0\} \leq c & \text{if } x \in [3/4, 1] \end{cases} \quad (2.22)$$

for sufficiently small  $U$ . In particular, it follows that  $|f_\psi(x) - x| \geq |c - x|$  for all  $x \in [0, 1]$  and, because  $\phi$  is centered, that

$$\mathcal{L}(\psi) \geq \int_0^1 (c - x)^2 dx \geq \int_0^1 (\frac{1}{2} - x)^2 dx = \mathcal{L}(\phi). \quad (2.23)$$

Thus,  $\phi$  is a local minimum. □

The proof of the next lemma revolves, for the most part, around the argument (2.9), presented in Sect. 2.4. The last statement of the lemma paired with Lemma 2.14 shows that saddle points with an affine realization are also points of differentiability of  $\mathcal{L}$ .

**Lemma 2.18** *Suppose  $\phi \in \mathbb{R}^{3N+1}$  is a critical point or a local extremum of  $\mathcal{L}$  but not a global minimum and that  $f_\phi$  is affine on  $[0, 1]$ . Then,  $\phi$  is centered and does not have any active or non-flat semi-active neurons, so, in particular,  $f_\phi \equiv 1/2$ . Moreover, if  $\phi$  is a saddle point, then it also does not have any non-flat degenerate neurons.*

**Proof** We know from Lemma 2.13 that  $\mathcal{L}$  is differentiable in those coordinates that correspond to non-degenerate neurons and its partial derivatives must vanish at  $\phi$ . Thus, the argument using (2.9) shows that  $\phi$  does not have any type-1-active or non-flat semi-active neurons. If  $\phi$  had a non-flat type-2-active neuron, say the  $j$ th, then we could, using the same argument with  $I_j$  in place of  $[0, 1]$ , conclude that  $f_\phi(x) = x$  on  $I_j$ . But since  $f_\phi$  was assumed to be affine, this could only be true if  $\phi$  were a global minimum. Having no type-1-active or non-flat type-2-active neurons,  $f_\phi$  must be constant. By the fourth equation of (2.7), this constant is  $1/2$ , so  $\phi$  is centered.

Next, suppose that the  $j$ th hidden neuron is flat type-2-active. In particular,  $I_j = [0, t_j]$  or  $I_j = [t_j, 1]$ , where  $t_j = -b_j/w_j \in (0, 1)$  is the breakpoint. After dividing by  $2w_j$ , the integral in the third equation of (2.7) evaluates to

$$0 = \int_{I_j} (x - t_j)(\frac{1}{2} - x) dx = \begin{cases} -\frac{1}{6}t_j^2(\frac{3}{2} - t_j) & \text{if } I_j = [0, t_j] \\ -\frac{1}{6}(1 - t_j)^2(t_j + \frac{1}{2}) & \text{if } I_j = [t_j, 1] \end{cases} \neq 0, \quad (2.24)$$

yielding a contradiction. Lastly, suppose  $\phi$  is a saddle point. If there were a non-flat degenerate neuron, then  $\mathcal{G}(\phi) = 0$  would imply  $0 = \int_0^1 x(f_\phi(x) - x) dx$ . But since we know that  $f_\phi(x) \equiv 1/2$ , this cannot be. □

The next lemma serves as the basis of Theorem 2.4.(IV.a). However, note that we also consider the possibility of a non-flat degenerate neuron, whereas Theorem 2.4.(IV.a.iv) requires the degenerate neuron to be flat. This generalization is needed in

the proof of Theorem 2.4.(III), which will be given later by way of contradiction. In addition, Lemma 2.19 shows that non-global local minima with an affine realization cannot have non-flat degenerate neurons and, hence, are points of differentiability of  $\mathcal{L}$  by Lemma 2.14. Together with the preceding lemma and Lemmas 2.11 and 2.14, we conclude that all critical points and local extrema with an affine realization are points of differentiability.

**Lemma 2.19** *Suppose  $\phi \in \mathbb{R}^{3N+1}$  is a critical point or a local extremum of  $\mathcal{L}$  but not a global minimum and that  $f_\phi$  is affine on  $[0, 1]$ . Suppose further that at least one of its hidden neurons satisfies one of the properties (IV.a.i)–(IV.a.iii) in Theorem 2.4 or is degenerate. Then,  $\phi$  is a saddle point.*

**Proof** Since, by Lemma 2.8,  $\mathcal{L}$  cannot have any local maxima, it is enough to show that  $\mathcal{L}$  is strictly decreasing along some direction starting from  $\phi$ . First, assume that the  $j$ th hidden neuron of  $\phi$  is flat semi-active. Then, Lemma 2.12 asserts smoothness of the loss in the coordinates of the  $j$ th hidden neuron and

$$\begin{aligned} \frac{\partial}{\partial w_j} \frac{\partial}{\partial w_j} \mathcal{L}(\phi) &= 2v_j \int_0^1 x \frac{\partial}{\partial w_j} f_\phi(x) dx = 0, \\ \frac{\partial}{\partial v_j} \frac{\partial}{\partial w_j} \mathcal{L}(\phi) &= 2v_j \int_0^1 x \frac{\partial}{\partial v_j} f_\phi(x) dx + 2 \int_0^1 x (f_\phi(x) - x) dx \\ &= 2 \int_0^1 x (f_\phi(x) - x) dx =: R, \\ \frac{\partial}{\partial v_j} \frac{\partial}{\partial v_j} \mathcal{L}(\phi) &= 2 \int_0^1 (w_j x + b_j) \frac{\partial}{\partial v_j} f_\phi(x) dx =: S, \end{aligned} \tag{2.25}$$

where we used that the  $j$ th hidden neuron is flat. Since  $2 \int_0^1 (f_\phi(x) - x) dx = \frac{\partial}{\partial c} \mathcal{L}(\phi) = 0$ , we must have  $R \neq 0$  for otherwise  $\phi$  would be a global minimum by the argument (2.9). This yields

$$\det \begin{pmatrix} \frac{\partial}{\partial w_j} \frac{\partial}{\partial w_j} \mathcal{L}(\phi) & \frac{\partial}{\partial w_j} \frac{\partial}{\partial v_j} \mathcal{L}(\phi) \\ \frac{\partial}{\partial v_j} \frac{\partial}{\partial w_j} \mathcal{L}(\phi) & \frac{\partial}{\partial v_j} \frac{\partial}{\partial v_j} \mathcal{L}(\phi) \end{pmatrix} = \det \begin{pmatrix} 0 & R \\ R & S \end{pmatrix} = -R^2 < 0. \tag{2.26}$$

In particular, this matrix must have a strictly negative eigenvalue, and a second order expansion of the loss restricted to  $(w_j, v_j)$  shows that  $\mathcal{L}$  is strictly decreasing along the direction of an eigenvector associated to this negative eigenvalue.

Next, assume that the  $j$ th hidden neuron is semi-inactive with  $I_j = \{0\}$  and  $v_j \leq 0$  (case one) or that it is degenerate with  $v_j \leq 0$  (case two). In either case, note that  $b_j = 0$  and consider the perturbation  $\phi_s = (w^s, b^s, v^s, c^s)$ ,  $s \in [0, 1]$ , of  $\phi = \phi_0$  given by  $w_j^s = w_j - s$ ,  $b_j^s = -s w_j^s$ , and  $v_j^s = v_j - s$  (all other coordinates coincide with those of  $\phi$ ). Note that we have  $w_j^s < 0$  and  $v_j^s < 0$  for all  $s \in (0, 1]$  in both cases. For simplicity, denote  $a^s = v_j^s w_j^s$ . By Lemma 2.18, we already know that  $\phi$  is centered and does not have any active or non-flat semi-active neurons. Thus, for every



$s, x \in [0, 1]$ , we can write

$$\begin{aligned} f_{\phi_s}(x) &= c + v_j^s \max\{w_j^s x + b_j^s, 0\} = c + v_j^s \max\{w_j^s(x - s), 0\} \\ &= \frac{1}{2} + a^s(x - s)\mathbb{1}_{[0,s]}(x). \end{aligned} \tag{2.27}$$

Using this formula, we have for all  $s \in [0, 1]$

$$\begin{aligned} \mathcal{L}(\phi_s) - \mathcal{L}(\phi) &= \int_0^s [a^s(x - s)]^2 dx - \int_0^s 2a^s(x - s)(x - \frac{1}{2}) dx \\ &= \frac{1}{3}a^s(a^s + 1)s^3 - \frac{1}{2}a^s s^2 \\ &= \begin{cases} -\frac{1}{2}v_j w_j s^2 + \mathcal{O}(s^3) & \text{if } w_j \neq 0 \neq v_j \\ -\frac{1}{2}|v_j + w_j|s^3 + \mathcal{O}(s^4) & \text{if } w_j \neq 0 = v_j \text{ or } w_j = 0 \neq v_j \\ -\frac{1}{2}s^4 + \mathcal{O}(s^5) & \text{if } w_j = 0 = v_j, \end{cases} \end{aligned} \tag{2.28}$$

which is strictly negative for small  $s > 0$ . Hence,  $\phi$  is a saddle point.

Lastly, assume that the  $j$ th hidden neuron is semi-inactive with  $I_j = \{1\}$  and  $v_j \geq 0$  (case one) or that it is degenerate with  $v_j > 0$  (case two). This is dealt with the same way as the previous step. Let  $\phi_s \in \mathbb{R}^{3N+1}$ ,  $s \in [0, 1]$ , be given by  $w_j^s = w_j + s$ ,  $b_j^s = -(1 - s)w_j^s$ , and  $v_j^s = v_j + s$ . This time, we have  $w_j^s > 0$  and  $a^s = v_j^s w_j^s > 0$  for all  $s \in (0, 1]$  in both cases. The realization of  $\phi_s$  on  $[0, 1]$  is given for all  $s, x \in [0, 1]$  by

$$f_{\phi_s}(x) = c + v_j^s \max\{w_j^s x + b_j^s, 0\} = \frac{1}{2} + a^s(x - 1 + s)\mathbb{1}_{[1-s,1]}(x). \tag{2.29}$$

Essentially by the same computation as in the previous step,

$$\begin{aligned} \mathcal{L}(\phi_s) - \mathcal{L}(\phi) &= \frac{1}{3}a^s(a^s + 1)s^3 - \frac{1}{2}a^s s^2 \\ &= \begin{cases} -\frac{1}{2}v_j w_j s^2 + \mathcal{O}(s^3) & \text{if } w_j \neq 0 \neq v_j \\ -\frac{1}{2}(v_j + w_j)s^3 + \mathcal{O}(s^4) & \text{if } w_j \neq 0 = v_j \text{ or } w_j = 0 \neq v_j \\ -\frac{1}{2}s^4 + \mathcal{O}(s^5) & \text{if } w_j = 0 = v_j, \end{cases} \end{aligned} \tag{2.30}$$

from which we conclude that  $\phi$  is a saddle point. □

This finishes the treatment of the affine case, and we now tend to the more involved non-affine case in the next section.

### 2.7 Critical Points of the Loss Function with Non-affine Realization

The following lemma is the main tool for this section. It generalizes the argument (2.9) that we presented in Sect. 2.4; see Lemma 2.20.(vi) below. This lemma captures the combinatorics of piecewise affine functions satisfying conditions of the form (2.8).

**Lemma 2.20** *Let  $n \in \mathbb{N}_0$ ,  $A_0, \dots, A_n, B_0, \dots, B_n, q_0, \dots, q_{n+1} \in \mathbb{R}$  satisfy  $q_0 < \dots < q_{n+1}$ , and consider a function  $f \in C([q_0, q_{n+1}], \mathbb{R})$  satisfying for all  $i \in \{0, \dots, n\}$ ,  $x \in [q_i, q_{i+1}]$  that  $f(x) = A_i x + B_i$  and  $\int_{q_i}^{q_{i+1}} (f(y) - y) dy = 0$ . Then,*

(i) *we have for all  $i \in \{0, \dots, n\}$  that*

$$\begin{aligned} A_i - 1 &= (-1)^i \frac{q_1 - q_0}{q_{i+1} - q_i} (A_0 - 1), \\ B_i &= (-1)^{i+1} \frac{q_{i+1} + q_i}{2} \frac{q_1 - q_0}{q_{i+1} - q_i} (A_0 - 1), \end{aligned} \tag{2.31}$$

(ii) *we have  $f = \text{id}_{[q_0, q_{n+1}]}$   $\iff \forall i \in \{0, \dots, n\}: A_i = 1, B_i = 0$*

$$\begin{aligned} &\iff \exists i \in \{0, \dots, n\}: A_i = 1, B_i = 0 \\ &\iff \exists i \in \{0, \dots, n\}: f|_{[q_i, q_{i+1}]} = \text{id}_{[q_i, q_{i+1}]}, \end{aligned}$$

(iii) *for all  $i \in \{0, \dots, n\}$  we have  $\text{sign}(A_i - 1) = (-1)^i \text{sign}(A_0 - 1)$ .*

*If, in addition,  $0 = \int_{q_0}^{q_{n+1}} x(f(x) - x) dx$ , then*

- (iv) *we have  $0 = (A_0 - 1) \sum_{i=0}^n (-1)^i (q_{i+1} - q_i)^2$ ,*
- (v) *if  $f \neq \text{id}_{[q_0, q_{n+1}]}$ , then  $0 = \sum_{i=0}^n (-1)^{i+1} (q_{i+1} - q_i)^2$ ,*
- (vi) *if  $n = 0$ , then  $f = \text{id}_{[q_0, q_1]}$ .*

**Proof** First note that we must have  $A_i q_{i+1} + B_i = A_{i+1} q_{i+1} + B_{i+1}$  for all  $i \in \{0, \dots, n - 1\}$ . Moreover, the assumption  $0 = \int_{q_i}^{q_{i+1}} (f(x) - x) dx$  is equivalent to  $B_i = -\frac{1}{2}(q_{i+1} + q_i)(A_i - 1)$ . Combining these yields

$$A_{i+1} - 1 = -\frac{q_{i+1} - q_i}{q_{i+2} - q_{i+1}} (A_i - 1) \tag{2.32}$$

for all  $i \in \{0, \dots, n - 1\}$ . Induction then proves the formula for  $A_i - 1$ , and the formula for  $B_i$  follows. Lastly, by plugging the formulas for  $A_i$  and  $B_i$  into  $f(x)$ , we compute

$$\begin{aligned} \int_{q_0}^{q_{n+1}} x(f(x) - x) dx &= \sum_{i=0}^n \int_{q_i}^{q_{i+1}} x((A_i - 1)x + B_i) dx \\ &= \frac{q_1 - q_0}{12} (A_0 - 1) \sum_{i=0}^n (-1)^i (q_{i+1} - q_i)^2. \end{aligned} \tag{2.33}$$

The remaining items follow immediately. □

In order to apply this lemma later on, let us verify that our network always satisfies the condition  $\int_{q_i}^{q_{i+1}} (f(y) - y) dx = 0$  for suitable choices of  $q_i$  and  $q_{i+1}$ .

**Lemma 2.21** *Suppose  $\phi \in \mathbb{R}^{3N+1}$  is a critical point or a local extremum of  $\mathcal{L}$  and denote by  $0 = q_0 < q_1 < \dots < q_n < q_{n+1} = 1$ , for  $n \in \mathbb{N}_0$ , the roughest partition such that  $f_\phi$  is affine on all subintervals  $[q_i, q_{i+1}]$ . Then, we have for all  $i \in \{0, \dots, n\}$  that*

$$\int_{q_i}^{q_{i+1}} (f_\phi(x) - x)dx = 0. \tag{2.34}$$

**Proof** First, note that  $\phi$  must have a non-flat type-2-active neuron whose breakpoint is  $q_i$ , for all  $i \in \{1, \dots, n\}$ . From the fourth line of (2.7), we know that  $\int_0^1 (f_\phi(x) - x)dx = 0$ . This and the second line of (2.7) imply, for any non-flat type-2-active neuron  $j$ ,

$$\int_{I_j} (f_\phi(x) - x)dx = 0 = \int_{[0,1] \setminus I_j} (f_\phi(x) - x)dx. \tag{2.35}$$

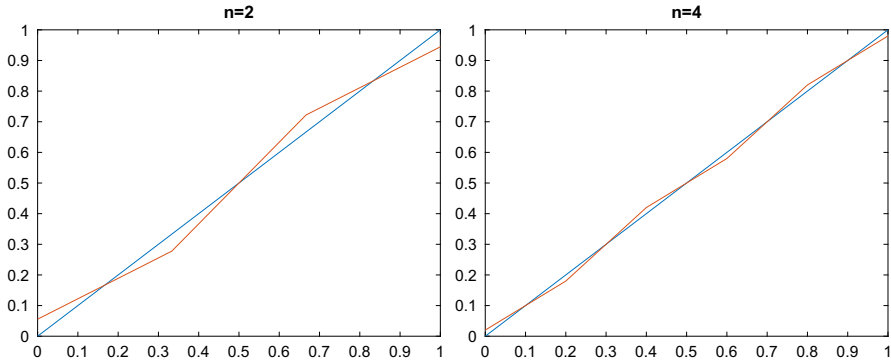
Since either  $I_j = [0, t_j]$  or  $[0, 1] \setminus I_j = [0, t_j]$ , it follows that  $\int_0^{q_i} (f_\phi(x) - x)dx = 0$ , for all  $i \in \{0, \dots, n + 1\}$ . Taking differences of these integrals yields the desired statement. □

Next, as a first application of Lemma 2.20, we prove that only global minima can have type-1-active or non-flat semi-active neurons. We already established this in Lemma 2.18 in the affine case, but now we extend it to the non-affine case. The statement from Lemma 2.18 about saddle points not having non-flat degenerate neurons also holds in the non-affine case, but we will not see this until later in Sect. 2.8.

**Lemma 2.22** *Suppose  $\phi \in \mathbb{R}^{3N+1}$  is a critical point or a local extremum of  $\mathcal{L}$  but not a global minimum. Then,  $\phi$  does not have any type-1-active or non-flat semi-active neurons.*

**Proof** For affine  $f_\phi$ , the result has been established in Lemma 2.18. Thus, suppose  $f_\phi$  is not affine on  $[0, 1]$  and that  $\phi$  has a type-1-active or non-flat semi-active neuron. Denote by  $0 = q_0 < q_1 < \dots < q_n < q_{n+1} = 1$ , for  $n \in \mathbb{N}$ , the roughest partition such that  $f_\phi$  is affine on all subintervals  $[q_i, q_{i+1}]$ . We know from Lemma 2.21 that  $\int_{q_0}^{q_1} (f_\phi(x) - x)dx = 0$ , and we claim that also  $\int_{q_0}^{q_1} x(f_\phi(x) - x)dx = 0$ . To prove this, note that  $\phi$  must have at least one non-flat type-2-active neuron (without loss of generality the first) with breakpoint  $-b_1/w_1 = q_1$ . Moreover, (2.8) shows that  $0 = \int_0^1 x(f_\phi(x) - x)dx$  if applied with the type-1-active or non-flat semi-active neuron. Using this and  $\frac{\partial}{\partial w_1} \mathcal{L}(\phi) = 0$ , one deduces the claim as in the proof of Lemma 2.21. Hence, we conclude  $f_\phi|_{[q_0, q_1]} = \text{id}_{[q_0, q_1]}$  with the argument (2.9). But then we also get  $f_\phi = \text{id}_{[q_0, q_{n+1}]}$  by Lemma 2.20.(ii) and Lemma 2.21, yielding a contradiction. □

We now turn to the proof of Theorem 2.4.(IV.b). More precisely, we show that critical points and local extrema whose realizations are not affine must take a very specific form. The only degree of freedom of their realization functions is a single parameter varying over the set of even integers in  $\{1, \dots, N\}$ . Examples of the possible



**Fig. 2** Examples of the network realizations (red) in Lemma 2.23 for the cases  $n = 2$  and  $n = 4$ . The blue line is the target function (identity function) (Color figure online)

realizations are shown in Fig. 2, which illustrates that the degree of freedom is reflected by the number of breakpoints. Once this number is fixed, the shape of the function is uniquely determined: the breakpoints are equally spaced in the interval  $[0, 1]$ , and the slope of the realization on each affine segment alternates between two given values in such a way that the function symmetrically oscillates around the diagonal. In addition, we deduce in Lemma 2.23 that critical points and local extrema can realize these functions only in a very specific way, limited by few combinatorial choices.

**Lemma 2.23** *Suppose  $\phi \in \mathbb{R}^{3N+1}$  is a critical point or a local extremum of  $\mathcal{L}$  but not a global minimum and that  $f_\phi$  is not affine on  $[0, 1]$ . Denote by  $0 = q_0 < q_1 < \dots < q_n < q_{n+1} = 1$ , for  $n \in \mathbb{N}$ , the roughest partition such that  $f_\phi$  is affine on all subintervals  $[q_i, q_{i+1}]$ , and denote by  $K_i \subseteq \{1, \dots, N\}$  the set of all type-2-active neurons of  $\phi$  whose breakpoint is  $q_i$ . Then, the following hold:*

- (i)  $n$  is even,
- (ii)  $q_i = \frac{i}{n+1}$  for all  $i \in \{1, \dots, n\}$ ,
- (iii)  $-b_j/w_j \in \{q_1, \dots, q_n\}$  for all type-2-active neurons  $j \in \{1, \dots, N\}$  of  $\phi$ ,
- (iv)  $\text{sign}(w_j) = (-1)^{i+1}$  for all  $i \in \{1, \dots, n\}$ ,  $j \in K_i$ ,
- (v)  $\sum_{j \in K_i} v_j w_j = 2/(n + 1)$  for all  $i \in \{1, \dots, n\}$ ,
- (vi)  $\phi$  is centered,
- (vii)  $f_\phi(x) = x - \frac{(-1)^i}{n+1} (x - \frac{i+1/2}{n+1})$  for all  $i \in \{0, \dots, n\}$ ,  $x \in [q_i, q_{i+1}]$ .

The proof of this lemma requires a successive application of Lemma 2.20. We prove the statements of the lemma in a different order than stated. First of all, Lemma 2.20. (ii) will enforce the correct sign for each  $w_j$ ,  $j \in K_i$ . That  $n$  is even will be a consequence of these signs. It will also follow from the signs together with Lemma 2.20. (v) that  $q_i = \frac{i}{n+1}$ . Afterward, we use the formulas (2.31) from Lemma 2.20 to verify that any type-2-active neuron must have as breakpoint one of  $q_1, \dots, q_n$ . Once this has been shown, we obtain a more explicit version of those formulas and deduce  $\sum_{k \in K_i} v_k w_k = 2/(n + 1)$ . That  $f_\phi$  takes exactly the form in Lemma 2.23. (vii) is a by-product of the last derivation, and that  $\phi$  is centered is shown last.

**Proof of Lemma 2.23** We begin by noting that none of the sets  $K_i, i \in \{1, \dots, n\}$ , can be empty. Furthermore, the third equation of (2.7) and Lemma 2.21 imply that (2.8) holds for all neurons in  $\bigcup_i K_i$  even if they are flat. Applying Lemma 2.20.(ii), which we can do by Lemma 2.21, ensures that  $f_\phi|_{[q_i, q_{i+1}]} \neq \text{id}_{[q_i, q_{i+1}]}$  for all  $i \in \{0, \dots, n\}$ . In particular, (2.8) and the argument (2.9) show for all  $i \in \{1, \dots, n - 1\}$  and  $j_0 \in K_i, j_1 \in K_{i+1}$  that  $\text{sign}(w_{j_0}) \neq \text{sign}(w_{j_1})$  for otherwise we would have  $I_{j_0} \setminus I_{j_1} = [q_i, q_{i+1}]$  or  $I_{j_1} \setminus I_{j_0} = [q_i, q_{i+1}]$  (depending on the sign) and, hence,

$$\int_{q_i}^{q_{i+1}} (f_\phi(x) - x)dx = 0 = \int_{q_i}^{q_{i+1}} x(f_\phi(x) - x)dx. \tag{2.36}$$

Likewise, we must have  $\int_0^{q_1} x(f_\phi(x) - x)dx \neq 0$  and, hence,  $w_j > 0$  for any  $j \in K_1$ . Combining the previous two arguments establishes  $\text{sign}(w_j) = (-1)^{i+1}$  for any  $i \in \{1, \dots, n\}, j \in K_i$ . Just like  $w_j > 0$  for any  $j \in K_1$ , we must also have  $w_j < 0$  for any  $j \in K_n$ . Thus,  $-1 = \text{sign}(w_j) = (-1)^{n+1}$  for all  $j \in K_n$ , so  $n$  is even. Now that we know the sign of each parameter  $w_j$  for neurons  $j \in \bigcup_i K_i$ , we can use (2.8) again to find that  $\int_{q_i}^{q_{i+2}} x(f_\phi(x) - x)dx = 0$  for all  $i \in \{0, \dots, n - 1\}$ . Then, Lemma 2.20.(v) (with the partition  $q_i, q_{i+1}, q_{i+2}$ ) tells us

$$0 = (q_{i+2} - q_{i+1})^2 - (q_{i+1} - q_i)^2. \tag{2.37}$$

This can only hold for all  $i \in \{0, \dots, n - 1\}$  if the points  $q_1, \dots, q_n$  are equidistributed, which means  $q_i = i/(n + 1)$ . Next, if we denote  $f_\phi(x) = A_i x + B_i$  on  $[q_i, q_{i+1}]$ , then the formulas (2.31) must hold for all  $i \in \{0, \dots, n\}$ . Since  $q_1, \dots, q_n$  are equidistributed, the formulas simplify to

$$A_i - 1 = (-1)^i(A_0 - 1) \quad \text{and} \quad B_i = (-1)^{i+1} \frac{i + \frac{1}{2}}{n + 1}(A_0 - 1) \tag{2.38}$$

for all  $i \in \{0, \dots, n\}$ . Using (2.38), one can verify that any type-2-active neuron of  $\phi$  must have as breakpoint one of the points  $q_1, \dots, q_n$ . If this were not the case, say the  $j$ th hidden neuron were type-2-active with breakpoint  $t_j = -b_j/w_j$ , then one could choose  $i \in \{0, \dots, n\}$  such that  $q_i < t_j < q_{i+1}$ . Using (2.8), (2.38), and Lemma 2.21, the integral from the third line of (2.7) reads (after dividing by  $2w_j$ )

$$\begin{aligned} & \int_{I_j} (x - t_j)(f_\phi(x) - x)dx \\ &= \int_{[q_i, q_{i+1}] \cap I_j} (x - t_j)(f_\phi(x) - x)dx - \begin{cases} 0 & \text{if } i \text{ is even} \\ \int_{q_i}^{q_{i+1}} x(f_\phi(x) - x)dx & \text{if } i \text{ is odd} \end{cases} \\ &= \begin{cases} \frac{1}{6}(A_0 - 1)(t_j - q_i)^2(q_{i+1} - t_j + \frac{1}{2(n+1)}) & \text{if } I_j = [0, t_j] \text{ and } i \text{ is even} \\ \text{or if } I_j = [t_j, 1] \text{ and } i \text{ is odd} \\ \frac{1}{6}(A_0 - 1)(q_{i+1} - t_j)^2(t_j - q_i + \frac{1}{2(n+1)}) & \text{if } I_j = [0, t_j] \text{ and } i \text{ is odd} \\ \text{or if } I_j = [t_j, 1] \text{ and } i \text{ is even.} \end{cases} \end{aligned} \tag{2.39}$$

So, the partial derivative of  $\mathcal{L}$  with respect to  $v_j$  does not vanish, yielding a contradiction. This proves that all type-2-active neurons lie in  $\bigcup_i K_i$ . In particular, we can write

$$A_l = \sum_{\substack{i=1 \\ i \text{ odd}}}^l \sum_{j \in K_i} v_j w_j + \sum_{\substack{i=l+1 \\ i \text{ even}}}^n \sum_{j \in K_i} v_j w_j \tag{2.40}$$

for all  $l \in \{0, \dots, n\}$  because  $\phi$  does not have any type-1-active neurons by Lemma 2.22. We can combine this formula with (2.38) to find for all  $i \in \{0, \dots, n - 1\}$

$$\begin{aligned} -(A_0 - 1) &= (-1)^i (A_{i+1} - 1) = (-1)^i (A_i - 1) + \sum_{j \in K_{i+1}} v_j w_j \\ &= A_0 - 1 + \sum_{j \in K_{i+1}} v_j w_j. \end{aligned} \tag{2.41}$$

Thus, the quantity  $a := \sum_{j \in K_i} v_j w_j$  is independent of  $i \in \{1, \dots, n\}$ . Consequently, we obtain  $A_i = an/2$  for even  $i$  (including  $i = 0$ ) and  $A_i = a(1 + n/2)$  for odd  $i$ . The identity  $A_1 - 1 = 1 - A_0$  then forces  $a = 2/(n + 1)$ . That  $\phi$  has to be centered follows from  $f_\phi(0) = B_0$ .  $\square$

As our final building block for the proof of Theorem 2.4, we show that the networks from Lemma 2.23 are saddle points of the loss function. To achieve this, we will find a set of coordinates in which  $\mathcal{L}$  is twice differentiable and calculate the determinant of the Hessian of  $\mathcal{L}$  restricted to these coordinates. It will turn out to be strictly negative, from which it follows that we deal with a saddle point.

**Lemma 2.24** *Suppose  $\phi \in \mathbb{R}^{3N+1}$  is a critical point or a local extremum of  $\mathcal{L}$  but not a global minimum and that  $f_\phi$  is not affine on  $[0, 1]$ . Then,  $\phi$  is a saddle point of  $\mathcal{L}$ .*

**Proof** Take  $n \in \mathbb{N}$  satisfying the assumptions of Lemma 2.23 and let  $K_1 \subseteq \{1, \dots, N\}$  denote the set of those type-2-active neurons with breakpoint  $1/(n + 1)$ . Denote by  $K_1^- \subseteq K_1$  the set of all those hidden neurons  $j \in K_1$  with  $v_j < 0$ . It may happen that  $K_1^-$  is empty. However, the complement  $K_1 \setminus K_1^-$  is never empty since  $\sum_{j \in K_1} v_j w_j = 2/(n + 1)$  and  $\text{sign}(w_j) = 1$  for all  $j \in K_1$  by Lemma 2.23. Let  $j_1 \in K_1$  be any hidden neuron with  $v_{j_1} > 0$  and denote by  $j_2, \dots, j_l$ , for  $l \in \{1, \dots, N\}$ , an enumeration of  $K_1^-$ . Moreover, let  $k \in \{1, \dots, N\}$  be any type-2-active neuron with breakpoint  $t_k = 2/(n + 1)$ .

We know from Lemma 2.15 that  $\mathcal{L}$  is twice continuously differentiable in the coordinates of type-2-active neurons and in  $(v, c)$ . We will show that the Hessian  $H$  of  $\mathcal{L}$  restricted to  $(b_{j_1}, \dots, b_{j_l}, v_k, c)$  has a strictly negative determinant.

In order to compute this determinant, we introduce some shorthand notation. For  $i \in \{1, \dots, l\}$ , denote  $\lambda_i = \frac{n+1}{2} v_{j_i} w_{j_i}$  so that  $\sum_{i=1}^l \lambda_i \leq 1$  by the choice of neurons in the collection  $\{j_1, \dots, j_l\}$ . Define  $\mu = \frac{n+1}{2n}$  and the vectors  $u_1 = (v_{j_1}, \dots, v_{j_l})$ ,  $u_2 = (\frac{-1}{4n^2\mu} w_k, 1)$ , and  $u = (u_1, u_2)$ . Furthermore, let  $D$  be the diagonal matrix with entries  $-v_{j_i}^2 / (4\lambda_i n)$ ,  $i \in \{1, \dots, l\}$ , let  $A$  be the Hessian of  $\mathcal{L}$  restricted to  $(v_k, c)$ , let

$B = \mu A - u_2 u_2^T$ , and let  $E$  be the diagonal block matrix with blocks  $D$  and  $B$ . Then,  $H = \frac{1}{\mu}(E + uu^T)$  and, hence,

$$\det(H) = \mu^{-(l+2)}(1 + u^T E^{-1}u) \det(E) \tag{2.42}$$

once we verified that  $E$  is invertible. We calculate directly

$$\det(A) = \det \left( \begin{array}{cc} \frac{2}{3(n\mu)^3} w_k^2 & \frac{-1}{(n\mu)^2} w_k \\ \frac{-1}{(n\mu)^2} w_k & \frac{n+1}{n\mu} \end{array} \right) = \frac{2n-1}{3(n\mu)^4} w_k^2 > 0. \tag{2.43}$$

Next, we compute

$$\Gamma := \frac{1}{\mu} u_2^T A^{-1} u_2 = \frac{32n^2 - 21n + 3}{16n(2n-1)} \in (0, 1). \tag{2.44}$$

Using  $\Gamma$ , we obtain  $\det(B) = \mu^2(1 - \Gamma) \det(A) > 0$  and  $B^{-1} = \frac{1}{\mu} A^{-1} + \frac{1}{\mu^2(1-\Gamma)} A^{-1} u_2 u_2^T A^{-1}$ . In particular,  $E$  is invertible. Using  $u_2^T B^{-1} u_2 = \frac{\Gamma}{1-\Gamma}$ , we can write

$$u^T E^{-1} u = u_1^T D^{-1} u_1 + u_2^T B^{-1} u_2 = -4n \sum_{i=1}^l \lambda_i + \frac{\Gamma}{1-\Gamma}. \tag{2.45}$$

The determinant of  $D$  is  $-(4n)^{-l} \prod_{i=1}^l v_{j_i}^2 |\lambda_i|^{-1} < 0$  so that

$$\Delta := -\mu^{-(l+2)}(1 - \Gamma)^{-1} \det(D) \det(B) \tag{2.46}$$

is strictly positive. Summing up, we obtain that the determinant of  $H$  is

$$\det(H) = \Delta \left( 4n(1 - \Gamma) \sum_{i=1}^l \lambda_i - 1 \right). \tag{2.47}$$

We already mentioned that  $\sum_{i=1}^l \lambda_i \leq 1$ . Finally, we compute  $4n(1 - \Gamma) = \frac{5n-3}{8n-4} < 1$  to conclude  $\det(H) < 0$ , which finishes the proof.  $\square$

We now have constructed all the tools needed to prove Theorem 2.4 in the special case in which the target function is the identity on  $[0, 1]$ . This will be done in the next section.

### 2.8 Classification of the Critical Points if the Target Function is the Identity

In this section, we gather the results of the previous two sections to prove the main theorem in the case where the target function is the identity on  $[0, 1]$ .

**Proposition 2.25** *Let  $\phi = (w, b, v, c) \in \mathbb{R}^{3N+1}$ . Then, the following hold:*

- (I)  $\phi$  is not a local maximum of  $\mathcal{L}$ .
- (II) If  $\phi$  is a critical point or a local extremum of  $\mathcal{L}$ , then  $\mathcal{L}$  is differentiable at  $\phi$  with gradient  $\nabla\mathcal{L}(\phi) = 0$ .
- (III)  $\phi$  is a non-global local minimum of  $\mathcal{L}$  if and only if  $\phi$  is centered and, for all  $j \in \{1, \dots, N\}$ , the  $j$ th hidden neuron of  $\phi$  is
  - (a) inactive,
  - (b) semi-inactive with  $I_j = \{0\}$  and  $v_j > 0$ , or
  - (c) semi-inactive with  $I_j = \{1\}$  and  $v_j < 0$ .
- (IV)  $\phi$  is a saddle point of  $\mathcal{L}$  if and only if  $\phi$  is centered,  $\phi$  does not have any type-1-active neurons,  $\phi$  does not have any non-flat semi-active neurons,  $\phi$  does not have any non-flat degenerate neurons, and exactly one of the following two items holds:
  - (a)  $\phi$  does not have any type-2-active neurons and there exists  $j \in \{1, \dots, N\}$  such that the  $j$ th hidden neuron of  $\phi$  is
    - (i) flat semi-active,
    - (ii) semi-inactive with  $I_j = \{0\}$  and  $v_j \leq 0$ ,
    - (iii) semi-inactive with  $I_j = \{1\}$  and  $v_j \geq 0$ , or
    - (iv) flat degenerate.
  - (b) There exists  $n \in \{2, 4, 6, \dots\}$  such that  $(\bigcup_{j \in \{1, \dots, N\}, w_j \neq 0} \{-\frac{b_j}{w_j}\}) \cap (0, 1) = \bigcup_{i=1}^n \{\frac{i}{n+1}\}$  and, for all  $j \in \{1, \dots, N\}$ ,  $i \in \{1, \dots, n\}$  with  $w_j \neq 0 = b_j + \frac{iw_j}{n+1}$ , it holds that  $\text{sign}(w_j) = (-1)^{i+1}$  and  $\sum_{k \in \{1, \dots, N\}, w_k \neq 0 = b_k + \frac{iw_k}{n+1}} v_k w_k = \frac{2}{n+1}$ .
- (V) If  $\phi$  is a non-global local minimum of  $\mathcal{L}$  or a saddle point of  $\mathcal{L}$  without type-2-active neurons, then  $f_\phi(x) = 1/2$  for all  $x \in [0, 1]$ .
- (VI) If  $\phi$  is a saddle point of  $\mathcal{L}$  with at least one type-2-active neuron, then there exists  $n \in \{2, 4, 6, \dots\}$  such that  $n \leq N$  and, for all  $i \in \{0, \dots, n\}$ ,  $x \in [\frac{i}{n+1}, \frac{i+1}{n+1}]$ , one has

$$f_\phi(x) = x - \frac{(-1)^i}{n+1} \left( x - \frac{i + \frac{1}{2}}{n+1} \right). \tag{2.48}$$

**Proof** Statement (I) follows from Lemma 2.8 and the ‘if’ part of the ‘if and only if’ statement in (III) is the content of Lemma 2.17. Moreover, if  $\phi$  is as in (IV.a), then it is a critical point because it satisfies (2.7) and it is a saddle point by Lemma 2.19. Next, denote  $q_i = i/(n + 1)$  for all  $i \in \{0, \dots, n + 1\}$ . If  $\phi$  is as in (IV.b), then its realization on  $[0, 1]$  is given by

$$f_\phi(x) = \frac{1}{2} + \frac{2}{n+1} \sum_{i=1}^n (-1)^{i+1} \max\{(-1)^{i+1}(x - q_i), 0\}. \tag{2.49}$$

which coincides with the formula (2.48). In particular, we have  $\int_{q_i}^{q_{i+1}} (f_\phi(x) - x) dx = 0$  for all  $i \in \{0, \dots, n\}$  and  $\int_{q_i}^{q_{i+2}} x(f_\phi(x) - x) dx = 0$  for all  $i \in \{0, \dots, n - 1\}$ . The



latter asserts that  $\int_{q_i}^1 x(f_\phi(x) - x)dx = 0$  for odd  $i$  and  $\int_0^{q_i} x(f_\phi(x) - x)dx = 0$  for even  $i$ . Thus,  $\phi$  satisfies (2.7) and, hence, is a critical point. Furthermore, it is a saddle point by Lemma 2.24. This proves the ‘if’ part of the ‘if and only if’ statement in (IV).

Now, suppose  $\phi$  is a non-global local minimum. Then,  $f_\phi$  is affine by Lemma 2.24. Lemma 2.18 asserts that  $\phi$  is centered and does not have any active or non-flat semi-active neurons. Furthermore, for each hidden neuron, Lemma 2.19 rules out all possibilities except (III.a)–(III.c). This proves the ‘only if’ part of (III).

Next, suppose  $\phi$  is a saddle point. If  $f_\phi$  is affine, then  $\phi$  is centered and does not have any active, non-flat semi-active, or non-flat degenerate neurons by Lemma 2.18. If there is no hidden neuron as in (IV.a.i)–(IV.a.iv), then all hidden neurons satisfy one of the conditions in (III.a)–(III.c). But this contradicts Lemma 2.17. This proves (IV.a). If  $f_\phi$  is not affine, then it still does not admit any type-1-active or non-flat semi-active neurons by Lemma 2.22. Moreover, Lemma 2.23 shows that  $\phi$  is centered and its type-2-active neurons satisfy (IV.b). We need to argue that  $\phi$  does not have any non-flat degenerate neurons in this case either. If there were a non-flat degenerate neuron, then  $\mathcal{G}(\phi) = 0$  implies  $0 = \int_0^1 x(f_\phi(x) - x)dx$ . But Lemma 2.20.(v) and Lemma 2.23 ensure that this integral is different from zero. This finishes the proof of the ‘only if’ part of (IV).

Next, we prove (II). If  $\phi$  is a saddle point, then it does not have any non-flat degenerate neurons by (IV). If  $\phi$  is a non-global local extremum, then (I) and (III) imply that  $\phi$  does not have any non-flat degenerate neurons either. Thus,  $\mathcal{L}$  is differentiable at  $\phi$  by Lemma 2.14. If  $\phi$  is a global minimum, then  $\phi$  is point of differentiability by Lemma 2.11.

Statement (V) follows immediately from (III) and (IV.a). The remaining statement (VI) is implied by (IV.b) and (2.49). □

### 2.9 Completion of the Proof of Theorem 2.4

In this section, we show that Theorem 2.4 can always be reduced to its special case, Proposition 2.25, by employing a transformation of the parameter space.

**Proof of Theorem 2.4** First, we assume that  $T = (0, 1)$ . Consider the transformation  $P: \mathbb{R}^{3N+1} \rightarrow \mathbb{R}^{3N+1}$  of the parameter space given by  $P(w, b, v, c) = (w, b, \frac{v}{\alpha}, \frac{c-\beta}{\alpha})$ . We then have  $\mathcal{L}_{N,T,\mathcal{A}}(\phi) = \alpha^2 \mathcal{L} \circ P(\phi)$  for all  $\phi \in \mathbb{R}^{3N+1}$ . Since the coordinates  $w$  and  $b$  remain unchanged and the vector  $v$  only gets scaled under the transformation  $P$ , the transformation  $P$  does not change the types of the hidden neurons. Moreover, a network  $\phi \in \mathbb{R}^{3N+1}$  is  $(T, \mathcal{A})$ -centered if and only if  $P(\phi)$  is centered. The map  $P$  clearly is a smooth diffeomorphism and, hence, Theorem 2.4 with  $T = (0, 1)$  is exactly what we obtain from Proposition 2.25 under the transformation  $P$ .

Now, we deduce Theorem 2.4 for general  $T$ . This time, set  $\mathcal{B} = (\alpha(T_1 - T_0), \alpha T_0 + \beta)$  and denote by  $Q: \mathbb{R}^{3N+1} \rightarrow \mathbb{R}^{3N+1}$  the transformation  $Q(w, b, v, c) = ((T_1 - T_0)w, T_0w + b, v, c)$ . Then,  $\mathcal{L}_{N,T,\mathcal{A}}(\phi) = (T_1 - T_0)\mathcal{L}_{N,(0,1),\mathcal{B}} \circ Q(\phi)$  for any  $\phi \in \mathbb{R}^{3N+1}$ . As above, the transformation  $Q$  does not change the types of the hidden

neurons. Note for the breakpoints that

$$-\frac{b_j}{w_j} = T_0 + \frac{i(T_1 - T_0)}{n + 1} \iff -\frac{T_0 w_j + b_j}{(T_1 - T_0) w_j} = \frac{i}{n + 1}. \tag{2.50}$$

Also,  $\phi \in \mathbb{R}^{3N+1}$  is  $(T, \mathcal{A})$ -centered if and only if  $\mathcal{Q}(\phi)$  is  $((0, 1), \mathcal{B})$ -centered. Since we have shown the theorem to hold for  $T = (0, 1)$ , the smooth diffeomorphism  $\mathcal{Q}$  yields Theorem 2.4 in the general case.  $\square$

### 3 From ReLU to Leaky ReLU

In this section, we attempt to derive Theorem 2.4 for leaky ReLU activation, given by  $x \mapsto \max\{x, \gamma x\}$  for a parameter  $\gamma \in (0, 1)$ . We denote the realization  $f_\phi^\gamma \in C(\mathbb{R}, \mathbb{R})$  of a network  $\phi = (w, b, v, c) \in \mathbb{R}^{3N+1}$  with this activation by

$$f_\phi^\gamma(x) = c + \sum_{j=1}^N v_j \max\{w_j x + b_j, \gamma(w_j x + b_j)\}. \tag{3.1}$$

Analogously to the ReLU case, given  $\mathcal{A} = (\alpha, \beta) \in \mathbb{R}^2$  and  $T = (T_0, T_1) \in \mathbb{R}^2$ , the loss function  $\mathcal{L}_{N,T,\mathcal{A}}^\gamma \in C(\mathbb{R}^{3N+1}, \mathbb{R})$  is the  $L^2$ -loss given by

$$\mathcal{L}_{N,T,\mathcal{A}}^\gamma(\phi) = \int_{T_0}^{T_1} (f_\phi^\gamma(x) - \alpha x - \beta)^2 dx. \tag{3.2}$$

Again, we call a point a critical point of  $\mathcal{L}_{N,T,\mathcal{A}}^\gamma$  if it is a zero of the generalized gradient defined by right-hand partial derivatives. The notions about types of neurons remain the same as in Definition 2.3. Strictly speaking, the notions ‘inactive’ and ‘semi-inactive’ are no longer suitable for leaky ReLU activation, but it is convenient to stick to the same terminology. We will deduce the classification for leaky ReLU by reducing it to the ReLU case in some instances and deal with other instances directly.

#### 3.1 Partial Reduction to the ReLU Case

As before, we first consider the special case where the target function is the identity on  $[0, 1]$ . Let us abbreviate  $\mathcal{L}^\gamma = \mathcal{L}_{N,(0,1),(1,0)}^\gamma$  and  $\mathcal{L} = \mathcal{L}_{2N,(0,1),(1,0)}$ . Let  $P: \mathbb{R}^{3N+1} \rightarrow \mathbb{R}^{6N+1}$  denote the smooth map  $P(w, b, v, c) = (w, -w, b, -b, v, -\gamma v, c)$ . Then,  $f_\phi^\gamma = f_{P(\phi)}$  and  $\mathcal{L}^\gamma = \mathcal{L} \circ P$ . Hence, if  $\mathcal{L}$  is differentiable at  $P(\phi)$ , then  $\mathcal{L}^\gamma$  is differentiable at  $\phi$ , so differentiability properties of  $\mathcal{L}$  convert to  $\mathcal{L}^\gamma$ . The partial derivatives of  $\mathcal{L}^\gamma$  at any network  $\phi$  and any non-degenerate or flat degenerate neuron

$j$  are given by

$$\begin{aligned}
 \frac{\partial}{\partial w_j} \mathcal{L}^\gamma(\phi) &= \left(\frac{\partial}{\partial w_j} \mathcal{L}\right)(P(\phi)) - \left(\frac{\partial}{\partial w_{j+N}} \mathcal{L}\right)(P(\phi)), \\
 \frac{\partial}{\partial b_j} \mathcal{L}^\gamma(\phi) &= \left(\frac{\partial}{\partial b_j} \mathcal{L}\right)(P(\phi)) - \left(\frac{\partial}{\partial b_{j+N}} \mathcal{L}\right)(P(\phi)), \\
 \frac{\partial}{\partial v_j} \mathcal{L}^\gamma(\phi) &= \left(\frac{\partial}{\partial v_j} \mathcal{L}\right)(P(\phi)) - \gamma \left(\frac{\partial}{\partial v_{j+N}} \mathcal{L}\right)(P(\phi)), \\
 \frac{\partial}{\partial c} \mathcal{L}^\gamma(\phi) &= \left(\frac{\partial}{\partial c} \mathcal{L}\right)(P(\phi)).
 \end{aligned}
 \tag{3.3}$$

We can also write these in explicit formulas. To do so, we complement the notation  $I_j$  by the intervals  $\hat{I}_j = \{x \in [0, 1] : w_j x + b_j < 0\} = [0, 1] \setminus I_j$ . Then,

$$\begin{aligned}
 \frac{\partial}{\partial w_j} \mathcal{L}^\gamma(\phi) &= 2v_j \int_{I_j} x(f_\phi^\gamma(x) - x)dx + 2\gamma v_j \int_{\hat{I}_j} x(f_\phi^\gamma(x) - x)dx, \\
 \frac{\partial}{\partial b_j} \mathcal{L}^\gamma(\phi) &= 2v_j \int_{I_j} (f_\phi^\gamma(x) - x)dx + 2\gamma v_j \int_{\hat{I}_j} (f_\phi^\gamma(x) - x)dx, \\
 \frac{\partial}{\partial v_j} \mathcal{L}^\gamma(\phi) &= 2 \int_{I_j} (w_j x + b_j)(f_\phi^\gamma(x) - x)dx + 2\gamma \int_{\hat{I}_j} (w_j x + b_j)(f_\phi^\gamma(x) - x)dx, \\
 \frac{\partial}{\partial c} \mathcal{L}^\gamma(\phi) &= 2 \int_0^1 (f_\phi^\gamma(x) - x)dx.
 \end{aligned}
 \tag{3.4}$$

This notation allows to treat non-flat degenerate neurons. For such neurons, the right-hand partial derivatives of  $\mathcal{L}^\gamma$  are also given by the above formulas. We now show how the reduction to the ReLU case works.

**Lemma 3.1** *Suppose  $\phi \in \mathbb{R}^{3N+1}$  is a critical point or a local extremum of  $\mathcal{L}^\gamma$  but not a global minimum and that  $\int_0^1 x(f_\phi^\gamma(x) - x)dx = 0$ . Then, all neurons of  $\phi$  are flat semi-active, flat inactive with  $w_j = 0$ , or flat degenerate.*

**Proof** We first show that  $P(\phi)$  is a critical point of  $\mathcal{L}$  and then apply Theorem 2.4 to  $P(\phi)$ . Since the partial derivative of  $\mathcal{L}^\gamma$  with respect to  $c$  exists and must be zero, we have

$$\frac{1}{2} \frac{\partial}{\partial c} \mathcal{L}^\gamma(\phi) = \int_0^1 (f_{P(\phi)}(x) - x)dx = 0 = \int_0^1 x(f_{P(\phi)}(x) - x)dx. \tag{3.5}$$

This shows that the (right-hand) partial derivatives of  $\mathcal{L}$  are zero at  $P(\phi)$  with respect to coordinates corresponding to inactive, semi-inactive, semi-active, type-1-active, and degenerate neurons. We need to verify that also partial derivatives of  $\mathcal{L}$  with respect to type-2-active neurons vanish at  $P(\phi)$ . To see this, note that, for a type-2-active neuron

$j$  of  $\phi$ , the partial derivative of  $\mathcal{L}^\gamma$  with respect to  $w_j$  exists at  $\phi$  and

$$0 = \frac{\partial}{\partial w_j} \mathcal{L}^\gamma(\phi) = 2(1 - \gamma)v_j \int_{I_j} x(f_\phi^\gamma(x) - x)dx. \tag{3.6}$$

Thus,

$$\begin{aligned} 0 &= 2v_j \int_{I_j} x(f_\phi^\gamma(x) - x)dx = \left(\frac{\partial}{\partial w_j} \mathcal{L}\right)(P(\phi)), \\ 0 &= -2\gamma v_j \int_{\hat{I}_j} x(f_\phi^\gamma(x) - x)dx = \left(\frac{\partial}{\partial w_{j+N}} \mathcal{L}\right)(P(\phi)), \end{aligned} \tag{3.7}$$

and analogously for the coordinates  $b_j, b_{j+N}, v_j, v_{j+N}$ . This concludes that  $P(\phi)$  is a critical point of  $\mathcal{L}$ . By Theorem 2.4,  $P(\phi)$  does not have any type-1-active, non-flat semi-active, or non-flat degenerate neurons. By definition of the map  $P$ , it follows that  $\phi$  does not have any type-1-active, non-flat semi-active, or non-flat degenerate neurons, nor does it have any semi-inactive, non-flat inactive, or inactive neurons with  $w_j \neq 0$  for otherwise  $P(\phi)$  would have one of the former types. Further, by definition of  $P$ , any type-2-active neuron of  $\phi$  gives rise to two type-2-active neurons of  $P(\phi)$  with the same breakpoint but with opposite signs of the  $w$ -coordinate. This is not possible by (IV.b) of Theorem 2.4, so  $\phi$  cannot have any type-2-active neurons. In summary,  $\phi$  can only have flat semi-active, flat degenerate, or flat inactive neurons with  $w_j = 0$ . □

The condition  $\int_0^1 x(f_\phi^\gamma(x) - x)dx = 0$  in the previous lemma is easily converted into a condition about existence of certain types of neurons. This is done in the first part of the next lemma. For the second part, we recycle some arguments we learned from the ReLU case.

**Lemma 3.2** *Suppose  $\phi \in \mathbb{R}^{3N+1}$  is a critical point or a local extremum of  $\mathcal{L}^\gamma$  but not a global minimum. Then, all neurons of  $\phi$  are flat semi-active, flat inactive with  $w_j = 0$ , degenerate, or type-2-active. Moreover, if  $\phi$  does not have any non-flat type-2-active neurons, then  $\phi$  is a saddle point and it also does not have any flat type-2-active or non-flat degenerate neurons.*

**Proof** Suppose  $\phi$  had a neuron of a different type than in the first statement of this lemma, say the  $j$ th. Note that one of the intervals  $I_j$  and  $\hat{I}_j$  is empty and the other one is  $[0, 1]$  (up to possibly a singleton). Since the  $j$ th neuron is non-degenerate,  $\mathcal{L}^\gamma$  is differentiable with respect to the coordinates of the  $j$ th neuron, so  $\int_0^1 x(f_\phi^\gamma(x) - x)dx = 0$ . This contradicts Lemma 3.1.

The remainder of the proof is similar to the ones of Lemmas 2.18 and 2.19. Assume  $\phi$  does not have any non-flat type-2-active neurons. Then,  $f_\phi^\gamma$  is constant on  $[0, 1]$ , and this constant is  $1/2$  since  $\frac{\partial}{\partial c} \mathcal{L}^\gamma(\phi) = 0$ . We claim that  $\phi$  cannot have any flat type-2-active neurons. Suppose for contradiction the  $j$ th neuron was that. Let  $\tau = \text{sign}(w_j)$

and  $t_j = -b_j/w_j \in (0, 1)$ . Then,  $\frac{\partial}{\partial v_j} \mathcal{L}^\gamma(\phi) = 0$  implies

$$\begin{aligned} 0 &= \int_{t_j}^1 (x - t_j)\left(\frac{1}{2} - x\right)dx + \gamma^\tau \int_0^{t_j} (x - t_j)\left(\frac{1}{2} - x\right)dx \\ &= \frac{1}{12}(-1 + (1 - \gamma^\tau)(3 - 2t_j)t_j^2). \end{aligned} \tag{3.8}$$

But, for any  $\gamma, t \in (0, 1), \tau \in \{-1, 1\}$ , we have  $-1 + (1 - \gamma^\tau)(3 - 2t)t^2 < 0$ , which is a contradiction. Thus, all neurons of  $\phi$  are flat semi-active, flat inactive with  $w_j = 0$ , or degenerate. With an argument analogous to the proof of Lemma 2.19, we find that  $\phi$  is a saddle point of  $\mathcal{L}^\gamma$ . Indeed, if there is a flat semi-active or flat inactive neuron  $j$  with  $w_j = 0$ , then, with  $\tau = 1 - \text{sign}(b_j)$ ,

$$\det \left( \begin{array}{cc} \frac{\partial}{\partial w_j} \frac{\partial}{\partial w_j} \mathcal{L}^\gamma(\phi) & \frac{\partial}{\partial w_j} \frac{\partial}{\partial v_j} \mathcal{L}^\gamma(\phi) \\ \frac{\partial}{\partial v_j} \frac{\partial}{\partial w_j} \mathcal{L}^\gamma(\phi) & \frac{\partial}{\partial v_j} \frac{\partial}{\partial v_j} \mathcal{L}^\gamma(\phi) \end{array} \right) = - \left( 2\gamma^{\tau/2} \int_0^1 x\left(\frac{1}{2} - x\right)dx \right)^2 = -\frac{1}{36}\gamma^\tau < 0. \tag{3.9}$$

Instead, if there is a degenerate neuron  $j$ , then, for the perturbation  $\phi^s, s \in [0, 1]$ , in the coordinates of the  $j$ th neuron given by  $w_j^s = \tau s, b_j^s = -\tau s^2$ , and  $v_j^s = v_j + \tau s$  with  $\tau = 1$  if  $v_j \geq 0$  and  $\tau = -1$  if  $v_j < 0$ , we have

$$\begin{aligned} \mathcal{L}^\gamma(\phi^s) - \mathcal{L}^\gamma(\phi) &= \frac{1}{6}v_j^s w_j^s \gamma^{(1-\tau)/2}(-1 + (1 - \gamma^\tau)(3 - 2s)s^2) \\ &\quad + \frac{1}{3}(v_j^s w_j^s)^2 \gamma^{1-\tau}((1 - s)^3 + \gamma^{2\tau} s^3) \\ &= -\frac{1}{6}s(|v_j| + s)\gamma^{(1-\tau)/2} + \frac{1}{3}|v_j|^2 s^2 \gamma^{1-\tau} + \mathcal{O}(s^3), \end{aligned} \tag{3.10}$$

which is strictly negative for small  $s > 0$ . This concludes that  $\phi$  is a saddle point. In particular, any degenerate neuron  $j$  must be flat because

$$0 = \frac{\partial^+}{\partial w_j} \mathcal{L}^\gamma(\phi) = 2v_j \int_0^1 x\left(\frac{1}{2} - x\right)dx = -\frac{v_j}{6}. \tag{3.11}$$

□

We finished dealing with critical points of  $\mathcal{L}^\gamma$  that have a constant realization function. In the next section, we find saddle points of  $\mathcal{L}^\gamma$  analogous to the ones in Theorem 2.4.(IV.b). For these, we cannot reduce the analysis entirely to the known ReLU case. However, the arguments are analogous to the ones developed in Lemmas 2.23 and 2.24, and we can use a shortcut for small  $\gamma$  by arguing that we approximate the ReLU case in a suitable sense.

### 3.2 Explicit Analysis for Leaky ReLU

The following is the analog of Lemma 2.23 in the leaky ReLU case. Informally, one recovers Lemma 2.23 from Lemma 3.3 in the limit  $\gamma \rightarrow 0$ . We will discuss this in more detail after having proved the lemma.

**Lemma 3.3** *Suppose  $\phi \in \mathbb{R}^{3N+1}$  is a critical point or a local extremum of  $\mathcal{L}^\gamma$  but not a global minimum and that  $\phi$  has a type-2-active neuron. Denote by  $0 = q_0 < q_1 < \dots < q_n < q_{n+1} = 1$ , for  $n \in \mathbb{N}_0$ , the roughest partition such that  $f_\phi^\gamma$  is affine on all subintervals  $[q_i, q_{i+1}]$ , and denote by  $K_i \subseteq \{1, \dots, N\}$  the set of all type-2-active neurons of  $\phi$  whose breakpoint is  $q_i$ . Then,  $n \geq 1$  and there exists  $\sigma \in \{-1, 1\}$  such that, abbreviating*

$$\delta = \gamma^{(1-\sigma)/4} + \gamma^{(1-\sigma(-1)^n)/4} + (n - 1)\sqrt{1 + \gamma}, \tag{3.12}$$

the following hold:

- (i) (a)  $q_i = q_1 + \frac{(i-1)(q_n - q_1)}{n-1}$  for all  $i \in \{2, \dots, n - 1\}$ ,
- (b)  $q_1 = \delta^{-1}\gamma^{(1-\sigma)/4}$ , and  $q_n = 1 - \delta^{-1}\gamma^{(1-\sigma(-1)^n)/4}$ , and  $q_n - q_1 = \delta^{-1}(n - 1)\sqrt{1 + \gamma}$ ,
- (ii)  $-b_j/w_j \in \{q_1, \dots, q_n\}$  for all type-2-active neurons  $j \in \{1, \dots, N\}$  of  $\phi$ ,
- (iii)  $\text{sign}(w_j) = \sigma(-1)^{i+1}$  for all  $i \in \{1, \dots, n\}$ ,  $j \in K_i$ ,

$$\text{(iv) (a) } \sum_{j \in K_i} v_j w_j = \begin{cases} \gamma^{-1/2} & \text{if } i = 1 = n, \\ \frac{1}{\delta} \left( \frac{1}{\sqrt{1+\gamma}} + \frac{1}{\gamma^{(1-\sigma)/4}} \right) & \text{if } i = 1 \neq n, \\ \frac{1}{\delta} \frac{2}{\sqrt{1+\gamma}} & \text{if } 2 \leq i \leq n - 1, \\ \frac{1}{\delta} \left( \frac{1}{\sqrt{1+\gamma}} + \frac{1}{\gamma^{(1-\sigma(-1)^n)/4}} \right) & \text{if } i = n \neq 1, \end{cases}$$

(v)  $\phi$  is centered,

$$\text{(vi) } f_\phi^\gamma(x) - x = \frac{-\sigma(-1)^i(1-\gamma)}{\delta} \cdot \begin{cases} \frac{x}{\gamma^{(1-\sigma)/4}} - \frac{1}{2\delta} & \text{if } i = 0, \\ \frac{x}{\sqrt{1+\gamma}} - \frac{i-1/2}{\delta} - \frac{\gamma^{(1-\sigma)/4}}{\delta\sqrt{1+\gamma}} & \text{if } 1 \leq i \leq n - 1, \\ \frac{x}{\gamma^{(1-\sigma(-1)^n)/4}} + \frac{1}{2\delta} - \frac{1}{\gamma^{(1-\sigma(-1)^n)/4}} & \text{if } i = n \end{cases}$$

for all  $i \in \{0, \dots, n\}$ ,  $x \in [q_i, q_{i+1}]$ .

**Proof** First, note that  $\phi$  must have at least one non-flat type-2-active neuron by Lemma 3.2. For any such neuron  $j$ ,

$$0 = \frac{1}{2v_j} \frac{\partial}{\partial w_j} \mathcal{L}^\gamma(\phi) = (1 - \gamma) \int_{I_j} x(f_\phi^\gamma(x) - x)dx + \gamma \int_0^1 x(f_\phi^\gamma(x) - x)dx, \tag{3.13}$$

so the two integrals

$$\begin{aligned} \int_{I_j} x(f_\phi^\gamma(x) - x)dx &= \frac{-\gamma}{1 - \gamma} \int_0^1 x(f_\phi^\gamma(x) - x)dx, \\ \int_{I_j} x(f_\phi^\gamma(x) - x)dx &= \frac{1}{1 - \gamma} \int_0^1 x(f_\phi^\gamma(x) - x)dx \end{aligned} \tag{3.14}$$

are independent of the non-flat type-2-active neuron  $j$ . Doing the same with the coordinate  $b_j$  and using that  $2 \int_0^1 (f_\phi^\gamma(x) - x)dx = \frac{\partial}{\partial c} \mathcal{L}^\gamma(\phi) = 0$ , we find

$$\int_{I_j} (f_\phi^\gamma(x) - x)dx = 0 = \int_{\hat{I}_j} (f_\phi^\gamma(x) - x)dx. \tag{3.15}$$

The function  $f_\phi^\gamma$  cannot be affine for otherwise we could apply Lemma 2.20 with the partition  $0 < t_j < 1$  for the breakpoint  $t_j$  of any non-flat type-2-active neuron  $j$  and obtain a contradiction with  $\phi$  not being a global minimum. In other words,  $n \neq 0$ . Moreover, since each  $K_i, i \in \{1, \dots, n\}$ , must contain a non-flat neuron, we deduce from (3.15) that  $\int_{q_i}^{q_{i+1}} (f_\phi^\gamma(x) - x)dx = 0$  for all  $i \in \{0, \dots, n\}$ . It follows from this and  $\frac{\partial}{\partial v} \mathcal{L}^\gamma(\phi) = 0$  that (3.14) holds even for flat neurons  $j \in \bigcup_i K_i$ . Also, Lemma 2.20 implies that the two integrals in (3.14) are not zero. In particular,

$$\int_{I_j} x(f_\phi^\gamma(x) - x)dx \neq \int_{\hat{I}_j} x(f_\phi^\gamma(x) - x)dx \tag{3.16}$$

for any  $j \in \bigcup_i K_i$  and, hence,  $\text{sign}(w_{j_0}) = \text{sign}(w_{j_1})$  if  $j_0$  and  $j_1$  belong to the same set  $K_i$ . Furthermore, we find from (3.14) that  $\text{sign}(w_{j_0}) \neq \text{sign}(w_{j_1})$  for all  $i \in \{1, \dots, n-1\}$  and  $j_0 \in K_i, j_1 \in K_{i+1}$  by taking differences of the integrals  $\int_{I_j} x(f_\phi^\gamma(x) - x)dx$  for different  $j$ . This establishes item (iii). Consequently, we obtain from Lemma 2.20.(v) (with the partition  $q_i, q_{i+1}, q_{i+2}$ ) that

$$0 = (q_{i+2} - q_{i+1})^2 - (q_{i+1} - q_i)^2, \tag{3.17}$$

for all  $i \in \{1, \dots, n-2\}$  (note that we do not obtain this equality for  $i = 0$  or  $i = n-1$ ). Thus, the points  $q_1, \dots, q_n$  are equidistributed in  $[q_1, q_n]$  (but not necessarily in  $[0, 1]$ ), which is exactly item (i.a). Next, we prove item (i.b). To do so, we distinguish between even  $n$  and odd  $n$ . In the former case,  $\text{sign}(w_{j_1}) \neq \text{sign}(w_{j_n})$  for all  $j_1 \in K_1, j_n \in K_n$  by item (iii) and, hence, by (3.14),

$$\int_0^{q_1} x(f_\phi^\gamma(x) - x)dx = \int_{q_n}^1 x(f_\phi^\gamma(x) - x)dx. \tag{3.18}$$

Write  $f_\phi^\gamma(x) = A_i x + B_i$  on  $[q_i, q_{i+1}]$ , for all  $i \in \{0, \dots, n\}$ , so that the formulas in (2.31) hold. We compute

$$\begin{aligned} \frac{1}{12}(A_0 - 1)q_1^3 &= \int_0^{q_1} x(f_\phi^\gamma(x) - x)dx = \int_{q_n}^1 x(f_\phi^\gamma(x) - x)dx \\ &= \frac{(-1)^n}{12}(A_0 - 1)q_1(1 - q_n)^2. \end{aligned} \tag{3.19}$$

Thus,  $q_1 = 1 - q_n$  and, by (i.a),

$$\begin{aligned} \int_0^1 x(f_\phi^\gamma(x) - x)dx &= \frac{1}{12}(A_0 - 1)q_1 \sum_{k=0}^n (-1)^k (q_{k+1} - q_k)^2 \\ &= \frac{1}{12}(A_0 - 1)q_1 \left( 2q_1^2 - \left( \frac{1 - 2q_1}{n - 1} \right)^2 \right). \end{aligned} \tag{3.20}$$

Hence, it follows from (3.14) and item (iii) that

$$q_1^2 = \frac{\sigma\gamma^{(1-\sigma)/2}}{1 - \gamma} \left( 2q_1^2 - \left( \frac{1 - 2q_1}{n - 1} \right)^2 \right). \tag{3.21}$$

Solving this as a quadratic equation in  $q_1$  under the constraint  $q_1 \in (0, 1/2)$  yields  $q_1 = \delta^{-1}\gamma^{(1-\sigma)/4}$ . Now, assume  $n$  is odd. Recall that  $\int_{q_i}^{q_{i+2}} x(f_\phi^\gamma(x) - x)dx = 0$  for all  $i \in \{1, \dots, n - 2\}$ . In particular,  $\int_{q_1}^{q_n} x(f_\phi^\gamma(x) - x)dx = 0$ . Note that  $\sigma$  is already determined as the sign of  $w_j$  for any  $j \in K_1$ . The partial derivative with respect to  $w_j$  being zero for a non-flat neuron  $j \in K_1$  implies

$$\begin{aligned} 0 &= \int_{q_n}^1 x(f_\phi^\gamma(x) - x)dx + \gamma^\sigma \int_0^{q_1} x(f_\phi^\gamma(x) - x)dx \\ &= -\frac{1}{12}(A_0 - 1)q_1((1 - q_n)^2 - \gamma^\sigma q_1^2). \end{aligned} \tag{3.22}$$

Thus,  $1 - q_n = \gamma^{\sigma/2}q_1$ . From this, the formula for  $q_1$  follows in the case  $n = 1$ . If  $n \neq 1$ , then we use that the partial derivative with respect to  $w_j$  for a non-flat neuron  $j \in K_2$  is zero to calculate

$$\begin{aligned} 0 &= \int_{q_2}^1 x(f_\phi^\gamma(x) - x)dx + \gamma^{-\sigma} \int_0^{q_2} x(f_\phi^\gamma(x) - x)dx \\ &= \int_{q_{n-1}}^1 x(f_\phi^\gamma(x) - x)dx + \gamma^{-\sigma} \int_0^{q_2} x(f_\phi^\gamma(x) - x)dx \\ &= \frac{1}{12}(A_0 - 1)q_1 \left[ \gamma^{-\sigma} q_1^2 - (1 - q_n)^2 + (1 - \gamma^{-\sigma}) \left( \frac{q_n - q_1}{n - 1} \right)^2 \right]. \end{aligned} \tag{3.23}$$

Using  $1 - q_n = \gamma^{\sigma/2}q_1$ , the term in the rectangular brackets becomes a quadratic polynomial in  $q_1$ , and solving for  $q_1$  leads to  $q_1 = \delta^{-1}\gamma^{(1-\sigma)/4}$ . This finishes item (i.b). From here on, we no longer treat even  $n$  and odd  $n$  separately. Next, we show item (ii). Given any type-2-active neuron  $j \in \{1, \dots, N\}$ , take  $i \in \{0, \dots, n\}$  with



$q_i \leq t_j \leq q_{i+1}$  and denote  $\tau = \text{sign}(w_j)$ . Then,  $\frac{\partial}{\partial v_j} \mathcal{L}(\phi) = 0$  implies

$$\begin{aligned}
 0 &= \int_{t_j}^{q_{i+1}} (x - t_j)(f_\phi^\gamma(x) - x)dx + \gamma^\tau \int_{q_i}^{t_j} (x - t_j)(f_\phi^\gamma(x) - x)dx \\
 &+ \int_{q_{i+1}}^1 x(f_\phi^\gamma(x) - x)dx + \gamma^\tau \int_0^{q_i} x(f_\phi^\gamma(x) - x)dx.
 \end{aligned}
 \tag{3.24}$$

A direct computation with the formulas in (2.31) yields

$$\begin{aligned}
 &\int_{t_j}^{q_{i+1}} (x - t_j)(f_\phi^\gamma(x) - x)dx + \gamma^\tau \int_{q_i}^{t_j} (x - t_j)(f_\phi^\gamma(x) - x)dx \\
 &= \frac{(-1)^i}{12} (A_0 - 1) \frac{q_1}{q_{i+1} - q_i} \left[ (q_{i+1} - q_i)^3 - (1 - \gamma^\tau)(t_j - q_i)^2(3q_{i+1} - 2t_j - q_i) \right].
 \end{aligned}
 \tag{3.25}$$

Furthermore, if  $i \neq 0$  and  $\tau = \sigma(-1)^{i+1}$ , then

$$\begin{aligned}
 &\int_{q_{i+1}}^1 x(f_\phi^\gamma(x) - x)dx + \gamma^\tau \int_0^{q_i} x(f_\phi^\gamma(x) - x)dx \\
 &= - \int_{q_i}^{q_{i+1}} x(f_\phi^\gamma(x) - x)dx + \int_{q_i}^1 x(f_\phi^\gamma(x) - x)dx + \gamma^\tau \int_0^{q_i} x(f_\phi^\gamma(x) - x)dx \\
 &= - \int_{q_i}^{q_{i+1}} x(f_\phi^\gamma(x) - x)dx = - \frac{(-1)^i}{12} (A_0 - 1) q_1 (q_{i+1} - q_i)^2,
 \end{aligned}
 \tag{3.26}$$

where the second-last equality is implied by  $\frac{\partial}{\partial w_k} \mathcal{L}(\phi) = 0$  for a non-flat type-2-active neuron  $k \in K_i$ . Similarly, if  $i \neq n$  and  $\tau = \sigma(-1)^{i+2}$ , then

$$\begin{aligned}
 &\int_{q_{i+1}}^1 x(f_\phi^\gamma(x) - x)dx + \gamma^\tau \int_0^{q_i} x(f_\phi^\gamma(x) - x)dx \\
 &= -\gamma^\tau \int_{q_i}^{q_{i+1}} x(f_\phi^\gamma(x) - x)dx = -\gamma^\tau \frac{(-1)^i}{12} (A_0 - 1) q_1 (q_{i+1} - q_i)^2.
 \end{aligned}
 \tag{3.27}$$

The remaining cases are  $i \in \{0, n\}$  with  $\tau = -\sigma$ , respectively,  $\tau = \sigma(-1)^n$ , for which

$$\int_{q_{i+1}}^1 x(f_\phi^\gamma(x) - x)dx + \gamma^\tau \int_0^{q_i} x(f_\phi^\gamma(x) - x)dx = \frac{(-1)^{n-i}}{12}(A_0 - 1)\gamma^{i\tau/n}q_1 \cdot \begin{cases} \gamma^{(n-i)\sigma/n}q_1^2 & \text{if } n \text{ is odd,} \\ q_1^2 - (q_2 - q_1)^2 & \text{if } n \text{ is even.} \end{cases} \tag{3.28}$$

In conclusion, we obtain from (3.24) and (3.25) that

$$0 = \begin{cases} -(t_j - q_i)^2(3q_{i+1} - 2t_j - q_i) & \text{if } i \neq 0 \text{ and } \tau = \sigma(-1)^{i+1}, \\ (q_{i+1} - q_i)^3 - (t_j - q_i)^2(3q_{i+1} - 2t_j - q_i) & \text{if } i \neq n \text{ and } \tau = \sigma(-1)^{i+2}, \\ (1 - \gamma^\sigma)q_1^3 - (1 - \gamma^{-\sigma})t_j^2(3q_1 - 2t_j) & \text{if } n \text{ is odd, } i = 0, \text{ and } \tau = -\sigma, \\ (1 + \gamma^\sigma)(1 - q_n)q_1^2 - (t_j - q_n)^2(3 - 2t_j - q_n) & \text{if } n \text{ is odd, } i = n, \text{ and } \tau = -\sigma, \\ 2q_1^3 - q_1(q_2 - q_1)^2 - (1 - \gamma^{-\sigma})t_j^2(3q_1 - 2t_j) & \text{if } n \text{ is even, } i = 0, \text{ and } \tau = -\sigma, \\ (1 + \gamma^\sigma)q_1^3 - \gamma^\sigma q_1(q_2 - q_1)^2 & \text{if } n \text{ is even, } i = n, \text{ and } \tau = \sigma. \end{cases} \tag{3.29}$$

In the first case, we must have  $t_j = q_i$ . In the second case, the term can be rewritten as  $(q_{i+1} - t_j)^2(q_{i+1} + 2t_j - 3q_i)$ , so we must have  $t_j = q_{i+1}$ . In the third case, the two summands always have opposite signs, so their difference is always strictly positive or strictly negative but not zero. In the fourth case, the right hand side is lower bounded by  $(1 - q_n)q_1^2$ , so it cannot be zero. In the fifth case, after plugging in  $q_1$  and  $q_2$ , we find that  $t_j$  must satisfy

$$0 = \sqrt{\gamma}\gamma^{(1+\sigma)/4} + t_j^2\delta^2(3\gamma^{(1-\sigma)/4} - 2t_j\delta). \tag{3.30}$$

However, there is no solution  $t_j$  to this equation with  $t_j \in [0, q_1]$ . Lastly, in the sixth case,  $1 - t_j$  must satisfy the same equation, which is incompatible with  $t_j \in [q_n, 1]$ . This proves item (ii). Now, we tend to item (iv). Since  $\bigcup_i K_i$  contains all type-2-active neurons of  $\phi$  and there are no type-1-active neurons by Lemma 3.2, we can write the slopes of  $f_\phi^\gamma$  as

$$A_l = \sum_{i=1}^l \gamma^{\frac{1+\sigma(-1)^i}{2}} \sum_{j \in K_i} v_j w_j + \sum_{i=l+1}^n \gamma^{\frac{1-\sigma(-1)^i}{2}} \sum_{j \in K_i} v_j w_j, \tag{3.31}$$

for all  $l \in \{0, \dots, n\}$ , by item (iii). With this, we find, for all  $i \in \{1, \dots, n\}$ ,

$$\begin{aligned} -\frac{q_1}{q_{i+1} - q_i}(A_0 - 1) &= (-1)^{i-1}(A_i - 1) \\ &= (-1)^{i-1}(A_{i-1} - 1) + \sigma(1 - \gamma) \sum_{j \in K_i} v_j w_j \\ &= \frac{q_1}{q_i - q_{i-1}}(A_0 - 1) + \sigma(1 - \gamma) \sum_{j \in K_i} v_j w_j. \end{aligned} \tag{3.32}$$

Thus, for all  $i \in \{1, \dots, n\}$ ,

$$\sum_{j \in K_i} v_j w_j = \frac{-\sigma}{1 - \gamma} (A_0 - 1) q_1 \frac{q_{i+1} - q_{i-1}}{(q_{i+1} - q_i)(q_i - q_{i-1})}. \tag{3.33}$$

Combining this with the formula (3.31) for  $A_0$  yields

$$\frac{-\sigma(1 - \gamma)}{A_0 - 1} = \sigma(1 - \gamma) + q_1 \sum_{i=1}^n \gamma^{\frac{1-\sigma(-1)^i}{2}} \frac{q_{i+1} - q_{i-1}}{(q_{i+1} - q_i)(q_i - q_{i-1})} = \gamma^{(1-\sigma)/4} \delta. \tag{3.34}$$

Plugging this back into the formula for  $\sum_{j \in K_i} v_j w_j$ , we obtain for  $n = 1$  that  $\sum_{j \in K_1} v_j w_j = \gamma^{-1/2}$  and for  $n \geq 2, i \in \{1, \dots, n\}$  that

$$\begin{aligned} \sum_{j \in K_i} v_j w_j &= \frac{1}{\delta^2} \frac{q_{i+1} - q_{i-1}}{(q_{i+1} - q_i)(q_i - q_{i-1})} \\ &= \begin{cases} \delta^{-1}((1 + \gamma)^{-1/2} + \gamma^{-(1-\sigma)/4}) & \text{if } i = 1, \\ 2\delta^{-1}(1 + \gamma)^{-1/2} & \text{if } 2 \leq i \leq n - 1, \\ \delta^{-1}((1 + \gamma)^{-1/2} + \gamma^{-(1-\sigma(-1)^n)/4}) & \text{if } i = n. \end{cases} \end{aligned} \tag{3.35}$$

This establishes item (iv). By the formulas in (2.31) and (3.34),

$$A_i - 1 = \sigma(-1)^{i+1} (1 - \gamma) \delta^{-1} \cdot \begin{cases} \gamma^{-(1-\sigma)/4} & \text{if } i = 0, \\ (1 + \gamma)^{-1/2} & \text{if } 1 \leq i \leq n - 1, \\ \gamma^{-(1-\sigma(-1)^n)/4} & \text{if } i = n \end{cases} \tag{3.36}$$

and

$$B_i = \frac{1}{2} \sigma(-1)^i (1 - \gamma) \delta^{-2} \cdot \begin{cases} 1 & \text{if } i = 0, \\ 2i - 1 + 2\gamma^{(1-\sigma)/4} (1 + \gamma)^{-1/2} & \text{if } 1 \leq i \leq n - 1, \\ 2\gamma^{-(1-\sigma(-1)^n)/4} \delta - 1 & \text{if } i = n. \end{cases} \tag{3.37}$$

In particular, item (vi) holds. Lastly, we know from Lemma 3.2 and item (iii) that

$$0 = f_\phi^\gamma(0) - B_0 = c - \sum_{i=1}^n \gamma^{\frac{1-\sigma(-1)^i}{2}} q_i \sum_{j \in K_i} v_j w_j - B_0. \tag{3.38}$$

After plugging in the formulas for  $B_0, \delta, q_i$ , and  $\sum_{j \in K_i} v_j w_j$ , a lengthy but straightforward computation results in  $c = 1/2$ . Thus,  $\phi$  is centered, which concludes the proof. □

We make a few remarks about the relationship between the previous lemma and Lemma 2.23. The quantity  $\delta$  in Lemma 3.3 replaces the factor  $n + 1$  that appears throughout Lemma 2.23. In the limit  $\gamma \rightarrow 0$ ,

$$\delta \rightarrow \begin{cases} n & \text{if } n \text{ is odd,} \\ n + 1 & \text{if } n \text{ is even and } \sigma = 1, \\ n - 1 & \text{if } n \text{ is even and } \sigma = -1. \end{cases} \tag{3.39}$$

Thus, in order to match Lemma 2.23 with the limit case  $\gamma \rightarrow 0$  of Lemma 3.3, one would need to apply the former lemma with

$$\tilde{n} = \begin{cases} n - 1 & \text{if } n \text{ is odd,} \\ n & \text{if } n \text{ is even and } \sigma = 1, \\ n - 2 & \text{if } n \text{ is even and } \sigma = -1 \end{cases} \tag{3.40}$$

in place of  $n$  so that  $\delta \rightarrow \tilde{n} + 1$ . One would hope that the quantities in Lemma 3.3 converge to their counterparts from Lemma 2.23 with  $\tilde{n}$  as  $\gamma \rightarrow 0$ . Although the number of breakpoints in each lemma is different in most cases (i.e.  $n \neq \tilde{n}$ ), this convergence actually happens: on the one hand, if  $n$  is odd and  $\sigma = 1$ , then  $q_n \rightarrow 1$  ‘degenerates’ into the endpoint of the interval  $[0, 1]$  and only the  $(n - 1)$ -many breakpoints  $q_1, \dots, q_{n-1}$  remain, which converge to  $\frac{i}{\tilde{n}+1}, i \in \{1, \dots, \tilde{n}\}$ . Similarly, if  $n$  is odd and  $\sigma = -1$ , then  $q_1 \rightarrow 0$  degenerates and  $q_2, \dots, q_n$  remain and converge to the correct breakpoints  $\frac{i}{\tilde{n}+1}, i \in \{1, \dots, \tilde{n}\}$ . On the other hand, if  $n$  is even and  $\sigma = 1$ , then none of the breakpoints degenerate and  $q_1, \dots, q_n$  remain and converge. Lastly, if  $n$  is even and  $\sigma = -1$ , then both  $q_1 \rightarrow 0$  and  $q_n \rightarrow n$ , and we are left with  $q_2, \dots, q_{n-1}$ , which converge.

In addition, note that the parity of the  $w$ -coordinate of the type-2-active neurons match in each lemma even though these are  $\sigma(-1)^{i+1}$  and  $(-1)^{i+1}$ , respectively. They match because  $q_1$  can only degenerate into 0 if  $\sigma = -1$ . Lastly, note that the quantities  $\sum_{j \in K_i} v_j w_j$  also converge to their counterparts as  $\gamma \rightarrow 0$ .

**Lemma 3.4** *Suppose  $\phi \in \mathbb{R}^{3N+1}$  is a critical point or a local extremum of  $\mathcal{L}^\gamma$  but not a global minimum and that  $\phi$  has a type-2-active neuron. There exists  $\gamma_0 \in (0, 1]$  depending only on  $N$  such that if  $\gamma < \gamma_0$ , then  $\phi$  is a saddle point of  $\mathcal{L}^\gamma$ .*

Recall that, in the proof of Lemma 2.24, we studied the Hessian of  $\mathcal{L}^\gamma$  restricted to a suitable set of coordinates, taken from type-2-active neurons with breakpoints  $\frac{i}{\bar{n}+1}, i \in \{1, 2\}$ . To prove Lemma 3.4, we proceed analogously, which works for sufficiently small  $\gamma$  by the above observation about Lemmas 2.23 and 3.3. More precisely, if  $n \neq 1$  and  $\sigma = 1$ , then we will be able to work with the same set of coordinates because  $q_1 \rightarrow \frac{1}{\bar{n}+1}$  and  $q_2 \rightarrow \frac{2}{\bar{n}+1}$ . On the other hand, if  $n \geq 3$  and  $\sigma = -1$ , then  $q_1 \rightarrow 0$  but  $q_2 \rightarrow \frac{1}{\bar{n}+1}$  and  $q_3 \rightarrow \frac{2}{\bar{n}+1}$ . In this case, we will use the analogous set of coordinates with  $q_2$  and  $q_3$  in place of  $q_1$  and  $q_2$ . However, the argument does not work if  $n = 1$  or if  $n = 2$  and  $\sigma = -1$  because then  $q_1 \rightarrow 0, q_2 \rightarrow 1$ , and  $f_\phi^\gamma$  becomes an affine function as  $\gamma \rightarrow 0$ . We will treat these two cases separately.

**Proof of Lemma 3.4** Take  $n, \delta, q_1, \dots, q_n$ , and  $\sigma$  from Lemma 3.3. First, assume  $n = 2$  with  $\sigma = 1$  or  $n \geq 3$ . Abbreviate  $\tau = (3 - \sigma)/2 \in \{1, 2\}$ . Similar to the proof of Lemma 2.24, let  $K_\tau \subseteq \{1, \dots, N\}$  denote the set of those type-2-active neurons with breakpoint  $q_\tau$ , and let  $K_\tau^- \subseteq K_\tau$  be the subset of those neurons  $j \in K_\tau$  with  $v_j < 0$ . Let  $j_1 \in K_\tau$  with  $v_{j_1} > 0$ , which exists since  $a := \sum_{j \in K_\tau} v_j w_j > 0$  and  $w_j > 0$  for all  $j \in K_\tau$ , and let  $j_2, \dots, j_l$ , for  $l \in \{1, \dots, N\}$ , be an enumeration of  $K_\tau^-$ . Moreover, let  $k \in \{1, \dots, N\}$  be any type-2-active neuron with breakpoint  $q_{\tau+1}$ . As in the ReLU case, we consider the Hessian  $H$  of  $\mathcal{L}^\gamma$  restricted to  $(b_{j_1}, \dots, b_{j_l}, v_k, c)$ .

We again introduce some shorthand notation. For all  $i \in \{1, \dots, l\}$ , denote  $\lambda_i = a^{-1} v_{j_i} w_{j_i}$  so that  $\sum_{i=1}^l \lambda_i \leq 1$ . Define  $\mu = \frac{1}{2}(1 - (1 - \gamma^2)q_\tau)^{-1} > 0$  and the vectors  $u_1 = (v_{j_1}, \dots, v_{j_l})$ ,

$$u_2 = \mu \begin{pmatrix} w_k(\gamma(1 - 2q_{\tau+1}) - (1 - \gamma)(q_{\tau+1} - q_\tau)^2) \\ 2(1 - (1 - \gamma)q_\tau) \end{pmatrix}, \tag{3.41}$$

and  $u = (u_1, u_2)$ . Furthermore, let  $D$  be the diagonal matrix with entries  $-\mu(1 - \gamma)^2 v_{j_i}^2 / (a\delta^2 \lambda_i)$ ,  $i \in \{1, \dots, l\}$ , let  $A$  be the Hessian of  $\mathcal{L}^\gamma$  restricted to  $(v_k, c)$ , let  $B = \mu A - u_2 u_2^T$ , and let  $E$  be the diagonal block matrix with blocks  $D$  and  $B$ . Then,  $H = \frac{1}{\mu}(E + uu^T)$ . The matrix  $A$  is

$$A = \begin{pmatrix} \frac{2}{3} w_k^2 (q_{\tau+1}^3 + \gamma^2(1 - q_{\tau+1})^3) - w_k (q_{\tau+1}^2 - \gamma(1 - q_{\tau+1})^2) \\ -w_k (q_{\tau+1}^2 - \gamma(1 - q_{\tau+1})^2) & 2 \end{pmatrix}, \tag{3.42}$$

of which both the determinant and the upper left entry are strictly positive. In particular,  $A$  is positive definite and, hence,  $\Gamma := \frac{1}{\mu} u_2^T A^{-1} u_2$  is strictly positive. If  $\Gamma < 1$ , then the same considerations as in the proof of Lemma 2.24 show that  $B$  and  $E$  are invertible and

$$\begin{aligned} \det(H) &= \mu^{-(l+2)} (1 + u_1^T D^{-1} u_1 + u_2^T B^{-1} u_2) \det(E) \\ &= \Delta \left( \frac{a}{\mu} \left( \frac{\delta}{1 - \gamma} \right)^2 (1 - \Gamma) \sum_{i=1}^l \lambda_i - 1 \right), \end{aligned} \tag{3.43}$$

where  $\Delta = -\mu^{(l+2)}(1 - \Gamma)^{-1} \det(D) \det(B) > 0$ . So far, we did not impose any restrictions on  $\gamma$ . To verify that  $\Gamma < 1$ , we use the limit argument to reduce the calculation to the one we performed in the proof of Lemma 2.24. To this end, we point out that  $\Gamma$  is independent of  $w_k$  and that  $\delta, q_\tau, q_{\tau+1}$ , and  $\mu$  only depend on  $n$  and  $\gamma$ . For fixed  $n$ , if we let  $\gamma$  tend to zero, then  $\delta \rightarrow \tilde{n} + 1$ ,  $q_\tau \rightarrow \frac{1}{\tilde{n}+1}$ ,  $q_{\tau+1} \rightarrow \frac{2}{\tilde{n}+1}$ , and  $\mu \rightarrow \frac{\tilde{n}+1}{2\tilde{n}}$ , where we take  $\tilde{n} = n - 1 + \sigma$  if  $n$  is even and  $\tilde{n} = n - 1$  if  $n$  is odd. These limits coincide with the corresponding objects from the proof of Lemma 2.24 with  $\tilde{n}$  in place of  $n$  as discussed prior to stating Lemma 3.4. The same goes for the limits of  $a, u_2$ , and  $A$ . Thus, we find from (2.44) that, for sufficiently small  $\gamma$ ,

$$\Gamma \approx \frac{32\tilde{n}^2 - 21\tilde{n} + 3}{16\tilde{n}(2\tilde{n} - 1)} < 1 \quad \text{and} \quad \frac{a}{\mu} \left( \frac{\delta}{1 - \gamma} \right)^2 (1 - \Gamma) \approx 4\tilde{n}(1 - \Gamma) \approx \frac{5\tilde{n} - 3}{8\tilde{n} - 4} < 1. \tag{3.44}$$

This concludes the existence of a  $\gamma_0 \in (0, 1]$  such that if  $\gamma < \gamma_0$ , then  $\det(H) < 0$ . This  $\gamma_0$  depends only on  $n$ . Since  $n \leq N$ , we can shrink  $\gamma_0$  if necessary so that it depends only on  $N$ .

It remains to treat the cases  $n = 1$  and  $n = 2$  with  $\sigma = -1$ . Assume  $n = 1$ . This time, let  $j_1 \in \{1, \dots, N\}$  be any type-2-active neuron with  $\text{sign}(v_{j_1}) = \sigma$ , and let  $j_2, \dots, j_l$ , for  $l \in \{1, \dots, N\}$ , be an enumeration of all type-2-active neurons with  $\text{sign}(v_{j_l}) = -\sigma$ . As before, let  $a = \gamma^{-1/2}$ ,  $\lambda_i = a^{-1}v_{j_i}w_{j_i}$ ,  $\mu = \frac{1}{2}\gamma^{-1/2}(1 - \sqrt{\gamma} + \gamma)^{-1}$ ,  $D_i = -\mu(1 - \gamma)^2v_{j_i}^2/(a\delta^2\lambda_i)$ , and  $u_1 = (v_{j_1}, \dots, v_{j_l})$  so that  $\sum_{i=1}^l \lambda_i \leq 1$  and  $\det(D) < 0$ . On the other hand, let  $u_2 = \sigma\mu\sqrt{\gamma}(1 - \gamma)\lambda_1/(\delta^2v_{j_1})$  and  $B = \mu\frac{\partial^2}{\partial v_{j_1}^2}\mathcal{L}^\gamma(\phi) - u_2^2 = \frac{1}{3}\mu^2\gamma\lambda_1^2v_{j_1}^{-2} > 0$ . Then, the Hessian of  $\mathcal{L}^\gamma$  restricted to the coordinates  $(b_{j_1}, \dots, b_{j_l}, v_{j_1})$  is  $H = \frac{1}{\mu}(E + uu^T)$ , where  $E$  is the diagonal block matrix with blocks  $D$  and  $B$ . Hence,

$$\begin{aligned} \det(H) &= \mu^{-(l+1)}B \det(D)(1 + u_1^T D^{-1}u_1 + u_2^2/B) \\ &= -\mu^{-(l+1)}B \det(D)\frac{4(1 - \sqrt{\gamma} + \gamma)}{(1 + \sqrt{\gamma})^2} \left( \frac{1}{2} \left( \frac{1 + \sqrt{\gamma}}{1 - \sqrt{\gamma}} \right)^2 \sum_{i=1}^l \lambda_i - 1 \right). \end{aligned} \tag{3.45}$$

In particular,  $\det(H) < 0$  for sufficiently small  $\gamma$ .

Lastly, assume  $n = 2$  and  $\sigma = -1$ . Similar as in the beginning, let  $K_1 \subseteq \{1, \dots, N\}$  denote the set of those type-2-active neurons with breakpoint  $q_1$ , and let  $K_1^+ \subseteq K_1$  be the subset of those neurons  $j \in K_1$  with  $v_j > 0$ . Let  $j_1 \in K_1$  with  $v_{j_1} < 0$ , which exists since  $a = \sum_{j \in K_1} v_j w_j > 0$  and  $w_j < 0$  for all  $j \in K_1$ , and let  $j_2, \dots, j_l$ , for  $l \in \{1, \dots, N\}$ , be an enumeration of  $K_1^+$ . Further, denote the same shorthand  $\lambda_i = a^{-1}v_{j_i}w_{j_i}$  and  $u_1 = (v_{j_1}, \dots, v_{j_l})$  but set  $\mu = \frac{3}{2}(q_1^3 + \gamma^2 - \gamma^2q_1^3)^{-1}$  and  $D_i = -\mu(1 - \gamma)^2q_1^2v_{j_i}^2/(a\delta^2\lambda_i)$ . Then, the Hessian of  $\mathcal{L}^\gamma$  restricted to  $(w_{j_1}, \dots, w_{j_l})$  is  $H = \frac{1}{\mu}(D + u_1u_1^T)$  with determinant

$$\det(H) = \mu^{-l}(1 + u_1^T D^{-1}u_1) \det(D) = -\mu^{-l} \det(D) \left( \frac{a\delta^2}{\mu(1 - \gamma)^2q_1^2} \sum_{i=1}^l \lambda_i - 1 \right). \tag{3.46}$$

By construction,  $\sum_{i=1}^l \lambda_i \leq 1$  and, by plugging in the formulas for  $a$ ,  $q_1$ , and  $\delta$  from Lemma 3.3,

$$\frac{a\delta^2}{\mu(1 - \gamma)^2q_1^2} = \frac{2}{3} \frac{\sqrt{1 + \gamma} + \sqrt{\gamma}}{(1 - \gamma)^2\sqrt{1 + \gamma}} (1 + \sqrt{\gamma}\delta^3 - \gamma^2) = \frac{2}{3} \frac{1}{(1 - \gamma)^2} + \mathcal{O}(\sqrt{\gamma}). \tag{3.47}$$

In particular,  $\det(H) < 0$  for small  $\gamma$ . □

### 3.3 Classification for Leaky ReLU Activation

In the following, we state the classification of critical points of the  $L^2$ -loss for leaky ReLU networks. It is almost analogous to Theorem 2.4, but the main difference is the absence of non-global local minima. These critical points vanish for leaky ReLU because they were caused solely by dead ReLU neurons.

**Theorem 3.5** *Let  $N \in \mathbb{N}$ ,  $\phi = (w, b, v, c) \in \mathbb{R}^{3N+1}$ ,  $\mathcal{A} = (\alpha, \beta) \in \mathbb{R}^2$ , and  $T = (T_0, T_1) \in \mathbb{R}^2$  satisfy  $\alpha \neq 0$  and  $0 \leq T_0 < T_1$ . Then, there exists  $\gamma_0 \in (0, 1]$  such that for all  $\gamma \in (0, \gamma_0)$  the following hold:*

- (I)  $\phi$  is not a local maximum of  $\mathcal{L}_{N,T,\mathcal{A}}^\gamma$ .
- (II) If  $\phi$  is a critical point or a local extremum of  $\mathcal{L}_{N,T,\mathcal{A}}^\gamma$ , then  $\mathcal{L}_{N,T,\mathcal{A}}^\gamma$  is differentiable at  $\phi$  with gradient  $\nabla \mathcal{L}_{N,T,\mathcal{A}}^\gamma(\phi) = 0$ .
- (III)  $\phi$  is not a non-global local minimum of  $\mathcal{L}_{N,T,\mathcal{A}}^\gamma$ .
- (IV)  $\phi$  is a saddle point of  $\mathcal{L}_{N,T,\mathcal{A}}^\gamma$  if and only if  $\phi$  is  $(T, \mathcal{A})$ -centered, for all  $j \in \{1, \dots, N\}$  the  $j$ th hidden neuron of  $\phi$  is flat semi-active, flat inactive with  $w_j = 0$ , flat degenerate, or type-2-active, and exactly one of the following two items holds:
  - (a)  $\phi$  does not have any type-2-active neurons.
  - (b) There exist  $\sigma \in \{-1, 1\}$ ,  $n \in \mathbb{N}$  such that if  $\delta = \gamma^{(1-\sigma)/4} + \gamma^{(1-\sigma(-1)^n)/4} + (n-1)\sqrt{1+\gamma}$  and  $q_i = T_0 + \frac{T_1-T_0}{\delta}(\gamma^{(1-\sigma)/4} + (i-1)\sqrt{1+\gamma})$ ,  $i \in \{1, \dots, n\}$ , then  $\bigcup_{j \in \{1, \dots, N\}, w_j \neq 0} \{-\frac{b_j}{w_j}\} = \{q_1, \dots, q_n\}$  and, for all  $j \in \{1, \dots, N\}$ ,  $i \in \{1, \dots, n\}$  with  $w_j \neq 0 = b_j + w_j q_i$ , it holds that  $\text{sign}(w_j) = \sigma(-1)^{i+1}$  and

$$\sum_{\substack{k \in \{1, \dots, N\}, \\ w_k \neq 0 = b_k + w_k q_i}} v_k w_k = \begin{cases} \frac{\alpha}{\sqrt{\gamma}} & \text{if } i = 1 = n, \\ \frac{\alpha}{\delta} \left( \frac{1}{\sqrt{1+\gamma}} + \frac{1}{\gamma^{(1-\sigma)/4}} \right) & \text{if } i = 1 \neq n, \\ \frac{\alpha}{\delta} \frac{2}{\sqrt{1+\gamma}} & \text{if } 2 \leq i \leq n-1, \\ \frac{\alpha}{\delta} \left( \frac{1}{\sqrt{1+\gamma}} + \frac{1}{\gamma^{(1-\sigma(-1)^n)/4}} \right) & \text{if } i = n \neq 1. \end{cases} \tag{3.48}$$

- (V) If  $\phi$  is a saddle point of  $\mathcal{L}_{N,T,\mathcal{A}}^\gamma$  without type-2-active neurons, then  $f_\phi^\gamma(x) = \frac{\alpha}{2}(T_0 + T_1) + \beta$  for all  $x \in [T_0, T_1]$ .
- (VI) If  $\phi$  is a saddle point of  $\mathcal{L}_{N,T,\mathcal{A}}^\gamma$  with at least one type-2-active neuron, then there exist  $\sigma \in \{-1, 1\}$ ,  $n \in \mathbb{N}$  such that  $n \leq N$  and, for all  $i \in \{0, \dots, n\}$ ,  $x \in [q_i, q_{i+1}]$ , one has

$$\begin{aligned}
 f_\phi^\gamma(x) - \alpha x - \beta &= \left[ \frac{-\sigma(-1)^i(1-\gamma)\alpha}{\delta} \right] \\
 \times \begin{cases} \frac{x-T_0}{\gamma^{(1-\sigma)/4}} - \frac{T_1-T_0}{2\delta} & \text{if } i = 0, \\ \frac{x-T_0}{\sqrt{1+\gamma}} - \frac{(i-1/2)(T_1-T_0)}{\delta} - \frac{\gamma^{(1-\sigma)/4}(T_1-T_0)}{\delta\sqrt{1+\gamma}} & \text{if } 1 \leq i \leq n-1, \\ \frac{x-T_0}{\gamma^{(1-\sigma(-1)^n)/4}} + \frac{T_1-T_0}{2\delta} - \frac{T_1-T_0}{\gamma^{(1-\sigma(-1)^n)/4}} & \text{if } i = n, \end{cases}
 \end{aligned}
 \tag{3.49}$$

where  $\delta$  and  $q_1, \dots, q_n$  are the same as in item (IV.b).

**Proof** We prove Theorem 3.5 in the special case  $\mathcal{A} = (1, 0)$  and  $T = (0, 1)$ . The general case follows from this the same way as Theorem 2.4 followed from Proposition 2.25 in Sect. 2.9. The first item is shown in Lemma 2.8; see Remark 2.9.

Suppose  $\phi$  is a critical point or a local extremum of  $\mathcal{L}^\gamma$  but not a global minimum. By Lemma 3.2, all neurons of  $\phi$  are flat semi-active, flat inactive with  $w_j = 0$ , degenerate, or type-2-active. If, in addition,  $\phi$  does not have any type-2-active neurons, then it also does not have any non-flat degenerate neurons, it is a saddle point, and  $\phi$  must be centered since  $\frac{\partial}{\partial c} \mathcal{L}^\gamma(\phi) = 0$ . If, on the other hand,  $\phi$  has a type-2-active neuron, then  $\phi$  is as in item (IV.b) by Lemma 3.3 apart from potentially having non-flat degenerate neurons, and  $\phi$  is a saddle point by Lemma 3.4. However, a posteriori,  $\phi$  cannot have non-flat degenerate neurons because, by Lemma 3.3.(vi),

$$\int_0^1 x(f_\phi^\gamma(x) - x)dx = -\frac{(1-\gamma)^2}{12\delta^4} < 0,
 \tag{3.50}$$

so  $\frac{\partial^+}{\partial w_j} \mathcal{L}^\gamma(\phi)$  could not be zero for a non-flat degenerate neuron  $j$ . This proves item (III) and the ‘only if’ part in item (IV). This also implies that any critical point or local extremum of  $\mathcal{L}^\gamma$  is a global minimum or does not have any non-flat degenerate neurons. Hence, the relation  $\mathcal{L}^\gamma = \mathcal{L} \circ P$  with the smooth map  $P$  and the differentiability properties of  $\mathcal{L}$  assert item (II).

If  $\phi$  is as in item (IV.a), then it clearly is a critical point of  $\mathcal{L}^\gamma$ , and it is a saddle point by Lemma 3.2. If  $\phi$  is as in item (IV.b), then  $f_\phi^\gamma$  is given by the formula in item (VI). We can calculate  $\int_{q_i}^{q_{i+1}} (f_\phi^\gamma(x) - x)dx = 0$  for all  $i \in \{0, \dots, n\}$  and

$$\int_{q_i}^1 x(f_\phi^\gamma(x) - x)dx + \gamma^{\sigma(-1)^{i+1}} \int_0^{q_i} x(f_\phi^\gamma(x) - x)dx = 0
 \tag{3.51}$$

for all  $i \in \{1, \dots, n\}$ . It follows from this that  $\phi$  is a critical point of  $\mathcal{L}^\gamma$ , and it is a saddle point by Lemma 3.4. This proves the ‘if’ part in item (IV). Item (V) is immediate and the last item was implicit in the previous step.  $\square$

**Remark 3.6** The restriction on  $\gamma$  to lie in  $(0, \gamma_0)$  is only needed in the proof of Lemma 3.4. All other proofs were carried out for general  $\gamma \in (0, 1)$ . We believe that, in fact, one can take  $\gamma_0 = 1$  in Lemma 3.4 and, hence, that Theorem 3.5 also holds for general  $\gamma \in (0, 1)$ .



### 4 Classification for Quadratic Activation

As the last case, we consider the quadratic activation function. The realization  $f_\phi^{\text{quad}} \in C(\mathbb{R}, \mathbb{R})$  of a network  $\phi = (w, b, v, c) \in \mathbb{R}^{3N+1}$  with the quadratic activation is

$$f_\phi^{\text{quad}}(x) = c + \sum_{j=1}^N v_j(w_jx + b_j)^2. \tag{4.1}$$

Given  $\mathcal{A} = (\alpha, \beta) \in \mathbb{R}^2$  and  $T = (T_0, T_1) \in \mathbb{R}^2$ , the loss function  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}$  is the  $L^2$ -loss given by

$$\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi) = \int_{T_0}^{T_1} (f_\phi^{\text{quad}}(x) - \alpha x - \beta)^2 dx. \tag{4.2}$$

This time, there are no issues with differentiability since  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}$  is infinitely times differentiable, even analytic, everywhere. The classification turns out to be simpler than in the ReLU and leaky ReLU case as there are no local extrema and only saddle points with a constant realization function.

**Theorem 4.1** *Let  $N \in \mathbb{N}$ ,  $\phi = (w, b, v, c) \in \mathbb{R}^{3N+1}$ ,  $\mathcal{A} = (\alpha, \beta) \in \mathbb{R}^2$ , and  $T = (T_0, T_1) \in \mathbb{R}^2$  satisfy  $\alpha \neq 0$  and  $T_0 < T_1$ . Then, the following hold:*

- (I)  $\phi$  is not a local maximum of  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}$ .
- (II)  $\phi$  is not a non-global local minimum of  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}$ .
- (III)  $\phi$  is a global minimum of  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}$  if and only if  $N \geq 2$  and  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi) = 0$ .
- (IV)  $\phi$  is a saddle point of  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}$  if and only if  $\phi$  is  $(T, \mathcal{A})$ -centered and, for all  $j \in \{1, \dots, N\}$ , the  $j$ th hidden neuron of  $\phi$  satisfies  $v_j b_j = 0 = w_j$  or  $w_j \neq v_j = 0 = b_j + \frac{1}{2}(T_0 + T_1)w_j$ .
- (V) If  $\phi$  is a saddle point of  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}$ , then  $f_\phi^{\text{quad}}(x) = \frac{\alpha}{2}(T_0 + T_1) + \beta$  for all  $x \in [T_0, T_1]$ .

**Proof** As for the other activation functions, the first item is shown in Lemma 2.8; see Remark 2.9. Now, suppose  $\phi$  is a critical point of  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}$  and  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi) > 0$ . Since  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}$  is smooth, we have, for any  $j \in \{1, \dots, N\}$ ,

$$\begin{aligned} 0 &= \frac{\partial}{\partial w_j} \mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi) = 4v_j \int_{T_0}^{T_1} x(w_jx + b_j)(f_\phi^{\text{quad}}(x) - \alpha x - \beta) dx, \\ 0 &= \frac{\partial}{\partial b_j} \mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi) = 4v_j \int_{T_0}^{T_1} (w_jx + b_j)(f_\phi^{\text{quad}}(x) - \alpha x - \beta) dx, \\ 0 &= \frac{\partial}{\partial v_j} \mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi) = 2 \int_{T_0}^{T_1} (w_jx + b_j)^2 (f_\phi^{\text{quad}}(x) - \alpha x - \beta) dx, \\ 0 &= \frac{\partial}{\partial c} \mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi) = 2 \int_{T_0}^{T_1} (f_\phi^{\text{quad}}(x) - \alpha x - \beta) dx. \end{aligned} \tag{4.3}$$

Thus, if there exists  $j \in \{1, \dots, N\}$  with  $v_j \neq 0 \neq w_j$ , then  $\int_{T_0}^{T_1} x^m (f_\phi^{\text{quad}}(x) - \alpha x - \beta) dx = 0$  for all  $m \in \{0, 1, 2\}$ . However, note that the zero polynomial is the only polynomial  $p$  of degree at most two satisfying  $\int_{T_0}^{T_1} x^m p(x) dx = 0$  for all  $m \in \{0, 1, 2\}$ . Hence, since  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi) > 0$ , we must have  $v_j = 0$  or  $w_j = 0$  for all neurons. In particular,  $f_\phi^{\text{quad}}$  is constant and  $\int_{T_0}^{T_1} x (f_\phi^{\text{quad}}(x) - \alpha x - \beta) dx \neq 0$ . Thus, for all  $j$ , if  $v_j \neq 0$ , then  $b_j = 0$ . So far, we have shown that all neurons must satisfy  $v_j = 0$  or  $w_j = 0 = b_j$ . It follows that  $\phi$  is  $(T, \mathcal{A})$ -centered. For a neuron  $j$  with  $w_j \neq 0$  and  $t_j = -b_j/w_j$ , we have

$$0 = 2 \int_{T_0}^{T_1} (w_j x + b_j)^2 (c - \alpha x - \beta) dx = -2\alpha w_j^2 \int_{T_0}^{T_1} (x - t_j)^2 (x - \frac{T_0+T_1}{2}) dx, \tag{4.4}$$

which is true if and only if  $t_j = (T_0 + T_1)/2$ . This proves the ‘only if’ direction in (IV). Next, we show that  $\phi$  must be a saddle point. We will pick a path  $\phi_s = (w^s, b^s, v^s, c^s)$ ,  $s \in (-1, 1)$ , through  $\phi = \phi_0$ , which differs only in the coordinates of the first neuron and in

$$c^s = c - v_1^s (b_1^s)^2 - \frac{1}{3} A_s (T_0^2 + T_0 T_1 + T_1^2) - B_s (T_0 + T_1), \tag{4.5}$$

where  $A_s = v_1^s (w_1^s)^2$  and  $B_s = v_1^s w_1^s b_1^s$ . Then,

$$\begin{aligned} \frac{\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi_s) - \mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi_0)}{(T_1 - T_0)^3} &= \frac{1}{45} A_s^2 (4T_0^2 + 7T_0 T_1 + 4T_1^2) + \frac{1}{3} A_s B_s (T_0 + T_1) \\ &\quad + \frac{1}{3} B_s^2 - \frac{\alpha}{6} (A_s (T_0 + T_1) + 2B_s). \end{aligned} \tag{4.6}$$

We distinguish three cases. First, if  $v_1 = 0 \neq w_1$ , then we use  $w_1^s = w_1, b_1^s = b_1 - s w_1$ , and  $v_1^s = -\text{sign}(\alpha) s^2$ . In this case,  $B_s = -\frac{1}{2} A_s (T_0 + T_1) - s A_s$  and, hence,

$$\frac{\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi_s) - \mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi_0)}{(T_1 - T_0)^3} = -\frac{|\alpha|}{3} w_1^2 s^3 + \mathcal{O}(s^4). \tag{4.7}$$

This is strictly negative for sufficiently small  $s > 0$ , so  $\phi$  is a saddle point. Secondly, if  $v_1 \neq 0 = w_1$ , then we use  $w_1^s = s, b_1^s = -\frac{1}{2} (T_0 + T_1) s + \text{sign}(\alpha v_1) s^2$ , and  $v_1^s = v_1$ . In this case,

$$\frac{\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi_s) - \mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi_0)}{(T_1 - T_0)^3} = -\frac{|\alpha|}{3} |v_1| s^3 + \mathcal{O}(s^4). \tag{4.8}$$

In the last case, namely  $v_1 = 0 = w_1$ , we use  $w_1^s = sb_1^s$ ,  $b_1^s = b_1 + s$ , and  $v_1^s = \text{sign}(\alpha)s^3(b_1^s)^{-2}$ . Then,

$$\frac{\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi_s) - \mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi_0)}{T_1 - T_0} = -\frac{|\alpha|}{3}s^4 + \mathcal{O}(s^5). \tag{4.9}$$

We have shown that if  $\phi$  is a critical point with  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi) > 0$ , then it is a saddle point. This establishes item (II) and it also implies that if  $\phi$  is a global minimum, then  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi) = 0$ . The latter is only possible if  $N \geq 2$ . Conversely, if  $N \geq 2$ , then there are networks with zero loss, so item (III) holds. If  $\phi$  is  $(T, \mathcal{A})$ -centered and all of its neurons are as in item (IV), then  $\nabla \mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi) = 0$  and  $\phi$  is a saddle point since clearly  $\mathcal{L}_{N,T,\mathcal{A}}^{\text{quad}}(\phi) > 0$ . This finishes (IV), and (V) follows.  $\square$

The conditions in Theorem 4.1.(IV) are equivalent to all neurons being flat semi-active, flat inactive with  $w_j = 0$ , flat type-2-active with breakpoint  $-b_j/w_j = (T_0 + T_1)/2$ , or degenerate. However, for the quadratic activation, the notions of in-/active neurons seem no longer appropriate.

**Remark 4.2** In Theorem 4.1, the case  $N = 1$  of a single neuron is special due to the absence of global minima. The loss can still be arbitrarily small, but there is no network achieving the infimum. Indeed, for all  $(w, b) \in \mathbb{R}^2$  with  $w \neq 0$ ,

$$\inf_{(v,c) \in \mathbb{R}^2} \mathcal{L}_{1,T,\mathcal{A}}^{\text{quad}}(w, b, v, c) = \frac{1}{12}\alpha^2(T_1 - T_0)^3 \times \left( 1 - \frac{60 \left( \frac{T_0+T_1}{2} + \frac{b}{w} \right)^2}{(T_1 - T_0)^2 + 60 \left( \frac{T_0+T_1}{2} + \frac{b}{w} \right)^2} \right) \xrightarrow[\text{monotone}]{\frac{T_0+T_1}{2} + \frac{b}{w} \rightarrow \pm\infty} 0. \tag{4.10}$$

**Acknowledgements** The second author acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy EXC 2044-390685587, Mathematics Muenster: Dynamics-Geometry-Structure and by the startup fund project of Shenzhen Research Institute of Big Data under grant No. T00120220001.

**Funding** Open access funding provided by Swiss Federal Institute of Technology Zurich.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Allen-Zhu, Z., Li, Y., Song, Z.: A convergence theory for deep learning via over-parameterization. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR, , vol. 97, pp. 242–252 (2019)
- Baldi, P., Hornik, K.: Neural networks and principal component analysis: Learning from examples without local minima. *Neural Netw.* **2**(1), 53–58 (1989)
- Cheridito, P., Jentzen, A., Riekert, A., Rossmannek, F.: A proof of convergence for gradient descent in the training of artificial neural networks for constant target functions. *J. Complexity.* **72**, 101646 (2022). <https://doi.org/10.1016/j.jco.2022.101646>
- Chizat, L., Bach, F.: Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In: Abernethy, J., Agarwal, S. (eds.) Proceedings of Thirty Third Conference on Learning Theory. Proceedings of Machine Learning Research, PMLR, vol. 125, pp. 1305–1338 (2020)
- Chizat, L., Oyallon, E., Bach, F.: On lazy training in differentiable programming. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 2937–2947. Curran Associates Inc. (2019)
- Choromanska, A., Henaff, M., Mathieu, M., Ben Arous, G., LeCun, Y.: The Loss Surfaces of Multilayer Networks. In: Lebanon, G., Vishwanathan, S.V.N. (eds.) Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, PMLR, vol. 38, pp. 192–204 (2015)
- Choromanska, A., LeCun, Y., Ben Arous, G.: Open problem: the landscape of the loss surfaces of multilayer networks. In: Günwald, P., Hazan, E., Kale, S. (eds.) Proceedings of the 28th Conference on Learning Theory. Proceedings of Machine Learning Research, PMLR, vol. 40, pp. 1756–1760 (2015)
- Dauphin, Y.N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., Bengio, Y.: Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 27, pp. 2933–2941. Curran Associates, Inc. (2014)
- Du, S., Lee, J.: On the power of over-parametrization in neural networks with quadratic activation. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning (Stockholmsmässan, Stockholm Sweden). Proceedings of Machine Learning Research, PMLR, vol. 80, pp. 1329–1338 (2018)
- Eberle, S., Jentzen, A., Riekert, A., Weiss, G.S.: Existence, uniqueness, and convergence rates for gradient flows in the training of artificial neural networks with ReLU activation. [arXiv:2108.08106v1](https://arxiv.org/abs/2108.08106v1) (2021)
- Fehrman, B., Gess, B., Jentzen, A.: Convergence rates for the stochastic gradient descent method for non-convex objective functions. *J. Mach. Learn. Res.* **21**(136), 1–48 (2020)
- Fukumizu, K., Amari, S.: Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Netw.* **13**(3), 317–327 (2000)
- Jentzen, A., Riekert, A.: Convergence analysis for gradient flows in the training of artificial neural networks with ReLU activation. [arXiv:2107.04479v1](https://arxiv.org/abs/2107.04479v1) (2021)
- Kawaguchi, K.: Deep learning without poor local minima. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 29, pp. 586–594. Curran Associates, Inc. (2016)
- LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015)
- Lee, J.D., Simchowitz, M., Jordan, M.I., Recht, B.: Gradient descent only converges to minimizers. In: Feldman, V., Rakhlin, A., Shamir, O. (eds.) 29th Annual Conference on Learning Theory. Proceedings of Machine Learning Research, PMLR, vol. 49, pp. 1246–1257 (2016)
- Lee, J.D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M.I., Recht, B.: First-order methods almost always avoid strict saddle points. *Math. Program.* **176**(1), 311–337 (2019)
- Livni, R., Shalev-Shwartz, S., Shamir, O.: On the computational efficiency of training neural networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 27, pp. 855–863. Curran Associates, Inc. (2014)
- Sarao Mannelli, S., Vanden-Eijnden, E., Zdeborová, L.: Optimization and generalization of shallow neural networks with quadratic activation functions. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) In Proceedings NEURIPS2020\_9b8b50fb. Advances in Neural Information Processing Systems, vol. 33, pp. 13445–13455 (2020)

- Nguyen, Q., Hein, M.: The loss surface of deep and wide neural networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR, vol. 70, pp. 2603–2612 (2017)
- Panageas, I., Piliouras, G.: Gradient Descent Only Converges to Minimizers: Non-Isolated Critical Points and Invariant Regions. In: Papadimitriou, C.H. (ed.) 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Leibniz International Proceedings in Informatics (LIPIcs), vol. 67, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, pp. 2:1–2:12 (2017)
- Pennington, J., Bahri, Y.: Geometry of neural network loss surfaces via random matrix theory. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR, vol. 70, pp. 2798–2806 (2017)
- Petersen, P., Raslan, M., Voigtlaender, F.: Topological properties of the set of functions generated by neural networks of fixed size. *Foundations of Computational Mathematics* (2020)
- Safran, I., Shamir, O.: On the quality of the initial basin in overspecified neural networks. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of the 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR, vol. 48, pp. 774–782 (2016)
- Safran, I., Shamir, O.: Spurious local minima are common in two-layer ReLU neural networks. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR, vol. 80, pp. 4433–4441 (2018)
- Soltanolkotabi, M., Javanmard, A., Lee, J.D.: Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Trans. Inf. Theory* **65**(2), 742–769 (2019)
- Soudry, D., Carmon, Y.: No bad local minima: Data independent training error guarantees for multilayer neural networks. [arXiv:1605.08361v2](https://arxiv.org/abs/1605.08361v2) (2016)
- Soudry, D., Hoffer, E.: Exponentially vanishing sub-optimal local minima in multilayer neural networks. [arXiv:1702.05777v5](https://arxiv.org/abs/1702.05777v5) (2017)
- Venturi, L., Bandeira, A.S., Bruna, J.: Spurious valleys in one-hidden-layer neural network optimization landscapes. *J. Mach. Learn. Res.* **20**(133), 1–34 (2019)
- Wojtowytsch, S.: On the convergence of gradient descent training for two-layer relu-networks in the mean field regime. [arXiv:2005.13530v1](https://arxiv.org/abs/2005.13530v1) (2020)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.