Institut für Parallele und Verteilte Systeme

Abteilung Simulation Software Engineering

Universität Stuttgart
Universitätsstraße 38
70569 Stuttgart

Masterarbeit

Convergence Analysis of Neural Networks

David Holzmüller

Studiengang: M. Sc. Informatik

Prüfer: Prof. Dr. Dirk Pflüger

Betreuer: Prof. Dr. Ingo Steinwart

begonnen am: 15. Mai 2019

beendet am: 14. November 2019

Abstract

We prove that two-layer (Leaky)ReLU networks with one-dimensional input and output trained using gradient descent on a least-squares loss and He et al. [20] initialization are not universally consistent. Specifically, we define a submanifold of all data distributions on which gradient descent fails to spread the nonlinearities across the data with high probability, i.e. it only finds a bad local minimum or valley of the optimization landscape. In these cases, the network found by gradient descent essentially only performs linear regression. We provide numerical evidence that this happens in practical situations and that stochastic gradient descent exhibits similar behavior. We relate the speed of convergence to such a local optimum to a stable linear system whose eigenvalues have different asymptotics. We also provide an upper bound on the learning rate based on this observation. While we mainly operate in the underparameterized regime like most consistency results for classical algorithms, our proof also applies to certain overparameterized cases that are not covered by recent results showing convergence of overparameterized neural nets to a global optimum.

Zusammenfassung

Diese Masterarbeit beweist, dass zweischichtige neuronale Netze mit ReLU- oder LeakyReLU-Aktivierungsfunktionen und einem Input- und Output-Neuron, die mit Gradientenabstieg auf einem Least-Squares-Loss optimiert werden, nicht universell konsistent sind, solange das weit verbreitete Initialisierungsverfahren von He et al. [20] verwendet wird. Speziell beruht der Beweis auf der Beobachtung, dass ein solches neuronales Netz bei der Initialisierung nur in $x = 0$ nichtlinear ist und es sich unter gewissen Voraussetzungen wie ein lineares Regressionsverfahren verhält, anstatt auf den Daten hinreichend nichtlinear zu werden. Die Existenz entsprechender lokaler Minima wurde bereits von Yun et al. [39] gezeigt. Die Neuerung dieser Arbeit besteht darin, die Dynamik des Gradientenabstiegs zu untersuchen, um unter gewissen Voraussetzungen zu zeigen, dass diese lokalen Minima mit hoher Wahrscheinlichkeit erreicht werden.

Die Beweisidee besteht darin, dass sich die Änderung der Netzgewichte, also die Größe der Komponenten des Gradienten, in etwa proportional zur Abweichung der Netzfunktion von den optimalen linearen Regressionsgeraden verhält. Eine schnelle Konvergenz gegen die optimalen linearen Regressionsgeraden bedeutet, dass sich die Netzgewichte während des Trainings nur wenig ändern und sich daher die Nichtlinearitäten nicht weit vom Ursprung $x = 0$ entfernen können. Falls die optimalen linearen Regressionsgeraden annähernd durch den Ursprung $(0, 0)$ verlaufen, dann zeigt diese Arbeit, dass die Netzfunktion mit hoher Wahrscheinlichkeit schnell gegen ein solches Optimum konvergiert und daher nur ein suboptimales lokales Minimum findet. Tatsächlich wird bewiesen, dass sich diese Wahrscheinlichkeit wie $1 - O(n^{-\gamma})$ für alle $\gamma < 1/2$ verhält, wobei n die Anzahl der Neuronen in der verborgenen Schicht ist. Es wird gezeigt, dass sich die Konvergenz des Netzes wie ein vierdimensionales diskretes lineares System verhält. Die Annahme an die optimalen linearen Regressionsgeraden stellt sicher, dass die Initialisierung nahe an den stark negativen Eigenwerten liegt, sodass die Konvergenz schnell genug ist. Die in Abschnitt 5 eingeführte Theorie mündet in Abschnitt 5.5 in mehrere Inkonsistenzresultate. Die Eigenwertanalyse liefert eine effizient berechenbare obere Schranke an die Optimierungsschrittweite h, unterhalb derer das inkonsistente Verhalten auftritt.

Mithilfe der vorgestellten Theorie wird in Abschnitt 6 durch Monte-Carlo-Experimente die Wahrscheinlichkeit abgeschätzt, dass sich im obigen Szenario die Nichtlinearitäten über die Datenpunkte hinweg verteilen. Die Experimente bestätigen die theoretische Rate $O(n^{-\gamma})$ für alle $\gamma < 1/2$ und zeigen, dass diese Asymptotik bereits für realistische Netzgrößen eintritt. Sie zeigen auch ein analoges Verhalten von stochastischem Gradientenabstieg. Darüber hinaus wird diskutiert, inwieweit dieses Ergebnis auf ähnliche Initialisierungsverfahren zutrifft.

Das vorgestellte Inkonsistenzresultat steht im Kontrast zu anderen Arbeiten, die universelle Konsistenz unter der unrealistischen Annahme einer perfekten Optimierung [36, 13] oder Konvergenz gegen ein globales Optimum für gewisse überparametrisierte Netze [22, 11, 8, 1] zeigen.

Contents

1 Introduction

A learning method is called universally consistent if, regardless of which probability distribution the training data is sampled from, the learning method converges to some desired quantity of the distribution as the number of samples goes to infinity. For classical (nonparametric) learning methods such as histogram rules, kernel regression and k-nearest neighbor rules, universal consistency has already been shown in the late 20th century, as outlined in [7, 19]. In Deep Learning, the situation is more difficult. For certain classes of neural networks, universal consistency has also been shown for regression by White [36] and for classification by Faragó and Lugosi [13], for an overview we refer to [19]. However, these results depend on finding a global optimum of the neural network, which has been shown to be NP-hard for certain neural networks by Blum and Rivest [4]. While this does not imply that practical optimization algorithms for neural networks are inconsistent, it poses the question whether such consistency can be achieved using optimization algorithms such as (stochastic) gradient descent. Mücke and Steinwart [25] show that consistency for neural networks can be achieved in a computationally feasible way by imitating histograms with neural networks. However, this does not correspond to practical uses of neural networks with gradient-based optimization methods [17].

In general, gradient-based methods can get stuck in local minima. It has been shown for many neural network scenarios that non-global minima exist [32, 18, 38, 28], but usually the probability of reaching bad local minima with specific optimizers is not investigated. For overparameterized networks, i.e. networks with more parameters than data points, there are results showing that, under some assumptions on the data distribution, (stochastic) gradient descent reaches a global optimum with high probability and there exist some generalization guarantees [22, 11, 8, 1, 2].

The (leaky) rectified linear unit activation function, which is often abbreviated by (Leaky)ReLU, is a very popular activation function in Deep Learning [17]. Yun et al. [39] showed that (Leaky)ReLU networks with least-squares loss have non-strict spurious local minima where the neural network function corresponds to the linear regression optimum on the data points. Here, we show that under certain conditions, the probability of gradient descent getting stuck in such a minimum converges to one as the number of hidden neurons goes to infinity.

1.1 Contribution

In this thesis, we prove that optimizing two-layer ReLU or LeakyReLU networks with one-dimensional input and output using gradient descent on a least-squares loss does not yield an universally consistent estimator if the common initialization method by He et al. [20] is used. To this end, we show that under certain assumptions on the dataset, the nonlinearities (kinks) of the neural network properly spread across the dataset with probability $O(n^{-\gamma})$ for all $\gamma < 1/2$, where n is the number of hidden neurons. We present bounds on the speed of convergence and prove that the weights of the neural network only change little over training, similar to what has been proven for certain overparameterized networks [22, 11, 8, 1]. However, in our case, the net does not converge to a global optimum but to a "linear regression optimum" which is only a

non-strict local (non-global) minimum of the loss function (cf. [39]). We also provide an explicit bound on the step size for gradient descent (i.e. the learning rate) such that our result holds. This bound can be computed for a given dataset and initialization and behaves asymptotically like $\Theta(1/n)$ with high probability. Like Du et al. [11], we find that the evolution of the network function behaves similarly to a linear time-invariant system. However, a central finding here is that the eigenvalues of this system have different asymptotic behavior, which relates to the choice of the step size and also to conditions on the dataset and the initialization.

Monte Carlo experiments show that the proven phenomenon is present for practical numbers $n \in \{16, 32, \dots, 2048\}$ of hidden neurons and that stochastic gradient descent exhibits similar behavior. Moreover, we argue using Monte Carlo experiments and heuristic calculations that the initialization variances can be scaled differently such that the presented problem is unlikely to occur. The code for these experiments can be found at `https://github.com/dholzmueller/nn_inconsistency`.

1.2 Outline

The main theorem, intuition and proof ideas of this thesis are outlined in Section 2. In Section 3, we discuss related work and mention some similarities and differences to the present work. Section 4 introduces some notational conventions and known facts about matrices that will be relevant for parts of the thesis.

Section 5 contains the main theory: In Section 5.1, gradient descent equations are derived and reformulated in a fashion that is relevant for later stages of the proof. The gained insights are discussed in Section 5.2. Concentration inequalities proven in Appendix A then enter in Section 5.3 to obtain characteristics of data sampling and of the specific class of considered initialization methods. In Section 5.4, the reformulated gradient descent equations are analyzed asymptotically for large network and sample sizes using the concentration inequality results of the previous section and some more general results from Appendix B. Finally, the obtained bounds on gradient descent trajectories are leveraged in Section 5.5 to formulate different inconsistency results.

In Section 6, several Monte Carlo experiments on (stochastic) gradient descent are shown and the influence of the initialization method is discussed theoretically and experimentally. We conclude with remarks on interesting followup research questions in Section 7.

2 Overview

In this section, we will provide an overview over the main goal of the thesis, which is to prove an inconsistency result for a certain class of neural networks.

Definition 2.1. We consider a neural net with one input, one output, one hidden layer with n hidden neurons, and a LeakyReLU activation function

$$\varphi(x) \coloneqq \begin{cases} x & , x \geq 0 \\ \alpha x & , x \leq 0 \end{cases}$$

with fixed parameter $\alpha \in \mathbb{R}$. In the important special case $\alpha = 0$, we obtain the ReLU activation function. Such a net parameterizes a function $f_W : \mathbb{R} \to \mathbb{R}$ with parameters $W = (a, b, c, w) \in \mathbb{R}^{3n+1}$, where $a, b, w \in \mathbb{R}^n$ and $c \in \mathbb{R}$, via

$$f_W(x) = c + \sum_{i=1}^{n} w_i \varphi(a_i x + b_i) \ .$$

The parameters b, c are also called biases of the network.

Let $f : \mathbb{R} \to \mathbb{R}$ be a (measurable) function, let P be a probability distribution on $\mathbb{R} \times \mathbb{R}$ and let $D = ((x_1, y_1), \ldots, (x_N, y_N)) \in (\mathbb{R} \times \mathbb{R})^N$ be a dataset (with N data points). We define the (least-squares) risk of f with respect to P or D as

$$\begin{aligned} R_P(f) &\coloneqq \frac{1}{2} \mathbb{E}_{(x,y)\sim P} (y - f(x))^2 \\ R_D(f) &\coloneqq \frac{1}{2N} \sum_{j=1}^{N} (y_j - f(x_j))^2 \ . \end{aligned}$$

We define the loss of a parameter vector $W \in \mathbb{R}^{3n+1}$ with respect to P or D as

$$\begin{aligned} L_P(W) &\coloneqq R_P(f_W) \\ L_D(W) &\coloneqq R_D(f_W) \ . \end{aligned}$$ ◀

We can then optimize the neural net using gradient descent with step size $h > 0$ on W:

$$W_{k+1} \coloneqq W_k - h \nabla L_D(W_k) \ . \tag{2.1}$$

For $\alpha \neq 1$, L_D is not differentiable everywhere since φ is not differentiable in 0. In practice, this can be handled by defining $\varphi'(0) \coloneqq \alpha$ or $\varphi'(0) \coloneqq 1$ (cf. Goodfellow et al. [17], section 6.3). In the upcoming theory, the value of $\varphi'(0)$ will be irrelevant since we will show that with high probability, the inputs $a_i x + b_i$ to φ will not change their sign during training under certain conditions and hence will not be zero.

We will consider a generalized version of the initialization method by He et al. [20]. Applying the initialization method by He et al. [20], also called Kaiming initialization, to the network in Definition 2.1 means initializing the weights independently as

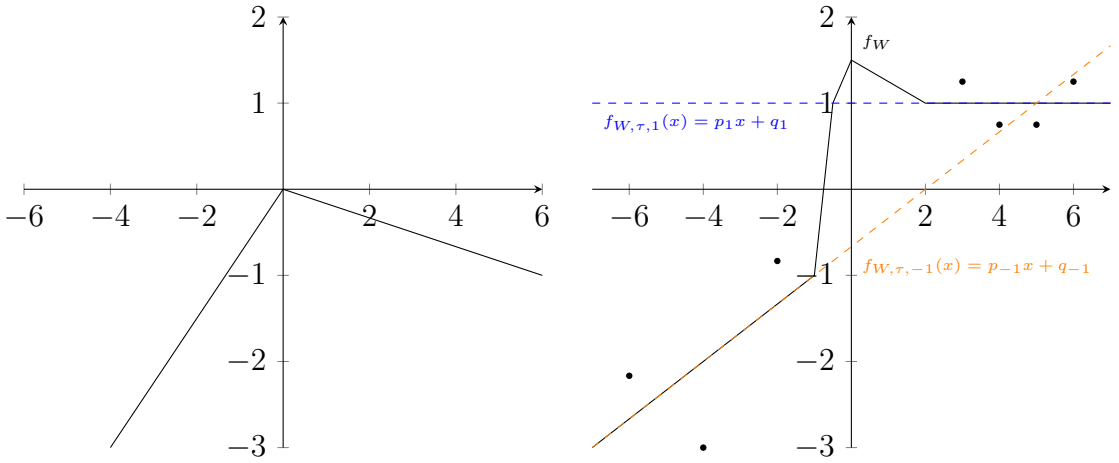
$$\begin{aligned} a_{i,0} &\sim \mathcal{N}(0, 2) \\ w_{i,0} &\sim \mathcal{N}(0, 2/n) \end{aligned}$$

$$b_{i,0} = 0$$
$$c_0 = 0 \ .$$

The constant 2 in the variances was specifically computed for ReLU activations by He et al. [20]. In order to account for differently chosen constants for LeakyReLU, we allow $\mathrm{Var}(a_{i,0}) = c_a$ and $\mathrm{Var}(w_{i,0}) = c_w/n$ for fixed $c_a, c_w > 0$ in our proofs.
Since φ is a continuous piecewise linear function, the functions f_W that can be represented by neural networks in Definition 2.1 are piecewise affine and continuous. A single hidden neuron i represents a function

$$x \mapsto \varphi(a_i x + b_i) \ ,$$

which is affine if $a_i = 0$. If $a_i \neq 0$, it has a non-differentiable point for $a_i x + b_i = 0$, i.e. at the point $x = -b_i/a_i$. The non-differentiable points of the overall function f_W are thus a subset of the points $\{-b_i/a_i \mid a_i \neq 0\}$, which we also call *kinks* or *knots* of the neural network [34]. Since we assume that $b_i = 0$ after the initialization of the network, all kinks are initially located at zero. This is displayed in Figure 1a. During gradient descent, these kinks may move around. To the right of the rightmost kink, the function f_W is affine. Under some assumptions, the affine continuation of this rightmost affine part of f_W is exactly the function $f_{W,\tau,1}$ that we will define later. Similarly, the affine continuation of the leftmost affine part of f_W will (under some assumptions) correspond to the function $f_{W,\tau,-1}$ defined later. A possible neural network and the affine continuations of both parts are shown in Figure 1b.



(a) Typical initial network.

(b) Typical trained network with (at least) four kinks.

Figure 1: Examples of untrained and trained network functions f_W.

Empirically, it can be observed that in some cases, the kinks only move a little so that they never reach the data points x_j. This means that on the set $\{(x_j, y_j) \mid x_j > 0\}$ of positive data points, the function f_W is affine and thus may not fit the data points well. The same holds true for the set $\{(x_j, y_j) \mid x_j < 0\}$ of negative data points. The goal of this thesis is to prove that for certain probability distributions P, the probability of sampling a dataset D and an initialization W_0 such that a kink crosses the data points

during gradient descent converges to zero as the number n of hidden neurons converges to infinity. This means that this setup for training of neural networks is inconsistent, i.e. the loss does not converge to the optimum achievable loss as $n, N \to \infty$. In our case, consistency means the following:

Definition 2.2 (Consistency, cf. Definition 6.4 in [35])**.** Let P be a probability distribution on $\mathbb{R} \times \mathbb{R}$ and let $(\mathcal{L}_N)_{N\in\mathbb{N}}$ with (measurable) $\mathcal{L}_N : (\mathbb{R} \times \mathbb{R})^N \to (\mathbb{R} \to \mathbb{R})$ be a sequence of estimators, i.e. given a dataset $D \in (\mathbb{R} \times \mathbb{R})^N$, $f_D \coloneqq \mathcal{L}_N(D) : \mathbb{R} \to \mathbb{R}$ is a regression function. We call $(\mathcal{L}_N)_{N\in\mathbb{N}}$ consistent for P (with respect to the least-squares loss), if for all $\varepsilon > 0$,

$$\lim_{N\to\infty} P^N \left(D \in (\mathbb{R} \times \mathbb{R})^N : R_P(f_D) \geq \inf_{f:\mathbb{R}\to\mathbb{R}} R_P(f) + \varepsilon \right) = 0 \, .$$

Here, P^N means that the components (x_j, y_j) of D are sampled independently from P. We call $(\mathcal{L}_N)_{N\in\mathbb{N}}$ universally consistent (with respect to the least-squares loss) if this holds for all probability distributions P on $\mathbb{R} \times \mathbb{R}$. We define consistency and universal consistency analogously if the estimators $\mathcal{L}_N$ are random instead of deterministic. ◀

We will prove several inconsistency results in Section 5.5 which imply the following result:

Theorem 2.3. *Let $(\mathcal{L}_N)_{N\in\mathbb{N}}$ be a sequence of estimators such that $\mathcal{L}_N$ takes a dataset $D \in (\mathbb{R} \times \mathbb{R})^N$ sampled randomly according to a probability distribution P on $\mathbb{R} \times \mathbb{R}$ and outputs a trained neural network with:*

- *one input neuron, one hidden layer with n_N neurons, and one output neuron such that $n_N = O(\sqrt{N})$ and $\lim_{N\to\infty} n_N = \infty$,*
- *ReLU or LeakyReLU activation function with $|\alpha| \neq 1$ applied to the hidden layer,*
- *gradient descent on the least-squares loss function with step size $0 < h_N = o\left(n_N^{-1}\right)$,*
- *an initialization method matching Definition 5.21, e.g. the one by He et al. [20],*
- *any stopping criterion.*

Then, for all probability distributions P that satisfy Assumption 5.16, $(\mathcal{L}_N)_{N\in\mathbb{N}}$ is not consistent. In particular, $(\mathcal{L}_N)_{N\in\mathbb{N}}$ is not universally consistent.

Proof. Let P satisfy Assumption 5.16. By Corollary 5.43, there exists $C > 0$ such that

$$R_P(\mathcal{L}_N(D)) \geq \inf_{f:\mathbb{R}\to\mathbb{R}} R_P(f) + C$$

with probability $1 - O(n_N^{-\gamma})$ for all $\gamma < 1/2$. Since $n_N \to \infty$ by assumption, this means that $(\mathcal{L}_N)_{N\in\mathbb{N}}$ is not consistent for P. Example 5.18 shows that such a probability distribution P exists. □

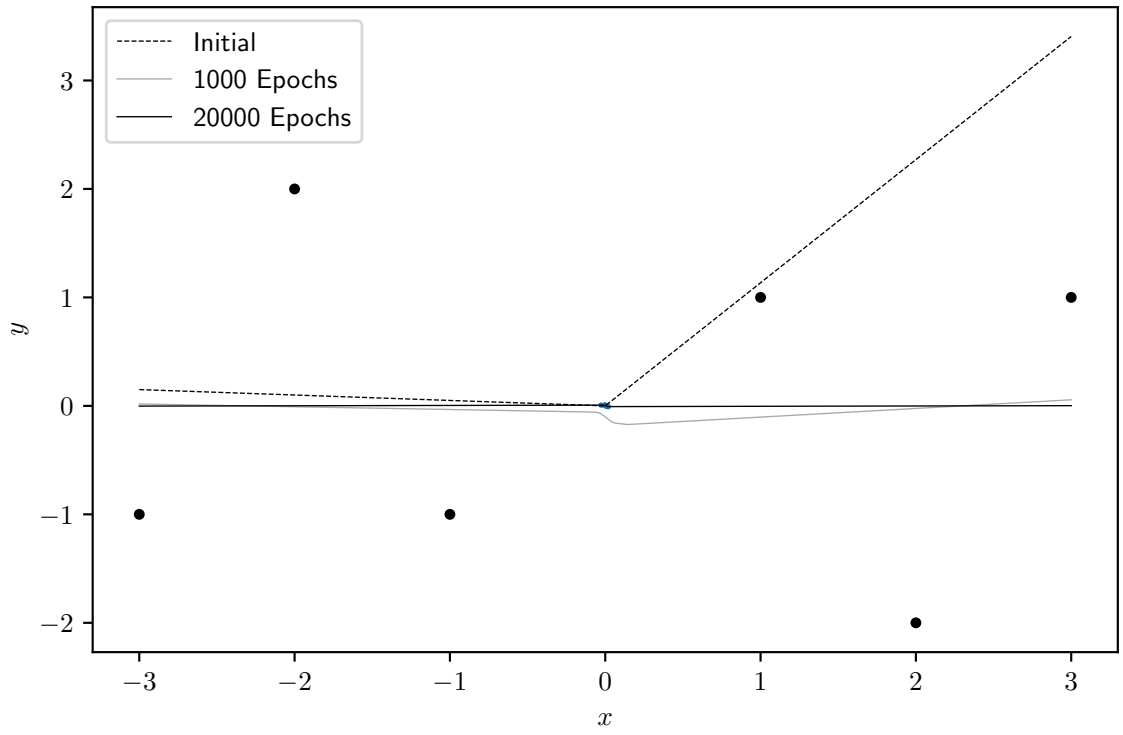
Remark 2.4. Assumption 5.16 essentially restricts P to a submanifold of the set of all probability distributions which still contains continuous distributions with observation noise. For most of these P, the condition $n_N \to \infty$ is necessary for consistency since the nets need to converge to the "optimal regression function", which is in general highly nonlinear. We show in Remark 5.38 that the condition $h = o\left(n_N^{-1}\right)$ can be weakened to $h \in (0, \tilde{C} n_N^{-1})$ for sufficiently small $\tilde{C} > 0$.
In Corollary 5.39, we prove a different version of Theorem 2.3, which shows that we can choose P as a symmetric uniform distribution on six data points such that

- $R_P(\mathcal{L}_N(D)) \geq 1$ with probability $1 - O(n^{-\gamma})$ for all $\gamma < 1/2$,
- for $n \geq 5$, there exists $W \in \mathbb{R}^{3n+1}$ with $R_P(f_W) = 0$.

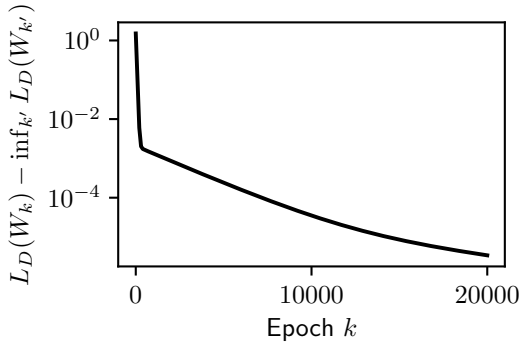
The computations in our proofs could be simplified by instead investigating probability distributions P such that all x_j are positive. This would however exclude symmetric and normalized distributions. These simplifications are discussed in Remark 5.14 and Remark 5.44. ◀

The main idea behind the proof is to show that the affine functions $f_{W,\tau,1}$ and $f_{W,\tau,-1}$ converge so quickly towards the optimal linear regression lines on the positive and negative parts of the dataset that the kinks do not have enough "time" to reach the data points. An example is shown in Figure 2. In this example, the dataset is chosen such that the optimal linear regression lines for both parts of the dataset are $y = 0$. We can see in Figure 2a that the neural network function f_W converges to this linear regression optimum. Figure 2b shows that the difference of the loss to the linear regression optimum converges to zero and has fast and slow dynamics: It first decays very quickly and then more slowly. As we will see, the reason is that the convergence of f_W to the linear regression optimum is driven by a linear system whose initialization is close to the eigenspace spanned by eigenvectors to strongly negative eigenvalues. This property of the initialization is true under the assumption that the optimal linear regression lines corresponding to both parts of the dataset have intercept zero, i.e. contain the point $(0, 0)$. In Figure 2a, this is true since both optimal linear regression lines are given by the equation $y = 0$. In Figure 2c, we can see that with decreasing loss, the kinks move more slowly (which is due to a smaller gradient). If the loss decreases quickly enough, the kinks do not reach the dataset.
The proof will proceed roughly as follows:

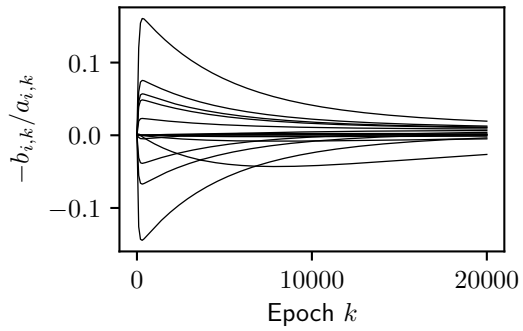
- Since the function φ is linear on $(0, \infty)$ and $(-\infty, 0)$, we can replace each occurrence of φ by a linearized version in the loss function L_D, yielding a modified loss function $L_{D,\tau}$. We will see that as long as no kink crosses a datapoint, we have $\nabla L_D(W_k) = \nabla L_{D,\tau}(W_k)$ and gradient descent behaves identically on both loss functions. The following steps outline the proof that this indeed happens with probability $1 - o(1)$.
- We can use concentration inequalities to show that the initialization behaves in a certain fashion with high probability. For now, we only consider the case where the initialization behaves according to these concentration inequalities. We also



(a) Neural network function f_{W_k} after $k = 0, 1000$ and 20000 epochs. The small blue points represent the kinks after 20000 epochs. The large black points constitute the training set.



(b) Evolution of the loss difference $L_D(W_k) - \inf_{k'} L_D(W_{k'})$ during training.



(c) Evolution of the kinks $-b_{i,k}/a_{i,k}$ during training.

Figure 2: Real data from training a neural network as in Definition 2.1 with $\alpha = 0$, $n = 16$, $h = 0.001$ and dataset $D = ((-3, -1), (-2, 2), (-1, -1), (1, 1), (2, -2), (3, 1))$.

ignore subpolynomial factors in our asymptotic notation. In Section 5, we will use a precise asymptotic notation that does not ignore these aspects.

- The linearized version of gradient descent remains accurate whenever none of the terms $a_i x_j + b_i$ crosses zero, since this is where the nonlinearity of φ occurs. We assume that the points x_j satisfy $|x_j| \geq m_P$ for some constant $m_P > 0$, i.e. there should be no data points near zero. For the initial condition, we will see that $\min_{i,j} |a_i(0)x_j + b_i(0)| = \Theta(1/n)$. We want to show that a_i and b_i change

by $o(1/n)$, which will then yield that $a_i x_j + b_i$ does not cross zero with high probability for large enough n. This in turn means that the kink $-b_i/a_i$ never reaches x_j.

- We analyze how the slopes p_1, p_{-1} and intercepts q_1, q_{-1} of $f_{W,\tau,1}$ and $f_{W,\tau,-1}$ shown in Figure 1b behave. A constant modification $\overline{v}_k = v_k - v^{\text{opt}}$ of the vector $v_k = (p_{1,k}, p_{-1,k}, q_{1,k}, q_{-1,k})$ satisfies a linear iteration equation $\overline{v}_{k+1} = \overline{v}_k - hA_kM\overline{v}_k$. The matrix A_k depends on the parameters W_k while M depends on the data points. Since M does not depend on the initialization of the system, it can be regarded as constant. Moreover, under certain assumptions, A_k and M are symmetric and positive definite. Typical values might look like this:

$$A_0 = \begin{pmatrix} n/2 & & & \\ & n/2 & & \\ & & 3/2 & 1 \\ & & 1 & 3/2 \end{pmatrix}, \qquad M = \begin{pmatrix} 7/3 & & 1 & \\ & 7/3 & & -1 \\ 1 & & 1/2 & \\ & -1 & & 1/2 \end{pmatrix} .$$

  It turns out that the matrix A_0M has two eigenvalues of order $\Theta(n)$ and two eigenvalues of order $\Theta(1)$. If $\overline{v}_0 = (*, *, 0, 0)^\top$ (which translates to the assumption that both optimal linear regression lines should have intercept zero), we will prove that $\overline{v}_0$ is close to the subspace spanned by the first two eigenvectors and we should have $h\sum_{k=0}^{\infty} \|\overline{v}_k\| = O(1/n)$. Establishing this behavior is the purpose of Appendix B.

- Back to our goal of proving that a_i and b_i change by $o(1/n)$: For a_i, we can show that $a_{i,k+1} - a_{i,k} = hr_{\sigma,k}w_{i,k}$, where $r_{\sigma,k}$ is a linear function of $\overline{v}_k$. Hence, a_i changes at most by

$$\sum_{k=0}^{\infty} |a_{i,k+1} - a_{i,k}| \leq \left(\sup_{k\geq 0} |w_{i,k}|\right) \cdot h \sum_{k=0}^{\infty} |r_{\sigma,k}| = O(n^{-1/2}) \cdot O(1/n) = o(1/n) \ .$$

  We can use a similar argument for b_i. While concentration inequalities yield that $\max_i |w_{i,0}| = O(n^{-1/2})$ up to a subpolynomial factor, we also need to show that $|w_{i,k}|$ remains close to $|w_{i,0}|$. Similarly, we have to ensure that the matrix A_k, which depends on the parameters W_k, remains close to A_0 such that our assertion $h\sum_{k=0}^{\infty} \|\overline{v}_k\| = O(1/n)$ is true. To this end, we argue that as long as $h\sum_{l=0}^{k} \|\overline{v}_l\|$ is small, W_k is close to W_0 and vice versa. Since each of these conditions implies the other, we can argue by induction that both must hold for all $k \in \mathbb{N}_0$. The two directions of the argument are handled in Proposition 5.31 and Proposition 5.33, respectively.

The proof is structured as follows: In Section 5.1, we simplify the gradient descent equations and introduce quantities like $f_{W,\tau,\sigma}, \overline{v}$ and p_σ, q_σ for $\sigma \in \{-1, 1\}$. We also derive update equations like $\overline{v}_{k+1} = \overline{v}_k - hA_kM\overline{v}_k$ for these new quantities. In Section 5.2, we give several remarks about the newly derived quantities and their interaction. Section 5.3 then investigates stochastic properties of initialization and dataset sampling. It also formally introduces asymptotic notation adapted to our proofs. We especially rely on this notation in Section 5.4, where we analyze how much the weight vector W_k changes with k and how large the sum $h\sum_{k=0}^{\infty} \|\overline{v}_k\|$ is. The inconsistency results are then presented and proved in Section 5.5. Some proofs are deferred to the appendix.

3 Related Work

Several different branches of research are related to the presented theory: Universal consistency of globally optimized neural networks [36, 13] and NP-hardness of neural network optimization [4] has already been investigated in the 1980s and 1990s. Around this time, researchers also began a closer investigation of the loss landscape of neural networks [32, 18]. In the meantime, researchers have found more characterizations regarding which types of networks contain (bad) non-global minima or (bad) saddle points [33, 38, 14, 28].

For example, Soudry and Carmon [33] show that in overparameterized deep neural networks with ReLU-type activation functions, least-squares loss, fixed Gaussian dropout, without bias terms and for almost every dataset, each differentiable local minimum is also a global one. Their strategy is to show that if the gradient of the loss function is zero, then $Ge = 0$, where e is a vector containing the errors on the data points and G is a matrix depending on the data points and the network's activations. They show that G almost surely has full rank for sufficient overparameterization. However, Gaussian dropout is typically not used. Safran and Shamir [28] find non-global minima in non-overparameterized ReLU networks and observe empirically that as the number of hidden units and data points increases, the probability of reaching a global minimum can become very small. Yun et al. [39] construct examples where a ReLU network has differentiable local minima that are not global minima. In this thesis, we show conditions for convergence to non-global minima of this type.

Another line of work studies the dynamics of specific gradient-based optimization algorithms on various (mostly overparameterized) network types. Saxe et al. [29] analyze exact solutions of training deep linear networks with negative gradient flow under certain conditions. For example, they assume that the input data $(x_1, \dots, x_n)$ is whitened (normalized) and the initialization satisfies an orthogonality condition. Many works try to characterize the behavior of gradient descent on certain classes of neural networks when the labels stem from a neural network of the same class and/or when the inputs x_j follow a Gaussian distribution, e.g. [9, 10, 23, 5, 30]. Li and Liang [22] investigate training overparameterized two-layer ReLU networks without biases using SGD on a cross-entropy classification loss with softmax output activation. Under certain assumptions on the training data, they prove that the parameters found by SGD have a low generalization error with high probability. However, they only optimize over the first layer. Similar to this thesis, they fix the activation pattern of the ReLU activation functions to the pattern at initialization inside the proof. They also find that in the overparameterized setting, the network weights do not change much.

Perhaps the most related work to ours is a paper by Du et al. [11]. In this paper, the authors consider two-layer overparameterized ReLU networks without biases and apply gradient descent to optimize a least-squares loss. Similar to the present work, they directly try to analyze the dynamics of gradient descent in function space and not only in weight space. Another similarity is that they analyze the dynamics of the loss. Using a Gramian-based approach, they show that the training loss converges to zero at a certain rate. Like Li and Liang [22], they observe that most activation patterns do not change and the network weights remain close to their initialization. Their gram matrix H exhibits some similarity to our matrices A and M. However, $H \in \mathbb{R}^{N \times N}$,

whereas $A, M \in \mathbb{R}^{4\times 4}$. They achieve $H \succ 0$ (i.e. H symmetric and positive definite) via overparameterization and requiring that no two points x_j are parallel, while we achieve $A, M \succ 0$ by placing multiple points $\{x_j \mid j \in J_\sigma\}$ into the same columns/rows of the matrices. They use an induction to show that H and W do not change much and the loss decays at a certain rate. This is similar to our induction showing that A and W do not change much and $\overline{v}$ decays at a certain rate. However, in their case, the induction can continue when an activation pattern changes. Meanwhile, in our case, the matrix A is initialized in a more special fashion and we depend on a two-phase loss decay with fast and slow dynamics. In their setting, our matrix A would be considerably simpler since they initialize the second-layer weights as $w_i \sim \mathcal{U}\{-1, 1\}$ and they do not use biases. The intuition behind their paper is also different than ours: In higher-dimensional input spaces, the kinks are not points but hyperplanes of codimension one. Since Du et al. [11] do not use biases, all of their hyperplanes pass through the origin. In the overparameterized regime with their assumption of non-parallel points x_j, it is still likely that these hyperplanes separate all data points so that they do not have to move before the network can fit the data. In contrast, we use a one-dimensional input space, where their assumption of no training points being parallel can only be satisfied if there is at most one training point.

In another paper, Du et al. [8] consider overparameterized feedforward and residual networks using certain smooth activation functions trained with least-squares loss. For wide enough layers, they can use the condition on the activation functions to show that a Gramian matrix H is positive definite and its minimal eigenvalue yields a bound on the speed of convergence.

Jacot et al. [21] investigate the behavior of negative gradient flow on neural networks in the infinite-width limit. They observe that it relates to a kernel which they call Neural Tangent Kernel. Allen-Zhu et al. [1] build on the work by Jacot et al. [21] and show convergence of overparameterized deep networks trained with gradient descent or stochastic gradient descent to a global optimum with high probability. They use an initialization similar to He et al. [20] but they do not use biases. Since they append a fixed component to their input vector, the associated weights can be interpreted as biases, but these biases are initialized using a Gaussian distribution, while we initialize biases to zero as suggested by He et al. [20]. Their result even holds if the last layer is fixed, which addresses another phenomenon: In the overparameterized regime, learning can also happen by just updating the last layer, which for the least-squares loss is equivalent to a (convex) linear regression problem [27, 37]. In our scenario, we find that mainly the last layer is updated since the other layer has higher initialization variance, cf. Corollary 5.35 and Remark 5.36.

Arora et al. [2] consider overparameterized two-layer ReLU networks without biases optimized using gradient descent on a least-squares loss. Building on the theory by Du et al. [11], they find that the convergence speed of gradient descent is data-dependent and provides a bound on the generalization error. We also find such a data-dependence but with different interpretation and implications. A further similarity is explained in Remark 5.46.

Some very recent works try to reduce the amount of overparameterization that is assumed in proofs of convergence to a global optimum [12, 26, 40]. While our main results such as Theorem 2.3 are stated for networks with $n = O(\sqrt{N})$, i.e. in the underparameterized

regime, this assumption is only required because the intercept of the optimal linear regression lines of the sampled dataset must be sufficiently close to zero. If one chooses a fixed dataset where the optimal linear regression lines pass exactly through zero, our negative result holds for an arbitrarily large amount of overparameterization as discussed in Remark 5.41. This is consistent with other results on overparameterized networks since, as discussed above, their assumptions are not satisfied in our scenario. Moreover, since the neural networks basically only perform linear regression in our case with high probability, they can also be seen as overparameterized for this task. This "overparameterization" is important to the proof because it allows to apply concentration inequalities that characterize likely properties of the initialization.

We also want to mention some other research which is related to this thesis. ReLU kink movement appears to be rarely discussed in the literature, but some examples can be found in papers by Maennel et al. [24] and Steinwart [34]. Steinwart [34] notices experimentally that there are examples where kinks of ReLU networks fail to distribute across training data points. He proposes to choose a data-dependent initialization of the network such that the kinks are well-distributed among the data. Mücke and Steinwart [25] prove that there exists a consistent and an inconsistent training method for ReLU networks such that both are empirical risk minimizers (ERMs), i.e. find global minima of the training loss. However, they use overparameterized neural networks and do not investigate gradient descent methods. Finally, this work is related to initialization methods for neural networks, which are discussed for example in Section 8.4 in [17]. A popular initialization method called Xavier initialization was introduced by Glorot and Bengio [15] and is constructed for tanh and similar activation functions. In another celebrated work, He et al. [20] calculated different initialization variances for ReLU activations. We use a generalized version of their initialization method in this thesis.

4 Notation and Matrix Algebra Basics

In this section, we will introduce some notation that is used throughout the thesis. We will also list some results about matrices, especially involving matrix norms, eigenvalues and singular values, cf. e.g. [3, 16].

Definition 4.1. Let $A, B \in \mathbb{R}^{m\times m}$ and $C \in \mathbb{R}^{n\times m}$.

(1) We denote the sign of a real number $x \in \mathbb{R}$ by

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{, if } x > 0 \\ 0 & \text{, if } x = 0 \\ -1 & \text{, if } x < 0 \ . \end{cases}$$

(2) For a set S, we denote its indicator function by $\mathbb{1}_S$, i.e.

$$\mathbb{1}_S(x) = \begin{cases} 1 & \text{, if } x \in S \\ 0 & \text{, otherwise.} \end{cases}$$

(3) We write $A \succ B$ iff A, B are symmetric and $A - B$ is positive definite. Similarly, we write $A \succeq B$ iff A, B are symmetric and $A - B$ is positive semidefinite. Especially, $A \succ 0$ iff A is symmetric and positive definite, while $A \succeq 0$ iff A is symmetric and positive semidefinite. We define $\preceq$ and $\prec$ analogously.

(4) Let $\operatorname{eig}(A)$ denote the set of eigenvalues of A. If $\operatorname{eig}(A) \subseteq \mathbb{R}$, we define

$$\begin{aligned} \lambda_{\max}(A) &\coloneqq \max \operatorname{eig}(A) \\ \lambda_{\min}(A) &\coloneqq \min \operatorname{eig}(A) \ . \end{aligned}$$

Especially, $\lambda_{\max}$ and $\lambda_{\min}$ are defined for symmetric matrices.

(5) It is well-known that each real rectangular matrix $C \in \mathbb{R}^{n\times m}$ has a singular value decomposition $C = UDV^\top$, where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ are orthogonal matrices and $D \in \mathbb{R}^{n\times m}$ is diagonal with nonnegative entries. The set of singular values of C is unique and defined as

$$\begin{cases} \{D_{ii} \mid 1 \le i \le \min\{n, m\}\} & \text{, if } n = m \\ \{D_{ii} \mid 1 \le i \le \min\{n, m\}\} \cup \{0\} & \text{, if } n \neq m. \end{cases}$$

We use $\sigma_{\max}(C)$ and $\sigma_{\min}(C)$ to denote the maximum singular value and the minimum singular value, respectively.

(6) A square matrix A is invertible iff $\sigma_{\min}(A) > 0$. In this case, if $A = UDV^\top$ is a singular value decomposition of A, then $A^{-1} = VD^{-1}U^\top$. This shows $\sigma_{\max}(A^{-1}) = \sigma_{\min}(A)^{-1}$ and $\sigma_{\min}(A^{-1}) = \sigma_{\max}(A)^{-1}$. Similarly, if A has real eigenvalues, then $\lambda_{\max}(A^{-1}) = \lambda_{\min}(A)^{-1}$ and $\lambda_{\min}(A^{-1}) = \lambda_{\max}(A)^{-1}$. If A is symmetric, the singular values are the absolute values of the eigenvalues. If $A \succeq 0$, then A has a singular value decomposition $A = UDU^\top$ which is also an orthogonal diagonalization of A. We can then define the (symmetric) square root of A as $A^{1/2} \coloneqq UD^{1/2}U^\top$, where $D^{1/2}$ contains the square roots of the entries of D. Note that $\lambda_{\max}(A^{1/2}) = \lambda_{\max}(A)^{1/2}$ and $\lambda_{\min}(A^{1/2}) = \lambda_{\min}(A)^{1/2}$.

(7) As matrix norms, we use the Frobenius norm as well as the induced 2- and ∞-norms:

$$\begin{aligned}\|C\|_F &= \left(\sum_{i,j} C_{i,j}^2\right)^{1/2} \\ \|C\|_2 &= \sup_{x\neq 0} \frac{\|Cx\|_2}{\|x\|_2} = \sigma_{\max}(C) \\ \|C\|_\infty &= \sup_{x\neq 0} \frac{\|Cx\|_\infty}{\|x\|_\infty} = \max_i \sum_j |C_{ij}| \ .\end{aligned}$$

These satisfy the following inequalities (cf. e.g. Section 2.3 in [16]):

$$\begin{aligned}\|C\|_2 \leq \|C\|_F \leq \sqrt{m}\|C\|_2 \\ \frac{1}{\sqrt{m}}\|C\|_\infty \leq \|C\|_2 \leq \sqrt{n}\|C\|_\infty \ .\end{aligned}$$

(8) We define the condition number of an invertible matrix $A \in \mathbb{R}^{m\times m}$ by

$$\operatorname{cond}(A) \coloneqq \|A\|_2 \cdot \|A^{-1}\|_2 = \sigma_{\max}(A)\sigma_{\max}(A^{-1}) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} \ .$$

If $A \succ 0$, then

$$\operatorname{cond}(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \ .$$

(9) We occasionally use element-wise operations on matrices. For example, $|A|$ is the matrix containing as entries the absolute values of the entries of A and $\sup_s A(s)$ consists of the element-wise suprema. Also, $A \leq B$ means that $A_{ij} \leq B_{ij}$ for all i, j. ◀

There are some more facts about matrices that we will use during some proofs. We show some typical arguments here:

- We will use the fact that for symmetric A,

$$\lambda_{\max}(A) = \sup_{\|v\|_2=1} v^\top A v = \|A\|_2, \qquad \lambda_{\min}(A) = \inf_{\|v\|_2=1} v^\top A v \ ,$$

which is a special case of the Courant-Fischer-Weyl min-max principle (e.g. Corollary III.1.2 in [3]). This shows $A \succeq 0 \Leftrightarrow \lambda_{\min}(A) \geq 0$. Moreover, for symmetric $A, B \in \mathbb{R}^{m\times m}$, we have

$$A \succeq B \quad \Leftrightarrow \quad \forall v \in \mathbb{R}^m : v^\top A v \geq v^\top B v \ .$$

For example, for $A, M \succeq 0$ and $v \in \mathbb{R}^m$,

$$\begin{aligned}v^\top M^{1/2} A M^{1/2} v &= (M^{1/2}v)^\top A (M^{1/2}v) \\ &\leq \lambda_{\max}(A)(M^{1/2}v)^\top (M^{1/2}v)\end{aligned}$$

$$
\begin{aligned}
&= \lambda_{\max}(A) v^\top M v \\
&\leq \lambda_{\max}(A)\lambda_{\max}(M) v^\top v \ .
\end{aligned}
$$

We can thus conclude that $\lambda_{\max}(M^{1/2}AM^{1/2}) \leq \lambda_{\max}(A)\lambda_{\max}(M)$. Moreover, $M^{1/2}AM^{1/2} \preceq \lambda_{\max}(A)M \preceq \lambda_{\max}(A)\lambda_{\max}(M)I$, where I is the identity matrix. For $A, B \succeq 0$, we can also use such an argument to show that

$$
\begin{aligned}
\lambda_{\max}(A+B) &\leq \lambda_{\max}(A) + \lambda_{\max}(B) \\
\lambda_{\min}(A+B) &\geq \lambda_{\min}(A) + \lambda_{\min}(B) \ .
\end{aligned}
$$

- If

$$
M = \begin{pmatrix} M_{11} & M_{12} \\ M_{12}^\top & M_{22} \end{pmatrix} \succeq 0 \ ,
$$

we know that

$$
x^\top M_{11} x = \begin{pmatrix} x \\ 0 \end{pmatrix}^\top M \begin{pmatrix} x \\ 0 \end{pmatrix} \geq \lambda_{\min}(M)\|x\|_2^2 \ ,
$$

hence $M_{11} \succeq 0$ with $\lambda_{\min}(M_{11}) \geq \lambda_{\min}(M)$. Similarly, $\lambda_{\max}(M_{11}) \leq \lambda_{\max}(M)$ and analogous identities hold for M_{22}. A similar argument also shows that

$$
\begin{pmatrix} M_1 & \\ & M_2 \end{pmatrix} \succ 0 \text{ iff } M_1, M_2 \succ 0 \ .
$$

If $M_1 = U_1 D_1 V_1^\top$ and $M_2 = U_2 D_2 V_2^\top$ are singular value decompositions, then

$$
M \coloneqq \begin{pmatrix} M_1 & \\ & M_2 \end{pmatrix} = \begin{pmatrix} U_1 & \\ & U_2 \end{pmatrix} \begin{pmatrix} D_1 & \\ & D_2 \end{pmatrix} \begin{pmatrix} V_1 & \\ & V_2 \end{pmatrix}^\top
$$

is a singular value decomposition. Hence $\sigma_{\max}(M) = \max\{\sigma_{\max}(M_1), \sigma_{\max}(M_2)\}$.

- We have

$$
\sigma_{\min}(C) = \inf_{\|v\|_2=1} \|Cv\|_2
$$

and hence for $A \succeq 0$,

$$
\begin{aligned}
\lambda_{\max}(C^\top AC) &= \sup_{\|v\|_2=1} v^\top C^\top ACv \leq \sigma_{\max}(C)^2 \sup_{\|w\|_2=1} w^\top Aw = \sigma_{\max}(C)^2 \lambda_{\max}(A) \\
\lambda_{\min}(C^\top AC) &= \inf_{\|v\|_2=1} v^\top C^\top ACv \geq \sigma_{\min}(C)^2 \inf_{\|w\|_2=1} w^\top Aw = \sigma_{\min}(C)^2 \lambda_{\min}(A) \ .
\end{aligned}
$$

- If $\hat{C}$ is a submatrix of C (with some columns and/or rows removed), then there exist orthogonal projections P, Q with $\hat{C} = PCQ^\top$ and hence $\|\hat{C}\|_2 \leq \|P\|_2\|C\|_2\|Q\|_2 \leq \|C\|_2$.

5 Theory

In this section, the main inconsistency results are derived, although some proofs are deferred to the appendix.

5.1 System Decomposition

First, we are concerned with analyzing the quantities from Definition 2.1 to obtain new insights. To this end, we want to "linearize" f_W. If $a_i x + b_i \neq 0$, our choice of φ satisfies

$$\varphi(a_i x + b_i) = \varphi'(\operatorname{sgn}(a_i x + b_i)) \cdot (a_i x + b_i) \ ,$$

where $\varphi'(\operatorname{sgn}(a_i x + b_i))$ remains constant for small changes of W and x. Moreover, if the weight vector W and the dataset D satisfy $a_i, x_j \neq 0$ and $b_i = 0$ for all i, j, which is a typical case for the initial weight vector $W = W_0$, then $\operatorname{sgn}(a_i x_j + b_i) = \operatorname{sgn}(a_i)\operatorname{sgn}(x_j)$. This motivates the following definition:

Definition 5.1 (Sign patterns)**.** Let $I \coloneqq \{1, \dots, n\}$ and $J \coloneqq \{1, \dots, N\}$. A given weight vector W and a dataset D with $a_i, x_j \neq 0$ for all $i \in I, j \in J$ induce partitions $I = I_1(W) \cup I_{-1}(W), J = J_1(W) \cup J_{-1}(W)$ and a sign vector $\tau(W) \in \mathbb{R}^n$ via

$$\begin{aligned} \tau_i(W) &\coloneqq \operatorname{sgn}(a_i) \\ I_\sigma(W) &\coloneqq \{i \in I \mid \operatorname{sgn}(a_i) = \sigma\} \\ J_\sigma(D) &\coloneqq \{j \in J \mid \operatorname{sgn}(x_j) = \sigma\} \end{aligned}$$

for $\sigma \in \{-1, 1\}$. ◀

Now, we are able to consider "linearized" versions of f_W and L_D:

Definition 5.2 (Linearization of the problem)**.** For a given (fixed) sign vector $\tau \in \{-1, 1\}^n$, corresponding partitions $I_\sigma = \{i \in I \mid \tau_i = \sigma\}$ and a sign $\sigma \in \{-1, 1\}$, we define the modified neural net function

$$f_{W,\tau,\sigma}(x) \coloneqq c + \sum_{i \in I} w_i \varphi'(\tau_i \cdot \sigma)(a_i x + b_i) = c + \sum_{i \in I_\sigma} w_i (a_i x + b_i) + \alpha \sum_{i \in I_{-\sigma}} w_i (a_i x + b_i)$$

and the modified loss

$$L_{D,\tau}(W) \coloneqq \frac{1}{2N} \sum_{j \in J} (y_j - f_{W,\tau,\operatorname{sgn}(x_j)}(x_j))^2 \ . \qquad ◀$$

Note that while $f_{W,\tau,\sigma}(x)$ is linear in x, it is not linear in W. Instead, $f_{W,\tau,\sigma}(x)$ is polynomial in W, containing up to second-order terms. Hence, $L_{D,\tau}(W)$ contains up to fourth-order terms in W. Unlike L_D, $L_{D,\tau}$ is differentiable everywhere. We obtain the derivatives

$$\begin{aligned} \frac{\partial L_{D,\tau}}{\partial a_i}(W) &= \frac{1}{N} \sum_{j \in J} (f_{W,\tau,\operatorname{sgn}(x_j)}(x_j) - y_j) \varphi'(\tau_i \operatorname{sgn}(x_j)) w_i x_j \\ \frac{\partial L_{D,\tau}}{\partial b_i}(W) &= \frac{1}{N} \sum_{j \in J} (f_{W,\tau,\operatorname{sgn}(x_j)}(x_j) - y_j) \varphi'(\tau_i \operatorname{sgn}(x_j)) w_i \end{aligned} \tag{5.1}$$

$$\frac{\partial L_{D,\tau}}{\partial w_i}(W) = \frac{1}{N}\sum_{j\in J}(f_{W,\tau,\operatorname{sgn}(x_j)}(x_j) - y_j)\varphi'(\tau_i \operatorname{sgn}(x_j))(a_i x_j + b_i)$$
$$\frac{\partial L_{D,\tau}}{\partial c}(W) = \frac{1}{N}\sum_{j\in J}(f_{W,\tau,\operatorname{sgn}(x_j)}(x_j) - y_j) \ .$$

Now, we consider gradient descent for a fixed sign vector τ. We will later choose $\tau \coloneqq \tau(W_0)$.

Definition 5.3 (Gradient descent)**.** For a given initial vector W_0 and step size $h > 0$, we recursively define

$$W_{k+1} \coloneqq W_k - h\nabla L_{D,\tau}(W_k) \ .$$

Moreover, we write $W_k = (a_{\cdot,k}, b_{\cdot,k}, c_k, w_{\cdot,k})$ and we may implicitly omit the index k when deriving identities that hold for each $k \in \mathbb{N}_0$. For any derived quantity $\xi \coloneqq g(W)$, define

$$\delta\xi \coloneqq \delta g(W) \coloneqq g(W - h\nabla L_{D,\tau}(W)) - g(W)$$

such that

$$\xi_{k+1} = g(W_{k+1}) = g(W_k) + (g(W_{k+1}) - g(W_k)) = \xi_k + \delta\xi_k$$

and hence

$$\delta g(W) = g(W + \delta W) - g(W) \ . \qquad \blacktriangleleft$$

We can now write iteration rules differently: Instead of

$$W_{k+1} = W_k - h\nabla L_{D,\tau}(W_k) \ ,$$

we will use the more convenient notation

$$\delta W = -h\nabla L_{D,\tau}(W)$$

which suppresses the iteration index k and reads more like the negative gradient flow ODE

$$\dot{W} = -h\nabla L_{D,\tau}(W) \ .$$

Lemma 5.4 (Differential calculus for δ)**.** *Let $g : \mathbb{R}^{3n+1} \to \mathbb{R}^m$ for some $n, m \geq 1$.*

(a) If g is linear, then $\delta g(W) = g(\delta W) = -hg(\nabla L_{D,\tau}(W))$.

(b) If g is constant, then $\delta g = 0$.

(c) If $g_1, g_2 : \mathbb{R}^{3n+1} \to \mathbb{R}$ are linear, then

$$\delta(g_1 \cdot g_2) = (\delta g_1) \cdot g_2 + g_1 \cdot (\delta g_2) + (\delta g_1) \cdot (\delta g_2) \ .$$

(d) If $g_1, g_2 : \mathbb{R}^{3n+1} \to \mathbb{R}^m$, then

$$\delta(g_1 + g_2) = \delta g_1 + \delta g_2 \ .$$

(e) If $g_2 : \mathbb{R}^{3n+1} \to \mathbb{R}^m, g_1 : \mathbb{R}^m \to \mathbb{R}^{m'}$ and g_1 is linear, then

$$\delta(g_1 \circ g_2) = g_1 \circ (\delta g_2) \ .$$

(f) If $g_1, \ldots, g_m : \mathbb{R}^{3n+1} \to \mathbb{R}$, then

$$\delta \begin{pmatrix} g_1 \\ \vdots \\ g_m \end{pmatrix} = \begin{pmatrix} \delta g_1 \\ \vdots \\ \delta g_m \end{pmatrix} \ .$$

Proof.

(a) If g is linear, then

$$\delta g(W) = g(W + \delta W) - g(W) = g(\delta W) = g(-h \nabla L_{D,\tau}(W)) = -hg(\nabla L_{D,\tau}(W)) \ .$$

(b) Trivial.

(c) In this case,

$$\begin{aligned} \delta g(W) &= g(W + \delta W) - g(W) \\ &= g_1(W) g_2(\delta W) + g_1(\delta W) g_2(W) \\ &\quad + g_1(\delta W) g_2(\delta W) \\ &\overset{\text{(a)}}{=} \delta g_1(W) g_2(W) + g_1(W) \delta g_2(W) + \delta g_1(W) \delta g_2(W) \ . \end{aligned}$$

(d) We have

$$\begin{aligned} \delta(g_1 + g_2)(W) &= (g_1 + g_2)(W + \delta W) - (g_1 + g_2)(W) \\ &= (g_1(W + \delta W) - g_1(W)) + (g_2(W + \delta W) - g_2(W)) \\ &= \delta g_1(W) + \delta g_2(W) \ . \end{aligned}$$

(e) For $W \in \mathbb{R}^{3n+1}$,

$$\begin{aligned} \delta(g_1 \circ g_2)(W) &= g_1(g_2(W + \delta W)) - g_1(g_2(W)) = g_1(g_2(W + \delta W) - g_2(W)) \\ &= g_1(\delta g_2(W)) \ . \end{aligned}$$

(f) This follows from

$$\begin{pmatrix} g_1 \\ \vdots \\ g_m \end{pmatrix} (W + \delta W) - \begin{pmatrix} g_1 \\ \vdots \\ g_m \end{pmatrix} (W) = \begin{pmatrix} g_1(W + \delta W) - g_1(W) \\ \vdots \\ g_m(W + \delta W) - g_1(W) \end{pmatrix} \ . \qquad \square$$

The following definition introduces essential terms that will be used throughout this thesis.

Definition 5.5 (Derived quantities)**.**

(a) For $\sigma \in \{-1, 1\}$, we write $\Sigma_{\sigma,a^2} \coloneqq \sum_{i \in I_\sigma} a_i^2$, $\Sigma_{\sigma,wa} \coloneqq \sum_{i \in I_\sigma} w_i a_i$ and so on.

(b) Let $k_1 < \ldots < k_{|J_\sigma|}$ such that $J_\sigma = \{k_1, \ldots, k_{|J_\sigma|}\}$. Define

$$X_\sigma \coloneqq \begin{pmatrix} x_{k_1} & 1 \\ \vdots & \vdots \\ x_{k_{|J_\sigma|}} & 1 \end{pmatrix} \in \mathbb{R}^{|J_\sigma| \times 2}, \qquad Y_\sigma \coloneqq \begin{pmatrix} y_{k_1} \\ \vdots \\ y_{k_{|J_\sigma|}} \end{pmatrix} \in \mathbb{R}^{|J_\sigma|},$$
$$M_\sigma \coloneqq \frac{1}{N} X_\sigma^\top X_\sigma = \frac{1}{N} \begin{pmatrix} \sum_{j \in J_\sigma} x_j^2 & \sum_{j \in J_\sigma} x_j \\ \sum_{j \in J_\sigma} x_j & \sum_{j \in J_\sigma} 1 \end{pmatrix} \in \mathbb{R}^{2 \times 2} .$$

Since $M_\sigma = \frac{1}{N} X_\sigma^\top X_\sigma$, we have $M_\sigma \succeq 0$. Moreover, if $|\{x_j \mid j \in J_\sigma\}| \geq 2$, then $\operatorname{rank}(X_\sigma) = 2$ and hence M_σ is invertible, which implies $M_\sigma \succ 0$.

(c) The matrix M_σ helps in relating different interesting quantities. For $M_\sigma \succ 0$, let

$$\hat{v}_\sigma \coloneqq \begin{pmatrix} \hat{p}_\sigma \\ \hat{q}_\sigma \end{pmatrix} \coloneqq \begin{pmatrix} \Sigma_{\sigma,wa} \\ \Sigma_{\sigma,wb} \end{pmatrix} \qquad v_\sigma \coloneqq \begin{pmatrix} p_\sigma \\ q_\sigma \end{pmatrix} \coloneqq \begin{pmatrix} \hat{p}_\sigma + \alpha \hat{p}_{-\sigma} \\ c + \hat{q}_\sigma + \alpha \hat{q}_{-\sigma} \end{pmatrix}$$
$$\hat{u}_\sigma \coloneqq \begin{pmatrix} \hat{r}_\sigma \\ \hat{s}_\sigma \end{pmatrix} \coloneqq \begin{pmatrix} -\frac{1}{N} \sum_{j \in J_\sigma} (f_{W,\tau,\sigma}(x_j) - y_j) x_j \\ -\frac{1}{N} \sum_{j \in J_\sigma} (f_{W,\tau,\sigma}(x_j) - y_j) \end{pmatrix} \quad u_\sigma \coloneqq \begin{pmatrix} r_\sigma \\ s_\sigma \end{pmatrix} \coloneqq \begin{pmatrix} \hat{r}_\sigma + \alpha \hat{r}_{-\sigma} \\ \hat{s}_\sigma + \alpha \hat{s}_{-\sigma} \end{pmatrix}$$

and

$$\hat{u}_\sigma^0 \coloneqq \frac{1}{N} X_\sigma^\top Y_\sigma, \qquad v_\sigma^{\text{opt}} \coloneqq M_\sigma^{-1} \hat{u}_\sigma^0, \qquad \overline{v}_\sigma \coloneqq v_\sigma - v_\sigma^{\text{opt}} .$$

We will show in Lemma 5.8 that $\hat{u}_\sigma = -M_\sigma \overline{v}_\sigma$. The u-vectors are interesting since their components can be used to simplify δW. As we will see in Lemma 5.8, v_σ is interesting since $f_{W,\tau,\sigma}(x) = p_\sigma x + q_\sigma$ for $j \in J_\sigma, x \in \mathbb{R}$ (cf. Figure 1b). The notation of the different variants is motivated as follows: Expressions with a hat such as $\hat{v}_\sigma$ and $\hat{u}_\sigma$ only sum over one sign σ. We will prove in Proposition 5.10 that $L_{D,\tau}$ is minimal iff $\overline{v}_\sigma = 0$ for $\sigma \in \{-1, 1\}$. Hence, in the optimum $\overline{v}_\sigma = 0$, we have $v_\sigma = v_\sigma^{\text{opt}} = (X_\sigma^\top X_\sigma)^{-1} X_\sigma^\top Y_\sigma$, which is a vector containing the slope and intercept of the optimal least-squares regression line through $\{(x_j, y_j) \mid j \in J_\sigma\}$ (see e.g. Section 5.1.4 in [17]).

We will also use the matrices

$$G_\sigma^{\text{w}} \coloneqq \begin{pmatrix} \Sigma_{\sigma,w^2} & 0 \\ 0 & \Sigma_{\sigma,w^2} \end{pmatrix}, \quad G_\sigma^{\text{ab}} \coloneqq \begin{pmatrix} \Sigma_{\sigma,a^2} & \Sigma_{\sigma,ab} \\ \Sigma_{\sigma,ab} & \Sigma_{\sigma,b^2} \end{pmatrix}, \quad G_\sigma^{\text{wab}} \coloneqq (r_\sigma \Sigma_{\sigma,wa} + s_\sigma \Sigma_{\sigma,wb}) I_2 ,$$

where I_2 is the 2×2 identity matrix.

(d) For any two vectors $z_1, z_{-1} \in \mathbb{R}^2$ defined in step (c) and any two matrices $F_1, F_{-1} \in \mathbb{R}^{2 \times 2}$ defined in step (b), we define

$$\tilde{z} \coloneqq \begin{pmatrix} z_1 \\ z_{-1} \end{pmatrix} \in \mathbb{R}^4, \quad \tilde{F} \coloneqq \begin{pmatrix} F_1 & \\ & F_{-1} \end{pmatrix} \in \mathbb{R}^{4 \times 4} .$$

For example, this means that

$$\tilde{u} = \begin{pmatrix} u_1 \\ u_{-1} \end{pmatrix} = \begin{pmatrix} r_1 \\ s_1 \\ r_{-1} \\ s_{-1} \end{pmatrix} .$$

In addition, we define new matrices

$$\tilde{C} \coloneqq \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \quad \tilde{B} \coloneqq \begin{pmatrix} 1 & 0 & \alpha & 0 \\ 0 & 1 & 0 & \alpha \\ \alpha & 0 & 1 & 0 \\ 0 & \alpha & 0 & 1 \end{pmatrix} = \begin{pmatrix} I_2 & \alpha I_2 \\ \alpha I_2 & I_2 \end{pmatrix},$$
$$\tilde{A} \coloneqq \tilde{B}(\tilde{G}^{\mathrm{w}} + \tilde{G}^{\mathrm{ab}} + h\tilde{G}^{\mathrm{wab}})\tilde{B} + \tilde{C} .$$

We will prove in Proposition 5.9 that $\delta\tilde{\overline{v}} = h\tilde{A}\tilde{\hat{u}} = -h\tilde{A}\tilde{M}\tilde{\overline{v}}$.

(e) For any vector $\tilde{z} \in \mathbb{R}^4$ and any matrix $\tilde{F} \in \mathbb{R}^{4\times 4}$ defined in step (d), we define

$$z \coloneqq \tilde{P}\tilde{z}, \quad F \coloneqq \tilde{P}\tilde{F}\tilde{P}^{-1} = \tilde{P}\tilde{F}\tilde{P} ,$$

where $\tilde{P} = \tilde{P}^\top = \tilde{P}^{-1}$ is the permutation matrix

$$\tilde{P} \coloneqq \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} .$$

For example, this yields

$$u = \tilde{P}\tilde{u} = \begin{pmatrix} r_1 \\ r_{-1} \\ s_1 \\ s_{-1} \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & \alpha & & \\ \alpha & 1 & & \\ & & 1 & \alpha \\ & & \alpha & 1 \end{pmatrix} =: \begin{pmatrix} \hat{B} & \\ & \hat{B} \end{pmatrix} .$$

We see that this change of basis by $\tilde{P}$ makes the matrices $\tilde{B}$ and $\tilde{C}$ block-diagonal while it destroys the block-diagonal structure of $\tilde{G}^{\mathrm{ab}}$ and $\tilde{M}$. We will see in Lemma 5.27 that G^{ab} is still (block-)diagonal at initialization. We will use the tilde quantities as an intermediate step to derive equations for the non-tilde quantities, since the latter will be more suitable for us to analyze eigenvectors and eigenvalues.

Elementary arguments show that

$$(M_1 \succ 0 \text{ and } M_{-1} \succ 0) \Leftrightarrow \tilde{M} \succ 0 \Leftrightarrow M = \tilde{P}\tilde{M}\tilde{P}^\top \succ 0 .$$

Therefore, we need to require $M \succ 0$ so that v^{opt} and $\overline{v}$ can be defined.

(f) Many of the quantities above depend on the dataset D, which we may highlight later by indexing them with D. For example, we may write u_D instead of u.

(g) Finally, let

$$\theta_i \coloneqq \begin{pmatrix} a_i \\ b_i \\ w_i \end{pmatrix}, \quad \Sigma_\sigma \coloneqq \sum_{i \in I_\sigma} \theta_i \theta_i^\top = \begin{pmatrix} \Sigma_{\sigma,a^2} & \Sigma_{\sigma,ab} & \Sigma_{\sigma,wa} \\ \Sigma_{\sigma,ab} & \Sigma_{\sigma,b^2} & \Sigma_{\sigma,wb} \\ \Sigma_{\sigma,wa} & \Sigma_{\sigma,wb} & \Sigma_{\sigma,w^2} \end{pmatrix}, \quad Q_\sigma \coloneqq \begin{pmatrix} 0 & 0 & r_\sigma \\ 0 & 0 & s_\sigma \\ r_\sigma & s_\sigma & 0 \end{pmatrix}.$$

These quantities will be analyzed in the next proposition. ◀

Proposition 5.6. *For $i \in I_\sigma, \sigma \in \{-1, 1\}$, we have*

$$\begin{aligned} \delta\theta_i &= hQ_\sigma\theta_i \\ \delta c &= h(\hat{s}_1 + \hat{s}_{-1}) \\ \delta\Sigma_\sigma &= hQ_\sigma\Sigma_\sigma + h\Sigma_\sigma Q_\sigma + h^2 Q_\sigma \Sigma_\sigma Q_\sigma \end{aligned}$$

and the latter identity can also be written as

$$\Sigma_{\sigma,k+1} = (I + hQ_{\sigma,k})\Sigma_{\sigma,k}(I + hQ_{\sigma,k}) \ .$$

Proof. The first two equations can also be written as

$$\begin{aligned} \delta a_i &= hr_\sigma w_i \\ \delta b_i &= hs_\sigma w_i \\ \delta w_i &= hr_\sigma a_i + hs_\sigma b_i \\ \delta c &= h(\hat{s}_1 + \hat{s}_{-1}) \ . \end{aligned}$$

We will prove the first of these equations, the other ones follow similarly. Set $g(W) \coloneqq a_i$. With Lemma 5.4 (a), we obtain

$$\begin{aligned} \delta a_i &= \delta g(W) = -hg(\nabla L_{D,\tau}(W)) = -h\frac{\partial L_{D,\tau}}{\partial a_i}(W) \\ &\overset{(5.1)}{=} -h\frac{1}{N}\sum_{j \in J}(f_{W,\tau,\mathrm{sgn}(x_j)}(x_j) - y_j)\varphi'(\tau_i \cdot \mathrm{sgn}(x_j))w_i x_j \\ &= -h\frac{1}{N}\left(\sum_{j \in J_\sigma}(f_{W,\tau,\sigma}(x_j) - y_j)w_i x_j + \alpha \sum_{j \in J_{-\sigma}}(f_{W,\tau,-\sigma}(x_j) - y_j)w_i x_j\right) \\ &= h(\hat{r}_\sigma + \alpha\hat{r}_{-\sigma})w_i = hr_\sigma w_i \ . \end{aligned}$$

Now for Σ_σ: Since $Q_\sigma = Q_\sigma^\top$, we have

$$\begin{aligned} \Sigma_{\sigma,k+1} &= \sum_{i \in I_\sigma} \theta_{i,k+1}\theta_{i,k+1}^\top = \sum_{i \in I_\sigma}(I + hQ_{\sigma,k})\theta_{i,k}\theta_{i,k}^\top(I + hQ_{\sigma,k})^\top \\ &= (I + hQ_{\sigma,k})\left(\sum_{i \in I_\sigma}\theta_{i,k}\theta_{i,k}^\top\right)(I + hQ_{\sigma,k})^\top = (I + hQ_{\sigma,k})\Sigma_{\sigma,k}(I + hQ_{\sigma,k}) \ , \end{aligned}$$

which means that

$$\delta\Sigma_k = \Sigma_{k+1} - \Sigma_k = hQ_{\sigma,k}\Sigma_{\sigma,k} + h\Sigma_{\sigma,k}Q_{\sigma,k} + h^2 Q_{\sigma,k}\Sigma_{\sigma,k}Q_{\sigma,k} \ . \qquad \square$$

Remark 5.7. The term $h^2 Q_\sigma \Sigma_\sigma Q_\sigma^\top$ in Proposition 5.6 corresponds to the term $\delta g_1 \cdot \delta g_2$ in the "product rule" for δ (Lemma 5.4 (c)). It vanishes when using negative gradient flow. We will see that for small enough h, this term does not affect the qualitative behavior of gradient descent. ◀

The following lemma shows relations between several quantities from Definition 5.5.

Lemma 5.8. *Let $M \succ 0$. For $\sigma \in \{-1, 1\}$ and $x \in \mathbb{R}$, we have*

$$\begin{aligned} f_{W,\tau,\sigma}(x) &= p_\sigma x + q_\sigma \\ \hat{u}_\sigma &= -M_\sigma \overline{v}_\sigma \ . \end{aligned}$$

Moreover,

$$\tilde{u} = \tilde{B}\tilde{\hat{u}}, \quad \tilde{\hat{u}} = -\tilde{M}\tilde{\overline{v}}, \quad \tilde{v} = \tilde{B}\tilde{\hat{v}} + \begin{pmatrix} 0 \\ c \\ 0 \\ c \end{pmatrix} \ .$$

Proof. For $x \in \mathbb{R}$,

$$\begin{aligned} f_{W,\tau,\sigma}(x) &= c + \sum_{i \in I} w_i \varphi'(\tau_i \sigma)(a_i x + b_i) \\ &= c + \sum_{i \in I_\sigma} (w_i a_i x + w_i b_i) + \alpha \sum_{i \in I_{-\sigma}} (w_i a_i x + w_i b_i) = p_\sigma x + q_\sigma \ . \end{aligned}$$

Therefore,

$$\begin{aligned} \hat{u}_\sigma = \begin{pmatrix} \hat{r}_\sigma \\ \hat{s}_\sigma \end{pmatrix} &= -\frac{1}{N} \begin{pmatrix} \sum_{j \in J_\sigma} (f_{W,\tau,\sigma}(x_j) - y_j) x_j \\ \sum_{j \in J_\sigma} (f_{W,\tau,\sigma}(x_j) - y_j) \end{pmatrix} \\ &= -\frac{1}{N} \begin{pmatrix} \sum_{j \in J_\sigma} (p_\sigma x_j + q_\sigma - y_j) x_j \\ \sum_{j \in J_\sigma} (p_\sigma x_j + q_\sigma - y_j) \end{pmatrix} \\ &= -\frac{1}{N} \begin{pmatrix} p_\sigma \sum_{j \in J_\sigma} x_j^2 + q_\sigma \sum_{j \in J_\sigma} x_j - \sum_{j \in J_\sigma} x_j y_j \\ p_\sigma \sum_{j \in J_\sigma} x_j + q_\sigma \sum_{j \in J_\sigma} 1 - \sum_{j \in J_\sigma} y_j \end{pmatrix} \\ &= -M_\sigma \begin{pmatrix} p_\sigma \\ q_\sigma \end{pmatrix} + \frac{1}{N} X_\sigma^\top Y_\sigma = -M_\sigma v_\sigma + \hat{u}_\sigma^0 = -M_\sigma (v_\sigma - v_\sigma^{\mathrm{opt}}) = -M_\sigma \overline{v}_\sigma \ . \end{aligned}$$

We now obtain

$$\begin{aligned} \tilde{u} = \begin{pmatrix} r_1 \\ s_1 \\ r_{-1} \\ s_{-1} \end{pmatrix} &= \begin{pmatrix} \hat{r}_1 + \alpha \hat{r}_{-1} \\ \hat{s}_1 + \alpha \hat{s}_{-1} \\ \hat{r}_{-1} + \alpha \hat{r}_1 \\ \hat{s}_{-1} + \alpha \hat{s}_1 \end{pmatrix} = \tilde{B}\tilde{\hat{u}} \\ \tilde{\hat{u}} = \begin{pmatrix} \hat{u}_1 \\ \hat{u}_{-1} \end{pmatrix} &= \begin{pmatrix} M_1 & \\ & M_{-1} \end{pmatrix} \begin{pmatrix} \overline{v}_1 \\ \overline{v}_{-1} \end{pmatrix} = \tilde{M}\tilde{\overline{v}} \\ \tilde{v} = \begin{pmatrix} p_1 \\ q_1 \\ p_{-1} \\ q_{-1} \end{pmatrix} &= \begin{pmatrix} \hat{p}_1 + \alpha \hat{p}_{-1} \\ c + \hat{q}_1 + \alpha \hat{q}_{-1} \\ \hat{p}_{-1} + \alpha \hat{p}_1 \\ c + \hat{q}_{-1} + \alpha \hat{q}_1 \end{pmatrix} = \tilde{B}\tilde{\hat{v}} + \begin{pmatrix} 0 \\ c \\ 0 \\ c \end{pmatrix} \ . \end{aligned}$$

□

This enables us to compute another iteration equation:

Proposition 5.9. *Let $M \succ 0$. Then,*

$$\delta\tilde{\overline{v}} = -h\tilde{A}\tilde{M}\tilde{\overline{v}} = -h(\tilde{B}(\tilde{G}^{\mathrm{w}} + \tilde{G}^{\mathrm{ab}} + h\tilde{G}^{\mathrm{wab}})\tilde{B} + \tilde{C})\tilde{M}\tilde{\overline{v}}$$
$$\delta\overline{v} = -hAM\overline{v} = -h(B(G^{\mathrm{w}} + G^{\mathrm{ab}} + hG^{\mathrm{wab}})B + C)M\overline{v} \ .$$

Hence,

$$\overline{v}_{k+1} = \overline{v}_k + \delta\overline{v}_k = (I - hA_kM)\overline{v}_k \ .$$

Proof. Consider

$$\hat{v}_\sigma = \begin{pmatrix} \hat{p}_\sigma \\ \hat{q}_\sigma \end{pmatrix} = \begin{pmatrix} \Sigma_{\sigma,wa} \\ \Sigma_{\sigma,wb} \end{pmatrix} = \begin{pmatrix} I_2 & 0 \end{pmatrix} \Sigma_\sigma \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \ .$$

Using Proposition 5.6 and Lemma 5.4, we obtain

$$\begin{aligned}
\delta\hat{v}_\sigma &= \begin{pmatrix} I_2 & 0 \end{pmatrix} \left(hQ_\sigma\Sigma_\sigma + h\Sigma_\sigma Q_\sigma + h^2 Q_\sigma\Sigma_\sigma Q_\sigma \right) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \\
&= h \begin{pmatrix} 0 & 0 & r_\sigma \\ 0 & 0 & s_\sigma \end{pmatrix} \begin{pmatrix} \Sigma_{\sigma,wa} \\ \Sigma_{\sigma,wb} \\ \Sigma_{\sigma,w^2} \end{pmatrix} + h \begin{pmatrix} \Sigma_{\sigma,a^2} & \Sigma_{\sigma,ab} & \Sigma_{\sigma,wa} \\ \Sigma_{\sigma,ab} & \Sigma_{\sigma,b^2} & \Sigma_{\sigma,wb} \end{pmatrix} \begin{pmatrix} r_\sigma \\ s_\sigma \\ 0 \end{pmatrix} \\
&\quad + h^2 \begin{pmatrix} 0 & 0 & r_\sigma \\ 0 & 0 & s_\sigma \end{pmatrix} \begin{pmatrix} \Sigma_{\sigma,a^2} & \Sigma_{\sigma,ab} & \Sigma_{\sigma,wa} \\ \Sigma_{\sigma,ab} & \Sigma_{\sigma,b^2} & \Sigma_{\sigma,wb} \\ \Sigma_{\sigma,wa} & \Sigma_{\sigma,wb} & \Sigma_{\sigma,w^2} \end{pmatrix} \begin{pmatrix} r_\sigma \\ s_\sigma \\ 0 \end{pmatrix} \\
&= h \begin{pmatrix} \Sigma_{\sigma,w^2} & 0 \\ 0 & \Sigma_{\sigma,w^2} \end{pmatrix} u_\sigma + h \begin{pmatrix} \Sigma_{\sigma,a^2} & \Sigma_{\sigma,ab} \\ \Sigma_{\sigma,ab} & \Sigma_{\sigma,b^2} \end{pmatrix} u_\sigma + h^2(r_\sigma\Sigma_{\sigma,wa} + s_\sigma\Sigma_{\sigma,wb})u_\sigma \\
&= h \left(\tilde{G}^{\mathrm{w}}_\sigma + \tilde{G}^{\mathrm{ab}}_\sigma + h\tilde{G}^{\mathrm{wab}}_\sigma \right) u_\sigma \ .
\end{aligned}$$

Therefore, $\delta\tilde{\hat{v}} = h\left(\tilde{G}^{\mathrm{w}} + \tilde{G}^{\mathrm{ab}} + h\tilde{G}^{\mathrm{wab}}\right)\tilde{u}$. Also,

$$\delta \begin{pmatrix} 0 \\ c \\ 0 \\ c \end{pmatrix} = h \begin{pmatrix} 0 \\ \hat{s}_1 + \hat{s}_{-1} \\ 0 \\ \hat{s}_1 + \hat{s}_{-1} \end{pmatrix} = h \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{r}_1 \\ \hat{s}_1 \\ \hat{r}_{-1} \\ \hat{s}_{-1} \end{pmatrix} = h\tilde{C}\tilde{\hat{u}} \ .$$

We can now use the identities from Lemma 5.8 and the fact that $\tilde{v} - \tilde{\overline{v}} = \tilde{v}^{\mathrm{opt}}$ is constant to compute

$$\begin{aligned}
\delta\tilde{\overline{v}} = \delta\tilde{v} &= \tilde{B}\delta\tilde{\hat{v}} + \delta \begin{pmatrix} 0 \\ c \\ 0 \\ c \end{pmatrix} = \tilde{B}h\left(\tilde{G}^{\mathrm{w}} + \tilde{G}^{\mathrm{ab}} + h\tilde{G}^{\mathrm{wab}}\right)\tilde{u} + h\tilde{C}\tilde{\hat{u}} \\
&= h(\tilde{B}(\tilde{G}^{\mathrm{w}} + \tilde{G}^{\mathrm{ab}} + h\tilde{G}^{\mathrm{wab}})\tilde{B} + \tilde{C})\tilde{\hat{u}} = h\tilde{A}\tilde{\hat{u}} = -h\tilde{A}\tilde{M}\tilde{\overline{v}} \ .
\end{aligned}$$

Since $\tilde{P}^2 = I$, it follows that

$$\delta\overline{v} = \delta(\tilde{P}\tilde{\overline{v}}) = \tilde{P}\delta\tilde{\overline{v}} = -h\tilde{P}\tilde{A}\tilde{M}\tilde{\overline{v}} = -h\tilde{P}\tilde{A}\tilde{P}\tilde{P}\tilde{M}\tilde{P}\tilde{P}\tilde{\overline{v}} = -hAM\overline{v}$$

and

$$\begin{aligned} A = \tilde{P}\tilde{A}\tilde{P} &= \tilde{P}\Big(\tilde{B}\tilde{P}\tilde{P}(\tilde{G}^{\mathrm{w}} + \tilde{G}^{\mathrm{ab}} + h\tilde{G}^{\mathrm{wab}})\tilde{P}\tilde{P}\tilde{B} + \tilde{C}\Big)\tilde{P} \\ &= B(G^{\mathrm{w}} + G^{\mathrm{ab}} + hG^{\mathrm{wab}})B + C \ . \end{aligned}$$

□

The next proposition explains the relation of the previously investigated quantities to the loss. It is not necessary for investigating the dynamics of gradient descent but will help to establish the suboptimality of the trained networks in Corollary 5.39.

Proposition 5.10. *The (modified) loss $L_{D,\tau}$ is quadratic in $\overline{v}$:*

$$L_{D,\tau}(W) = \frac{1}{2}\overline{v}^\top M\overline{v} + \frac{1}{2N}\sum_{\sigma\in\{-1,1\}} Y_\sigma^\top (I - X_\sigma(X_\sigma^\top X_\sigma)^{-1}X_\sigma^\top)Y_\sigma \ .$$

Proof. The loss can be rewritten by completing the square:

$$\begin{aligned} 2L_{D,\tau}(W) &= \frac{1}{N}\sum_{\sigma\in\{-1,1\}}\sum_{j\in J_\sigma}(f_{W,\tau,\sigma}(x_j) - y_j)^2 \\ &\overset{5.8}{=} \frac{1}{N}\sum_{\sigma\in\{-1,1\}}\sum_{j\in J_\sigma}(p_\sigma x_j + q_\sigma - y_j)^2 \\ &= \frac{1}{N}\sum_{\sigma\in\{-1,1\}}\sum_{j=1}^{|J_\sigma|}(X_\sigma v_\sigma - Y_\sigma)_j^2 \\ &= \frac{1}{N}\sum_{\sigma\in\{-1,1\}}(X_\sigma v_\sigma - Y_\sigma)^\top(X_\sigma v_\sigma - Y_\sigma) \\ &= \frac{1}{N}\sum_{\sigma\in\{-1,1\}}\Big(Nv_\sigma^\top M_\sigma v_\sigma - v_\sigma^\top(X_\sigma^\top Y_\sigma) - (X_\sigma^\top Y_\sigma)^\top v_\sigma + Y_\sigma^\top Y_\sigma\Big) \\ &= \sum_{\sigma\in\{-1,1\}}\Big(v_\sigma - (X_\sigma^\top X_\sigma)^{-1}X_\sigma^\top Y_\sigma\Big)^\top M_\sigma\Big(v_\sigma - (X_\sigma^\top X_\sigma)^{-1}X_\sigma^\top Y_\sigma\Big) \\ &\quad + \frac{1}{N}\sum_{\sigma\in\{-1,1\}} Y_\sigma^\top(I - X_\sigma(X_\sigma^\top X_\sigma)^{-1}X_\sigma^\top)Y_\sigma \\ &\overset{5.5\,(c)}{=} \sum_{\sigma\in\{-1,1\}}\overline{v}_\sigma^\top M_\sigma\overline{v}_\sigma + \frac{1}{N}\sum_{\sigma\in\{-1,1\}} Y_\sigma^\top(I - X_\sigma(X_\sigma^\top X_\sigma)^{-1}X_\sigma^\top)Y_\sigma \ . \end{aligned}$$

Since $\tilde{P} = \tilde{P}^\top = \tilde{P}^{-1}$, we conclude

$$\sum_{\sigma\in\{-1,1\}}\overline{v}_\sigma^\top M_\sigma\overline{v}_\sigma = \begin{pmatrix}\overline{v}_1\\ \overline{v}_{-1}\end{pmatrix}^\top\begin{pmatrix}M_1 & \\ & M_{-1}\end{pmatrix}\begin{pmatrix}\overline{v}_1\\ \overline{v}_{-1}\end{pmatrix} = \tilde{\overline{v}}^\top\tilde{M}\tilde{\overline{v}} = \tilde{\overline{v}}^\top\tilde{P}\tilde{P}\tilde{M}\tilde{P}\tilde{P}\tilde{\overline{v}} = \overline{v}^\top M\overline{v} \ ,$$

which yields the claim. □

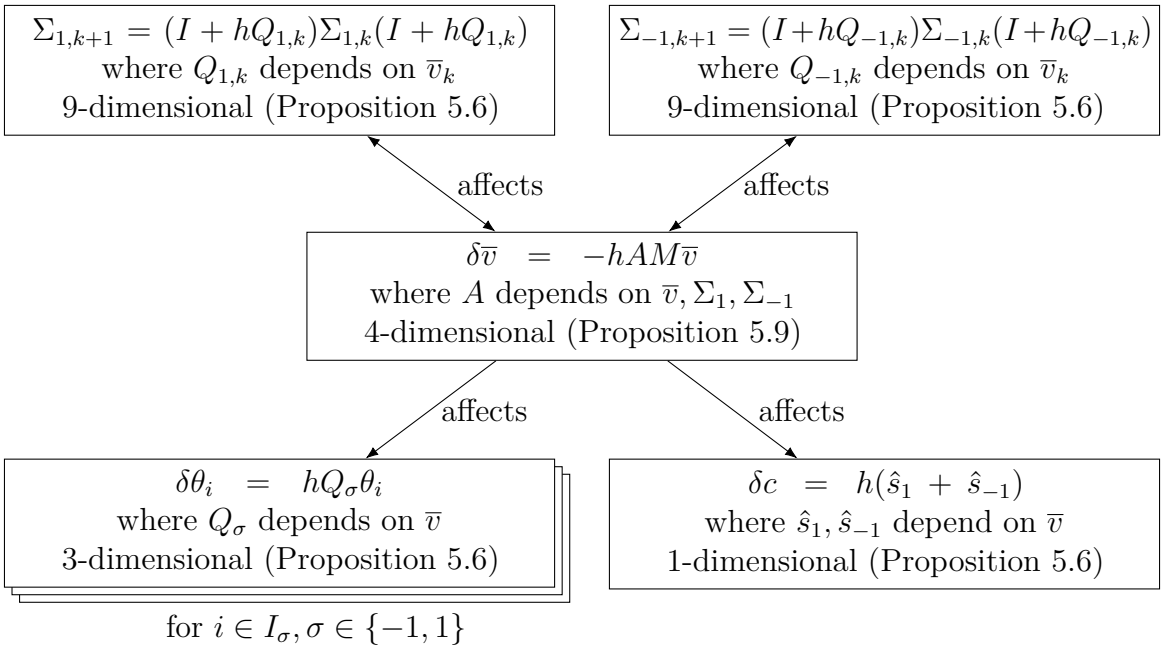


Figure 3: Decomposition into different systems that can be used to analyze the behavior of gradient descent.

5.2 Comments

Remark 5.11 (System decomposition)**.** We have so far derived different "systems", i.e. results on how quantities evolve during gradient descent. These systems and their dependencies are depicted in Figure 3. In particular, we see that the systems for Σ_1, Σ_{-1} and $\overline{v}$ together yield a 22-dimensional system that does not depend on any other quantities. This 22-dimensional system describes some central properties of the neural network parameters W although its dimension does not depend on n. These properties include:

- The loss $L_{D,\tau}(W)$ (cf. Proposition 5.10).
- Slope p_σ and intercept q_σ for both signs $\sigma \in \{-1, 1\}$.
- We can derive upper bounds on the weights W: For $i \in I_\sigma$, we obtain $|a_i| \leq \sqrt{\Sigma_{\sigma,a^2}}$ and similarly for b, w. Since c occurs in q_1, q_{-1}, we can use $|q_1|$ or $|q_{-1}|$ to derive an upper bound on c as well.

While this system has a dimension independent of n, the probability distribution over its initialization may well depend on n. If its evolution is known, the evolution $(W_k)_{k\in\mathbb{N}_0}$ can be determined by solving n independent three-dimensional systems and the one-dimensional system $\delta c = h(\hat{s}_1 + \hat{s}_{-1})$. In this thesis, we will proceed along similar lines (cf. Remark 5.15): We will first analyze the behavior of the 22-dimensional system and then apply our results to the three-dimensional systems.

In fact, the 22-dimensional system can be reduced to a 14-dimensional system: The matrices Σ_σ are always symmetric and thus effectively 6-dimensional, which reduces

the dimension from 22 to 16. Moreover, we always have

$$\begin{pmatrix} p_1 \\ p_{-1} \end{pmatrix} = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix} \begin{pmatrix} \Sigma_{1,wa} \\ \Sigma_{-1,wa} \end{pmatrix} .$$

However, removing these redundancies is not beneficial for our analysis. ◀

Remark 5.12. The components of the equation $\delta\overline{v} = -hAM\overline{v}$ in Proposition 5.9 can be interpreted as follows: Recall that

$$G_\sigma^{\mathrm{w}} = \begin{pmatrix} \Sigma_{\sigma,w^2} & 0 \\ 0 & \Sigma_{\sigma,w^2} \end{pmatrix}, \quad G_\sigma^{\mathrm{ab}} = \begin{pmatrix} \Sigma_{\sigma,a^2} & \Sigma_{\sigma,ab} \\ \Sigma_{\sigma,ab} & \Sigma_{\sigma,b^2} \end{pmatrix}, \quad G^{\mathrm{wab}} = (r_\sigma \Sigma_{\sigma,wa} + s_\sigma \Sigma_{\sigma,wb}) I_2,$$

$$\tilde{B} = \begin{pmatrix} I_2 & \alpha I_2 \\ \alpha I_2 & I_2 \end{pmatrix}, \qquad \tilde{C} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix},$$

$$\tilde{A} = \tilde{B} \begin{pmatrix} G_1^{\mathrm{w}} + G_1^{\mathrm{ab}} + hG_1^{\mathrm{wab}} & \\ & G_{-1}^{\mathrm{w}} + G_{-1}^{\mathrm{ab}} + hG_{-1}^{\mathrm{wab}} \end{pmatrix} \tilde{B} + \tilde{C} .$$

- The matrix $G_\sigma^{\mathrm{w}} \succeq 0$ describes the improvement of $\overline{v}_\sigma$ by updating the weights a_i and b_i. The larger $|w_i|$, the larger the gradients $\frac{\partial L_{D,\tau}}{\partial a_i}, \frac{\partial L_{D,\tau}}{\partial b_i}$ and the more effect does a change in a_i, b_i have on the overall function $f_{W,\tau,\sigma}$.

- The matrix G_σ^{ab} is also positive semidefinite since $\mathrm{tr}(G_\sigma^{\mathrm{ab}}) \geq 0$ and $\det(G_\sigma^{\mathrm{ab}}) = \Sigma_{\sigma,a^2}\Sigma_{\sigma,b^2} - \Sigma_{\sigma,ab}^2 \geq 0$ due to Cauchy-Schwarz. It describes the improvement of $\overline{v}_\sigma$ by updating the weights $(w_i)_{i\in I_\sigma}$. Larger values of $|a_i|, |b_i|$ mean stronger effects of changing w_i. If the vectors $(a_i)_{i\in I_\sigma}$ and $(b_i)_{i\in I_\sigma}$ are linearly dependent (perfectly correlated), then $\tilde{G}_\sigma^{\mathrm{ab}}$ only has rank one and changing the w_i cannot independently update both components of $\overline{v}_\sigma$. Recall that the components of $\overline{v}_\sigma$ are the differences of the slope and intercept of $f_{W,\tau,\sigma}$ to the optimal linear regression slope and intercept, respectively.

- The matrix $\tilde{B}$ causes an interaction between both signs $\sigma \in \{-1, 1\}$ if the leaky parameter α is nonzero. If it is zero, the hidden neurons are only active for one sign σ and do only indirectly interact via the bias c.

- The matrix $\tilde{C}$ describes the improvement of v_σ by updating the bias c. It is not block-diagonal since c is active for both signs $\sigma \in \{-1, 1\}$. However, $\tilde{C}$ only has rank one since changing c can only change q_1 and q_{-1} by the same amount. $\tilde{C}$ is positive semidefinite since it is symmetric and it has eigenvectors $e_1, e_3, (0, 1, 0, -1)$ to the eigenvalue 0 and $(0, 1, 0, 1)$ to the eigenvalue 2.

- The matrix $\tilde{G}^{\mathrm{wab}}$ represents parts of the error that (discrete) gradient descent makes when trying to approximate negative gradient flow. It arises from the additional term $\delta g_1 \cdot \delta g_2$ in the product rule for δ (Lemma 5.4 (c)) and does not need to be positive semidefinite. If h is too large, the matrix $\tilde{A}$ might therefore not be positive semidefinite. This corresponds to the fact that the loss might increase during gradient descent if the step size h is too large. We will now derive why $\tilde{A} \succeq 0$ is related to a nonincreasing loss.

By Proposition 5.10, $L_{D,\tau}(W) - \overline{v}^\top M \overline{v}$ is constant. Hence,

$$\begin{aligned}\delta L_{D,\tau}(W) &= (\delta\overline{v})^\top M\overline{v} + \overline{v}M(\delta\overline{v}) + (\delta\overline{v})^\top M(\delta\overline{v}) \\ &= -h\overline{v}^\top MAM\overline{v} - h\overline{v}^\top MAM\overline{v} + h^2\overline{v}^\top MAMAM\overline{v} \ . \end{aligned} \tag{5.2}$$

Now assume $A \succeq 0$. We have

$$MAMAM \preceq \lambda_{\max}(M)\lambda_{\max}(A)MAM \tag{5.3}$$

and hence if

$$h\lambda_{\max}(M)\lambda_{\max}(A) \leq 2 \ ,$$

we can combine Eq. (5.2) and Eq. (5.3) to obtain $\delta L_{D,\tau}(W) \leq 0$, hence the loss is non-increasing. Similarly, $A \succ 0$ and $h\lambda_{\max}(M)\lambda_{\max}(A) < 2$ yields strictly decreasing loss. In these cases, the loss and hence the function $\overline{v} \mapsto \overline{v}^\top M\overline{v}$ are Lyapunov functions. By the above argument, we find that the matrix M and hence the data points affect the evolution of $\overline{v}$ although (for small enough step size) it does not prevent that $\overline{v}$ converges to zero. However, if M is badly conditioned, the speed of convergence might deteriorate. ◀

Remark 5.13 (Discretization error)**.** We have already seen that the systems for Σ_σ and $\overline{v}$ are affected by terms that arise from the term $\delta g_1 \cdot \delta g_2$ in the "discrete product rule" of Lemma 5.4 (c). We will see that in our scenario (with small enough step size), these "disturbances" are small enough to not influence the qualitative behavior of gradient descent. There is also an invariant that holds when using negative gradient flow but breaks down when using gradient descent: In the former case, $a_i^2 + b_i^2 - w_i^2$ remains constant during the optimization for each $i \in I$. An analogous identity for linear networks has been observed in [29]. ◀

Remark 5.14 (Alternative systems)**.** In some special cases, the approach presented here only works if we modify the systems. For example, the assumption $M \succ 0$ is not satisfied if the dataset is contained in $(0, \infty)$ since this implies $M_{-1} = 0$. In this case, the system $\delta\overline{v} = -hAM\overline{v}$ could be reduced to a two-dimensional system since p_{-1} and q_{-1} are irrelevant for the loss. We will also see that the argument here does not work for $|\alpha| = 1$ since this renders the matrix B singular. The case $\alpha = 1$ corresponds to a linear activation function $\varphi(x) = x$, which implies $p_1 = p_{-1}$ and $q_1 = q_{-1}$. Similarly, the case $\alpha = -1$ corresponds to $\varphi(x) = |x|$, which implies $p_1 = -p_{-1}$. In both cases, the dimension of v could be reduced. ◀

Remark 5.15 (Proof idea)**.** We can now formulate the proof idea more precisely. We will roughly show that under certain circumstances and with high probability for a randomly sampled initialization and dataset, the following statements (up to subpolynomial factors) can be argued:

- For a modified system $\delta\overline{v} = -hA^{\text{ref}}M\overline{v}$, where $A^{\text{ref}} = A_0 - hBG^{\text{wab}}B$ will be introduced in Definition 5.26[1], we prove in Proposition 5.29 that $h\sum_{k=0}^{\infty} \|\overline{v}_k\| = O(1/n)$.

[1] We could also have used $A^{\text{ref}} \coloneqq A_0$ but this would not guarantee $A^{\text{ref}} \succeq 0$ for large h.

- The following two statements can be proven together by induction on k:
  - For the original system $\delta\overline{v} = -hAM\overline{v}$, we have $h\sum_{l=0}^{k}\|\overline{v}_l\| = O(1/n)$.
  - For $\sigma \in \{-1,1\}$, $|\Sigma_{\sigma,k} - \Sigma_{\sigma,0}|$ is "small".

  Essentially, Σ_σ does not change much because $\delta\Sigma_\sigma = hQ_\sigma\Sigma_\sigma + h\Sigma_\sigma Q_\sigma + h^2 Q_\sigma\Sigma_\sigma Q_\sigma$, where

$$h\sum_{l=0}^{k}\|Q_{\sigma,l}\| = O\left(h\sum_{l=0}^{k}\|\overline{v}_l\|\right) = O(1/n)\ .$$

  Conversely, the system $\delta\overline{v} = -hAM\overline{v}$ does not deviate much from $\delta\overline{v} = -hA^{\mathrm{ref}}M\overline{v}$ since A depends on Σ_1, Σ_{-1} which do not change much. (Also, A depends on $\overline{v}$ via $\tilde{G}^{\mathrm{quad}}$, but with an additional factor h which is assumed to be small.) One part of the induction is handled in Proposition 5.31, the overall induction is then handled in Proposition 5.33.

- In Corollary 5.35, we prove that the quantities a_i and b_i only change by $O(n^{-3/2})$: For example, for $i \in I_\sigma$, we have $\delta a_i = hr_\sigma w_i$. We will see that the initialization roughly implies $\sup_{k\in\mathbb{N}_0}|w_{i,k}| = O(n^{-1/2})$. Moreover, $\sum_{k=0}^{\infty}|r_{\sigma,k}| = O\left(\sum_{k=0}^{\infty}\|\overline{v}_k\|\right) = O(1/n)$. By a simple argument, this yields $\sup_{k\in\mathbb{N}_0}|a_{i,k} - a_{i,0}| = O(n^{-3/2})$.

- We will see in Definition 5.23 and Theorem 5.25 that the initialization roughly implies $\min_{i\in I}|a_{i,0}| = \Omega(1/n)$. Since $b_{i,0}$ is initialized to zero, this means that for n being large enough, $\operatorname{sgn}(a_{i,k}x_j + b_{i,k})$ is constant in k for all i and hence the kinks do not cross the data points. ◀

5.3 Stochastic Properties of Initialization and Dataset

When training a neural network, there are usually two sources of randomness: The dataset and the initialization. We will investigate both of them in this section, starting with the dataset.

Assumption 5.16. Let P be a probability distribution on $\mathbb{R}\times\mathbb{R}$. Let P_X be the marginal distribution of P with respect to the first component of $\mathbb{R}\times\mathbb{R}$. We impose some constraints on P:

(P1) $\int(|x|^p + |y|^p)\,\mathrm{d}P(x,y) < \infty$ for all $p \in (0,\infty)$.

(P2) For $\sigma \in \{-1,1\}$ and all $x \in \mathbb{R}$, $P_X(\sigma(0,\infty)\setminus\{x\}) > 0$, where $\sigma(0,\infty) \coloneqq \{\sigma x \mid x \in (0,\infty)\}$.

(P3) There exists a value $m_P > 0$ with $P_X((-m_P, m_P)) = 0$.

Let

$$M_{P,\sigma} \coloneqq \mathbb{E}_{x\sim P_X}\mathbb{1}_{(0,\infty)}(\sigma x)\begin{pmatrix} x^2 & x \\ x & 1\end{pmatrix}, \quad \hat{u}^0_{P,\sigma} \coloneqq \mathbb{E}_{(x,y)\sim P}\mathbb{1}_{(0,\infty)}(\sigma x)\begin{pmatrix} xy \\ y\end{pmatrix}\ .$$

These expected values are well-defined because of assumption (P1). Moreover, the matrices

$$M_x := \begin{pmatrix} x^2 & x \\ x & 1 \end{pmatrix}$$

for $x \in \mathbb{R}$ are positive semidefinite and their kernels $\operatorname{Span}\{(1,-x)^\top\}$ are disjoint for different x. Hence, for each $0 \neq v \in \mathbb{R}^2$, $v^\top M_x v$ can only be zero for one value of x. Therefore,

$$v^\top M_{P,\sigma} v = \mathbb{E}_{x \sim P_X} \mathbb{1}_{\sigma(0,\infty)}(x) v^\top M_x v > 0$$

because of assumption (P2). This shows $M_P \succ 0$. Hence, we can define

$$v_{P,\sigma}^{\text{opt}} := M_{P,\sigma}^{-1} \hat{u}_{P,\sigma}^0, \quad \tilde{v}_P^{\text{opt}} := \begin{pmatrix} v_{P,1}^{\text{opt}} \\ v_{P,-1}^{\text{opt}} \end{pmatrix}, \quad v_P^{\text{opt}} := \tilde{P} \tilde{v}_P^{\text{opt}} ,$$

where $\tilde{P}$ is the permutation matrix from Definition 5.5.
We can now formulate a fourth assumption on P:

(P4) For $\sigma \in \{-1, 1\}$, the second component of $v_{P,\sigma}^{\text{opt}} \in \mathbb{R}^2$ is zero.

Let $\mathcal{F}_{\text{aff}}$ be the class of all functions of the form

$$f_v : \mathbb{R} \to \mathbb{R}, x \mapsto \begin{cases} p_1 x + q_1 & , \text{ if } x \geq 0 \\ p_{-1} x + q_{-1} & , \text{ if } x < 0, \end{cases}$$

where $v = (p_1, p_{-1}, q_1, q_{-1})^\top \in \mathbb{R}^4$. Our last assumption is the following one:

(P5) We have

$$\inf_{f \in \mathcal{F}_{\text{aff}}} R_P(f) > \inf_{f : \mathbb{R} \to \mathbb{R}} R_P(f) .$$

Remark 5.17. In Assumption 5.16, (P1) is a technical condition ensuring the existence of all moments. Condition (P2) requires that the support of P_X intersects the positive and negative parts of the x-axis in more than one point, respectively. Condition (P3) requires that no data points can be sampled in a neighborhood of $x = 0$.
The main limitation of Assumption 5.16 lies in condition (P4), which requires the intercepts of the optimal linear regression lines for $P((X,Y) \mid \operatorname{sgn}(X) = \sigma)$ to be zero for both signs $\sigma \in \{-1, 1\}$. This can be interpreted as restricting P to a submanifold of codimension 2 of the set of all probability distributions on $\mathbb{R} \times \mathbb{R}$. However, P may still be a continuous distribution with observation noise.
Condition (P5) ensures that a neural network which is affine on $(-\infty, -m_P)$ and on (m_P, ∞) cannot achieve a close-to-optimal training loss, which is required for showing the inconsistency of a neural network estimator. ◀

We now show a particularly simple example of a probability distribution that satisfies Assumption 5.16:

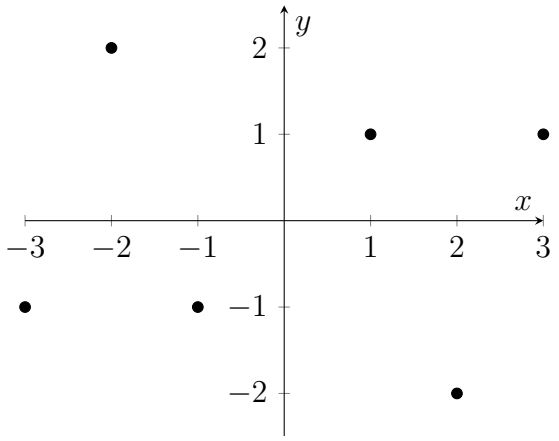


Figure 4: Example of a set of points where the uniform distribution on this set satisfies Assumption 5.16.

Example 5.18. Let P be the uniform distribution on

$$D_P \coloneqq \{(-3,-1),(-2,2),(-1,-1),(1,1),(2,-2),(3,1)\} \ .$$

These points are depicted in Figure 4. We verify the conditions from Assumption 5.16:

(P1) Since any $(x,y) \in D_P$ satisfies $|x| \leq 3$ and $|y| \leq 2$, we have

$$\int (|x|^p + |y|^p) \, \mathrm{d}P(x,y) \leq 3^p + 2^p < \infty \ .$$

(P2) The measure P_X is uniformly distributed on $\{-3,-2,-1,1,2,3\}$. For $\sigma \in \{-1,1\}$, we hence have $P(\sigma(0,\infty)) = 1/2$ and

$$P(\sigma(0,\infty) \setminus \{x\}) \geq \frac{1}{2} - P(\{x\}) \geq \frac{1}{2} - \frac{1}{6} = \frac{1}{3} > 0 \ .$$

We can also directly compute

$$M_{P,1} = \begin{pmatrix} 7/3 & 1 \\ 1 & 1/2 \end{pmatrix}, \quad M_{P,-1} = \begin{pmatrix} 7/3 & -1 \\ -1 & 1/2 \end{pmatrix}, \quad M_P = \begin{pmatrix} 7/3 & 0 & 1 & 0 \\ 0 & 7/3 & 0 & -1 \\ 1 & 0 & 1/2 & 0 \\ 0 & -1 & 0 & 1/2 \end{pmatrix}$$

and all of these matrices are positive definite.

(P3) We can choose $m_P = 1$.

(P4) We compute

$$\begin{aligned} \hat{u}^0_{P,1} &= \frac{1}{6} \begin{pmatrix} 1 \cdot 1 + 2 \cdot (-2) + 3 \cdot 1 + 0 + 0 + 0 \\ 1 + (-2) + 1 + 0 + 0 + 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \hat{u}^0_{P,-1} &= \frac{1}{6} \begin{pmatrix} 0 + 0 + 0 + (-1) \cdot (-1) + (-2) \cdot 2 + (-3) \cdot (-1) \\ 0 + 0 + 0 + (-1) + 2 + (-1) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \ , \end{aligned}$$

which implies $\hat{u}^0_P = 0$ and hence $v^{\mathrm{opt}}_P = M_P^{-1} \hat{u}^0_P = 0$.

(P5) This is intuitively clear. A formal proof can be done using Proposition 5.10 as shown in the proof of Corollary 5.39. ◀

Weights of neural networks are often initialized from uniform or Gaussian distributions (cf. e.g. [17], Section 8.4). We allow a more general class of distributions:

Assumption 5.19. Let Q^{wa} be a probability distribution on $\mathbb{R}$ that satisfies the following assumptions:

(Q1) Q^{wa} has a bounded symmetric probability density function p_Q^{wa}, i.e.

- There exists $B_Q^{\mathrm{wa}} \in \mathbb{R}$ such that $0 \leq p_Q^{\mathrm{wa}}(x) \leq B_Q^{\mathrm{wa}}$ for all $x \in \mathbb{R}$,
- $p_Q^{\mathrm{wa}}(x) = p_Q^{\mathrm{wa}}(-x)$ for all $x \in \mathbb{R}$,
- p_Q^{wa} is a probability density function of Q^{wa}:

$$Q^{\mathrm{wa}}(E) = \int_E p_Q^{\mathrm{wa}}(x)\,\mathrm{d}x$$

for all events $E \subseteq \mathbb{R}$.

(Q2) Q^{wa} is p-integrable for all $p \in (0, \infty)$, i.e.

$$\int_{\mathbb{R}} |x|^p p_Q^{\mathrm{wa}}(x)\,\mathrm{d}x < \infty$$

for all $p \in (0, \infty)$.

(Q3) Q^{wa} has variance 1:

$$\int_{\mathbb{R}} x^2 p_Q^{\mathrm{wa}}(x)\,\mathrm{d}x = 1\ .$$

This "normalization condition" does not impose any restriction on the initialization since, as we will see later, the variables can be scaled.

Example 5.20. Two examples for distributions that satisfy the conditions from Assumption 5.19 are:

- The standard normal distribution $Q^{\mathrm{wa}} = \mathcal{N}(0, 1)$.
- The uniform distribution on $[-\sqrt{3}, \sqrt{3}]$, which has variance 1. ◀

Next, we want to define probability spaces which involve random sampling and random initialization:

Definition 5.21. Let P be a probability distribution on $\mathbb{R} \times \mathbb{R}$ satisfying Assumption 5.16 and let Q^{wa} be a probability distribution on $\mathbb{R}$ satisfying Assumption 5.19. Furthermore, let $c_a, c_w > 0$ be arbitrary constants.

For a number $n \geq 1$ of neurons and $N \geq 1$ of data points, we consider probability spaces $(\Omega_{n,N}, \mathcal{F}_{n,N}, P_{n,N})$ with independent random variables D and W_0 distributed as follows:

- The dataset $D = ((x_1, y_1), \ldots, (x_N, y_N)) \in (\mathbb{R} \times \mathbb{R})^N$ consists of i.i.d. pairs (x_j, y_j) that are distributed according to P. In other words, $D \sim P^N$.
- The components of the vector $W_0 = (a_{\cdot,0}, b_{\cdot,0}, c_0, w_{\cdot,0}) \in \mathbb{R}^{3n+1}$ are independent and distributed as follows:

$$\begin{aligned} \sqrt{\frac{1}{c_a}} a_{i,0} &\sim Q^{\mathrm{wa}} \\ \sqrt{\frac{n}{c_w}} w_{i,0} &\sim Q^{\mathrm{wa}} \\ b_{i,0} &= 0 \\ c_0 &= 0 \ . \end{aligned}$$

  We denote the distribution of W_0 on $\mathbb{R}^{3n+1}$ by P_n.

The initialization W_0 induces a sign pattern $\tau = \tau(W_0)$ as defined in Definition 5.1. We can now define the random variables $W_k, k \in \mathbb{N}_0$, recursively as usual:

$$W_{k+1} \coloneqq W_k - h \nabla L_{D,\tau}(W_k) \ . \qquad \blacktriangleleft$$

Example 5.22. In the popular initialization scheme proposed by He et al. [20], the components of W_0 are sampled independently as follows:

$$\begin{aligned} a_{i,0} &\sim \mathcal{N}(0, 2) \\ w_{i,0} &\sim \mathcal{N}(0, 2/n) \\ b_{i,0} &= 0 \\ c_0 &= 0 \ . \end{aligned}$$

This is covered by Definition 5.21 by setting $Q^{\mathrm{wa}} = \mathcal{N}(0, 1)$ as in Example 5.20 and choosing $c_a = c_w = 2$. $\blacktriangleleft$

In the following, we will define an event that is likely to occur and guarantees some important properties of the initialization and dataset:

Definition 5.23. For $\varepsilon, \gamma > 0$ and $n \geq 1$, let $E^{\mathrm{W}}_{n,\varepsilon,\gamma}$ denote the set of all $W_0 \in \mathbb{R}^{3n+1}$ where the following properties hold:

(W1) $b_{i,0} = c_0 = 0$,

(W2) $\max_i |w_{i,0}| \leq n^{-1/2+\varepsilon}$,

(W3) $\max_i |a_{i,0}| \leq n^{\varepsilon}$,

(W4) $\min_i |a_{i,0}| \geq n^{-(1+\gamma)}$,

(W5) $\Sigma_{\sigma,a^2,0} \in [nc_a/4, nc_a]$ for all $\sigma \in \{-1, 1\}$,

(W6) $\Sigma_{\sigma,w^2,0} \in [c_w/4, c_w]$ for all $\sigma \in \{-1, 1\}$,

(W7) $|\Sigma_{\sigma,wa,0}| \leq n^{\varepsilon}$ for all $\sigma \in \{-1, 1\}$.

For $\varepsilon > 0$ and $N \geq 1$, let $E^{\mathrm{D}}_{N,\varepsilon}$ denote the set of all datasets $D \in (\mathbb{R} \times \mathbb{R})^N$ where the following properties hold:

(D1) v_D^{opt} is well-defined, i.e. M_D is invertible, and $\|v_P^{\mathrm{opt}} - v_D^{\mathrm{opt}}\|_\infty \leq N^{(\varepsilon-1)/2}$.

(D2) $\lambda_{\min}(M_D) \geq \frac{1}{2}\lambda_{\min}(M_P)$ and $\lambda_{\max}(M_D) \leq 2\lambda_{\max}(M_P)$.

(D3) $\min_j |x_j| \geq m_P$.

Finally, for $\varepsilon, \gamma > 0$ and $n, N \geq 1$, define the event where all of the previous properties hold:

$$E_{n,N,\varepsilon,\gamma} \coloneqq \{\omega \in \Omega_{n,N} \mid W_0(\omega) \in E^{\mathrm{W}}_{n,\varepsilon,\gamma} \text{ and } D(\omega) \in E^{\mathrm{D}}_{N,\varepsilon}\} \; . \qquad \blacktriangleleft$$

In order to bound $P_{n,N}(E_{n,N,\varepsilon,\gamma})$ asymptotically and bound many other quantities later on, we use asymptotic notation to simplify our calculations:

Definition 5.24 (Asymptotic notation)**.** We want to prove what happens in the limit as $n \to \infty$ and $N \to \infty$. Some other parameters may vary along with n, N while other parameters remain constant:

- The variable parameters θ_{var} comprise the number n of hidden neurons, the number N of data points, the step size $h > 0$, step count variables such as $k, l \in \mathbb{N}_0$ and the randomness $\omega \in \Omega_{n,N}$.
- The fixed parameters θ_{const} comprise the probability distributions P and Q^{wa}, the constants c_a, c_w in Definition 5.21, the LeakyReLU parameter α and further parameters used in the proof such as ε and γ.

Given a domain $\mathcal{D}_{\mathrm{var}}$ for θ_{var} and any assignment of θ_{const}, we may be given a function $f_{\theta_{\mathrm{const}}} : \mathcal{D}_{\mathrm{var}} \to \overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{-\infty, \infty\}, \theta_{\mathrm{var}} \mapsto f_{\theta_{\mathrm{const}}}(\theta_{\mathrm{var}})$. We then define complexity classes as follows:

$$\begin{aligned}
O(f_{\theta_{\mathrm{const}}}) &\coloneqq \{g : \mathcal{D}_{\mathrm{var}} \to \overline{\mathbb{R}} \mid \exists C_{\theta_{\mathrm{const}}} > 0, n_0 \in \mathbb{N}_0, N_0 \in \mathbb{N}_0 \text{ s.t.} \\
&\qquad \text{for all } \theta_{\mathrm{var}} \in \mathcal{D}_{\mathrm{var}} \text{ with } n \geq n_0, N \geq N_0\text{:} \\
&\qquad g(\theta_{\mathrm{var}}) \leq C_{\theta_{\mathrm{const}}} f_{\theta_{\mathrm{const}}}(\theta_{\mathrm{var}})\} \\
\Theta(f_{\theta_{\mathrm{const}}}) &\coloneqq \{g : \mathcal{D}_{\mathrm{var}} \to \overline{\mathbb{R}} \mid g \in O(f_{\theta_{\mathrm{const}}}) \text{ and } f_{\theta_{\mathrm{const}}} \in O(g)\} \\
o(f_{\theta_{\mathrm{const}}}) &\coloneqq \{g : \mathcal{D}_{\mathrm{var}} \to \overline{\mathbb{R}} \mid \forall C > 0 \; \exists n_0 \in \mathbb{N}_0, N_0 \in \mathbb{N}_0 \text{ s.t.} \\
&\qquad \text{for all } \theta_{\mathrm{var}} \in \mathcal{D}_{\mathrm{var}} \text{ with } n \geq n_0, N \geq N_0\text{:} \\
&\qquad g(\theta_{\mathrm{var}}) \leq C f_{\theta_{\mathrm{const}}}(\theta_{\mathrm{var}})\} \\
\Omega(f_{\theta_{\mathrm{const}}}) &\coloneqq \{g : \mathcal{D}_{\mathrm{var}} \to \overline{\mathbb{R}} \mid f_{\theta_{\mathrm{const}}} \in O(g)\} \; .
\end{aligned}$$

Note that the constant $C_{\theta_{\mathrm{const}}}$ in the O-notation may depend on θ_{const} but not on θ_{var}. As usual, we write $f(\theta_{\mathrm{var}}) = O(g(\theta_{\mathrm{var}}))$ instead of $f \in O(g)$ and similarly for Θ, o and Ω. We may also write $f(\theta_{\mathrm{var}}) = 1 - O(g(\theta_{\mathrm{var}}))$ instead of $f \in 1 - O(g)$, which is equivalent to $1 - f \in O(g)$.

If no domain $\mathcal{D}_{\text{var}}$ is specified, we allow $n, N \in \mathbb{N}_{\geq 1}, k, l \in \mathbb{N}_0, h > 0, \omega \in \Omega_{n,N}$. However, we frequently impose further restrictions on the domain. For example, if we restrict the domain $\mathcal{D}_{\text{var}}$ by requiring $\omega \in E_{n,N,\varepsilon,\gamma}$, property (W5) implies

$$\Sigma_{\sigma,a^2,0} = \Theta(n) \ .$$

Note that $\Sigma_{\sigma,a^2,0}$ implicitly depends on n and ω. Without the restriction to $\omega \in E_{n,N,\varepsilon,\gamma}$, this asymptotic statement would not be true.
As an other example, we will later require $N \geq \varrho n^2$ for a fixed parameter $\varrho > 0$. Then, we have $n = O(\sqrt{N})$. ◀

The following theorem contains all stochastic properties that we will need here:

Theorem 5.25. *Let $\varepsilon, \gamma > 0$. Then,*

$$P_{n,N}(E_{n,N,\varepsilon,\gamma}) = 1 - O(n^{-\gamma} + N^{-\gamma'})$$

for all $\gamma' > 0$.[2]

Proof. This is proven in Appendix A:

- Proposition A.3 shows $P_{n,N}(W \notin E^{\text{W}}_{n,\varepsilon,\gamma}) = O(n^{-\gamma})$.
- Proposition A.5 shows $P_{n,N}(D \notin E^{\text{D}}_{N,\varepsilon}) = O(N^{-\gamma'})$.

By the union bound, it follows that

$$\begin{aligned} 1 - P_{n,N}(E_{n,N,\varepsilon,\gamma}) &= P_{n,N}(W \notin E^{\text{W}}_{n,\varepsilon,\gamma} \text{ or } D \notin E^{\text{D}}_{N,\varepsilon}) \\ &\leq P_{n,N}(W \notin E^{\text{W}}_{n,\varepsilon,\gamma}) + P_{n,N}(W \notin E^{\text{D}}_{N,\varepsilon}) \\ &= O(n^{-\gamma} + N^{-\gamma'}) \ . \end{aligned}$$

□

5.4 Interactions between Systems

In the following, we prove asymptotic results using O-notation as defined in Definition 5.24 about the quantities defined in Definition 5.5 conditioned on the event $E_{n,N,\varepsilon,\gamma}$ from Definition 5.23. First, we want to investigate a reference matrix that is close to A_0:

Definition 5.26. Let $A^{\text{ref}} := B(G_0^{\text{w}} + G_0^{\text{ab}})B + C$, where $B, G^{\text{w}}, G^{\text{ab}}, C$ are defined in Definition 5.5. ◀

Lemma 5.27. *Let $|\alpha| \neq 1$ and $\varepsilon, \gamma > 0$. Then, for $\omega \in E_{n,N,\varepsilon,\gamma}$, the matrix A^{ref} is of the form*

$$A^{\text{ref}} = \begin{pmatrix} A_1^{\text{ref}} & \\ & A_2^{\text{ref}} \end{pmatrix}$$

[2]This means that the term corresponding to N decreases faster than the inverse of any positive polynomial.

with $0 \prec A_1^{\mathrm{ref}}, A_2^{\mathrm{ref}} \in \mathbb{R}^{2\times 2}$ *and*

$$\begin{aligned}\mathrm{eig}(A_1^{\mathrm{ref}}) &\subseteq \left[(1-|\alpha|)^2\frac{nc_a+c_w}{4}, (1+|\alpha|)^2(nc_a+c_w)\right]\\ \mathrm{eig}(A_2^{\mathrm{ref}}) &\subseteq \left[(1-|\alpha|)^2\frac{c_w}{4}, (1+|\alpha|)^2c_w+2\right] \ .\end{aligned}$$

In the asymptotic notation of Definition 5.24, this means

$$\lambda_{\min}(A_1^{\mathrm{ref}}) = \Theta(n), \quad \lambda_{\max}(A_1^{\mathrm{ref}}) = \Theta(n), \quad \lambda_{\min}(A_2^{\mathrm{ref}}) = \Theta(1), \quad \lambda_{\max}(A_2^{\mathrm{ref}}) = \Theta(1) \ .$$

Proof. Since $b_{i,0} = 0$ by initialization property (W1) in Definition 5.23, we have $\Sigma_{\sigma,ab,0} = \Sigma_{\sigma,b^2,0} = 0$. This yields

$$G_{\sigma,0}^{\mathrm{w}} + G_{\sigma,0}^{\mathrm{ab}} = \begin{pmatrix}\Sigma_{\sigma,w^2,0}+\Sigma_{\sigma,a^2,0} & \\ & \Sigma_{\sigma,w^2,0}\end{pmatrix} =: \begin{pmatrix}\xi_\sigma & \\ & \zeta_\sigma\end{pmatrix} \ .$$

Hence,

$$\begin{aligned}G_0^{\mathrm{w}} + G_0^{\mathrm{ab}} &= \tilde{P}(\tilde{G}_0^{\mathrm{w}} + \tilde{G}_0^{\mathrm{ab}})\tilde{P} = \tilde{P}\begin{pmatrix}G_{1,0}^{\mathrm{w}}+G_{1,0}^{\mathrm{ab}} & \\ & G_{-1,0}^{\mathrm{w}}+G_{-1,0}^{\mathrm{ab}}\end{pmatrix}\tilde{P}\\ &= \tilde{P}\begin{pmatrix}\xi_1 & & & \\ & \zeta_1 & & \\ & & \xi_{-1} & \\ & & & \zeta_{-1}\end{pmatrix}\tilde{P} = \begin{pmatrix}\xi_1 & & & \\ & \xi_{-1} & & \\ & & \zeta_1 & \\ & & & \zeta_{-1}\end{pmatrix} =: \begin{pmatrix}G_1 & \\ & G_2\end{pmatrix} \ .\end{aligned}$$

We have seen in Definition 5.5 that

$$B = \begin{pmatrix}\hat{B} & \\ & \hat{B}\end{pmatrix}, \quad \hat{B} = \begin{pmatrix}1 & \alpha\\ \alpha & 1\end{pmatrix}, \quad C = \begin{pmatrix}0 & 0 & & \\ 0 & 0 & & \\ & & 1 & 1\\ & & 1 & 1\end{pmatrix} =: \begin{pmatrix}0 & \\ & \hat{C}\end{pmatrix} \ .$$

Using the previous results, we obtain

$$A^{\mathrm{ref}} = \begin{pmatrix}\hat{B}G_1\hat{B} & \\ & \hat{B}G_2\hat{B}+\hat{C}\end{pmatrix} =: \begin{pmatrix}A_1^{\mathrm{ref}} & \\ & A_2^{\mathrm{ref}}\end{pmatrix} \ .$$

It remains to investigate the eigenvalues of these symmetric blocks. Properties (W5) and (W6) yield

$$\mathrm{eig}(G_1) = \{\xi_1, \xi_{-1}\} \subseteq \left[\frac{nc_a+c_w}{4}, nc_a+c_w\right], \quad \mathrm{eig}(G_2) = \{\zeta_1, \zeta_{-1}\} \subseteq \left[\frac{c_w}{4}, c_w\right] \ .$$

The matrix $\hat{B}$ has eigenvectors $(1,1)$ to the eigenvalue $1+\alpha$ and $(1,-1)$ to the eigenvalue $1-\alpha$. The eigenvalues of $\hat{B}$ are thus $1+|\alpha|$ and $1-|\alpha|$. Since $\hat{B}$ is symmetric, its singular values are the absolute values of the eigenvalues. Hence $\sigma_{\max}(\hat{B}) = 1+|\alpha|$ and $\sigma_{\min}(\hat{B}) = |1-|\alpha|| > 0$ since we assumed $|\alpha| \neq 1$. Finally, $\mathrm{eig}(\hat{C}) = \{0,2\}$. This yields

$$\begin{aligned}\mathrm{eig}(A_1^{\mathrm{ref}}) &= \mathrm{eig}(\hat{B}G_1\hat{B}) \in \left[(1-|\alpha|)^2\frac{nc_a+c_w}{4}, (1+|\alpha|)^2(nc_a+c_w)\right]\\ \mathrm{eig}(A_2^{\mathrm{ref}}) &= \mathrm{eig}(\hat{B}G_2\hat{B}+\hat{C}) \in \left[(1-|\alpha|)^2\frac{c_w}{4}, (1+|\alpha|)^2c_w+2\right] \ .\end{aligned}$$

□

The following lemma extends the facts from Section 4 in order to handle a quantity that will occur in many theorems afterwards.

Lemma 5.28. *Let $A \in \mathbb{R}^{m\times m}$ be symmetric and let $0 \prec M \in \mathbb{R}^{m\times m}$. Then,* $\mathrm{eig}(AM) \subseteq \mathbb{R}$ *and*

$$\begin{aligned}\lambda_{\max}(AM) &= \lambda_{\max}(M^{1/2}AM^{1/2}) \leq \lambda_{\max}(A)\lambda_{\max}(M)\\ \lambda_{\min}(AM) &= \lambda_{\min}(M^{1/2}AM^{1/2}) \geq \lambda_{\min}(A)\lambda_{\min}(M)\ .\end{aligned}$$

Proof. Since $M \succ 0$, there exists a square root $M^{1/2} \succ 0$ of M. Thus, AM is similar to the symmetric matrix $M^{1/2}AM^{1/2} = M^{1/2}AMM^{-1/2}$ which has real eigenvalues. Hence, AM has real eigenvalues and

$$\lambda_{\max}(AM) = \lambda_{\max}(M^{1/2}AM^{1/2}), \quad \lambda_{\min}(AM) = \lambda_{\min}(M^{1/2}AM^{1/2})\ .$$

Moreover, for $v \in \mathbb{R}^m$, we have

$$\begin{aligned}v^\top M^{1/2}AM^{1/2}v &\leq v^\top M^{1/2}\lambda_{\max}(A)M^{1/2}v = \lambda_{\max}(A)v^\top Mv \leq \lambda_{\max}(A)\lambda_{\max}(M)v^\top v\\ v^\top M^{1/2}AM^{1/2}v &\geq v^\top M^{1/2}\lambda_{\min}(A)M^{1/2}v = \lambda_{\min}(A)v^\top Mv \geq \lambda_{\min}(A)\lambda_{\min}(M)v^\top v\ ,\end{aligned}$$

which shows $\lambda_{\max}(AM) \leq \lambda_{\max}(A)\lambda_{\max}(M)$ and $\lambda_{\min}(AM) \geq \lambda_{\min}(A)\lambda_{\min}(M)$. □

Proposition 5.29. *Let $|\alpha| \neq 1$, $\varepsilon \in (0,1)$, $\gamma > 0$ and $\varrho > 0$. Then, for $\omega \in E_{n,N,\varepsilon,\gamma}$ and*

$$h \leq \frac{1}{\lambda_{\max}(A^{\mathrm{ref}}M_D)}\ ,$$

with $\lambda_{\max}(A^{\mathrm{ref}}M_D)$ as in Lemma 5.28 and $M_D = M$ as in Definition 5.5, we have for $N \geq \varrho n^2$:

$$\begin{aligned}h\sum_{k=0}^{\infty}\|(I - hA^{\mathrm{ref}}M_D)^k\overline{v}_0\|_\infty &= O(n^{\varepsilon-1})\\ h\sum_{k=0}^{\infty}\|(I - hA^{\mathrm{ref}}M_D)^k\|_\infty &= O(1)\ .\end{aligned}$$

Proof. By definition, we have $\overline{v}_0 = v_0 - v^{\mathrm{opt}}$, where

$$|v_0| = \begin{pmatrix}|p_{1,0}|\\ |p_{-1,0}|\\ |q_{1,0}|\\ |q_{-1,0}|\end{pmatrix} = \begin{pmatrix}|\Sigma_{1,wa,0} + \alpha\Sigma_{-1,wa,0}|\\ |\Sigma_{-1,wa,0} + \alpha\Sigma_{1,wa,0}|\\ |\Sigma_{1,wb,0} + \alpha\Sigma_{-1,wb,0}|\\ |\Sigma_{-1,wb,0} + \alpha\Sigma_{1,wb,0}|\end{pmatrix} \overset{\text{(W1), (W7)}}{\leq} \begin{pmatrix}(1+|\alpha|)n^\varepsilon\\ (1+|\alpha|)n^\varepsilon\\ 0\\ 0\end{pmatrix}$$

and $v^{\mathrm{opt}} = v_D^{\mathrm{opt}} = v_P^{\mathrm{opt}} + (v_D^{\mathrm{opt}} - v_P^{\mathrm{opt}})$ with

$$|v_P^{\mathrm{opt}}| \overset{\text{(P4)}}{=} \begin{pmatrix}O(1)\\ O(1)\\ 0\\ 0\end{pmatrix}, \quad \|v_D^{\mathrm{opt}} - v_P^{\mathrm{opt}}\|_\infty \overset{\text{(D1)}}{\leq} N^{(\varepsilon-1)/2} \overset{\varepsilon-1<0}{\leq} \varrho^{(\varepsilon-1)/2}n^{\varepsilon-1} = O(n^{\varepsilon-1})\ .$$

Thus, we can group

$$\overline{v}_0 = \begin{pmatrix} \overline{v}_{0,1} \\ \overline{v}_{0,2} \end{pmatrix}$$

with $\overline{v}_{0,1}, \overline{v}_{0,2} \in \mathbb{R}^2$ and $\|\overline{v}_{0,1}\|_\infty = O(n^\varepsilon)$, $\|\overline{v}_{0,2}\|_\infty = O(n^{\varepsilon-1})$.
Now, we want to apply Proposition B.2 using the eigenvalue bounds from Lemma 5.27: Because of (D2),

$$\lambda_{\min}(M_D), \lambda_{\max}(M_D) \in \left[\frac{1}{2}\lambda_{\min}(M_P), 2\lambda_{\max}(M_P)\right]$$

and therefore $\lambda_{\min}(M_D) = \Theta(1)$, $\lambda_{\max}(M_D) = \Theta(1)$. Thus,

$$\begin{aligned}
\lambda &\coloneqq \lambda_{\min}(A_1^{\text{ref}})\lambda_{\min}(M_D) = \Theta(n)\Theta(1) = \Theta(n) \\
\beta &\coloneqq \frac{\lambda_{\max}(A_2^{\text{ref}})\lambda_{\max}(M_D)}{\lambda - \lambda_{\max}(A_2^{\text{ref}})\lambda_{\max}(M_D)} = \frac{\Theta(1)\Theta(1)}{\Theta(n) - \Theta(1)\Theta(1)} = \Theta(1/n)
\end{aligned}$$

and the conditions of Proposition B.2 are satisfied for n large enough (with $A = A^{\text{ref}}, M = M_D, m_1 = m_2 = 2, m = 4$):

$$\begin{aligned}
\lambda &\geq \left(1 + 2\sqrt{m_1}\sqrt{\text{cond}(M_D)}\right)\lambda_{\max}(A_2^{\text{ref}})\lambda_{\max}(M_D) = \Theta(1) \\
\beta &\leq \frac{1}{2\sqrt{m_1}\sqrt{\text{cond}(M_D)}} = \Theta(1) \\
h &\leq \frac{1}{\lambda_{\max}(AM)} \ .
\end{aligned}$$

Observe that

$$\lambda_{\min}(A^{\text{ref}}M_D) \overset{\text{Lemma 5.28}}{\geq} \lambda_{\min}(A^{\text{ref}})\lambda_{\min}(M_D) = \Theta(1) \ .$$

Hence, Proposition B.2 yields for n large enough:

$$\begin{aligned}
h\sum_{k=0}^{\infty}\|(I - hA^{\text{ref}}M_D)^k\|_2 &\leq \frac{\sqrt{\text{cond}(M_D)}}{\lambda_{\min}(A^{\text{ref}}M_D)} = O(1) \\
h\sum_{k=0}^{\infty}\|(I - hA^{\text{ref}}M_D)^k\overline{v}_0\|_2 &\leq 2\sqrt{\text{cond}(M_D)}\left(\left(\frac{1}{\lambda} + \frac{2\sqrt{m_1}\sqrt{\text{cond}(M)}\beta}{\lambda_{\min}(A^{\text{ref}}M_D)}\right)\|\overline{v}_{0,1}\|_2\right. \\
&\qquad \left. + \frac{1}{\lambda_{\min}(A^{\text{ref}}M_D)}\|\overline{v}_{0,2}\|_2\right) \\
&= \Theta(1)\left((O(1/n) + O(1/n))\,O(n^\varepsilon) + O(1)O(n^{\varepsilon-1})\right) \\
&= O(n^{\varepsilon-1}) \ .
\end{aligned}$$

Since $\|\cdot\|_2 \leq \sqrt{4}\|\cdot\|_\infty$ on $\mathbb{R}^{4\times 4}$ as mentioned in Definition 4.1, the claim follows. □

Now, we want to investigate how much θ_i and Σ_σ change during gradient descent. We first derive a general result without restriction to $E_{n,N,\varepsilon,\gamma}$ and then derive further results that hold on $E_{n,N,\varepsilon,\gamma}$.

Definition 5.30. Let

$$\kappa_{u,k} \coloneqq h\sum_{l=0}^{k}\|u_l\|_\infty, \qquad \tilde{Q} \coloneqq \begin{pmatrix}0&0&1\\0&0&1\\1&1&0\end{pmatrix},$$
$$\mathbf{1}_m \coloneqq \begin{pmatrix}1\\\vdots\\1\end{pmatrix} \in \mathbb{R}^m, \quad \mathbf{1}_{m\times m} \coloneqq \begin{pmatrix}1&\dots&1\\\vdots&&\vdots\\1&\dots&1\end{pmatrix} \in \mathbb{R}^{m\times m}\ . \qquad \blacktriangleleft$$

Now, we can state the general result, which resembles a first-order Taylor approximation:[3]

Proposition 5.31. *Here, $|\cdot|$ and $\leq$ for matrices and vectors should be understood component-wise. Let $k \in \mathbb{N}_0$, $\sigma \in \{-1,1\}$ and $i \in I_\sigma$. Then,*

$$|\theta_{i,k} - \theta_{i,0}| \leq \kappa_{u,k}\tilde{Q}|\theta_{i,0}| + 2\kappa_{u,k}^2 e^{2\kappa_{u,k}}\|\theta_{i,0}\|_\infty \mathbf{1}_3$$
$$|\Sigma_{\sigma,k} - \Sigma_{\sigma,0}| \leq \kappa_{u,k}(\tilde{Q}|\Sigma_{\sigma,0}| + |\Sigma_{\sigma,0}|\tilde{Q}) + 8\kappa_{u,k}^2 e^{4\kappa_{u,k}}\|\Sigma_{\sigma,0}\|_\infty \mathbf{1}_{3\times 3}\ .$$

Proof. We will use the following identities for arbitrary matrices $A, B \in \mathbb{R}^{m\times m}$ and vectors $v \in \mathbb{R}^m$ that are easy to verify:

$$|AB| \leq |A|\cdot|B|, \quad |A+B| \leq |A| + |B|, \quad |Av| \leq |A|\cdot|v|,$$
$$|A| \leq \|A\|_\infty \cdot \mathbf{1}_{m\times m}, \quad |v| \leq \|v\|_\infty \cdot \mathbf{1}_m\ .$$

Define

$$\begin{aligned}
\tilde{Q}_k &\coloneqq h\sum_{l=0}^{k} Q_{\sigma,l}\\
s_k &\coloneqq h\sum_{l=0}^{k}\|Q_{\sigma,l}\|_\infty\\
E_k &\coloneqq \Big((I+hQ_{\sigma,k})\cdot\ldots\cdot(I+hQ_{\sigma,0}) - \tilde{Q}_k - I\Big)\\
F_k &\coloneqq (I+hQ_{\sigma,k})\cdot\ldots\cdot(I+hQ_{\sigma,0})\Sigma_{\sigma,0}(I+hQ_{\sigma,0})\cdot\ldots\cdot(I+hQ_{\sigma,k})\\
&\quad - \tilde{Q}_k\Sigma_{\sigma,0} - \Sigma_{\sigma,0}\tilde{Q}_k - \Sigma_{\sigma,0}\ .
\end{aligned}$$

Then, by Proposition 5.6, we have for $k \geq 1$:

$$\begin{aligned}
|\theta_{i,k} - \theta_{i,0}| &= |((I+hQ_{\sigma,k-1})\cdot\ldots\cdot(I+hQ_{\sigma,0}) - I)\theta_{i,0}|\\
&= \left|\tilde{Q}_{k-1}\theta_{i,0} + E_{k-1}\theta_{i,0}\right| \leq |\tilde{Q}_{k-1}||\theta_{i,0}| + \|E_{k-1}\|_\infty\|\theta_{i,0}\|_\infty\mathbf{1}_3 \qquad (5.4)
\end{aligned}$$

and

$$\begin{aligned}
&|\Sigma_{\sigma,k} - \Sigma_{\sigma,0}|\\
&= |(I+hQ_{\sigma,k-1})\cdot\ldots\cdot(I+hQ_{\sigma,0})\Sigma_{\sigma,0}(I+hQ_{\sigma,0})\cdot\ldots\cdot(I+hQ_{\sigma,k-1}) - \Sigma_{\sigma,0}|\\
&= \left|\tilde{Q}_{k-1}\Sigma_{\sigma,0} + \Sigma_{\sigma,0}\tilde{Q}_{k-1} + F_{k-1}\right|
\end{aligned}$$

[3] In the "first-order term", the matrices are still sparse. "Higher-order" approximations are not useful for our purpose.

$$\leq |\tilde{Q}_{k-1}||\Sigma_{\sigma,0}| + |\Sigma_{\sigma,0}||\tilde{Q}_{k-1}| + \|F_{k-1}\|_\infty \mathbf{1}_{3\times 3} \ . \tag{5.5}$$

Observe that for $k \geq 0$,

$$(I + hQ_{\sigma,k}) \cdot \ldots \cdot (I + hQ_{\sigma,0}) = \sum_{l=0}^{k+1} \sum_{0\leq k_1<\ldots<k_l\leq k} hQ_{k_l} \cdot \ldots \cdot hQ_{k_1}$$

and hence

$$\begin{aligned}
\|E_k\|_\infty &= \left\| \sum_{l=2}^{k+1} \sum_{0\leq k_1<\ldots<k_l\leq k} hQ_{\sigma,k_l} \cdot \ldots \cdot hQ_{\sigma,k_1} \right\|_\infty \\
&\leq \sum_{l=2}^{k+1} \sum_{0\leq k_1<\ldots<k_l\leq k} \|hQ_{\sigma,k_l}\|_\infty \cdot \ldots \cdot \|hQ_{\sigma,k_1}\|_\infty \\
&= (1 + \|hQ_{\sigma,k}\|_\infty) \cdot \ldots \cdot (1 + \|hQ_{\sigma,0}\|_\infty) - s_k - 1 \\
&\leq e^{\|hQ_{\sigma,k}\|_\infty} \cdot \ldots \cdot e^{\|hQ_{\sigma,0}\|_\infty} - s_k - 1 \\
&= e^{s_k} - s_k - 1 = \sum_{l=2}^{\infty} \frac{s_k^l}{l!} = s_k^2 \sum_{l=0}^{\infty} \frac{s_k^l}{(l+2)!} \overset{(l+2)!\geq 2(l!)}{\leq} \frac{1}{2} s_k^2 e^{s_k} \ .
\end{aligned}$$

Similarly, we can use the index set $\mathcal{I} := \{0,\ldots,k+1\}^2 \setminus \{(0,0),(1,0),(0,1)\}$ to derive

$$\begin{aligned}
\|F_k\|_\infty &= \left\| \sum_{(l,l')\in\mathcal{I}} \sum_{\substack{0\leq k_1<\ldots<k_l\leq k \\ 0\leq k_1'<\ldots<k_{l'}'\leq k}} hQ_{\sigma,k_l} \cdots hQ_{\sigma,k_1} \Sigma_{\sigma,0} hQ_{\sigma,k_1'} \cdots hQ_{\sigma,k_{l'}'} \right\|_\infty \\
&\leq \sum_{(l,l')\in\mathcal{I}} \sum_{\substack{0\leq k_1<\ldots<k_l\leq k \\ 0\leq k_1'<\ldots<k_{l'}'\leq k}} h\|Q_{\sigma,k_l}\|_\infty \cdots h\|Q_{\sigma,k_1}\|_\infty \|\Sigma_{\sigma,0}\|_\infty h\|Q_{\sigma,k_1'}\|_\infty \cdots h\|Q_{\sigma,k_{l'}'}\|_\infty \\
&\leq e^{s_k}\|\Sigma_{\sigma,0}\|_\infty e^{s_k} - s_k\|\Sigma_{\sigma,0}\|_\infty - \|\Sigma_{\sigma,0}\|_\infty s_k - \|\Sigma_{\sigma,0}\|_\infty \\
&= (e^{2s_k} - 2s_k - 1)\|\Sigma_{\sigma,0}\|_\infty \\
&\leq 2s_k^2 e^{2s_k}\|\Sigma_{\sigma,0}\|_\infty \ .
\end{aligned}$$

Obviously,

$$|\tilde{Q}_k| \leq h \sum_{l=0}^{k} |Q_{\sigma,l}| = h \begin{pmatrix} 0 & 0 & \sum_{l=0}^k |r_{\sigma,l}| \\ 0 & 0 & \sum_{l=0}^k |s_{\sigma,l}| \\ \sum_{l=0}^k |r_{\sigma,l}| & \sum_{l=0}^k |s_{\sigma,l}| & \end{pmatrix} \leq \kappa_{u,k}\tilde{Q} \ .$$

We also have $s_k \leq 2\kappa_{u,k}$ since

$$\|Q_{\sigma,l}\|_\infty = \max_{i\in\{1,\ldots,3\}} \sum_{j=1}^{3} |(Q_{\sigma,l})_{ij}| = |r_{\sigma,l}| + |s_{\sigma,l}| \leq 2\|u_k\|_\infty \ .$$

We now obtain for $k \geq 1$:

$$\begin{aligned}
|\theta_{i,k} - \theta_{i,0}| &\overset{(5.4)}{\leq} |\tilde{Q}_{k-1}||\theta_{i,0}| + \|E_{k-1}\|_\infty \|\theta_{i,0}\|_\infty \mathbf{1}_3 \\
&\leq \kappa_{u,k-1}\tilde{Q}|\theta_{i,0}| + \frac{1}{2} s_{k-1}^2 e^{s_{k-1}} \|\theta_{i,0}\|_\infty \mathbf{1}_3 \\
&\leq \kappa_{u,k-1}\tilde{Q}|\theta_{i,0}| + 2\kappa_{u,k-1}^2 e^{2\kappa_{u,k-1}} \|\theta_{i,0}\|_\infty \mathbf{1}_3
\end{aligned}$$

$$\begin{aligned}|\Sigma_{\sigma,k}-\Sigma_{\sigma,0}| &\overset{(5.5)}{\leq} |\tilde{Q}_{k-1}||\Sigma_{\sigma,0}|+|\Sigma_{\sigma,0}||\tilde{Q}_{k-1}|+\|F_{k-1}\|_\infty \mathbf{1}_{3\times 3}\\ &\leq \kappa_{u,k-1}(\tilde{Q}|\Sigma_{\sigma,0}|+|\Sigma_{\sigma,0}|\tilde{Q})+2s_{k-1}^2 e^{2s_{k-1}}\|\Sigma_{\sigma,0}\|_\infty \mathbf{1}_{3\times 3}\\ &\leq \kappa_{u,k-1}(\tilde{Q}|\Sigma_{\sigma,0}|+|\Sigma_{\sigma,0}|\tilde{Q})+8\kappa_{u,k-1}^2 e^{4\kappa_{u,k-1}}\|\Sigma_{\sigma,0}\|_\infty \mathbf{1}_{3\times 3}\ .\end{aligned}$$

The claim then follows for $k \geq 1$ using $\kappa_{u,k-1} \leq \kappa_{u,k}$. For $k = 0$, the claim is trivial. □

Using the properties of $E_{n,N,\varepsilon,\gamma}$, we obtain the following result:

Corollary 5.32. *Let $\varepsilon, \gamma > 0$ with $\varepsilon \leq 1$. For $\omega \in E_{n,N,\varepsilon,\gamma}$, we have*

(a) $$\sup_i |\theta_{i,k}-\theta_{i,0}| \leq \kappa_{u,k}\begin{pmatrix} n^{\varepsilon-1/2}\\ n^{\varepsilon-1/2}\\ n^{\varepsilon}\end{pmatrix} + 2\kappa_{u,k}^2 e^{2\kappa_{u,k}} n^{\varepsilon}\mathbf{1}_3,$$

(b) $$|\Sigma_{\sigma,k}-\Sigma_{\sigma,0}| = \kappa_{u,k}\begin{pmatrix} O(n^\varepsilon) & O(n^\varepsilon) & O(n)\\ O(n^\varepsilon) & 0 & O(1)\\ O(n) & O(1) & O(n^\varepsilon)\end{pmatrix} + 8\kappa_{u,k}^2 e^{4\kappa_{u,k}} O(n)\mathbf{1}_{3\times 3}.$$

Proof.

(a) By properties (W1), (W2) and (W3) in Definition 5.23, we have

$$|\theta_{i,0}| = \begin{pmatrix}|a_{i,0}|\\ |b_{i,0}|\\ |w_{i,0}|\end{pmatrix} \leq \begin{pmatrix} n^\varepsilon\\ 0\\ n^{\varepsilon-1/2}\end{pmatrix}\ .$$

We can now apply Proposition 5.31 to obtain

$$|\theta_{i,k}-\theta_{i,0}| \leq \kappa_{u,k}\tilde{Q}|\theta_{i,0}| + 2\kappa_{u,k}^2 e^{2\kappa_{u,k}}\|\theta_{i,0}\|_\infty \mathbf{1}_3 \leq \kappa_{u,k}\begin{pmatrix} n^{\varepsilon-1/2}\\ n^{\varepsilon-1/2}\\ n^{\varepsilon}\end{pmatrix} + 2\kappa_{u,k}^2 e^{2\kappa_{u,k}} n^\varepsilon \mathbf{1}_3\ .$$

(b) By properties (W1), (W5), (W6) and (W7) in Definition 5.23, we have

$$|\Sigma_{\sigma,0}| = \begin{pmatrix}|\Sigma_{\sigma,a^2,0}| & |\Sigma_{\sigma,ab,0}| & |\Sigma_{\sigma,wa,0}|\\ |\Sigma_{\sigma,ab,0}| & |\Sigma_{\sigma,b^2,0}| & |\Sigma_{\sigma,wb,0}|\\ |\Sigma_{\sigma,wa,0}| & |\Sigma_{\sigma,wb,0}| & |\Sigma_{\sigma,w^2,0}|\end{pmatrix} = \begin{pmatrix} O(n) & 0 & O(n^\varepsilon)\\ 0 & 0 & 0\\ O(n^\varepsilon) & 0 & O(1)\end{pmatrix}\ .$$

We can now apply Proposition 5.31 to obtain

$$|\Sigma_{\sigma,k}-\Sigma_{\sigma,0}| \leq \kappa_{u,k}(\tilde{Q}|\Sigma_{\sigma,0}|+|\Sigma_{\sigma,0}|\tilde{Q}) + 8\kappa_{u,k}^2 e^{4\kappa_{u,k}}\|\Sigma_{\sigma,0}\|_\infty \mathbf{1}_{3\times 3}\ .$$

Since $0 < \varepsilon \leq 1$, we can conclude $\|\Sigma_{\sigma,0}\|_\infty = O(n)$ and

$$\begin{aligned}\tilde{Q}|\Sigma_{\sigma,0}|+|\Sigma_{\sigma,0}|\tilde{Q} &= \begin{pmatrix} O(n^\varepsilon) & 0 & O(1)\\ O(n^\varepsilon) & 0 & O(1)\\ O(n) & 0 & O(n^\varepsilon)\end{pmatrix} + \begin{pmatrix} O(n^\varepsilon) & O(n^\varepsilon) & O(n)\\ 0 & 0 & 0\\ O(1) & O(1) & O(n^\varepsilon)\end{pmatrix}\\ &= \begin{pmatrix} O(n^\varepsilon) & O(n^\varepsilon) & O(n)\\ O(n^\varepsilon) & 0 & O(1)\\ O(n) & O(1) & O(n^\varepsilon)\end{pmatrix}\ .\end{aligned}$$

□

The following result uses an induction argument to show that for large enough n, Σ_σ stays almost constant and $\kappa_{u,k}$ is small. It is one of the main steps in this section.

Proposition 5.33. *Let $|\alpha| \neq 1$, $\varepsilon \in (0, 1/2)$, $\gamma > 0$ and $\varrho > 0$. Then, for $N \geq \varrho n^2$, $\omega \in E_{n,N,\varepsilon,\gamma}$ and*

$$h \leq \frac{1}{\lambda_{\max}(A^{\mathrm{ref}} M_D)} ,$$

with $\lambda_{\max}(A^{\mathrm{ref}} M_D)$ as in Lemma 5.28, we have

$$\kappa_{u,k} = O(n^{\varepsilon-1}) ,$$

where $\kappa_{u,k}$ was defined in Definition 5.30 and the bound $O(n^{\varepsilon-1})$ is independent of $k \in \mathbb{N}_0$.

Proof. By Proposition 5.9, we know that $\overline{v}_{k+1} = (I - hA_k M_D)\overline{v}_k$. We want to bound $\kappa_{u,k} = h\sum_{l=0}^{k} \|u_l\|_\infty = h\sum_{l=0}^{k} \|BM_D\overline{v}_l\|_\infty$ by comparing it to the reference system $\delta\overline{v} = -hA^{\mathrm{ref}} M_D \overline{v}$ using Lemma B.3. Hence, we define

$$\begin{aligned} \delta_k &\coloneqq \sum_{l=0}^{k} \|(I - hA^{\mathrm{ref}} M_D)^l\|_\infty \cdot \sup_{0\leq l\leq k} \|(I - hA^{\mathrm{ref}} M_D) - (I - hA_l M_D)\|_\infty \\ &= h\sum_{l=0}^{k} \|(I - hA^{\mathrm{ref}} M_D)^l\|_\infty \cdot \sup_{0\leq l\leq k} \|(A_l - A^{\mathrm{ref}})M_D\|_\infty . \end{aligned} \tag{5.6}$$

For n large enough, we want to prove by induction that $\delta_k \leq 1/2$ for all $k \in \mathbb{N}_0$. Trivially, $\delta_{-1} = 0 \leq 1/2$. Now let $k \in \mathbb{N}_0$ with $\delta_{k-1} \leq 1/2$.

(1) By Lemma 5.8, we have $\tilde{u}_k = -\tilde{B}\tilde{M}_D\tilde{\overline{v}}_k$ and hence $u_k = -BM_D\overline{v}_k$. Thus,

$$\kappa_{u,k} = h\sum_{l=0}^{k} \|u_l\|_\infty \leq \|B\|_\infty \|M_D\|_\infty \cdot h\sum_{l=0}^{k} \|\overline{v}_l\|_\infty .$$

Because $\delta_{k-1} \leq 1/2$, we can apply Lemma B.3 and obtain

$$\begin{aligned} h\sum_{l=0}^{k} \|\overline{v}_l\|_\infty &\leq \frac{1}{1-\delta_{k-1}} h\sum_{l=0}^{k} \|(I - hA^{\mathrm{ref}} M_D)^l \overline{v}_0\|_\infty \\ &\leq 2h\sum_{l=0}^{\infty} \|(I - hA^{\mathrm{ref}} M_D)^l \overline{v}_0\|_\infty \\ &\overset{\text{Proposition 5.29}}{=} O(n^{\varepsilon-1}) . \end{aligned}$$

Moreover, as mentioned in Definition 4.1, we know that $\|M_D\|_\infty \leq \sqrt{4}\|M_D\|_2$ since $M_D \in \mathbb{R}^{4\times 4}$ and thus

$$\|M_D\|_\infty \leq 2\|M_D\|_2 = 2\lambda_{\max}(M_D) \overset{\text{(D2)}}{\leq} 4\lambda_{\max}(M_P) = O(1) . \tag{5.7}$$

Hence, we can write

$$\kappa_{u,k} = O(n^{\varepsilon-1}) , \tag{5.8}$$

where, in accordance with Definition 5.24, the constant in $O(n^{\varepsilon-1})$ does not depend on the induction step k.

(2) Let us investigate the components of Eq. (5.6):

$$\begin{aligned} h\sum_{l=0}^{k}\|(I-hA^{\mathrm{ref}}M_D)^l\|_\infty &\le h\sum_{l=0}^{\infty}\|(I-hA^{\mathrm{ref}}M_D)^l\|_\infty \overset{\text{Proposition 5.29}}{=} O(1)\\ (A_l-A^{\mathrm{ref}})M_D &= B\left((G_l^{\mathrm{w}}-G_0^{\mathrm{w}})+(G_l^{\mathrm{ab}}-G_0^{\mathrm{ab}})+hG_l^{\mathrm{wab}}\right)BM_D\\ \Rightarrow \|(A_l-A^{\mathrm{ref}})M_D\|_\infty &\overset{(5.7)}{=} O(1)\cdot(\|G_l^{\mathrm{w}}-G_0^{\mathrm{w}}\|_\infty+\|G_l^{\mathrm{ab}}-G_0^{\mathrm{ab}}\|_\infty+h\|G_l^{\mathrm{wab}}\|_\infty). \end{aligned}$$

First of all, for $0\le l\le k$,

$$\begin{aligned} \|G_l^{\mathrm{w}}-G_0^{\mathrm{w}}\|_\infty &= \max_{\sigma\in\{-1,1\}}|\Sigma_{\sigma,w^2,l}-\Sigma_{\sigma,w^2,0}|\\ &\overset{\text{Corollary 5.32}}{=} \kappa_{u,l}O(n^\varepsilon)+8\kappa_{u,l}^2e^{4\kappa_{u,l}}O(n)\\ &\overset{(5.8),\ \kappa_{u,l}\le\kappa_{u,k}}{=} O(n^{\varepsilon-1})O(n^\varepsilon)+8O(n^{2\varepsilon-2})e^{O(n^{\varepsilon-1})}O(n)\overset{\varepsilon\le1}{=}O(n^{2\varepsilon-1})\ . \end{aligned}$$

Similarly, for $0\le l\le k$,

$$\begin{aligned} \|G_l^{\mathrm{ab}}-G_0^{\mathrm{ab}}\|_\infty &\le \max_{\sigma\in\{-1,1\}}|\Sigma_{\sigma,a^2,l}-\Sigma_{\sigma,a^2,0}|+|\Sigma_{\sigma,ab,l}-\Sigma_{\sigma,ab,0}|+|\Sigma_{\sigma,b^2,l}-\Sigma_{\sigma,b^2,0}|\\ &= \kappa_{u,l}O(n^\varepsilon)+8\kappa_{u,l}^2e^{4\kappa_{u,l}}O(n)\\ &= \ldots = O(n^{2\varepsilon-1})\ . \end{aligned}$$

Observe that

$$h|r_{\sigma,l}|\le h\|u_l\|_\infty\le h\sum_{l'=0}^{k}\|u_{l'}\|_\infty=\kappa_{u,k}\overset{(5.8)}{=}O(n^{\varepsilon-1})$$

and similarly $h|s_{\sigma,l}|=O(n^{\varepsilon-1})$. Thus, we find

$$\begin{aligned} h\|G_l^{\mathrm{wab}}\|_\infty &= \max_{\sigma\in\{-1,1\}}h|r_{\sigma,l}\Sigma_{\sigma,wa,l}+s_{\sigma,l}\Sigma_{\sigma,wb,l}|\\ &= O(n^{\varepsilon-1})\cdot\left(\max_{\sigma\in\{-1,1\}}|\Sigma_{\sigma,wa,l}|+|\Sigma_{\sigma,wb,l}|\right)\ . \end{aligned}$$

Similar to the other calculations, we can compute for $0\le l\le k$

$$\begin{aligned} |\Sigma_{\sigma,wa,l}| &\le |\Sigma_{\sigma,wa,0}|+|\Sigma_{\sigma,wa,l}-\Sigma_{\sigma,wa,0}|\\ &\overset{(\mathrm{W7})}{=} O(n^\varepsilon)+O(n^{\varepsilon-1})O(n)+O(n^{2\varepsilon-2})O(n)\overset{\varepsilon\in(0,1)}{=}O(n^\varepsilon)\\ |\Sigma_{\sigma,wb,l}| &\le |\Sigma_{\sigma,wb,0}|+|\Sigma_{\sigma,wb,l}-\Sigma_{\sigma,wb,0}|\\ &= 0+O(n^{\varepsilon-1})O(1)+O(n^{2\varepsilon-2})O(n)=O(n^{2\varepsilon-1})\overset{\varepsilon\le1}{=}O(n^\varepsilon)\ , \end{aligned}$$

which yields $h\|G_l^{\mathrm{wab}}\|_\infty=O(n^{2\varepsilon-1})$.

We can now revisit the beginning of step (2) and obtain $\|(A_l-A^{\mathrm{ref}})M_D\|_\infty=O(n^{2\varepsilon-1})$ and

$$\begin{aligned} \delta_k &\overset{(5.6)}{=} h\sum_{l=0}^{k}\|(I-hA^{\mathrm{ref}}M_D)^l\|_\infty\cdot\sup_{0\le l\le k}\|(A_l-A^{\mathrm{ref}})M_D\|_\infty\\ &= O(1)\cdot O(n^{2\varepsilon-1})=O(n^{2\varepsilon-1})\ . \end{aligned}$$

We have shown that $\delta_{k-1} \leq 1/2$ implies $\delta_k = O(n^{2\varepsilon-1})$, where the constant in $O(n^{2\varepsilon-1})$ does not depend on k. Since $\varepsilon < 1/2$, $n^{2\varepsilon-1} \to 0$ as $n \to \infty$ and there exists $n_0 \in \mathbb{N}_0$ such that for all $n \geq n_0$ and $k \in \mathbb{N}_0$, $\delta_{k-1} \leq 1/2$ implies $\delta_k \leq 1/2$ and the induction works. Thus, for all $n \geq n_0$ and $k \in \mathbb{N}_0$, we know that $\delta_{k-1} \leq 1/2$ and we can apply step (1) to obtain

$$\kappa_{u,k} = O(n^{\varepsilon-1}) \ . \qquad \square$$

Remark 5.34. Proposition 5.33 can be formulated and proved in a non-asymptotic manner that does not use condition (P4) from Assumption 5.16 (which is used in Proposition 5.29). In this non-asymptotic formulation, one would show that if $h\sum_{k=0}^{\infty} \|(I - hA^{\mathrm{ref}}M_D)^k \overline{v}_0\|_\infty$, $h\sum_{k=0}^{\infty} \|(I - hA^{\mathrm{ref}}M_D)^k\|_\infty$, $|\Sigma_{\sigma,0}|$ and h are small enough in some sense, then

$$\sup_{k\in\mathbb{N}_0} \kappa_{u,k} \leq 2\|B\|_\infty \|M_D\|_\infty \cdot h \sum_{k=0}^{\infty} \|(I - hA^{\mathrm{ref}}M_D)^k\|_\infty \ . \qquad \blacktriangleleft$$

Corollary 5.35. *Let $|\alpha| \neq 1$, $\varepsilon \in (0, 1/2), \gamma > 0$ and $\varrho > 0$. Then, for $N \geq \varrho n^2$, $\omega \in E_{n,N,\varepsilon,\gamma}$ and*

$$h \leq \frac{1}{\lambda_{\max}(A^{\mathrm{ref}}M_D)} \ ,$$

with $\lambda_{\max}(A^{\mathrm{ref}}M_D)$ as in Lemma 5.28, we have

$$\sup_i |\theta_{i,k} - \theta_{i,0}| = \begin{pmatrix} O(n^{2\varepsilon-3/2}) \\ O(n^{2\varepsilon-3/2}) \\ O(n^{2\varepsilon-1}) \end{pmatrix}, \quad |\Sigma_{\sigma,k} - \Sigma_{\sigma,0}| = \begin{pmatrix} O(n^{2\varepsilon-1}) & O(n^{2\varepsilon-1}) & O(n^{\varepsilon}) \\ O(n^{2\varepsilon-1}) & O(n^{2\varepsilon-1}) & O(n^{2\varepsilon-1}) \\ O(n^{\varepsilon}) & O(n^{2\varepsilon-1}) & O(n^{2\varepsilon-1}) \end{pmatrix} .$$

Proof. This follows directly from inserting $\kappa_{u,k} = O(n^{\varepsilon-1})$ (Proposition 5.33) into Corollary 5.32. $\square$

Remark 5.36. Corollary 5.35 shows that $\Sigma_{\sigma,wa}$ changes by at most $O(n^{\varepsilon})$, while other quantities like Σ_{σ,a^2} change by at most $O(n^{2\varepsilon-1})$ (which is much less since $\varepsilon \in (0, 1/2)$). Since $\Sigma_{\sigma,wa} = \hat{p}_\sigma$ is related to the slope of $f_{W,\tau,\sigma}$, it must change by a nontrivial amount since the slope should converge to the optimal slope. Our proof of Proposition 5.33 works since $\Sigma_{\sigma,wa}$ does not occur in G^{w} and G^{ab}, hence G^{w} and G^{ab} only change by $O(n^{2\varepsilon-1})$. While $\Sigma_{\sigma,wa}$ does occur in hG^{wab}, choosing a small h mitigates the "large" change in $\Sigma_{\sigma,wa}$. $\blacktriangleleft$

5.5 Inconsistency Results

In the following, we derive inconsistency results in several formulations. We start with a general formulation and then proceed to introduce versions that are more directly related to universal consistency as defined in Definition 2.2.

Theorem 5.37 (General inconsistency result)**.** *Let $|\alpha| \neq 1$, $\gamma \in (0, 1/2)$, $\varepsilon \in (0, 1/4 - \gamma/2)$ and $\varrho > 0$. Let P be a probability distribution that satisfies conditions (P1) – (P4) in Assumption 5.16 with a constant $m_P > 0$. With this distribution P, define*

the random variables W_0 and D as in Definition 5.21 and the event $E_{n,N,\varepsilon,\gamma}$ as in Definition 5.23. Moreover, define $(\tilde{W}_k)_{k\in\mathbb{N}_0}$ for $k \geq 0$ by $\tilde{W}_0 \coloneqq W_0$ and the gradient descent iteration

$$\tilde{W}_{k+1} = \tilde{W}_k - h\nabla L_D(\tilde{W}_k)$$

for a step width

$$0 < h \leq \frac{1}{\lambda_{\max}(A^{\mathrm{ref}} M_D)}$$

with $\lambda_{\max}(A^{\mathrm{ref}} M_D)$ as in Lemma 5.28.
Then there exists an $n_0 \in \mathbb{N}_0$ such that for all $n \geq n_0$, $N \geq \varrho n^2$, $k \in \mathbb{N}_0$ and $\omega \in E_{n,N,\varepsilon,\gamma}$ (cf. Definition 5.23),

- $f_{\tilde{W}_k(\omega)}(x) = f_{\tilde{W}_k(\omega),\tau(W_0(\omega)),\mathrm{sgn}(x)}(x)$ *for $|x| \geq m_P$,*
- $f_{\tilde{W}_k(\omega)}|_{(-\infty,m_P]}$ *and* $f_{\tilde{W}_k(\omega)}|_{[m_P,\infty)}$ *are affine functions.*

Proof.

(1) Let W_0, D be random variables as in Definition 5.21 and $\tau \coloneqq \tau(W_0)$, cf. Definition 5.1. Define the random variables $W_k, k \in \mathbb{N}_0$, recursively by

$$W_{k+1} \coloneqq W_k - h\nabla L_{D,\tau}(W_k) \ .$$

We want to show that $\nabla L_{D,\tau}(W_k) = \nabla L_D(W_k)$ for all $k \in \mathbb{N}_0, n \geq n_0, N \geq \varrho n^2, \omega \in E_{n,N,\varepsilon,\gamma}$ for some $n_0 \in \mathbb{N}_0$ that has yet to be chosen.

(2) Recall that due to the special structure of φ,

$$\begin{aligned}
L_D(W) &= \frac{1}{2N}\sum_{j\in J}(y_j - f_W(x_j))^2 \\
L_{D,\tau}(W) &= \frac{1}{2N}\sum_{j\in J}(y_j - f_{W,\tau,\mathrm{sgn}(x_j)}(x_j))^2 \\
f_W(x_j) &= c + \sum_{i\in I}\varphi'(\mathrm{sgn}(a_i x_j + b_i))w_i(a_i x_j + b_i) \\
f_{W,\tau,\mathrm{sgn}(x_j)}(x_j) &= c + \sum_{i\in I}\varphi'(\tau_i \cdot \mathrm{sgn}(x_j))w_i(a_i x_j + b_i) \ .
\end{aligned}$$

Define the set

$$\mathcal{S}_{W_0} \coloneqq \{(a,b,c,w) \in \mathbb{R}^{3n+1} \mid \forall i \in I, |x| \geq m_P : \mathrm{sgn}(a_i x + b_i) = \mathrm{sgn}(a_{i,0} x + b_{i,0})\} \ ,$$

which is open since (W4) and (W1) yield $a_{i,0} \neq 0, b_{i,0} = 0$ and hence $a_{i,0}x + b_{i,0} \neq 0$ for $|x| \geq m_P > 0$. Since $\tau_i \cdot \mathrm{sgn}(x) = \mathrm{sgn}(a_{i,0}) \cdot \mathrm{sgn}(x) = \mathrm{sgn}(a_{i,0}x + b_{i,0})$, we conclude

$$f_W(x) = f_{W,\tau,\mathrm{sgn}(x)}(x) \qquad \text{for all } W \in \mathcal{S}_{W_0}, |x| \geq m_P. \tag{5.9}$$

Since $|x_j| \geq m_P$ by (D3), $L_D(W) = L_{D,\tau}(W)$ for all $W \in \mathcal{S}_{W_0}$. Since $\mathcal{S}_{W_0}$ is open, this also means $\nabla L_D(W) = \nabla L_{D,\tau}(W)$ for all $W \in \mathcal{S}_{W_0}$. Hence, we need to show that $W_k \in \mathcal{S}_{W_0}$ for all $k \in \mathbb{N}_0, n \geq n_0, \omega \in E_{n,N,\varepsilon,\gamma}$ for a suitable $n_0 \in \mathbb{N}_0$.

(3) As usual, assume $\omega \in E_{n,N,\varepsilon,\gamma}$. Then, Corollary 5.35 yields

$$\sup_{i,k} |a_{i,k} - a_{i,0}| = O(n^{2\varepsilon-3/2})$$
$$\sup_{i,k} |b_{i,k} - b_{i,0}| = O(n^{2\varepsilon-3/2}) \ .$$

Moreover, properties (W1) and (W4) of Definition 5.23 yield

$$\min_i |a_{i,0}| \geq n^{-(1+\gamma)}$$
$$\max_i |b_{i,0}| = 0 \ .$$

For $x \in \mathbb{R}$, we thus have

$$m_x \coloneqq \min_i |a_{i,0}x + b_{i,0}| \geq n^{-(1+\gamma)}|x|$$
$$\delta_x \coloneqq \sup_{i,k} |(a_{i,0}x_j + b_{i,0}) - (a_{i,k}x_j + b_{i,k})| = O(n^{2\varepsilon-3/2})|x| + O(n^{2\varepsilon-3/2}) \ .$$

and for $|x| \geq m_P$, where m_P was defined in Assumption 5.16, we obtain

$$\frac{\delta_x}{m_x} = O\left(n^{2\varepsilon-3/2+1+\gamma}\right) \underbrace{\left(\frac{1}{|x|}+1\right)}_{\leq \frac{1}{m_P}+1} = O\left(n^{2(\varepsilon-(1/4-\gamma/2))}\right) = o(1)$$

since $\varepsilon \in (0, 1/4 - \gamma/2)$ by assumption. Hence, there exists an $n_0 \in \mathbb{N}_0$ such that $\delta_x/m_x < 1$ for $n \geq n_0$ and $N \geq \varrho n^2$. But $\delta_x/m_x < 1$ means $\delta_x < m_x$ and thus

$$\min_i |a_{i,0}x + b_{i,0}| > \sup_{i,k} |(a_{i,0}x + b_{i,0}) - (a_{i,k}x + b_{i,k})|$$

for $|x| \geq m_P$. This means $W_k \in \mathcal{S}_{W_0}$ for all $k \in \mathbb{N}_0$. Therefore, $\nabla L_{D,\tau}(W_k) = \nabla L_D(W_k)$ for all $k \in \mathbb{N}_0$ and $(W_k)_{k\in\mathbb{N}_0}$ satisfies the original gradient descent iteration:

$$W_{k+1} = W_k - h\nabla L_D(W_k) \ .$$

By induction, we conclude $W_k = \tilde{W}_k$ for all $\omega \in E_{n,N,\varepsilon,\gamma}$ and all $k \in \mathbb{N}_0$. Moreover, we have shown in Eq. (5.9) that f_{W_k} and $f_{W_k,\tau,\sigma}$ agree on $\sigma[m_P, \infty)$. Since $f_{W_k,\tau,\sigma}$ is affine, the claim follows. □

Remark 5.38. The right-hand side of the bound

$$h \leq \frac{1}{\lambda_{\max}(A^{\text{ref}} M_D)}$$

in Theorem 5.37 can be computed for any specific dataset D and initialization W_0, but it depends on D and W_0. Note that for $\omega \in E_{n,N,\varepsilon,\gamma}$, we have

$$\lambda_{\max}(A^{\text{ref}} M_D) \overset{\text{Lemma 5.28}}{\leq} \lambda_{\max}(A^{\text{ref}})\lambda_{\max}(M_D)$$
$$\overset{\text{Lemma 5.27, (D2)}}{\leq} (1+|\alpha|)^2(c_w + \max\{2, nc_a\}) \cdot 2\lambda_{\max}(M_P) = \Theta(n) \ .$$

In the proof of Proposition B.2, it is shown that indeed $\lambda_{\max}(A^{\mathrm{ref}} M_D) = \Theta(n)$, but knowing $\lambda_{\max}(A^{\mathrm{ref}} M_D) = O(n)$ is sufficient here. This means that in Theorem 5.37, it is sufficient for the step width h to satisfy

$$h \leq \frac{1}{(1+|\alpha|)^2(c_w + \max\{2, nc_a\}) \cdot 2\lambda_{\max}(M_P)} = \Theta(1/n) \ , \tag{5.10}$$

although this bound cannot be computed in general since P might be unknown. An alternative is to choose $h = o(n^{-1})$, which ensures that Eq. (5.10) is satisfied for sufficiently large n. ◀

For the next inconsistency result, recall from Definition 2.1 that

$$L_D(W) = \frac{1}{2N} \sum_{j=1}^{N} (y_j - f_W(x_j))^2$$
$$L_P(W) = \frac{1}{2} \mathbb{E}_{(x,y)\sim P} (y - f_W(x))^2 \ .$$

Corollary 5.39 (Explicit inconsistency result)**.** *Let $|\alpha| \neq 1$ and $\varrho > 0$. Let P be the (symmetric) uniform probability distribution on the finite set*

$$\{(-3,-1), (-2,2), (-1,-1), (1,1), (2,-2), (3,1)\} \subseteq \mathbb{R} \times \mathbb{R} \ .$$

With this distribution P, define the random variables W_0 and D on the probability spaces $(\Omega_{n,N}, \mathcal{F}_{n,N}, P_{n,N})$ as in Definition 5.21 and the events $E_{n,N,\varepsilon,\gamma} \subseteq \Omega_{n,N}$ as in Definition 5.23. Moreover, define $(\tilde{W}_k)_{k\in\mathbb{N}_0}$ by $\tilde{W}_0 \coloneqq W_0$ and the gradient descent iteration

$$\tilde{W}_{k+1} = \tilde{W}_k - h \nabla L_D(\tilde{W}_k)$$

for a step width

$$0 < h \leq \frac{1}{\lambda_{\max}(A^{\mathrm{ref}} M_D)}$$

with $\lambda_{\max}(A^{\mathrm{ref}} M_D)$ as in Lemma 5.28.
Then, for $N \geq \varrho n^2$,

$$P_{n,N}(\exists k \in \mathbb{N}_0 : L_P(\tilde{W}_k) < 1) = O(n^{-\gamma})$$

for all $\gamma \in (0, 1/2)$, even though for all $n \geq 5$, there exists a $W \in \mathbb{R}^{3n+1}$ such that $L_P(W) = 0$.

Proof. By Example 5.18, P satisfies conditions (P1) – (P4) of Assumption 5.16 with $m_P = 1$. Let $\gamma \in (0, 1/2)$. Choose $\varepsilon \in (0, 1/4 - \gamma/2)$ arbitrarily. By Theorem 5.37, there exists $n_0 \in \mathbb{N}_0$ such that for all $n \geq n_0, k \in \mathbb{N}, \omega \in E_{n,N,\varepsilon,\gamma}$ and $|x| \geq m_P = 1$, we have

$$f_{\tilde{W}_k(\omega),\tau,\mathrm{sgn}(x)}(x) = f_{\tilde{W}_k(\omega)}(x) \ .$$

For the dataset

$$\tilde{D} \coloneqq ((\tilde{x}_1, \tilde{y}_1), \dots, (\tilde{x}_6, \tilde{y}_6)) \coloneqq ((-3,-1),(-2,2),(-1,-1),(1,1),(2,-2),(3,1)) \ ,$$

we have $L_P = L_{\tilde{D}}$ and hence, since $|x_j| \geq m_P$ by (D3),

$$\begin{aligned} L_P(\tilde{W}_k(\omega)) = L_{\tilde{D}}(\tilde{W}_k(\omega)) &= \frac{1}{2 \cdot 6} \sum_{j=1}^{6} (\tilde{y}_j - f_{\tilde{W}_k(\omega)}(\tilde{x}_j))^2 \\ &= \frac{1}{2 \cdot 6} \sum_{j=1}^{6} (\tilde{y}_j - f_{\tilde{W}_k(\omega),\tau,\operatorname{sgn}(\tilde{x}_j)}(\tilde{x}_j))^2 = L_{\tilde{D},\tau}(\tilde{W}_k(\omega)) \ . \end{aligned}$$

Moreover, by Proposition 5.10, any weight vector $W \in \mathbb{R}^{3n+1}$ satisfies

$$L_{\tilde{D},\tau}(W) = \frac{1}{2} \overline{v}^\top M_{\tilde{D}} \overline{v} + \frac{1}{2N} \sum_{\sigma \in \{-1,1\}} Y_{\tilde{D},\sigma}^\top (I - X_{\tilde{D},\sigma} (X_{\tilde{D},\sigma}^\top X_{\tilde{D},\sigma})^{-1} X_{\tilde{D},\sigma}^\top) Y_{\tilde{D},\sigma} \ ,$$

where $\overline{v}^\top M_{\tilde{D}} \overline{v} \geq 0$ and

$$X_{\tilde{D},1}^\top Y_{\tilde{D},1} = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{pmatrix}^\top \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad X_{\tilde{D},-1}^\top Y_{\tilde{D},-1} = \begin{pmatrix} -3 & 1 \\ -2 & 1 \\ -1 & 1 \end{pmatrix}^\top \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \ .$$

Hence,

$$L_{\tilde{D},\tau}(W) \geq \frac{1}{2N} \sum_{\sigma \in \{-1,1\}} Y_{\tilde{D},\sigma}^\top Y_{\tilde{D},\sigma} = \frac{1}{2 \cdot 6}(6+6) = 1 \ .$$

This shows

$$L_P(\tilde{W}_k(\omega)) \geq 1$$

for all $n \geq n_0, k \in \mathbb{N}_0$ and $\omega \in E_{n,N,\varepsilon,\gamma}$. By Theorem 5.25 with $\gamma' \coloneqq \gamma > 0$ we have

$$\begin{aligned} P_{n,N}(\exists k \in \mathbb{N}_0 : L_P(\tilde{W}_k) < 1) &\overset{n \geq n_0}{\leq} P_{n,N}((E_{n,N,\varepsilon,\gamma})^c) = O(n^{-\gamma} + N^{-\gamma'}) \\ &= O(n^{-\gamma} + (\varrho n^2)^{-\gamma}) = O(n^{-\gamma}) \ . \end{aligned}$$

By Lemma C.1, for $n \geq 5$, there exists $W \in \mathbb{R}^{3n+1}$ such that f_W interpolates $\tilde{D}$ and hence $L_P(W) = 0$. □

Remark 5.40. If we consider the case $\varepsilon \to 0$, we essentially have

$$\begin{aligned} \min_i |a_{i,0} x_j + b_{i,0}| &= \Omega(n^{-1-\gamma}) \\ \sup_{i,k} |(a_{i,k} x_j + b_{i,0}) - (a_{i,0} x_j + b_{i,0})| &= O(n^{-3/2}) \ . \end{aligned}$$

If we choose $\gamma \in (0, 1/2)$ to be small, there is still almost $O(n^{-1/2})$ of room left that can be used to strengthen some part of Theorem 5.37 or Corollary 5.39. In Corollary 5.39, we chose to allow larger values of γ, which makes the probability of reaching the global optimum converge to zero almost like $O(n^{-1/2})$. We could also weaken condition (P3) in Assumption 5.16 instead, allowing for a small probability of generating data points near zero (i.e. $\min_j |x_j| = \Omega(n^{\varepsilon - 1/2})$). Another possibility would be to allow n to grow faster than $O(\sqrt{N})$, up to $O(N^{1-\varepsilon})$ for some $\varepsilon > 0$. ◀

Remark 5.41. In Corollary 5.39, it is possible to use the fixed dataset[4]

$$D = ((-3,-1),(-2,2),(-1,-1),(1,1),(2,-2),(3,1))$$

instead of sampling D from the corresponding probability distribution P: Since $M_D = M_P$ and $v_P^{\mathrm{opt}} = v_D^{\mathrm{opt}}$, the conditions (D1) – (D3) in Definition 5.23 are always satisfied, hence $D \in E_{N,\varepsilon}^{\mathrm{D}}$. This means that even if we fix D, we have

$$P_{n,N}(E_{n,N,\varepsilon,\gamma}) = 1 - O(n^{-\gamma}) \ .$$

Theorem 5.37 now shows that for $n \geq n_0$ and all $\omega \in E_{n,N,\varepsilon,\gamma}$, the kinks will not reach the data points. ◀

Finally, we want to formulate a result that relates more directly to (universal) consistency as defined in Definition 2.2. To this end, we formally introduce neural network estimators.

Definition 5.42. Let Q^{wa} satisfy Assumption 5.19 and let $n, N \in \mathbb{N}$, $\alpha \in \mathbb{R}$, $h > 0$, $c_a, c_w > 0$ and $D \in (\mathbb{R} \times \mathbb{R})^N$. A NN-algorithm is a random function

$$\mathcal{A}_{n,N,\alpha,h,Q^{\mathrm{wa}},c_a,c_w} : D \mapsto \mathcal{A}_{n,N,\alpha,h,Q^{\mathrm{wa}},c_a,c_w}(D) \coloneqq (f_{W_k})_{k\in\mathbb{N}_0} \ ,$$

where $W_{k+1} = W_k - h\nabla L_D(W_k)$ and the components of $W_0 = (a_{\cdot,0}, b_{\cdot,0}, c_0, w_{\cdot,0})$ are independent and distributed as

$$\begin{aligned} \sqrt{\frac{1}{c_a}} a_{i,0} &\sim Q^{\mathrm{wa}} \\ \sqrt{\frac{n}{c_w}} w_{i,0} &\sim Q^{\mathrm{wa}} \\ b_{i,0} &= 0 \\ c_0 &= 0 \ . \end{aligned}$$ ◀

Corollary 5.43 (High-level inconsistency result)**.** *Let P be a probability distribution on $\mathbb{R} \times \mathbb{R}$ satisfying conditions (P1) – (P5) from Assumption 5.16. Let Q^{wa} satisfy Assumption 5.19, let $\alpha \in \mathbb{R} \setminus \{-1, 1\}$, let $c_a, c_w > 0$ and let $(n_N)_{N\in\mathbb{N}}, (h_N)_{N\in\mathbb{N}}$ be sequences with*

- *$n_N \in \mathbb{N}$, $n_N = O(\sqrt{N})$,*
- *$h_N \in (0, \infty)$, $h_N = o(n_N^{-1})$.*

For $D \in (\mathbb{R} \times \mathbb{R})^N$, let $(f_{D,k})_{k\in\mathbb{N}_0} \coloneqq \mathcal{A}_{n_N,N,\alpha,h_N,Q^{\mathrm{wa}},c_a,c_w}(D)$. Then, there exist $C, \tilde{C} > 0$ independent of n_N, N, h_N such that

$$P_{n_N,N}\left(\inf_{k\in\mathbb{N}_0} R_P\left(f_{D,k}\right) \leq \inf_{f:\mathbb{R}\to\mathbb{R}} R_P(f) + C\right) \leq \tilde{C} n_N^{-\gamma}$$

for all $\gamma < 1/2$.

[4]Of course, we can then also drop the assumption $N \geq \varrho n^2$.

Proof. Let $\gamma \in (0, 1/2)$. Choose $\varepsilon \in (0, 1/4 - \gamma/2)$ arbitrarily. Since

$$\frac{1}{\lambda_{\max}(A^{\mathrm{ref}} M_D)} = \Omega(n_N^{-1})$$

for $\omega \in E_{n_N,N,\varepsilon,\gamma}$ by Remark 5.38 and since we assumed $h_N = o(n_N^{-1})$, it follows that

$$h_N \leq \frac{1}{\lambda_{\max}(A^{\mathrm{ref}} M_D)}$$

whenever n_N is sufficiently large. Since $n_N = O(\sqrt{N})$, there exists $\varrho > 0$ with $N \geq \varrho n_N^2$ for all $N \in \mathbb{N}$. By Theorem 5.37, there exists $n_0 \in \mathbb{N}_0$ such that whenever $n_N \geq n_0$, it holds true for all $k \in \mathbb{N}$, $\omega \in E_{n_N,N,\varepsilon,\gamma}$ and $|x| \geq m_P$ that

$$f_{\tilde{W}_k(\omega)}(x) = f_{\tilde{W}_k(\omega),\tau,\mathrm{sgn}(x)}(x) \ .$$

Hence, whenever $n_N \geq n_0$, $\omega \in E_{n_N,N,\varepsilon,\gamma}$, $k \in \mathbb{N}_0$, there exists a function $f \in \mathcal{F}_{\mathrm{aff}}$ (with $\mathcal{F}_{\mathrm{aff}}$ defined in Assumption 5.16) such that $f(x) = f_{\tilde{W}_k(\omega)}(x)$ for all $|x| \geq m_P$. By (P5), we have $C \coloneqq \inf_{f \in \mathcal{F}_{\mathrm{aff}}} R_P(f) - \inf_{f:\mathbb{R}\to\mathbb{R}} R_P(f) > 0$. By (P3), this implies

$$R_P(f_{\tilde{W}_k(\omega)}) \geq \inf_{f \in \mathcal{F}_{\mathrm{aff}}} R_P(f) = \inf_{f:\mathbb{R}\to\mathbb{R}} R_P(f) + C \ .$$

For sufficiently large n_N, we hence obtain with $\gamma' \coloneqq \gamma > 0$:

$$\begin{aligned} P_{n_N,N}\left(R_P(f_{\tilde{W}_k(\omega)}) \geq \inf_{f:\mathbb{R}\to\mathbb{R}} R_P(f) + C\right) &\leq P((E_{n_N,N,\varepsilon,\gamma})^c) \\ &\overset{\text{Theorem 5.25}}{=} O\left(n_N^{-\gamma} + N^{-\gamma'}\right) \\ &= O\left(n_N^{-\gamma} + (\varrho n_N^2)^{-\gamma}\right) \\ &= O\left(n_N^{-\gamma}\right) \ . \end{aligned}$$

□

Remark 5.44. The presented proofs could have been simplified by considering distributions P on $[m_P, \infty)$ instead. This would lead to $M_{-1} = 0$, allowing us to reduce the four-dimensional system $\delta\overline{v} = -hAM\overline{v}$ to a two-dimensional system and thus to simplify many formulas. Here, we chose to prove the results for distributions on $(-\infty, m_P] \cup [m_P, \infty)$ since this allows for distributions that are normalized (even symmetric) around zero, which arguably strengthens the results and also provides more insights about how the "signs" interact. ◀

Remark 5.45. For the negative gradient flow equation

$$\dot{W}(t) = -\nabla L_{D,\tau}(W(t)) \ ,$$

we can use similar arguments to show that a_i and b_i only change by $o(1/n)$ with high probability. These arguments are partially simpler because using $\dot{W}$ instead of δW yields slightly simpler formulas. However, the induction argument in Section 5.4 becomes slightly more complicated and, more importantly, bounding the error between $W(kh)$ and W_k (the discrete gradient descent version) yields a very bad bound on the maximum step size h for which the argument works. ◀

Remark 5.46. A crucial step in the inconsistency proof was to obtain the L^1 bound

$$h \sum_{k=0}^{\infty} \|(I - hA^{\text{ref}}M)^k \overline{v}_0\|_\infty = O(1/n)$$

in Proposition 5.29 using the eigenvector analysis in Proposition B.2. In the continuous case outlined in Remark 5.45, this corresponds to

$$\int_0^\infty \|e^{-A^{\text{ref}}Mt}\overline{v}(0)\|_\infty \,\mathrm{d}t = O(1/n) \ .$$

The trajectory $t \mapsto e^{-A^{\text{ref}}Mt}\overline{v}_0$ satisfies the differential equation $\dot{\overline{v}}(t) = -A^{\text{ref}}M\overline{v}(t)$ and serves as an approximation of the true trajectory $\overline{v}(t)$ with $\dot{\overline{v}}(t) = -A(t)M\overline{v}(t)$. A L^2 bound would have been much easier to obtain but is not useful for our proof:[5]

$$\begin{aligned}
\int_0^\infty (L_{D,\tau}(W(t)) - \inf_{t'} L_{D,\tau}(W(t'))) \,\mathrm{d}t &\overset{5.10}{=} \int_0^\infty \overline{v}(t)^\top M \overline{v}(t) \,\mathrm{d}t \\
&\approx \int_0^\infty \left(e^{-A^{\text{ref}}Mt}\overline{v}_0\right)^\top M (e^{-A^{\text{ref}}Mt}\overline{v}_0) \,\mathrm{d}t \\
&= \overline{v}_0^\top \left(\int_0^\infty \left(e^{-A^{\text{ref}}Mt}\right)^\top M e^{-A^{\text{ref}}Mt} \,\mathrm{d}t \right) \overline{v}_0 \ ,
\end{aligned}$$

where

$$\begin{aligned}
&\int_0^\infty \left(e^{-A^{\text{ref}}Mt}\right)^\top M e^{-A^{\text{ref}}Mt} \,\mathrm{d}t \\
&= \int_0^\infty \left(e^{-A^{\text{ref}}Mt}\right)^\top \left((A^{\text{ref}}M)^\top \cdot \frac{1}{2}(A^{\text{ref}})^{-1} + \frac{1}{2}(A^{\text{ref}})^{-1} \cdot (A^{\text{ref}}M) \right) e^{-A^{\text{ref}}Mt} \,\mathrm{d}t \\
&= \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t} \left[- \left(e^{-A^{\text{ref}}Mt}\right)^\top \left(\frac{1}{2}(A^{\text{ref}})^{-1} \right) e^{-A^{\text{ref}}Mt} \right] \mathrm{d}t \\
&= \left[- \left(e^{-A^{\text{ref}}Mt}\right)^\top \left(\frac{1}{2}(A^{\text{ref}})^{-1} \right) e^{-A^{\text{ref}}Mt} \right]_0^\infty \\
&= -0 + I \cdot \frac{1}{2}(A^{\text{ref}})^{-1} \cdot I \\
&= \frac{1}{2}(A^{\text{ref}})^{-1} \ .
\end{aligned}$$

The resulting term $\frac{1}{2}\overline{v}_0^\top (A^{\text{ref}})^{-1}\overline{v}_0^\top$ resembles the term $y^\top (H^\infty)^{-1} y$ that describes the convergence speed of the neural network found by Arora et al. [2]. ◀

[5] At the core of this calculation is the observation that $X = \frac{1}{2}(A^{\text{ref}})^{-1}$ is the solution to the Lyapunov equation $(-A^{\text{ref}}M)^\top X + X(-A^{\text{ref}}M) + M = 0$. Solutions of Lyapunov equations can be used to compute such integrals (cf. e.g. Theorem 18 in [31]).

6 Empirical Results

In this section, we empirically estimate the probability that no kink crosses a data point when the datasets are sampled from the distribution P specified in Example 5.18. Our implementation is provided at `https://github.com/dholzmueller/nn_inconsistency`. We use $N = n^2$, which does not render the computations infeasible because this particular distribution P only samples from six distinct points. Training on a dataset sampled from P can be represented by weighting these six points appropriately in the training objective. Specifically, we estimate

$$P_{n,N}(\exists i, k : |b_{i,k}/a_{i,k}| \geq 1)$$

for $\alpha = 0$, $N = n^2$ and $n \in \{16, 32, \dots, 2048\}$ with 10^4 Monte Carlo samples for each n. We consider two different optimizers:

- Gradient Descent (GD) as used previously, and
- Stochastic Gradient Descent (SGD) with batch size 16 on batches that are sampled independently from the training data.

We use two ways of choosing the step size:

- $h = \frac{1}{\lambda_{\max}(A^{\text{ref}} M_D)}$, which we find to be approximately $0.4 \cdot n^{-1}$ for our choice of P.
- $h = 0.01 \cdot n^{-1}$ as a smaller choice for SGD.

We stop the training process whenever a kink crosses 1 or -1. In order to also stop the runs where no kink crosses a data point, we consider two stopping criteria:

- In some settings, we use an early stopping criterion. Every 1000 epochs (GD) or 1000 minibatches (SGD), the criterion checks the validation loss on an independently drawn but fixed validation set. If the validation loss does not decrease by more than 10^{-8} within the last ten of these checks, the criterion stops the training process. This is equivalent to using early stopping in Keras [6] with `patience` $= 10$ and `min_delta` $= 10^{-8}$.
- We employ the theory developed in this thesis to obtain a sufficient criterion for stopping. This means that if the criterion is satisfied, the theory guarantees (for GD) that the kinks will never reach 1 or -1. It works as follows:
  - Treat the current weight vector W_k as if it was the initialization W_0.
  - Compute the matrices A^{ref} and M.
  - Compute upper bounds on

    $$h \sum_{k=0}^{\infty} \|(I - hM^{1/2} A^{\text{ref}} M^{1/2})^k\|_\infty \text{ and } h \sum_{k=0}^{\infty} \|(I - hM^{1/2} A^{\text{ref}} M^{1/2})^k M^{1/2} \overline{v}_0\|_\infty$$

    based on an orthogonal diagonalization of $M^{1/2} A^{\text{ref}} M^{1/2}$.

– As in the proof of Proposition 5.33, assume $\delta_{k-1} \leq 1/2$ and use the previous step and Proposition 5.31 to obtain bounds on $\kappa_{u,k}$ and on the differences $A_k - A^{\text{ref}}$. Then, obtain a new bound for δ_k. If this bound, which is independent of k, is at most $1/2$, then the induction argument works and we can employ Proposition 5.31 to find bounds on $|\theta_{i,k} - \theta_{i,0}|$. If these bounds imply that no kink will reach a data point, stop the training process.

For $n \in \{1, 2, 4, 8\}$, we observed some randomly sampled (W_0, D) pairs where all a_i had the same sign or the condition $M_D \succ 0$ was violated. In these cases, the sufficient criterion will never be satisfied since this yields $\lambda_{\min}(A^{\text{ref}} M_D) = 0$, which implies $h \sum_{k=0}^{\infty} \|(I - hA^{\text{ref}} M_D)^k\|_\infty = \infty$. For $n \geq 16$, this problem did not occur in our experiments.

Finally, we also consider shifted versions P_Δ of the distribution P. We specify with Δ how much this distribution is shifted upwards, i.e. if $(x, y) \sim P$, then $(x, y + \Delta) \sim P_\Delta$. Note that P_Δ satisfies condition (P4) of Assumption 5.16 only if $\Delta = 0$. Since Corollary 5.39 predicts $P_{n,N}(\exists i, k : |b_{i,k}/a_{i,k}| \geq 1) = O(n^{-\gamma})$ for $\Delta = 0$ and all $\gamma < 1/2$, the function $n \mapsto 2n^{-1/2}$ is plotted as a comparison. (The factor 2 is used for visual reasons.) Our plots provide evidence that for small $|\Delta|$ and practical hidden layer sizes, the probability that the kinks reach 1 or -1 is still rather low.

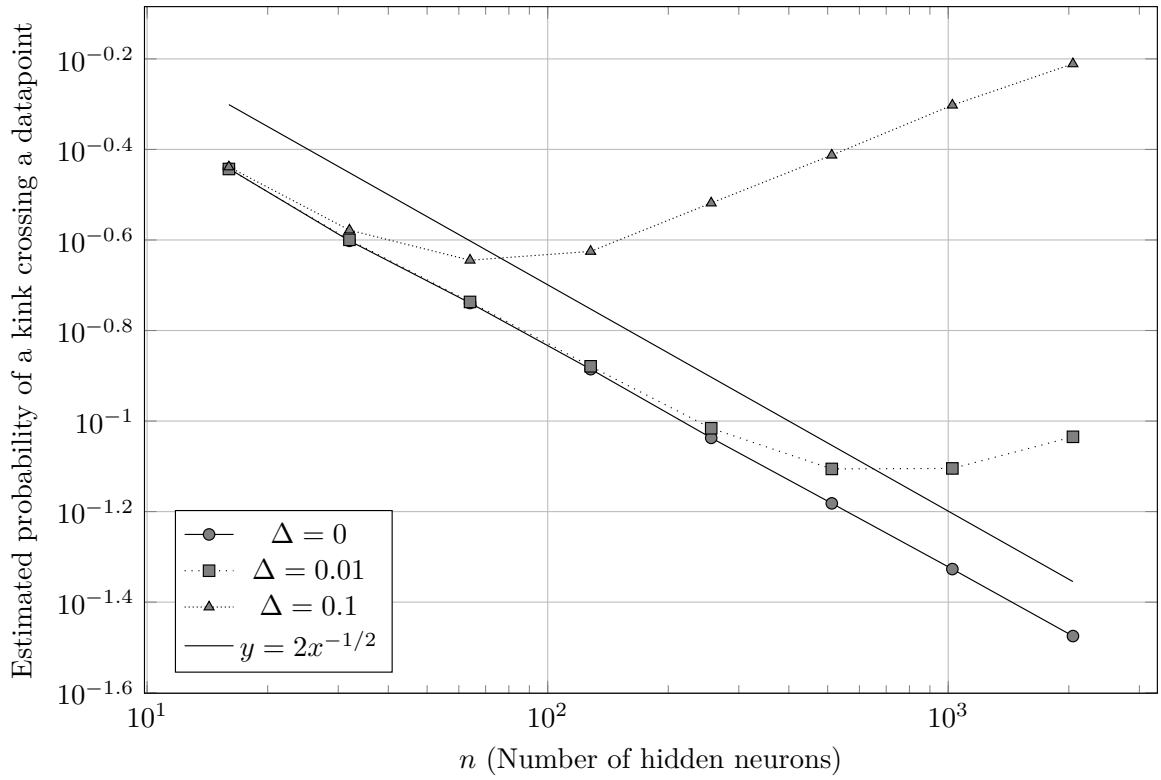


Figure 5: Monte Carlo estimates for the probability that a kink crosses $x = 1$ or $x = -1$ when using gradient descent, $h = \frac{1}{\lambda_{\max}(A^{\text{ref}} M_D)}$ and only the sufficient stopping criterion. Data is sampled from P_Δ for different values of Δ.

Figure 5 shows the probability estimates for gradient descent with the larger step size.

For $\Delta = 0$, the probabilities behave similarly to $n^{-1/2}$ while for $\Delta \neq 0$, they start to deviate from this behavior once n is large enough. This is due to the fact that

$$v_{P_\Delta}^{\text{opt}} = (0, 0, \Delta, \Delta)^\top \ .$$

Inserting this $v_{P_\Delta}^{\text{opt}}$ into the proof of Proposition 5.29, one obtains (in the limit $\varepsilon \to 0$)

$$h \sum_{k=0}^{\infty} \|(I - hA^{\text{ref}} M_D)^k \overline{v}_0\|_\infty = O(n^{-1} + |\Delta|) \ .$$

Once the term $|\Delta|$ dominates the term n^{-1}, the probability of a kink crossing 1 or -1 gets significantly larger than the corresponding probability for $\Delta = 0$. It can also be seen in Figure 5 that the curve for $\Delta = 0.01$ has its minimum at an about ten times larger value of n than the curve for $\Delta = 0.1$.

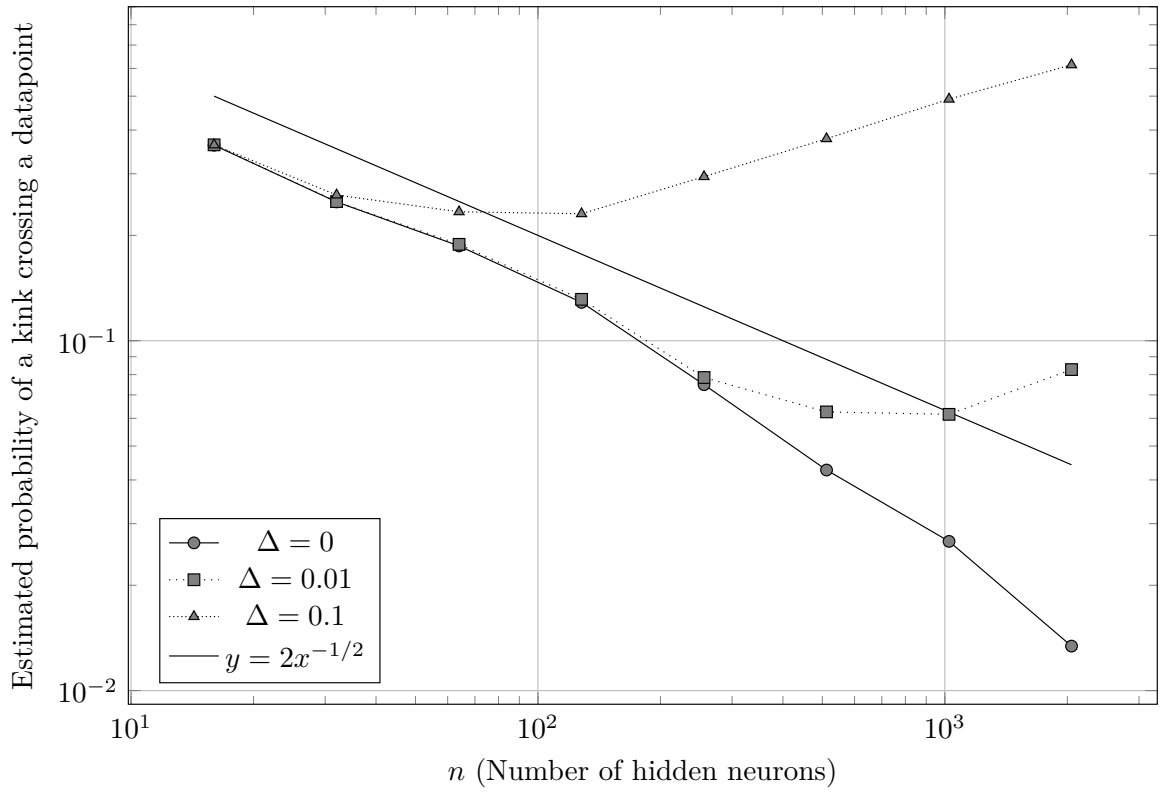


Figure 6: Monte Carlo estimates for the probability that a kink crosses $x = 1$ or $x = -1$ when using gradient descent, $h = \frac{1}{\lambda_{\max}(A^{\text{ref}} M_D)}$ and both early and sufficient stopping criteria. Training and validation data is sampled from P_Δ for different values of Δ.

Figure 6 shows that when additionally using the early stopping criterion, the probability estimates get slightly lower since some configurations get stopped even though the kinks would cross 1 or -1 later. When using SGD with early stopping as in Figure 7, we observe that the probabilities are generally higher than for gradient descent but show similar asymptotic behavior. The reason that the probabilities are higher might be that the step size $h = \frac{1}{\lambda_{\max}(A^{\text{ref}} M_D)}$ is too large for stochastic gradient descent to mimic gradient descent. If we reduce the step size to $h = 0.01 \cdot n^{-1}$, we observe significantly

reduced probabilities as shown in Figure 8. Another reason for the lower probabilities in Figure 8 may be that while we reduce the step size, we do not increase the number of batches between each early stopping check. Hence, the early stopping criterion might stop the training even earlier in terms of optimization progress.

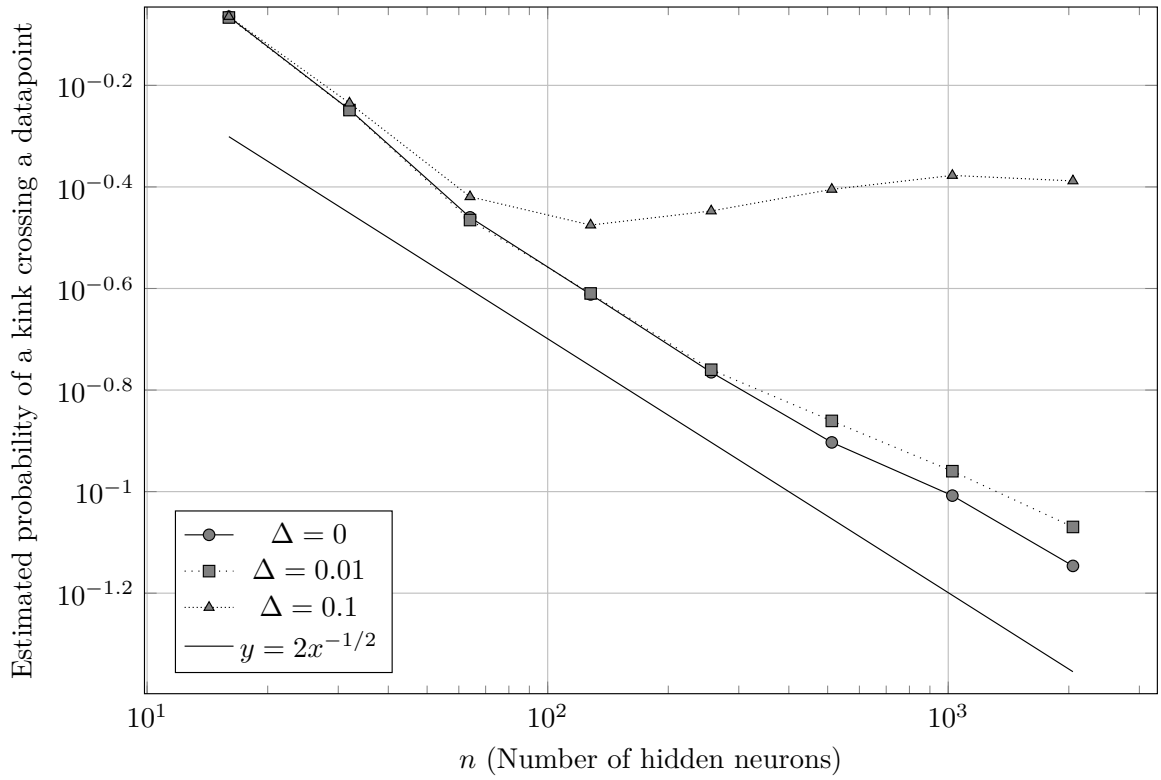


Figure 7: Monte Carlo estimates for the probability that a kink crosses $x = 1$ or $x = -1$ when using stochastic gradient descent, $h = \frac{1}{\lambda_{\max}(A^{\text{ref}} M_D)}$ and only the early stopping criterion. Training and validation data is sampled from P_Δ for different values of Δ.

Let us now roughly estimate what would happen if we chose the asymptotic variances of the initialization differently. Let us assume that

$$\operatorname{Var}(a_i) = \Theta(n^{-p}), \qquad \operatorname{Var}(w_i) = \Theta(n^{-q})$$

with $p, q \leq 1$. Let us also assume that we work with the dataset D corresponding to the probability distribution P in Example 5.18 so that $v_D^{\text{opt}} = 0$ and we do not have to care about sampling. In the following, we ignore factors like n^ε and the fact that A, w and Σ_σ might change significantly during gradient descent. We also require $\omega \in E_{n,N,\varepsilon,\gamma}$ as usual. For the case $p < q$, we have $\Sigma_{\sigma,w^2} = \Theta(n^{1-q}) = o(n^{1-p}) = o(\Sigma_{\sigma,a^2})$ and can therefore argue similar to Proposition 5.29: We apply Proposition B.2 (with $f \simeq g :\Leftrightarrow f = \Theta(g)$) to obtain

$$\begin{aligned} &h \sum_{k=0}^{\infty} \|(I - hA^{\text{ref}} M)^k \overline{v}_0\|_2 \\ &\leq 2\sqrt{\operatorname{cond}(M)} \left(\left(\frac{1}{\lambda} + \frac{2\sqrt{m_1}\sqrt{\operatorname{cond}(M)}\beta}{\lambda_{\min}(A^{\text{ref}} M)} \right) \|v_1\|_2 + \frac{1}{\lambda_{\min}(A^{\text{ref}} M)} \|v_2\|_2 \right) \end{aligned}$$

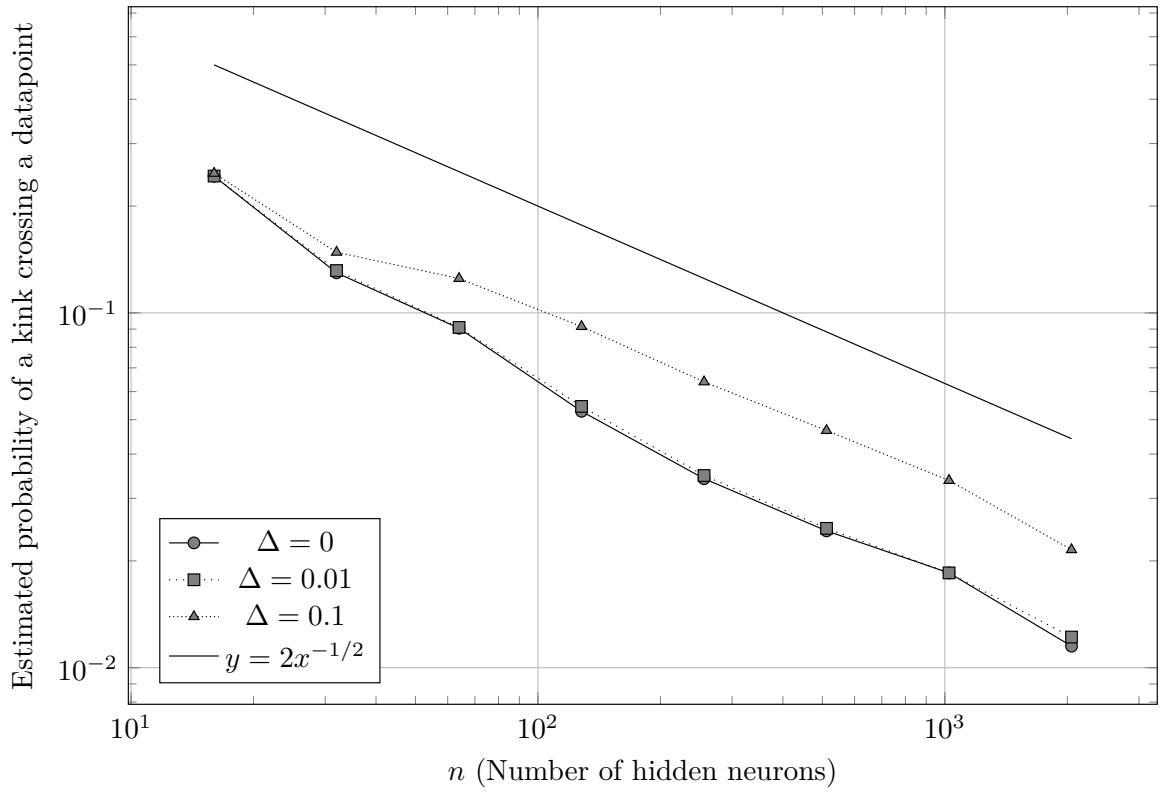


Figure 8: Monte Carlo estimates for the probability that a kink crosses $x = 1$ or $x = -1$ when using stochastic gradient descent, $h = 0.01 \cdot n^{-1}$ and only the early stopping criterion. Training and validation data is sampled from P_Δ for different values of Δ.

$$
\begin{aligned}
&\simeq \left(\frac{1}{\lambda} + \frac{\beta}{\lambda_{\min}(A^{\mathrm{ref}})}\right) \|\overline{v}_0\|_2 \\
&\simeq \left(\frac{1}{\Sigma_{\sigma,a^2} + \Sigma_{\sigma,w^2}} + \frac{1}{\Sigma_{\sigma,a^2} + \Sigma_{\sigma,w^2}}\right) |\Sigma_{\sigma,wa}| \\
&\simeq \frac{1}{n \cdot n^{-p}} \cdot n^{1/2-p/2-q/2} = n^{-1/2(1+q-p)} = n^{-1/2(1+|p-q|)} \ .
\end{aligned}
$$

In the case $p \geq q$, we may not be able to satisfy the conditions on λ and β, but we can still apply step (1) of the proof of Proposition B.2: We have $\lambda_{\min}(A^{\mathrm{ref}}), \lambda_{\max}(A^{\mathrm{ref}}) \simeq n^{1-q}$ and hence

$$
\begin{aligned}
h \sum_{k=0}^{\infty} \|(I - hA^{\mathrm{ref}}M)^k \overline{v}_0\|_2 &\leq \left(h \sum_{k=0}^{\infty} \|(I - hA^{\mathrm{ref}}M)^k\|_2\right) \cdot \|\overline{v}_0\|_2 \\
&\leq \frac{\sqrt{\mathrm{cond}(M)}}{\lambda_{\min}(A^{\mathrm{ref}}M)} \cdot \|\overline{v}_0\|_2 \\
&\simeq \frac{1}{n^{1-q}} n^{1/2-p/2-q/2} = n^{-1/2(1+p-q)} = n^{-1/2(1+|p-q|)} \ .
\end{aligned}
$$

Hence, in both cases we can perform the following heuristic calculation:

$$
\min_i |a_{i,0}| \simeq n^{-1-p/2}
$$

$$\begin{aligned}\max_i |a_{i,k} - a_{i,0}| &\simeq |w_{i,0}| \cdot h \sum_{k=0}^{\infty} \|(I - hA^{\text{ref}} M_D)^k \overline{v}_0\|_\infty \\ &\simeq n^{-q/2} \cdot n^{-1/2(1+|p-q|)} \\ &= n^{-1/2(1+q+|p-q|)} \ .\end{aligned}$$

The kinks now fail to reach the data points with high probability if $-1/2(1+q+|p-q|) < -1-p/2$ (cf. the proof of Theorem 5.37), which can be reformulated as

$$1 + q + |p - q| > 2 + p \Leftrightarrow q - p + |q - p| > 1 \Leftrightarrow q > p + \frac{1}{2} \ .$$

This heuristic criterion is satisfied by He et al. [20] asymptotics ($p = 0$, $q = 1$). If we choose $p = 1$ and $q = 0$ instead, we achieve the same initial distribution for $\Sigma_{\sigma,wa}$ and $\Sigma_{\sigma,wb}$ and hence for the vector v. However, our criterion $q > p + 1/2$ for convergence failure is not satisfied. Indeed, if we initialize the weights independently according to

$$\begin{aligned}a_i &\sim \mathcal{N}(0, 2/n) \\ w_i &\sim \mathcal{N}(0, 2) \\ b_i &= 0 \\ c &= 0 \ ,\end{aligned}$$

we find empirically that the probabilities of kinks crossing 1 or -1 are much higher, as shown in Figure 9. For Xavier initialization [15], which in our case yields $p = 1$ and $q = 1$, the criterion $q > p + 1/2$ is also not satisfied. However, Xavier initialization is not constructed for ReLU activations and yields a different initial distribution for v.

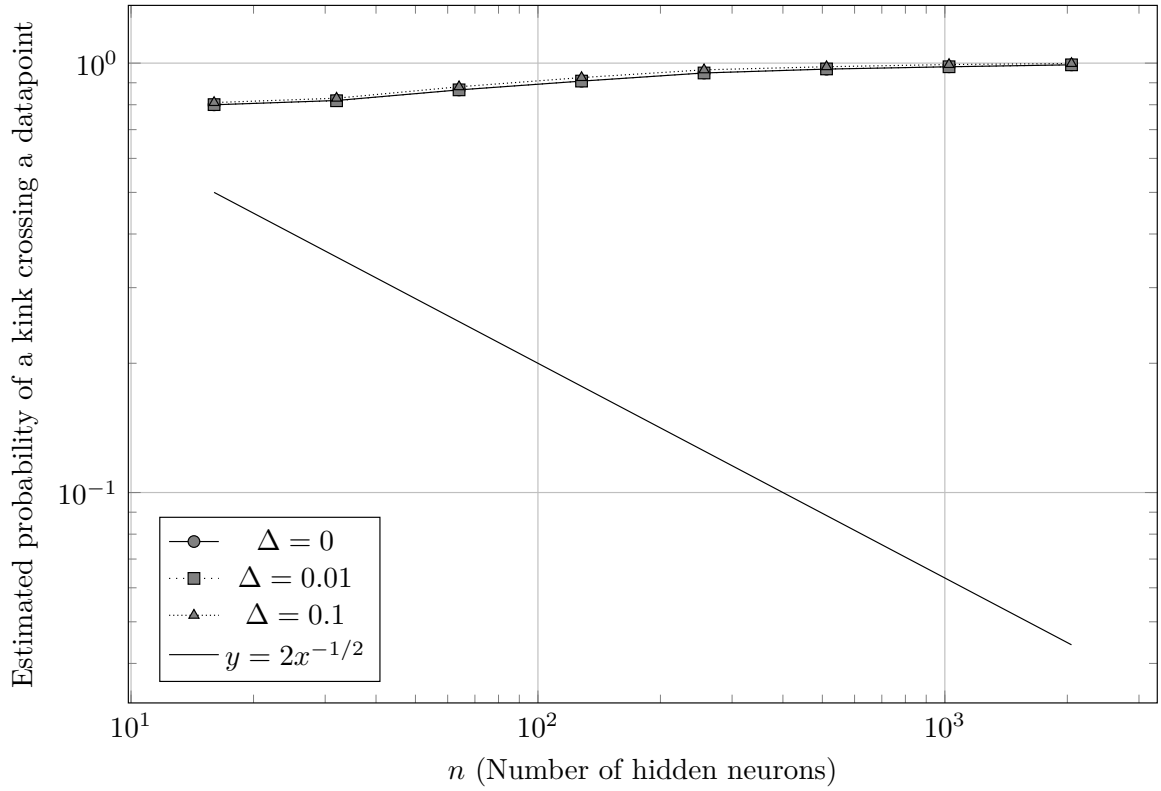


Figure 9: Monte Carlo estimates for the probability that a kink crosses $x = 1$ or $x = -1$ when using gradient descent, *switched variances of w and a at initialization*, $h = \frac{1}{\lambda_{\max}(A^{\text{ref}} M_D)}$ and only the sufficient stopping criterion. Data is sampled from P_Δ for different values of Δ.

7 Conclusion

In Theorem 5.37, Corollary 5.39 and Corollary 5.43, we have shown conditions under which two-layer (Leaky)ReLU networks fail to reach a good loss on a least-squares regression task. In Section 6, we provide numerical evidence that this failure also exists for stochastic gradient descent and practical network sizes. An interesting research question is therefore whether this result can be extended, for example to other optimizers, other initialization methods, deeper nets, higher input dimensions or more data distributions P.

Additionally, one may ask whether the proof can be simplified. Another proof approach would be to find positively invariant sets $S_{n,N} \subseteq \mathbb{R}^{3n+1} \times (\mathbb{R} \times \mathbb{R})^N$ with $P_{n,N}((W_0, D) \in S_{n,N}) \to 1$ and such that no $(W, D) \in S$ constitutes a "good" neural network with respect to the least-squares loss. Here, positively invariant means that $(W, D) \in S_{n,N}$ implies $(W - h\nabla L_D(W), D) \in S_{n,N}$. Such a set $S_{n,N}$ must exist for n, N large enough: The set

$$S_{n,N} \coloneqq \{(W_k(\omega), D(\omega)) \mid \omega \in E_{n,N,\varepsilon,\gamma}, k \in \mathbb{N}_0\}$$

with $W_{k+1} = W_k - h\nabla L_D(W_k)$ is positively invariant, we proved in Theorem 5.37 that it does not contain "good" parameters W and we proved in Theorem 5.25 that $P_{n,N}((W_0, D) \in S_{n,N}) \to 1$. The question is whether one can (explicitly) specify sets $S_{n,N}$ for which these properties are easier to prove, which might considerably reduce the length of the given proof. A disadvantage would be that this does not characterize the trajectories found by gradient descent very precisely.

The matrices A and M have been very useful in this thesis to understand the behavior of gradient descent. We have shown that $\lambda_{\max}(A^{\text{ref}}M)^{-1}$ yields a bound on the step size h such that gradient descent behaves similarly to gradient flow, which is discussed in Remark 5.38. This poses the question whether similar methods can be used to determine useful step sizes in other scenarios. It might also be interesting to investigate how data normalization affects the condition of the matrix M and thus the probabilities of not reaching a global optimum. It is also unclear whether the matrix $M^{1/2}$ or the condition of M have an intuitive interpretation.

In this thesis, it remains open what happens in the case where a kink reaches a data point, which is a necessary, but potentially not sufficient condition for gradient descent to find a global optimum. If gradient descent finds a global optimum, convergence might still be rather slow since a large absolute value of $-b_i/a_i$ suggests that the denominator a_i is rather small. Under certain assumptions, it has been shown that overparameterized networks converge to a global minimum with high probability although few kinks do ever cross data points [11]. Thus, the kinks must have already been well-spread across the data points at initialization. In the scenario investigated in this thesis, kinks do not spread well across the data. This poses the question whether kinks spread well in other (underparameterized) scenarios. Moreover, further research may attempt to clarify whether similar phenomena exist for smooth activation functions.

Acknowledgements I want to thank Ingo Steinwart for coming up with the idea for this topic and providing me with helpful ideas during supervision. I would also like to thank Carsten Scherer for an interesting discussion on analyzing interconnected systems and Dirk Pflüger for agreeing to examine this thesis.

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A Convergence Theory for Deep Learning via Over-Parameterization. In *International Conference on Machine Learning*, pages 242–252, 2019.

[2] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. In *International Conference on Machine Learning*, pages 322–332, 2019.

[3] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[4] Avrim Blum and Ronald L. Rivest. Training a 3-node neural network is NP-complete. In *Advances in neural information processing systems*, pages 494–501, 1989.

[5] Alon Brutzkus and Amir Globerson. Globally Optimal Gradient Descent for a ConvNet with Gaussian Inputs. In *Proceedings of the 34th International Conference on Machine Learning*, pages 605–614. JMLR. org, 2017.

[6] François Chollet et al. Keras. `https://keras.io`, 2015.

[7] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[8] Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient Descent Finds Global Minima of Deep Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 2019.

[9] Simon S. Du, Jason D. Lee, and Yuandong Tian. When is a Convolutional Filter Easy to Learn? In *6th International Conference on Learning Representations*. OpenReview.net, 2018.

[10] Simon S. Du, Jason D. Lee, Yuandong Tian, Aarti Singh, and Barnabas Poczos. Gradient Descent Learns One-hidden-layer CNN: Don't be Afraid of Spurious Local Minima. In *International Conference on Machine Learning*, pages 1338–1347, 2018.

[11] Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient Descent Provably Optimizes Over-parameterized Neural Networks. In *7th International Conference on Learning Representations*. OpenReview.net, 2019.

[12] Armin Eftekhari, ChaeHwan Song, and Volkan Cevher. Nearly Minimal Over-Parametrization of Shallow Neural Networks. *arXiv:1910.03948*, 2019.

[13] András Faragó and Gábor Lugosi. Strong universal consistency of neural network classifiers. *IEEE Transactions on Information Theory*, 39(4):1146–1151, 1993.

[14] Kenji Fukumizu and Shun-ichi Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks*, 13(3):317–327, 2000.

[15] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[16] Gene H. Golub and Charles F. Van Loan. Matrix computations. *The Johns Hopkins University Press, Baltimore, USA*, 1989.

[17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[18] Marco Gori and Alberto Tesi. On the Problem of Local Minima in Backpropagation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 14(1):76–86, 1992.

[19] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[21] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[22] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

[23] Yuanzhi Li and Yang Yuan. Convergence Analysis of Two-layer Neural Networks with ReLU Activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

[24] Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes ReLU network features. *arXiv:1803.08367*, 2018.

[25] Nicole Mücke and Ingo Steinwart. Global Minima of DNNs: The Plenty Pantry. *arXiv:1905.10686*, 2019.

[26] Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *arXiv:1902.04674*, 2019.

[27] Ali Rahimi and Benjamin Recht. Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pages 1313–1320, 2009.

[28] Itay Safran and Ohad Shamir. Spurious Local Minima are Common in Two-Layer ReLU Neural Networks. In *International Conference on Machine Learning*, pages 4430–4438, 2018.

[29] Andrew M. Saxe, James L. Mcclelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *In International Conference on Learning Representations*, 2014.

[30] Mahdi Soltanolkotabi. Learning ReLUs via Gradient Descent. In *Advances in Neural Information Processing Systems*, pages 2007–2017, 2017.

[31] Eduardo D. Sontag. *Mathematical control theory: deterministic finite dimensional systems*, volume 6. Springer Science & Business Media, 2013.

[32] Eduardo D. Sontag and Héctor J. Sussmann. Backpropagation can give rise to spurious local minima even for networks without hidden layers. *Complex Systems*, 3(1):91–106, 1989.

[33] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv:1605.08361*, 2016.

[34] Ingo Steinwart. A Sober Look at Neural Network Initializations. *arXiv:1903.11482*, 2019.

[35] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[36] Halbert White. Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural networks*, 3(5):535–549, 1990.

[37] Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *arXiv:1904.00687*, 2019.

[38] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. A critical view of global optimality in deep learning. *arXiv:1802.03487*, 2018.

[39] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small nonlinearities in activation functions create bad local minima in neural networks. *7th International Conference on Learning Representations*, 2019.

[40] Difan Zou and Quanquan Gu. An Improved Analysis of Training Over-parameterized Deep Neural Networks. *arXiv:1906.04688*, 2019.

A Stochastic Proofs

In this section, we prove the main theorem from Section 5.3, Theorem 5.25. In order to show that W_0 and D likely have certain properties, we employ concentration inequalities. Besides Markov's inequality, we use Hoeffding's inequality:

Theorem A.1 (Hoeffding's inequality, e.g. Theorem 6.10 in [35])**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $a < b, n \geq 1$ and $X_1, \ldots, X_n : \Omega \to [a, b]$ be independent random variables. Then, for $\tau \geq 0$, we have*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| \geq (b-a)\sqrt{\frac{\tau}{2n}}\right) \leq 2e^{-\tau} \ .$$

Using Markov and Hoeffding, we can prove an asymptotic concentration result. The intuition behind this result is that for random variables $X_1, \ldots, X_n$ with mean zero and finite variance, the value $n^{-1/2}(X_1 + \ldots + X_n)$ asymptotically has a Gaussian distribution by the central limit theorem. The tail of the Gaussian distribution decreases stronger than any inverse polynomial: If Φ is the CDF of a Gaussian distribution, then $\Phi(\beta n^{\varepsilon}) = O(n^{-\gamma})$ for all $\beta, \varepsilon, \gamma > 0$, where the constant in $O(n^{-\gamma})$ depends on $\beta, \varepsilon, \gamma$. However, the central limit theorem does not tell us how close the CDF of $n^{-1/2}(X_1 + \ldots + X_n)$ is to Φ, so we use Markov's and Hoeffding's inequalities instead.

Lemma A.2. *Let Q be a probability distribution on $\mathbb{R}$ with $\mu_p \coloneqq \int |x|^p \,\mathrm{d}Q(x) < \infty$ for all $p \in (1, \infty)$. For $n \in \mathbb{N}$, let $(\Omega_n, \mathcal{F}_n, P_n)$ be probability spaces with independent Q-distributed random variables $X_{n1}, X_{n2}, \ldots, X_{nn} : \Omega_n \to \mathbb{R}$. Then, the random variables $S_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_{ni}$ satisfy*

$$P_n\left(|S_n - \mathbb{E}S_n| \geq \beta n^{\varepsilon - 1/2}\right) = O(n^{-\gamma})$$

for all $\beta, \varepsilon, \gamma > 0$, where the constant in $O(n^{-\gamma})$ may depend on $\beta, \varepsilon, \gamma$ (cf. Definition 5.24).

Proof. Let $\beta, \varepsilon, \gamma > 0$ be fixed. For $n \in \mathbb{N}$ and $b > 0$ to be determined later, define $B \coloneqq \{\max_{1 \leq i \leq n} |X_{ni}| \leq b\}$. Then, for all $p > 0$,

$$P_n(B^c) \leq \sum_{i=1}^{n} P_n(|X_{ni}| \geq b) \leq n P_n(|X_{n1}|^p \geq b^p) \overset{\text{Markov}}{\leq} n\frac{\mathbb{E}_{P_n}|X_{n1}|^p}{b^p} = n\frac{\mu_p}{b^p} \ . \quad \text{(A.1)}$$

Since $S_n = S_n \mathbb{1}_B + S_n \mathbb{1}_{B^c}$, we can now bound

$$\begin{aligned} P_n(|S_n - \mathbb{E}S_n| \geq \beta n^{\varepsilon - 1/2}) \leq{}& \underbrace{P_n(|S_n \mathbb{1}_B - \mathbb{E}(S_n \mathbb{1}_B)| \geq \beta n^{\varepsilon - 1/2}/2)}_{\text{I}} \\ &+ \underbrace{P_n(|S_n \mathbb{1}_{B^c} - \mathbb{E}(S_n \mathbb{1}_{B^c})| \geq \beta n^{\varepsilon - 1/2}/2)}_{\text{II}} . \end{aligned}$$

With $\tau \coloneqq \gamma \log n$ and $b \coloneqq \beta n^{\varepsilon}\sqrt{\frac{1}{8\gamma \log n}}$, we have

$$(b - (-b))\sqrt{\frac{\tau}{2n}} = 2\beta n^{\varepsilon}\sqrt{\frac{1}{8\gamma \log n}} \cdot \sqrt{\frac{\gamma \log n}{2n}} = \beta n^{\varepsilon - 1/2}/2$$

and hence, Hoeffding (Theorem A.1) applied to $X_i \coloneqq X_{ni}\mathbb{1}_{|X_{ni}|\leq b}$ yields

$$\mathrm{I} \leq 2e^{-\tau} = 2n^{-\gamma} \ .$$

Moreover, we have

$$\begin{aligned} |\mathbb{E}_{P^n}(S_n \mathbb{1}_{B^c})| &\leq \|S_n \mathbb{1}_{B^c}\|_{\mathcal{L}_1(P_n)} \overset{\text{Hölder}}{\leq} \|S_n\|_{\mathcal{L}_2(P_n)} \|\mathbb{1}_{B^c}\|_{\mathcal{L}_2(P_n)} \\ &\leq \left(\frac{1}{n} \sum_{i=1}^{n} \|X_{ni}\|_{\mathcal{L}_2(P_n)} \right) \|\mathbb{1}_{B^c}\|_{\mathcal{L}_2(P_n)} = \sqrt{\mu_2} \sqrt{P_n(B^c)} \\ &\overset{(A.1)}{\leq} \sqrt{\mu_2} \sqrt{n \frac{\mu_p}{b^p}} = \sqrt{\frac{\mu_2 \mu_p}{\beta^p}} (8\gamma \log n)^{p/4} n^{(1-\varepsilon p)/2} \ . \end{aligned}$$

If we choose $p \geq 2/\varepsilon$, we have $(1-\varepsilon p)/2 \leq -1/2 < \varepsilon - 1/2$ and hence $|\mathbb{E}(S_n \mathbb{1}_{B^c})| < \beta n^{\varepsilon - 1/2}/2$ for n large enough. Now, let n be sufficiently large. For $\omega \in B$, we have $S_n(\omega)\mathbb{1}_{B^c}(\omega) = 0$ and hence $|S_n(\omega)\mathbb{1}_{B^c}(\omega) - \mathbb{E}(S_n \mathbb{1}_{B^c})| < \beta n^{\varepsilon - 1/2}/2$. Thus,

$$\mathrm{II} \leq P(B^c) \leq n \frac{\mu_p}{b^p} = \frac{\mu_p}{\beta^p} \cdot (8\gamma \log n)^{p/2} n^{1-\varepsilon p} \ .$$

If we choose $p > (1+\gamma)/\varepsilon$, then $1 - \varepsilon p < -\gamma$ and hence $\mathrm{II} = O(n^{-\gamma})$. □

Now, we can prove that certain properties of the initialization W_0 hold with high probability. We will see that in all properties except (W4), the tails of the probability distributions decrease so quickly that only the parameter γ in (W4) is relevant for the rate of convergence.

Proposition A.3. *Let $\varepsilon, \gamma > 0$. As in Definition 5.23, let $E^{\mathrm{W}}_{n,\varepsilon,\gamma} \subseteq \mathbb{R}^{3n+1}$ for $n \in \mathbb{N}$ denote the set of all $W_0 \in \mathbb{R}^{3n+1}$ for which the following properties hold:*

(W1) $b_{i,0} = c_0 = 0$,

(W2) $\max_i |w_{i,0}| \leq n^{-1/2+\varepsilon}$,

(W3) $\max_i |a_{i,0}| \leq n^{\varepsilon}$,

(W4) $\min_i |a_{i,0}| \geq n^{-(1+\gamma)}$,

(W5) $\Sigma_{\sigma,a^2,0} \in [nc_a/4, nc_a]$ *for all* $\sigma \in \{-1,1\}$,

(W6) $\Sigma_{\sigma,w^2,0} \in [c_w/4, c_w]$ *for all* $\sigma \in \{-1,1\}$,

(W7) $|\Sigma_{\sigma,wa,0}| \leq n^{\varepsilon}$ *for all* $\sigma \in \{-1,1\}$.

Then, $P_{n,N}(W_0 \notin E^{\mathrm{W}}_{n,\varepsilon,\gamma}) = O(n^{-\gamma})$, where the constant in $O(n^{-\gamma})$ may depend on ε and γ (cf. Definition 5.24).

Proof. We will show the statement for each of the properties (W1) – (W7) individually, the rest follows by the union bound. Note that $a_{i,0}$ and $\sqrt{n}w_{i,0}$ have a distribution independent of n. Recall from Definition 5.21 that P_n denotes the distribution of $W_0 \in \mathbb{R}^{3n+1}$.

By property (Q2) in Assumption 5.19,

$$\int_{\mathbb{R}} |x|^p p_Q^{\mathrm{wa}}(x)\,\mathrm{d}x < \infty$$

for all $p \in (0,\infty)$. All random variables X that will be used later in the argument are simple combinations of random variables that are distributed according to Q^{wa}. It can be shown (using the Minkowski and Hölder inequalities) that these random variables satisfy $\mathbb{E}|X|^p < \infty$ for all $p \in (0,\infty)$. We will repeatedly use this property.

(W1) True by Definition 5.21.

(W2) For $p > 0$, define

$$c_p \coloneqq \mathbb{E}_{P_n}|\sqrt{n}w_{i,0}|^p = c_w^{p/2}\mathbb{E}_{P_n}\Bigg|\underbrace{\sqrt{\frac{n}{c_w}}w_{i,0}}_{\sim Q^{\mathrm{wa}}}\Bigg|^p = c_w^{p/2}\int_{\mathbb{R}}|x|^p p_Q^{\mathrm{wa}}(x)\,\mathrm{d}x < \infty\ ,$$

hence c_p is independent of i and n. By the Markov inequality,

$$\begin{aligned} P_n\left(|w_{i,0}| \geq n^{-1/2+\varepsilon}\right) &= P_n\left(|w_{i,0}|^p \geq n^{(-1/2+\varepsilon)p}\right) \\ &\leq \frac{\mathbb{E}|w_{i,0}|^p}{n^{(-1/2+\varepsilon)p}} = \frac{\mathbb{E}|\sqrt{n}w_{i,0}|^p}{n^{p/2}n^{(-1/2+\varepsilon)p}} = c_p n^{-\varepsilon p}\ . \end{aligned}$$

By choosing $p = (1+\gamma)/\varepsilon$, we can use the union bound to conclude

$$P_n\left(\max_i |w_{i,0}| \geq n^{-1/2+\varepsilon}\right) \leq n \cdot c_p n^{-\varepsilon p} = c_p n^{1-\varepsilon p} = O(n^{-\gamma})\ .$$

(W3) Similar to (W2).

(W4) By property (Q1) of Assumption 5.19, Q^{wa} has a probability density p_Q^{wa} that is bounded by B_Q^{wa}. Thus, for all $\delta \geq 0$, we obtain

$$P_n(|a_{i,0}| \leq \delta) = \int_{-\delta}^{\delta} p_Q^{\mathrm{wa}}(x)\,\mathrm{d}x \leq 2\delta \cdot B_Q^{\mathrm{wa}}\ .$$

Therefore,

$$\begin{aligned} P_n\left(\min_i |a_{i,0}| \leq n^{-(1+\gamma)}\right) &\leq \sum_{i=1}^n P_n\left(|a_{i,0}| \leq n^{-(1+\gamma)}\right) \leq n \cdot 2B_Q^{\mathrm{wa}} n^{-(1+\gamma)} \\ &= O(n^{-\gamma})\ . \end{aligned}$$

(W5) For the next three properties, we need some preparation. Let

$$\begin{aligned} A_{\sigma,i} &\coloneqq \mathbb{1}_{(0,\infty)}(\sigma a_{i,0})a_{i,0} \\ W_{\sigma,i} &\coloneqq \mathbb{1}_{(0,\infty)}(\sigma a_{i,0})w_{i,0}\ . \end{aligned}$$

Note that the indicator function is applied to $\sigma a_{i,0}$ in both definitions. Then, $\Sigma_{\sigma,a^2,0} = \sum_{i\in I_\sigma} a_{i,0}^2 = \sum_{i=1}^n A_{\sigma,i}^2$ and similarly for $\Sigma_{\sigma,w^2,0}$ and $\Sigma_{\sigma,wa,0}$. We obtain

$$\mathbb{E}_{P_n} A_{\sigma,i}^2 \quad = \quad \int A_{\sigma,i}^2\,\mathrm{d}P_n = \int \left(\mathbb{1}_{(0,\infty)}(\sigma a_{i,0})a_{i,0}\right)^2\,\mathrm{d}P_n$$

$$
\begin{aligned}
&= \int_{\{\sigma a_{i,0}>0\}} a_{i,0}^2 \,\mathrm{d}P_n = \int_{(0,\infty)} (\sqrt{c_a}x)^2 p_Q^{\mathrm{wa}}(x)\,\mathrm{d}x \\
&\overset{(Q1)}{=} \frac{1}{2} c_a \int_{\mathbb{R}} x^2 p_Q^{\mathrm{wa}}(x)\,\mathrm{d}x \\
&\overset{(Q3)}{=} \frac{c_a}{2} \,. \\
\mathbb{E}_{P_n} W_{\sigma,i}^2 &= \mathbb{E}_{P_n}\left(\left(\mathbb{1}_{(0,\infty)}(\sigma a_{i,0})\right)^2 w_{i,0}^2\right) \\
&\overset{\text{indep.}}{=} \left(\mathbb{E}_{P_n}\left(\mathbb{1}_{(0,\infty)}(\sigma a_{i,0})\right)^2\right) \cdot \left(\mathbb{E}_{P_n} w_{i,0}^2\right) \\
&= P_n(\sigma a_{i,0} > 0) \cdot \int \left(\sqrt{\frac{c_w}{n}} x\right)^2 p_Q^{\mathrm{wa}}(x)\,\mathrm{d}x \\
&\overset{(Q1),\,(Q3)}{=} \frac{1}{2} \cdot \frac{c_w}{n} \,. \\
\mathbb{E}_{P_n} W_{\sigma,i} A_{\sigma,i} &= \mathbb{E}_{P_n} \mathbb{1}_{(0,\infty)}(\sigma a_{i,0}) w_{i,0} a_{i,0} \\
&\overset{\text{indep.}}{=} \Big(\underbrace{\mathbb{E}_{P_n} w_{i,0}}_{\overset{(Q1)}{=}0}\Big) \cdot \left(\mathbb{E}_{P_n} \mathbb{1}_{(0,\infty)}(\sigma a_{i,0}) a_{i,0}\right) \\
&= 0 \,.
\end{aligned}
$$

Now, define

$$
S_n \coloneqq \frac{\Sigma_{\sigma,a^2,0}}{n} = \frac{1}{n}\sum_{i=1}^n A_{\sigma,i}^2 \,,
$$

which is an average of n i.i.d. variables that are p-integrable for every $p > 0$. Then, $\mathbb{E}_{P_n} S_n = \mathbb{E}_{P_n} A_{\sigma,1}^2 = c_a/2$ and Lemma A.2 with $\varepsilon = 1/2, \beta = c_a/4$ yields:

$$
P_n\left(\left|S_n - \frac{c_a}{2}\right| \geq \frac{c_a}{4}\right) = O(n^{-\gamma}) \,.
$$

Hence,

$$
\begin{aligned}
P_n(\Sigma_{\sigma,a^2,0} \notin [nc_a/4, nc_a]) &= P_n\left(S_n \notin [c_a/4, c_a]\right) \leq P_n(S_n \notin [c_a/4, 3c_a/4]) \\
&= O(n^{-\gamma}) \,.
\end{aligned}
$$

(W6) An analogous argument yields $P_n(\Sigma_{\sigma,w^2,0} \notin [c_w/4, c_w]) = O(n^{-\gamma})$.

(W7) Let $S_n \coloneqq \frac{1}{n}\sum_{i=1}^n A_{\sigma,i} \cdot \sqrt{n} W_{\sigma,i} = \Sigma_{\sigma,wa,0}/\sqrt{n}$. Then, $\mathbb{E}_{P_n} S_n = 0$ and thus

$$
P_n(|\Sigma_{\sigma,wa,0}| \leq n^{\varepsilon}) = P_n(|S_n| \leq n^{\varepsilon - 1/2}) \overset{\text{Lemma A.2}}{=} O(n^{-\gamma}) \,. \qquad \square
$$

Now, we want to investigate stochastic properties of the dataset. In order to show that M_P^{-1} (as defined in Assumption 5.16) is likely close to M_D^{-1} (as defined in Definition 5.5), we need the following lemma, which is similar for example to Theorem 2.3.4 in [16]:

Lemma A.4. *Let $A, B \in \mathbb{R}^{m\times m}$ and let $\|\cdot\|$ be a matrix norm on $\mathbb{R}^{m\times m}$. If A is invertible and $\|A^{-1}\|\|A - B\| < 1$, then B is invertible with*

$$
\|B^{-1} - A^{-1}\| \leq \|A^{-1}\|\|A - B\|\|B^{-1}\|, \qquad \|B^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|A - B\|} \,.
$$

Proof. We have $B = A(I - A^{-1}(A-B))$ and since $\|A^{-1}(A-B)\| \leq \|A^{-1}\|\|A-B\| < 1$, the Neumann series implies that

$$(I - A^{-1}(A-B))^{-1} = \sum_{k=0}^{\infty} (A^{-1}(A-B))^k \ .$$

Hence B is invertible with $B^{-1} - A^{-1} = A^{-1}(A-B)B^{-1}$ and

$$B^{-1} = (I - A^{-1}(A-B))^{-1}A^{-1} = \sum_{k=0}^{\infty} (A^{-1}(A-B))^k A^{-1} \ ,$$

which yields both bounds using the submultiplicativity of $\|\cdot\|$. □

Now, we can show that for large N, a dataset $D \sim P^N$ likely has characteristics that are close to P.

Proposition A.5. *Let $\varepsilon, \gamma > 0$. As in Definition 5.23, let $E^{\mathrm{D}}_{N,\varepsilon} \subseteq (\mathbb{R}\times\mathbb{R})^N$ be the set of all datasets $D \in (\mathbb{R}\times\mathbb{R})^N$ for which the following properties are satisfied:*

(D1) v_D^{opt} is well-defined, i.e. M_D is invertible, and $\|v_P^{\mathrm{opt}} - v_D^{\mathrm{opt}}\|_\infty \leq N^{(\varepsilon-1)/2}$.

(D2) $\lambda_{\min}(M_D) \geq \frac{1}{2}\lambda_{\min}(M_P)$ and $\lambda_{\max}(M_D) \leq 2\lambda_{\max}(M_P)$.

(D3) $\min_j |x_j| \geq m_P$.

Then, $P_{n,N}(D \notin E^{\mathrm{D}}_{N,\varepsilon}) = P^N((E^{\mathrm{D}}_{N,\varepsilon})^c) = O(N^{-\gamma})$, where the constant in $O(N^{-\gamma})$ may depend on ε, γ (cf. Definition 5.24).

Proof. Again, we bound the probabilities for each property separately.

(D1) For $\sigma \in \{-1, 1\}$, define

$$S_N \coloneqq (M_{D,\sigma})_{11} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{(0,\infty)}(\sigma x_j) x_j^2 \ .$$

Then,

$$\mathbb{E}_{P^N} S_N = \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}_{P^N}(\mathbb{1}_{(0,\infty)}(\sigma x_j)x_j^2) = \mathbb{E}_{(x,y)\sim P}\mathbb{1}_{(0,\infty)}(\sigma x)x^2 = (M_{P,\sigma})_{11} \ .$$

Because of property (P1) from Assumption 5.16, P has finite moments and we can apply Lemma A.2: For all $\beta > 0$,

$$P^N\left(|(M_{D,\sigma})_{11} - (M_{P,\sigma})_{11}| \geq \beta N^{(\varepsilon-1)/2}\right) = O(N^{-\gamma}) \ .$$

We can get similar bounds for other entries of $M_{D,\sigma}$ and $\hat{u}^0_{D,\sigma}$. Since

$$M_D - M_P = \tilde{P}\begin{pmatrix} M_{D,1} - M_{P,1} & 0 \\ 0 & M_{D,-1} - M_{P,-1}\end{pmatrix}\tilde{P}, \quad \hat{u}^0_D = \tilde{P}\begin{pmatrix}\hat{u}^0_{D,1} \\ \hat{u}^0_{D,-1}\end{pmatrix},$$

and $\|\tilde{P}\|_\infty = 1$, the union bound implies that the following properties hold with probability $1 - O(n^{-\gamma})$:

$$\|M_D - M_P\|_\infty \leq 2\beta N^{(\varepsilon-1)/2}, \quad \|\hat{u}_D^0 - \hat{u}_P^0\|_\infty \leq \beta N^{(\varepsilon-1)/2} . \tag{A.2}$$

Now assume that (A.2) holds. Set $A \coloneqq M_P, B \coloneqq M_D, a \coloneqq \hat{u}_P^0, b \coloneqq \hat{u}_D^0$. As shown in Assumption 5.16, condition (P2) implies that A is invertible. Without loss of generality, we can assume $\varepsilon < 1/2$. Then, for N large enough,

$$\|A^{-1}\|_\infty \|A - B\|_\infty \leq \|A^{-1}\|_\infty 2\beta N^{(\varepsilon-1)/2} \leq \frac{1}{2} .$$

Hence, Lemma A.4 implies that $B = M_D$ is invertible with $\|B^{-1}\|_\infty \leq 2\|A^{-1}\|_\infty$ and

$$\begin{aligned}
\|v_D^{\mathrm{opt}} - v_P^{\mathrm{opt}}\|_\infty &= \|B^{-1}b - A^{-1}a\|_\infty \\
&\leq \|B^{-1}\|_\infty \|b - a\|_\infty + \|B^{-1} - A^{-1}\|_\infty \|a\|_\infty \\
&\leq \|B^{-1}\|_\infty \|b - a\|_\infty + \|A^{-1}\|_\infty \|A - B\|_\infty \|B^{-1}\|_\infty \|a\|_\infty \\
&\leq 2\|A^{-1}\|_\infty \left(\|b - a\|_\infty + \|A^{-1}\|_\infty \|a\|_\infty \|B - A\|_\infty \right) \\
&\overset{(A.2)}{\leq} 4\|A^{-1}\|_\infty (1 + \|A^{-1}\|_\infty \|a\|_\infty) \beta N^{(\varepsilon-1)/2} .
\end{aligned}$$

We can choose $\beta > 0$ such that $4\|A^{-1}\|_\infty(1 + \|A^{-1}\|_\infty\|a\|_\infty)\beta \leq 1$. Therefore,

$$\|v_{P,\sigma}^{\mathrm{opt}} - v_{D,\sigma}^{\mathrm{opt}}\|_\infty \leq N^{(\varepsilon-1)/2}$$

with probability $1 - O(N^{-\gamma})$.

(D2) For each $v \in \mathbb{R}^4$, we have

$$|v^\top M_D v - v^\top M_P v| \leq \|v\|_2 \|M_D - M_P\|_2 \|v\|_2 \leq \sqrt{4}\|M_D - M_P\|_\infty \|v\|_2^2$$

since $\|\cdot\|_2 \leq \sqrt{4}\|\cdot\|_\infty$ on $\mathbb{R}^{4\times 4}$ as mentioned in Definition 4.1. If we choose $\beta > 0$ small enough such that (A.2) implies $2\|M_D - M_P\|_\infty \leq \lambda_{\min}(M_P)/2$, it follows that

$$\begin{aligned}
\lambda_{\min}(M_D) &= \inf_{\|v\|_2=1} v^\top M_D v \geq \inf_{\|v\|_2=1} v^\top M_P v - |v^\top M_P v - v^\top M_D v| \\
&\geq \lambda_{\min}(M_P) - 2\|M_D - M_P\|_\infty \geq \lambda_{\min}(M_P)/2 .
\end{aligned}$$

Since (A.2) holds with probability $1 - O(N^{-\gamma})$, we have $\lambda_{\min}(M_D) \geq \lambda_{\min}(M_P)/2$ with probability $1 - O(N^{-\gamma})$. The probability for $\lambda_{\max}(M_D) \leq 2\lambda_{\max}(M_P)$ can be bounded similarly.

(D3) This holds with probability one due to property (P3) from Assumption 5.16. □

Remark A.6. In this section, we did not use property (P4) and (P5) from Assumption 5.16. Property (P4) is used in Proposition 5.29 and (P5) is used in Corollary 5.43. ◀

B L1 Bounds

In this section, we show how to bound $h\sum_{k=0}^{\infty}\|\overline{v}_k\|_\infty$, where $(\overline{v}_k)_{k\in\mathbb{N}_0}$ solves $\delta\overline{v}_k = -hA_kM\overline{v}_k$. We first show how to proceed for the reference system $\delta\overline{v}_k = -hA^{\mathrm{ref}}M\overline{v}_k$ and then how to apply this bound to the original system $\delta\overline{v}_k = -hA_kM\overline{v}_k$. For analyzing the reference system $\delta\overline{v}_k = -hA^{\mathrm{ref}}M\overline{v}_k$, we will use Cauchy's interlacing theorem:

Theorem B.1 (Cauchy's interlacing theorem)**.** *Let*

$$E = \begin{pmatrix} E_{11} & E_{12} \\ E_{12}^\top & E_{22} \end{pmatrix} \in \mathbb{R}^{m\times m}$$

be symmetric with $E_{11}\in\mathbb{R}^{m_1\times m_1}$. Let $\lambda_1(E)\geq\lambda_2(E)\geq\ldots\geq\lambda_m(E)$ be the eigenvalues of E and let $\lambda_1(E_{11})\geq\ldots\geq\lambda_{m_1}(E_{11})$ be the eigenvalues of E_{11}. Then,

$$\lambda_i(E)\geq\lambda_i(E_{11})\geq\lambda_{i+(m-m_1)}(E)$$

for all $i\in\{1,\ldots,m_1\}$.

Proof. See e.g. Corollary III.1.5 in [3]. □

The following proposition is used to analyze the reference system. Its proof uses the facts from Section 4. The idea is that A_1 should contain the "large" eigenvalues of A.

Proposition B.2. *Let $0\prec A, M\in\mathbb{R}^{m\times m}$ with*

$$A = \begin{pmatrix} A_1 & \\ & A_2 \end{pmatrix}, \quad M = \begin{pmatrix} M_{11} & M_{12} \\ M_{12}^\top & M_{22} \end{pmatrix}$$

and $A_1, M_{11}\in\mathbb{R}^{m_1\times m_1}, A_2, M_{22}\in\mathbb{R}^{m_2\times m_2}$. With $\lambda_{\min}$, $\lambda_{\max}$ and cond *defined in Definition 4.1, assume*

$$\lambda \coloneqq \lambda_{\min}(A_1)\lambda_{\min}(M) \geq \left(1+2\sqrt{m_1}\sqrt{\mathrm{cond}(M)}\right)\lambda_{\max}(A_2)\lambda_{\max}(M) \tag{B.1}$$

such that

$$\beta \coloneqq \frac{\lambda_{\max}(A_2)\lambda_{\max}(M)}{\lambda-\lambda_{\max}(A_2)\lambda_{\max}(M)} \leq \frac{1}{2\sqrt{m_1}\sqrt{\mathrm{cond}(M)}}\,. \tag{B.2}$$

Moreover, let $h>0$ with[6]

$$h \leq \frac{1}{\lambda_{\max}(AM)}\,. \tag{B.3}$$

Then, for $v=(v_1,v_2)^\top\in\mathbb{R}^4$ with $v_1,v_2\in\mathbb{R}^2$,

$$h\sum_{k=0}^{\infty}\|(I-hAM)^k\|_2 \leq \frac{\sqrt{\mathrm{cond}(M)}}{\lambda_{\min}(AM)}$$

$$h\sum_{k=0}^{\infty}\|(I-hAM)^kv\|_2 \leq 2\sqrt{\mathrm{cond}(M)}\left(\left(\frac{1}{\lambda}+\frac{2\sqrt{m_1}\sqrt{\mathrm{cond}(M)}\beta}{\lambda_{\min}(AM)}\right)\|v_1\|_2 + \frac{1}{\lambda_{\min}(AM)}\|v_2\|_2\right).$$

[6] Cf. Lemma 5.28.

Proof. The proof is divided in multiple steps.

(1) *Diagonalization yields a simple bound:*

As shown in Lemma 5.28, the matrix AM is similar to the symmetric matrix

$$\tilde{A} \coloneqq M^{1/2}AM^{1/2} = M^{1/2}(AM)M^{-1/2} \succ 0 \ .$$

The matrix $\tilde{A}$ can thus be orthogonally diagonalized as $\tilde{A} = UDU^\top$ with U orthogonal and D diagonal such that D contains the eigenvalues of $\tilde{A}$ in descending order. Then, $I - hD$ only contains non-negative entries due to (B.3) with its maximal entry being $1 - h\lambda_{\min}(AM)$. Thus, $\|(I - hD)^k\|_2 = (1 - h\lambda_{\min}(AM))^k$. By applying $(I - hAM)M^{-1/2} = M^{-1/2} - hAM^{-1/2} = M^{-1/2}(I - h\tilde{A})$ inductively, we find $(I - hAM)^k M^{-1/2} = M^{-1/2}(I - h\tilde{A})^k$. We can now compute

$$\begin{aligned}
h\sum_{k=0}^{\infty} \|(I - hAM)^k\|_2 &= h\sum_{k=0}^{\infty} \|M^{-1/2}(I - h\tilde{A})^k M^{1/2}\|_2 \\
&= h\sum_{k=0}^{\infty} \|M^{-1/2}U(I - hD)^k U^\top M^{1/2}\|_2 \\
&\leq \|M^{-1/2}\|_2 \|M^{1/2}\|_2 \cdot h\sum_{k=0}^{\infty} \|(I - hD)^k\|_2 \\
&= \operatorname{cond}(M^{1/2}) h \sum_{k=0}^{\infty} (1 - h\lambda_{\min}(AM))^k \\
&= \sqrt{\operatorname{cond}(M)} \frac{h}{1 - (1 - h\lambda_{\min}(AM))} \\
&= \frac{\sqrt{\operatorname{cond}(M)}}{\lambda_{\min}(AM)} \ .
\end{aligned} \tag{B.4}$$

(2) *AM has m_1 "large" eigenvalues:*

Let

$$M^{1/2} = \begin{pmatrix} \tilde{M}_{11} & \tilde{M}_{12} \\ \tilde{M}_{12}^\top & \tilde{M}_{22} \end{pmatrix}$$

be the block decomposition of $M^{1/2}$. Then,

$$M^{1/2}AM^{1/2} = \begin{pmatrix} \tilde{M}_{11}A_1\tilde{M}_{11} + \tilde{M}_{12}A_2\tilde{M}_{12}^\top & * \\ * & * \end{pmatrix}$$

and by Cauchy's interlacing theorem (Theorem B.1),

$$\begin{aligned}
\lambda_{m_1}(\tilde{A}) &\geq \lambda_{m_1}(\tilde{M}_{11}A_1\tilde{M}_{11} + \tilde{M}_{12}A_2\tilde{M}_{12}^\top) \\
&= \lambda_{\min}(\tilde{M}_{11}A_1\tilde{M}_{11} + \tilde{M}_{12}A_2\tilde{M}_{12}^\top) \\
&\geq \lambda_{\min}(\tilde{M}_{11}A_1\tilde{M}_{11}) \geq \lambda_{\min}(A_1)\lambda_{\min}(\tilde{M}_{11})^2 \\
&\geq \lambda_{\min}(A_1)\lambda_{\min}(M^{1/2})^2 = \lambda_{\min}(A_1)\lambda_{\min}(M) = \lambda \ .
\end{aligned} \tag{B.5}$$

(3) *Lower components of eigenvectors to large eigenvalues are small:*

Let $w = (w_1, w_2)^\top$ be an eigenvector of AM with eigenvalue $\lambda_w \geq \lambda$. The lower part of the identity $\lambda_w w = AMw$ reads as

$$\lambda_w w_2 = A_2 M_{12}^\top w_1 + A_2 M_{22} w_2 \ ,$$

which yields

$$\begin{aligned}\lambda\|w_2\|_2 \leq \lambda_w\|w_2\|_2 &\leq \|A_2\|_2\|M_{12}^\top\|_2\|w_1\|_2 + \|A_2\|_2\|M_{22}\|_2\|w_2\|_2 \\ &\leq \lambda_{\max}(A_2)\lambda_{\max}(M)\|w_1\|_2 + \lambda_{\max}(A_2)\lambda_{\max}(M)\|w_2\|_2\end{aligned}$$

and, since $\lambda - \lambda_{\max}(A_2)\lambda_{\max}(M) \overset{(B.1)}{>} 0$,

$$\|w_2\|_2 \leq \frac{\lambda_{\max}(A_2)\lambda_{\max}(M)}{\lambda - \lambda_{\max}(A_2)\lambda_{\max}(M)}\|w_1\|_2 = \beta\|w_1\|_2 \ . \tag{B.6}$$

(4) *The first m_1 eigenvectors of AM are "well-conditioned":*

Let

$$U = \begin{pmatrix} U_1 & U_2 \end{pmatrix} = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} \ .$$

Since $U^\top U = I_m$, we have $U_1^\top U_1 = I_{m_1}$. Moreover, let

$$F = \begin{pmatrix} F_1 & F_2 \end{pmatrix} \coloneqq U_1^\top M^{1/2}, \qquad W = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \coloneqq M^{-1/2} U_1 \ .$$

The columns of W are the eigenvectors of AM to the m_1 largest eigenvalues:

$$\begin{aligned}AMW &= M^{-1/2}M^{1/2}AM^{1/2}U_1 = M^{-1/2}UDU^\top U_1 = M^{-1/2}UD\begin{pmatrix} I_{m_1} \\ 0 \end{pmatrix} \\ &= M^{-1/2}U_1 D_1 = WD_1 \ ,\end{aligned} \tag{B.7}$$

where D_1 is the upper left $m_1 \times m_1$ block of D. Thus,

$$\begin{aligned}\|F\|_2 &\leq \|U_1^\top\|_2\|M^{1/2}\|_2 = 1 \cdot \lambda_{\max}(M^{1/2}) = \lambda_{\max}(M)^{1/2} \\ \|W\|_2 &\leq \|M^{-1/2}\|_2\|U_1\|_2 = \lambda_{\max}(M^{-1/2}) \cdot 1 = \lambda_{\min}(M)^{-1/2} \\ \|W_2\|_2 \leq \|W_2\|_F &\overset{(B.6)}{\leq} \beta\|W_1\|_F \leq \beta\|W\|_F \leq \beta\sqrt{m_1}\|W\|_2 \leq \beta\sqrt{m_1}\lambda_{\min}(M)^{-1/2} \ .\end{aligned}$$

We want to show that W_1^{-1} exists and $\|W_1^{-1}\|_2$ is sufficiently small. Observe that $I = U_1^\top U_1 = FW = F_1W_1 + F_2W_2$ and

$$\|F_2W_2\|_2 \leq \|F_2\|_2\|W_2\|_2 \leq \lambda_{\max}(M)^{1/2}\beta\sqrt{m_1}\lambda_{\min}(M)^{-1/2} \overset{(B.2)}{\leq} \frac{1}{2} \ .$$

Hence, $F_1W_1 = I - F_2W_2$ is invertible with

$$(F_1W_1)^{-1} = \sum_{k=0}^{\infty}(F_2W_2)^k, \qquad \|(F_1W_1)^{-1}\|_2 \leq \sum_{k=0}^{\infty}\|F_2W_2\|_2^k \leq 2 \ .$$

Since F_1W_1 has full rank, W_1 and F_1 must also have full rank. Hence, $(F_1W_1)^{-1} = W_1^{-1}F_1^{-1}$ and

$$\|W_1^{-1}\|_2 \leq \|(F_1W_1)^{-1}\|_2\|F_1\|_2 \leq 2\lambda_{\max}(M)^{1/2} \ .$$

(5) *Bound the sum for a "similar" initial vector:*

Note that for $\tilde{v}_2 \coloneqq W_2 W_1^{-1} v_1$, we have

$$W W_1^{-1} v_1 = \begin{pmatrix} I \\ W_2 W_1^{-1} \end{pmatrix} v_1 = \begin{pmatrix} v_1 \\ \tilde{v}_2 \end{pmatrix} \tag{B.8}$$

and $\tilde{v}_2$ is "small":

$$\|\tilde{v}_2\|_2 \leq \|W_2\|_2 \|W_1^{-1}\|_2 \|v_1\|_2 \leq \beta \sqrt{m_1} \lambda_{\min}(M)^{-1/2} \cdot 2\lambda_{\max}(M)^{1/2} \cdot \|v_1\|_2 \ .$$

By Eq. (B.7), we have $AMW = WD_1$, where D_1 is the upper left $m_1 \times m_1$ block of D. Therefore,

$$\begin{aligned} h \sum_{k=0}^{\infty} \|(I_m - hAM)^k W W_1^{-1} v_1\|_2 &= h \sum_{k=0}^{\infty} \|W (I_{m_1} - hD_1)^k W_1^{-1} v_1\|_2 \\ &\leq \|W\|_2 \|W_1^{-1}\|_2 \|v_1\|_2 \cdot h \sum_{k=0}^{\infty} \|(I_{m_1} - hD_1)^k\|_2 \ , \end{aligned}$$

where

$$\|W\|_2 \|W_1^{-1}\|_2 \|v_1\|_2 \leq 2\lambda_{\max}(M)^{1/2} \lambda_{\min}(M)^{-1/2} \|v_1\|_2 = 2\sqrt{\operatorname{cond}(M)} \|v_1\|_2$$

and we can compute the remaining sum similarly as in step (1):

$$\begin{aligned} h \sum_{k=0}^{\infty} \|(I - hD_1)^k\|_2 &= h \sum_{k=0}^{\infty} (1 - h\lambda_{m_1}(AM))^k \\ &\overset{\text{(B.5)}}{\leq} h \sum_{k=0}^{\infty} (1 - h\lambda)^k = \frac{h}{1 - (1 - h\lambda)} = \frac{1}{\lambda} \ . \end{aligned}$$

(6) *Bound the original sum:*

Using $v = W W_1^{-1} v_1 + \begin{pmatrix} 0 \\ v_2 - \tilde{v}_2 \end{pmatrix}$, we obtain

$$\begin{aligned} & h \sum_{k=0}^{\infty} \|(I - hAM)^k v\|_2 \\ \overset{\text{(B.8)}}{\leq} \ & h \sum_{k=0}^{\infty} \|(I - hAM)^k W W_1^{-1} v_1\|_2 + h \sum_{k=0}^{\infty} \left\| (I - hAM)^k \begin{pmatrix} 0 \\ v_2 - \tilde{v}_2 \end{pmatrix} \right\|_2 \\ \leq \ & h \sum_{k=0}^{\infty} \|(I - hAM)^k W W_1^{-1} v_1\|_2 + h \sum_{k=0}^{\infty} \|(I - hAM)^k\|_2 \cdot (\|v_2\|_2 + \|\tilde{v}_2\|_2) \\ \overset{\text{(5), (B.4)}}{\leq} \ & 2 \frac{\sqrt{\operatorname{cond}(M)}}{\lambda} \|v_1\|_2 + 2 \frac{\sqrt{\operatorname{cond}(M)}}{\lambda_{\min}(AM)} \left(\|v_2\|_2 + 2\beta \sqrt{m_1} \sqrt{\operatorname{cond}(M)} \|v_1\|_2 \right) \\ = \ & 2\sqrt{\operatorname{cond}(M)} \left(\left(\frac{1}{\lambda} + \frac{2\sqrt{m_1} \sqrt{\operatorname{cond}(M)} \beta}{\lambda_{\min}(AM)} \right) \|v_1\|_2 + \frac{1}{\lambda_{\min}(AM)} \|v_2\|_2 \right) \ . \quad \square \end{aligned}$$

The bound from Proposition B.2 can be transferred to the original system via the next lemma:

Lemma B.3. *Let $\|\cdot\|$ denote an arbitrary vector norm on $\mathbb{R}^m$ and its induced matrix norm. Let $k \in \mathbb{N}_0$, $K_0, \dots, K_{k-1} \in \mathbb{R}^{m\times m}$ and $\tilde{K} \in \mathbb{R}^{m\times m}$. If*

$$\delta_{k-1} \coloneqq \sum_{l=0}^{k-1} \|\tilde{K}^l\| \cdot \sup_{l\in\{0,\dots,k-1\}} \|K_l - \tilde{K}\| < 1 \ ,$$

where $\delta_{-1} \coloneqq 0$, then each sequence $v_0, \dots, v_k$ with $v_{l+1} = K_l v_l$ for all $l \in \{0, \dots, k-1\}$ satisfies

$$\sum_{l=0}^{k} \|v_l\| \le \frac{1}{1-\delta_{k-1}} \sum_{l=0}^{k} \|\tilde{K}^l v_0\| \ .$$

Proof. Clearly, for $l \in \{0, \dots, k-1\}$,

$$v_{l+1} = \tilde{K} v_l + (K_l - \tilde{K}) v_l$$

and hence, by induction on l,

$$v_l = \tilde{K}^l v_0 + \sum_{l'=0}^{l-1} \tilde{K}^{l-1-l'} (K_{l'} - \tilde{K}) v_{l'}$$

for all $l \in \{0, \dots, k\}$. Summing norms on both sides yields

$$\begin{aligned}
\sum_{l=0}^{k} \|v_l\| &\le \sum_{l=0}^{k} \|\tilde{K}^l v_0\| + \sum_{l=0}^{k} \sum_{l'=0}^{l-1} \|\tilde{K}^{l-1-l'} (K_{l'} - \tilde{K}) v_{l'}\| \\
&= \sum_{l=0}^{k} \|\tilde{K}^l v_0\| + \sum_{l'=0}^{k-1} \sum_{l=l'+1}^{k} \|\tilde{K}^{l-1-l'} (K_{l'} - \tilde{K}) v_{l'}\| \\
&\le \sum_{l=0}^{k} \|\tilde{K}^l v_0\| + \sum_{l'=0}^{k-1} \left(\sum_{l=0}^{k-1-l'} \|\tilde{K}^l\| \right) \cdot \sup_{l\in\{0,\dots,k-1\}} \|K_l - \tilde{K}\| \cdot \|v_{l'}\| \\
&\le \sum_{l=0}^{k} \|\tilde{K}^l v_0\| + \delta_{k-1} \sum_{l'=0}^{k} \|v_{l'}\| \ .
\end{aligned}$$

Hence $(1-\delta_{k-1}) \sum_{l=0}^{k} \|v_l\| \le \sum_{l=0}^{k} \|\tilde{K}^l v_0\|$ and since $\delta_{k-1} < 1$, the inequality is preserved when dividing by $1 - \delta_{k-1}$. □

C Interpolation with LeakyReLU Networks

The following lemma shows that N data points can be interpolated using $N-1$ hidden neurons:

Lemma C.1. *Let $\alpha \neq 1$ and let $D = ((x_1, y_1), \ldots, (x_N, y_N)) \in (\mathbb{R} \times \mathbb{R})^N$ with $x_j \neq x_{j'}$ for $j \neq j'$. Then, for $n \geq N-1$, there exists $W = (a, b, c, w) \in \mathbb{R}^{3n+1}$ such that*

$$f_W(x_j) = y_j$$

for all $j \in \{1, \ldots, N\}$.

Proof. Without loss of generality, assume $x_1 < \ldots < x_N$. Using $N-1$ hidden neurons, we can represent a function $\tilde{f}_w$ of the form

$$\tilde{f}_w(x) \coloneqq \sum_{i=1}^{N-1} w_i \varphi(1 \cdot x + (-x_i)) \ .$$

Then, $\tilde{f}_w$ is continuous and for $x \in (x_j, x_{j+1})$, we have

$$\tilde{f}_w'(x) = \sum_{i=1}^{j} w_i + \sum_{i=j+1}^{N-1} \alpha w_i \ ,$$

i.e.

$$\begin{pmatrix} \tilde{f}_w'|_{(x_1,x_2)} \\ \vdots \\ \tilde{f}_w'|_{(x_{N-1},x_N)} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & \alpha & \ldots & \alpha \\ 1 & 1 & & \vdots \\ \vdots & & \ddots & \alpha \\ 1 & 1 & \ldots & 1 \end{pmatrix}}_{=:\tilde{M} \in \mathbb{R}^{(N-1)\times(N-1)}} \begin{pmatrix} w_1 \\ \vdots \\ w_{N-1} \end{pmatrix} \ .$$

By subtracting α times the first column of $\tilde{M}$ of the other columns of $\tilde{M}$, we see that $\tilde{M}$ can be transformed into a triangular matrix with nonzero diagonal entries. Hence, $\tilde{M}$ is invertible. Choose w such that

$$\tilde{f}_w'|_{(x_j,x_{j+1})} = \frac{y_{j+1} - y_j}{x_{j+1} - x_j} \ .$$

Then, the function

$$f(x) = \underbrace{(y_1 - \tilde{f}_w(x_1))}_{=:c} + \tilde{f}_w(x) + \sum_{i=N}^{n} 0 \cdot \varphi(0 \cdot x + 0)$$

corresponds to an interpolating neural net with n hidden neurons. □

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Datum und Unterschrift:

Declaration

I hereby declare that the work presented in this thesis is entirely my own. I did not use any other sources and references that the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

Date and Signature: