Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

On the Existence of Global Minima and Convergence Analyses for Gradient Descent Methods in the Training of Deep Neural Networks

Arnulf Jentzen * [1,2] and Adrian Riekert † [2]

[1]School of Data Science and Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China.
[2]Applied Mathematics: Institute for Analysis and Numerics, University of Münster, Germany.

Abstract. Although gradient descent (GD) optimization methods in combination with rectified linear unit (ReLU) artificial neural networks (ANNs) often supply an impressive performance in real world learning problems, till this day it remains – in all practically relevant scenarios – an open problem of research to rigorously prove (or disprove) the conjecture that such GD optimization methods do converge in the training of ANNs with ReLU activation.

In this article we study fully-connected feedforward deep ReLU ANNs with an arbitrarily large number of hidden layers and we prove convergence of the risk of the GD optimization method with random initializations in the training of such ANNs under the assumption that the unnormalized probability density function $\mathfrak{p}\colon [a,b]^d \to [0,\infty)$ of the probability distribution of the input data of the considered supervised learning problem is piecewise polynomial, under the assumption that the target function $f\colon [a,b]^d \to \mathbb{R}^\delta$ (describing the relationship between input data and the output data) is piecewise polynomial, and under the assumption that the risk function of the considered supervised learning problem admits at least one regular global minimum. In addition, in the special situation of shallow ANNs with just one hidden layer and one-dimensional input we also verify this assumption by proving in the training of such shallow ANNs that for every Lipschitz continuous target function there exists a global minimum in the risk landscape. Finally, in the training of deep ANNs with ReLU activation we also study solutions of gradient flow (GF) differential equations and we prove by proving that every non-divergent GF trajectory converges with a polynomial rate of convergence to a critical point (in the sense of limiting Fréchet subdifferentiability).

Our mathematical convergence analysis builds up on ideas from our previous article [S. Eberle, A. Jentzen, A. Riekert, & G. Weiss, Existence, uniqueness, and convergence rates for gradient flows in the training of artificial neural networks with ReLU activation. *arXiv:2108.08106* (2021)], on tools from real algebraic geometry such as the concept of semi-algebraic functions and generalized Kurdyka-Łojasiewicz inequalities, on tools from functional analysis such as the Arzelà–Ascoli theorem on the relative compactness of uniformly bounded and equicontinuous sequences of continuous functions, on tools from nonsmooth analysis such as the concept of limiting Fréchet subgradients, as well as on the fact that the set of realization functions of shallow ReLU ANNs with fixed architecture forms a closed subset of the set of continuous functions revealed in [P. Petersen, M. Raslan, & F. Voigtlaender, Topological properties of the set of functions generated by neural networks of fixed size. *Found. Comput. Math.* **21** (2021), no. 2, 375–444].

Keywords:
Deep learning,
Global minimum,
Kurdyka-Łojasiewicz inequality,
Gradient descent,
Optimization.

Article Info.:
Volume: 1
Number: 2
Pages: 141- 246
Date: June/2022
doi.org/10.4208/jml.220114a

Article History:
Received: 14/1/2022
Accepted: 23/6/2022

Communicated by:
Weinan E

*Corresponding author. ajentzen@cuhk.edu.cn, ajentzen@uni-muenster.de.
†ariekert@uni-muenster.de.

1 Introduction and main results

Even though gradient descent (GD) type optimization methods in combination with artificial neural networks (ANNs) often supply an impressive performance in real world learning problems, till this day it remains – in all practically relevant scenarios – an open problem of research to rigorously prove (or disprove) the conjecture that such GD optimization methods do converge in the training of ANNs. Moreover, in the case of ANNs with the widely-used rectified linear unit (ReLU) activation function, this problem of research receives additional difficulty due to the lack of differentiability of the rectifier function $\mathbb{R} \ni x \mapsto \max\{x, 0\} \in \mathbb{R}$.

Although the convergence analysis for GD type optimization methods in the training of ANNs remains a fundamental open problem of research, there are several auspicious approaches in the scientific literature which provide interesting first steps in this area of research. To briefly introduce the reader to this topic of research, we now highlight/mention some of those findings in a short way and refer to the below mentioned references for further details.

In particular, we refer, for example, to [2, 17, 24, 25, 27, 37, 38, 72, 75] for convergence results for gradient flow (GF) and GD processes in the training of ANNs in the so-called overparametrized regime, where the number of ANN parameters highly exceeds the number of considered input-output training data pairs. As the number of neurons goes to infinity, the corresponding GF processes converge to a measure-valued process called Wasserstein gradient flow; cf., for instance, [12, 15, 16], [26, Section 5.1], and the references mentioned therein.

Regarding abstract results on the convergence of GF and GD processes we refer, for example, to [5, 39, 55, 56, 62] for the case of convex objective functions, we refer, for instance, to [1, 3, 4, 10, 21, 45, 48, 49, 52, 53, 57] for convergence results for GF and GD processes under Łojasiewicz type conditions, and we refer, for instance, to [7, 30, 50, 60] and the references mentioned therein for further results without convexity conditions. In general, without global assumptions on the objective function such as convexity, gradient-based methods may converge to non-global local minima or saddle points. It therefore becomes important to analyze critical points of the objective function in the training of ANNs and we refer, for example, to [14,65,68,73,74] for articles which study the appearance of critical points in the risk landscape in the training of ANNs. The question under which conditions gradient-based optimization algorithms cannot converge to saddle points was investigated, for example, in [32, 48, 49, 58, 59]. For more detailed overviews and additional references on GD optimization schemes we mention, for instance, Bottou et al. [11], Fehrman et al. [30, Section 1.1], [39, Section 1], and Ruder [64].

In this article we study the training of fully-connected feedforward ANNs with ReLU activation by means of GD type optimization methods (we also refer to Figure 1.1 and Figure 1.2 in this introductory section below for graphical illustrations of two example architectures for the ANNs investigated in this work). In particular, one of the key contributions of this work is rigorously verify, under the assumption that the unnormalized probability density function $\mathfrak{p} \colon [a, b]^d \to [0, \infty)$ of the probability distribution of the input data of the considered supervised learning problem is piecewise polynomial (see Definition 5.1 in Section 5 for our precise meaning of a piecewise polynomial function), and the assumption that the target function $f = (f_1, \dots, f_\delta) \colon [a, b]^d \to \mathbb{R}^\delta$ (the function describing the relationship between the input data and the output data which one intends to learn approximately) is piecewise polynomial, it holds in the training of *deep ReLU ANNs with an arbitrarily large number of hidden layers* that the risk function (the function which is to be minimized) and its associated generalized gradient function satisfy at *every point* of the ANN parameter space a *generalized Kurdyka-Łojasiewicz inequality*; see Proposition 6.2 in Subsection 6.4 for the precise statement. In the previous sentence the quantity $d \in \mathbb{N} = \{1, 2, 3, \dots\}$ is an arbitrarily large natural number which describes the dimension of the input data, the quantity $\delta \in \mathbb{N}$ is a natural number which describes the dimension of the output data, and the quantities $a, b \in \mathbb{R}$ with $a < b$ are real numbers which border the region $[a, b]^d$ in which the input data takes values in. Proposition 6.2 in Subsection 6.4 in this work generalizes Proposition 5.1 in our previous article Eberle et al. [28] where such generalized Kurdyka-Łojasiewicz inequalities have been established in training of ReLU ANNs with one hidden layer. The proof of Proposition 6.2 relies on the fact that the considered risk function is semi-algebraic, which we establish in Corollary 5.1 below, and the abstract Kurdyka-Łojasiewicz inequality in Bolte et al. [10, Theorem 3.1]. A similar result regarding semi-algebraicity of the empirical risk, which is measured with respect to a finite set of input-output data pairs, is already known, cf. Davis et al. [20, Corollary 5.11].

We then use the established generalized Kurdyka-Łojasiewicz inequalities in Proposition 6.2 to prove convergence of GD type optimization methods in the training of deep ReLU ANNs where we first focus on *time-continuous* GD optimization methods (see Section 7) and, thereafter, investigate *time-discrete* GD optimization methods (see Section 8).

Specifically, in the time-continuous situation (see Section 7 and Subsection 1.3 in this introductory section) we establish in the training of deep ReLU ANNs, under the assumption that the unnormalized probability density function $\mathfrak{p} \colon [a, b]^d \to [0, \infty)$ and the target function $f \colon [a, b]^d \to \mathbb{R}$ are both piecewise polynomial, that every non-divergent solution of the associated *gradient flow (GD) differential equation* converges with a strictly positive rate of convergence to a *generalized critical point* of the risk function (in the sense of the limiting Fréchet subdifferential; see Definition 3.1 in Subsection 3.6) and also that the risk of the GF solution converges with rate 1 to the risk of the generalized critical point (see Theorem 7.1 in Subsection 7.3 below and Theorem 1.3 in Subsection 1.3 in this introductory section below, respectively, for the precise statements). This generalizes the approach in Eberle et al. [28, Subsection 5.2] from shallow ReLU ANNs to deep ReLU ANNs.

Moreover, in the time-discrete situation (see Section 8 and Subsections 1.1 and 1.4 in this introductory section) we establish in the training of deep ReLU DNNs, under the assumption that $\mathfrak{p}$ and f are piecewise polynomial and that the risk function of the considered deep supervised learning problem admits at least one regular global minimum point, that the risk of the plain vanilla GD optimization method with random initializations converges in the training of deep ReLU ANNs to 0 as the number of GD steps increases to ∞, as the number of random initializations increases to ∞, as the step size of the GD method (the learning rate of the GD method) decreases to 0, and as the width of the ANNs increases to ∞; see Theorem 8.1 in Section 8 below and Theorem 1.4 in Subsection 1.4 in this introductory section below, respectively, for the precise statement.

Another key contribution of this work (see Section 2) is to prove in the special situation of shallow ReLU ANNs with just one hidden layer and one-dimensional input and output (corresponding to the case $d = \delta = 1$) that for every Lipschitz continuous target function $f\colon [a,b] \to \mathbb{R}$ we have that there exist global minimum points of the risk function; see Theorem 2.2 in Subsection 2.6 and Theorem 1.1 in Subsection 1.1 in this introductory section below, respectively, for the precise statement. In the case of shallow ANNs we thereby verify the above mentioned assumption that the risk function of the considered supervised learning problem admits at least one regular global minimum point; cf. Corollary 2.6 in Subsection 2.7 below and Theorem 1.2 in Subsection 1.2 in this introductory section below, respectively.

To elucidate the findings of this work more clearly, we now present 4 selected specific results (which have already been briefly outlined in the above introductory paragraphs) regarding the training of ReLU ANNs, Theorem 1.1 in Subsection 1.1, Theorem 1.2 in Subsection 1.2, Theorem 1.3 in Subsection 1.3, and Theorem 1.4 in Subsection 1.4, with all details in a self-contained fashion. Theorem 1.1 and Theorem 1.2 deal with shallow ReLU ANNs with just one hidden layer and one-dimensional input and output ($d = \delta = 1$) and Theorem 1.3 and Theorem 1.4 treat the situation of deep ReLU ANNs with an arbitrarily large number of hidden layers and multi-dimensional input and output ($d, \delta \in \mathbb{N}$).

1.1 Existence of global minima for shallow artificial neural networks (ANNs)

Maybe the most basic question that one can ask regarding the training of ANNs is the existence of global minimum points in the risk landscape. In particular, without the existence of a global minimum point, one can not hope for a GD type optimization method to converge to a global minimum point. Surprisingly, there is almost no result in the scientific literature which actually establishes the existence of global minimum points of risk functions in the training of ANNs and in our perspective this subject is a very important direct of future research.

Theorem 1.1 below proves in the training of shallow ANNs with ReLU activation that for every distribution $\mu\colon \mathcal{B}([a,b]) \to [0,\infty]$ of the input data of the considered supervised learning problem and every Lipschitz continuous target function $f\colon [a,b] \to \mathbb{R}$ it holds that there exists a global minimum point of the risk function. The natural number $\mathfrak{h} \in \mathbb{N}$ in Theorem 1.1 specifies the number of neurons on the hidden layer of the ANN (the dimensionality of the hidden layer of the ANN), the natural number $\mathfrak{d} \in \mathbb{N}$ in Theorem 1.1 specifies the overall number of real parameters used to described to the considered ANNs, and the set $\mathcal{B}([a,b])$ is the Borel sigma-algebra on the real interval $[a,b] \subseteq \mathbb{R}$.

In Theorem 1.1 we thus consider ANNs with 1 neuron on the input layer (with a 1-dimensional input layer), with $\mathfrak{h}$ neurons on the hidden layer (with $\mathfrak{h}$-dimensional hidden layer), and with 1 neuron on the output layer (with a 1-dimensional output layer). There are hence $\mathfrak{h}$ real weight parameters and $\mathfrak{h}$ real bias parameters to describe the affine linear transformation between the 1-dimensional input layer and the $\mathfrak{h}$-dimensional hidden layer and $\mathfrak{h}$ real weight parameters and 1 real bias parameter to describe the affine linear transformation between the $\mathfrak{h}$-dimensional hidden layer and the 1-dimensional output layer. The overall number $\mathfrak{d} \in \mathbb{N}$ of real ANN parameters in Theorem 1.1 therefore satisfies $\mathfrak{d} = (\mathfrak{h} + \mathfrak{h}) + (\mathfrak{h} + 1) = 3\mathfrak{h} + 1$. We also refer to Figure 1.1 for a graphical illustration of an example architecture for the ANNs considered in Theorem 1.1.

The function $\mathcal{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ in (1.1) in Theorem 1.1 is the risk function in the considered supervised learning problem and the finite measure $\mu \colon \mathcal{B}([a, b]) \to [0, \infty]$ is the unnormalized probability distribution of the input data of the considered supervised learning problem. In Theorem 1.1 we considered ReLU ANNs and the ReLU activation function $\mathbb{R} \ni x \mapsto \max\{x, 0\} \in \mathbb{R}$ appears on the right hand side of (1.1). In this set-up of shallow ReLU ANNs Theorem 1.1 reveals the existence of a global minimum point $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ of the risk function $\mathcal{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$.

Theorem 1.1. *Let $\mathfrak{h}, \mathfrak{d} \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in [a, \infty)$ satisfy $\mathfrak{d} = 3\mathfrak{h} + 1$, let $f \colon [a, b] \to \mathbb{R}$ be Lipschitz continuous, let $\mu \colon \mathcal{B}([a, b]) \to [0, \infty]$ be a finite measure, and let $\mathcal{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{L}(\theta) = \int_a^b \bigl(f(x) - \theta_{\mathfrak{d}} - \textstyle\sum_{j=1}^{\mathfrak{h}} \theta_{2\mathfrak{h}+j} \max\{\theta_{\mathfrak{h}+j} + \theta_j x, 0\}\bigr)^2 \,\mu(\mathrm{d}x). \tag{1.1}$$

Then there exists $\theta \in \mathbb{R}^{\mathfrak{d}}$ such that $\mathcal{L}(\theta) = \inf_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(\vartheta)$.

Theorem 1.1 is an immediate consequence of Theorem 2.2 in Subsection 2.6 below. Theorem 1.1 proves that there exists an ANN parameter vector $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ which satisfies that the risk $\mathcal{L}(\theta)$ of θ coincides with the infimum over all risk values $\inf_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(\vartheta)$.

As observed in Petersen et al. [61], the existence of global minima has direct implications for the training of ANNs. In particular, if there is no global minimum then any sequence $(\theta_n)_{n \in \mathbb{N}} \subseteq \mathbb{R}^{\mathfrak{d}}$ with $\lim_{n \to \infty} \mathcal{L}(\theta_n) = \inf_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(\vartheta)$ necessarily diverges to infinity. This behavior is highly undesirable in numerical computations. If the target function f is not continuous, this divergence phenomenon can indeed be observed in practice, as the results and numerical examples in [31] show. On the other hand, using our existence result for global minima we are able to establish convergence of GD with random initializations in the training of shallow ANNs if the assumptions of Theorem 1.1 are satisfied, see the next subsection for details.

In the scientific literature a similar existence result for ANNs with the Heaviside activation function $\mathbb{R} \ni x \mapsto \mathbb{1}_{[0,\infty)}(x) \in \mathbb{R}$ was established in Kainen et al. [43]. Moreover, we would like to point out that Theorem 1.1 does in general not hold without the Lipschitz continuity assumption on f. Indeed, Petersen et al. [61, Theorem 3.1] implies in the case where $\mathfrak{h} \geq 2$ and where the measure μ is non-atomic in the sense that its support is uncountable that the set of realization functions

$$\begin{aligned}&\bigl\{v \in C([0,1], \mathbb{R}) \colon \\ &\bigl(\exists\, \theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}} \colon \forall\, x \in [0,1] \colon v(x) = \theta_{\mathfrak{d}} + \textstyle\sum_{j=1}^{\mathfrak{h}} \theta_{2\mathfrak{h}+j} \max\{\theta_{\mathfrak{h}+j} + \theta_j x, 0\}\bigr)\bigr\}\end{aligned} \tag{1.2}$$

is not closed in the L^2-space $L^2([0,1], \mu)$. Specifically, Petersen et al. [61, Theorem 3.1] shows that there exists $f \in L^\infty([0,1], \mu)$ such that $\inf_{\theta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(\theta) = 0$ and $\{\theta \in \mathbb{R}^{\mathfrak{d}} \colon \mathcal{L}(\theta) = 0\} = \varnothing$. The function f constructed in Petersen et al. [61] is a step function of the form $f(x) = \mathbb{1}_{(x^*,1]}(x)$ for some suitable $x^* \in (0,1)$ depending on the measure μ and, thus, does not have a continuous representative. A similar non-closedness statement for the logistic

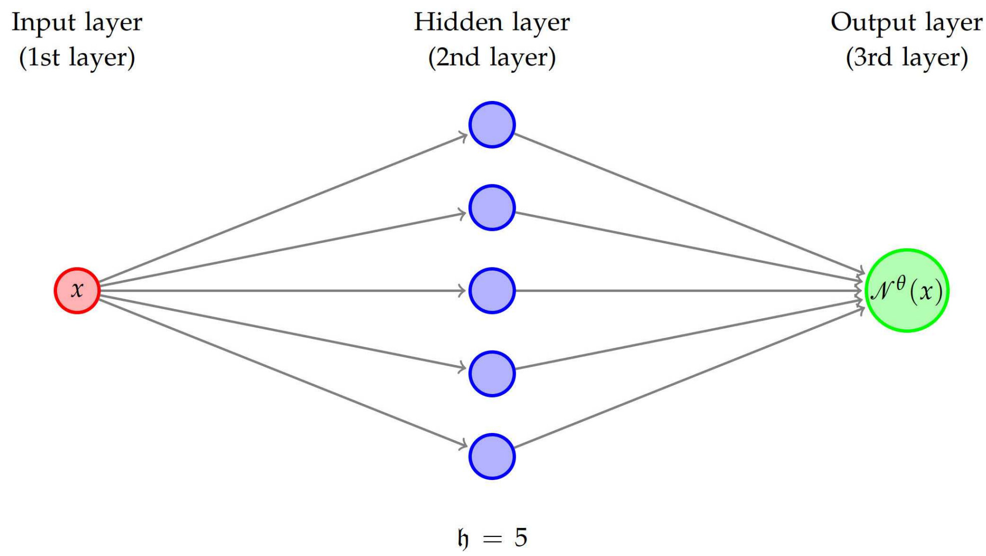

Figure 1.1: Graphical illustration of the considered shallow ANN architecture in Theorems 1.1 and 1.2 in the special case of an ANN with $\mathfrak{h} = 5$ neurons on the hidden layer. In this situation we have for every ANN parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}} = \mathbb{R}^{16}$ that the realization function $\mathbb{R} \ni x \mapsto \mathcal{N}^\theta(x) \in \mathbb{R}$ of the considered ANN maps the scalar input $x \in [a, b]$ to the scalar output $\mathcal{N}^\theta(x) = \theta_{\mathfrak{d}} + \sum_{j=1}^{\mathfrak{h}} \theta_{2\mathfrak{h}+j} \max\{\theta_j x + \theta_{\mathfrak{h}+j}, 0\} \in \mathbb{R}$.

activation function was proved earlier in Girosi & Poggio [33]. In Theorem 1.1 we assume that $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous. We guess that the statement remains true if one only assumes that f is continuous.

1.2 Gradient descent (GD) with random initializations in the training of shallow ANNs

In Theorem 1.2 below we employ Theorem 1.1 above to establish in the training of shallow ReLU ANNs (with 1 neuron on the input layer, $\mathfrak{h} \in \mathbb{N}$ neurons on the hidden layer, and 1 neuron on the output layer) that the risk of the plain vanilla GD optimization method with random initializations *converges in probability* to 0 as the number $n \in \mathbb{N}$ of GD steps increases to ∞, as the number $K \in \mathbb{N}$ of random initializations increases to ∞, as the step size $\gamma \in (0, \infty)$ of the GD optimization method (the learning rate of the GD optimization method) decreases to 0, and as the width $\mathfrak{h} \in \mathbb{N}$ of the considered ANNs increases to ∞; see (1.6) in Theorem 1.2 below for the precise statement.

In Theorem 1.2 we consider the GD optimization method with random initializations and the triple $(\Omega, \mathcal{F}, \mathbb{P})$ in Theorem 1.2 serves as the underlying probability space for the random initializations. Note that the function which maps random variables $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ to the real number

$$\mathbb{E}\big[\min\{|X - Y|, 1\}\big] \tag{1.3}$$

is nothing else but one commonly used metric which characterizes *convergence in probabil-*

ity (cf., e.g., Klenke [46, Theorem 6.7 in Chapter 6]) and (1.6) in Theorem 1.2 thus indeed establishes convergence in probability of the risk of the GD optimization method to 0.

Theorem 1.2. *Let $N \in \mathbb{N}$, ${}_0, {}_1, \ldots, {}_N, a, b \in \mathbb{R}$, satisfy $a = {}_0 < {}_1 < \cdots < {}_N = b$, let $f \in C([a,b], \mathbb{R})$, let $\mathfrak{p} \colon [a,b] \to [0,\infty)$ be a function, assume for all $n \in \{1, \ldots, N\}$ that $f|_{({}_{n-1}, {}_n)}$ and $\mathfrak{p}|_{({}_{n-1}, {}_n)}$ are polynomials, for every $\mathfrak{h} \in \mathbb{N}$ let $\mathcal{L}_\mathfrak{h} \colon \mathbb{R}^{3\mathfrak{h}+1} \to \mathbb{R}$ satisfy for all $\theta = (\theta_1, \ldots, \theta_{3\mathfrak{h}+1}) \in \mathbb{R}^{3\mathfrak{h}+1}$ that*

$$\mathcal{L}_\mathfrak{h}(\theta) = \textstyle\int_a^b \bigl(f(x) - \theta_{\mathfrak{d}} - \sum_{j=1}^{\mathfrak{h}} \theta_{2\mathfrak{h}+j} \max\{\theta_j x + \theta_{\mathfrak{h}+j}, 0\} \bigr)^2 \mathfrak{p}(x) \, \mathrm{d}x, \tag{1.4}$$

for every $\mathfrak{h} \in \mathbb{N}$ let $\mathcal{G}_\mathfrak{h} \colon \mathbb{R}^{3\mathfrak{h}+1} \to \mathbb{R}^{3\mathfrak{h}+1}$ satisfy for all $\theta \in \{\vartheta \in \mathbb{R}^{3\mathfrak{h}+1} \colon \mathcal{L}_\mathfrak{h}$ is differentiable at $\vartheta\}$ that $\mathcal{G}_\mathfrak{h}(\theta) = (\nabla \mathcal{L}_\mathfrak{h})(\theta)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $n, \mathfrak{h}, K \in \mathbb{N}_0$, $\gamma \in \mathbb{R}$ let $\Theta_{\mathfrak{h},n}^{K,\gamma} \colon \Omega \to \mathbb{R}^{3\mathfrak{h}+1}$ and $\mathbf{k}_{\mathfrak{h},n}^{K,\gamma} \colon \Omega \to \mathbb{N}$ be random variables, assume for all $\mathfrak{h} \in \mathbb{N}$, $\gamma \in \mathbb{R}$ that $\Theta_{\mathfrak{h},0}^{K,\gamma}$, $K \in \mathbb{N}$, are i.i.d., assume for all $\mathfrak{h} \in \mathbb{N}$, $\gamma, r \in (0,1)$, $\theta \in \mathbb{R}^{3\mathfrak{h}+1}$ that $\mathbb{P}(\|\Theta_{\mathfrak{h},0}^{1,\gamma} - \theta\| < r) > 0$, and assume for all $n, \mathfrak{h} \in \mathbb{N}_0$, $K \in \mathbb{N}$, $\gamma \in \mathbb{R}$, $\omega \in \Omega$ that

$$\Theta_{\mathfrak{h},n+1}^{K,\gamma}(\omega) = \Theta_{\mathfrak{h},n}^{K,\gamma}(\omega) - \gamma \mathcal{G}_\mathfrak{h}(\Theta_{\mathfrak{h},n}^{K,\gamma}(\omega)), \tag{1.5a}$$

$$\mathbf{k}_{\mathfrak{h},n}^{K,\gamma}(\omega) \in \operatorname{arg\,min}_{\kappa \in \{1,\ldots,K\}} \mathcal{L}_\mathfrak{h}(\Theta_{\mathfrak{h},n}^{\kappa,\gamma}(\omega)). \tag{1.5b}$$

Then

$$\limsup_{\mathfrak{h}\to\infty} \limsup_{\gamma \searrow 0} \limsup_{K\to\infty} \limsup_{n\to\infty} \mathbb{E}\bigl[\min\bigl\{\mathcal{L}_\mathfrak{h}(\Theta_n^{\mathfrak{h}, \mathbf{k}_n^{\mathfrak{h},K,\gamma}, \gamma}), 1\bigr\}\bigr] = 0. \tag{1.6}$$

Theorem 1.2 is a direct consequence of item (ii) in Corollary 8.6 in Subsection 8.7 below and the reversed version of Fatou's lemma. Note that in (1.5) above the random index $\mathbf{k}_{\mathfrak{h},n}^{K,\gamma}(\omega) \in \mathbb{N}$ selects the trajectory with the minimal risk after $n \in \mathbb{N}$ gradient steps among the first $K \in \mathbb{N}$ random initializations. Observe that (1.6) demonstrates that the risk $\mathcal{L}_\mathfrak{h}(\Theta_n^{\mathfrak{h}, \mathbf{k}_n^{\mathfrak{h},K,\gamma}, \gamma})$ of the GD optimization method with random initializations converges in probability (see (1.3) above) to 0 as the number n of GD steps increases to ∞, as the number K of random initializations increases to ∞, as the learning rate γ decreases to 0, and as the number $\mathfrak{h}$ of neurons on the hidden layer increases to ∞.

Roughly speaking, the proof of Theorem 1.2 consists of the following steps.

(I) We strengthen the existence result for global minima from Theorem 2.2 by proving in Corollary 2.6 that each of the risk functions $\mathcal{L}_\mathfrak{h}$, $\mathfrak{h} \in \mathbb{N}$, admits a global minimum around which suitable regularity conditions are satisfied.

(II) We establish in Corollary 5.1 that the considered risk functions are semi-algebraic. As a consequence, we show in Proposition 6.2 a generalized Kurdyka-Łojasiewicz inequality for the risk functions.

(III) In Corollary 8.4 below we show an abstract local convergence result to local minima for GD under a Kurdyka-Łojasiewicz type assumption and a suitable regularity assumption. Specifically, we assume that the considered local minimum admits a neighborhood on which the objective function is differentiable with a Lipschitz continuous gradient.

(IV) As a consequence, we obtain in Corollary 8.5 an abstract convergence result for GD processes with random initializations. Due to the first two steps, Corollary 8.5 is applicable to each risk function $\mathcal{L}_{\mathfrak{h}}$, $\mathfrak{h} \in \mathbb{N}$, under the assumptions of Theorem 1.2.

In [42, Theorem 1.1] a GD convergence result related to Theorem 1.2 above has been established. Roughly speaking, in [42, Theorem 1.1] a convergence result similar to (1.6) has been obtained in the situation where the learning rate $\gamma \in (0, \infty)$ must be sufficiently small but may be chosen to be independent of the number $\mathfrak{h} \in \mathbb{N}$ of neurons on the hidden layer, where the target function $f\colon [a, b] \to \mathbb{R}$ must not only be piecewise polynomial but even piecewise affine linear, and where the unnormalized probability density function $\mathfrak{p}\colon [a, b] \to [0, \infty)$ does not necessarily have to be piecewise polynomial but instead must be strictly positive and Lipschitz continuous.

The convergence analysis in [42] follows a completely different strategy than the convergence analysis in this work. In particular, in contrast to the proof of Theorem 1.2 above, the proof in [42, Theorem 1.1] does not at all use generalized Kurdyka-Łojasiewicz inequalities but instead is based on differential geometric arguments and analyses of the Hessian matrices of the risk function (cf. [30]).

1.3 Gradient flows (GFs) in the training of deep ANNs

In Theorem 1.3 below we demonstrate in the training of deep ReLU ANNs with an arbitrarily large number of hidden layers, under the assumption that the unnormalized probability density function $\mathfrak{p}\colon [a, b]^{\ell_0} \to [0, \infty)$ and the target function $f\colon [a, b]^{\ell_0} \to \mathbb{R}^{\ell_L}$ are piecewise polynomial (see (1.8) below for details), that every non-divergent solution Θ_t, $t \in [0, \infty)$, of the associated GF differential equation converges with a strictly positive rate of convergence to a *generalized critical point* ϑ (in the sense of the limiting Fréchet subdifferential; see Definition 3.1 in Subsection 3.6) and also that the risk $\mathcal{L}_\infty(\Theta_t)$, $t \in [0, \infty)$, of the GF solution converges with rate 1 to the risk $\mathcal{L}_\infty(\vartheta)$ of the generalized critical point; see (1.12) in Theorem 1.3 below for the precise statement.

The natural number $L \in \mathbb{N}$ in Theorem 1.3 specifies the number of affine linear transformations in the considered deep ANNs (the considered deep ANNs in Theorem 1.3 thus consist of $L - 1$ hidden layer and, including input and output layers, $L + 1$ layers overall) and the natural numbers $\ell_0, \ell_1, \ell_2, \ldots \in \mathbb{N}$ in Theorem 1.3 specify the number of neurons of the layers in the sense that there are ℓ_0 neurons on the input layer (the input layer is ℓ_0-dimensional), that for every $i \in \{1, \ldots, L - 1\}$ there are ℓ_i neurons on the ith hidden layer (the ith hidden layer is ℓ_i-dimensional), and that there are ℓ_L neurons on the output layer (the output layer is ℓ_L-dimensional). In the deep ANNs considered in Theorem 1.3, we thus have for every $k \in \{1, \ldots, L\}$ that there are $\ell_k \ell_{k-1}$ real weight parameters and ℓ_k real bias parameters to describe the affine linear transformation between the $(k - 1)$st and the k-th layer. The overall number $\mathfrak{d} \in \mathbb{N}$ of real ANN parameters in Theorem 1.3 thus satisfies

$$\mathfrak{d} = \textstyle\sum_{k=1}^{L} (\ell_k \ell_{k-1} + \ell_k) = \sum_{k=1}^{L} \ell_k (\ell_{k-1} + 1). \tag{1.7}$$

We also refer to Figure 1.2 for a graphical illustration of an example architecture for the ANNs considered in Theorem 1.3.

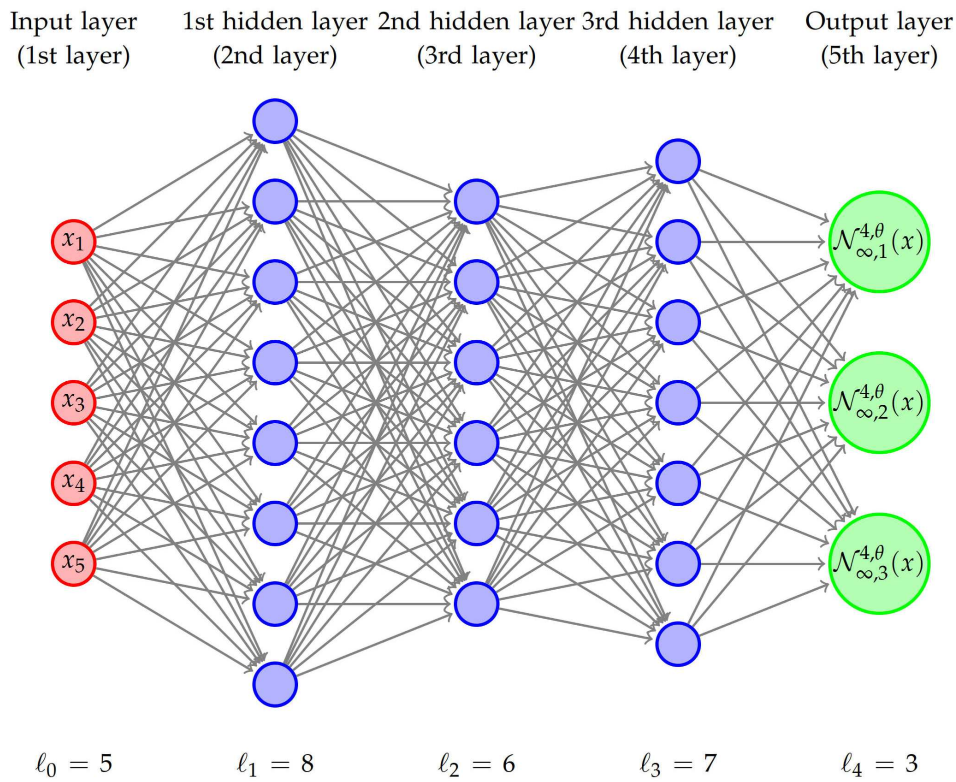

Figure 1.2: Graphical illustration of the considered deep ANN architecture in Theorem 1.3 in the special case of a deep ANN with 3 hidden layers (corresponding to $L = 4$ affine linear transformations), with 5 neurons on the input layer (corresponding to $\ell_0 = 5$), 8 neurons on the 1st hidden layer (corresponding to $\ell_1 = 8$), 6 neurons on the 2nd hidden layer (corresponding to $\ell_2 = 6$), 7 neurons on the 3rd hidden layer (corresponding to $\ell_3 = 7$), and 3 neurons on the output layer (corresponding to $\ell_4 = 3$). In this situation the dimension $\mathfrak{d}$ of the ANN parameter space satisfies $\mathfrak{d} = \sum_{i=1}^{4} \ell_i(\ell_{i-1} + 1) = 6 \cdot 8 + 9 \cdot 6 + 7 \cdot 7 + 3 \cdot 8 = 176$. Note that we have for every ANN parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}} = \mathbb{R}^{176}$ that the realization function $\mathbb{R}^5 \ni x \mapsto \mathcal{N}^{4,\theta}_{\infty}(x) \in \mathbb{R}^3$ of the considered deep ANN maps the 5-dimensional input vector $x = (x_1, x_2, x_3, x_4, x_5) \in [a, b]^5$ to the 3-dimensional output vector $\mathcal{N}^{4,\theta}_{\infty}(x) = (\mathcal{N}^{4,\theta}_{\infty,1}, \mathcal{N}^{4,\theta}_{\infty,2}, \mathcal{N}^{4,\theta}_{\infty,3}) \in \mathbb{R}^3$.

Because of the lack of differentiability of the ReLU activation function, the risk function $\mathcal{L}_\infty \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ in Theorem 1.3 is in general not continuously differentiable. In order to define an appropriately generalized gradient we approximate, as in [13, 36, 40, 41], the ReLU function through continuously differentiable functions $\mathfrak{R}_r \in C^1(\mathbb{R}, \mathbb{R})$, $r \in [1, \infty]$ (see (1.10) below for details). For every $\theta \in \mathbb{R}^{\mathfrak{d}}$, $r \in [1, \infty]$ we define the approximate realization function $\mathcal{N}^{L,\theta}_r \colon \mathbb{R}^{\ell_0} \to \mathbb{R}^{\ell_L}$ and the corresponding risk function $\mathcal{L}_r \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$. For every parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}}$ which satisfies that the approximate gradients $(\nabla \mathcal{L}_r)(\theta)$, $r \in [1, \infty)$, are convergent as $r \to \infty$ we define the generalized gradient $\mathcal{G}(\theta) \in \mathbb{R}^{\mathfrak{d}}$ as the limit $\lim_{r \to \infty} (\nabla \mathcal{L}_r)(\theta)$. In Proposition 3.1 below we verify that this limit, in fact, exists for every $\theta \in \mathbb{R}^{\mathfrak{d}}$, and thus the generalized gradient $\mathcal{G}(\theta)$ is uniquely defined for every $\theta \in \mathbb{R}^{\mathfrak{d}}$. Furthermore, we derive in Proposition 3.1 an explicit formula for the generalized gradient, which agrees with the standard implementation of the gradient via backpropagation.

Theorem 1.3. *Let* $L, \mathfrak{d}, \mathfrak{q} \in \mathbb{N}$, $(\ell_k)_{k\in\mathbb{N}_0} \subseteq \mathbb{N}$, $a \in \mathbb{R}$, $b \in [a,\infty)$ *satisfy* $\mathfrak{d} = \sum_{k=1}^{L} \ell_k(\ell_{k-1}+1)$, *for every* $i \in \{1,\dots,\mathfrak{q}\}$ *let* $\alpha_i \in \mathbb{R}^{\mathfrak{q}\times d}$, *let* $\beta_i \in \mathbb{R}^{\mathfrak{q}}$, *and let* $P_i\colon \mathbb{R}^{\ell_0} \to \mathbb{R}^{\ell_L+1}$ *be a polynomial, let* $f = (f_1,\dots,f_{\ell_L})\colon [a,b]^{\ell_0} \to \mathbb{R}^{\ell_L}$ *and* $\mathfrak{p}\colon [a,b]^{\ell_0} \to [0,\infty)$ *satisfy for all* $x \in [a,b]^{\ell_0}$ *that*

$$(f_1(x), f_2(x), \dots, f_{\ell_L}(x), \mathfrak{p}(x)) = \textstyle\sum_{i=1}^{\mathfrak{q}} P_i(x)\mathbb{1}_{[0,\infty)^{\mathfrak{q}}}(\alpha_i x + \beta_i), \tag{1.8}$$

for every $\theta = (\theta_1,\dots,\theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ *let* $\mathfrak{w}^{k,\theta} = (\mathfrak{w}^{k,\theta}_{i,j})_{(i,j)\in\{1,\dots,\ell_k\}\times\{1,\dots,\ell_{k-1}\}} \in \mathbb{R}^{\ell_k\times\ell_{k-1}}$, $k \in \mathbb{N}$, *and* $\mathfrak{b}^{k,\theta} = (\mathfrak{b}^{k,\theta}_1,\dots,\mathfrak{b}^{k,\theta}_{\ell_k}) \in \mathbb{R}^{\ell_k}$, $k \in \mathbb{N}$, *satisfy for all* $k \in \{1,\dots,L\}$, $i \in \{1,\dots,\ell_k\}$, $j \in \{1,\dots,\ell_{k-1}\}$ *that*

$$\mathfrak{w}^{k,\theta}_{i,j} = \theta_{(i-1)\ell_{k-1}+j+\sum_{h=1}^{k-1}\ell_h(\ell_{h-1}+1)} \qquad \text{and} \qquad \mathfrak{b}^{k,\theta}_i = \theta_{\ell_k\ell_{k-1}+i+\sum_{h=1}^{k-1}\ell_h(\ell_{h-1}+1)}, \tag{1.9}$$

let $\mathfrak{R}_r\colon \mathbb{R} \to \mathbb{R}$, $r \in [1,\infty]$, *satisfy for all* $r \in [1,\infty)$, $x \in (-\infty, 2^{-1}r^{-1}]$, $y \in \mathbb{R}$, $z \in [r^{-1},\infty)$ *that*

$$\mathfrak{R}_r \in C^1(\mathbb{R},\mathbb{R}), \quad \mathfrak{R}_r(x) = 0, \quad 0 \le \mathfrak{R}_r(y) \le \mathfrak{R}_\infty(y) = \max\{y,0\}, \quad \text{and} \quad \mathfrak{R}_r(z) = z, \tag{1.10}$$

assume $\sup_{r\in[1,\infty)}\sup_{x\in\mathbb{R}} |(\mathfrak{R}_r)'(x)| < \infty$, *for every* $r \in [1,\infty]$ *let* $\mathfrak{M}_r\colon (\cup_{n\in\mathbb{N}}\mathbb{R}^n) \to (\cup_{n\in\mathbb{N}}\mathbb{R}^n)$ *satisfy for all* $n \in \mathbb{N}$, $x = (x_1,\dots,x_n) \in \mathbb{R}^n$ *that* $\mathfrak{M}_r(x) = (\mathfrak{R}_r(x_1),\dots,\mathfrak{R}_r(x_n))$, *for every* $r \in [1,\infty]$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ *let* $\mathcal{N}^{k,\theta}_r\colon \mathbb{R}^{\ell_0} \to \mathbb{R}^{\ell_k}$, $k \in \mathbb{N}$, *satisfy for all* $k \in \mathbb{N}$, $x \in \mathbb{R}^{\ell_0}$ *that*

$$\mathcal{N}^{1,\theta}_r(x) = \mathfrak{b}^{1,\theta} + \mathfrak{w}^{1,\theta}x \qquad \text{and} \qquad \mathcal{N}^{k+1,\theta}_r(x) = \mathfrak{b}^{k+1,\theta} + \mathfrak{w}^{k+1,\theta}(\mathfrak{M}_{r^{1/k}}(\mathcal{N}^{k,\theta}_r(x))), \tag{1.11}$$

for every $r \in [1,\infty]$ *let* $\mathcal{L}_r\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ *satisfy for all* $\theta \in \mathbb{R}^{\mathfrak{d}}$ *that* $\mathcal{L}_r(\theta) = \int_{[a,b]^{\ell_0}} \|\mathcal{N}^{L,\theta}_r(x) - f(x)\|^2\,\mathfrak{p}(x)\,\mathrm{d}x$, *let* $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $\theta \in \{\vartheta \in \mathbb{R}^{\mathfrak{d}}\colon ((\nabla\mathcal{L}_r)(\vartheta))_{r\in[1,\infty)} \text{ is convergent}\}$ *that* $\mathcal{G}(\theta) = \lim_{r\to\infty}(\nabla\mathcal{L}_r)(\theta)$, *and*[1] *let* $\Theta \in C([0,\infty),\mathbb{R}^{\mathfrak{d}})$ *satisfy* $\liminf_{t\to\infty}\|\Theta_t\| < \infty$ *and* $\forall\, t \in [0,\infty)\colon \Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s$. *Then*[2] *there exist* $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathfrak{C}, \beta \in (0,\infty)$ *with* $0 \in (\mathbb{D}\mathcal{L}_\infty)(\vartheta)$ *such that for all* $t \in [0,\infty)$ *it holds that*

$$\|\Theta_t - \vartheta\| \le \mathfrak{C}(1+t)^{-\beta} \qquad \text{and} \qquad |\mathcal{L}_\infty(\Theta_t) - \mathcal{L}_\infty(\vartheta)| \le \mathfrak{C}(1+t)^{-1}. \tag{1.12}$$

Theorem 1.3 is an immediate consequence of Theorem 7.1 in Subsection 7.3 below. Note that the first inequality in (1.12) in Theorem 1.3 above assures that the standard norm $\|\Theta_t - \vartheta\|$ of the difference of the GF solution at time t and the generalized critical point ϑ converges with rate $\beta \in (0,\infty)$ to 0 and note that the second inequality in (1.12) in Theorem 1.3 above assures that the absolute value $|\mathcal{L}_\infty(\Theta_t) - \mathcal{L}_\infty(\vartheta)|$ of the difference of the risks of the GF solution at time t and the generalized critical point ϑ converges with rate 1 to 0.

[1] Throughout this article we denote by $\|\cdot\|\colon (\cup_{n\in\mathbb{N}}\mathbb{R}^n) \to \mathbb{R}$ and $\langle\cdot,\cdot\rangle\colon (\cup_{n\in\mathbb{N}}(\mathbb{R}^n\times\mathbb{R}^n)) \to \mathbb{R}$ the functions which satisfy for all $n \in \mathbb{N}$, $x = (x_1,\dots,x_n)$, $y = (y_1,\dots,y_n) \in \mathbb{R}^n$ that $\|x\| = (\sum_{i=1}^n |x_i|^2)^{1/2}$ and $\langle x,y\rangle = \sum_{i=1}^n x_i y_i$.

[2] In the conclusion of Theorem 1.3 we denote by $(\mathbb{D}\mathcal{L}_\infty)(\vartheta)$ the limiting Fréchet subdifferential of $\mathcal{L}_\infty\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ at $\vartheta \in \mathbb{R}^{\mathfrak{d}}$; see Definition 3.1 below for details.

In our proof of Theorem 1.3 we combine the generalized KL-inequality for the risk function in Proposition 6.2 with the abstract convergence results for GF processes in Section 7. The main regularity condition we need is the chain rule for the risk function $\mathcal{L}_\infty$, which was established in [36]. The fact that the limit $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ is a generalized critical point in the sense that 0 is an element of the limiting Fréchet subdifferential of $\mathcal{L}_\infty \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ at ϑ is a consequence of the fact that the generalized gradient we define is an element of the limiting Fréchet subdifferential, which we show in Proposition 3.3 below.

The assumption that the trajectory $(\Theta_t)_{t\in[0,\infty)}$ is bounded is necessary and is not implied by the other conditions. Indeed, in [31] we show that there are piecewise polynomial target functions for which GF trajectories with certain initialization do diverge to infinity.

In [28, Theorem 1.2] a GF convergence result related to Theorem 1.3 above has been obtained in the case of shallow ANNs with just one hidden layer. More specifically, in [28, Theorem 1.2] a GF convergence result similar to (1.12) has been established in the situation where the target function is additionally continuous and where the considered ANNs are not deep but shallow and just consist of 3 layers (input layer, output layer, and one hidden layer).

1.4 Gradient descent (GD) with random initializations in the training of deep ANNs

In Theorem 1.4 below we establish in the training of deep ReLU ANNs with an arbitrarily large number of hidden layers, under the assumption that the unnormalized probability density function $\mathfrak{p} \colon [a,b]^d \to [0,\infty)$ and the target function $f \colon [a,b]^d \to \mathbb{R}^\delta$ are piecewise polynomial (see (1.14) below for details) and that the risk function of the considered deep supervised learning problem admits at least one regular global minimum point, that the risk of the plain vanilla GD optimization method with random initializations converges in probability to 0 as the number of GD steps increases to ∞, as the number of random initializations increases to ∞, as the step size of the GD method (the learning rate of the GD method) decreases to 0, and as the width of the ANNs increases to ∞ (see (1.13) and (1.18) below for details).

Theorem 1.4. *Let $d, \delta, \mathfrak{q} \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in [a,\infty)$, $(\rho_{\mathfrak{a}})_{\mathfrak{a}\in\mathbb{N}} \subseteq (\mathbb{N} \cap (1,\infty))$, let $\ell^{\mathfrak{a}} = (\ell^{\mathfrak{a}}_0, \ell^{\mathfrak{a}}_1, \dots, \ell^{\mathfrak{a}}_{\rho_{\mathfrak{a}}}) \in \{d\} \times \mathbb{N}^{\rho_{\mathfrak{a}}-1} \times \{\delta\}$, $\mathfrak{a} \in \mathbb{N}$, satisfy*

$$\liminf_{\mathfrak{a}\to\infty} \min\{\ell^{\mathfrak{a}}_1, \ell^{\mathfrak{a}}_2, \dots, \ell^{\mathfrak{a}}_{\rho_{\mathfrak{a}}-1}\} = \infty, \tag{1.13}$$

for every $\mathfrak{a} \in \mathbb{N}$ let $\mathfrak{d}_{\mathfrak{a}} = \sum_{k=1}^{\rho_{\mathfrak{a}}} \ell^{\mathfrak{a}}_k(\ell^{\mathfrak{a}}_{k-1}+1)$, for every $i \in \{1,\dots,\mathfrak{q}\}$ let $\alpha_i \in \mathbb{R}^{\mathfrak{q}\times d}$, let $\beta_i \in \mathbb{R}^{\mathfrak{q}}$, and let $P_i \colon \mathbb{R}^d \to \mathbb{R}^{\delta+1}$ be a polynomial, let $f \colon [a,b]^d \to \mathbb{R}^\delta$ and $\mathfrak{p} \colon [a,b]^d \to [0,\infty)$ satisfy for all $x \in [a,b]^d$ that

$$(f_1(x), f_2(x), \dots, f_\delta(x), \mathfrak{p}(x)) = \textstyle\sum_{i=1}^{\mathfrak{q}} P_i(x) \mathbb{1}_{[0,\infty)^{\mathfrak{q}}}(\alpha_i x + \beta_i), \tag{1.14}$$

for every $\mathfrak{a} \in \mathbb{N}$, $k \in \{1,\dots,\rho_{\mathfrak{a}}\}$, $\theta = (\theta_1,\dots,\theta_{\mathfrak{d}_{\mathfrak{a}}}) \in \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}$ let $\mathfrak{w}^{k,\theta}_{\mathfrak{a}} = (\mathfrak{w}^{k,\theta}_{\mathfrak{a},i,j})_{(i,j)\in\{1,\dots,\ell^{\mathfrak{a}}_k\}\times\{1,\dots,\ell^{\mathfrak{a}}_{k-1}\}} \in \mathbb{R}^{\ell^{\mathfrak{a}}_k \times \ell^{\mathfrak{a}}_{k-1}}$ and $\mathfrak{b}^{k,\theta}_{\mathfrak{a}} = (\mathfrak{b}^{k,\theta}_{\mathfrak{a},1},\dots,\mathfrak{b}^{k,\theta}_{\mathfrak{a},\ell^{\mathfrak{a}}_k}) \in \mathbb{R}^{\ell^{\mathfrak{a}}_k}$ satisfy for all

$i \in \{1, \ldots, \ell_k^{\mathfrak{a}}\}$, $j \in \{1, \ldots, \ell_{k-1}^{\mathfrak{a}}\}$ *that*

$$\mathfrak{w}_{\mathfrak{a},i,j}^{k,\theta} = \theta_{(i-1)\ell_{k-1}^{\mathfrak{a}}+j+\sum_{h=1}^{k-1}\ell_h^{\mathfrak{a}}(\ell_{h-1}^{\mathfrak{a}}+1)} \qquad \text{and} \qquad \mathfrak{b}_{\mathfrak{a},i}^{k,\theta} = \theta_{\ell_k^{\mathfrak{a}}\ell_{k-1}^{\mathfrak{a}}+i+\sum_{h=1}^{k-1}\ell_h^{\mathfrak{a}}(\ell_{h-1}^{\mathfrak{a}}+1)}, \tag{1.15}$$

let $\mathfrak{M}\colon (\cup_{n\in\mathbb{N}}\mathbb{R}^n) \to (\cup_{n\in\mathbb{N}}\mathbb{R}^n)$ *satisfy for all* $n \in \mathbb{N}$, $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ *that* $\mathfrak{M}(x) = (\max\{x_1, 0\}, \ldots, \max\{x_n, 0\})$, *for every* $\mathfrak{a} \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}$ *let* $\mathcal{N}_{\mathfrak{a}}^{k,\theta}\colon \mathbb{R}^d \to \mathbb{R}^{\ell_k^{\mathfrak{a}}}$, $k \in \mathbb{N} \cap [1, \rho_{\mathfrak{a}}]$, *satisfy for all* $k \in \mathbb{N} \cap [1, \rho_{\mathfrak{a}})$, $x \in \mathbb{R}^d$ *that*

$$\mathcal{N}_{\mathfrak{a}}^{1,\theta}(x) = \mathfrak{b}_{\mathfrak{a}}^{1,\theta} + \mathfrak{w}_{\mathfrak{a}}^{1,\theta}x \qquad \text{and} \qquad \mathcal{N}_{\mathfrak{a}}^{k+1,\theta}(x) = \mathfrak{b}_{\mathfrak{a}}^{k+1,\theta} + \mathfrak{w}_{\mathfrak{a}}^{k+1,\theta}\big(\mathfrak{M}(\mathcal{N}_{\mathfrak{a}}^{k,\theta}(x))\big), \tag{1.16}$$

for every $\mathfrak{a} \in \mathbb{N}$ *let* $\mathcal{L}_{\mathfrak{a}}\colon \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}} \to \mathbb{R}$ *satisfy for all* $\theta \in \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}$ *that* $\mathcal{L}_{\mathfrak{a}}(\theta) = \int_{[a,b]^d} \|\mathcal{N}_{\mathfrak{a}}^{\rho_{\mathfrak{a}},\theta}(x) - f(x)\|^2 \,\mathfrak{p}(x)\,\mathrm{d}x$, *for every* $\mathfrak{a} \in \mathbb{N}$ *let* $\vartheta_{\mathfrak{a}} \in (\mathcal{L}_{\mathfrak{a}})^{-1}(\{\inf_{\theta\in\mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}} \mathcal{L}_{\mathfrak{a}}(\theta)\})$, $\varepsilon_{\mathfrak{a}} \in (0,1)$ *satisfy that* $\mathcal{L}_{\mathfrak{a}}|_{\{\theta\in\mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}\colon \|\theta-\vartheta_{\mathfrak{a}}\|<\varepsilon_{\mathfrak{a}}\}}$ *has a Lipschitz continuous derivative, for every* $\mathfrak{a} \in \mathbb{N}$ *let* $\mathcal{G}_{\mathfrak{a}}\colon \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}} \to \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}$ *satisfy for all* $\theta \in \{\vartheta \in \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}\colon \mathcal{L}_{\mathfrak{a}}$ *is differentiable at* $\vartheta\}$ *that* $\mathcal{G}_{\mathfrak{a}}(\theta) = (\nabla\mathcal{L}_{\mathfrak{a}})(\theta)$, *let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space, for every* $n, \mathfrak{a}, K \in \mathbb{N}_0$, $\gamma \in \mathbb{R}$ *let* $\Theta_{\mathfrak{a},n}^{K,\gamma}\colon \Omega \to \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}$ *and* $\mathbf{k}_{\mathfrak{a},n}^{K,\gamma}\colon \Omega \to \mathbb{N}$ *be random variables, assume for all* $\mathfrak{a} \in \mathbb{N}$, $\gamma \in \mathbb{R}$ *that* $\Theta_{\mathfrak{a},0}^{K,\gamma}$, $K \in \mathbb{N}$, *are i.i.d., assume for all* $\mathfrak{a} \in \mathbb{N}$, $\gamma, r \in (0,1)$, $\theta \in \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}$ *that* $\mathbb{P}(\|\Theta_{\mathfrak{a},0}^{1,\gamma} - \theta\| < r) > 0$, *and assume for all* $n \in \mathbb{N}_0$, $\mathfrak{a}, K \in \mathbb{N}$, $\gamma \in \mathbb{R}$, $\omega \in \Omega$ *that*

$$\Theta_{\mathfrak{a},n+1}^{K,\gamma}(\omega) = \Theta_{\mathfrak{a},n}^{K,\gamma}(\omega) - \gamma\mathcal{G}_{\mathfrak{a}}(\Theta_{\mathfrak{a},n}^{K,\gamma}(\omega)), \tag{1.17a}$$

$$\mathbf{k}_{\mathfrak{a},n}^{K,\gamma}(\omega) \in \arg\min\nolimits_{\kappa\in\{1,\ldots,K\}} \mathcal{L}_{\mathfrak{a}}(\Theta_{\mathfrak{a},n}^{\kappa,\gamma}(\omega)). \tag{1.17b}$$

Then

$$\limsup_{\mathfrak{a}\to\infty} \limsup_{\gamma\searrow 0} \limsup_{K\to\infty} \limsup_{n\to\infty} \mathbb{E}\big[\min\{\mathcal{L}_{\mathfrak{a}}\big(\Theta_{\mathfrak{a},n}^{\mathbf{k}_{\mathfrak{a},n}^{K,\gamma},\gamma}\big), 1\big] = 0. \tag{1.18}$$

Theorem 1.4 follows immediately from item (ii) in Theorem 8.1 in Subsection 8.6 below and the reversed version of Fatou's lemma. Observe that (1.18) above shows that the risk $\mathcal{L}_{\mathfrak{a}}(\Theta_{\mathfrak{a},n}^{\mathbf{k}_{\mathfrak{a},n}^{K,\gamma},\gamma})$ of the GD optimization method with random initializations converges in probability (see (1.3) above) to 0 as the number n of GD steps increases to ∞, as the number K of random initializations increases to ∞, as the learning rate γ decreases to 0, and as the width of the ANN increases to ∞ in the sense of (1.13) above.

The proof of Theorem 1.4 is mostly analogous to the proof of Theorem 1.2. The main difference is that in the general setting of deep ANNs the existence of global minima is not known. This is the reason why we assume in Theorem 1.4 for every $\mathfrak{a} \in \mathbb{N}$ that the parameter vector $\vartheta_{\mathfrak{a}} \in \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}$ is a global minimum of the risk function $\mathcal{L}_{\mathfrak{a}}\colon \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}} \to \mathbb{R}$. Additionally, we assume that for every $\mathfrak{a} \in \mathbb{N}$ there exists $\varepsilon_{\mathfrak{a}} \in (0,1)$ which satisfies that the restriction of $\mathcal{L}_{\mathfrak{a}}$ to the neighborhood $\{\theta \in \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}\colon \|\theta - \vartheta_{\mathfrak{a}}\| < \varepsilon_{\mathfrak{a}}\}$ is differentiable with a Lipschitz continuous derivative.

2 Existence of global minima for shallow ANNs

In this section we establish in Theorem 2.2 in Subsection 2.6 below in the case where the target function $f\colon [a,b] \to \mathbb{R}$ is Lipschitz continuous and where the considered ReLU

ANNs just consist of a one-dimensional input layer, a multi-dimensional hidden layer, and a one-dimensional output layer that there exists a global minimum point of the risk function. Theorem 1.1 in the introduction is a direct consequence of Theorem 2.2.

In Corollary 2.6 in Subsection 2.7 we strengthen Theorem 2.2 by showing that there also exists a global minimum point of the risk function such that the risk function is continuously differentiable on a neighborhood around the global minimum point. Our proof of Corollary 2.6 is based on an application of Theorem 2.2 as well as on applications of some basic regularity results from our earlier article Eberle et al. [28, Proposition 2.3 and Corollary 2.7].

Our proof of Theorem 2.2 can, roughly speaking, be divided into three parts.

(I) In Corollary 2.2 in Subsection 2.3 below we establish an explicit characterization for the functions $f\colon [0,1] \to \mathbb{R}$ which can be exactly represented by a shallow ReLU ANN with $\mathfrak{h} \in \mathbb{N}$ neurons on the hidden layer.

(II) Thereafter, we employ Corollary 2.2 to prove in Corollary 2.5 in Subsection 2.5 below in the case where the target function $f\colon [a,b] \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant $L \in \mathbb{R}$ and where the considered ReLU ANNs consist of a one-dimensional input layer, an $\mathfrak{h}$-dimensional hidden layer, and a one-dimensional output layer that, roughly speaking, for every ANN parameter vector $\theta \in \mathbb{R}^{3\mathfrak{h}+1}$ there exists an ANN parameter vector $\vartheta \in \mathbb{R}^{3\mathfrak{h}+1}$ whose realization function approximates f at least as well as the realization function of θ but is additionally also Lipschitz continuous with Lipschitz constant at most $\mathfrak{h}L$.

(III) Finally, we combine Corollary 2.5 with the Arzelà–Ascoli theorem and the fact that the set of realization functions of shallow ReLU ANNs with fixed architecture forms a closed subset of the set of continuous functions revealed in Petersen et al. [61, Theorem 3.8] to prove Theorem 2.2.

The question which functions $f\colon \mathbb{R}^d \to \mathbb{R}$ can be represented by a shallow ReLU ANN with a fixed number of neurons on the hidden layer has also been investigated in the article Dereich & Kassing [23] and in Theorem 3.2 in [23] a similar result as Corollary 2.2 has been established.

Our proofs of Corollary 2.2 and Corollary 2.5 also use the elementary results and notions regarding piecewise linear functions in Subsection 2.2 as well as the elementary Lemma 2.3 and Lemma 2.4, and only for completeness we include in this section also detailed proofs for these results.

2.1 Mathematical framework for shallow ANNs with ReLU activation

In Setting 2.1 we present our framework for shallow ANNs with ReLU activation which will be employed during the remainder of this section.

Setting 2.1. Let $\mathfrak{h}, \mathfrak{d} \in \mathbb{N}$, $L \in \mathbb{R}$, $f \in C([0,1],\mathbb{R})$ satisfy for all $x,y \in [0,1]$ that $|f(x) - f(y)| \leq L|x-y|$ and $\mathfrak{d} = 3\mathfrak{h}+1$, let $\mathfrak{w} = ((\mathfrak{w}^{\theta}_j)_{j\in\{1,\ldots,\mathfrak{h}\}})_{\theta\in\mathbb{R}^{\mathfrak{d}}}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{h}}$, $\mathfrak{b} = ((\mathfrak{b}^{\theta}_j)_{j\in\{1,\ldots,\mathfrak{h}\}})_{\theta\in\mathbb{R}^{\mathfrak{d}}}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{h}}$, $\mathfrak{v} = ((\mathfrak{v}^{\theta}_j)_{j\in\{1,\ldots,\mathfrak{h}\}})_{\theta\in\mathbb{R}^{\mathfrak{d}}}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{h}}$, $\mathfrak{c} = (\mathfrak{c}^{\theta})_{\theta\in\mathbb{R}^{\mathfrak{d}}}\colon \mathbb{R}^{\mathfrak{d}} \to$

$\mathbb{R}$, and $\mathfrak{q} = ((\mathfrak{q}_j^\theta)_{j\in\{1,\dots,\mathfrak{h}\}})\colon \mathbb{R}^\mathfrak{d} \to (-\infty,\infty]^\mathfrak{h}$ satisfy for all $\theta = (\theta_1,\dots,\theta_\mathfrak{d}) \in \mathbb{R}^\mathfrak{d}$, $j \in \{1,2,\dots,\mathfrak{h}\}$ that $\mathfrak{w}_j^\theta = \theta_j$, $\mathfrak{b}_j^\theta = \theta_{\mathfrak{h}+j}$, $\mathfrak{v}_j^\theta = \theta_{2\mathfrak{h}+j}$, $\mathfrak{c}^\theta = \theta_\mathfrak{d}$, and

$$\mathfrak{q}_j^\theta = \begin{cases} -\mathfrak{b}_j^\theta/\mathfrak{w}_j^\theta, & \mathfrak{w}_j^\theta \neq 0, \\ \infty, & \mathfrak{w}_j^\theta = 0, \end{cases} \tag{2.1}$$

let $\mu\colon \mathcal{B}([0,1]) \to [0,\infty]$ be a finite measure, and let $\mathscr{N} = (\mathscr{N}^\theta)_{\theta\in\mathbb{R}^\mathfrak{d}}\colon \mathbb{R}^\mathfrak{d} \to C([0,1],\mathbb{R})$ and $\mathcal{L}\colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$, $x \in [0,1]$ that

$$\mathscr{N}^\theta(x) = \mathfrak{c}^\theta + \textstyle\sum_{j=1}^{\mathfrak{h}} \mathfrak{v}_j^\theta \max\{\mathfrak{b}_j^\theta + \mathfrak{w}_j^\theta x, 0\} \tag{2.2}$$

and $\mathcal{L}(\theta) = \int_0^1 (\mathscr{N}^\theta(y) - f(y))^2 \, \mu(\mathrm{d}y)$.

2.2 Properties of the breakpoint function

Definition 2.1 (Breakpoint function)**.** *We denote by $Q\colon C([0,1],\mathbb{R}) \to [0,\infty]$ the function which satisfies for all $f \in C([0,1],\mathbb{R})$ that*

$$\begin{aligned} Q(f) = \min\Big(\{\infty\} \cup \{n \in \mathbb{N}_0\colon [\exists\, \mathfrak{A}_1,\mathfrak{A}_2,\dots,\mathfrak{A}_{n+1},\mathfrak{B}_1,\mathfrak{B}_2,\dots,\mathfrak{B}_{n+1},\mathfrak{q}_0,\mathfrak{q}_1,\dots,\mathfrak{q}_{n+1} \in \mathbb{R}\colon \\ ([0 = \mathfrak{q}_0 < \mathfrak{q}_1 < \dots < \mathfrak{q}_{n+1} = 1], \\ [\forall\, j \in \mathbb{N} \cap [1,n+1],\, x \in [\mathfrak{q}_{j-1},\mathfrak{q}_j]\colon f(x) = \mathfrak{A}_j x + \mathfrak{B}_j])]\}\Big). \end{aligned} \tag{2.3}$$

Definition 2.2 (Piecewise affine linear functions)**.** *We denote by $\mathscr{L} \subseteq C([0,1],\mathbb{R})$ the set given by*

$$\mathscr{L} = \{f \in C([0,1],\mathbb{R})\colon Q(f) < \infty\} \tag{2.4}$$

(cf. Definition 2.1).

Definition 2.3 (Slopes and axis intercepts for piecewise affine linear functions)**.** *Let $f \in \mathscr{L}$ (cf. Definition 2.2). Then we denote by $A_1(f), A_2(f), \dots, A_{Q(f)+1}(f)$, $B_1(f), B_2(f), \dots, B_{Q(f)+1}(f), q_0(f), q_1(f), \dots, q_{Q(f)+1}(f) \in \mathbb{R}$ the real numbers which satisfy $0 = q_0(f) < q_1(f) < \dots < q_{Q(f)+1}(f) = 1$ and*

$$\forall\, j \in \mathbb{N} \cap [1, Q(f)+1],\, x \in [q_{j-1}(f), q_j(f)]\colon f(x) = A_j(f)x + B_j(f) \tag{2.5}$$

(cf. Definition 2.1).

Proposition 2.1. *Let $f \in \mathscr{L}$, $i \in \{1,2,\dots,Q(f)\}$ (cf. Definitions 2.1 and 2.2). Then*

(i) it holds that $A_{i+1}(f) \neq A_i(f)$,

(ii) it holds that $B_{i+1}(f) = B_i(f) - (A_{i+1}(f) - A_i(f))q_i(f)$, and

(iii) it holds that $B_{i+1}(f) = B_1(f) - \sum_{j=1}^{i}(A_{j+1}(f) - A_j(f))q_j(f)$

(cf. Definitions 2.3 and 2.4).

Proof of Proposition 2.1. Observe that (2.3) ensures that $A_i(f) \neq A_{i+1}(f)$. Next note that the fact that for all $j \in \{1, 2, \ldots, Q(f) + 1\}$, $x \in [q_{j-1}(f), q_j(f)]$ it holds that $f(x) = A_j(f)x + B_j(f)$ proves that for all $j \in \{1, 2, \ldots, Q(f)\}$ we have that

$$A_j(f)q_j(f) + B_j(f) = A_{j+1}(f)q_j(f) + B_{j+1}(f). \tag{2.6}$$

Hence, we obtain for all $j \in \{1, 2, \ldots, Q(f)\}$ that $B_{j+1}(f) = B_j(f) - (A_{j+1}(f) - A_j(f))q_j(f)$. Induction hence establishes that for all $j \in \{1, 2, \ldots, Q(f)\}$ it holds that $B_{j+1}(f) = B_1(f) - \sum_{k=1}^{j}(A_{k+1}(f) - A_k(f))q_k(f)$. The proof of Proposition 2.1 is thus complete. □

Lemma 2.1 (Subadditivity of the breakpoint function). *Let $f, g \in C([0,1], \mathbb{R})$. Then*

$$Q(f+g) \leq Q(f) + Q(g) \tag{2.7}$$

(cf. Definition 2.1).

Proof of Lemma 2.1. Throughout this proof assume without loss of generality that

$$Q(f) + Q(g) < \infty. \tag{2.8}$$

Observe that (2.8) implies that there exist $N \in \mathbb{N}_0 \cap [0, Q(f) + Q(g)]$, $\mathfrak{q}_0, \mathfrak{q}_1, \ldots, \mathfrak{q}_{N+1} \in \mathbb{R}$ which satisfy

$$0 = \mathfrak{q}_0 < \mathfrak{q}_1 < \cdots < \mathfrak{q}_{N+1} = 1 \tag{2.9}$$

and

$$\{\mathfrak{q}_0, \mathfrak{q}_1, \ldots, \mathfrak{q}_{N+1}\} = \{q_0(f), q_1(f), \ldots, q_{Q(f)+1}(f)\} \cup \{q_0(g), q_1(g), \ldots, q_{Q(g)+1}(g)\} \tag{2.10}$$

(cf. Definition 2.3). Note that (2.9) and (2.10) ensure that for all $i \in \{0, 1, \ldots, N\}$ it holds that $(f+g)|_{[\mathfrak{q}_i, \mathfrak{q}_{i+1}]}$ is affine linear. Hence, we obtain that $Q(f+g) \leq N \leq Q(f) + Q(g)$. The proof of Lemma 2.1 is thus complete. □

Corollary 2.1. *Let $f, g \in \mathscr{L}$ (cf. Definition 2.2). Then*

(i) it holds that $Q(f+g) \leq Q(f) + Q(g)$ and

(ii) it holds that $f + g \in \mathscr{L}$

(cf. Definition 2.1).

Proof of Corollary 2.1. Observe that Lemma 2.1 and the assumption that $f, g \in \mathscr{L}$ assure that $Q(f+g) \leq Q(f) + Q(g) < \infty$. This completes the proof of Corollary 2.1. □

Definition 2.4 (Lipschitz constant). *We denote by* $\text{Lip}\colon C([0,1], \mathbb{R}) \to [0, \infty]$ *the function which satisfies for all $f \in C([0,1], \mathbb{R})$ that*

$$\text{Lip}(f) = \sup_{\substack{x,y \in [0,1], \\ x \neq y}} \left(\frac{|f(x) - f(y)|}{|x - y|} \right). \tag{2.11}$$

Lemma 2.2. *Let $f \in \mathscr{L}$ (cf. Definition 2.2). Then*

$$\operatorname{Lip}(f) = \max_{i \in \{1,2,\ldots,Q(f)+1\}} |A_i(f)| \tag{2.12}$$

(cf. Definitions 2.1, 2.3, and 2.4).

Proof of Lemma 2.2. Note that the fact that $0 = q_0(f) < q_1(f) < \cdots < q_{Q(f)+1}(f) = 1$ and the fact that for all $j \in \{1,2,\ldots,Q(f)+1\}$, $x \in [q_{j-1}(f), q_j(f)]$ it holds that $f(x) = A_j(f)x + B_j(f)$ ensure that

$$\sup_{x \in [a,b] \setminus \{q_0(f), q_1(f), \ldots, q_{Q(f)+1}(f)\}} |f'(x)| = \max_{i \in \{1,2,\ldots,Q(f)+1\}} |A_i(f)|. \tag{2.13}$$

This and the fundamental theorem of calculus establish (2.12). The proof of Lemma 2.2 is thus complete. □

2.3 Characterization results for realization functions of shallow ANNs

The objective of this subsection is to establish Corollary 2.2, which provides a complete characterization of all functions in $C([0,1],\mathbb{R})$ that can be represented by a shallow ANN with ReLU activation and $\mathfrak{h} \in \mathbb{N}$ hidden neurons. We first prove in Lemma 2.3 a simple necessary condition: All representable functions are piecewise linear with at most $\mathfrak{h} \in \mathbb{N}$ breakpoints.

Lemma 2.3. *Assume Setting 2.1 and let $\theta \in \mathbb{R}^{\mathfrak{d}}$. Then*

(i) it holds that $\mathscr{N}^\theta \in \mathscr{L}$ and

(ii) it holds that $Q(\mathscr{N}^\theta) \leq \mathfrak{h}$

(cf. Definitions 2.1 and 2.2).

Proof of Lemma 2.3. Throughout this proof let $g_j \in C([0,1],\mathbb{R})$, $j \in \{0,1,\ldots,\mathfrak{h}\}$, satisfy for all $j \in \{1,2,\ldots,\mathfrak{h}\}$, $x \in [0,1]$ that

$$g_j(x) = \mathfrak{v}_j^\theta \max\{\mathfrak{w}_j^\theta x + \mathfrak{b}_j^\theta, 0\} \qquad \text{and} \qquad g_0(x) = \mathfrak{c}^\theta. \tag{2.14}$$

Observe that (2.14) ensures for all $j \in \{1,2,\ldots,\mathfrak{h}\}$ that

$$g_j \in \mathscr{L} \qquad \text{and} \qquad Q(g_j) \in \{0,1\}. \tag{2.15}$$

Furthermore, note that (2.14) demonstrates that $g_0 \in \mathscr{L}$ and $Q(g_0) = 0$. Combining this, the fact that for all $x \in [0,1]$ it holds that $\mathscr{N}^\theta(x) = \sum_{j=0}^{\mathfrak{h}} g_j(x)$, Corollary 2.1, and induction establishes items (i) and (ii). The proof of Lemma 2.3 is thus complete. □

Moreover, every piecewise linear function with at most $\mathfrak{h} - 1$ breakpoints is representable, as we show in Lemma 2.4.

Lemma 2.4. *Assume Setting 2.1 and let $g \in \mathscr{L}$ satisfy $Q(g) \leq \mathfrak{h} - 1$ (cf. Definitions 2.1 and 2.2). Then there exists $\theta \in \mathbb{R}^{\mathfrak{d}}$ such that $\mathscr{N}^\theta = g$.*

Proof of Lemma 2.4. Throughout this proof let $\theta \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $j \in \{1, 2, \dots, \mathfrak{h}\}$ that

$$\mathfrak{w}_j^\theta = \begin{cases} 1 & : j \le Q(g)+1, \\ 0 & : j > Q(g)+1, \end{cases} \qquad \mathfrak{b}_j^\theta = \begin{cases} -q_{j-1}(g) & : j \le Q(g)+1, \\ 0 & : j > Q(g)+1, \end{cases} \tag{2.16}$$

$$\mathfrak{v}_j^\theta = \begin{cases} A_1(g) & : j = 1, \\ A_j(g) - A_{j-1}(g) & : 1 < j \le Q(g)+1, \\ 0 & : j > Q(g)+1, \end{cases} \tag{2.17}$$

and $\mathfrak{c}^\theta = B_1(g)$. Observe that (2.2), (2.16), and (2.17) ensure for all $x \in [0,1]$ that

$$\begin{aligned} \mathscr{N}^\theta(x) &= \mathfrak{c}^\theta + \sum_{j=1}^{\mathfrak{h}} \mathfrak{v}_j^\theta \max\{\mathfrak{w}_j^\theta x + \mathfrak{b}_j^\theta, 0\} \\ &= B_1(g) + \sum_{j=1}^{Q(g)+1} \mathfrak{v}_j^\theta \max\{\mathfrak{w}_j^\theta x + \mathfrak{b}_j^\theta, 0\} \\ &= B_1(g) + \sum_{j=1}^{Q(g)+1} \mathfrak{v}_j^\theta \max\{x - q_{j-1}(g), 0\} \\ &= B_1(g) + A_1(g)x + \sum_{j=2}^{Q(g)+1} (A_j(g) - A_{j-1}(g)) \max\{x - q_{j-1}(g), 0\} \\ &= B_1(g) + A_1(g)x + \sum_{j=1}^{Q(g)} (A_{j+1}(g) - A_j(g)) \max\{x - q_j(g), 0\}. \end{aligned} \tag{2.18}$$

Combining this with Proposition 2.1 establishes for all $i \in \{0, 1, \dots, Q(g)\}$, $x \in [q_i(g), q_{i+1}(g)]$ that

$$\begin{aligned} \mathscr{N}^\theta(x) &= B_1(g) + A_1(g)x + \sum_{j=1}^{i} (A_{j+1}(g) - A_j(g))(x - q_j(g)) \\ &= A_{i+1}(g)x + B_1(g) - \sum_{j=1}^{i} (A_{j+1}(g) - A_j(g)) q_j(g) \\ &= A_{i+1}(g)x + B_{i+1}(g) = g(x). \end{aligned} \tag{2.19}$$

The proof of Lemma 2.4 is thus complete. □

For piecewise linear functions with exactly $\mathfrak{h}$ breakpoints, the situation is more involved: They are only representable by a shallow ANN with $\mathfrak{h}$ hidden neurons if the slopes fulfill a certain linear relation; see (2.21) below for details. In Lemma 2.5 we establish that this condition is necessary for a function to be representable, and afterwards we show in Lemma 2.6 that it is also sufficient. Both proofs proceed by induction on the number of breakpoints.

Lemma 2.5. *For every $\mathfrak{h} \in \mathbb{N}_0$, $\theta = (\theta_1, \dots, \theta_{3\mathfrak{h}+1}) \in \mathbb{R}^{3\mathfrak{h}+1}$ let $\mathscr{N}^\theta \colon [0,1] \to \mathbb{R}$ satisfy for all $x \in [0,1]$ that*

$$\mathscr{N}^\theta(x) = \theta_{3\mathfrak{h}+1} + \sum_{j=1}^{\mathfrak{h}} \theta_{2\mathfrak{h}+j} \max\{\theta_{\mathfrak{h}+j} + \theta_j x, 0\}, \tag{2.20}$$

for every $\mathfrak{h} \in \mathbb{N}_0$ let $\mathbf{R}_\mathfrak{h} \subseteq C([0,1],\mathbb{R})$ satisfy $\mathbf{R}_\mathfrak{h} = \{f \in Q^{-1}(\{\mathfrak{h}\}) \colon [\exists\, \theta \in \mathbb{R}^{3\mathfrak{h}+1} \colon f = \mathscr{N}^\theta]\}$, and for every $\mathfrak{h} \in \mathbb{N}_0$ let $\mathbf{S}_\mathfrak{h} \subseteq C([0,1],\mathbb{R})$ satisfy

$$\begin{split}\mathbf{S}_\mathfrak{h} = \Big\{ f \in Q^{-1}(\{\mathfrak{h}\}) \colon \Big(\exists\, k \in \mathbb{N}, i_1, i_2, \dots, i_k \in \mathbb{N} \colon \\ \big[\big(\tfrac{k}{2} \notin \mathbb{N} \big), \big(i_1 < i_2 < \dots < i_k \le \mathfrak{h}+1 \big), \big(\textstyle\sum_{j=1}^k (-1)^j A_{i_j}(f) = 0 \big) \big] \Big) \Big\}\end{split} \tag{2.21}$$

(cf. Definitions 2.1 and 2.3). Then it holds for all $\mathfrak{h} \in \mathbb{N}_0$ that

$$\mathbf{R}_\mathfrak{h} \subseteq \mathbf{S}_\mathfrak{h}. \tag{2.22}$$

Proof of Lemma 2.5. Throughout this proof let $\operatorname{sgn} \colon \mathbb{R} \to \mathbb{R}$ satisfy for all $x \in (0,\infty)$, $k \in \{-1,0,1\}$ that $\operatorname{sgn}(kx) = k$ and for every $\mathfrak{h} \in \mathbb{N}$, $\theta = (\theta_1, \dots, \theta_{3\mathfrak{h}+1}) \in \mathbb{R}^{3\mathfrak{h}+1}$, $j \in \{1,2,\dots,\mathfrak{h}\}$ let $\mathfrak{q}_j^\theta \in (-\infty, \infty]$ satisfy

$$\mathfrak{q}_j^\theta = \begin{cases} -\frac{\theta_{\mathfrak{h}+j}}{\theta_j} & \colon \theta_j \neq 0, \\ \infty & \colon \theta_j = 0. \end{cases} \tag{2.23}$$

Observe that (2.23) ensures that for all $\mathfrak{h} \in \mathbb{N}$, $\theta \in \mathbb{R}^{3\mathfrak{h}+1}$ it holds that

$$Q(\mathscr{N}^\theta) \le \big| \big(\{\mathfrak{q}_1^\theta, \mathfrak{q}_2^\theta, \dots, \mathfrak{q}_\mathfrak{h}^\theta\} \cap \mathbb{R} \big) \big| \le \mathfrak{h}. \tag{2.24}$$

We prove (2.22) by induction on $\mathfrak{h} \in \mathbb{N}_0$. For the base case $\mathfrak{h} = 0$ observe that for all $\theta \in \mathbb{R}$, $x \in [0,1]$ it holds that $\mathscr{N}^\theta(x) = \theta$. Hence, we obtain that for all $\theta \in \mathbb{R}$ that $Q(\mathscr{N}^\theta) = 0$, $A_1(\mathscr{N}^\theta) = 0$, and $B_1(\mathscr{N}^\theta) = \theta$. Therefore, we obtain that $\mathbf{R}_0 = (\cup_{\theta \in \mathbb{R}} \{\mathscr{N}^\theta\}) \subseteq \mathbf{S}_0$. This establishes (2.22) in the base case $\mathfrak{h} = 0$. For the induction step let $\mathfrak{h} \in \mathbb{N}_0$ satisfy $\mathbf{R}_\mathfrak{h} \subseteq \mathbf{S}_\mathfrak{h}$ and let $F \in \mathbf{R}_{\mathfrak{h}+1}$. We intend to prove that $F \in \mathbf{S}_{\mathfrak{h}+1}$. Observe that the fact that $F \in \mathbf{R}_{\mathfrak{h}+1}$ ensures that there exists $\Xi \in \mathbb{R}^{3(\mathfrak{h}+1)+1} = \mathbb{R}^{3\mathfrak{h}+4}$ which satisfies $\mathscr{N}^\Xi = F$. Note that (2.24) and the fact that $Q(F) = \mathfrak{h}+1$ demonstrate that $\mathfrak{q}_1^\Xi, \mathfrak{q}_2^\Xi, \dots, \mathfrak{q}_{\mathfrak{h}+1}^\Xi \in \mathbb{R}$ and $|\{\mathfrak{q}_1^\Xi, \mathfrak{q}_2^\Xi, \dots, \mathfrak{q}_{\mathfrak{h}+1}^\Xi\}| = \mathfrak{h}+1$. This shows that there exists a bijective $p \colon \{1,2,\dots,\mathfrak{h}+1\} \to \{1,2,\dots,\mathfrak{h}+1\}$ which satisfies

$$-\infty < \mathfrak{q}_{p(1)}^\Xi < \mathfrak{q}_{p(2)}^\Xi < \dots < \mathfrak{q}_{p(\mathfrak{h}+1)}^\Xi < \infty. \tag{2.25}$$

In the following let $\Theta = (\Theta_1, \dots, \Theta_{3\mathfrak{h}+4})$ satisfy for all $j \in \{1,2,\dots,\mathfrak{h}+1\}$ that

$$\Theta_j = \Xi_{p(j)}, \quad \Theta_{\mathfrak{h}+1+j} = \Xi_{\mathfrak{h}+1+p(j)}, \quad \Theta_{2\mathfrak{h}+2+j} = \Xi_{2\mathfrak{h}+2+p(j)}, \quad \text{and} \quad \Theta_{3\mathfrak{h}+4} = \Xi_{3\mathfrak{h}+4}. \tag{2.26}$$

Observe that (2.25), (2.26), and the fact that $F = \mathscr{N}^{\Xi}$ ensure that

$$\mathscr{N}^{\Theta} = \mathscr{N}^{\Xi} = F \qquad \text{and} \qquad -\infty < \mathfrak{q}_1^{\Theta} < \mathfrak{q}_2^{\Theta} < \cdots < \mathfrak{q}_{\mathfrak{h}+1}^{\Theta} < \infty. \tag{2.27}$$

In the following let $\theta = (\theta_1, \ldots, \theta_{3\mathfrak{h}+1}) \in \mathbb{R}^{3\mathfrak{h}+1}$ satisfy

$$\theta = (\Theta_1, \ldots, \Theta_{\mathfrak{h}}, \Theta_{\mathfrak{h}+2}, \ldots, \Theta_{2\mathfrak{h}+1}, \Theta_{2\mathfrak{h}+3}, \ldots, \Theta_{3\mathfrak{h}+2}, \Theta_{3\mathfrak{h}+4}) \tag{2.28}$$

and let $f \in C([0,1], \mathbb{R})$ satisfy $f = \mathscr{N}^{\theta}$. Note that (2.28) ensures for all $x \in [0,1]$ that

$$\begin{aligned} f(x) &= \theta_{3\mathfrak{h}+1} + \sum_{j=1}^{\mathfrak{h}} \theta_{2\mathfrak{h}+j} \max\{\theta_{\mathfrak{h}+j} + \theta_j x, 0\} \\ &= F(x) - \Theta_{3\mathfrak{h}+3} \max\{\Theta_{2\mathfrak{h}+2} + \Theta_{\mathfrak{h}+1} x, 0\}. \end{aligned} \tag{2.29}$$

Next observe that (2.27) assures that $Q(f) = \mathfrak{h}$ and $-\infty < \mathfrak{q}_1^{\theta} = \mathfrak{q}_1^{\Theta} < \mathfrak{q}_2^{\theta} = \mathfrak{q}_2^{\Theta} < \cdots < \mathfrak{q}_{\mathfrak{h}}^{\theta} = \mathfrak{q}_{\mathfrak{h}}^{\Theta} < \mathfrak{q}_{\mathfrak{h}+1}^{\Theta} < \infty$. Combining this with the fact that $f = \mathscr{N}^{\theta}$ demonstrates that $f \in \mathbf{R}_{\mathfrak{h}}$. The induction hypothesis that $\mathbf{R}_{\mathfrak{h}} \subseteq \mathbf{S}_{\mathfrak{h}}$ therefore assures that $f \in \mathbf{S}_{\mathfrak{h}}$. This proves that there exist $k \in \mathbb{N}$, $i_1, i_2, \ldots, i_k \in \mathbb{N}$ which satisfy

$$\tfrac{k}{2} \notin \mathbb{N}, \qquad i_1 < i_2 < \cdots < i_k \leq \mathfrak{h} + 1, \qquad \text{and} \qquad \textstyle\sum_{j=1}^{k} (-1)^j A_{i_j}(f) = 0. \tag{2.30}$$

Next let $K \in \mathbb{N}$, $I_1, I_2, \ldots, I_K \in \mathbb{N}$ satisfy

$$\{I_1, I_2, \ldots, I_K\} = \begin{cases} \{i_1, i_2, \ldots, i_k\} & : \Theta_{\mathfrak{h}+1} > 0, \\ (\cup_{l=1}^{k-1} \{i_l\}) \cup \{\mathfrak{h}+2\} & : \Theta_{\mathfrak{h}+1} < 0 = \mathfrak{h} + 1 - i_k, \\ \{i_1, \ldots, i_k, \mathfrak{h}+1, \mathfrak{h}+2\} & : \Theta_{\mathfrak{h}+1} < 0 < \mathfrak{h} + 1 - i_k. \end{cases} \tag{2.31}$$

Note that (2.30) and (2.31) ensure that $\frac{K}{2} \notin \mathbb{N}$ and $I_1 < I_2 < \cdots < I_K \leq \mathfrak{h} + 2$. In order to prove that $F \in \mathbf{S}_{\mathfrak{h}+1}$, it is thus sufficient to verify that

$$\textstyle\sum_{j=1}^{K} (-1)^j A_{I_j}(F) = 0. \tag{2.32}$$

For this observe that (2.29) assures for all $x \in [0,1]$ that

$$\begin{aligned} F(x) &= f(x) + |\Theta_{\mathfrak{h}+1}| \Theta_{3\mathfrak{h}+3} \max\{|\Theta_{\mathfrak{h}+1}|^{-1} \Theta_{2\mathfrak{h}+2} + |\Theta_{\mathfrak{h}+1}|^{-1} \Theta_{\mathfrak{h}+1} x, 0\} \\ &= f(x) + \Theta_{\mathfrak{h}+1} \Theta_{3\mathfrak{h}+3} \operatorname{sgn}(\Theta_{\mathfrak{h}+1}) \max\{(x - \mathfrak{q}_{\mathfrak{h}+1}^{\Theta}) \operatorname{sgn}(\Theta_{\mathfrak{h}+1}), 0\}. \end{aligned} \tag{2.33}$$

In the following we distinguish between the case $\Theta_{\mathfrak{h}+1} > 0$, the case $\Theta_{\mathfrak{h}+1} < 0 = \mathfrak{h} + 1 - i_k$, and the case $\Theta_{\mathfrak{h}+1} < 0 < \mathfrak{h} + 1 - i_k$. We first prove (2.32) in the case

$$\Theta_{\mathfrak{h}+1} > 0. \tag{2.34}$$

Note that (2.33) and (2.34) demonstrate for all $x \in [0,1]$ that $F(x) = f(x) + \Theta_{\mathfrak{h}+1} \Theta_{3\mathfrak{h}+3} \max\{x - \mathfrak{q}_{\mathfrak{h}+1}^{\Theta}, 0\}$. Hence, we obtain for all $j \in \{1, 2, \ldots, \mathfrak{h}+1\}$ that $A_j(F) =$

$A_j(f)$. Combining this with (2.31) implies that $\sum_{j=1}^{K}(-1)^j A_{I_j}(F) = \sum_{j=1}^{k}(-1)^j A_{i_j}(f) = 0$. This establishes (2.32) in the case $\Theta_{\mathfrak{h}+1} > 0$. In the next step we prove (2.32) in the case

$$\Theta_{\mathfrak{h}+1} < 0 = \mathfrak{h} + 1 - i_k. \tag{2.35}$$

Observe that (2.33) and (2.35) show for all $x \in [0,1]$ that $F(x) = f(x) + \Theta_{\mathfrak{h}+1}\Theta_{3\mathfrak{h}+3}\min\{x - \mathfrak{q}^{\Theta}_{\mathfrak{h}+1}, 0\}$. Therefore, we obtain for all $j \in \{1, 2, \dots, \mathfrak{h}+1\}$ that $A_j(F) = A_j(f) + \Theta_{\mathfrak{h}+1}\Theta_{3\mathfrak{h}+3}$ and $A_{\mathfrak{h}+2}(F) = A_{\mathfrak{h}+1}(f)$. Combining this with (2.31), (2.35), and the fact that $\frac{k}{2} \notin \mathbb{N}$ shows that

$$\begin{aligned}
\textstyle\sum_{j=1}^{K}(-1)^j A_{I_j}(F) &= \Big[\textstyle\sum_{j=1}^{K-1}(-1)^j A_{I_j}(F)\Big] + (-1)^K A_{I_K}(F) \\
&= \Big[\textstyle\sum_{j=1}^{k-1}(-1)^j \Big(A_{i_j}(f) + \Theta_{\mathfrak{h}+1}\Theta_{3\mathfrak{h}+3}\Big)\Big] - A_{\mathfrak{h}+2}(F) \\
&= \Big[\textstyle\sum_{j=1}^{k-1}(-1)^j A_{i_j}(f)\Big] + \Big[\textstyle\sum_{j=1}^{k-1}(-1)^j \Theta_{\mathfrak{h}+1}\Theta_{3\mathfrak{h}+3}\Big] - A_{\mathfrak{h}+1}(f) \\
&= \Big[\textstyle\sum_{j=1}^{k-1}(-1)^j A_{i_j}(f)\Big] + \Big[\textstyle\sum_{j=1}^{k-1}(-1)^j\Big] \Theta_{\mathfrak{h}+1}\Theta_{3\mathfrak{h}+3} + (-1)^k A_{i_k}(f) \\
&= \textstyle\sum_{j=1}^{k}(-1)^j A_{i_j}(f) = 0.
\end{aligned} \tag{2.36}$$

This establishes (2.32) in the case $\Theta_{\mathfrak{h}+1} < 0 = \mathfrak{h} + 1 - i_k$. Next we prove (2.32) in the case

$$\Theta_{\mathfrak{h}+1} < 0 < \mathfrak{h} + 1 - i_k. \tag{2.37}$$

Note that (2.33) and (2.37) demonstrate for all $x \in [0,1]$ that $F(x) = f(x) + \Theta_{\mathfrak{h}+1}\Theta_{3\mathfrak{h}+3}\min\{x - \mathfrak{q}^{\Theta}_{\mathfrak{h}+1}, 0\}$. Hence, we obtain for all $j \in \{1, 2, \dots, \mathfrak{h}+1\}$ that $A_j(F) = A_j(f) + \Theta_{\mathfrak{h}+1}\Theta_{3\mathfrak{h}+3}$ and $A_{\mathfrak{h}+2}(F) = A_{\mathfrak{h}+1}(f)$. Combining this with (2.31), (2.37), and the fact that $\frac{k}{2} \notin \mathbb{N}$ shows that

$$\begin{aligned}
&\textstyle\sum_{j=1}^{K}(-1)^j A_{I_j}(F) \\
&= \Big[\textstyle\sum_{j=1}^{k}(-1)^j A_{i_j}(F)\Big] + (-1)^{k+1} A_{I_{k+1}}(F) + (-1)^{k+2} A_{I_{k+2}}(F) \\
&= \Big[\textstyle\sum_{j=1}^{k}(-1)^j \Big(A_{i_j}(f) + \Theta_{\mathfrak{h}+1}\Theta_{3\mathfrak{h}+3}\Big)\Big] + A_{I_{k+1}}(F) - A_{I_{k+2}}(F) \\
&= \Big[\textstyle\sum_{j=1}^{k}(-1)^j A_{i_j}(f)\Big] + \Big[\textstyle\sum_{j=1}^{k}(-1)^j \Theta_{\mathfrak{h}+1}\Theta_{3\mathfrak{h}+3}\Big] + A_{\mathfrak{h}+1}(F) - A_{\mathfrak{h}+2}(F) \\
&= \Big[\textstyle\sum_{j=1}^{k}(-1)^j\Big] \Theta_{\mathfrak{h}+1}\Theta_{3\mathfrak{h}+3} + A_{\mathfrak{h}+1}(f) + \Theta_{\mathfrak{h}+1}\Theta_{3\mathfrak{h}+3} - A_{\mathfrak{h}+1}(f) = 0.
\end{aligned} \tag{2.38}$$

This establishes (2.32) in the case $\Theta_{\mathfrak{h}+1} < 0 < \mathfrak{h} + 1 - i_k$. Observe that (2.32), the fact that $\frac{K}{2} \notin \mathbb{N}$, and the fact that $I_1 < I_2 < \cdots < I_K \leq \mathfrak{h} + 2$ prove that $F \in \mathbf{S}_{\mathfrak{h}+1}$. Induction thus establishes (2.22). The proof of Lemma 2.5 is thus complete. □

Lemma 2.6. *For every* $\mathfrak{h} \in \mathbb{N}_0$, $\theta = (\theta_1, \dots, \theta_{3\mathfrak{h}+1}) \in \mathbb{R}^{3\mathfrak{h}+1}$ *let* $\mathscr{N}^\theta \colon [0,1] \to \mathbb{R}$ *satisfy for all* $x \in [0,1]$ *that*

$$\mathscr{N}^\theta(x) = \theta_{3\mathfrak{h}+1} + \sum_{j=1}^{\mathfrak{h}} \theta_{2\mathfrak{h}+j} \max\{\theta_{\mathfrak{h}+j} + \theta_j x, 0\}, \tag{2.39}$$

for every $\mathfrak{h} \in \mathbb{N}_0$ *let* $\mathbf{R}_\mathfrak{h} \subseteq C([0,1],\mathbb{R})$ *satisfy* $\mathbf{R}_\mathfrak{h} = \{f \in Q^{-1}(\{\mathfrak{h}\}) \colon [\exists\, \theta \in \mathbb{R}^{3\mathfrak{h}+1} \colon f = \mathscr{N}^\theta]\}$*, and for every* $\mathfrak{h} \in \mathbb{N}_0$ *let* $\mathbf{S}_\mathfrak{h} \subseteq C([0,1],\mathbb{R})$ *satisfy*

$$\mathbf{S}_\mathfrak{h} = \Big\{ f \in Q^{-1}(\{\mathfrak{h}\}) \colon \Big(\exists\, k \in \mathbb{N}, i_1, i_2, \dots, i_k \in \mathbb{N} \colon \\ \big[\big(\tfrac{k}{2} \notin \mathbb{N} \big), \big(i_1 < i_2 < \dots < i_k \leq \mathfrak{h} + 1 \big), \big(\textstyle\sum_{j=1}^{k} (-1)^j A_{i_j}(f) = 0 \big) \big] \Big) \Big\} \tag{2.40}$$

(cf. Definitions 2.1 and 2.3). Then it holds for all $\mathfrak{h} \in \mathbb{N}_0$ *that*

$$\mathbf{S}_\mathfrak{h} \subseteq \mathbf{R}_\mathfrak{h}. \tag{2.41}$$

Proof of Lemma 2.6. We prove (2.41) by induction on $\mathfrak{h} \in \mathbb{N}_0$. For the base case $\mathfrak{h} = 0$ note that (2.40) ensures that

$$\mathbf{S}_0 = \Big\{ f \in Q^{-1}(\{0\}) \colon A_1(f) = 0 \Big\} = \{ f \in C([0,1],\mathbb{R}) \colon (\forall\, x \in [0,1] \colon f(x) = f(0)) \} \\ = \cup_{\theta \in \mathbb{R}} \{\mathscr{N}^\theta\}. \tag{2.42}$$

This establishes (2.40) in the base case $\mathfrak{h} = 0$. For the induction step let $\mathfrak{h} \in \mathbb{N}_0$ satisfy $\mathbf{S}_\mathfrak{h} \subseteq \mathbf{R}_\mathfrak{h}$ and let $F \in \mathbf{S}_{\mathfrak{h}+1}$. We intend to prove that $F \in \mathbf{R}_{\mathfrak{h}+1}$. Note that (2.40) ensures that there exist $K \in \mathbb{N}$, $I_1, I_2, \dots, I_K \in \mathbb{N}$ which satisfy

$$Q(F) = \mathfrak{h} + 1, \quad \tfrac{K}{2} \notin \mathbb{N}, \quad I_1 < I_2 < \dots < I_K \leq \mathfrak{h} + 2, \quad \text{and} \quad \textstyle\sum_{j=1}^{K} (-1)^j A_{I_j}(F) = 0. \tag{2.43}$$

Next let $f \colon [0,1] \to \mathbb{R}$ satisfy for all $x \in [0,1]$ that

$$f(x) = \begin{cases} F(x) - (A_{\mathfrak{h}+2}(F) - A_{\mathfrak{h}+1}(F)) \max\{x - q_{\mathfrak{h}+1}(F), 0\} & \colon I_K - 2 \neq \mathfrak{h}, \\ F(x) - (A_{\mathfrak{h}+2}(F) - A_{\mathfrak{h}+1}(F)) \max\{q_{\mathfrak{h}+1}(F) - x, 0\} & \colon I_K - 2 = \mathfrak{h}, \end{cases} \tag{2.44}$$

and let $k \in \mathbb{N}$, $i_1, i_2, \dots, i_k \in \mathbb{N}$ satisfy

$$\{i_1, i_2, \dots, i_k\} = \begin{cases} \{I_1, I_2, \dots, I_K\} & \colon I_K - 2 \neq \mathfrak{h}, \\ \{I_1, I_2, \dots, I_{K-2}\} & \colon I_K - 2 = \mathfrak{h} < \min\{I_{\max\{K-1,1\}}, K + \mathfrak{h} - 1, \} \\ (\cup_{l=1}^{K-1} \{I_l\}) \cup \{\mathfrak{h}+1\} & \colon I_K - 2 = \mathfrak{h} \geq \min\{I_{\max\{K-1,1\}}, K + \mathfrak{h} - 1\}. \end{cases} \tag{2.45}$$

Observe that (2.43) and (2.45) assure that $\frac{k}{2} \notin \mathbb{N}$ and $i_1 < i_2 < \dots < i_k \leq \mathfrak{h} + 1$. Moreover, note that (2.43) and (2.44) ensure that $f \in C([0,1],\mathbb{R})$, $Q(f) = \mathfrak{h}$, and

$$\left(\forall\, i \in \{1, 2, \dots, \mathfrak{h}+1\} \colon A_i(f) = \begin{cases} A_i(F) & \colon I_K - 2 \neq \mathfrak{h} \\ A_i(F) + A_{\mathfrak{h}+2}(F) - A_{\mathfrak{h}+1}(F) & \colon I_K - 2 = \mathfrak{h} \end{cases} \right). \tag{2.46}$$

In the next step we prove that

$$\textstyle\sum_{j=1}^{k} (-1)^j A_{i_j}(f) = 0. \tag{2.47}$$

In the following we distinguish between the case $I_K - 2 \neq \mathfrak{h}$, the case $I_K - 2 = \mathfrak{h} < \min\{I_{\max\{K-1,1\}}, K + \mathfrak{h} - 1\}$, and the case $I_K - 2 = \mathfrak{h} \geq \min\{I_{\max\{K-1,1\}}, K + \mathfrak{h} - 1\}$. We first prove (2.47) in the case

$$I_K - 2 \neq \mathfrak{h}. \tag{2.48}$$

Observe that (2.45) and (2.48) ensure that

$$\textstyle\sum_{j=1}^{k} (-1)^j A_{i_j}(f) = \sum_{j=1}^{K} (-1)^j A_{I_j}(F) = 0. \tag{2.49}$$

This establishes (2.47) in the case $I_K - 2 \neq \mathfrak{h}$. In the next step we prove (2.47) in the case

$$I_K - 2 = \mathfrak{h} < \min\{I_{\max\{K-1,1\}}, K + \mathfrak{h} - 1\}. \tag{2.50}$$

Note that (2.43), (2.45), (2.46), and (2.50) assure that

$$\begin{aligned}
&\textstyle\sum_{j=1}^{k} (-1)^j A_{i_j}(f) = \sum_{j=1}^{K-2} (-1)^j A_{I_j}(f) \\
&= \textstyle\sum_{j=1}^{K-2} (-1)^j \Big(A_{I_j}(F) + A_{\mathfrak{h}+2}(F) - A_{\mathfrak{h}+1}(F) \Big) \\
&= \Big[\textstyle\sum_{j=1}^{K-2} (-1)^j A_{I_j}(F) \Big] + \Big[\textstyle\sum_{j=1}^{K-2} (-1)^j \Big] \big[A_{\mathfrak{h}+2}(F) - A_{\mathfrak{h}+1}(F) \big] \\
&= \Big[\textstyle\sum_{j=1}^{K} (-1)^j A_{I_j}(F) \Big] - \Big[\textstyle\sum_{j=K-1}^{K} (-1)^j A_{I_j}(F) \Big] - \big[A_{\mathfrak{h}+2}(F) - A_{\mathfrak{h}+1}(F) \big] \\
&= - \big[A_{I_{K-1}}(F) - A_{I_K}(F) \big] - \big[A_{\mathfrak{h}+2}(F) - A_{\mathfrak{h}+1}(F) \big] \\
&= - \big[A_{\mathfrak{h}+1}(F) - A_{\mathfrak{h}+2}(F) \big] - \big[A_{\mathfrak{h}+2}(F) - A_{\mathfrak{h}+1}(F) \big] = 0.
\end{aligned} \tag{2.51}$$

This establishes (2.47) in the case $I_K - 2 = \mathfrak{h} < \min\{I_{\max\{K-1,1\}}, K + \mathfrak{h} - 1\}$. In the next step we prove (2.47) in the case

$$I_K - 2 = \mathfrak{h} \geq \min\{I_{\max\{K-1,1\}}, K + \mathfrak{h} - 1\}. \tag{2.52}$$

Observe that (2.43), (2.45), (2.46), and (2.52) assure that

$$\begin{aligned}
&\textstyle\sum_{j=1}^{k} (-1)^j A_{i_j}(f) = \Big[\textstyle\sum_{j=1}^{K-1} (-1)^j A_{I_j}(f) \Big] + (-1)^k A_{\mathfrak{h}+1}(f) \\
&= \Big[\textstyle\sum_{j=1}^{K-1} (-1)^j \big(A_{I_j}(F) + A_{\mathfrak{h}+2}(F) - A_{\mathfrak{h}+1}(F) \big) \Big] - \big[A_{\mathfrak{h}+1}(F) + A_{\mathfrak{h}+2}(F) - A_{\mathfrak{h}+1}(F) \big] \\
&= \Big[\textstyle\sum_{j=1}^{K-1} (-1)^j A_{I_j}(F) \Big] + \Big[\textstyle\sum_{j=1}^{K-1} (-1)^j \Big] \big(A_{\mathfrak{h}+2}(F) - A_{\mathfrak{h}+1}(F) \big) - A_{\mathfrak{h}+2}(F) \\
&= \Big[\textstyle\sum_{j=1}^{K} (-1)^j A_{I_j}(F) \Big] + A_{I_K}(F) - A_{\mathfrak{h}+2}(F) = A_{\mathfrak{h}+2}(F) - A_{\mathfrak{h}+2}(F) = 0.
\end{aligned} \tag{2.53}$$

This establishes (2.47) in the case $I_K - 2 = \mathfrak{h} \geq \min\{I_{\max\{K-1,1\}}, K + \mathfrak{h} - 1\}$. Note that (2.47) implies that $f \in \mathbf{S}_{\mathfrak{h}}$. The induction hypothesis that $\mathbf{S}_{\mathfrak{h}} \subseteq \mathbf{R}_{\mathfrak{h}}$ hence assures that $f \in \mathbf{R}_{\mathfrak{h}}$. Combining this with (2.44) and the fact that $Q(F) = \mathfrak{h} + 1$ shows that $F \in \mathbf{R}_{\mathfrak{h}+1}$. Induction thus establishes (2.41). The proof of Lemma 2.6 is thus complete. □

Finally, in Corollary 2.2 we combine the previous results to obtain the promised characterization.

Corollary 2.2. *Let $\mathfrak{h} \in \mathbb{N}_0$, for every $\theta = (\theta_1, \ldots, \theta_{3\mathfrak{h}+1}) \in \mathbb{R}^{3\mathfrak{h}+1}$ let $\mathscr{N}^\theta \colon [0,1] \to \mathbb{R}$ satisfy $x \in [0,1]$ that $\mathscr{N}^\theta(x) = \theta_{3\mathfrak{h}+1} + \sum_{j=1}^{\mathfrak{h}} \theta_{2\mathfrak{h}+j} \max\{\theta_{\mathfrak{h}+j} + \theta_j x, 0\}$, and let $f \in C([0,1], \mathbb{R})$. Then the following two statements are equivalent:*

(i) Then exists $\theta \in \mathbb{R}^{3\mathfrak{h}+1}$ such that $\mathscr{N}^\theta = f$.

(ii) There exist $k \in \mathbb{N}$, $i_1, i_2, \ldots, i_k \in \mathbb{N}$ such that $\frac{k}{2} \notin \mathbb{N}$, $i_1 < i_2 < \cdots < i_k \leq \mathfrak{h} + 1$, $Q(f) \leq \mathfrak{h}$, and

$$(\mathfrak{h} - Q(f) - 1)\left|\textstyle\sum_{j=1}^{k} (-1)^j A_{\min\{i_j, Q(f)+1\}}(f)\right| \geq 0 \tag{2.54}$$

(cf. Definitions 2.1 and 2.3).

Proof of Corollary 2.2. Observe that Lemma 2.3 and Lemma 2.5 prove that (item (i)→ item (ii)). Furthermore, note that Lemma 2.4 and Lemma 2.6 establish that (item (ii)→ item (i)). The proof of Corollary 2.2 is thus complete. □

2.4 Structure preserving approximations for piecewise affine linear functions

The next elementary lemma is an immediate consequence of the definitions in Subsection 2.2. It will be employed in the sequel to switch the endpoints of the domain $[0,1]$ and thereby make some simplifying assumptions.

Lemma 2.7. *Let $L \in \mathbb{R}$, $f \in C([0,1], \mathbb{R})$ satisfy for all $x, y \in [0,1]$ that $|f(x) - f(y)| \leq L|x - y|$, let $g \in \mathscr{L}$, $i \in \{1, 2, \ldots, Q(g) + 1\}$, $\mathbf{a} \in \mathbb{R}$ satisfy $L \leq |\mathbf{a}| \leq |A_i(g)|$ and $\mathbf{a}A_i(g) > 0$, let $F \colon [0,1] \to \mathbb{R}$ and $G \colon [0,1] \to \mathbb{R}$ satisfy for all $x \in [0,1]$ that $F(x) = -f(1-x)$ and $G(x) = -g(1-x)$, and let $I \in \mathbb{N}$ satisfy $I = Q(g) + 2 - i$ (cf. Definitions 2.1, 2.2, and 2.3). Then*

(i) it holds that $F \in C([0,1], \mathbb{R})$,

(ii) it holds for all $x, y \in [0,1]$ that $|F(x) - F(y)| \leq L|x - y|$,

(iii) it holds that $G \in \mathscr{L}$,

(iv) it holds that $Q(G) = Q(g)$,

(v) it holds that $I \in \{1, 2, \ldots, Q(G) + 1\}$,

(vi) it holds for all $j \in \{0, 1, \ldots, Q(g) + 1\}$ that $q_j(G) = 1 - q_{Q(g)+1-j}(g)$,

(vii) it holds for all $j \in \{1, 2, \ldots, Q(g) + 1\}$ that $A_j(G) = A_{Q(g)+2-j}(g)$,

(viii) it holds that $L \leq |\mathbf{a}| \leq |A_I(G)| = |A_i(g)|$, and

(ix) it holds that $\mathbf{a}A_I(G) = \mathbf{a}A_i(g) > 0$.

Proof of Lemma 2.7. Observe that (2.3) and (2.5) establish items (i), (ii), (iii), (iv), (v), (vi), (vii), (viii), and (ix). The proof of Lemma 2.7 is thus complete. □

Our next goal is to prove in Lemma 2.11 below that for any piecewise linear function $g \in \mathscr{L}$ there exists a piecewise linear $h \in \mathscr{L}$ which has at most as many breakpoints as g, approximates a given Lipschitz continuous target function $f \in C([0,1],\mathbb{R})$ as least as well as g, and has a Lipschitz constant bounded by the Lipschitz constant of f. To show this we will, roughly speaking, adjust the slopes of the piecewise linear function g one by one and apply induction. Loosely speaking, the following three results, Lemmas 2.8, 2.9, and 2.10 all consider different cases depending on the slope to be adjusted in each step. The various cases are illustrated in the figures.

Lemma 2.8. *Let $L \in (0,\infty)$, $f \in C([0,1],\mathbb{R})$ satisfy for all $x,y \in [0,1]$ that $|f(x)-f(y)| \leq L|x-y|$, let $g \in \mathscr{L}$, $i \in \{1,2,\ldots,Q(g)+1\}$, $\mathbf{a} \in \mathbb{R}$ satisfy $L \leq \mathbf{a} \leq A_i(g)$, assume for all $x \in (q_{i-1}(g),q_i(g))$ that $g(x) \neq f(x)$, and let $\mu\colon \mathcal{B}([0,1]) \to [0,\infty]$ be a finite measure (cf. Definitions 2.1, 2.2, and 2.3). Then there exists $h \in \mathscr{L}$ such that $\int_0^1 (h(y)-f(y))^2\,\mu(\mathrm{d}y) \leq \int_0^1 (g(y)-f(y))^2\,\mu(\mathrm{d}y)$, $Q(h) \leq Q(g)$, and*

$$(Q(g)-Q(h)-1)\left(\textstyle\sum_{j=1}^{Q(h)+1} |A_j(h)-A_j(g)\mathbb{1}_{\mathbb{N}\setminus\{i\}}(j)-\mathbf{a}\mathbb{1}_{\{i\}}(j)|\right) \geq 0. \tag{2.55}$$

Proof of Lemma 2.8. Throughout this proof assume without loss of generality that $\mathbf{a} < A_i(g)$, let $\mathfrak{q}_0,\mathfrak{q}_1,\ldots,\mathfrak{q}_{Q(g)+1} \in \mathbb{R}$ satisfy for all $j \in \{0,1,\ldots,Q(g)+1\}$ that $\mathfrak{q}_j = q_j(g)$, and assume without loss of generality[3] that $\forall\, x \in (\mathfrak{q}_{i-1},\mathfrak{q}_i)\colon f(x) < g(x)$. Note that the fact that f and g are continuous proves that $f(\mathfrak{q}_{i-1}) \leq g(\mathfrak{q}_{i-1})$ and $f(\mathfrak{q}_i) \leq g(\mathfrak{q}_i)$. In the following we distinguish between several cases:

(I) We first prove (2.55) in the case

$$i = Q(g)+1 \tag{2.56}$$

(cf. Figure 2.1). Let $h \in \mathscr{L}$ satisfy for all $x \in [0,\mathfrak{q}_{Q(g)}]$, $y \in [\mathfrak{q}_{Q(g)},1]$ that $h(x)=g(x)$ and $h(y) = g(\mathfrak{q}_{Q(g)}) + \mathbf{a}(y-\mathfrak{q}_{Q(g)})$. Observe that for all $j \in \{1,2,\ldots,Q(g)\}$ that $h|_{[\mathfrak{q}_{j-1},\mathfrak{q}_j]}$ is affine-linear with slope $A_j(g)$. Furthermore, note that $h|_{[\mathfrak{q}_{Q(g)},1]}$ is affine-linear with slope $\mathbf{a}$. Moreover, observe that $\big((A_{Q(g)}(g)=\mathbf{a}) \to (Q(h)=Q(g)-1 < Q(g))\big)$ and

$$\begin{aligned}\big((A_{Q(g)}(g) \neq \mathbf{a}) \to [(Q(h)=Q(g)) \wedge (A_{Q(g)+1}(h)=\mathbf{a})\\ \wedge (\forall\, j \in \mathbb{N}\cap[1,Q(h)]\colon A_j(h)=A_j(g))]\big).\end{aligned} \tag{2.57}$$

In addition, note that the fact that $\forall\, x,y \in [0,1]\colon |f(x)-f(y)| \leq L|x-y|$, the fact that $\forall\, y \in [\mathfrak{q}_{Q(g)},1]\colon g(y) = g(\mathfrak{q}_{Q(g)}) + A_i(g)(y-\mathfrak{q}_{Q(g)})$, and the fact that $\mathbf{a} \in [L,A_i(g))$ ensure that for all $y \in [\mathfrak{q}_{Q(g)},1]$ we have that

$$f(y) \leq f(\mathfrak{q}_{Q(g)}) + L(y-\mathfrak{q}_{Q(g)}) \leq g(\mathfrak{q}_{Q(g)}) + \mathbf{a}(y-\mathfrak{q}_{Q(g)}) = h(y) \leq g(y). \tag{2.58}$$

This implies that $\int_0^1 (h(y)-f(y))^2\,\mu(\mathrm{d}y) \leq \int_0^1 (g(y)-f(y))^2\,\mu(\mathrm{d}y)$, which establishes (2.55) in the case $i = Q(g)+1$.

[3]Otherwise the fact that f and g are continuous ensures that $\forall\, x \in (\mathfrak{q}_{i-1},\mathfrak{q}_i)\colon f(x) > g(x)$ and we can consider $f \curvearrowleft ([0,1] \ni x \mapsto -f(1-x) \in \mathbb{R})$, $g \curvearrowleft ([0,1] \ni x \mapsto -g(1-x) \in \mathbb{R})$, and $i \curvearrowleft Q(g)+2-i$ (cf. Lemma 2.7).

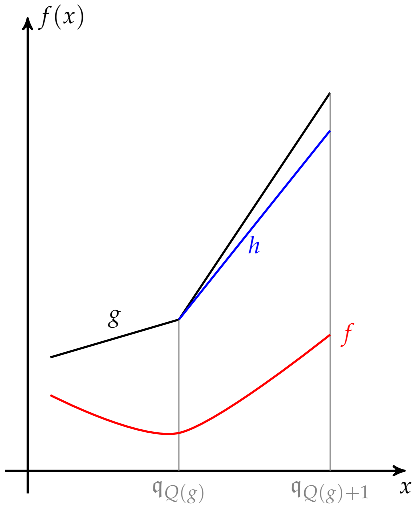

Figure 2.1: Case (I) in Lemma 2.8. Note that $\mathfrak{q}_i = \mathfrak{q}_{Q(g)+1} = 1$. The new function $h \in \mathscr{L}$ is linear on $[\mathfrak{q}_{i-1}, \mathfrak{q}_i] = [\mathfrak{q}_{Q(g)}, \mathfrak{q}_{Q(g)+1}]$ with slope $\mathbf{a}$ and agrees with g on $[0, \mathfrak{q}_{Q(g)}]$.

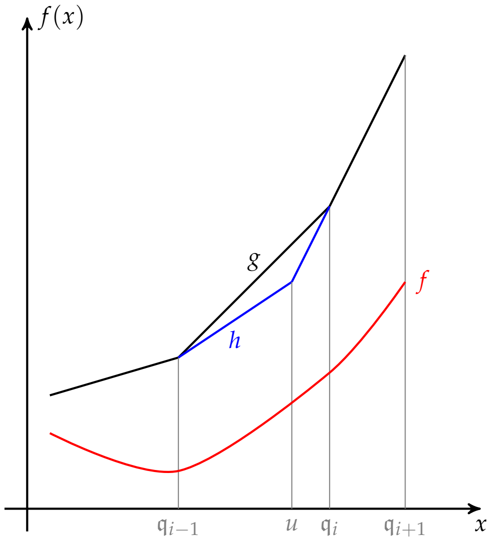

Figure 2.2: Case (II) in Lemma 2.8. The new function $h \in \mathscr{L}$ is linear on $[u, \mathfrak{q}_i]$ with slope $A_{i+1}(g)$, linear on $[\mathfrak{q}_{i-1}, u]$ with slope $\mathbf{a}$, and agrees with g outside of $[\mathfrak{q}_{i-1}, \mathfrak{q}_i]$.

(II) Next we prove (2.55) in the case

$$(i < Q(g) + 1) \wedge (A_{i+1}(g) > A_i(g)) \wedge (g(\mathfrak{q}_{i+1}) - g(\mathfrak{q}_{i-1}) \geq \mathbf{a}(\mathfrak{q}_{i+1} - \mathfrak{q}_{i-1})) \quad (2.59)$$

(cf. Figure 2.2). Observe that the fact that $\mathbf{a} \in [L, A_i(g))$ shows that

$$
\begin{aligned}
&g(\mathfrak{q}_{i+1}) + A_{i+1}(g)(\mathfrak{q}_i - \mathfrak{q}_{i+1}) - g(\mathfrak{q}_{i-1}) - \mathbf{a}(\mathfrak{q}_i - \mathfrak{q}_{i-1}) \\
&= g(\mathfrak{q}_i) - g(\mathfrak{q}_{i-1}) - \mathbf{a}(\mathfrak{q}_i - \mathfrak{q}_{i-1}) > 0.
\end{aligned} \tag{2.60}
$$

Furthermore, note that the fact that $A_{i+1}(g) > A_i(g)$ shows that

$$
A_{i+1}(g) > \left[\tfrac{\mathfrak{q}_{i+1}-\mathfrak{q}_i}{\mathfrak{q}_{i+1}-\mathfrak{q}_{i-1}}\right] A_{i+1}(g) + \left[\tfrac{\mathfrak{q}_i-\mathfrak{q}_{i-1}}{\mathfrak{q}_{i+1}-\mathfrak{q}_{i-1}}\right] A_i(g) = \tfrac{g(\mathfrak{q}_{i+1})-g(\mathfrak{q}_{i-1})}{\mathfrak{q}_{i+1}-\mathfrak{q}_{i-1}}. \tag{2.61}
$$

Hence, we obtain $g(\mathfrak{q}_{i+1}) + A_{i+1}(g)(\mathfrak{q}_{i-1} - \mathfrak{q}_{i+1}) - g(\mathfrak{q}_{i-1}) < 0$. The intermediate value theorem and (2.60) therefore assure that there exists $u \in (\mathfrak{q}_{i-1}, \mathfrak{q}_i)$ which satisfies

$$
g(\mathfrak{q}_{i-1}) + \mathbf{a}(u - \mathfrak{q}_{i-1}) = g(\mathfrak{q}_{i+1}) + A_{i+1}(g)(u - \mathfrak{q}_{i+1}). \tag{2.62}
$$

Let $h \in \mathscr{L}$ satisfy for all $x \in [0, \mathfrak{q}_{i-1}] \cup [\mathfrak{q}_{i+1}, 1]$, $y \in [\mathfrak{q}_{i-1}, u]$, $z \in [u, \mathfrak{q}_{i+1}]$ that $h(x) = g(x)$, $h(y) = g(\mathfrak{q}_{i-1}) + \mathbf{a}(y - \mathfrak{q}_{i-1})$, and $h(z) = g(\mathfrak{q}_{i+1}) + A_{i+1}(g)(z - \mathfrak{q}_{i+1})$. Observe that

$$
\begin{aligned}
\big([(i = 1) \vee (A_{\max\{i-1,1\}}(g) \neq \mathbf{a})] \to [(Q(h) = Q(g)) \\
\wedge (\forall\, j \in (\mathbb{N} \cap [1, Q(h)+1]) \backslash \{i\} \colon A_j(h) = A_j(g))]\big)
\end{aligned} \tag{2.63}
$$

and $\big([(i > 1) \wedge (A_{\max\{i-1,1\}}(g) = \mathbf{a})] \to (Q(h) < Q(g))\big)$. Moreover, note that the fact that $\mathbf{a} \in [L, A_i(g)]$ implies for all $y \in [\mathfrak{q}_{i-1}, \mathfrak{q}_{i+1}]$ that $h(y) \le g(y)$. In addition, observe that the fact that $f(\mathfrak{q}_{i-1}) \le g(\mathfrak{g}_{i-1})$, the fact that $L \le \mathbf{a} \le A_{i+1}(g)$, and the fact that $\forall\, x, y \in [0,1] \colon |f(x) - f(y)| \le L|x - y|$ prove for all $y \in [\mathfrak{q}_{i-1}, \mathfrak{q}_{i+1}]$ that $f(y) \le h(y)$. Hence, we obtain $\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \le \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$. This establishes (2.55) in the case $(i < Q(g) + 1) \wedge (A_{i+1}(g) > A_i(g)) \wedge (g(\mathfrak{q}_{i+1}) - g(\mathfrak{q}_{i-1}) \ge \mathbf{a}(\mathfrak{q}_{i+1} - \mathfrak{q}_{i-1}))$.

(III) Next we prove (2.55) in the case

$$
(i < Q(g) + 1) \wedge (A_{i+1}(g) < A_i(g)) \wedge (g(\mathfrak{q}_{i+1}) - g(\mathfrak{q}_{i-1}) \ge \mathbf{a}(\mathfrak{q}_{i+1} - \mathfrak{q}_{i-1})) \tag{2.64}
$$

(cf. Figure 2.3). Let $h \in \mathscr{L}$ satisfy for all $x \in [0, \mathfrak{q}_{i-1}] \cup [\mathfrak{q}_{i+1}, 1]$, $y \in [\mathfrak{q}_{i-1}, \mathfrak{q}_{i+1}]$ that $h(x) = g(x)$ and $h(y) = g(\mathfrak{q}_{i-1}) + \left[\tfrac{g(\mathfrak{q}_{i+1})-g(\mathfrak{q}_{i-1})}{\mathfrak{q}_{i+1}-\mathfrak{q}_{i-1}}\right](y - \mathfrak{q}_{i-1})$. Clearly, we have that $h \in \mathscr{L}$ and $Q(h) < Q(g)$. Furthermore, note that the fact that $A_{i+1}(g) < A_i(g)$ shows that

$$
A_{i+1}(g) \le \left[\tfrac{\mathfrak{q}_{i+1}-\mathfrak{q}_i}{\mathfrak{q}_{i+1}-\mathfrak{q}_{i-1}}\right] A_{i+1}(g) + \left[\tfrac{\mathfrak{q}_i-\mathfrak{q}_{i-1}}{\mathfrak{q}_{i+1}-\mathfrak{q}_{i-1}}\right] A_i(g) = \tfrac{g(\mathfrak{q}_{i+1})-g(\mathfrak{q}_{i-1})}{\mathfrak{q}_{i+1}-\mathfrak{q}_{i-1}} \le A_i(g). \tag{2.65}
$$

Therefore, we obtain for all $y \in [\mathfrak{q}_{i-1}, \mathfrak{q}_{i+1}]$ that $h(y) \le g(y)$. Moreover, observe that the fact that $f(\mathfrak{q}_{i-1}) \le g(\mathfrak{g}_{i-1})$, the fact that $L \le \frac{g(\mathfrak{q}_{i+1})-g(\mathfrak{g}_{i-1})}{\mathfrak{q}_{i+1}-\mathfrak{g}_{i-1}}$, and the fact that $\forall\, x, y \in [0,1] \colon |f(x) - f(y)| \le L|x - y|$ prove for all $y \in [\mathfrak{q}_{i-1}, \mathfrak{q}_{i+1}]$ that $f(y) \le h(y)$. Hence, we obtain $\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \le \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$. This establishes (2.55) in the case $(i < Q(g) + 1) \wedge (A_{i+1}(g) > A_i(g)) \wedge (g(q_{i+1}) - g(\mathfrak{q}_{i-1}) \ge \mathbf{a}(\mathfrak{q}_{i+1} - \mathfrak{q}_{i-1}))$.

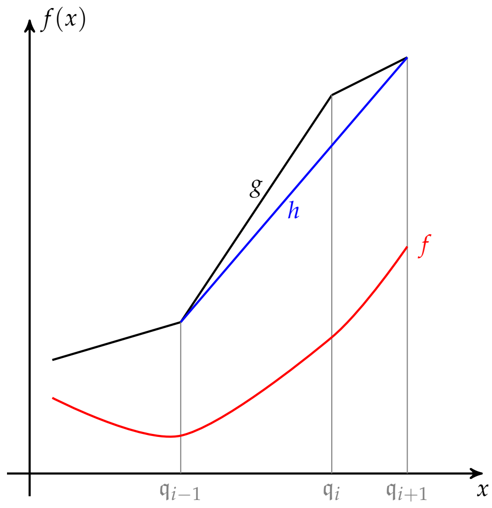

Figure 2.3: Case (III) in Lemma 2.8. The new function $h \in \mathscr{L}$ is linear on $[\mathfrak{q}_{i-1}, \mathfrak{q}_{i+1}]$ with slope $\frac{g(\mathfrak{q}_{i+1})-g(\mathfrak{q}_{i-1})}{\mathfrak{q}_{i+1}-\mathfrak{q}_{i-1}} \geq \mathbf{a}$ and agrees with g outside of $[\mathfrak{q}_{i-1}, \mathfrak{q}_{i+1}]$. It thus satisfies $Q(h) < Q(g)$.

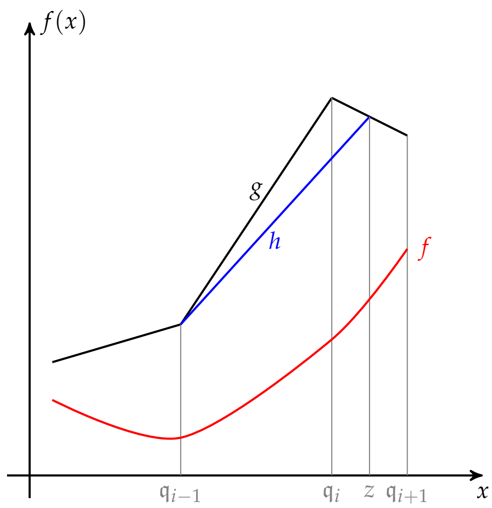

Figure 2.4: Case (IV) in Lemma 2.8. The new function $h \in \mathscr{L}$ is linear on $[\mathfrak{q}_{i-1}, z]$ with slope $\mathbf{a}$ and agrees with g outside of $[\mathfrak{q}_{i-1}, z]$.

(IV) Finally, we prove (2.55) in the case

$$(i < Q(g) + 1) \wedge (g(\mathfrak{q}_{i+1}) - g(\mathfrak{q}_{i-1}) < \mathbf{a}(\mathfrak{q}_{i+1} - \mathfrak{q}_{i-1})) \tag{2.66}$$

(cf. Figure 2.4). Note that the fact that $g(\mathfrak{q}_i) - g(\mathfrak{q}_{i-1}) = A_i(g)(\mathfrak{q}_i - \mathfrak{q}_{i-1}) > \mathbf{a}(\mathfrak{q}_i - \mathfrak{q}_{i-1})$ and the intermediate value theorem demonstrate that there exists $z \in (\mathfrak{q}_i, \mathfrak{q}_{i+1})$ which satisfies $g(z) = g(\mathfrak{q}_{i-1}) + \mathbf{a}(z - \mathfrak{q}_{i-1})$. Let $h \in \mathscr{L}$ satisfy for all $x \in [0, \mathfrak{q}_{i-1}] \cup$

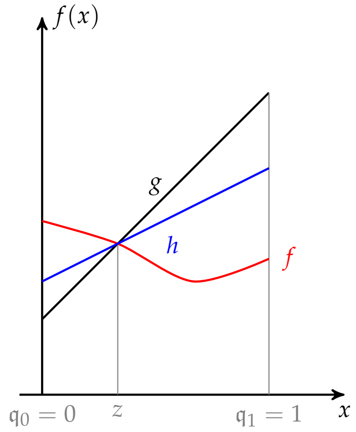

Figure 2.5: Case (I) in Lemma 2.9. Here $Q(g) = 0$, so g is linear. The new function $h \in \mathscr{L}$ is linear on $[\mathfrak{q}_0, \mathfrak{q}_1] = [0, 1]$ with slope $\mathbf{a}$ and satisfies $h(z) = g(z) = f(z)$.

$[z, 1]$, $y \in [\mathfrak{q}_{i-1}, z]$ that $h(x) = g(x)$ and $h(y) = g(\mathfrak{q}_{i-1}) + \mathbf{a}(y - \mathfrak{q}_{i-1})$. Observe that $Q(h) = Q(g)$, $A_i(h) = \mathbf{a}$, and $\forall\, j \in \{1, 2, \ldots, Q(g)+1\} \backslash \{i\} \colon A_j(h) = A_j(g)$. In addition, note that the fact that $f(\mathfrak{q}_{i-1}) \leq g(\mathfrak{g}_{i-1})$, the fact that $L \leq \mathbf{a}$, and the fact that $\forall\, x, y \in [0, 1] \colon |f(x) - f(y)| \leq L|x - y|$ prove for all $y \in [\mathfrak{q}_{i-1}, z]$ that $f(y) \leq h(y) \leq g(y)$. Therefore, we obtain $\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$. This establishes (2.55) in the case $(i < Q(g) + 1) \wedge (g(\mathfrak{q}_{i+1}) - g(\mathfrak{q}_{i-1}) < \mathbf{a}(\mathfrak{q}_{i+1} - \mathfrak{q}_{i-1}))$.

The proof of Lemma 2.8 is thus complete. □

Lemma 2.9. *Let $L \in (0, \infty)$, $f \in C([0, 1], \mathbb{R})$ satisfy for all $x, y \in [0, 1]$ that $|f(x) - f(y)| \leq L|x - y|$, let $g \in \mathscr{L}$, $i \in \{1, Q(g) + 1\}$, $\mathbf{a} \in \mathbb{R}$ satisfy $L \leq \mathbf{a} \leq A_i(g)$, let $z \in (q_{i-1}(g), q_i(g))$ satisfy $g(z) = f(z)$, and let $\mu \colon \mathcal{B}([0, 1]) \to [0, \infty]$ be a finite measure. Then there exists $h \in \mathscr{L}$ such that $\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$, $Q(h) \leq Q(g)$, and*

$$(Q(g) - Q(h) - 1) \left(\textstyle\sum_{j=1}^{Q(h)+1} |A_j(h) - A_j(g) \mathbb{1}_{\mathbb{N} \backslash \{i\}}(j) - \mathbf{a} \mathbb{1}_{\{i\}}(j)| \right) \geq 0. \tag{2.67}$$

Proof of Lemma 2.9. Throughout this proof assume without loss of generality that $\mathbf{a} < A_i(g)$, assume without loss of generality that $i = 1$ (cf. Lemma 2.7), and let $\mathfrak{q}_0, \mathfrak{q}_1, \ldots, \mathfrak{q}_{Q(g)+1} \in \mathbb{R}$ satisfy for all $j \in \{0, 1, \ldots, Q(g) + 1\}$ that $\mathfrak{q}_j = q_j(g)$. In the following we distinguish between several cases:

(I) We first prove (2.67) in the case

$$1 = i = Q(g) + 1 \tag{2.68}$$

(cf. Figure 2.5). Let $h \in \mathscr{L}$ satisfy for all $x \in [0, 1]$ that $h(x) = g(z) + \mathbf{a}(x - z)$. Observe that $h \in \mathscr{L}$, $Q(h) = 0 = Q(g)$, and $A_1(h) = \mathbf{a}$. Furthermore, note that the assumption that $A_1(g) > \mathbf{a} \geq L$ and the fact that $\forall\, x, y \in [0, 1] \colon |f(x) - f(y)| \leq L|x - y|$ prove for all $x \in [0, z]$, $y \in [z, 1]$ that $g(x) \leq h(x) \leq f(x)$ and $f(y) \leq$

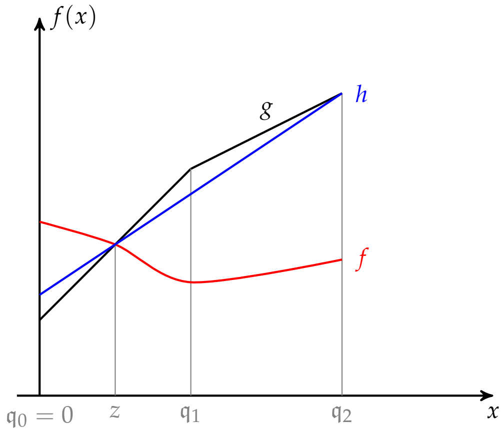

Figure 2.6: Case (II) in Lemma 2.9. The new function $h \in \mathscr{L}$ satisfies $h(z) = f(z) = g(z)$, is linear on $[\mathfrak{q}_0, \mathfrak{q}_2]$ with slope $\frac{g(\mathfrak{q}_2)-g(z)}{\mathfrak{q}_2-z} \geq \mathbf{a}$, and agrees with g on $[\mathfrak{q}_2, 1]$.

$h(y) \leq g(y)$. This implies $\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$. This establishes (2.67) in the case $1 = i = Q(g) + 1$.

(II) Next we prove (2.67) in the case

$$(1 = i < Q(g) + 1) \wedge (A_2(g) < A_1(g)) \wedge (g(\mathfrak{q}_2) - g(z) \geq \mathbf{a}(\mathfrak{q}_2 - z)) \tag{2.69}$$

(cf. Figure 2.6). Let $h \in \mathscr{L}$ satisfy for all $x \in [0, \mathfrak{q}_2]$, $y \in [\mathfrak{q}_2, 1]$ that $h(x) = g(z) + \big[\frac{g(\mathfrak{q}_2)-g(z)}{\mathfrak{q}_2-z}\big](x - z)$ and $h(y) = g(y)$. Clearly, we have that $h \in \mathscr{L}$ and $Q(h) < Q(g)$. Moreover, observe that the fact that

$$\begin{aligned} A_1(g) > \big[\tfrac{\mathfrak{q}_2-\mathfrak{q}_1}{\mathfrak{q}_2-z}\big] A_2(g) + \big[\tfrac{\mathfrak{q}_1-z}{\mathfrak{q}_2-z}\big] A_1(g) &= \tfrac{g(\mathfrak{q}_2)-g(z)}{\mathfrak{q}_2-z} \\ &\geq \max\{A_2(g), \mathbf{a}\} \geq \max\{A_2(g), L\} \end{aligned} \tag{2.70}$$

and the fact $\forall\, x, y \in [0, 1]\colon |f(x) - f(y)| \leq L|x - y|$ prove that for all $x \in [0, z]$, $y \in [z, \mathfrak{q}_2]$ we have that $g(x) \leq h(x) \leq f(x)$ and $f(y) \leq h(y) \leq g(y)$. Hence, we obtain $\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$. This establishes (2.67) in the case $(1 = i < Q(g) + 1) \wedge (A_2(g) < A_1(g)) \wedge (g(\mathfrak{q}_2) - g(z) \geq \mathbf{a}(\mathfrak{q}_2 - z))$.

(III) Next we prove (2.67) in the case

$$(1 = i < Q(g) + 1) \wedge (A_2(g) > A_1(g)) \wedge (g(\mathfrak{q}_2) - g(z) \geq \mathbf{a}(\mathfrak{q}_2 - z)) \tag{2.71}$$

(cf. Figure 2.7). Note that the intermediate value theorem ensures that there exists $u \in [z, \mathfrak{q}_1]$ which satisfies $g(z) + \mathbf{a}(u - z) = g(\mathfrak{q}_2) + A_2(g)(u - \mathfrak{q}_2)$. Let $h \in \mathscr{L}$ satisfy for all $x \in [0, u]$, $y \in [u, \mathfrak{q}_1]$, $z \in [\mathfrak{q}_1, 1]$ that $h(x) = g(z) + \mathbf{a}(x - z)$, $h(y) = h(u) + A_2(g)(y - u)$, and $h(z) = g(z)$. Observe that $Q(h) = Q(g)$, $A_1(h) = \mathbf{a}$, and $\forall\, j \in \{2, 3, \ldots, Q(g) + 1\}\colon A_j(h) = A_j(g)$. In addition, note that for all $x \in [0, z]$, $y \in [z, \mathfrak{q}_1]$ it holds that $g(x) \leq h(x) \leq f(x)$ and $f(y) \leq h(y) \leq g(y)$. Therefore, we

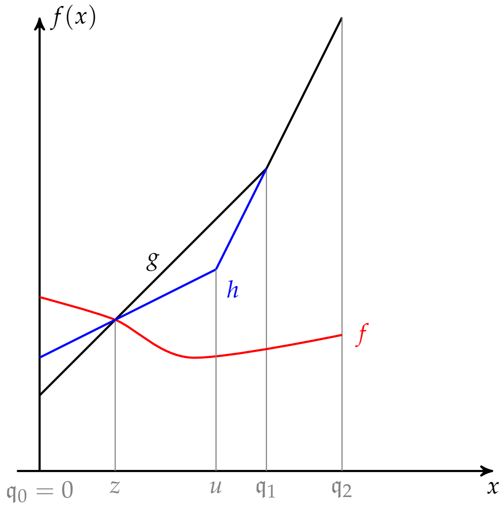

Figure 2.7: Case (III) in Lemma 2.9. The new function $h \in \mathscr{L}$ is linear on $[0, u]$ with slope $\mathbf{a}$, linear on $[u, \mathfrak{q}_1]$ with slope $A_2(g)$, and agrees with g on $[\mathfrak{q}_1, 1]$.

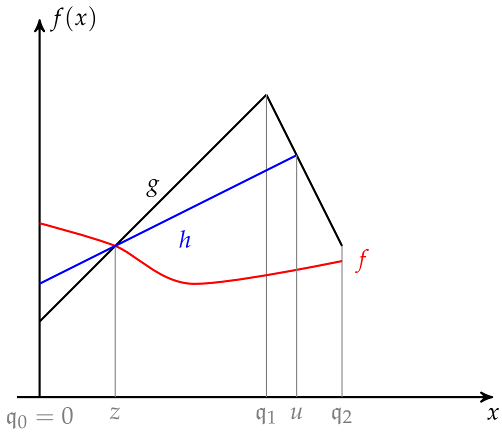

Figure 2.8: Case (IV) in Lemma 2.9. The new function $h \in \mathscr{L}$ is linear on $[0, u]$ with slope $\mathbf{a}$ and agrees with g on $[u, 1]$.

obtain $\int_0^1 (h(y) - f(y))^2 \,\mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \,\mu(\mathrm{d}y)$. This establishes (2.67) in the case $(1 = i < Q(g) + 1) \wedge (A_2(g) > A_1(g)) \wedge (g(\mathfrak{q}_2) - g(z) \geq \mathbf{a}(\mathfrak{q}_2 - z))$.

(IV) Finally, we prove (2.67) in the case

$$(1 = i < Q(g) + 1) \wedge (g(\mathfrak{q}_2) - g(z) < \mathbf{a}(\mathfrak{q}_2 - z)) \tag{2.72}$$

(cf. Figure 2.8). Observe that the fact that $g(\mathfrak{q}_1) - g(z) = A_1(g)(\mathfrak{q}_1 - z) > \mathbf{a}(\mathfrak{q}_1 - z)$ and the intermediate value theorem demonstrate that there exists $u \in (\mathfrak{q}_1, \mathfrak{q}_2)$ which satisfies $g(u) = g(z) + \mathbf{a}(u - z)$. Let $h \in \mathscr{L}$ satisfy for all $x \in [0, u]$, $y \in [u, 1]$ that

$h(x) = g(z) + \mathbf{a}(x - z)$ and $h(y) = g(y)$. Note that $\big((A_2(g) = \mathbf{a}) \to (Q(h) < Q(g))\big)$ and

$$\big((A_2(g) \neq \mathbf{a}) \to [(Q(h) = Q(g)) \wedge (A_1(h) = \mathbf{a}) \wedge (\forall\, j \in \mathbb{N} \cap (1, Q(g) + 1] : A_j(h) = A_j(g))]\big). \quad (2.73)$$

Furthermore, observe that the assumption that $A_1(g) > \mathbf{a}$ and the fact that $\forall\, x, y \in [0,1]\colon |f(x) - f(y)| \leq L|x - y|$ prove that for all $x \in [0, z]$, $y \in [z, u]$ we have that $g(x) \leq h(x) \leq f(x)$ and $f(y) \leq h(y) \leq g(y)$. This implies $\int_0^1 (h(y) - f(y))^2 \,\mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \,\mu(\mathrm{d}y)$. This establishes (2.67) in the case $(1 = i < Q(g) + 1) \wedge (g(\mathfrak{q}_2) - g(z) < \mathbf{a}(\mathfrak{q}_2 - z))$.

The proof of Lemma 2.9 is thus complete. □

Lemma 2.10. *Let $L \in (0, \infty)$, $f \in C([0,1], \mathbb{R})$ satisfy for all $x, y \in [0,1]$ that $|f(x) - f(y)| \leq L|x - y|$, let $g \in \mathscr{L}$, $i \in \mathbb{N} \cap (1, Q(g)]$, $\mathbf{a} \in \mathbb{R}$ satisfy $L \leq \mathbf{a} \leq A_i(g)$, let $z \in (q_{i-1}(g), q_i(g))$ satisfy $g(z) = f(z)$, and let $\mu\colon \mathcal{B}([0,1]) \to [0, \infty]$ be a finite measure. Then there exists $h \in \mathscr{L}$ such that $\int_0^1 (h(y) - f(y))^2 \,\mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \,\mu(\mathrm{d}y)$, $Q(h) \leq Q(g)$, and*

$$(Q(g) - Q(h) - 1)\big(\textstyle\sum_{j=1}^{Q(h)+1} |A_j(h) - A_j(g)\mathbb{1}_{\mathbb{N}\setminus\{i\}}(j) - \mathbf{a}\mathbb{1}_{\{i\}}(j)|\big) \geq 0. \quad (2.74)$$

Proof of Lemma 2.10. Throughout this proof assume without loss of generality that $\mathbf{a} < A_i(g)$ and let $\mathfrak{q}_0, \mathfrak{q}_1, \ldots, \mathfrak{q}_{Q(g)+1} \in \mathbb{R}$ satisfy for all $j \in \{0, 1, \ldots, Q(g) + 1\}$ that $\mathfrak{q}_j = q_j(g)$. In the following we distinguish between several cases:

(I) We first prove (2.74) in the case

$$A_i(g) < \min\{A_{i-1}(g), A_{i+1}(g)\} \quad (2.75)$$

(cf. Figure 2.9). Note that the fact that $\mathbf{a} < A_i(g)$ and the intermediate value theorem assure that there exist $u \in (\mathfrak{q}_{i-1}, z)$, $v \in (z, \mathfrak{q}_i)$ which satisfy $g(\mathfrak{q}_{i-2}) + A_{i-1}(g)(u - \mathfrak{q}_{i-2}) = g(z) + \mathbf{a}(u - z)$ and $g(\mathfrak{q}_{i+1}) + A_{i+1}(g)(v - \mathfrak{q}_{i+1}) = g(z) + \mathbf{a}(v - z)$. Let $h \in \mathscr{L}$ satisfy for all $x \in [0, \mathfrak{q}_{i-1}] \cup [\mathfrak{q}_{i+1}, 1]$, $y_1 \in [\mathfrak{q}_{i-1}, u]$, $y_2 \in [u, v]$, $y_3 \in [v, \mathfrak{q}_i]$ that $h(x) = g(x)$, $h(y_1) = g(\mathfrak{q}_{i-1}) + A_{i-1}(g)(y_1 - \mathfrak{q}_{i-1})$, $h(y_2) = g(z) + \mathbf{a}(y_2 - z)$, and $h(y_3) = g(\mathfrak{q}_i) + A_{i+1}(g)(y_3 - \mathfrak{q}_i)$. Observe that $Q(h) = Q(g)$, $A_i(h) = \mathbf{a}$, and $\forall\, j \in \{1, 2, \ldots, Q(g) + 1\} \backslash \{i\}\colon A_j(h) = A_j(g)$. Furthermore, note that the assumption that $A_i(g) > \mathbf{a}$ and the fact that $\forall\, x, y \in [0,1]\colon |f(x) - f(y)| \leq L|x - y|$ demonstrate for all $x \in [\mathfrak{q}_{i-1}, z]$, $y \in [z, \mathfrak{q}_i]$ that $g(x) \leq h(x) \leq f(x)$ and $f(y) \leq h(y) \leq g(y)$. Hence, we obtain $\int_0^1 (h(y) - f(y))^2 \,\mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \,\mu(\mathrm{d}y)$. This establishes (2.74) in the case $A_i(g) < \min\{A_{i-1}(g), A_{i+1}(g)\}$.

(II) Next we prove (2.74) in the case

$$\max\{\tfrac{g(\mathfrak{q}_{i+1}) - g(z)}{\mathfrak{q}_{i+1} - z}, \tfrac{g(\mathfrak{q}_{i-2}) - g(z)}{\mathfrak{q}_{i-2} - z}\} < \mathbf{a} \quad (2.76)$$

(cf. Figure 2.10). Observe that the fact that $A_i(g) > \mathbf{a}$ proves that $\max\{A_{i-1}(g), A_{i+1}(g)\} < \mathbf{a}$. Moreover, note that the fact that $A_i(g) > \mathbf{a}$ and the

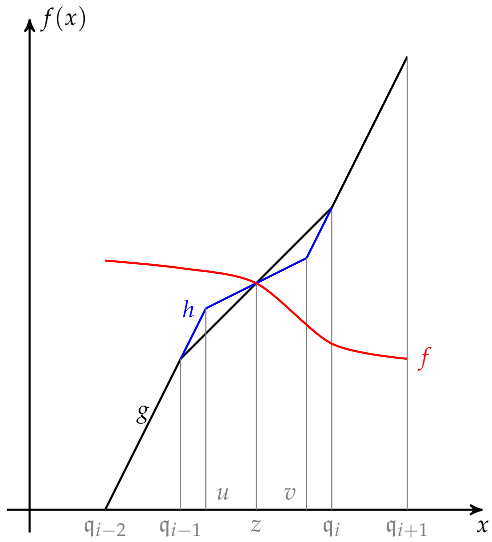

Figure 2.9: Case (I) in Lemma 2.10. The new function $h \in \mathscr{L}$ is linear on $[\mathfrak{q}_{i-1}, u]$ with slope $A_{i-1}(g)$, linear on $[u, v]$ with slope $\mathbf{a}$, linear on $[v, \mathfrak{q}_i]$ with slope $A_{i+1}(g)$, and agrees with g outside of $[\mathfrak{q}_{i-1}, \mathfrak{q}_i]$.

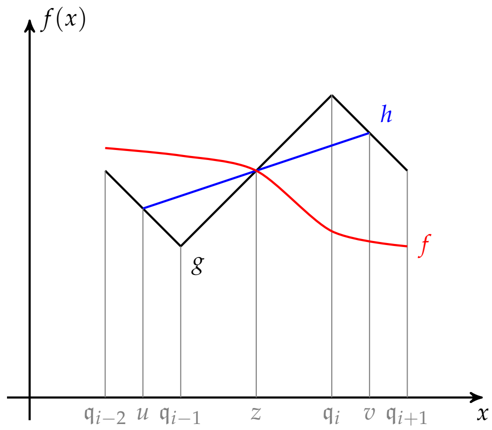

Figure 2.10: Case (II) in Lemma 2.10. The new function $h \in \mathscr{L}$ is linear on $[u, v]$ with slope $\mathbf{a}$ and agrees with g outside of $[u, v]$.

intermediate value theorem assure that there exist $u \in (\mathfrak{q}_{i-2}, \mathfrak{q}_{i-1})$, $v \in (\mathfrak{q}_i, \mathfrak{q}_{i+1})$ which satisfy $\frac{g(u)-g(z)}{u-z} = \frac{g(v)-g(z)}{v-z} = \mathbf{a}$. Let $h \in \mathscr{L}$ satisfy for all $x \in [0, u] \cup [v, 1]$, $y \in [u, v]$ that $h(x) = g(x)$ and $h(y) = g(z) + \mathbf{a}(y - z)$. Observe that the fact that $A_{i-1}(g) \neq \mathbf{a}$ and the fact that $A_{i+1}(g) \neq \mathbf{a}$ show that $Q(h) = Q(g)$, $A_i(h) = \mathbf{a}$, and $\forall\, j \in \{1, 2, \ldots, Q(g) + 1\} \backslash \{i\} \colon A_j(h) = A_j(g)$. In addition, note that the assumption that $A_i(g) > \mathbf{a}$ and the fact that $\forall\, x, y \in [0, 1] \colon |f(x) - f(y)| \leq L|x - y|$ demonstrate for all $x \in [u, z]$, $y \in [z, v]$ that $g(x) \leq h(x) \leq f(x)$ and $f(y) \leq h(y) \leq g(y)$. Therefore, we obtain $\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$. This establishes

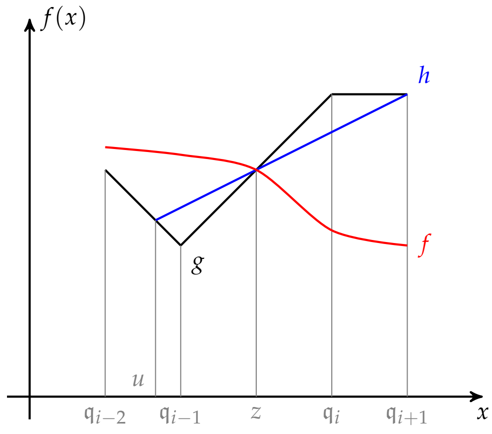

Figure 2.11: Case (III) in Lemma 2.10. The new function $h \in \mathscr{L}$ is linear on $[u, \mathfrak{q}_{i+1}]$ with slope $\frac{g(\mathfrak{q}_{i+1})-g(z)}{\mathfrak{q}_{i+1}-z} \geq \mathbf{a}$ and agrees with g outside of $[u, \mathfrak{q}_{i+1}]$.

(2.74) in the case $\max\{\frac{g(\mathfrak{q}_{i+1})-g(z)}{\mathfrak{q}_{i+1}-z}, \frac{g(\mathfrak{q}_{i-2})-g(z)}{\mathfrak{q}_{i-2}-z}\} < \mathbf{a}$.

(III) Next we prove (2.74) in the case

$$\big(A_i(g) > \max\{A_{i-1}(g), A_{i+1}(g)\}\big) \wedge \big(\max\{\tfrac{g(\mathfrak{q}_{i+1})-g(z)}{\mathfrak{q}_{i+1}-z}, \tfrac{g(\mathfrak{q}_{i-2})-g(z)}{\mathfrak{q}_{i-2}-z}\} \geq \mathbf{a}\big) \tag{2.77}$$

(cf. Figure 2.11). In the following we assume without loss of generality that $\frac{g(\mathfrak{q}_{i+1})-g(z)}{\mathfrak{q}_{i+1}-z} \geq \frac{g(\mathfrak{q}_{i-2})-g(z)}{\mathfrak{q}_{i-2}-z}$ (cf. Lemma 2.7). Observe that the fact that $A_i(g) > \max\{A_{i-1}(g), A_{i+1}(g)\}$ shows that

$$A_i(g) = \frac{g(\mathfrak{q}_{i-1}) - g(z)}{\mathfrak{q}_{i-1} - z} > \frac{g(\mathfrak{q}_{i+1}) - g(z)}{\mathfrak{q}_{i+1} - z} \geq \frac{g(\mathfrak{q}_{i-2}) - g(z)}{\mathfrak{q}_{i-2} - z}. \tag{2.78}$$

The intermediate value theorem hence proves that there exists $u \in [\mathfrak{q}_{i-2}, \mathfrak{q}_{i-1})$ which satisfies $g(u) = g(z) + \big[\frac{g(\mathfrak{q}_{i+1})-g(z)}{\mathfrak{q}_{i+1}-z}\big](u - z)$. Let $h \in \mathscr{L}$ satisfy for all $x \in [0, u] \cup [\mathfrak{q}_{i+1}, 1]$, $y \in [u, \mathfrak{q}_{i+1}]$ that $h(x) = g(x)$ and $h(y) = g(z) + \big[\frac{g(\mathfrak{q}_{i+1})-g(z)}{\mathfrak{q}_{i+1}-z}\big](y - z)$. Note that $Q(h) < Q(g)$. Furthermore, observe that the assumption that $A_i(g) > \mathbf{a}$ and the fact that $\forall\, x, y \in [0, 1] \colon |f(x) - f(y)| \leq L|x - y|$ demonstrate for all $x \in [u, z]$, $y \in [z, \mathfrak{q}_{i+1}]$ that $g(x) \leq h(x) \leq f(x)$ and $f(y) \leq h(y) \leq g(y)$. Therefore, we obtain $\int_0^1 (h(y) - f(y))^2\, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2\, \mu(\mathrm{d}y)$. This establishes (2.74) in the case $\big(A_i(g) > \max\{A_{i-1}(g), A_{i+1}(g)\}\big) \wedge \big(\max\{\frac{g(\mathfrak{q}_{i+1})-g(z)}{\mathfrak{q}_{i+1}-z}, \frac{g(\mathfrak{q}_{i-2})-g(z)}{\mathfrak{q}_{i-2}-z}\} \geq \mathbf{a}\big)$.

(IV) Next we prove (2.74) in the case

$$\big(\max\{A_{i-1}(g), A_{i+1}(g)\} > A_i(g) > \min\{A_{i-1}(g), A_{i+1}(g)\}\big) \wedge \big(\min\{\tfrac{g(\mathfrak{q}_{i+1})-g(z)}{\mathfrak{q}_{i+1}-z}, \tfrac{g(\mathfrak{q}_{i-2})-g(z)}{\mathfrak{q}_{i-2}-z}\} \geq \mathbf{a}\big) \tag{2.79}$$

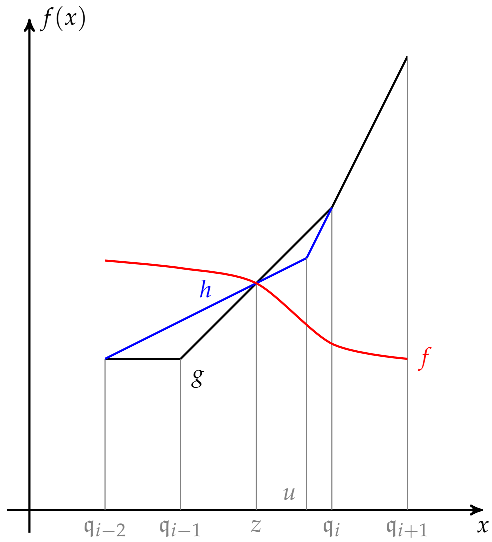

Figure 2.12: Case (IV) in Lemma 2.10. The new function $h \in \mathscr{L}$ is linear on $[\mathfrak{q}_{i-2}, u]$ with slope $\frac{g(z)-g(\mathfrak{q}_{i-2})}{z-\mathfrak{q}_{i-2}} \geq \mathbf{a}$, linear on $[u, \mathfrak{q}_i]$ with slope $A_{i+1}(g)$, and agrees with g outside of $[\mathfrak{q}_{i-2}, \mathfrak{q}_i]$.

(cf. Figure 2.12). In the following we assume without loss of generality that $A_{i-1}(g) < A_i(g) < A_{i+1}(g)$ (cf. Lemma 2.7). Note that the fact $\frac{g(\mathfrak{q}_{i-2})-g(z)}{\mathfrak{q}_{i-2}-z} < A_i(g)$ and the intermediate value theorem imply that there exists $u \in (z, \mathfrak{q}_i)$ which satisfies $g(z) + \big[\frac{g(\mathfrak{q}_{i-2})-g(z)}{\mathfrak{q}_{i-2}-z}\big](u-z) = g(q_i) + A_{i+1}(g)(u - \mathfrak{q}_i)$. Let $h \in \mathscr{L}$ satisfy for all $x \in [0, \mathfrak{q}_{i-2}] \cup [\mathfrak{q}_i, 1]$, $y_1 \in [\mathfrak{q}_{i-2}, u]$, $y_2 \in [u, \mathfrak{q}_i]$ that $h(x) = g(x)$, $h(y_1) = g(\mathfrak{q}_{i-2}) + \big[\frac{g(\mathfrak{q}_{i-2})-g(z)}{\mathfrak{q}_{i-2}-z}\big](y_1 - \mathfrak{q}_{i-2})$, and $h(y_2) = g(\mathfrak{q}_i) + A_{i+1}(g)(y_2 - \mathfrak{q}_i)$. Observe that $Q(h) < Q(g)$ and $\int_0^1 (h(y) - f(y))^2 \,\mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \,\mu(\mathrm{d}y)$. This establishes (2.74) in the case $\big(\max\{A_{i-1}(g), A_{i+1}(g)\} > A_i(g) > \min\{A_{i-1}(g), A_{i+1}(g)\}\big) \wedge \big(\min\{\frac{g(\mathfrak{q}_{i+1})-g(z)}{\mathfrak{q}_{i+1}-z}, \frac{g(\mathfrak{q}_{i-2})-g(z)}{\mathfrak{q}_{i-2}-z}\} \geq \mathbf{a}\big)$.

(V) Finally, we prove (2.74) in the case

$$\big(\max\{A_{i-1}(g), A_{i+1}(g)\} > A_i(g) > \min\{A_{i-1}(g), A_{i+1}(g)\}\big) \wedge \big(\min\{\tfrac{g(\mathfrak{q}_{i+1})-g(z)}{\mathfrak{q}_{i+1}-z}, \tfrac{g(\mathfrak{q}_{i-2})-g(z)}{\mathfrak{q}_{i-2}-z}\} < \mathbf{a}\big) \tag{2.80}$$

(cf. Figure 2.13). In the following we assume without loss of generality that

$$A_{i-1}(g) < A_i(g) < A_{i+1}(g) \tag{2.81}$$

(cf. Lemma 2.7). Note that (2.80) and (2.81) show that $\frac{g(\mathfrak{q}_{i-2})-g(z)}{\mathfrak{q}_{i-2}-z} = \min\{\frac{g(\mathfrak{q}_{i+1})-g(z)}{\mathfrak{q}_{i+1}-z}, \frac{g(\mathfrak{q}_{i-2})-g(z)}{\mathfrak{q}_{i-2}-z}\} < \mathbf{a}$. The intermediate value theorem therefore implies that there exist $u \in (\mathfrak{q}_{i-2}, \mathfrak{q}_{i-1})$, $v \in (z, \mathfrak{q}_i)$ which satisfy $\frac{g(u)-g(z)}{u-z} = \mathbf{a}$ and $g(z) + \mathbf{a}(v - z) = g(\mathfrak{q}_i) +$

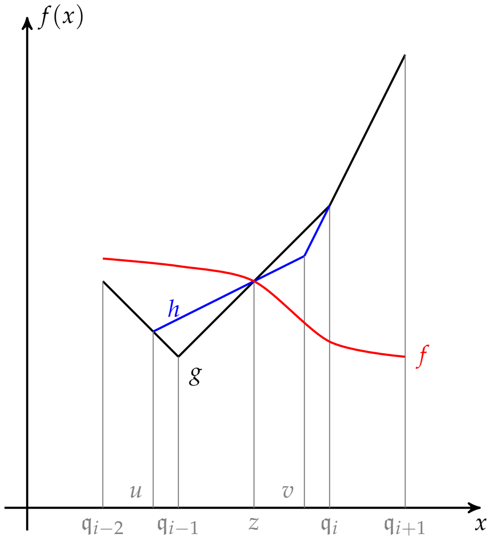

Figure 2.13: Case (V) in Lemma 2.10. The new function $h \in \mathscr{L}$ is linear on $[u, v]$ with slope $\mathbf{a}$, linear on $[v, \mathfrak{q}_i]$ with slope $A_{i+1}(g)$, and agrees with g outside of $[v, \mathfrak{q}_i]$.

$A_{i+1}(g)(v - \mathfrak{q}_i)$. Let $h \in \mathscr{L}$ satisfy for all $x \in [0, u] \cup [\mathfrak{q}_i, 1]$, $y_1 \in [u, v]$, $y_2 \in [v, \mathfrak{q}_i]$ that $h(x) = g(x)$, $h(y_1) = g(z) + \mathbf{a}(y_1 - z)$, and $h(y_2) = g(\mathfrak{q}_i) + A_{i+1}(g)(y_2 - \mathfrak{q}_i)$. Observe that $Q(h) = Q(g)$, $A_i(h) = \mathbf{a}$, and $\forall\, j \in \{1, 2, \ldots, Q(g) + 1\} \backslash \{i\} \colon A_j(h) = A_j(g)$. Moreover, note that the fact that $\forall\, x, y \in [0, 1] \colon |f(x) - f(y)| \leq L|x - y|$ and the fact that $\mathbf{a} < A_i(g)$ demonstrate that $\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$. This establishes (2.74) in the case $\big(\max\{A_{i-1}(g), A_{i+1}(g)\} > A_i(g) > \min\{A_{i-1}(g), A_{i+1}(g)\}\big) \wedge \big(\min\{\frac{g(\mathfrak{q}_{i+1}) - g(z)}{\mathfrak{q}_{i+1} - z}, \frac{g(\mathfrak{q}_{i-2}) - g(z)}{\mathfrak{q}_{i-2} - z}\} < \mathbf{a}\big)$.

The proof of Lemma 2.10 is thus complete. □

Next, we summarize Lemmas 2.8, 2.9, and 2.10 in the following corollary.

Corollary 2.3. *Let $L \in (0, \infty)$, $f \in C([0, 1], \mathbb{R})$ satisfy for all $x, y \in [0, 1]$ that $|f(x) - f(y)| \leq L|x - y|$, let $g \in \mathscr{L}$, $i \in \{1, 2, \ldots, Q(g) + 1\}$, $\mathbf{a} \in \mathbb{R}$ satisfy $L \leq |\mathbf{a}| \leq |A_i(g)|$ and $\mathbf{a}A_i(g) > 0$, and let $\mu \colon \mathcal{B}([0, 1]) \to [0, \infty]$ be a finite measure (cf. Definitions 2.1, 2.2, and 2.3). Then there exists $h \in \mathscr{L}$ such that $\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$, $Q(h) \leq Q(g)$, and*

$$(Q(g) - Q(h) - 1)\left(\textstyle\sum_{j=1}^{Q(h)+1} |A_j(h) - A_j(g) \mathbb{1}_{\mathbb{N} \backslash \{i\}}(j) - \mathbf{a} \mathbb{1}_{\{i\}}(j)|\right) \geq 0. \tag{2.82}$$

Proof of Corollary 2.3. Throughout this proof assume without loss of generality[4] that $L \leq \mathbf{a} \leq A_i(g)$. Observe that Lemma 2.8 establishes (2.82) in the case

$$\big[\forall\, x \in (q_{i-1}(g), q_i(g)) \colon f(x) \neq g(x)\big]. \tag{2.83}$$

[4]Otherwise we consider $f \curvearrowleft -f$, $g \curvearrowleft -g$, $\mathbf{a} \curvearrowleft -\mathbf{a}$.

Furthermore, note that Lemma 2.9 establishes (2.82) in the case

$$\big[(\exists\, z \in (q_{i-1}(g), q_i(g))\colon f(z) = g(z)) \wedge (i \in \{1, Q(g)+1\})\big]. \tag{2.84}$$

Moreover, observe that Lemma 2.10 establishes (2.82) in the case

$$\big[(\exists\, z \in (q_{i-1}(g), q_i(g))\colon f(z) = g(z)) \wedge (i \notin \{1, Q(g)+1\})\big]. \tag{2.85}$$

The proof of Corollary 2.3 is thus complete. □

The following two results are a consequence of Corollary 2.4 and induction.

Corollary 2.4. *Let $L \in (0,\infty)$, $f \in C([0,1],\mathbb{R})$ satisfy for all $x, y \in [0,1]$ that $|f(x) - f(y)| \leq L|x-y|$, let $g \in \mathscr{L}$, let $\mathbb{A} \subseteq \{1, 2, \ldots, Q(g)+1\}$ be a set, let $\mathbf{a} = (\mathbf{a}_j)_{j\in\mathbb{A}}\colon \mathbb{A} \to \mathbb{R}$ satisfy for all $j \in \mathbb{A}$ that $L \leq |\mathbf{a}_j| \leq |A_j(g)|$ and $\mathbf{a}_j A_j(g) > 0$, and let $\mu\colon \mathcal{B}([0,1]) \to [0,\infty]$ be a finite measure (cf. Definitions 2.1, 2.2, and 2.3). Then there exists $h \in \mathscr{L}$ such that $\int_0^1 (h(y) - f(y))^2\, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2\, \mu(\mathrm{d}y)$, $Q(h) \leq Q(g)$, and*

$$(Q(g) - Q(h) - 1)\,\big(\textstyle\sum_{j=1}^{Q(h)+1} |A_j(h) - A_j(g)\mathbb{1}_{\mathbb{N}\setminus\mathbb{A}}(j) - \mathbf{a}_j \mathbb{1}_{\mathbb{A}}(j)|\big) \geq 0. \tag{2.86}$$

Proof of Corollary 2.4. Note that induction and Corollary 2.3 establish (2.86). The proof of Corollary 2.4 is thus complete. □

Lemma 2.11. *Let $f \in C([0,1],\mathbb{R})$ and let $\mu\colon \mathcal{B}([0,1]) \to [0,\infty]$ be a finite measure. Then for all $g \in \mathscr{L}$ there exists $h \in \mathscr{L}$ such that*

$$Q(h) \leq Q(g), \quad \mathrm{Lip}(h) \leq \mathrm{Lip}(f), \tag{2.87a}$$

$$\int_0^1 (h(y) - f(y))^2\, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2\, \mu(\mathrm{d}y) \tag{2.87b}$$

(cf. Definitions 2.1, 2.2, 2.3, and 2.4).

Proof of Lemma 2.11. Throughout this proof assume without loss of generality that $0 < \mathrm{Lip}(f) < \infty$ and let $\mathbb{Q}\colon \mathscr{L} \to \mathbb{N}_0$ satisfy for all $g \in \mathscr{L}$ that

$$\mathbb{Q}(g) = (Q(g)+1)^2 + \#(\{i \in \{1, 2, \ldots, Q(g)+1\}\colon |A_i(g)| > \mathrm{Lip}(f)\}). \tag{2.88}$$

Observe that (2.88) assures for all $g_1, g_2 \in \mathscr{L}$ with $Q(g_1) < Q(g_2)$ that

$$\begin{aligned} \mathbb{Q}(g_1) &\leq (Q(g_1)+1)^2 + Q(g_1) + 1 < (Q(g_1)+1)^2 + 2(Q(g_1)+1) + 1 \\ &= (Q(g_1)+2)^2 \leq (Q(g_2)+1)^2 \leq \mathbb{Q}(g_2). \end{aligned} \tag{2.89}$$

Next we claim that for all $k \in \mathbb{N}_0$, $g \in \mathbb{Q}^{-1}(\{k\})$ there exists $h \in \mathscr{L}$ such that

$$Q(h) \leq Q(g), \quad \mathrm{Lip}(h) \leq \mathrm{Lip}(f), \tag{2.90a}$$

$$\int_0^1 (h(y) - f(y))^2\, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2\, \mu(\mathrm{d}y). \tag{2.90b}$$

We now prove (2.90) by induction on $k \in \mathbb{N}_0$. For the base case $k = 0$ we note that $\mathbb{Q}^{-1}(\{0\}) = \varnothing$. This establishes (2.90) in the base case $k = 0$. For the induction step let $k \in \mathbb{N}_0$ satisfy for all $g \in \mathbb{Q}^{-1}(\{0, 1, \ldots, k\})$ that there exists $h \in \mathscr{L}$ such that

$$Q(h) \leq Q(g), \quad \mathrm{Lip}(h) \leq \mathrm{Lip}(f), \tag{2.91a}$$

$$\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y) \tag{2.91b}$$

and let $g \in \mathbb{Q}^{-1}(\{k+1\})$ satisfy

$$\mathrm{Lip}(g) > \mathrm{Lip}(f). \tag{2.92}$$

Observe that Lemma 2.2 and (2.92) ensure that there exists $i \in \{1, 2, \ldots, Q(g) + 1\}$ which satisfies

$$|A_i(g)| > \mathrm{Lip}(f). \tag{2.93}$$

Note that (2.93) shows that there exists $\mathbf{a} \in \mathbb{R}$ which satisfies

$$\mathrm{Lip}(f) = |\mathbf{a}| \leq |A_i(g)| \qquad \text{and} \qquad \mathbf{a} A_i(g) > 0. \tag{2.94}$$

Observe that (2.94), the fact that $\mathrm{Lip}(f) \in (0, \infty)$, and Corollary 2.3 demonstrate that there exists $\mathfrak{g} \in \mathscr{L}$ which satisfies $\int_0^1 (\mathfrak{g}(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$, $Q(\mathfrak{g}) \leq Q(g)$, and

$$(Q(g) - Q(\mathfrak{g}) - 1) \left(\textstyle\sum_{j=1}^{Q(\mathfrak{g})+1} |A_j(\mathfrak{g}) - A_j(g) \mathbb{1}_{\mathbb{N} \setminus \{i\}}(j) - \mathbf{a} \mathbb{1}_{\{i\}}(j)| \right) \geq 0. \tag{2.95}$$

Note that (2.89) and (2.95) assure that $\mathbb{Q}(\mathfrak{g}) < \mathbb{Q}(g) = k + 1$. Hence, we obtain $\mathfrak{g} \in \mathbb{Q}^{-1}(\{0, 1, \ldots, k\})$. Combining this with (2.91) and (2.95) demonstrates that there exists $h \in \mathscr{L}$ such that $Q(h) \leq Q(\mathfrak{g}) \leq Q(g)$, $\mathrm{Lip}(h) \leq \mathrm{Lip}(f)$, and

$$\textstyle\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (\mathfrak{g}(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y). \tag{2.96}$$

Induction thus establishes (2.90). Observe that (2.90) implies (2.87). The proof of Lemma 2.11 is thus complete. □

Lemma 2.11 is not yet sufficient to establish Proposition 2.2 since, as mentioned before, not every piecewise linear function with at most $\mathfrak{h} \in \mathbb{N}$ breakpoints is representable by an ANN with $\mathfrak{h}$ hidden neurons. Thus we need to ensure that the linear relation for the slopes (cf. Corollary 2.2) is also preserved by our inductive construction. This is the content of Lemma 2.12, which is again a consequence of Corollary 2.3 and induction.

Lemma 2.12. *Let $f\colon [0,1] \to \mathbb{R}$ be Lipschitz continuous and let $\mu\colon \mathcal{B}([0,1]) \to [0, \infty]$ be a finite measure. Then for all $g \in \mathscr{L}$, $k \in \mathbb{N}$, $i_1, i_2, \ldots, i_k \in \mathbb{N}$ with $\frac{k}{2} \notin \mathbb{N}$, $i_1 < i_2 < \cdots < i_k \leq Q(g) + 1$, and $\sum_{j=1}^{k} (-1)^j A_{i_j}(g) = 0$ there exists $h \in \mathscr{L}$ such that $\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$, $Q(h) \leq Q(g)$, and*

$$(Q(g) - Q(h) - 1) \left(\left| \textstyle\sum_{j=1}^{k} (-1)^j A_{\min\{i_j, Q(h)+1\}}(h) \right| + \max\{\mathrm{Lip}(h) - Q(g) \, \mathrm{Lip}(f), 0\} \right) \geq 0 \tag{2.97}$$

(cf. Definitions 2.1, 2.2, and 2.3).

Proof of Lemma 2.12. Throughout this proof assume without loss of generality that $\operatorname{Lip}(f) > 0$, let $\operatorname{sgn}\colon \mathbb{R} \to \mathbb{R}$ satisfy for all $x \in (0,\infty)$ that $\operatorname{sgn}(x) = 1$, $\operatorname{sgn}(-x) = -1$, and $\operatorname{sgn}(0) = 0$, and let $\mathbb{Q}\colon \mathscr{L} \to \mathbb{N}_0$ satisfy for all $g \in \mathscr{L}$ that

$$\mathbb{Q}(g) = (Q(g)+1)^2 + \#\big(\{i \in \{1,2,\dots,Q(g)+1\}\colon |A_i(g)| > Q(g)\operatorname{Lip}(f)\}\big) \tag{2.98}$$

(cf. Definitions 2.1, 2.2, and 2.3). Note that (2.98) assures for all $g_1, g_2 \in \mathscr{L}$ with $Q(g_1) < Q(g_2)$ that

$$\begin{aligned}\mathbb{Q}(g_1) &\le (Q(g_1)+1)^2 + Q(g_1) + 1 < (Q(g_1)+1)^2 + 2(Q(g_1)+1) + 1\\ &= (Q(g_1)+2)^2 \le (Q(g_2)+1)^2 \le \mathbb{Q}(g_2).\end{aligned} \tag{2.99}$$

We claim that for all $n \in \mathbb{N}_0$, $g \in \mathbb{Q}^{-1}(\{n\})$, $k \in \mathbb{N}$, $i_1, i_2, \dots, i_k \in \mathbb{N}$ with $\frac{k}{2} \notin \mathbb{N}$, $i_1 < i_2 < \dots < i_k \le Q(g)+1$, and $\sum_{j=1}^k (-1)^j A_{i_j}(g) = 0$ there exists $h \in \mathscr{L}$ such that $\int_0^1 (h(y)-f(y))^2\,\mu(\mathrm{d}y) \le \int_0^1 (g(y)-f(y))^2\,\mu(\mathrm{d}y)$, $Q(h) \le Q(g)$, and

$$(Q(g)-Q(h)-1)\big(\big|\textstyle\sum_{j=1}^k (-1)^j A_{\min\{i_j, Q(h)+1\}}(h)\big| + \max\{\operatorname{Lip}(h) - Q(g)\operatorname{Lip}(f), 0\}\big) \ge 0. \tag{2.100}$$

We now prove (2.100) by induction on $n \in \mathbb{N}_0$. For the base case $n = 0$ observe that $\mathbb{Q}^{-1}(\{0\}) = \varnothing$. This establishes (2.100) in the base case $n = 0$. For the induction step let $n \in \mathbb{N}_0$ satisfy for all $g \in \mathbb{Q}^{-1}(\{0,1,\dots,n\})$, $k \in \mathbb{N}$, $i_1, i_2, \dots, i_k \in \mathbb{N}$ with $\frac{k}{2} \notin \mathbb{N}$, $i_1 < i_2 < \dots < i_k \le Q(g)+1$, and $\sum_{j=1}^k (-1)^j A_{i_j}(g) = 0$ that there exists $h \in \mathscr{L}$ such that $\int_0^1 (h(y)-f(y))^2\,\mu(\mathrm{d}y) \le \int_0^1 (g(y)-f(y))^2\,\mu(\mathrm{d}y)$, $Q(h) \le Q(g)$, and

$$(Q(g)-Q(h)-1)\big(\big|\textstyle\sum_{j=1}^k (-1)^j A_{\min\{i_j, Q(h)+1\}}(h)\big| + \max\{\operatorname{Lip}(h) - Q(g)\operatorname{Lip}(f), 0\}\big) \ge 0, \tag{2.101}$$

and[5] let $g \in \mathbb{Q}^{-1}(\{n+1\})$, $k \in \mathbb{N}$, $i_1, i_2, \dots, i_k \in \mathbb{N}$ satisfy

$$\tfrac{k}{2} \notin \mathbb{N}, \qquad i_1 < i_2 < \dots < i_k \le Q(g)+1, \qquad \textstyle\sum_{j=1}^k (-1)^j A_{i_j}(g) = 0, \tag{2.102}$$

and $\operatorname{Lip}(g) > Q(g)\operatorname{Lip}(f)$. We now prove that there exists $h \in \mathscr{L}$ such that $\int_0^1 (h(y) - f(y))^2\,\mu(\mathrm{d}y) \le \int_0^1 (g(y)-f(y))^2\,\mu(\mathrm{d}y)$, $Q(h) \le Q(g)$, and

$$(Q(g)-Q(h)-1)\big(\big|\textstyle\sum_{j=1}^k (-1)^j A_{\min\{i_j, Q(h)+1\}}(h)\big| + \max\{\operatorname{Lip}(h) - Q(g)\operatorname{Lip}(f), 0\}\big) \ge 0. \tag{2.103}$$

Observe that Lemma 2.2 and the fact that $\operatorname{Lip}(g) > Q(g)\operatorname{Lip}(f)$ ensure that there exist $\mathfrak{I} \in \{1,2,\dots,Q(g)+1\}$, $s \in \{-1,1\}$ which satisfy

$$sA_{\mathfrak{I}}(g) = |A_{\mathfrak{I}}(g)| > Q(g)\operatorname{Lip}(f). \tag{2.104}$$

In the following we distinguish between the case $\mathfrak{I} \notin \{i_1, i_2, \dots, i_k\}$ and the case $\mathfrak{I} \in \{i_1, i_2, \dots, i_k\}$. We first prove (2.103) in the case

$$\mathfrak{I} \notin \{i_1, i_2, \dots, i_k\}. \tag{2.105}$$

[5] Note that we could choose $h = g$ in (2.103) if we would have $\operatorname{Lip}(g) \le Q(g)\operatorname{Lip}(f)$.

Note that (2.104) and Corollary 2.3 assure that there exists $\mathfrak{g} \in \mathscr{L}$ which satisfies $\int_0^1 (\mathfrak{g}(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$, $Q(\mathfrak{g}) \leq Q(g)$, and

$$(Q(g) - Q(\mathfrak{g}) - 1) \left(\sum_{j=1}^{Q(\mathfrak{g})+1} |A_j(\mathfrak{g}) - A_j(g) \mathbb{1}_{\mathbb{N} \setminus \{\mathfrak{I}\}}(j) - sQ(g)\mathrm{Lip}(f) \mathbb{1}_{\{\mathfrak{I}\}}(j)| \right) \geq 0. \tag{2.106}$$

Moreover, observe that (2.102) ensures that $\mathbb{Q}(g) = n + 1$. Combining this with (2.98), (2.99), and (2.104) demonstrates that $\mathbb{Q}(\mathfrak{g}) < \mathbb{Q}(g) = n + 1$. Therefore, we obtain

$$\mathfrak{g} \in \mathbb{Q}^{-1}(\{0, 1, \ldots, n\}). \tag{2.107}$$

In addition, note that (2.102), (2.105), and (2.106) show that $Q(\mathfrak{g}) \leq Q(g)$, $\int_0^1 (\mathfrak{g}(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$, and

$$(Q(g) - Q(\mathfrak{g}) - 1) \left| \sum_{j=1}^{k} (-1)^j A_{\min\{i_j, Q(\mathfrak{g})+1\}}(\mathfrak{g}) \right| \geq 0. \tag{2.108}$$

Combining (2.101), (2.102), and (2.107) hence[6] establishes (2.103) in the case $\mathfrak{I} \notin \{i_1, i_2, \ldots, i_k\}$. In the next step we prove (2.103) in the case

$$\mathfrak{I} \in \{i_1, i_2, \ldots, i_k\}. \tag{2.109}$$

Note that (2.109) demonstrates that there exist $\mathscr{J} \in \{1, 2, \ldots, k\}$, $\mathfrak{S} \in \{-1, 1\}$ which satisfy

$$i_{\mathscr{J}} = \mathfrak{I} \qquad \text{and} \qquad \mathfrak{S} = s(-1)^{\mathscr{J}} = \mathrm{sgn}((-1)^{\mathscr{J}} A_{i_{\mathscr{J}}}) = \mathrm{sgn}((-1)^{\mathscr{J}} A_{\mathfrak{I}}). \tag{2.110}$$

In the following let $\alpha_v \in \mathbb{R}$, $v \in \{-1, 1\}$, satisfy for all $v \in \{-1, 1\}$ that

$$\alpha_v = \sum_{j \in \mathbb{N} \cap [1,k], \, \mathrm{sgn}((-1)^j A_{i_j}(g)) = v\mathfrak{S}} |A_{i_j}(g)|. \tag{2.111}$$

Observe that (2.102) and (2.111) ensure that

$$\begin{aligned}
&\mathfrak{S}\,(\alpha_1 - \alpha_{-1}) = \mathfrak{S}\alpha_1 - \mathfrak{S}\alpha_{-1} = \sum_{v \in \{-1,1\}} [v\mathfrak{S}\alpha_v] \\
&= \sum_{v \in \{-1,1\}} \sum_{j \in \mathbb{N} \cap [1,k], \, \mathrm{sgn}((-1)^j A_{i_j}(g)) = v\mathfrak{S}} \left[v\mathfrak{S} |A_{i_j}(g)| \right] \\
&= \sum_{v \in \{-1,1\}} \sum_{j \in \mathbb{N} \cap [1,k], \, \mathrm{sgn}((-1)^j A_{i_j}(g)) = v\mathfrak{S}} \left[v\mathfrak{S} |(-1)^j A_{i_j}(g)| \right] \\
&= \sum_{v \in \{-1,1\}} \sum_{j \in \mathbb{N} \cap [1,k], \, \mathrm{sgn}((-1)^j A_{i_j}(g)) = v\mathfrak{S}} \left[(-1)^j A_{i_j}(g) \right] \\
&= \sum_{j \in \mathbb{N} \cap [1,k]} (-1)^j A_{i_j}(g) = 0.
\end{aligned} \tag{2.112}$$

Hence, we obtain $\alpha_1 = \alpha_{-1}$. Next note that (2.110) and (2.111) assure that $\alpha_1 \geq |A_{\mathfrak{I}}(g)|$. Combining this with (2.104) and the fact that $\alpha_{-1} = \alpha_1$ demonstrates that

$$\alpha_{-1} \geq \left(|A_{\mathfrak{I}}(g)| - Q(g)\mathrm{Lip}(f) \right) + Q(g)\mathrm{Lip}(f) > Q(g)\mathrm{Lip}(f). \tag{2.113}$$

[6]Observe that we can choose $h = \mathfrak{g}$ in (2.103) in the case where $Q(\mathfrak{g}) < Q(g)$.

Therefore, we obtain that there exist $l \in \mathbb{N}$, $J_1, J_2, \ldots, J_l \in \mathbb{N}$, $r_1, r_2, \ldots, r_l \in [0, \infty)$ which satisfy for all $v \in \{1, 2, \ldots, l\}$ that

$$J_1 < J_2 < \cdots < J_l \leq k, \qquad \operatorname{sgn}((-1)^{J_v} A_{i_{J_v}}(g)) = -\mathfrak{S}, \qquad |A_{i_{J_v}}(g)| - r_v \geq \operatorname{Lip}(f), \tag{2.114}$$

and $r_1 + r_2 + \ldots + r_l = |A_{\mathfrak{I}}(g)| - Q(g)\operatorname{Lip}(f)$. In the following let $\mathbb{A} \subseteq \{1, 2, \ldots, Q(g) + 1\}$ satisfy

$$\mathbb{A} = \{i_{J_1}, i_{J_2}, \ldots, i_{J_l}\} \cup \{\mathfrak{I}\} \tag{2.115}$$

and let $\mathbf{a} = (\mathbf{a}_j)_{j \in \mathbb{A}} \colon \mathbb{A} \to \mathbb{R}$ satisfy for all $v \in \{1, 2, \ldots, l\}$ that

$$\mathbf{a}_{i_{J_v}} = \Big(|A_{i_{J_v}}| - r_v\Big) \operatorname{sgn}(A_{i_{J_v}}) \qquad \text{and} \qquad \mathbf{a}_{\mathfrak{I}} = sQ(g)\operatorname{Lip}(f). \tag{2.116}$$

Note that (2.114) and (2.116) ensure for all $v \in \{1, 2, \ldots, l\}$ that

$$\begin{aligned} \mathbf{a}_{i_{J_v}} A_{i_{J_v}}(g) &= \Big(|A_{i_{J_v}}(g)| - r_v\Big) \operatorname{sgn}(A_{i_{J_v}}(g)) A_{i_{J_v}}(g) = \Big(|A_{i_{J_v}}(g)| - r_v\Big) |A_{i_{J_v}(g)}| \\ &\geq \operatorname{Lip}(f)|A_{i_{J_v}}(g)| \geq \operatorname{Lip}(f)(|A_{i_{J_v}}(g)| - r_v) \geq [\operatorname{Lip}(f)]^2 > 0. \end{aligned} \tag{2.117}$$

Next observe that (2.104) implies that $|A_{\mathfrak{I}}(g)| > Q(g)\operatorname{Lip}(f) \geq 0$. Hence, we obtain $A_{\mathfrak{I}}(g) \neq 0$. This and (2.102) prove that $Q(g) > 0$. Combining (2.104), (2.114), and (2.116) therefore shows that

$$\begin{aligned} \mathbf{a}_{\mathfrak{I}} A_{\mathfrak{I}}(g) &= \big[sQ(g)\operatorname{Lip}(f)\big]\big[s^{-1}|A_{\mathfrak{I}}(g)|\big] = |A_{\mathfrak{I}}(g)|Q(g)\operatorname{Lip}(f) \\ &> [Q(g)\operatorname{Lip}(f)]^2 > 0. \end{aligned} \tag{2.118}$$

This and (2.117) assure for all $j \in \mathbb{A}$ that

$$\mathbf{a}_j A_j(g) > 0. \tag{2.119}$$

Furthermore, note that (2.104), (2.114), (2.116), and the fact that $Q(g) \geq 1$ demonstrate for all $v \in \{1, 2, \ldots, l\}$ that

$$\operatorname{Lip}(f) \leq |A_{i_{J_v}}| - r_v = |\mathbf{a}_{i_{J_v}}| \leq |A_{i_{J_v}}|, \tag{2.120a}$$

$$\operatorname{Lip}(f) \leq Q(g)\operatorname{Lip}(f) = |\mathbf{a}_{\mathfrak{I}}| < |A_{\mathfrak{I}}(g)|. \tag{2.120b}$$

Therefore, we obtain for all $j \in \mathbb{A}$ that $\operatorname{Lip}(f) \leq |\mathbf{a}_j| \leq A_j(g)$. Combining this with (2.119) enables us to apply Corollary 2.4 to obtain that there exists $\mathfrak{g} \in \mathscr{L}$ which satisfies $\int_0^1 (\mathfrak{g}(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$, $Q(\mathfrak{g}) \leq Q(g)$, and

$$(Q(g) - Q(\mathfrak{g}) - 1)\left(\textstyle\sum_{j=1}^{Q(\mathfrak{g})+1} |A_j(\mathfrak{g}) - A_j(g)\mathbb{1}_{\mathbb{N}\setminus\mathbb{A}}(j) - \mathbf{a}_j\mathbb{1}_{\mathbb{A}}(j)|\right) \geq 0. \tag{2.121}$$

Observe that (2.98), (2.99), (2.104), (2.116), and (2.121) show that $\mathbb{Q}(\mathfrak{g}) < \mathbb{Q}(g)$. This and (2.102) show that $\mathbb{Q}(\mathfrak{g}) \leq \mathbb{Q}(g) - 1 = (n+1) - 1 = n$. Hence, we obtain

$$\mathfrak{g} \in \mathbb{Q}^{-1}(\{0, 1, \ldots, n\}). \tag{2.122}$$

Moreover, note that (2.102), (2.116), (2.121), and the fact that $r_1 + r_2 + \ldots + r_l = |A_{\mathfrak{J}}(g)| - Q(g)\operatorname{Lip}(f)$ assure that $\int_0^1 (\mathfrak{g}(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y)$, $Q(\mathfrak{g}) \leq Q(g)$, and

$$(Q(g) - Q(\mathfrak{g}) - 1) \, |\textstyle\sum_{j=1}^{k} (-1)^j A_{\min\{i_j, Q(\mathfrak{g})+1\}}(\mathfrak{g})| \geq 0. \tag{2.123}$$

Combining (2.101), (2.102), and (2.108) hence[7] establishes (2.103) in the case $\mathfrak{I} \in \{i_1, i_2, \ldots, i_k\}$. Induction thus proves (2.100). Note that (2.100) establishes (2.97). The proof of Lemma 2.12 is thus complete. □

2.5 Structure preserving approximations for realization functions of shallow ANNs

In this subsection we employ Lemma 2.11 and Lemma 2.12 above to prove in Proposition 2.2 the announced result about the existence of a better ANN approximation which is additionally Lipschitz with a constant depending only on the width $\mathfrak{h} \in \mathbb{N}$ and the target function f.

Lemma 2.13. *Assume Setting 2.1 and let $\theta \in \mathbb{R}^{\mathfrak{d}}$ satisfy $(\mathfrak{h} - Q(\mathscr{N}^{\theta}) - 1)\max\{\operatorname{Lip}(\mathscr{N}^{\theta}) - \mathfrak{h}L, 0\} \geq 0$ (cf. Definitions 2.1 and 2.2). Then there exists $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ such that*

$$\mathcal{L}(\vartheta) \leq \mathcal{L}(\theta), \qquad Q(\mathscr{N}^{\vartheta}) \leq Q(\mathscr{N}^{\theta}), \qquad \text{and} \qquad \operatorname{Lip}(\mathscr{N}^{\vartheta}) \leq \mathfrak{h}L \tag{2.124}$$

(cf. Definition 2.4).

Proof of Lemma 2.13. In the following we distinguish between the case $Q(\mathscr{N}^{\theta}) = \mathfrak{h}$ and the case $Q(\mathscr{N}^{\theta}) < \mathfrak{h}$. We first prove (2.124) in the case

$$Q(\mathscr{N}^{\theta}) = \mathfrak{h}. \tag{2.125}$$

Observe that (2.125) and the assumption that $(\mathfrak{h} - Q(\mathscr{N}^{\theta}) - 1)\max\{\operatorname{Lip}(\mathscr{N}^{\theta}) - \mathfrak{h}L, 0\} \geq 0$ ensure that $-\max\{\operatorname{Lip}(\mathscr{N}^{\theta}) - \mathfrak{h}L, 0\} \geq 0$. Therefore, we obtain $\operatorname{Lip}(\mathscr{N}^{\theta}) \leq \mathfrak{h}L$. This establishes (2.124) in the case $Q(\mathscr{N}^{\theta}) = \mathfrak{h}$. In the next step we prove (2.124) in the case

$$Q(\mathscr{N}^{\theta}) < \mathfrak{h}. \tag{2.126}$$

Note that (2.126) and Lemma 2.11 prove that there exists $h \in \mathscr{L}$ which satisfies $Q(h) \leq Q(\mathscr{N}^{\theta}) \leq \mathfrak{h} - 1$, $\operatorname{Lip}(h) \leq L \leq \mathfrak{h}L$, and

$$\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (\mathscr{N}^{\theta}(y) - f(y))^2 \, \mu(\mathrm{d}y) = \mathcal{L}(\theta). \tag{2.127}$$

Observe that Lemma 2.4 and the fact that $Q(h) \leq \mathfrak{h} - 1$ show that there exists $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ which satisfies

$$\mathscr{N}^{\vartheta} = h. \tag{2.128}$$

Note that (2.127) and (2.128) ensure that $\mathcal{L}(\vartheta) \leq \mathcal{L}(\theta)$, $Q(\mathscr{N}^{\vartheta}) = Q(h) \leq Q(\mathscr{N}^{\theta})$, and $\operatorname{Lip}(\mathscr{N}^{\vartheta}) = \operatorname{Lip}(h) \leq \mathfrak{h}L$. This establishes (2.129) in the case $Q(\mathscr{N}^{\theta}) < \mathfrak{h}$. The proof of Lemma 2.13 is thus complete. □

[7] Observe that we can choose $h = \mathfrak{g}$ in (2.103) in the case where $Q(\mathfrak{g}) < Q(g)$.

Proposition 2.2. *Assume Setting 2.1 and let $\theta \in \mathbb{R}^{\mathfrak{d}}$. Then there exists $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ such that*

$$\mathcal{L}(\vartheta) \leq \mathcal{L}(\theta), \qquad Q(\mathscr{N}^{\vartheta}) \leq Q(\mathscr{N}^{\theta}), \qquad \textit{and} \qquad \operatorname{Lip}(\mathscr{N}^{\vartheta}) \leq \mathfrak{h}L \tag{2.129}$$

(cf. Definitions 2.1 and 2.4).

Proof of Proposition 2.2. Observe that Lemma 2.3 proves that $\mathscr{N}^{\theta} \in \mathscr{L}$ and $Q(\mathscr{N}^{\theta}) \leq \mathfrak{h}$ (cf. Definition 2.2). In the following we distinguish between the case $Q(\mathscr{N}^{\theta}) < \mathfrak{h}$ and the case $Q(\mathscr{N}^{\theta}) = \mathfrak{h}$. We first prove (2.129) in the case

$$Q(\mathscr{N}^{\theta}) < \mathfrak{h}. \tag{2.130}$$

Note that (2.130) and Lemma 2.13 show that there exists $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ such that $\mathcal{L}(\vartheta) \leq \mathcal{L}(\theta)$, $Q(\mathscr{N}^{\vartheta}) \leq Q(\mathscr{N}^{\theta})$, and $\operatorname{Lip}(\mathscr{N}^{\vartheta}) \leq \mathfrak{h}L$. This establishes (2.129) in the case $Q(\mathscr{N}^{\theta}) < \mathfrak{h}$. In the next step we prove (2.129) in the case

$$Q(\mathscr{N}^{\theta}) = \mathfrak{h}. \tag{2.131}$$

Observe that (2.131) and Corollary 2.2 imply that there exists $k \in \mathbb{N}$, $i_1, i_2, \ldots, i_k \in \mathbb{N}$ which satisfy $\frac{k}{2} \notin \mathbb{N}$, $1 \leq i_1 < i_2 < \cdots < i_k \leq \mathfrak{h} + 1$, and $\sum_{j=1}^{k}(-1)^j A_{i_j}(\mathscr{N}^{\theta})$ (cf. Definition 2.3). Combining this with Lemma 2.12 ensures that there exists $h \in \mathscr{L}$ which satisfies

$$\textstyle\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \int_0^1 (\mathscr{N}^{\theta}(y) - f(y))^2 \, \mu(\mathrm{d}y) = \mathcal{L}(\theta), \qquad Q(h) \leq Q(\mathscr{N}^{\theta}), \tag{2.132}$$

and

$$(Q(\mathscr{N}^{\theta}) - Q(h) - 1)\big(\big|\textstyle\sum_{j=1}^{k}(-1)^j A_{i_j}(h)\big| + \max\{\operatorname{Lip}(h) - Q(\mathscr{N}^{\theta})L, 0\}\big) \geq 0. \tag{2.133}$$

Hence, we obtain that

$$\textstyle\int_0^1 (h(y) - f(y))^2 \, \mu(\mathrm{d}y) \leq \mathcal{L}(\theta), \qquad Q(h) \leq \mathfrak{h}, \tag{2.134}$$

$$(\mathfrak{h} - Q(h) - 1)\big(\big|\textstyle\sum_{j=1}^{k}(-1)^j A_{i_j}(h)\big| + \max\{\operatorname{Lip}(h) - \mathfrak{h}L, 0\}\big) \geq 0. \tag{2.135}$$

Therefore, we get that $(\mathfrak{h} - Q(h) - 1)\big|\sum_{j=1}^{k}(-1)^j A_{i_j}(h)\big| \geq 0$. Combining Corollary 2.2 and (2.134) hence shows that there exist $\psi \in \mathbb{R}^{\mathfrak{d}}$ which satisfies

$$\mathscr{N}^{\psi} = h. \tag{2.136}$$

Note that (2.134), (2.135), and (2.136) demonstrate that

$$\mathcal{L}(\psi) \leq \mathcal{L}(\theta), \quad Q(\mathscr{N}^{\psi}) \leq \mathfrak{h}, \quad \text{and} \quad (\mathfrak{h} - Q(\mathscr{N}^{\psi}) - 1)\max\{\operatorname{Lip}(\mathscr{N}^{\psi}) - \mathfrak{h}L, 0\} \geq 0. \tag{2.137}$$

Lemma 2.13 hence implies that there exists $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ which satisfies

$$\mathcal{L}(\vartheta) \leq \mathcal{L}(\psi) \leq \mathcal{L}(\theta), \quad Q(\mathscr{N}^{\vartheta}) \leq Q(\mathscr{N}^{\psi}) \leq \mathfrak{h} = Q(\mathscr{N}^{\theta}), \quad \text{and} \quad \operatorname{Lip}(\mathscr{N}^{\vartheta}) \leq \mathfrak{h}L. \tag{2.138}$$

This proves (2.129) in the case $Q(\mathscr{N}^{\theta}) = \mathfrak{h}$. The proof of Proposition 2.2 is thus complete. □

As a simple consequence of Proposition 2.2 we obtain in Corollary 2.5 below that the new network parameter vector $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ can be chosen in such a way that in addition to the Lipschitz constant also the supremum norm of its realization function $\mathscr{N}^{\vartheta}$ is bounded by a constant depending only on $\mathfrak{h}$ and the target function f.

Corollary 2.5. *Assume Setting 2.1 and let $\theta \in \mathbb{R}^{\mathfrak{d}}$. Then there exists $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ such that $\mathcal{L}(\vartheta) \leq \mathcal{L}(\theta)$, $Q(\mathscr{N}^{\vartheta}) \leq Q(\mathscr{N}^{\theta})$, $\sup_{x\in[0,1]}|\mathscr{N}^{\vartheta}(x)| \leq \mathfrak{h}L + \sup_{x\in[0,1]}|f(x)|$, and $\operatorname{Lip}(\mathscr{N}^{\vartheta}) \leq \mathfrak{h}L$ (cf. Definition 2.4).*

Proof of Corollary 2.5. Note that Proposition 2.2 establishes that there exist $\psi \in \mathbb{R}^{\mathfrak{d}}$, $r \in [0,\infty)$ which satisfy

$$\mathcal{L}(\psi) \leq \mathcal{L}(\theta), \quad Q(\mathscr{N}^{\psi}) \leq Q(\mathscr{N}^{\theta}), \tag{2.139a}$$

$$\operatorname{Lip}(\mathscr{N}^{\psi}) \leq \mathfrak{h}L, \quad r = \inf_{x\in[0,1]}|\mathscr{N}^{\psi}(x) - f(x)|. \tag{2.139b}$$

Observe that (2.139) assures that there exist $y \in [0,1]$, $k \in \{-1,1\}$ which satisfy $\mathscr{N}^{\psi}(y) - f(y) = kr$. In the following let $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $i \in \{1,2,\ldots,3\mathfrak{h}\}$ that $\vartheta_i = \psi_i$ and $\vartheta_{\mathfrak{d}} = \psi_{\mathfrak{d}} - kr$. Note that the fact that for all $x \in [0,1]$ it holds that $\mathscr{N}^{\vartheta}(x) = \mathscr{N}^{\psi}(x) - kr$ and (2.139) show that $\operatorname{Lip}(\mathscr{N}^{\vartheta}) = \operatorname{Lip}(\mathscr{N}^{\psi}) \leq \mathfrak{h}L$ and $Q(\mathscr{N}^{\vartheta}) = Q(\mathscr{N}^{\psi}) \leq Q(\mathscr{N}^{\theta})$. The fact that $\mathscr{N}^{\vartheta}(y) = \mathscr{N}^{\psi}(y) - kr = f(y)$ and the triangle inequality therefore imply for all $z \in [0,1]$ that

$$\begin{aligned}|\mathscr{N}^{\vartheta}(z)| &\leq |\mathscr{N}^{\vartheta}(y)| + |\mathscr{N}^{\vartheta}(z) - \mathscr{N}^{\vartheta}(y)| = |f(y)| + |\mathscr{N}^{\vartheta}(z) - \mathscr{N}^{\vartheta}(y)| \\ &\leq \sup\nolimits_{x\in[0,1]}|f(x)| + \mathfrak{h}L|z-y| \leq \sup\nolimits_{x\in[0,1]}|f(x)| + \mathfrak{h}L.\end{aligned} \tag{2.140}$$

It remains to prove that $\mathcal{L}(\vartheta) \leq \mathcal{L}(\psi)$. For this we assume without loss of generality that

$$r > 0. \tag{2.141}$$

Observe that (2.139), (2.141), and the fact that $[0,1] \ni x \mapsto \mathscr{N}^{\psi}(x) - f(x) \in \mathbb{R}$ is continuous imply for all $x \in [0,1]$ that $k(\mathscr{N}^{\psi}(x) - f(x)) \geq r$. Hence, we obtain for all $x \in [0,1]$ that

$$\begin{aligned}|\mathscr{N}^{\vartheta}(x) - f(x)| &= |\mathscr{N}^{\psi}(x) - f(x) - kr| = |k(\mathscr{N}^{\psi}(x) - f(x)) - r| \\ &= k(\mathscr{N}^{\psi}(x) - f(x)) - r \leq k(\mathscr{N}^{\psi}(x) - f(x)) \\ &\leq |\mathscr{N}^{\psi}(x) - f(x)|.\end{aligned} \tag{2.142}$$

This demonstrates that $\mathcal{L}(\vartheta) \leq \mathcal{L}(\psi)$. The proof of Corollary 2.5 is thus complete. □

2.6 Existence of global minima for shallow ANNs

In this subsection we establish in Proposition 2.3 the existence of a global minimizer of the risk function under the assumptions of Setting 2.1. For the proof, we combine Corollary 2.5 with the Arzelà–Ascoli theorem to extract a convergent subsequence from a minimizing sequence. Due to the fact that the set of realization functions of shallow ReLU ANNs with fixed architecture is closed in the set of continuous functions with respect to the supremum norm (cf. Petersen et al. [61, Theorem 3.8]) the limit is again equal to the realization function of a suitable ANN.

Proposition 2.3. *Assume Setting 2.1. Then there exists* $\theta \in \mathbb{R}^{\mathfrak{d}}$ *such that* $\mathcal{L}(\theta) = \inf_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(\vartheta)$, $\operatorname{Lip}(\mathscr{N}^{\theta}) \leq \mathfrak{h}L$, *and* $\sup_{x \in [0,1]} |\mathscr{N}^{\theta}(x)| \leq \mathfrak{h}L + \sup_{x \in [0,1]} |f(x)|$ *(cf. Definition 2.4).*

Proof of Proposition 2.3. Note that there exists $\phi = (\phi_n)_{n \in \mathbb{N}} \colon \mathbb{N} \to \mathbb{R}^{\mathfrak{d}}$ which satisfies

$$\limsup_{n \to \infty} \mathcal{L}(\phi_n) = \inf_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(\vartheta). \tag{2.143}$$

Observe that Corollary 2.5 implies that there exists $\psi = (\psi_n)_{n \in \mathbb{N}} \colon \mathbb{N} \to \mathbb{R}^{\mathfrak{d}}$ which satisfies for all $n \in \mathbb{N}$ that

$$\mathcal{L}(\psi_n) \leq \mathcal{L}(\phi_n), \quad \sup_{x \in [0,1]} |\mathscr{N}^{\psi_n}(x)| \leq \mathfrak{h}L + \sup_{x \in [0,1]} |f(x)|, \quad \text{and} \quad \operatorname{Lip}(\mathscr{N}^{\psi_n}) \leq \mathfrak{h}L. \tag{2.144}$$

Note that (2.143) and (2.144) show that

$$\inf_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(\vartheta) \leq \limsup_{n \to \infty} \mathcal{L}(\psi_n) \leq \limsup_{n \to \infty} \mathcal{L}(\phi_n) = \inf_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(\vartheta). \tag{2.145}$$

Hence, we obtain that $\lim_{n \to \infty} \mathcal{L}(\psi_n) = \inf_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(\vartheta)$. Furthermore, observe that (2.144) and the Arzela–Ascoli theorem demonstrate that there exist $g \in C([0,1], \mathbb{R})$ and a strictly increasing $k \colon \mathbb{N} \to \mathbb{N}$ such that

$$\limsup_{n \to \infty} \sup_{x \in [0,1]} |\mathscr{N}^{\psi_{k(n)}}(x) - g(x)| = 0. \tag{2.146}$$

Next note that Petersen et al. [61, Theorem 3.8] assures that $\{h \in C([0,1], \mathbb{R}) \colon (\exists \, \vartheta \in \mathbb{R}^{\mathfrak{d}} \colon \mathscr{N}^{\vartheta} = h)\}$ is a closed subset of $C([0,1], \mathbb{R})$ with respect to the supremum norm on $C([0,1], \mathbb{R})$. Combining this with (2.146) implies that there exists $\theta \in \mathbb{R}^{\mathfrak{d}}$ which satisfies

$$\mathscr{N}^{\theta} = g. \tag{2.147}$$

Observe that (2.146), (2.147), and Lebesgue's theorem of dominated convergence ensure that

$$\begin{aligned}
\mathcal{L}(\theta) &= \int_0^1 (\mathscr{N}^{\theta}(y) - f(y))^2 \, \mu(\mathrm{d}y) = \int_0^1 (g(y) - f(y))^2 \, \mu(\mathrm{d}y) \\
&= \int_0^1 \Big[\lim_{n \to \infty} (\mathscr{N}^{\psi_{k(n)}}(y) - f(y))^2 \Big] \, \mu(\mathrm{d}y) = \lim_{n \to \infty} \bigg[\int_0^1 (\mathscr{N}^{\psi_{k(n)}}(y) - f(y))^2 \, \mu(\mathrm{d}y) \bigg] \\
&= \lim_{n \to \infty} \mathcal{L}(\psi_{k(n)}) = \inf_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(\vartheta).
\end{aligned} \tag{2.148}$$

Furthermore, note that (2.144), (2.146), and (2.147) demonstrate that $\sup_{x \in [0,1]} |\mathscr{N}^{\theta}(x)| \leq \mathfrak{h}L + \sup_{x \in [0,1]} |f(x)|$ and $\operatorname{Lip}(\mathscr{N}^{\theta}) \leq \mathfrak{h}L$. The proof of Proposition 2.3 is thus complete. □

Proposition 2.3 is formulated only for the input domain $[0,1]$. In Theorem 2.2 we generalize this result to a general input interval $[a,b] \subseteq \mathbb{R}$ by employing a suitable coordinate transformation.

Theorem 2.2. *Let $\mathfrak{h}, \mathfrak{d} \in \mathbb{N}$, $L, a \in \mathbb{R}$, $b \in (a, \infty)$, $f \in C([a,b], \mathbb{R})$ satisfy for all $x, y \in [a,b]$ that $\mathfrak{d} = 3\mathfrak{h} + 1$ and $|f(x) - f(y)| \leq L|x-y|$, let $\mu\colon \mathcal{B}([a,b]) \to [0,\infty]$ be a finite measure, for every $\theta = (\theta_1, \ldots, \theta_\mathfrak{d}) \in \mathbb{R}^\mathfrak{d}$ let $\mathscr{N}^\theta\colon \mathbb{R} \to \mathbb{R}$ satisfy for all $x \in \mathbb{R}$ that*

$$\mathscr{N}^\theta(x) = \theta_\mathfrak{d} + \textstyle\sum_{j=1}^{\mathfrak{h}} \theta_{2\mathfrak{h}+j} \max\{\theta_{\mathfrak{h}+j} + \theta_j x, 0\}, \tag{2.149}$$

and let $\mathcal{L}\colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$ that $\mathcal{L}(\theta) = \int_a^b (f(x) - \mathscr{N}^\theta(x))^2 \, \mu(\mathrm{d}x)$. Then there exists $\theta \in \mathbb{R}^\mathfrak{d}$ such that $\mathcal{L}(\theta) = \inf_{\vartheta \in \mathbb{R}^\mathfrak{d}} \mathcal{L}(\vartheta)$, $\sup_{x\in[a,b]}|\mathscr{N}^\theta(x)| \leq \mathfrak{h}L(b-a) + \sup_{x\in[a,b]}|f(x)|$, and

$$\sup_{x,y\in[a,b],x\neq y} \left[\tfrac{|\mathscr{N}^\theta(x) - \mathscr{N}^\theta(y)|}{|x-y|}\right] \leq \mathfrak{h}L. \tag{2.150}$$

Proof of Theorem 2.2. Throughout this proof let $\mathbf{f}\colon [0,1] \to [a,b]$ and $\mathbf{F}\colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}^\mathfrak{d}$ satisfy for all $x \in [0,1]$, $\theta = (\theta_1, \ldots, \theta_\mathfrak{d}) \in \mathbb{R}^\mathfrak{d}$ that $\mathbf{f}(x) = a + (b-a)x$ and

$$\mathbf{F}(\theta) = ((b-a)\theta_1, \ldots, (b-a)\theta_\mathfrak{h}, \theta_{\mathfrak{h}+1} + a\theta_1, \ldots, \theta_{2\mathfrak{h}} + a\theta_\mathfrak{h}, \theta_{2\mathfrak{h}+1}, \ldots, \theta_{3\mathfrak{h}}, \theta_{3\mathfrak{h}+1}), \tag{2.151}$$

let $g \in C([0,1], \mathbb{R})$ satisfy for all $x \in [0,1]$ that $g(x) = f(\mathbf{f}(x))$, and let $\nu\colon \mathcal{B}([0,1]) \to [0,\infty]$ satisfy for all $E \in \mathcal{B}([0,1])$ that $\nu(E) = \mu(\mathbf{f}(E))$. Observe that $\mathbf{f}$ and $\mathbf{F}$ are bijective. Moreover, note that for all $x \in [0,1]$, $\theta = (\theta_1, \ldots, \theta_\mathfrak{d}) \in \mathbb{R}^\mathfrak{d}$ it holds that

$$\begin{aligned} \mathscr{N}^{\mathbf{F}(\theta)}(x) &= \theta_\mathfrak{d} + \textstyle\sum_{j=1}^{\mathfrak{h}} \theta_{2\mathfrak{h}+j} \max\{(b-a)\theta_j x + \theta_{\mathfrak{h}+j} + a\theta_j, 0\} \\ &= \theta_\mathfrak{d} + \textstyle\sum_{j=1}^{\mathfrak{h}} \theta_{2\mathfrak{h}+j} \max\{\theta_j \mathbf{f}(x) + \theta_{\mathfrak{h}+j}, 0\} = \mathscr{N}^\theta(\mathbf{f}(x)). \end{aligned} \tag{2.152}$$

In addition, observe that for all $x, y \in [0,1]$ we have that

$$|g(x) - g(y)| = |f(\mathbf{f}(x)) - f(\mathbf{f}(y))| \leq L|\mathbf{f}(x) - \mathbf{f}(y)| = L(b-a)|x-y|. \tag{2.153}$$

Proposition 2.3 hence demonstrates that there exists $\psi \in \mathbb{R}^\mathfrak{d}$ which satisfies $\mathrm{Lip}(\mathscr{N}^\psi) \leq \mathfrak{h}L(b-a)$, $\sup_{x\in[0,1]}|\mathscr{N}^\psi(x)| \leq \mathfrak{h}L(b-a) + \sup_{x\in[0,1]}|g(x)|$, and

$$\int_0^1 (\mathscr{N}^\psi(x) - g(x))^2 \, \nu(\mathrm{d}x) = \inf_{\vartheta\in\mathbb{R}^\mathfrak{d}} \left[\int_0^1 (\mathscr{N}^\vartheta(x) - g(x))^2 \, \nu(\mathrm{d}x)\right] \tag{2.154}$$

(cf. Definition 2.4). In the following let $\theta \in \mathbb{R}^\mathfrak{d}$ satisfy $\theta = \mathbf{F}^{-1}(\psi)$. Note that (2.152) and the integral transformation theorem assure for all $\vartheta \in \mathbb{R}^\mathfrak{d}$ that

$$\begin{aligned} \mathcal{L}(\vartheta) &= \int_a^b (\mathscr{N}^\vartheta(x) - f(x))^2 \, \mu(\mathrm{d}x) = \int_0^1 (\mathscr{N}^\vartheta(\mathbf{f}(x)) - f(\mathbf{f}(x)))^2 \, \nu(\mathrm{d}x) \\ &= \int_0^1 (\mathscr{N}^{\mathbf{F}(\vartheta)}(x) - g(x))^2 \, \nu(\mathrm{d}x). \end{aligned} \tag{2.155}$$

Combining this with (2.154) and the fact that $\mathbf{F}$ is bijective shows that

$$\begin{aligned} \mathcal{L}(\theta) &= \int_0^1 (\mathscr{N}^\psi(x) - g(x))^2 \, \nu(\mathrm{d}x) = \inf_{\vartheta\in\mathbb{R}^\mathfrak{d}} \left[\int_0^1 (\mathscr{N}^\vartheta(x) - g(x))^2 \, \nu(\mathrm{d}x)\right] \\ &= \inf_{\vartheta\in\mathbb{R}^\mathfrak{d}} \left[\int_0^1 (\mathscr{N}^{\mathbf{F}(\vartheta)}(x) - g(x))^2 \, \nu(\mathrm{d}x)\right] = \inf_{\vartheta\in\mathbb{R}^\mathfrak{d}} \mathcal{L}(\vartheta). \end{aligned} \tag{2.156}$$

In addition, observe that (2.152) ensures for all $x \in [a, b]$ that

$$|\mathscr{N}^\theta(x)| = |\mathscr{N}^\psi(\mathbf{f}^{-1}(x))| \le \mathfrak{h}L(b-a) + \sup_{y\in[0,1]}|g(y)| \\ = \mathfrak{h}L(b-a) + \sup_{y\in[a,b]}|f(y)|. \tag{2.157}$$

Finally, note that (2.152) demonstrates for all $x, y \in [a, b]$ that

$$|\mathscr{N}^\theta(x) - \mathscr{N}^\theta(y)| = |\mathscr{N}^\psi(\mathbf{f}^{-1}(x)) - \mathscr{N}^\psi(\mathbf{f}^{-1}(y))| \le \mathfrak{h}L(b-a)|\mathbf{f}^{-1}(x) - \mathbf{f}^{-1}(y)| \\ = \mathfrak{h}L|x-y|. \tag{2.158}$$

The proof of Theorem 2.2 is thus complete. □

2.7 Existence of regular global minima for shallow ANNs

In the final result of this section, Corollary 2.6, we strengthen Theorem 2.2 by showing that there also exists a global minimizer of the risk function which admits a neighborhood on which the risk function is continuously differentiable. Furthermore, the gradient on this neighborhood can be obtained from a sequence of approximate realization functions using suitable differentiable approximations of the ReLU function, as outlined in the introduction. The proof relies on regularity results from our previous article Eberle et al. [28].

Corollary 2.6. *Let $\mathfrak{h}, \mathfrak{d} \in \mathbb{N}$, $L, a \in \mathbb{R}$, $b \in (a, \infty)$, $f \in C([a, b], \mathbb{R})$ satisfy for all $x, y \in [a, b]$ that $\mathfrak{d} = 3\mathfrak{h} + 1$ and $|f(x) - f(y)| \le L|x - y|$, let $\mathfrak{p}\colon [a, b] \to [0, \infty)$ be bounded and measurable, let $\mathfrak{R}_r\colon \mathbb{R} \to \mathbb{R}$, $r \in \mathbb{N} \cup \{\infty\}$, satisfy for all $x \in \mathbb{R}$ that $(\cup_{r\in\mathbb{N}}\{\mathfrak{R}_r\}) \subseteq C^1(\mathbb{R}, \mathbb{R})$, $\mathfrak{R}_\infty(x) = \max\{x, 0\}$, $\sup_{r\in\mathbb{N}} \sup_{y\in[-|x|,|x|]} |(\mathfrak{R}_r)'(y)| < \infty$, and*

$$\limsup_{r\to\infty}(|\mathfrak{R}_r(x) - \mathfrak{R}_\infty(x)| + |(\mathfrak{R}_r)'(x) - \mathbb{1}_{(0,\infty)}(x)|) = 0, \tag{2.159}$$

for every $r \in \mathbb{N} \cup \{\infty\}$ let $\mathcal{L}_r\colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}$ satisfy for all $\theta = (\theta_1, \ldots, \theta_\mathfrak{d}) \in \mathbb{R}^\mathfrak{d}$ that

$$\mathcal{L}_r(\theta) = \int_a^b (f(x) - \theta_\mathfrak{d} - \textstyle\sum_{j=1}^{\mathfrak{h}} \theta_{2\mathfrak{h}+j}[\mathfrak{R}_r(\theta_{\mathfrak{h}+j} + \theta_j x)])^2\, \mathfrak{p}(x)\, \mathrm{d}x, \tag{2.160}$$

let $U \subseteq \mathbb{R}^\mathfrak{d}$ satisfy

$$U = \{\theta = (\theta_1, \ldots, \theta_\mathfrak{d}) \in \mathbb{R}^\mathfrak{d}\colon (\forall\, i \in \{1, 2, \ldots, \mathfrak{h}\}\colon |\theta_i| + |\theta_{\mathfrak{h}+i}| > 0)\}, \tag{2.161}$$

and let $\mathcal{G}\colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}^\mathfrak{d}$ satisfy for all $\theta \in \{\vartheta \in \mathbb{R}^\mathfrak{d}\colon ((\nabla\mathcal{L}_r)(\vartheta))_{r\in\mathbb{N}}$ is convergent$\}$ that $\mathcal{G}(\theta) = \lim_{r\to\infty}(\nabla\mathcal{L}_r)(\theta)$. Then

(i) it holds that U is open,

(ii) it holds that $(\mathcal{L}_\infty)|_U \in C^1(U, \mathbb{R})$,

(iii) it holds that $U \ni \theta \mapsto (\nabla\mathcal{L}_\infty)(\theta) \in \mathbb{R}^\mathfrak{d}$ is locally Lipschitz continuous,

(iv) it holds for all $\theta \in U$ that $(\nabla\mathcal{L}_\infty)(\theta) = \mathcal{G}(\theta)$, and

(v) it holds that there exists $\theta \in U$ *such that* $\mathcal{L}_\infty(\theta) = \inf_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}_\infty(\vartheta)$, $\sup_{x \in [a,b]} |\mathscr{N}^\theta(x)| \leq \mathfrak{h} L(b-a) + \sup_{x \in [a,b]} |f(x)|$, *and* $\sup_{x,y \in [a,b], x \neq y} (|x-y|^{-1} |\mathscr{N}^\theta(x) - \mathscr{N}^\theta(y)|) \leq \mathfrak{h} L$.

Proof of Corollary 2.6. Throughout this proof for every $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $r \in \mathbb{N} \cup \{\infty\}$ let $\mathscr{N}_r^\theta : \mathbb{R} \to \mathbb{R}$ satisfy for all $x \in \mathbb{R}$ that

$$\mathscr{N}_r^\theta(x) = \theta_{\mathfrak{d}} + \textstyle\sum_{j=1}^{\mathfrak{h}} \theta_{2\mathfrak{h}+j} [\mathfrak{R}_r(\theta_{\mathfrak{h}+j} + \theta_j x)]. \tag{2.162}$$

Observe that (2.161) proves item (i). Note that Eberle et al. [28, Proposition 2.3] establishes items (ii) and (iv). Observe that Eberle et al. [28, Corollary 2.7] and item (iv) prove item (iii). Note that Theorem 2.2 (applied with[8] $\mu \curvearrowleft (\mathcal{B}([a,b]) \ni E \mapsto \int_E \mathfrak{p}(x)\,\mathrm{d}x \in [0,\infty])$ in the notation of Theorem 2.2) proves that there exists $\psi \in \mathbb{R}^{\mathfrak{d}}$ which satisfies $\sup_{x \in [a,b]} |\mathscr{N}^\psi(x)| \leq \mathfrak{h} L(b-a) + \sup_{x \in [a,b]} |f(x)|$, $\sup_{x,y \in [a,b], x \neq y} (|x-y|^{-1} |\mathscr{N}^\psi(x) - \mathscr{N}^\psi(y)|) \leq \mathfrak{h} L$, and

$$\mathcal{L}_\infty(\psi) = \inf_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}_\infty(\vartheta). \tag{2.163}$$

In the following let $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}})$ satisfy for all $i \in \mathbb{N} \cap ([1,\mathfrak{h}] \cup (2\mathfrak{h}, \mathfrak{d}])$, $j \in \mathbb{N} \cap (\mathfrak{h}, 2\mathfrak{h}]$ that

$$\theta_i = \psi_i \qquad \text{and} \qquad \theta_j = \psi_j - \mathbb{1}_{\{0\}}(|\psi_{j-\mathfrak{h}}| + |\psi_j|). \tag{2.164}$$

Observe that (2.164) shows for all $i \in \{1, 2, \ldots, \mathfrak{h}\}$, $x \in \mathbb{R}$ that $\max\{\theta_{\mathfrak{h}+i} + \theta_i x, 0\} = \max\{\psi_{\mathfrak{h}+i} + \psi_i x, 0\}$. Therefore, we obtain for all $x \in \mathbb{R}$ that $\mathscr{N}_\infty^\theta(x) = \mathscr{N}_\infty^\psi(x)$. Combining this with (2.163) establishes item (v). The proof of Corollary 2.6 is thus complete. ☐

3 Regularity analysis for generalized gradients in the training of deep ANNs

In this section we introduce in Setting 3.1 in Subsection 3.1 below our mathematical framework for deep ReLU ANNs. As in [13, 36, 41] we approximate the ReLU activation function $\mathfrak{R}_\infty : \mathbb{R} \to \mathbb{R}$ through continuously differentiable functions $\mathfrak{R}_r : \mathbb{R} \to \mathbb{R}$, $r \in [1,\infty)$, in order to define an appropriate generalized gradient $\mathcal{G} : \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ of the risk function $\mathcal{L}_\infty : \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$; see (3.3) and (3.6) in Setting 3.1.

In Proposition 3.1 in Subsection 3.2 below (explicit representation and pointwise approximations for $\mathcal{G}$), in Lemma 3.1 in Subsection 3.3 below (local Lipschitz continuity of $\mathcal{L}_\infty$), and in Lemma 3.2 in Subsection 3.4 below (uniform local boundedness for $\nabla \mathcal{L}_r$, $r \in [1,\infty)$) we then state several important regularity properties of the risk function $\mathcal{L} : \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ and its generalized gradient function $\mathcal{G} : \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$. Proposition 3.1 is proved in Hutzenthaler et al. [36, Theorem 2.9], Lemma 3.1 follows, e.g., from Hutzenthaler et al. [36, Lemma 2.10], and Lemma 3.2 is a consequence from Hutzenthaler et al. [36, Lemma 3.6].

In Corollary 3.1 in Subsection 3.4 we show that the risk function $\mathcal{L}_\infty : \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ is weakly differentiable with the generalized gradient function $\mathcal{G} : \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ serving as a weak gradient function.

[8] Here and in the remainder of this article, when applying another theorem/lemma/proposition we use the notation $\curvearrowleft$ to indicate which values are assigned to the variables in the applied result. In this particular case, Theorem 2.2, where μ is an arbitrary finite measure on $[a,b]$, is applied with the specific measure $\mathcal{B}([a,b]) \ni E \mapsto \int_E \mathfrak{p}(x)\,\mathrm{d}x \in [0,\infty]$.

In Proposition 3.2 in Subsection 3.5 below we establish that the risk function $\mathcal{L}_\infty \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ is differentiable Lebesgue almost everywhere with its gradients agreeing Lebesgue almost everywhere with the generalized gradient function $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$. Our proof of Proposition 3.2 relies on Lemma 3.1, Corollary 3.1, and well-known results on weak derivatives of locally Lipschitz continuous functions (cf. Evans [29]).

In Subsection 3.6 below we gather several known notions and lemmas regarding Fréchet subdifferentials. In particular, in the scientific literature Definition 3.1 can be found, e.g., as Rockafellar & Wets [63, Definition 8.3] and Bolte et al. [10, Definition 2.10], items (iii), (iv), and (v) in Lemma 3.3 are proved, e.g., as [63, Theorem 8.6 and Exercise 8.8], and Lemma 3.4 is a reformulation of the well-known fact that the limiting Fréchet subdifferential of a continuous function has a closed graph (see, e.g., [63, Proposition 8.7]).

Finally, in Proposition 3.3 in Subsection 3.9 below (the main result of this section) we establish that for every ANN parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that the generalized gradient $\mathcal{G}(\theta)$ is a limiting subgradient of the risk function $\mathcal{L}_\infty$ at θ. Our proof of Proposition 3.3 relies on Proposition 3.2, on Lemma 3.3, as well as on the continuity type result for the generalized gradient function in Lemma 3.7 in Subsection 3.8. Our proof of Lemma 3.7, in turn, is based on local underestimate type result in Lemma 3.5 in Subsection 3.7 as well as on the conditional continuity result for the generalized gradient function in Lemma 3.6 in Subsection 3.8 below.

3.1 Mathematical framework for deep ANNs with ReLU activation

Setting 3.1. Let $a \in \mathbb{R}$, $b \in [a, \infty)$, $\in (0, \infty)$, $\in (, \infty)$, $(\ell_k)_{k \in \mathbb{N}_0} \subseteq \mathbb{N}$, $L, \mathfrak{d} \in \mathbb{N}$ satisfy $\mathfrak{d} = \sum_{k=1}^{L} \ell_k(\ell_{k-1} + 1)$, for every $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ let $\mathfrak{w}^{k,\theta} = (\mathfrak{w}^{k,\theta}_{i,j})_{(i,j) \in \{1,\ldots,\ell_k\} \times \{1,\ldots,\ell_{k-1}\}} \in \mathbb{R}^{\ell_k \times \ell_{k-1}}$, $k \in \mathbb{N}$, and $\mathfrak{b}^{k,\theta} = (\mathfrak{b}^{k,\theta}_1, \ldots, \mathfrak{b}^{k,\theta}_{\ell_k}) \in \mathbb{R}^{\ell_k}$, $k \in \mathbb{N}$, satisfy for all $k \in \{1, \ldots, L\}$, $i \in \{1, \ldots, \ell_k\}$, $j \in \{1, \ldots, \ell_{k-1}\}$ that

$$\mathfrak{w}^{k,\theta}_{i,j} = \theta_{(i-1)\ell_{k-1} + j + \sum_{h=1}^{k-1} \ell_h(\ell_{h-1}+1)} \qquad \text{and} \qquad \mathfrak{b}^{k,\theta}_i = \theta_{\ell_k \ell_{k-1} + i + \sum_{h=1}^{k-1} \ell_h(\ell_{h-1}+1)}, \tag{3.1}$$

for every $k \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ let $\mathcal{A}^\theta_k = (\mathcal{A}^\theta_{k,1}, \ldots, \mathcal{A}^\theta_{k,\ell_k}) \colon \mathbb{R}^{\ell_{k-1}} \to \mathbb{R}^{\ell_k}$ satisfy for all $x \in \mathbb{R}^{\ell_{k-1}}$ that

$$\mathcal{A}^\theta_k(x) = \mathfrak{b}^{k,\theta} + \mathfrak{w}^{k,\theta} x, \tag{3.2}$$

let $\mathfrak{R}_r \colon \mathbb{R} \to \mathbb{R}$, $r \in [1, \infty]$, satisfy for all $r \in [1, \infty)$, $x \in (-\infty, r^{-1}]$, $y \in \mathbb{R}$, $z \in [r^{-1}, \infty)$ that

$$\mathfrak{R}_r \in C^1(\mathbb{R}, \mathbb{R}), \quad \mathfrak{R}_r(x) = 0, \quad 0 \le \mathfrak{R}_r(y) \le \mathfrak{R}_\infty(y) = \max\{y, 0\}, \quad \text{and} \quad \mathfrak{R}_r(z) = z, \tag{3.3}$$

assume $\sup_{r \in [1,\infty)} \sup_{x \in \mathbb{R}} |(\mathfrak{R}_r)'(x)| < \infty$, for every $r \in [1, \infty]$, $k \in \mathbb{N}$ let $\mathfrak{M}_{r,k} \colon \mathbb{R}^{\ell_k} \to \mathbb{R}^{\ell_k}$ satisfy for all $x = (x_1, \ldots, x_{\ell_k}) \in \mathbb{R}^{\ell_k}$ that

$$\mathfrak{M}_{r,k}(x) = (\mathfrak{R}_r(x_1), \ldots, \mathfrak{R}_r(x_{\ell_k})), \tag{3.4}$$

for every $\theta \in \mathbb{R}^{\mathfrak{d}}$ let $\mathcal{N}_r^{k,\theta} = (\mathcal{N}_{r,1}^{k,\theta}, \dots, \mathcal{N}_{r,\ell_k}^{k,\theta}) \colon \mathbb{R}^{\ell_0} \to \mathbb{R}^{\ell_k}$, $k \in \mathbb{N}$, $r \in [1,\infty]$, and $\mathcal{X}_i^{k,\theta} \subseteq \mathbb{R}^{\ell_0}$, $k, i \in \mathbb{N}$, satisfy for all $k \in \mathbb{N}$, $r \in [1,\infty]$, $i \in \{1, \dots, \ell_k\}$ that

$$\mathcal{N}_r^{1,\theta} = \mathcal{A}_1^{\theta}, \quad \mathcal{N}_r^{k+1,\theta} = \mathcal{A}_{k+1}^{\theta} \circ \mathfrak{M}_{r^{1/k},k} \circ \mathcal{N}_r^{k,\theta}, \quad \text{and} \quad \mathcal{X}_i^{k,\theta} = \{x \in [a,b]^{\ell_0} \colon \mathcal{N}_{\infty,i}^{k,\theta}(x) > 0\}, \tag{3.5}$$

let $f = (f_1, \dots, f_{\ell_L}) \colon [a,b]^{\ell_0} \to \mathbb{R}^{\ell_L}$ be measurable, let $\mu \colon \mathcal{B}([a,b]^{\ell_0}) \to [0,\infty]$ be a finite measure, for every $r \in [1,\infty]$ let $\mathcal{L}_r \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}_r(\theta) = \int_{[a,b]^{\ell_0}} \|\mathcal{N}_r^{L,\theta}(x) - f(x)\|^2 \, \mu(\mathrm{d}x), \tag{3.6}$$

and let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}) \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \{\vartheta \in \mathbb{R}^{\mathfrak{d}} \colon ((\nabla \mathcal{L}_r)(\vartheta))_{r \in [1,\infty)} \text{ is convergent}\}$ that $\mathcal{G}(\theta) = \lim_{r \to \infty} (\nabla \mathcal{L}_r)(\theta)$.

3.2 Explicit representations for the generalized gradient function

In Proposition 3.1 we show that the approximating sequence of gradients $(\nabla \mathcal{L}_r)(\theta)$, $r \in [1,\infty)$, in Setting 3.1 converges for every $\theta \in \mathbb{R}^{\mathfrak{d}}$. Furthermore, we derive in items (iii) and (iv) explicit formulae for the limit $\mathcal{G}(\theta)$. This explicit representation of $\mathcal{G}$ agrees with the standard generalized gradient obtained by formally defining the derivative of the ReLU as the left derivative $\mathbb{1}_{(0,\infty)}$ and applying the chain rule.

Proposition 3.1. *Assume Setting 3.1 and let $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$. Then*

(i) it holds for all $r \in [1,\infty)$ that $\mathcal{L}_r \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$,

(ii) it holds that $\limsup_{r \to \infty} (|\mathcal{L}_r(\theta) - \mathcal{L}_\infty(\theta)| + \|(\nabla \mathcal{L}_r)(\theta) - \mathcal{G}(\theta)\|) = 0$,

(iii) it holds for all $k \in \{1, \dots, L\}$, $i \in \{1, \dots, \ell_k\}$, $j \in \{1, \dots, \ell_{k-1}\}$ that

$$\begin{aligned} &\mathcal{G}_{(i-1)\ell_{k-1}+j+\sum_{h=1}^{k-1} \ell_h(\ell_{h-1}+1)}(\theta) \\ &= \sum_{\substack{v_k, v_{k+1}, \dots, v_L \in \mathbb{N}, \\ \forall\, w \in \mathbb{N} \cap [k,L] \colon v_w \le \ell_w}} \int_{[a,b]^{\ell_0}} 2 \Big[\mathfrak{R}_\infty(\mathcal{N}_{\infty,j}^{\max\{k-1,1\},\theta}(x)) \mathbb{1}_{(1,L]}(k) + x_j \mathbb{1}_{\{1\}}(k) \Big] \\ &\quad \cdot \Big[\mathbb{1}_{\{i\}}(v_k) \Big] \Big[\mathcal{N}_{\infty,v_L}^{L,\theta}(x) - f_{v_L}(x) \Big] \Big[\textstyle\prod_{n=k+1}^{L} \big(\mathfrak{w}_{v_n,v_{n-1}}^{n,\theta} \mathbb{1}_{\mathcal{X}_{v_{n-1}}^{n-1,\theta}}(x) \big) \Big] \, \mu(\mathrm{d}x), \end{aligned} \tag{3.7}$$

(iv) it holds for all $k \in \{1, \dots, L\}$, $i \in \{1, \dots, \ell_k\}$ that

$$\begin{aligned} &\mathcal{G}_{\ell_k \ell_{k-1}+i+\sum_{h=1}^{k-1} \ell_h(\ell_{h-1}+1)}(\theta) \\ &= \sum_{\substack{v_k, v_{k+1}, \dots, v_L \in \mathbb{N}, \\ \forall\, w \in \mathbb{N} \cap [k,L] \colon v_w \le \ell_w}} \int_{[a,b]^{\ell_0}} 2 \Big[\mathbb{1}_{\{i\}}(v_k) \Big] \\ &\quad \cdot \Big[\mathcal{N}_{\infty,v_L}^{L,\theta}(x) - f_{v_L}(x) \Big] \Big[\textstyle\prod_{n=k+1}^{L} \big(\mathfrak{w}_{v_n,v_{n-1}}^{n,\theta} \mathbb{1}_{\mathcal{X}_{v_{n-1}}^{n-1,\theta}}(x) \big) \Big] \, \mu(\mathrm{d}x). \end{aligned} \tag{3.8}$$

Proof of Proposition 3.1. Note that [36, Items (i), (iv), (v), and (vi) in Theorem 2.9] establishes items (i), (ii), (iii), and (iv) The proof of Proposition 3.1 is thus complete. □

3.3 Local Lipschitz continuity of the risk function

Lemma 3.1. *Assume Setting 3.1 and let $K \subseteq \mathbb{R}^{\mathfrak{d}}$ be compact. Then there exists $\mathfrak{C} \in \mathbb{R}$ such that for all $\theta, \vartheta \in K$ it holds that*

$$|\mathcal{L}_\infty(\theta) - \mathcal{L}_\infty(\vartheta)| + \big(\sup_{x \in [a,b]^{\ell_0}} \|\mathcal{N}_\infty^{L,\theta}(x) - \mathcal{N}_\infty^{L,\vartheta}(x)\|\big) \leq \mathfrak{C}\|\theta - \vartheta\|. \tag{3.9}$$

Proof of Lemma 3.1. Observe that, e.g., [36, Lemma 2.10] establishes (3.9). The proof of Lemma 3.1 is thus complete. □

3.4 Weak differentiability properties of the risk function

Lemma 3.2. *Assume Setting 3.1 and let $K \subseteq \mathbb{R}^{\mathfrak{d}}$ be non-empty and compact. Then*

$$\sup_{\theta \in K} \sup_{r \in [1,\infty)} (|\mathcal{L}_r(\theta)| + |\mathcal{L}_\infty(\theta)| + \|(\nabla \mathcal{L}_r)(\theta)\| + \|\mathcal{G}(\theta)\|) < \infty. \tag{3.10}$$

Proof of Lemma 3.2. Note that [36, Lemma 3.6] and item (i) in Proposition 3.1 show for all $s \in (0,\infty)$ that $\sup_{\theta \in \{\vartheta \in \mathbb{R}^{\mathfrak{d}} : \|\vartheta\| \leq s\}} \sup_{r \in [1,\infty)} \|(\nabla \mathcal{L}_r)(\theta)\| < \infty$. The fundamental theorem of calculus and the fact that for all $r \in [1,\infty)$ it holds that $\mathcal{L}_r(0) = \mathcal{L}_\infty(0)$ hence demonstrate that for all $s \in (0,\infty)$ we have that $\sup_{\theta \in \{\vartheta \in \mathbb{R}^{\mathfrak{d}} : \|\vartheta\| \leq s\}} \sup_{r \in [1,\infty)} (|\mathcal{L}_r(\theta)| + \|(\nabla \mathcal{L}_r)(\theta)\|) < \infty$. Combining this with item (ii) in Proposition 3.1 establishes (3.10). The proof of Lemma 3.2 is thus complete. □

As a consequence of Proposition 3.1 and the uniform boundedness result in Lemma 3.2 we obtain in Corollary 3.1 that the generalized gradient $\mathcal{G}$ serves as a weak gradient of the risk function $\mathcal{L}_\infty$.

Corollary 3.1 (Weak differentiability)**.** *Assume Setting 3.1, let $\varphi = (\varphi(\theta))_{\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}} : \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ be compactly supported and continuously differentiable, and let $i \in \{1, 2, \ldots, \mathfrak{d}\}$. Then $\int_{\mathbb{R}^{\mathfrak{d}}} |\mathcal{L}_\infty(\theta)\, (\frac{\partial}{\partial \theta_i} \varphi)(\theta)| + |\mathcal{G}_i(\theta)\, \varphi(\theta)| \, \mathrm{d}\theta < \infty$ and*

$$\int_{\mathbb{R}^{\mathfrak{d}}} \mathcal{L}_\infty(\theta)\, (\tfrac{\partial}{\partial \theta_i} \varphi)(\theta) \, \mathrm{d}\theta = -\int_{\mathbb{R}^{\mathfrak{d}}} \mathcal{G}_i(\theta)\, \varphi(\theta) \, \mathrm{d}\theta. \tag{3.11}$$

Proof of Corollary 3.1. Observe that the assumption that φ has a compact support ensures that there exists $R \in (0,\infty)$ which satisfies for all $\theta \in \mathbb{R}^{\mathfrak{d}} \backslash [-R,R]^{\mathfrak{d}}$ that

$$\varphi(\theta) = 0. \tag{3.12}$$

Note that Lemma 3.2 demonstrates that

$$\sup_{\theta \in [-R,R]^{\mathfrak{d}}} \sup_{r \in [1,\infty)} \big(|\mathcal{L}_r(\theta)| + |\mathcal{L}_\infty(\theta)| + \|(\nabla \mathcal{L}_r)(\theta)\| + \|\mathcal{G}(\theta)\|\big) < \infty. \tag{3.13}$$

This and (3.12) assure that for all $r \in [1,\infty)$, $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} &|\mathcal{L}_r(\theta)(\tfrac{\partial}{\partial \theta_i} \varphi)(\theta)| + |(\tfrac{\partial}{\partial \theta_i} \mathcal{L}_r)(\theta) \varphi(\theta)| \\ &\leq \big[\sup_{\vartheta \in [-R,R]^{\mathfrak{d}}} \sup_{s \in [1,\infty)} \big(|\mathcal{L}_s(\vartheta)| + \|(\nabla \mathcal{L}_s)(\vartheta)\| + |\varphi(\vartheta)| + \|(\nabla \varphi)(\vartheta)\|\big)\big] \mathbb{1}_{[-R,R]^{\mathfrak{d}}}(\theta) \\ &< \infty. \end{aligned} \tag{3.14}$$

Proposition 3.1, Lebesgue's dominated convergence theorem, (3.12), and integration by parts therefore ensure that

$$\int_{\mathbb{R}^{\mathfrak{d}}} \mathcal{L}_\infty(\theta)\,(\tfrac{\partial}{\partial\theta_i}\varphi)(\theta)\,\mathrm{d}\theta = \lim_{r\to\infty}\left[\int_{\mathbb{R}^{\mathfrak{d}}} \mathcal{L}_r(\theta)\,(\tfrac{\partial}{\partial\theta_i}\varphi)(\theta)\,\mathrm{d}\theta\right]$$
$$= \lim_{r\to\infty}\left[\int_{[-R,R]^{\mathfrak{d}}} \mathcal{L}_r(\theta)\,(\tfrac{\partial}{\partial\theta_i}\varphi)(\theta)\,\mathrm{d}\theta\right] = -\left(\lim_{r\to\infty}\left[\int_{[-R,R]^{\mathfrak{d}}} (\tfrac{\partial}{\partial\theta_i}\mathcal{L}_r)(\theta)\,\varphi(\theta)\,\mathrm{d}\theta\right]\right). \tag{3.15}$$

Proposition 3.1, (3.14), and Lebesgue's dominated convergence theorem hence show that

$$\int_{\mathbb{R}^{\mathfrak{d}}} \mathcal{L}_\infty(\theta)\,(\tfrac{\partial}{\partial\theta_i}\varphi)(\theta)\,\mathrm{d}\theta = -\int_{[-R,R]^{\mathfrak{d}}} \left[\lim_{r\to\infty}(\tfrac{\partial}{\partial\theta_i}\mathcal{L}_r)(\theta)\right]\varphi(\theta)\,\mathrm{d}\theta$$
$$= -\int_{\mathbb{R}^{\mathfrak{d}}} \mathcal{G}_i(\theta)\,\varphi(\theta)\,\mathrm{d}\theta. \tag{3.16}$$

This, (3.12), and (3.13) establish (3.11). The proof of Corollary 3.1 is thus complete. □

3.5 Strong differentiability properties of the risk function

We next establish in Proposition 3.2 that the risk function is a.e. strongly differentiable with gradient $\mathcal{G}$. The proof relies on Corollary 3.1, the local Lipschitz continuity result in Lemma 3.1, Rademacher's theorem, and the fact that locally Lipschitz continuous functions are weakly differentiable with the weak gradient a.e. equal to the strong gradient (cf. Evans [29]).

Proposition 3.2. *Assume Setting 3.1. Then there exists $E \in \mathcal{B}(\mathbb{R}^{\mathfrak{d}})$ such that*

(i) it holds that $\int_{\mathbb{R}^{\mathfrak{d}}\setminus E} 1\,\mathrm{d}\theta = 0$,

(ii) it holds for all $\theta \in E$ that $\mathcal{L}_\infty$ is differentiable at θ, and

(iii) it holds for all $\theta \in E$ that $(\nabla\mathcal{L}_\infty)(\theta) = \mathcal{G}(\theta)$.

Proof of Proposition 3.2. Throughout this proof let $G = (G_1, \ldots, G_{\mathfrak{d}})\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$G(\theta) = \begin{cases} (\nabla\mathcal{L}_\infty)(\theta) & \colon \mathcal{L}_\infty \text{ is differentiable at } \theta, \\ 0 & \colon \mathcal{L}_\infty \text{ is not differentiable at } \theta. \end{cases} \tag{3.17}$$

Observe that (3.17), the fact that for all measurable $g_n\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$, $n \in \mathbb{N}$, it holds that $\{\theta \in \mathbb{R}^{\mathfrak{d}}\colon (g_n(\theta))_{n\in\mathbb{N}} \text{ is a Cauchy sequence}\}$ is measurable, and the fact that for all measurable and pointwise convergent $g_n\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$, $n \in \mathbb{N}$, it holds that $\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \lim_{n\to\infty} g_n(\theta) \in \mathbb{R}^{\mathfrak{d}}$ is measurable establish that G is measurable. Furthermore, note that Lemma 3.1 ensures that $\mathcal{L}_\infty$ is locally Lipschitz continuous. Rademacher's theorem (cf. Evans [29, Theorem 5.8.6]) therefore demonstrates that there exists $\mathcal{E} \in \{A \in \mathcal{B}(\mathbb{R}^{\mathfrak{d}})\colon \int_{\mathbb{R}^{\mathfrak{d}}\setminus A} 1\,\mathrm{d}\theta = 0\}$ which satisfies for all $\theta \in \mathcal{E}$ that $\mathcal{L}_\infty$ is differentiable at θ. Lemma 3.1, Evans [29, Theorems

5.8.4 and 5.8.5], and (3.17) hence show for all compactly supported $\varphi \in C^\infty(\mathbb{R}^\mathfrak{d}, \mathbb{R})$ and all $i \in \{1, 2, \dots, \mathfrak{d}\}$ that $\int_{\mathbb{R}^\mathfrak{d}} |\mathcal{L}_\infty(\theta) \, (\frac{\partial}{\partial\theta_i}\varphi)(\theta)| + |G_i(\theta) \, \varphi(\theta)| \, \mathrm{d}\theta < \infty$ and

$$\int_{\mathbb{R}^\mathfrak{d}} \mathcal{L}_\infty(\theta) \, (\tfrac{\partial}{\partial\theta_i}\varphi)(\theta) \, \mathrm{d}\theta = -\int_{\mathbb{R}^\mathfrak{d}} G_i(\theta) \, \varphi(\theta) \, \mathrm{d}\theta. \tag{3.18}$$

Combining this with Corollary 3.1 and the fundamental lemma of calculus of variations (cf., e.g., Hörmander [34, Theorem 1.2.5]) implies that there exists $\mathbf{E} \in \{A \in \mathcal{B}(\mathbb{R}^\mathfrak{d}) \colon \int_{\mathbb{R}^\mathfrak{d} \setminus A} 1 \, \mathrm{d}\theta = 0\}$ which satisfies for all $\theta \in \mathbf{E}$ that

$$G(\theta) = \mathcal{G}(\theta). \tag{3.19}$$

Observe that (3.17), (3.19), and the fact that for all $\theta \in \mathcal{E}$ it holds that $\mathcal{L}_\infty$ is differentiable at θ assure that for all $\theta \in (\mathbf{E} \cap \mathcal{E})$ it holds that

$$\mathcal{G}(\theta) = (\nabla \mathcal{L}_\infty)(\theta). \tag{3.20}$$

Moreover, note that the fact that $\mathbf{E} \in \{A \in \mathcal{B}(\mathbb{R}^\mathfrak{d}) \colon \int_{\mathbb{R}^\mathfrak{d} \setminus A} 1 \, \mathrm{d}\theta = 0\}$ and the fact that $\mathcal{E} \in \{A \in \mathcal{B}(\mathbb{R}^\mathfrak{d}) \colon \int_{\mathbb{R}^\mathfrak{d} \setminus A} 1 \, \mathrm{d}\theta = 0\}$ ensure that $(\mathbf{E} \cap \mathcal{E}) \in \{A \in \mathcal{B}(\mathbb{R}^\mathfrak{d}) \colon \int_{\mathbb{R}^\mathfrak{d} \setminus A} 1 \, \mathrm{d}\theta = 0\}$. Combining this and the fact that for all $\theta \in (\mathbf{E} \cap \mathcal{E})$ it holds that $\mathcal{L}_\infty$ is differentiable at θ with (3.20) establishes items (i), (ii), and (iii). The proof of Proposition 3.2 is thus complete. □

3.6 Fréchet subdifferentials and limiting Fréchet subdifferentials

Definition 3.1 (Fréchet subdifferentials and limiting Fréchet subdifferentials)**.** *Let $n \in \mathbb{N}$, $f \in C(\mathbb{R}^n, \mathbb{R})$, $x \in \mathbb{R}^n$. Then we denote by $(\mathcal{D}f)(x) \subseteq \mathbb{R}^n$ the set given by*

$$(\mathcal{D}f)(x) = \left\{ y \in \mathbb{R}^n \colon \left[\liminf_{\mathbb{R}^n \setminus \{0\} \ni h \to 0} \left(\frac{f(x+h) - f(x) - \langle y, h \rangle}{\|h\|} \right) \geq 0 \right] \right\} \tag{3.21}$$

and we denote by $(\mathbb{D}f)(x) \subseteq \mathbb{R}^n$ the set given by

$$(\mathbb{D}f)(x) = \textstyle\bigcap_{\varepsilon \in (0,\infty)} \overline{\left[\cup_{y \in \{z \in \mathbb{R}^n \colon \|x - z\| < \varepsilon\}} (\mathcal{D}f)(y) \right]}. \tag{3.22}$$

Lemma 3.3 (Properties of Fréchet subdifferentials)**.** *Let $n \in \mathbb{N}$, $f \in C(\mathbb{R}^n, \mathbb{R})$. Then*

(i) it holds for all $x \in \mathbb{R}^n$ that

$$\begin{aligned}(\mathbb{D}f)(x) = \big\{ y \in \mathbb{R}^n \colon \big[\exists\, z = (z_1, z_2) \colon \mathbb{N} \to \mathbb{R}^n \times \mathbb{R}^n \colon & \\ \big(\big[\forall\, k \in \mathbb{N} \colon z_2(k) \in (\mathcal{D}f)(z_1(k)) \big] & \\ \wedge \big[\limsup\nolimits_{k \to \infty} (\|z_1(k) - x\| + \|z_2(k) - y\|) = 0 \big] \big) \big] \big\}, &\end{aligned} \tag{3.23}$$

(ii) it holds for all $x \in \mathbb{R}^n$ that $(\mathcal{D}f)(x) \subseteq (\mathbb{D}f)(x)$,

(iii) it holds for all $x \in \{y \in \mathbb{R}^n \colon f \text{ is differentiable at } y\}$ that $(\mathcal{D}f)(x) = \{(\nabla f)(x)\}$,

(iv) it holds for all $x \in \cup_{U \subseteq \mathbb{R}^n,\, U \text{ is open},\, f|_U \in C^1(U,\mathbb{R})} U$ *that* $(\mathbb{D}f)(x) = \{(\nabla f)(x)\}$*, and*

(v) it holds for all $x \in \mathbb{R}^n$ *that* $(\mathbb{D}f)(x)$ *is closed.*

(cf. Definition 3.1).

Proof of Lemma 3.3. Throughout this proof let $Z^{x,y} = (Z_1^{x,y}, Z_2^{x,y})\colon \mathbb{N} \to \mathbb{R}^n \times \mathbb{R}^n$, $x, y \in \mathbb{R}^n$, satisfy for all $x, y \in \mathbb{R}^n$, $k \in \mathbb{N}$ that

$$Z_1^{x,y}(k) = x \qquad \text{and} \qquad Z_2^{x,y}(k) = y. \tag{3.24}$$

Observe that (3.22) establishes item (i). Note that (3.24) shows for all $x \in \mathbb{R}^n$, $y \in (\mathcal{D}f)(x)$ that

$$\begin{aligned}&\Big[\forall\, k \in \mathbb{N}\colon \Big(Z_2^{x,y}(k) \in (\mathcal{D}f)(Z_1^{x,y}(k))\Big)\Big]\\ &\wedge \Big[\limsup_{k\to\infty}\big(\|Z_1^{x,y}(k) - x\| + \|Z_2^{x,y}(k) - y\|\big) = 0\Big].\end{aligned} \tag{3.25}$$

This establishes item (ii). Next observe that Rockafellar & Wets [63, Exercise 8.8] establishes items (iii) and (iv). Finally, [63, Theorem 8.6] establishes item (v). The proof of Lemma 3.3 is thus complete. □

Lemma 3.4 (Limits of limiting Fréchet subgradients)**.** *Let* $n \in \mathbb{N}$, $f \in C(\mathbb{R}^n, \mathbb{R})$, *let* $(x_k)_{k\in\mathbb{N}_0} \subseteq \mathbb{R}^n$ *and* $(y_k)_{k\in\mathbb{N}_0} \subseteq \mathbb{R}^n$ *satisfy* $\limsup_{k\to\infty}(\|x_k - x_0\| + \|y_k - y_0\|) = 0$, *and assume for all* $k \in \mathbb{N}$ *that* $y_k \in (\mathbb{D}f)(x_k)$ *(cf. Definition 3.1). Then* $y_0 \in (\mathbb{D}f)(x_0)$.

Proof of Lemma 3.4. Note that, e.g., [63, Proposition 8.7] implies that $y_0 \in (\mathbb{D}f)(x_0)$. The proof of Lemma 3.4 is thus complete. □

3.7 Local underestimates for the realization functions of DNNs

Next we establish in Lemma 3.5 a technical lemma that will be used in the proof of Lemma 3.7 below. Roughly speaking, since we work with the left derivative of the ReLU function we need to approximate the realization functions from below to obtain convergence of the generalized gradients.

Lemma 3.5. *Assume Setting 3.1 and let* $\theta \in \mathbb{R}^{\mathfrak{d}}$, $\varepsilon \in (0, \infty)$. *Then there exists a non-empty and open* $U \subseteq \mathbb{R}^{\mathfrak{d}}$ *such that for all* $\vartheta \in U$, $k \in \{1, 2, \ldots, L\}$, $i \in \{1, 2, \ldots, \ell_k\}$, $x \in [a, b]^{\ell_0}$ *it holds that*

$$\|\vartheta - \theta\| < \varepsilon \qquad \textit{and} \qquad \mathcal{N}_{\infty,i}^{k,\vartheta}(x) \le \mathcal{N}_{\infty,i}^{k,\theta}(x). \tag{3.26}$$

Proof of Lemma 3.5. Throughout this proof let $\mathfrak{C}_k \in (0, \infty)$, $k \in \mathbb{N}$, satisfy for all $k \in \mathbb{N}$ that $\mathfrak{C}_1 = \max\{\ell_0|a|, \ell_0|b|, 1\}$ and

$$\mathfrak{C}_{k+1} = 2\mathfrak{C}_k(k+1)\max\{1, |a|, |b|\}(\max\{1, \|\theta\| + 2\mathfrak{C}_k\})^k \big[\textstyle\prod_{j=0}^{k}(\ell_j + 1)\big], \tag{3.27}$$

let $\delta \in (0,\infty)$ satisfy $\delta = \min\{1, \varepsilon(2\mathfrak{C}_L\mathfrak{d})^{-1}\}$, and let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy

$$U = \Big\{\vartheta \in \mathbb{R}^{\mathfrak{d}} \colon \Big(\Big[\forall\, k \in \{1,\dots,L\}, i \in \{1,\dots,\ell_k\}, j \in \{1,\dots,\ell_{k-1}\} \colon |\mathfrak{w}^{k,\vartheta}_{i,j} - \mathfrak{w}^{k,\theta}_{i,j}| < \delta\Big] \\ \wedge \Big[\forall\, k \in \{1,\dots,L\}, i \in \{1,\dots,\ell_k\} \colon \mathfrak{b}^{k,\theta}_i - 2\mathfrak{C}_k\delta < \mathfrak{b}^{k,\vartheta}_i < \mathfrak{b}^{k,\theta}_i - \mathfrak{C}_k\delta\Big]\Big)\Big\}. \quad (3.28)$$

Observe that (3.28) ensures that $U \subseteq \mathbb{R}^{\mathfrak{d}}$ is non-empty and open. Furthermore, note that (3.27) shows for all $k \in \mathbb{N}$ that $\mathfrak{C}_{k+1} > 2\mathfrak{C}_k$ and $\mathfrak{C}_k \geq 1$. Combining this with (3.28) assures for all $\vartheta \in U$, $i \in \{1,2,\dots,\mathfrak{d}\}$ that

$$|\vartheta_i - \theta_i| < \max\{\delta, 2\mathfrak{C}_1\delta, 2\mathfrak{C}_2\delta, \dots, 2\mathfrak{C}_L\delta\} = 2\mathfrak{C}_L\delta \leq 2\mathfrak{C}_L\big(\varepsilon(2\mathfrak{C}_L\mathfrak{d})^{-1}\big) = \mathfrak{d}^{-1}\varepsilon. \quad (3.29)$$

Therefore, we obtain for all $\vartheta \in U$ that

$$\|\vartheta - \theta\| = \big[\textstyle\sum_{i=1}^{\mathfrak{d}} |\vartheta_i - \theta_i|^2\big]^{1/2} \leq \mathfrak{d}\big[\max_{i \in \{1,2,\dots,L\}} |\vartheta_i - \theta_i|\big] < \varepsilon. \quad (3.30)$$

Moreover, observe that (3.2), (3.5), and (3.28) ensure for all $\vartheta \in U$, $i \in \{1,2,\dots,\ell_1\}$, $x = (x_1,\dots,x_{\ell_0}) \in [a,b]^{\ell_0}$ that

$$\begin{aligned} \mathcal{N}^{1,\vartheta}_{\infty,i}(x) - \mathcal{N}^{1,\theta}_{\infty,i}(x) &= (\mathfrak{b}^{1,\vartheta}_i - \mathfrak{b}^{1,\theta}_i) + \textstyle\sum_{j=1}^{\ell_0} (\mathfrak{w}^{1,\vartheta}_{i,j} - \mathfrak{w}^{1,\theta}_{i,j}) x_j \\ &< -\mathfrak{C}_1\delta + \textstyle\sum_{j=1}^{\ell_0} |(\mathfrak{w}^{1,\vartheta}_{i,j} - \mathfrak{w}^{1,\theta}_{i,j}) x_j| \qquad (3.31) \\ &\leq -\mathfrak{C}_1\delta + \delta\big(\textstyle\sum_{j=1}^{\ell_0} |x_j|\big) \leq -\mathfrak{C}_1\delta + \ell_0\delta\max\{|a|,|b|\} \leq 0. \qquad (3.32) \end{aligned}$$

It thus remains to prove an analogous inequality for the subsequent layers. For this let $\vartheta \in U$, $k \in \mathbb{N} \cap (0,L)$, $i \in \{1,2,\dots,\ell_{k+1}\}$, $x \in [a,b]^{\ell_0}$, let $\mathbf{d} \in \mathbb{N}$ satisfy $\mathbf{d} = \ell_{k+1}\ell_k + 1 + \sum_{j=1}^{k} \ell_j(\ell_{j-1}+1)$, let $\mathbf{D} \in \mathbb{N}$ satisfy $\mathbf{D} = \sum_{j=1}^{k+1} \ell_j(\ell_{j-1}+1)$, and let $\psi \in \mathbb{R}^{\mathfrak{d}}$ satisfy

$$\psi = (\vartheta_1, \vartheta_2, \dots, \vartheta_{\mathbf{d}-1}, \theta_{\mathbf{d}}, \theta_{\mathbf{d}+1}, \dots, \theta_{\mathfrak{d}}). \quad (3.33)$$

Note that (3.2), (3.5), (3.28), and (3.33) show that

$$\begin{aligned} \mathcal{N}^{k+1,\vartheta}_{\infty,i}(x) &= \mathcal{N}^{k+1,\psi}_{\infty,i}(x) + (\mathfrak{b}^{k+1,\vartheta}_i - \mathfrak{b}^{k+1,\psi}_i) \\ &= \mathcal{N}^{k+1,\psi}_{\infty,i}(x) + (\mathfrak{b}^{k+1,\vartheta}_i - \mathfrak{b}^{k+1,\theta}_i) < \mathcal{N}^{k+1,\psi}_{\infty,i}(x) - \mathfrak{C}_{k+1}\delta. \qquad (3.34) \end{aligned}$$

Next note that, e.g., [6, Theorem 2.1] (applied with $a \curvearrowleft a$, $b \curvearrowleft b$, $d \curvearrowleft \mathbf{D}$, $L \curvearrowleft k+1$, $\ell \curvearrowleft (\ell_0,\ell_1,\dots,\ell_{k+1})$ in the notation of [6, Theorem 2.36]) demonstrates that

$$\begin{aligned} &|\mathcal{N}^{k+1,\theta}_{\infty,i}(x) - \mathcal{N}^{k+1,\psi}_{\infty,i}(x)| \\ &\leq (k+1)\max\{1,|a|,|b|\}\Big[\max\Big\{1, \max_{i\in\{1,2,\dots,\mathbf{D}\}}|\theta_i|, \max_{i\in\{1,2,\dots,\mathbf{D}\}}|\psi_i|\Big\}\Big]^k \\ &\quad \cdot \big[\textstyle\prod_{m=0}^{k}(\ell_m+1)\big]\Big[\max_{i\in\{1,2,\dots,\mathbf{D}\}}|\theta_i - \psi_i|\Big]. \qquad (3.35) \end{aligned}$$

In addition, observe that (3.28) ensures that

$$\begin{aligned}\max_{i\in\{1,2,\dots,\mathbf{D}\}}|\theta_i-\psi_i| &= \max_{i\in\{1,2,\dots,\mathbf{d}-1\}}|\theta_i-\vartheta_i|\\ &\le \max\{\delta,2\mathfrak{C}_1\delta,2\mathfrak{C}_2\delta,\dots,2\mathfrak{C}_k\delta\} = 2\mathfrak{C}_k\delta.\end{aligned} \tag{3.36}$$

Combining this with (3.27) and (3.35) proves that

$$\begin{aligned}&|\mathcal{N}^{k+1,\theta}_{\infty,i}(x)-\mathcal{N}^{k+1,\psi}_{\infty,i}(x)|\\ &\le (k+1)\max\{1,|a|,|b|\}\big[\max\big\{1,2\mathfrak{C}_k\delta+\max_{i\in\{1,2,\dots,\mathbf{D}\}}|\theta_i|,\max_{i\in\{1,2,\dots,\mathbf{D}\}}|\theta_i|\big\}\big]^k\\ &\quad\cdot\big[\textstyle\prod_{m=0}^{k}(\ell_m+1)\big]\big[2\mathfrak{C}_k\delta\big]\\ &\le 2\mathfrak{C}_k\delta(k+1)\max\{1,|a|,|b|\}\big[\textstyle\prod_{m=0}^{k}(\ell_m+1)\big]\big[\max\{1,\|\theta\|+2\mathfrak{C}_k\}\big]^k = \mathfrak{C}_{k+1}\delta.\end{aligned} \tag{3.37}$$

This and (3.34) assure that

$$\begin{aligned}\mathcal{N}^{k+1,\vartheta}_{\infty,i}(x) &< \mathcal{N}^{k+1,\psi}_{\infty,i}(x)-\mathfrak{C}_{k+1}\delta = \mathcal{N}^{k+1,\theta}_{\infty,i}(x)+\big(\mathcal{N}^{k+1,\psi}_{\infty,i}(x)-\mathcal{N}^{k+1,\theta}_{\infty,i}(x)\big)-\mathfrak{C}_{k+1}\delta\\ &\le \mathcal{N}^{k+1,\theta}_{\infty,i}(x)+|\mathcal{N}^{k+1,\psi}_{\infty,i}(x)-\mathcal{N}^{k+1,\theta}_{\infty,i}(x)|-\mathfrak{C}_{k+1}\delta \le \mathcal{N}^{k+1,\theta}_{\infty,i}(x).\end{aligned} \tag{3.38}$$

The proof of Lemma 3.5 is thus complete. □

3.8 Continuity properties for the generalized gradient function

Lemma 3.6 (Continuity points of the generalized gradient function)**.** *Assume Setting 3.1 and let $\theta=(\theta_n)_{n\in\mathbb{N}_0}\colon\mathbb{N}_0\to\mathbb{R}^{\mathfrak{d}}$ satisfy for all $k\in\{1,2,\dots,L\}$, $i\in\{1,2,\dots,\ell_k\}$, $x\in[a,b]^{\ell_0}$ that*

$$\limsup_{n\to\infty}\big(\|\theta_n-\theta_0\|+|\mathbb{1}_{(0,\infty)}(\mathcal{N}^{k,\theta_n}_{\infty,i}(x))-\mathbb{1}_{(0,\infty)}(\mathcal{N}^{k,\theta_0}_{\infty,i}(x))|\big)=0. \tag{3.39}$$

Then $\limsup_{n\to\infty}\|\mathcal{G}(\theta_n)-\mathcal{G}(\theta_0)\|=0$.

Proof of Lemma 3.6. Note that Lemma 3.1 (applied for every $k\in\{1,2,\dots,L\}$ with $L\curvearrowleft k$ in the notation of Lemma 3.1), (3.1), (3.2), and (3.5) assure that for all $k\in\{1,2,\dots,L\}$, $j\in\{1,2,\dots,\ell_k\}$ it holds that

$$\limsup_{n\to\infty}\sup_{x\in[a,b]^{\ell_0}}|\mathcal{N}^{k,\theta_n}_{\infty,j}(x)-\mathcal{N}^{k,\theta_0}_{\infty,j}(x)|=0. \tag{3.40}$$

Furthermore, observe that (3.5) and (3.39) ensure for all $x\in[a,b]^{\ell_0}$, $k\in\{1,2,\dots,L\}$, $i\in\{1,2,\dots,\ell_k\}$ that $\limsup_{n\to\infty}|\mathbb{1}_{\mathcal{X}^{k,\theta_n}_i}(x)-\mathbb{1}_{\mathcal{X}^{k,\theta_0}_i}(x)|=0$. Combining this, (3.39), and (3.40) with Proposition 3.1 and Lebesgue's dominated convergence theorem establishes that $\limsup_{n\to\infty}\|\mathcal{G}(\theta_n)-\mathcal{G}(\theta_0)\|=0$. The proof of Lemma 3.6 is thus complete. □

As a consequence of Lemmas 3.5 and 3.6 we show in Lemma 3.7 that, loosely speaking, the generalized gradient $\mathcal{G}(\theta)$ at an arbitrary point $\theta\in\mathbb{R}^{\mathfrak{d}}$ can be represented as the limit of generalized gradients of a sequence $\vartheta_n\to\theta$, even after removing an arbitrary set of zero measure.

Lemma 3.7. *Assume Setting 3.1 and let $\theta \in \mathbb{R}^{\mathfrak{d}}$, $E \in \mathcal{B}(\mathbb{R}^{\mathfrak{d}})$ satisfy $\int_{\mathbb{R}^{\mathfrak{d}} \setminus E} 1 \, d\vartheta = 0$. Then there exists $\vartheta = (\vartheta_n)_{n \in \mathbb{N}} \colon \mathbb{N} \to E$ such that*

$$\limsup_{n\to\infty} \big(\|\vartheta_n - \theta\| + \|\mathcal{G}(\vartheta_n) - \mathcal{G}(\theta)\| \big) = 0. \tag{3.41}$$

Proof of Lemma 3.7. Note that Lemma 3.5 assures that there exist non-empty and open $U_n \subseteq \mathbb{R}^{\mathfrak{d}}$, $n \in \mathbb{N}$, which satisfy for all $n \in \mathbb{N}$, $\vartheta \in U_n$, $k \in \{1, 2, \ldots, L\}$, $i \in \{1, 2, \ldots, \ell_k\}$, $x \in [a, b]^{\ell_0}$ that

$$\|\vartheta - \theta\| < \tfrac{1}{n} \qquad \text{and} \qquad \mathcal{N}^{k,\vartheta}_{\infty,i}(x) \leq \mathcal{N}^{k,\theta}_{\infty,i}(x). \tag{3.42}$$

Observe that the assumption that $\int_{\mathbb{R}^{\mathfrak{d}} \setminus E} 1 \, d\vartheta = 0$ implies for all $n \in \mathbb{N}$ that $(U_n \cap E) \neq \varnothing$. In the following let $\vartheta = (\vartheta_n)_{n \in \mathbb{N}} \colon \mathbb{N} \to E$ satisfy for all $n \in \mathbb{N}$ that

$$\vartheta_n \in U_n. \tag{3.43}$$

Note that (3.42) assures for all $n \in \mathbb{N}$ that $\|\vartheta_n - \theta\| < \frac{1}{n}$. Hence, we obtain that

$$\limsup_{n\to\infty} \|\vartheta_n - \theta\| = 0. \tag{3.44}$$

Lemma 3.1 (applied for every $k \in \{1, 2, \ldots, L\}$ with $L \curvearrowleft k$ in the notation of Lemma 3.1) therefore implies that for all $k \in \{1, 2, \ldots, L\}$, $i \in \{1, 2, \ldots, \ell_k\}$, $x \in [a, b]^{\ell_0}$ we have that

$$\limsup_{n\to\infty} |\mathcal{N}^{k,\vartheta_n}_{\infty,i}(x) - \mathcal{N}^{k,\theta}_{\infty,i}(x)| = 0. \tag{3.45}$$

Furthermore, observe that (3.42) and (3.43) assure for all $n \in \mathbb{N}$, $k \in \{1, 2, \ldots, L\}$, $i \in \{1, 2, \ldots, \ell_k\}$, $x \in [a, b]^{\ell_0}$ that $\mathcal{N}^{k,\vartheta_n}_{\infty,i}(x) \leq \mathcal{N}^{k,\theta}_{\infty,i}(x)$. Combining this and (3.45) with the fact that the function $\mathbb{R} \ni x \mapsto \mathbb{1}_{(0,\infty)}(x) \in \mathbb{R}$ is left continuous demonstrates for all $x \in [a, b]^{\ell_0}$, $k \in \{1, 2, \ldots, L\}$, $i \in \{1, 2, \ldots, \ell_k\}$ that $\limsup_{n\to\infty} |\mathbb{1}_{(0,\infty)}(\mathcal{N}^{k,\vartheta_n}_{\infty,i}(x)) - \mathbb{1}_{(0,\infty)}(\mathcal{N}^{k,\theta}_{\infty,i}(x))| = 0$. Lemma 3.6 and (3.44) hence show that $\limsup_{n\to\infty} \|\mathcal{G}(\vartheta_n) - \mathcal{G}(\theta)\| = 0$. Combining this with (3.44) establishes (3.41). The proof of Lemma 3.7 is thus complete. ☐

3.9 Generalized gradients as limiting Fréchet subdifferentials

We next employ the differentiability result from Proposition 3.2, the approximation result for the generalized gradient from Lemma 3.7, and the definition of the limiting Fréchet subdifferential to establish in Proposition 3.3 the main result of this section: For every $\theta \in \mathbb{R}^{\mathfrak{d}}$, the generalized gradient $\mathcal{G}(\theta)$ is an element of the limiting Fréchet subdifferential $(\mathbb{D}\mathcal{L}_\infty)(\theta)$.

Proposition 3.3. *Assume Setting 3.1 and let $\theta \in \mathbb{R}^{\mathfrak{d}}$. Then $\mathcal{G}(\theta) \in (\mathbb{D}\mathcal{L}_\infty)(\theta)$ (cf. Definition 3.1).*

Proof of Proposition 3.3. Note that Proposition 3.2 ensures that there exists $E \in \mathcal{B}(\mathbb{R}^{\mathfrak{d}})$ which satisfies $\int_{\mathbb{R}^{\mathfrak{d}} \setminus E} 1 \, d\vartheta = 0$, which satisfies for all $\vartheta \in E$ that $\mathcal{L}_\infty$ is differentiable at ϑ, and which satisfies for all $\vartheta \in E$ that

$$(\nabla \mathcal{L}_\infty)(\vartheta) = \mathcal{G}(\vartheta). \tag{3.46}$$

Observe that (3.46) and Lemma 3.3 prove for all $\vartheta \in E$ that

$$\mathcal{G}(\vartheta) \in (\mathcal{D}\mathcal{L}_\infty)(\vartheta). \tag{3.47}$$

Furthermore, note that the fact that $\int_{\mathbb{R}^{\mathfrak{d}} \backslash E} 1 \, \mathrm{d}\vartheta = 0$ and Lemma 3.7 imply that there exists $\vartheta = (\vartheta_n)_{n \in \mathbb{N}} \colon \mathbb{N} \to E$ which satisfies

$$\limsup_{n \to \infty} \big(\|\vartheta_n - \theta\| + \|\mathcal{G}(\vartheta_n) - \mathcal{G}(\theta)\| \big) = 0. \tag{3.48}$$

Observe that (3.47) and (3.48) demonstrate that $\mathcal{G}(\theta) \in (\mathbb{D}\mathcal{L}_\infty)(\theta)$. The proof of Proposition 3.3 is thus complete. ☐

Finally, as a consequence of Proposition 3.3 we show in Corollary 3.2 that on every open set on which the risk function $\mathcal{L}_\infty$ is continuously differentiable its gradient agrees with $\mathcal{G}$. This fact will be used in the convergence analysis of GD processes in Section 8.

Corollary 3.2. *Assume Setting 3.1. Then it holds for all* $\theta \in \cup_{U \subseteq \mathbb{R}^{\mathfrak{d}},\, U \text{ is open},\, (\mathcal{L}_\infty)|_U \in C^1(U, \mathbb{R})} U$ *that* $\mathcal{G}(\theta) = (\nabla \mathcal{L}_\infty)(\theta)$ *(cf. Definition 3.1).*

Proof of Corollary 3.2. Note that item (iv) in Lemma 3.3 (applied with $n \curvearrowleft \mathfrak{d}$, $f \curvearrowleft \mathcal{L}_\infty$ in the notation of Lemma 3.3) and Proposition 3.3 ensure that for all open $U \subseteq \mathbb{R}^{\mathfrak{d}}$ and all $\theta \in U$ with $(\mathcal{L}_\infty)|_U \in C^1(U, \mathbb{R})$ it holds that

$$\mathcal{G}(\theta) \in (\mathbb{D}\mathcal{L}_\infty)(\theta) = \{(\nabla \mathcal{L}_\infty)(\theta)\}. \tag{3.49}$$

Therefore, we obtain for all open $U \subseteq \mathbb{R}^{\mathfrak{d}}$ and all $\theta \in U$ with $(\mathcal{L}_\infty)|_U \in C^1(U, \mathbb{R})$ that $\mathcal{G}(\theta) = (\nabla \mathcal{L}_\infty)(\theta)$. The proof of Corollary 3.2 is thus complete. ☐

4 Suitable piecewise rational functions

In this section we identify in (4.1) in Definition 4.1 a suitable subclass of the class of semi-algebraic functions which is closed under integration (see Proposition 4.3 in Subsection 4.7 below) and which contains the realization functions of deep ReLU ANNs (see Proposition 4.4 in Subsection 4.9 below). The fact that functions in this class of suitable piecewise rational functions are semi-algebraic is established in Proposition 4.2 below. We also summarize in Subsection 4.3 some basic facts regarding semi-algebraic sets and functions. The results from this section will be employed in Section 5 below to establish that the considered risk function in the training of deep ANNs with ReLU activation are semi-algebraic.

Closedness under integration is not a trivial issue due to the fact that, in general, the integral of a semi-algebraic function is not necessarily semi-algebraic (in fact, in general not even globally subanalytic, see Kaiser [44]). Our analysis of the integrals of the functions considered in Definition 4.1 below crucially relies on the fact that they are piecewise rational on regions separated by hyperplanes in the x-component. This property is also satisfied by the realization functions of ANNs with ReLU activation.

The function class in Definition 4.1 and some of the results in this section are inspired by the findings in our previous article Eberle et al. [28, Section 4]. In particular, Definition 4.1 extends [28, Definition 4.6], Proposition 4.2 in Subsection 4.6 below extends [28, Lemma 4.7], and Proposition 4.3 in Subsection 4.7 below extends [28, Proposition 4.8].

4.1 Suitable piecewise rational functions

Definition 4.1 (Vector spaces of suitable piecewise rational functions)**.** *Let $m, n \in \mathbb{N}_0$, $\delta \in (0, \infty]$. Then we denote by $\mathscr{F}_{m,n,\delta}$ the $\mathbb{R}$-vector space given by*

$$\mathscr{F}_{m,n,\delta} = \operatorname{span}_{\mathbb{R}}\Bigg(\Bigg\{F\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}\colon \Bigg[\exists\, r \in \mathbb{N},\, R \in \mathscr{R}_{m,\delta},\, Q \in \{q \in \mathscr{P}_n\colon \deg(q) \le \delta\},$$
$$P = (P_{i,j})_{(i,j)\in\{1,2,\dots,r\}\times\{0,1,\dots,n\}} \subseteq \mathscr{P}_m\colon \Bigg(\forall\, \theta \in \mathbb{R}^m,\, x = (x_1, \dots, x_n) \in \mathbb{R}^n\colon$$
$$f(\theta, x) = R(\theta)Q(x)\left[\prod_{i=1}^{r} \mathbb{1}_{[0,\infty)}(P_{i,0}(\theta) + \textstyle\sum_{j=1}^{n} P_{i,j}(\theta)x_j)\right]\Bigg)\Bigg]\Bigg\}\Bigg) \tag{4.1}$$

(cf. Definitions 4.2, 4.5, and 4.6).

In (4.1) above we denote by $\operatorname{span}_{\mathbb{R}}$ the linear span of the given functions $F\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ with coefficients in $\mathbb{R}$, i.e., the $\mathbb{R}$-vectorspace generated by these functions.

Observe that functions in $\mathscr{F}_{m,n,\delta}$ depend on two vectors $\theta \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$. In the considered deep learning framework this will be applied in the situation where θ is the parameter vector of a suitable ANN, x is the input vector of the ANN, and $f(\theta, x)$ is the output.

4.2 Elementary properties of suitable piecewise rational functions

Lemma 4.1. *Let $m, n \in \mathbb{N}_0$. Then*

(i) it holds for all $\delta_1, \delta_2 \in (0, \infty]$ with $\delta_1 \le \delta_2$ that $\mathscr{R}_{n,\delta_1} \subseteq \mathscr{R}_{n,\delta_2}$,

(ii) it holds for all $\delta \in (0, 1]$ that $\mathscr{R}_{n,\delta} = \mathscr{P}_n$,

(iii) it holds for all $\delta_1, \delta_2 \in (0, \infty]$ with $\delta_1 \le \delta_2$ that $\mathscr{F}_{m,n,\delta_1} \subseteq \mathscr{F}_{m,n,\delta_2}$,

(iv) it holds that $\mathscr{F}_{m,n,1} \subseteq \mathscr{F}_{m,n,\infty}$,

(v) it holds for all $f, g \in \mathscr{F}_{m,n,\infty}$ that

$$\big(\mathbb{R}^m \times \mathbb{R}^n \ni (\theta, x) \mapsto f(\theta, x)g(\theta, x) \in \mathbb{R}\big) \in \mathscr{F}_{m,n,\infty}, \tag{4.2}$$

(vi) it holds for all $\delta \in (0, \infty]$ that

$$\mathscr{F}_{m,n,\delta} = \operatorname{span}_{\mathbb{R}}\Big(\Big\{F\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}\colon \Big[\exists\, r \in \mathbb{N},$$
$$A_1, A_2, \dots, A_r \in \{\{0\}, [0, \infty), (0, \infty)\}, R \in \mathscr{R}_{m,\delta},\, Q \in \{q \in \mathscr{P}_n\colon \deg(q) \le \delta\},$$
$$P = (P_{i,j})_{(i,j)\in\{1,2,\dots,r\}\times\{0,1,\dots,n\}} \subseteq \mathscr{P}_m\colon \Big(\forall\, \theta \in \mathbb{R}^m\colon \forall\, x = (x_1, \dots, x_n) \in \mathbb{R}^n\colon$$
$$f(\theta, x) = R(\theta)Q(x)\big[\textstyle\prod_{i=1}^{r} \mathbb{1}_{A_i}(P_{i,0}(\theta) + \sum_{j=1}^{n} P_{i,j}(\theta)x_j)\big]\Big)\Big]\Big\}\Big) \tag{4.3}$$

(cf. Definition 4.1).

Proof of Lemma 4.1. Note that (4.13) and the fact that for all $\delta_1, \delta_2 \in (0, \infty]$ with $\delta_1 \leq \delta_2$ it holds that

$$\{q \in \mathscr{P}_n \colon \deg(q) < \delta_1\} \subseteq \{q \in \mathscr{P}_n \colon \deg(q) < \delta_2\} \tag{4.4}$$

establish item (i). Observe that (4.13) and the fact that for all $\delta \in (0, 1]$ it holds that

$$\{q \in \mathscr{P}_n \colon \deg(q) < \delta\} = \{q \in \mathscr{P}_n \colon \deg(q) = 0\} \tag{4.5}$$

prove item (ii). Note that (4.1), item (i), and the fact that for all $\delta_1, \delta_2 \in (0, \infty]$ with $\delta_1 \leq \delta_2$ it holds that

$$\{q \in \mathscr{P}_n \colon \deg(q) \leq \delta_1\} \subseteq \{q \in \mathscr{P}_n \colon \deg(q) \leq \delta_2\} \tag{4.6}$$

establish item (iii). Observe that item (iii) proves item (iv). Note that (4.1) establishes item (v). Observe that the fact that $\forall\, y \in \mathbb{R} \colon \mathbb{1}_{\{0\}}(y) = \mathbb{1}_{(-\infty,0]\cap[0,\infty)}(y) = \mathbb{1}_{(-\infty,0]}(y)\mathbb{1}_{[0,\infty)}(y) = \mathbb{1}_{[0,\infty)}(y)\mathbb{1}_{[0,\infty)}(-y)$ shows that for all $P_0, P_1, \ldots, P_n \in \mathscr{P}_m$ it holds that

$$\begin{aligned} &\mathbb{1}_{\{0\}}(P_0(\theta) + \textstyle\sum_{j=1}^n P_j(\theta)x_j) \\ &= \mathbb{1}_{[0,\infty)}(P_0(\theta) + \textstyle\sum_{j=1}^n P_j(\theta)x_j)\mathbb{1}_{[0,\infty)}(-P_0(\theta) + \textstyle\sum_{j=1}^n (-P_j(\theta))x_j). \end{aligned} \tag{4.7}$$

Furthermore, note that the fact that $\forall\, y \in \mathbb{R} \colon \mathbb{1}_{(0,\infty)}(y) = 1 - \mathbb{1}_{(-\infty,0]}(y) = 1 - \mathbb{1}_{[0,\infty)}(-y)$ shows that for all $P_0, P_1, \ldots, P_n \in \mathscr{P}_m$ it holds that

$$\mathbb{1}_{(0,\infty)}(P_0(\theta) + \textstyle\sum_{j=1}^n P_j(\theta)x_j) = 1 - \mathbb{1}_{[0,\infty)}(-P_0(\theta) + \textstyle\sum_{j=1}^n (-P_j(\theta))x_j). \tag{4.8}$$

Combining (4.1) and (4.7) hence shows that for all $\delta \in (0, \infty]$, $r \in \mathbb{N}$, $A_1, A_2, \ldots, A_r \in \{\{0\}, (0, \infty), [0, \infty)\}$, $R \in \mathscr{R}_{m,\delta}$, $Q \in \{q \in \mathscr{P}_n \colon \deg(q) \leq \delta\}$, $P = (P_{i,j})_{\{1,2,\ldots,r\}\times\{0,1,\ldots,n\}} \subseteq \mathscr{P}_m$ it holds that

$$\big(\mathbb{R}^m \times \mathbb{R}^n \ni (\theta, x) \mapsto R(\theta)Q(x)\big[\textstyle\prod_{i=1}^r \mathbb{1}_{A_i}(P_0(\theta) + \textstyle\sum_{j=1}^n P_j(\theta)x_j)\big] \in \mathbb{R}\big) \in \mathscr{F}_{m,n,\delta}. \tag{4.9}$$

This establishes item (vi). The proof of Lemma 4.1 is thus complete. □

4.3 Semi-algebraic sets

In the following we gather several known definitions and elementary results regarding semi-algebraic sets and functions; cf., e.g., Bochnak et al. [9], Coste [18], Shiota [67], and Van den Dries & Miller [69].

Definition 4.2 (Set of polynomials)**.** *Let $n \in \mathbb{N}_0$. Then we denote by $\mathscr{P}_n \subseteq C(\mathbb{R}^n, \mathbb{R})$ the set*[9] *of all polynomials from $\mathbb{R}^n$ to $\mathbb{R}$.*

[9]Note that $\mathbb{R}^0 = \{0\}$, $C(\mathbb{R}^0, \mathbb{R}) = C(\{0\}, \mathbb{R})$, and $\#(C(\mathbb{R}^0, \mathbb{R})) = \#(C(\{0\}, \mathbb{R})) = \infty$. In particular, this shows for all $n \in \mathbb{N}_0$ that $\dim(\mathbb{R}^n) = n$ and $\#(C(\mathbb{R}^n, \mathbb{R})) = \infty$.

Definition 4.3 (Multidimensional semi-algebraic sets)**.** *Let $n \in \mathbb{N}$ and let $A \subseteq \mathbb{R}^n$ be a set. Then we say that A is an n-dimensional semi-algebraic set if and only if there exist $M, N \in \mathbb{N}$ and $(P_{i,j,k})_{(i,j,k)\in\{1,2,\ldots,M\}\times\{1,2,\ldots,N\}\times\{0,1\}} \subseteq \mathscr{P}_n$ such that*

$$A = \cup_{i=1}^{M}\big(\cap_{j=1}^{N}\{x \in \mathbb{R}^n : P_{i,j,0}(x) = 0 < P_{i,j,1}(x)\}\big) \tag{4.10}$$

(cf. Definition 4.2).

Note that in (4.10) we have that $\{x \in \mathbb{R}^n : P_{i,j,0}(x) = 0 < P_{i,j,1}(x)\} = \{x \in \mathbb{R}^n : [P_{i,j,0}(x) = 0 \wedge P_{i,j,1}(x) > 0]\} = \{x \in \mathbb{R}^n : P_{i,j,0}(x) = 0\} \cap \{x \in \mathbb{R}^n : P_{i,j,1}(x) > 0\}$.

The following properties of semi-algebraic sets are well-known and not hard to show from the definition; see, e.g., Shiota [67, (I.2.9)].

Proposition 4.1. *Let $m, n \in \mathbb{N}$. Then*

(i) it holds for all n-dimensional semi-algebraic sets A, B that $A \cup B$, $A \cap B$, and $\mathbb{R}^n \backslash A$ are n-dimensional semi-algebraic sets,

(ii) it holds for every n-dimensional semi-algebraic set A and every m-dimensional semi-algebraic set B that $A \times B$ is an $(m+n)$-dimensional semi-algebraic set,

(iii) it holds for every $P \in \mathscr{P}_n$ that $\{x \in \mathbb{R}^n : P(x) \geq 0\}$ is an n-dimensional semi-algebraic set,

(iv) it holds for all $a \in \mathbb{R}^n$ that $\{a\} \subseteq \mathbb{R}^n$ is an n-dimensional semi-algebraic set

(cf. Definitions 4.2 and 4.3).

4.4 Semi-algebraic functions

Definition 4.4 (Semi-algebraic functions)**.** *Let $m, n \in \mathbb{N}$ and let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a function. Then we say that f is a semi-algebraic function (we say that f is semi-algebraic) if and only if it holds that* $\operatorname{Graph}(f)$ *is an $(m+n)$-dimensional semi-algebraic set (cf. Definition 4.3).*

The next elementary result, Lemma 4.2, is a direct consequence of, e.g., [67, (I.2.9)] or [9, Proposition 2.2.6] (see, e.g., also Bierstone & Milman [8, Section 1]).

Lemma 4.2. *Let $n \in \mathbb{N}$ and let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ be semi-algebraic (cf. Definition 4.4). Then*

(i) it holds that $\mathbb{R}^n \ni x \mapsto f(x) + g(x) \in \mathbb{R}$ is semi-algebraic and

(ii) it holds that $\mathbb{R}^n \ni x \mapsto f(x)g(x) \in \mathbb{R}$ is semi-algebraic.

Lemma 4.3. *Let $n \in \mathbb{N}$ and let $A \subseteq \mathbb{R}^n$ be an n-dimensional semi-algebraic set (cf. Definition 4.3). Then $\mathbb{R}^n \ni x \mapsto \mathbb{1}_A(x) \in \mathbb{R}$ is semi-algebraic (cf. Definition 4.4).*

Proof of Lemma 4.3. Throughout this proof let $f : \mathbb{R}^n \to \mathbb{R}$ satisfy for all $x \in \mathbb{R}^n$ that

$$f(x) = \mathbb{1}_A(x). \tag{4.11}$$

Observe that (4.11) shows that

$$\mathrm{Graph}(f) = (A \times \{1\}) \cup ((\mathbb{R}^n \backslash A) \times \{0\}) \subseteq \mathbb{R}^{n+1}. \tag{4.12}$$

Furthermore, note that Proposition 4.1 ensures that $\{0\}$ and $\{1\}$ are 1-dimensional semi-algebraic sets and that $\mathbb{R}^n \backslash A$ is an n-dimensional semi-algebraic set. Combining this with Proposition 4.1 shows that $A \times \{1\}$ and $(\mathbb{R}^n \backslash A) \times \{0\}$ are $(n+1)$-dimensional semi-algebraic sets. Proposition 4.1 and (4.12) therefore show that $\mathrm{Graph}(f)$ is an $(n+1)$-dimensional semi-algebraic set. This establishes that f is semi-algebraic. The proof of Lemma 4.3 is thus complete. ☐

4.5 Rational functions as semi-algebraic functions

The next goal is to establish in Proposition 4.2 below that the functions in the classes $\mathscr{F}_{m,0,\infty}$, $m \in \mathbb{N}$, are semi-algebraic. As a preparation, we first recall in Lemma 4.4 below the simple fact that rational functions are semi-algebraic.

Definition 4.5 (Degree[10] of a polynomial)**.** *Let $n \in \mathbb{N}_0$, $P \in \mathscr{P}_n$ (cf. Definition 4.2). Then we denote by* $\deg(P) \in \mathbb{N}_0$ *the degree of P.*

Definition 4.6 (Sets of suitable rational functions)**.** *Let $n \in \mathbb{N}_0$, $\delta \in (0, \infty]$. Then we denote by $\mathscr{R}_{n,\delta}$ the set given by*

$$\mathscr{R}_{n,\delta} = \Bigg\{ R \colon \mathbb{R}^n \to \mathbb{R} \colon \bigg(\exists\, P \in \mathscr{P}_n,\, Q \in \{q \in \mathscr{P}_n \colon \deg(q) < \delta\} \colon \left[\forall\, x \in \mathbb{R}^n \colon R(x) = \begin{cases} [Q(x)]^{-1} P(x) & \colon Q(x) \neq 0 \\ 0 & \colon Q(x) = 0 \end{cases} \right] \bigg) \Bigg\} \tag{4.13}$$

(cf. Definition 4.2).

Lemma 4.4. *Let $n \in \mathbb{N}$, $R \in \mathscr{R}_{n,\infty}$. Then R is semi-algebraic.*

Proof of Lemma 4.4. Observe that the assumption that $R \in \mathscr{R}_{n,\infty}$ assures that there exist $P, Q \in \mathscr{P}_n$ which satisfy for all $x \in \mathbb{R}^n$ that

$$R(x) = \begin{cases} \frac{P(x)}{Q(x)} & \colon Q(x) \neq 0, \\ 0 & \colon Q(x) = 0. \end{cases} \tag{4.14}$$

Note that (4.14) ensures that

$$\begin{aligned} \mathrm{Graph}(R) &= \{(x,y) \in \mathbb{R}^n \times \mathbb{R} \colon (R(x) = y)\} \\ &= \Big\{(x,y) \in \mathbb{R}^n \times \mathbb{R} \colon \big[(R(x) = y), (Q(x) = 0)\big]\Big\} \\ &\quad \cup \Big\{(x,y) \in \mathbb{R}^n \times \mathbb{R} \colon \big[(R(x) = y), (Q(x) \neq 0)\big]\Big\} \end{aligned}$$

[10] Observe that $\deg(\mathbb{R}^2 \ni (x_1, x_2) \mapsto x_1 x_2 \in \mathbb{R}) = 2$. Furthermore, note that for all $P \in \mathscr{P}_0$, $x, y \in \mathbb{R}^0 = \{0\}$ it holds that $P(x) = P(y) = P(0)$ and $\deg(P) = 0$.

$$= \Big\{(x,y) \in \mathbb{R}^n \times \mathbb{R}\colon \big(y = Q(x) = 0\big)\Big\}$$
$$\cup \Big\{(x,y) \in \mathbb{R}^n \times \mathbb{R}\colon \big[\big(P(x) = yQ(x)\big), \big(Q(x) \neq 0\big)\big]\Big\}. \tag{4.15}$$

Hence, we obtain

$$\text{Graph}(R) = \Big[\Big\{(x,y) \in \mathbb{R}^n \times \mathbb{R}\colon \big(y = 0\big)\Big\} \cap \Big\{(x,y) \in \mathbb{R}^n \times \mathbb{R}\colon \big(Q(x) = 0\big)\Big\}\Big]$$
$$\cup \Big[\Big\{(x,y) \in \mathbb{R}^n \times \mathbb{R}\colon \big(P(x) - yQ(x) = 0 < [Q(x)]^2\big)\Big\}\Big]. \tag{4.16}$$

Combining this with (4.10) establishes that R is semi-algebraic. The proof of Lemma 4.4 is thus complete. □

4.6 Suitable piecewise rational functions as semi-algebraic functions

Proposition 4.2. *Let $m \in \mathbb{N}$, $f \in \mathscr{F}_{m,0,\infty}$ (cf. Definition 4.1). Then $\mathbb{R}^m \ni \theta \mapsto f(\theta, 0) \in \mathbb{R}$ is semi-algebraic (cf. Definition 4.4).*

Proof of Proposition 4.2. Observe that (4.1) and the assumption that $f \in \mathscr{F}_{m,0,\infty}$ assure that there exist $V \in \mathbb{N}$, $r_1, r_2, \ldots, r_V \in \mathbb{N}$, $R_1, R_2, \ldots, R_V \in \mathscr{R}_{m,\infty}$, $P^1 = (P^1_i)_{i\in\{1,2,\ldots,r_1\}} \subseteq \mathscr{P}_m$, $P^2 = (P^2_i)_{i\in\{1,2,\ldots,r_2\}} \subseteq \mathscr{P}_m$, $\ldots$, $P^V = (P^V_i)_{i\in\{1,2,\ldots,r_V\}} \subseteq \mathscr{P}_m$ which satisfy for all $\theta \in \mathbb{R}^m$ that

$$f(\theta, 0) = \sum_{v=1}^{V} \left[R_v(\theta) \left[\prod_{i=1}^{r_v} \mathbb{1}_{[0,\infty)}\big(P^v_i(\theta)\big) \right] \right]. \tag{4.17}$$

Note that (4.17) shows for all $\theta \in \mathbb{R}^m$ that

$$f(\theta, 0) = \sum_{v=1}^{V} \left[R_v(\theta) \left[\prod_{i=1}^{r_v} \mathbb{1}_{\{\vartheta\in\mathbb{R}^m\colon P^v_i(\vartheta)\geq 0\}}(\theta) \right] \right]. \tag{4.18}$$

Furthermore, observe that Proposition 4.1 and Lemma 4.3 prove that for all $v \in \{1, 2, \ldots, V\}$, $i \in \{1, 2, \ldots, r_v\}$ it holds that

$$\mathbb{R}^m \ni \theta \mapsto \mathbb{1}_{\{\vartheta\in\mathbb{R}^m\colon P^v_i(\vartheta)\geq 0\}}(\theta) \in \mathbb{R} \tag{4.19}$$

is semi-algebraic. Moreover, note that Lemma 4.4 assures that for all $v \in \{1, 2, \ldots, V\}$ it holds that R_v is semi-algebraic. Combining this and (4.19) with Lemma 4.2 shows that for all $v \in \{1, 2, \ldots, V\}$ it holds that

$$\mathbb{R}^m \ni \theta \mapsto R_v(\theta)\big[\textstyle\prod_{i=1}^{r_v} \mathbb{1}_{\{\vartheta\in\mathbb{R}^m\colon P^v_i(\vartheta)\geq 0\}}(\theta)\big] \in \mathbb{R} \tag{4.20}$$

is semi-algebraic. Lemma 4.2 and (4.18) therefore show that $\mathbb{R}^m \ni \theta \mapsto f(\theta, 0) \in \mathbb{R}$ is semi-algebraic. The proof of Proposition 4.2 is thus complete. □

4.7 Closedness under parametric integration of suitable piecewise rational functions

Proposition 4.3. *Let $m, n \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in (a, \infty)$, $f \in \mathscr{F}_{m,n,\infty}$ (cf. Definition 4.1). Then*

(i) it holds for all $\theta \in \mathbb{R}^m$, $x_1, x_2, \dots, x_{n-1} \in \mathbb{R}$ that $\int_a^b |f(\theta, x_1, x_2, \dots, x_{n-1}, x_n)| \, dx_n < \infty$ and

(ii) it holds that

$$\big(\mathbb{R}^m \times \mathbb{R}^{n-1} \ni (\theta, x_1, \dots, x_{n-1}) \mapsto \textstyle\int_a^b f(\theta, x_1, x_2, \dots, x_{n-1}, x_n) \, dx_n \in \mathbb{R}\big) \in \mathscr{F}_{m,n-1,\infty}. \tag{4.21}$$

Proof of Proposition 4.3. Observe that (4.1) and the fact that $\{q \in \mathscr{P}_n \colon \deg(q) \leq \infty\} = \mathscr{P}_n \subseteq C(\mathbb{R}^n, \mathbb{R})$ prove that for all $\theta \in \mathbb{R}^m$, $r \in (0, \infty)$ it holds that

$$\sup\nolimits_{x \in [-r,r]^n} |f(\theta, x)| < \infty. \tag{4.22}$$

This shows item (i). Furthermore, note that [28, Proposition 4.8] and item (vi) in Lemma 4.1 establish item (ii). The proof of Proposition 4.3 is thus complete. □

4.8 Closedness under rectification of suitable piecewise rational functions

The next result, Lemma 4.5, establishes that the function classes $\mathscr{F}_{m,n,1}$ introduced in Definition 4.1 above are closed under composition with the ReLU function. This will be used to show in Proposition 4.4 below that these function classes contain the realization functions of DNNs with ReLU activation.

Lemma 4.5. *Let $m, n \in \mathbb{N}$, $f \in \mathscr{F}_{m,n,1}$ (cf. Definition 4.1). Then*

$$\big(\mathbb{R}^m \times \mathbb{R}^n \ni v \mapsto \max\{f(v), 0\} \in \mathbb{R}\big) \in \mathscr{F}_{m,n,1}. \tag{4.23}$$

Proof of Lemma 4.5. Observe that (4.1) and the assumption that $f \in \mathscr{F}_{m,n,1}$ ensure that there exist $V \in \mathbb{N}$, $r_1, r_2, \dots, r_V \in \mathbb{N}$, $R_1, R_2, \dots, R_V \in \mathscr{R}_{m,1}$, $Q_1, Q_2, \dots, Q_V \in \{q \in \mathscr{P}_n \colon \deg(q) \leq 1\}$, $P^1 = (P^1_{i,j})_{(i,j) \in \{1,2,\dots,r_1\} \times \{0,1,\dots,n\}} \subseteq \mathscr{P}_m$, $P^2 = (P^2_{i,j})_{(i,j) \in \{1,2,\dots,r_2\} \times \{0,1,\dots,n\}} \subseteq \mathscr{P}_m$, $\dots$, $P^V = (P^V_{i,j})_{(i,j) \in \{1,2,\dots,r_V\} \times \{0,1,\dots,n\}} \subseteq \mathscr{P}_m$ which satisfy for all $\theta \in \mathbb{R}^m$, $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ that

$$f(\theta, x) = \sum_{v=1}^{V} \left[R_v(\theta) Q_v(x) \left[\prod_{i=1}^{r_v} \mathbb{1}_{[0,\infty)} \big(P^v_{i,0}(\theta) + \textstyle\sum_{j=1}^n P^v_{i,j}(\theta) x_j \big) \right] \right]. \tag{4.24}$$

In the following let $p_v \colon \mathbb{R}^m \times \mathbb{R}^n \to \{0, 1\}$, $v \in \{1, 2, \dots, V\}$, satisfy for all $v \in \{1, 2, \dots, V\}$, $\theta \in \mathbb{R}^m$, $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ that

$$p_v(\theta, x) = \prod_{i=1}^{r_v} \mathbb{1}_{[0,\infty)} \big(P^v_{i,0}(\theta) + \textstyle\sum_{j=1}^n P^v_{i,j}(\theta) x_j \big), \tag{4.25}$$

for every $W \subseteq \{1, 2, \dots, V\}$ let $\mathfrak{p}_W \in \{0, 1\}$ satisfy

$$\mathfrak{p}_W = \left[\prod_{v \in W} p_v(\theta, x)\right]\left[\prod_{v \in \{1,2,\dots,V\} \setminus W} (1 - p_v(\theta, x))\right], \tag{4.26}$$

and for every $\theta \in \mathbb{R}^m$, $x \in \mathbb{R}^n$ let $\mathscr{V}_{\theta,x} \subseteq \mathbb{N}$ satisfy

$$\mathscr{V}_{\theta,x} = \{v \in \{1, 2, \dots, V\} \colon p_v(\theta, x) = 1\}. \tag{4.27}$$

Note that (4.25), (4.26), and (4.27) assure that for all $W \subseteq \{1, 2, \dots, V\}$ it holds that

$$\mathfrak{p}_W = \begin{cases} 1 & : W = \mathscr{V}_{\theta,x}, \\ 0 & : W \neq \mathscr{V}_{\theta,x}. \end{cases} \tag{4.28}$$

Combining this with (4.24), (4.25), and (4.27) proves that for all $\theta \in \mathbb{R}^m$, $x \in \mathbb{R}^n$ it holds that

$$\begin{aligned} f(\theta, x) &= \sum_{v=1}^{V} R_v(\theta) Q_v(x) p_v(\theta, x) = \sum_{v \in \mathscr{V}_{\theta,x}} R_v(\theta) Q_v(x) p_v(\theta, x) \\ &= \sum_{v \in \mathscr{V}_{\theta,x}} R_v(\theta) Q_v(x) = \sum_{W \subseteq \{1,2,\dots,V\}} \left(\mathfrak{p}_W \left[\sum_{v \in W} R_v(\theta) Q_v(x)\right]\right). \end{aligned} \tag{4.29}$$

This and (4.28) show that for all $\theta \in \mathbb{R}^m$, $x \in \mathbb{R}^n$ it holds that

$$\max\{f(\theta, x), 0\} = \sum_{W \subseteq \{1,2,\dots,V\}} \left(\mathfrak{p}_W \max\left\{\sum_{v \in W} R_v(\theta) Q_v(x), 0\right\}\right). \tag{4.30}$$

The fact that for all $r \in \mathbb{R}$ it holds that $\max\{r, 0\} = r \mathbb{1}_{[0,\infty)}(r)$ hence demonstrates that for all $\theta \in \mathbb{R}^m$, $x \in \mathbb{R}^n$ it holds that

$$\begin{aligned} \max\{f(\theta, x), 0\} &= \sum_{W \subseteq \{1,2,\dots,V\}} \left(\mathfrak{p}_W \left[\sum_{v \in W} R_v(\theta) Q_v(x)\right] \mathbb{1}_{[0,\infty)}\left(\sum_{v \in W} R_v(\theta) Q_v(x)\right)\right) \\ &= \sum_{W \subseteq \{1,2,\dots,V\}} \sum_{w \in W} \left(R_w(\theta) Q_w(x) \left[\mathfrak{p}_W \mathbb{1}_{[0,\infty)}\left(\sum_{v \in W} R_v(\theta) Q_v(x)\right)\right]\right). \end{aligned} \tag{4.31}$$

Furthermore, observe that (4.26) shows that

$$\begin{aligned} \mathfrak{p}_W &= \left[\prod_{v \in W} p_v(\theta, x)\right]\left[\sum_{U \subseteq (\{1,2,\dots,V\} \setminus W)} (-1)^{\#(U)} \left[\textstyle\prod_{v \in U} p_v(\theta, x)\right]\right] \\ &= \sum_{U \subseteq (\{1,2,\dots,V\} \setminus W)} \left((-1)^{\#(U)} \left[\prod_{v \in W} p_v(\theta, x)\right]\left[\prod_{v \in U} p_v(\theta, x)\right]\right). \end{aligned} \tag{4.32}$$

Combining this and (4.31) proves that for all $\theta \in \mathbb{R}^m$, $x \in \mathbb{R}^n$ it holds that

$$\max\{f(\theta,x),0\} = \sum_{W\subseteq\{1,2,\dots,V\}} \sum_{w\in W} \sum_{U\subseteq(\{1,2,\dots,V\}\backslash W)} \biggl((-1)^{\#(U)} R_w(\theta) Q_w(x) \cdot \left[\mathbb{1}_{[0,\infty)}\Bigl(\textstyle\sum_{v\in W} R_v(\theta)Q_v(x) \Bigr) \right] \left[\textstyle\prod_{v\in W} p_v(\theta,x) \right] \left[\textstyle\prod_{v\in U} p_v(\theta,x) \right] \biggr). \tag{4.33}$$

Moreover, note that item (ii) in Lemma 4.1, (4.1), (4.25), and the fact that $R_1, R_2, \dots, R_V \in \mathscr{R}_{m,1}$ show that for all $W \subseteq \{1,2,\dots,V\}$, $U \subseteq (\{1,2,\dots,V\}\backslash W)$ and all $w \in W$ it holds that

$$\biggl(\mathbb{R}^m \times \mathbb{R}^n \ni (\theta,x) \mapsto R_w(\theta)Q_w(x) \left[\mathbb{1}_{[0,\infty)}\Bigl(\textstyle\sum_{v\in W} R_v(\theta)Q_v(x) \Bigr) \right] \cdot \left[\textstyle\prod_{v\in W} p_v(\theta,x) \right] \left[\textstyle\prod_{v\in U} p_v(\theta,x) \right] \in \mathbb{R} \biggr) \in \mathscr{F}_{m,n,1}. \tag{4.34}$$

Combining this and (4.33) with (4.1) establishes (4.23). The proof of Lemma 4.5 is thus complete. □

4.9 Realization functions of DNNs as suitable piecewise rational functions

Proposition 4.4. *Assume Setting 3.1. Then it holds for all $i \in \{1,2,\dots,\ell_L\}$ that*

$$\bigl(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\ell_0} \ni (\theta,x) \mapsto \mathcal{N}^{L,\theta}_{\infty,i}(x) \in \mathbb{R} \bigr) \in \mathscr{F}_{\mathfrak{d},\ell_0,1} \tag{4.35}$$

(cf. Definition 4.1).

Proof of Proposition 4.4. Observe that (3.5) ensures for all $k \in \mathbb{N}_0$, $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1,2,\dots,\ell_{k+1}\}$, $x = (x_1,\dots,x_{\ell_0}) \in \mathbb{R}^{\ell_0}$ that

$$\mathcal{N}^{k+1,\theta}_{\infty,i}(x) = \begin{cases} \mathfrak{b}^{k+1,\theta}_i + \sum_{j=1}^{\ell_k} \mathfrak{w}^{k+1,\theta}_{i,j} x_j. & : k = 0, \\ \mathfrak{b}^{k+1,\theta}_i + \sum_{j=1}^{\ell_k} \mathfrak{w}^{k+1,\theta}_{i,j} \max\{\mathcal{N}^{k,\theta}_{\infty,j}(x), 0\}. & : k > 0. \end{cases} \tag{4.36}$$

Next we claim that for all $k \in \{1,2,\dots,L\}$ it holds that

$$\bigl(\textstyle\bigcup_{i=1}^{\ell_k} \{ \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\ell_0} \ni (\theta,x) \mapsto \mathcal{N}^{k,\theta}_{\infty,i}(x) \in \mathbb{R} \} \bigr) \subseteq \mathscr{F}_{\mathfrak{d},\ell_0,1}. \tag{4.37}$$

In the following we prove (4.37) by induction on $k \in \{1,2,\dots,L\}$. For the base case $k = 1$ note that (4.36) assures that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1,2,\dots,\ell_1\}$, $x = (x_1,\dots,x_{\ell_0}) \in \mathbb{R}^{\ell_0}$ it holds that

$$\mathcal{N}^{1,\theta}_{\infty,i}(x) = \mathfrak{b}^{1,\theta}_i + \textstyle\sum_{j=1}^{\ell_0} \mathfrak{w}^{1,\theta}_{i,j} x_j. \tag{4.38}$$

This establishes (4.37) in the case $k = 1$. For the induction step observe that Lemma 4.5 implies that for all $k \in \mathbb{N} \cap (0, L)$, $j \in \{1, 2, \ldots, \ell_k\}$ with $(\bigcup_{i=1}^{\ell_k} \{\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\ell_0} \ni (\theta, x) \mapsto \mathcal{N}_{\infty,i}^{k,\theta}(x) \in \mathbb{R}\}) \subseteq \mathscr{F}_{\mathfrak{d},\ell_0,1}$ it holds that

$$\big(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\ell_0} \ni (\theta, x) \mapsto \max\{\mathcal{N}_{\infty,j}^{k,\theta}(x), 0\} \in \mathbb{R}\big) \in \mathscr{F}_{\mathfrak{d},\ell_0,1}. \tag{4.39}$$

Furthermore, note that (4.36) shows that for all $k \in \mathbb{N} \cap (0, L)$, $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \ldots, \ell_{k+1}\}$, $x \in \mathbb{R}^{\ell_0}$ we have that

$$\mathcal{N}_{\infty,i}^{k+1,\theta}(x) = \mathfrak{b}_i^{k+1,\theta} + \textstyle\sum_{j=1}^{\ell_k} \mathfrak{w}_{i,j}^{k+1,\theta} \max\{\mathcal{N}_{\infty,j}^{k,\theta}(x), 0\}. \tag{4.40}$$

Moreover, observe that (4.1) and (4.39) demonstrate for all $k \in \mathbb{N} \cap (0, L)$, $i \in \{1, 2, \ldots, \ell_{k+1}\}$, $j \in \{1, 2, \ldots, \ell_k\}$ with $(\bigcup_{v=1}^{\ell_k} \{\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\ell_0} \ni (\theta, x) \mapsto \mathcal{N}_{\infty,v}^{k,\theta}(x) \in \mathbb{R}\}) \subseteq \mathscr{F}_{\mathfrak{d},\ell_0,1}$ that

$$\big(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\ell_0} \ni (\theta, x) \mapsto \mathfrak{w}_{i,j}^{k+1,\theta} \max\{\mathcal{N}_{\infty,j}^{k,\theta}(x), 0\} \in \mathbb{R}\big) \in \mathscr{F}_{\mathfrak{d},\ell_0,1}. \tag{4.41}$$

The fact that $\mathscr{F}_{\mathfrak{d},\ell_0,1}$ is an $\mathbb{R}$-vector space and (4.1) hence show that for all $k \in \mathbb{N} \cap (0, L)$, $i \in \{1, 2, \ldots, \ell_{k+1}\}$ with $(\bigcup_{v=1}^{\ell_k} \{\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\ell_0} \ni (\theta, x) \mapsto \mathcal{N}_{\infty,v}^{k,\theta}(x) \in \mathbb{R}\}) \subseteq \mathscr{F}_{\mathfrak{d},\ell_0,1}$ it holds that

$$\big(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\ell_0} \ni (\theta, x) \mapsto \mathfrak{b}_i^{k+1,\theta} + \textstyle\sum_{j=1}^{\ell_k} \mathfrak{w}_{i,j}^{k+1,\theta} \max\{\mathcal{N}_{\infty,j}^{k,\theta}(x), 0\} \in \mathbb{R}\big) \in \mathscr{F}_{\mathfrak{d},\ell_0,1}. \tag{4.42}$$

This and (4.40) assure that for all $k \in \mathbb{N} \cap (0, L)$ with $(\bigcup_{i=1}^{\ell_k} \{\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\ell_0} \ni (\theta, x) \mapsto \mathcal{N}_{\infty,i}^{k,\theta}(x) \in \mathbb{R}\}) \subseteq \mathscr{F}_{\mathfrak{d},\ell_0,1}$ it holds that

$$\big(\textstyle\bigcup_{i=1}^{\ell_{k+1}} \{\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\ell_0} \ni (\theta, x) \mapsto \mathcal{N}_{\infty,i}^{k+1,\theta}(x) \in \mathbb{R}\}\big) \subseteq \mathscr{F}_{\mathfrak{d},\ell_0,1}. \tag{4.43}$$

Induction thus proves (4.37). Note that (4.37) establishes (4.35). The proof of Proposition 4.4 is thus complete. □

5 Piecewise polynomial functions

In this section we establish in Corollary 5.1 in Subsection 5.7 below that in the set-up of Setting 3.1 in Subsection 3.1 above we have, under the assumption that the measure μ is absolutely continuous with density $\mathfrak{p} \colon [a, b]^{\ell_0} \to \mathbb{R}$ and the assumption that the density function $\mathfrak{p} \colon [a, b]^{\ell_0} \to \mathbb{R}$ and every component of the target function $f = (f_1, \ldots, f_{\ell_L}) \colon [a, b]^{\ell_0} \to \mathbb{R}^{\ell_L}$ are piecewise polynomial in the sense of Definition 5.1 in Subsection 5.1 below, that the risk function $\mathcal{L}_\infty \colon [a, b]^{\ell_0} \to \mathbb{R}$ is semi-algebraic. In Section 6 below we will employ Corollary 5.1 to conclude that for every ANN parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}}$ we have that the risk function $\mathcal{L}_\infty \colon [a, b]^{\ell_0} \to \mathbb{R}$ satisfies a generalized Kurdyka-Łojasiewicz inequality on a neighbourhood of θ.

Throughout this work we consider fully connected feedforward ANNs. For different network architectures such as convolutional neural networks (CNNs) or residual neural networks it might be possible to establish analogous results by suitably adapting our arguments.

Our proof of Corollary 5.1 mainly relies on Proposition 4.3 in Subsection 4.7 above, on Proposition 4.4 in Subsection 4.9 above, as well as on the fact that for all $m \in \mathbb{N}$ it holds that functions in $\mathscr{F}_{m,0,\infty}$, are semi-algebraic according to Proposition 4.2 in Subsection 4.6 above.

Some of the concepts and results in this section are inspired by our previous article Eberle et al. [28, Section 4]. In particular, Definition 5.1 is a slight extension of [28, Definition 4.9] and Corollary 5.1 extends [28, Corollary 4.10] from the situation of shallow ReLU ANNs with one hidden layer to deep ReLU ANNs with an arbitrarily large number of hidden layers.

It should also be noted that Corollary 5.1 is a novel contribution mainly due to the fact that we consider the true risk defined by the integral over the entire input data. If one considers the empirical risk (calculated from a finite set of input data) an analogous result is already known, cf. Davis et al. [20, Corollary 5.11].

5.1 Piecewise polynomial functions

Definition 5.1 (Piecewise polynomial functions)**.** *Let $d \in \mathbb{N}$, let $A \subseteq \mathbb{R}^d$ and $B \subseteq \mathbb{R}$ be sets, and let $f\colon A \to B$ be a function. Then we say that f is piecewise polynomial if and only if there exist $n \in \mathbb{N}$, $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R}^{n\times d}$, $\beta_1, \beta_2, \ldots, \beta_n \in \mathbb{R}^n$, $P_1, P_2, \ldots, P_n \in \mathscr{P}_d$ such that for all $x \in A$ it holds that*

$$f(x) = \textstyle\sum_{i=1}^n P_i(x)\mathbb{1}_{[0,\infty)^n}(\alpha_i x + \beta_i) \tag{5.1}$$

(cf. Definition 4.2).

5.2 Characterization results for piecewise polynomial functions

The following results, Lemma 5.1 to Proposition 5.1, are elementary consequences of the definition of piecewise polynomial functions. They will be employed in the proof of Corollary 5.1 to show that the risk function is semi-algebraic if the density function and every component of the target function are piecewise polynomial.

Lemma 5.1. *Let $d \in \mathbb{N}$, let $A \subseteq \mathbb{R}^d$ be a set, and let $f\colon A \to \mathbb{R}$ be a function. Then the following three statements are equivalent:*

(i) It holds that f is piecewise polynomial (cf. Definition 5.1).

(ii) There exist $m, n \in \mathbb{N}$, $\alpha_1, \alpha_2, \ldots, \alpha_m \in \mathbb{R}^{n\times d}$, $\beta_1, \beta_2, \ldots, \beta_m \in \mathbb{R}^n$, $P_1, P_2, \ldots, P_m \in \mathscr{P}_d$ such that for all $x \in A$ it holds that

$$f(x) = \textstyle\sum_{i=1}^m P_i(x)\mathbb{1}_{[0,\infty)^n}(\alpha_i x + \beta_i). \tag{5.2}$$

(iii) There exist $n \in \mathbb{N}$, $m_1, m_2, \ldots, m_n \in \mathbb{N}$, $\alpha_1 \in \mathbb{R}^{m_1\times d}$, $\alpha_2 \in \mathbb{R}^{m_2\times d}$, $\ldots$, $\alpha_n \in \mathbb{R}^{m_n\times d}$, $\beta_1 \in \mathbb{R}^{m_1}$, $\beta_2 \in \mathbb{R}^{m_2}$, $\ldots$, $\beta_m \in \mathbb{R}^{m_n}$, $P_1, P_2, \ldots, P_n \in \mathscr{P}_d$ such that for all $x \in A$ it holds that

$$f(x) = \textstyle\sum_{i=1}^n P_i(x)\mathbb{1}_{[0,\infty)^{m_i}}(\alpha_i x + \beta_i). \tag{5.3}$$

Proof of Lemma 5.1. Throughout this proof for every $m \in \mathbb{N}$ let $\mathbf{e}_1^{(m)}, \mathbf{e}_2^{(m)}, \dots, \mathbf{e}_m^{(m)} \in \mathbb{R}^n$ satisfy $\mathbf{e}_1^{(m)} = (1, 0, \dots, 0)$, $\mathbf{e}_2^{(m)} = (0, 1, 0, \dots, 0)$, $\dots$, $\mathbf{e}_m^{(m)} = (0, \dots, 0, 1)$ and for every $m, n \in \mathbb{N}$ with $m \geq n$ let $A_{m,n} \in \mathbb{R}^{m \times n}$ satisfy for all $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ that

$$A_{m,n} x = \sum_{i=1}^{n} x_i \mathbf{e}_i^{(m)}. \tag{5.4}$$

Observe that (5.4) ensures that for all $n \in \mathbb{N}$, $m_1, m_2, \dots, m_n \in \mathbb{N}$, $\alpha_1 \in \mathbb{R}^{m_1 \times d}$, $\alpha_2 \in \mathbb{R}^{m_2 \times d}$, $\dots$, $\alpha_n \in \mathbb{R}^{m_n \times d}$, $\beta_1 \in \mathbb{R}^{m_1}$, $\beta_2 \in \mathbb{R}^{m_2}$, $\dots$, $\beta_m \in \mathbb{R}^{m_n}$, $P_1, P_2, \dots, P_n \in \mathscr{P}_d$, $x \in A$ it holds that

$$\begin{aligned}
&\sum_{i=1}^{n} P_i(x) \mathbb{1}_{[0,\infty)^{m_i}}(\alpha_i x + \beta_i) \\
&= \sum_{i=1}^{n} \big[P_{\min\{i,n\}}(x)\big] \big[\mathbb{1}_{[0,\infty)^{m_{\min\{i,n\}}}}\big(\alpha_{\min\{i,n\}} x + \beta_{\min\{i,n\}}\big)\big] \\
&= \sum_{i=1}^{m_1+\ldots+m_n} \big[P_{\min\{i,n\}}(x) \mathbb{1}_{[1,n]}(i)\big] \big[\mathbb{1}_{[0,\infty)^{m_{\min\{i,n\}}}}\big(\alpha_{\min\{i,n\}} x + \beta_{\min\{i,n\}}\big)\big] \\
&= \sum_{i=1}^{m_1+\ldots+m_n} \big[P_{\min\{i,n\}}(x) \mathbb{1}_{[1,n]}(i)\big] \\
&\quad \cdot \big[\mathbb{1}_{[0,\infty)^{m_1+\ldots+m_n}}\big(\big[A_{m_1+\ldots+m_n, m_{\min\{i,n\}}} \alpha_{\min\{i,n\}}\big] x + \big[A_{m_1+\ldots+m_n, m_{\min\{i,n\}}} \beta_{\min\{i,n\}}\big]\big)\big].
\end{aligned} \tag{5.5}$$

Combining this with (5.1) establishes that for all $n \in \mathbb{N}$, $m_1, m_2, \dots, m_n \in \mathbb{N}$, $\alpha_1 \in \mathbb{R}^{m_1 \times d}$, $\alpha_2 \in \mathbb{R}^{m_2 \times d}$, $\dots$, $\alpha_n \in \mathbb{R}^{m_n \times d}$, $\beta_1 \in \mathbb{R}^{m_1}$, $\beta_2 \in \mathbb{R}^{m_2}$, $\dots$, $\beta_m \in \mathbb{R}^{m_n}$, $P_1, P_2, \dots, P_n \in \mathscr{P}_d$ with $\forall\, x \in A \colon f(x) = \sum_{i=1}^{n} P_i(x) \mathbb{1}_{[0,\infty)^{m_i}}(\alpha_i x + \beta_i)$ it holds that f is piecewise polynomial. The proof of Lemma 5.1 is thus complete. ☐

5.3 Sums and products of piecewise polynomial functions

Lemma 5.2. *Let $d \in \mathbb{N}$, let $A \subseteq \mathbb{R}^d$ be a set, and let $f \colon A \to \mathbb{R}$ and $g \colon A \to \mathbb{R}$ be piecewise polynomial (cf. Definition 5.1). Then*

(i) it holds that $A \ni x \mapsto f(x) + g(x) \in \mathbb{R}$ is piecewise polynomial and

(ii) it holds that $A \ni x \mapsto f(x)g(x) \in \mathbb{R}$ is piecewise polynomial.

Proof of Lemma 5.2. Note that (5.1), the assumption that f is piecewise polynomial, and the assumption that g is piecewise polynomial ensure that there exist $n, m \in \mathbb{N}$, $\alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{R}^{n \times d}$, $\mathfrak{a}_1, \mathfrak{a}_2, \dots, \mathfrak{a}_m \in \mathbb{R}^{m \times d}$, $\beta_1, \beta_2, \dots, \beta_n \in \mathbb{R}^n$, $\mathfrak{b}_1, \mathfrak{b}_2, \dots, \mathfrak{b}_m \in \mathbb{R}^m$, $P_1, P_2, \dots, P_n, \mathfrak{P}_1, \mathfrak{P}_2, \dots, \mathfrak{P}_m \in \mathscr{P}_d$ which satisfy for all $x \in A$ that

$$f(x) = \textstyle\sum_{i=1}^{n} P_i(x) \mathbb{1}_{[0,\infty)^n}(\alpha_i x + \beta_i) \qquad \text{and} \qquad g(x) = \textstyle\sum_{i=1}^{m} \mathfrak{P}_i(x) \mathbb{1}_{[0,\infty)^m}(\mathfrak{a}_i x + \mathfrak{b}_i). \tag{5.6}$$

Observe that (5.6) assures for all $x \in A$ that

$$f(x) + g(x) = \big[\textstyle\sum_{i=1}^{n} P_i(x) \mathbb{1}_{[0,\infty)^n}(\alpha_i x + \beta_i)\big] + \big[\textstyle\sum_{i=1}^{m} \mathfrak{P}_i(x) \mathbb{1}_{[0,\infty)^m}(\mathfrak{a}_i x + \mathfrak{b}_i)\big] \tag{5.7}$$

and

$$\begin{aligned} f(x)g(x) &= \left[\textstyle\sum_{i=1}^{n} P_i(x)\mathbb{1}_{[0,\infty)^n}(\alpha_i x+\beta_i)\right]\left[\textstyle\sum_{i=1}^{m} \mathfrak{P}_i(x)\mathbb{1}_{[0,\infty)^m}(\mathfrak{a}_i x+\mathfrak{b}_i)\right] \\ &= \textstyle\sum_{(i,j)\in\{1,2,\ldots,n\}\times\{1,2,\ldots,m\}} \left[P_i(x)\mathfrak{P}_j(x)\right]\left[\mathbb{1}_{[0,\infty)^n}(\alpha_i x+\beta_i)\mathbb{1}_{[0,\infty)^m}(\mathfrak{a}_j x+\mathfrak{b}_j)\right]. \end{aligned} \tag{5.8}$$

Combining this with Lemma 5.1 establishes items (i) and (ii). The proof of Lemma 5.2 is thus complete. ☐

5.4 Indicator functions as piecewise polynomial functions

Lemma 5.3. *Let $d \in \mathbb{N}$, $a_1, a_2, \ldots, a_d \in \mathbb{R}$, $b_1 \in [a_1, \infty)$, $b_2 \in [a_2, \infty)$, $\ldots$, $b_d \in [a_d, \infty)$ and let $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfy for all $x \in \mathbb{R}^d$ that $f(x) = \mathbb{1}_{[a_1,b_1]\times\ldots\times[a_d,b_d]}(x)$. Then f is piecewise polynomial (cf. Definition 5.1).*

Proof of Lemma 5.3. Throughout this proof let $\alpha_1, \alpha_2, \ldots, \alpha_{2d} \in \mathbb{R}^{(2d)\times d}$ satisfy for all $i \in \{1, 2, \ldots, 2d\}$, $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ that

$$\alpha_i x = (x_1, x_2, \ldots, x_d, -x_1, -x_2, \ldots, -x_d), \tag{5.9}$$

let $\beta_1, \beta_2, \ldots, \beta_{2d} \in \mathbb{R}^{2d}$ satisfy for all $i \in \{1, 2, \ldots, 2d\}$ that

$$\beta_i = (-\alpha_1, -\alpha_2, \ldots, -\alpha_d, \beta_1, \beta_2, \ldots, \beta_d), \tag{5.10}$$

and let $P_i\colon \mathbb{R}^d \to \mathbb{R}$, $i \in \{1, 2, \ldots, 2d\}$, satisfy for all $i \in \mathbb{N} \cap (1, 2d]$, $x \in \mathbb{R}^d$ that $P_1(x) = 1$ and $P_i(x) = 0$. Note that (5.9) and (5.10) ensure that

$$\begin{aligned} &\times_{i=1}^{d} [a_i, b_i] \\ &= \Big\{x = (x_1, \ldots, x_d) \in \mathbb{R}^d \colon \Big(\forall\, i \in \{1, 2, \ldots, d\} \colon a_i \leq x_i \leq b_i\Big)\Big\} \\ &= \Big\{x = (x_1, \ldots, x_d) \in \mathbb{R}^d \colon \Big(\forall\, i \in \{1, 2, \ldots, d\} \colon \Big[\big(x_i - a_i \in [0, \infty)\big), \big(-x_i + b_i \in [0, \infty)\big)\Big]\Big)\Big\} \\ &= \big\{x \in \mathbb{R}^d \colon \big(\alpha_1 x + \beta_1 \in [0, \infty)^{2d}\big)\big\}. \end{aligned} \tag{5.11}$$

Therefore, we obtain for all $x \in \mathbb{R}^d$ that

$$\begin{aligned} f(x) &= \mathbb{1}_{\{y\in\mathbb{R}^d \colon \alpha_1 y+\beta_1\in[0,\infty)^{2d}\}}(x) = \mathbb{1}_{[0,\infty)^{2d}}(\alpha_1 x + \beta_1) = P_1(x)\mathbb{1}_{[0,\infty)^{2d}}(\alpha_1 x + \beta_1) \\ &= \sum_{i=1}^{2d} P_i(x)\mathbb{1}_{[0,\infty)^{2d}}(\alpha_i x + \beta_i). \end{aligned} \tag{5.12}$$

Combining this with the fact that $P_1, P_2, \ldots, P_{2d} \in \mathscr{P}_d$ establishes that f is piecewise polynomial. The proof of Lemma 5.3 is thus complete. ☐

5.5 Extensions of piecewise polynomial functions

Lemma 5.4. *Let $d \in \mathbb{N}$, let $A \subseteq \mathbb{R}^d$ be a set, and let $f\colon A \to \mathbb{R}$ be piecewise polynomial (cf. Definition 5.1). Then there exists a piecewise polynomial $F\colon \mathbb{R}^d \to \mathbb{R}$ such that $F|_A = f$.*

Proof of Lemma 5.3. Observe that (5.1) and the assumption that f is piecewise polynomial ensure that there exist $n \in \mathbb{N}$, $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R}^{n\times d}$, $\beta_1, \beta_2, \ldots, \beta_n \in \mathbb{R}^n$, $P_1, P_2, \ldots, P_n \in \mathscr{P}_d$ which satisfy for all $x \in A$ that

$$f(x) = \textstyle\sum_{i=1}^n P_i(x) \mathbb{1}_{[0,\infty)^n}(\alpha_i x + \beta_i). \tag{5.13}$$

In the following let $F\colon \mathbb{R}^d \to \mathbb{R}$ satisfy for all $x \in \mathbb{R}^d$ that

$$F(x) = \textstyle\sum_{i=1}^n P_i(x) \mathbb{1}_{[0,\infty)^n}(\alpha_i x + \beta_i). \tag{5.14}$$

Note that (5.1), (5.13), and (5.14) assure that F is piecewise polynomial. Furthermore, observe that (5.13) and (5.14) establish that $F|_A = f$. The proof of Lemma 5.3 is thus complete. ☐

Proposition 5.1. *Let $d \in \mathbb{N}$, $a_1, a_2, \ldots, a_d \in \mathbb{R}$, $b_1 \in [a_1, \infty)$, $b_2 \in [a_2, \infty)$, $\ldots$, $b_d \in [a_d, \infty)$, $A = [a_1, b_1] \times \ldots \times [a_d, b_d]$, let $f\colon A \to \mathbb{R}$ be piecewise polynomial, and let $F\colon \mathbb{R}^d \to \mathbb{R}$ satisfy for all $x \in A$, $y \in \mathbb{R}^d \backslash A$ that $F(x) = f(x)$ and $F(y) = 0$ (cf. Definition 5.1). Then F is piecewise polynomial.*

Proof of Proposition 5.1. Throughout this proof let $\mathfrak{g}\colon \mathbb{R}^d \to \mathbb{R}$ satisfy for all $x \in \mathbb{R}^d$ that $\mathfrak{g}(x) = \mathbb{1}_A(x)$. Note that Lemma 5.3 ensures that $\mathfrak{g}$ is piecewise polynomial. Furthermore, observe that Lemma 5.4 assures that there exists a piecewise polynomial $\mathfrak{f}\colon \mathbb{R}^d \to \mathbb{R}$ which satisfies

$$\mathfrak{f}|_A = f. \tag{5.15}$$

Note that (5.15) shows for all $x \in \mathbb{R}^d$ that

$$\mathfrak{f}(x)\mathfrak{g}(x) = \mathfrak{f}(x)\mathbb{1}_A(x) = F(x). \tag{5.16}$$

Moreover, observe that Lemma 5.2, the fact that $\mathfrak{f}$ is piecewise polynomial, and the fact that $\mathfrak{g}$ is piecewise polynomial demonstrate that $\mathbb{R}^d \ni x \mapsto \mathfrak{f}(x)\mathfrak{g}(x) \in \mathbb{R}$ is piecewise polynomial. Combining this with (5.16) establishes that F is piecewise polynomial. The proof of Proposition 5.1 is thus complete. ☐

5.6 Piecewise polynomial functions as suitable piecewise rational functions

We next establish in Proposition 5.2 that every piecewise polynomial function in the sense of Definition 5.1 is contained in a class of suitable piecewise rational functions introduced in Definition 4.1.

Proposition 5.2. *Let $m, n \in \mathbb{N}$ and let $f\colon \mathbb{R}^n \to \mathbb{R}$ be piecewise polynomial (cf. Definition 5.1). Then*

$$\big(\mathbb{R}^m \times \mathbb{R}^n \ni (\theta, x) \mapsto f(x) \in \mathbb{R}\big) \in \mathscr{F}_{m,n,\infty} \tag{5.17}$$

(cf. Definition 4.1).

Proof of Proposition 5.2. Note that the fact that for all $N \in \mathbb{N}$, $v = (v_1, \ldots, v_N) \in \mathbb{R}^N$ it holds that $\mathbb{1}_{[0,\infty)^N}(v) = \prod_{i=1}^N \mathbb{1}_{[0,\infty)}(v_i)$ assures that for all $N \in \mathbb{N}$, $\alpha = (\alpha_{i,j})_{(i,j)\in\{1,\ldots,N\}\times\{1,\ldots,n\}} \in \mathbb{R}^{N\times n}$, $\beta = (\beta_1, \ldots, \beta_N) \in \mathbb{R}^N$, $P \in \mathscr{P}_n$, $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ it holds that

$$P(x)\mathbb{1}_{[0,\infty)^N}(\alpha x + \beta) = P(x)\big[\textstyle\prod_{i=1}^N \mathbb{1}_{[0,\infty)}(\beta_i + \sum_{j=1}^n \alpha_{i,j} x_j)\big]. \tag{5.18}$$

Combining this with (4.1) demonstrates that for all $N \in \mathbb{N}$, $\alpha \in \mathbb{R}^{N\times n}$, $\beta \in \mathbb{R}^N$, $P \in \mathscr{P}_n$ it holds that

$$\big(\mathbb{R}^m \times \mathbb{R}^n \ni (\theta, x) \mapsto P(x)\mathbb{1}_{[0,\infty)^N}(\alpha x + \beta) \in \mathbb{R}\big) \in \mathscr{F}_{m,n,\infty}. \tag{5.19}$$

This and (4.1) assure that for all $N \in \mathbb{N}$, $\alpha_1, \alpha_2, \ldots, \alpha_N \in \mathbb{R}^{N\times n}$, $\beta_1, \beta_2, \ldots, \beta_N \in \mathbb{R}^N$, $P_1, P_2, \ldots, P_N \in \mathscr{P}_n$ it holds that

$$\big(\mathbb{R}^m \times \mathbb{R}^n \ni (\theta, x) \mapsto \textstyle\sum_{i=1}^N [P_i(x)\mathbb{1}_{[0,\infty)^N}(\alpha_i x + \beta_i)] \in \mathbb{R}\big) \in \mathscr{F}_{m,n,\infty}. \tag{5.20}$$

Combining this and (5.1) establishes (5.17). The proof of Proposition 5.2 is thus complete. □

5.7 Semi-algebraic risk functions in the training of deep ANNs

Finally, we combine the previous results to establish the main result of this section.

Corollary 5.1. *Assume Setting 3.1, assume for all $i \in \{1, 2, \ldots, \ell_L\}$ that f_i is piecewise polynomial, let $\mathfrak{p}\colon [a,b]^{\ell_0} \to \mathbb{R}$ be piecewise polynomial, and assume for all $E \in \mathcal{B}([a,b]^{\ell_0})$ that $\mu(E) = \int_E \mathfrak{p}(x)\,\mathrm{d}x$ (cf. Definition 5.1). Then $\mathcal{L}_\infty$ is semi-algebraic (cf. Definition 4.4).*

Proof of Corollary 5.1. Throughout this proof let $F = (F_1, \ldots, F_{\ell_L})\colon \mathbb{R}^{\ell_0} \to \mathbb{R}^{\ell_L}$ and $\mathfrak{P}\colon \mathbb{R}^{\ell_0} \to \mathbb{R}$ satisfy for all $i \in \{1, 2, \ldots, \ell_L\}$, $x \in \mathbb{R}^{\ell_0}$ that

$$F_i(x) = \begin{cases} f_i(x) & : x \in [a,b]^{\ell_0}, \\ 0 & : x \notin [a,b]^{\ell_0}, \end{cases} \qquad \text{and} \qquad \mathfrak{P}(x) = \begin{cases} \mathfrak{p}(x) & : x \in [a,b]^{\ell_0}, \\ 0 & : x \notin [a,b]^{\ell_0}. \end{cases} \tag{5.21}$$

Observe that Proposition 5.1 and (5.21) assure for all $i \in \{1, 2, \ldots, \ell_L\}$ that F_i and $\mathfrak{P}$ are piecewise polynomial. Proposition 5.2 hence ensures for all $i \in \{1, 2, \ldots, \ell_L\}$ that

$$\big\{\big(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\ell_0} \ni (\theta, x) \mapsto F_i(x) \in \mathbb{R}\big), \big(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\ell_0} \ni (\theta, x) \mapsto \mathfrak{P}(x) \in \mathbb{R}\big)\big\} \subseteq \mathscr{F}_{\mathfrak{d},\ell_0,\infty} \tag{5.22}$$

(cf. Definition 4.1). Furthermore, note that Proposition 4.4 and item (iv) in Lemma 4.1 demonstrate for all $i \in \{1, 2, \ldots, \ell_L\}$ that $(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\ell_0} \ni (\theta, x) \mapsto \mathcal{N}^{L,\theta}_{\infty,i}(x) \in \mathbb{R}) \in \mathscr{F}_{\mathfrak{d},\ell_0,\infty}$. Combining this, (4.1), and (5.22) with item (v) in Lemma 4.1 establish that

$$\big(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\ell_0} \ni (\theta, y) \mapsto \textstyle\sum_{i=1}^{\ell_L} \big[(\mathcal{N}^{L,\theta}_{\infty,i}(y) - F_i(y))^2 \mathfrak{P}(y)\big] \in \mathbb{R}\big) \in \mathscr{F}_{\mathfrak{d},\ell_0,\infty}. \tag{5.23}$$

Proposition 4.3 and induction therefore prove that

$$\begin{split}\big(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^0 \ni (\theta, x) \mapsto \textstyle\int_a^b \int_a^b \cdots \int_a^b \sum_{i=1}^{\ell_L} \big[(\mathcal{N}^{L,\theta}_{\infty,i}(y_1, \ldots, y_{\ell_0}) \\ - F_i(y_1, \ldots, y_{\ell_0}))^2 \mathfrak{P}(y_1, \ldots, y_{\ell_0})\big]\,\mathrm{d}y_{\ell_0} \cdots \mathrm{d}y_2\,\mathrm{d}y_1 \in \mathbb{R}\big) \in \mathscr{F}_{\mathfrak{d},0,\infty}.\end{split} \tag{5.24}$$

Moreover, observe that Fubini's theorem and (3.6) show for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\begin{aligned}
&\mathcal{L}_\infty(\theta) = \textstyle\int_{[a,b]^{\ell_0}} \|\mathcal{N}_\infty^{L,\theta}(y) - F(y)\|^2 \mathfrak{P}(y)\, \mathrm{d}y \\
&= \textstyle\int_a^b \int_a^b \cdots \int_a^b \|\mathcal{N}_\infty^{L,\theta}(y_1,\ldots,y_{\ell_0}) - F(y_1,\ldots,y_{\ell_0})\|^2 \mathfrak{P}(y_1,\ldots,y_{\ell_L})\, \mathrm{d}y_d \cdots \mathrm{d}y_2\, \mathrm{d}y_1 \\
&= \textstyle\int_a^b \int_a^b \cdots \int_a^b \sum_{i=1}^{\ell_L} \big[(\mathcal{N}_{\infty,i}^{L,\theta}(y_1,\ldots,y_{\ell_0}) - F_i(y_1,\ldots,y_{\ell_0}))^2 \mathfrak{P}(y_1,\ldots,y_{\ell_0})\big]\, \mathrm{d}y_{\ell_0} \cdots \mathrm{d}y_2\, \mathrm{d}y_1.
\end{aligned} \tag{5.25}$$

This and (5.24) disclose that $(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^0 \ni (\theta, x) \mapsto \mathcal{L}_\infty(\theta) \in \mathbb{R}) \in \mathscr{F}_{\mathfrak{d},0,\infty}$. Proposition 4.2 hence yields that $\mathcal{L}_\infty$ is semi-algebraic. The proof of Corollary 5.1 is thus complete. ☐

6 Generalized Kurdyka-Łojasiewicz inequalities for the training of deep ANNs

The main result of this section is Proposition 6.2 below, which reveals that under the assumption that the distribution of the input data has a piecewise polynomial density and that the target function is piecewise polynomial an appropriately generalized Kurdyka-Łojasiewicz inequality for the risk function is satisfied. We prove Proposition 6.2 by combining Bolte et al. [10, Theorem 3.1]) with the fact that the considered risk function $\mathcal{L}_\infty$ is semi-algebraic (cf. Corollary 5.1 above). Since [10, Theorem 3.1] is formulated for subanalytic functions (cf. Definition 6.4 below), we state in Proposition 6.1 below the well-known fact that every semi-algebraic function is subanalytic. We also formulate in Lemma 6.1 below the fact that the nonsmooth slope defined in (6.3) below is lower semi-continuous, which is well-known in the literature (see [10]). Only for completeness we include in this article a detailed proof of Lemma 6.1. As a simple consequence of Lemma 6.1 we show in Corollary 6.1 below that the Kurdyka-Łojasiewicz inequality always holds around noncritical points, which is also known (cf. Remark 3.2 in [10]).

For ANNs with analytic activation functions the risk function was shown to be analytic in Dereich & Kassing [21, Theorem 4.2] (for an arbitrary compactly supported input distribution). It therefore satisfies an analogous Kurdyka-Łojasiewicz inequality.

6.1 Semi-analytic and subanalytic sets

Definition 6.1 (Set of real analytic functions)**.** *Let $n \in \mathbb{N}$ and let $U \subseteq \mathbb{R}^n$ be an open set. Then we denote by $\mathscr{A}_U \subseteq C^\infty(U, \mathbb{R})$ the set of all real analytic functions from U to $\mathbb{R}$.*

For the next notions see, e.g., Bolte et al. [10, Definition 2.1] and Van den Dries & Miller [69].

Definition 6.2 (Multidimensional semi-analytic sets)**.** *Let $n \in \mathbb{N}$ and let $A \subseteq \mathbb{R}^n$ be a set. Then we say that A is an n-dimensional semi-analytic set if and only if for all $v \in \mathbb{R}^n$ there exist $M, N \in \mathbb{N}$, an open $U \subseteq \mathbb{R}^n$, and $(P_{i,j,k})_{(i,j,k) \in \{1,2,\ldots,M\} \times \{1,2,\ldots,N\} \times \{0,1\}} \subseteq \mathscr{A}_U$ such that $v \in U$*

and

$$A \cap U = \bigcup_{i=1}^{M} \left(\bigcap_{j=1}^{N} \{ x \in U \colon P_{i,j,0}(x) = 0 < P_{i,j,1}(x) \} \right) \tag{6.1}$$

(cf. Definition 6.1).

Definition 6.3 (Multidimensional subanalytic sets)**.** *Let $n \in \mathbb{N}$ and let $A \subseteq \mathbb{R}^n$ be a set. Then we say that A is an n-dimensional subanalytic set if and only if for all $v \in A$ there exist $m \in \mathbb{N}$, an open $U \subseteq \mathbb{R}^n$, and a bounded $(n+m)$-dimensional semi-analytic set $B \subseteq \mathbb{R}^{n+m}$ such that $v \in U$ and*

$$A \cap U = \{ x \in \mathbb{R}^n \colon (\exists\, y \in \mathbb{R}^m \colon (x, y) \in B) \} \tag{6.2}$$

(cf. Definition 6.2).

6.2 Subanalytic functions

Definition 6.4 (Subanalytic functions)**.** *Let $m, n \in \mathbb{N}$ and let $f \colon \mathbb{R}^m \to \mathbb{R}^n$ be a function. Then we say that f is a subanalytic function (we say that f is subanalytic) if and only if it holds that* $\operatorname{Graph}(f)$ *is an $(m+n)$-dimensional subanalytic set (cf. Definition 6.3).*

Proposition 6.1. *Let $m, n \in \mathbb{N}$ and let $f \colon \mathbb{R}^m \to \mathbb{R}^n$ be semi-algebraic (cf. Definition 4.4). Then f is subanalytic (cf. Definition 6.4).*

Proof of Proposition 6.1. Note that the assumption that f is semi-algebraic demonstrates that $\operatorname{Graph}(f)$ is an $(m+n)$-dimensional semi-algebraic set (cf. Definition 4.3). Moreover, it is well-known that every semi-algebraic set is subanalytic (cf., e.g., [69, Section 2.5]). Hence, we obtain that $\operatorname{Graph}(f)$ is an $(m+n)$-dimensional subanalytic set. The proof of Proposition 6.1 is thus complete. ☐

6.3 Lower semi-continuity of the nonsmooth slope

Lemma 6.1. *Let $\mathfrak{d} \in \mathbb{N}$, $f \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, let $\mathbf{M} \colon \mathbb{R}^{\mathfrak{d}} \to [0, \infty]$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathbf{M}(\theta) = \inf\left(\{ r \in \mathbb{R} \colon (\exists\, h \in (\mathbb{D}f)(\theta) \colon r = \|h\|) \} \cup \{\infty\} \right), \tag{6.3}$$

and let $\theta = (\theta_n)_{n \in \mathbb{N}_0} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy $\limsup_{n \to \infty} \|\theta_n - \theta_0\| = 0$ (cf. Definition 3.1). Then $\liminf_{n \to \infty} \mathbf{M}(\theta_n) \geq \mathbf{M}(\theta_0)$.

Proof of Lemma 6.1. Throughout this proof let $\mathbf{m} \in [0, \infty]$ satisfy $\mathbf{m} = \liminf_{n \to \infty} \mathbf{M}(\theta_n)$ and assume without loss of generality that

$$\mathbf{m} < \infty. \tag{6.4}$$

Observe that (6.4) assures that there exists a strictly increasing $N \colon \mathbb{N} \to \mathbb{N}$ which satisfies

$$\limsup_{n \to \infty} |\mathbf{M}(\theta_{N(n)}) - \mathbf{m}| = 0 \qquad \text{and} \qquad \sup_{n \in \mathbb{N}} \mathbf{M}(\theta_{N(n)}) < \infty. \tag{6.5}$$

Note that (6.3) and (6.5) prove that there exists $h = (h_n)_{n \in \mathbb{N}} \colon \mathbb{N} \to \mathbb{R}^{\mathfrak{d}}$ which satisfies for all $n \in \mathbb{N}$ that

$$h_n \in (\mathbb{D}f)(\theta_{N(n)}) \qquad \text{and} \qquad \|h_n\| \leq \mathbf{M}(\theta_{N(n)}) + n^{-1}. \tag{6.6}$$

Observe that (6.5) and (6.6) demonstrate that there exist $\mathbf{h} \in \mathbb{R}^{\mathfrak{d}}$ and a strictly increasing $M\colon \mathbb{N} \to \mathbb{N}$ which satisfy

$$\limsup_{n\to\infty} \|h_{M(n)} - \mathbf{h}\| = 0. \tag{6.7}$$

Note that (6.6), (6.7), and the assumption that $\limsup_{n\to\infty} \|\theta_n - \theta_0\| = 0$ demonstrate that $\limsup_{k\to\infty}(\|h_{M(k)} - \mathbf{h}\| + \|\theta_{N(M(k))} - \theta_0\|) = 0$ and $\forall\, k \in \mathbb{N}\colon h_{M(k)} \in (\mathbb{D}\mathcal{L}_\infty)(\theta_{N(M(k))})$. Combining this and Lemma 3.4 (applied with $n \curvearrowleft \mathfrak{d}$, $f \curvearrowleft f$, $x_0 \curvearrowleft \theta_0$, $(x_k)_{k\in\mathbb{N}} \curvearrowleft (\theta_{N(M(k))})_{k\in\mathbb{N}}$, $y_0 \curvearrowleft \mathbf{h}$, $(y_k)_{k\in\mathbb{N}} \curvearrowleft (h_{M(k)})_{k\in\mathbb{N}}$ in the notation of Lemma 3.4) proves that $\mathbf{h} \in (\mathbb{D}f)(\theta_0)$. This, (6.3), (6.5), (6.6), and (6.7) show that

$$\mathbf{M}(\theta_0) \le \|\mathbf{h}\| = \limsup_{n\to\infty} \|h_{M(n)}\| \le \limsup_{n\to\infty}(\mathbf{M}(\theta_{N(M(n))}) + [M(n)]^{-1}) = \mathbf{m}. \tag{6.8}$$

The proof of Lemma 6.1 is thus complete. □

Corollary 6.1. *Let $\mathfrak{d} \in \mathbb{N}$, $f \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, let $\mathbf{M}\colon \mathbb{R}^{\mathfrak{d}} \to [0,\infty]$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathbf{M}(\theta) = \inf\bigl(\{r \in \mathbb{R}\colon (\exists\, h \in (\mathbb{D}f)(\theta)\colon r = \|h\|)\} \cup \{\infty\}\bigr), \tag{6.9}$$

let $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ satisfy $0 \notin (\mathbb{D}f)(\vartheta)$, and let $\mathfrak{a} \in [0,1)$ (cf. Definition 3.1). Then there exist $\varepsilon, \mathfrak{C} \in (0,\infty)$ such that for all $\theta \in \{\psi \in \mathbb{R}^{\mathfrak{d}}\colon \|\psi - \vartheta\| < \varepsilon\}$ it holds that $|f(\theta) - f(\vartheta)|^{\mathfrak{a}} \le \mathfrak{C}\mathbf{M}(\theta)$.

Proof of Corollary 6.1. Observe that item (v) in Lemma 3.3, (6.9), and the assumption that $0 \notin (\mathbb{D}f)(\vartheta)$ prove that $\mathbf{M}(\vartheta) > 0$. Combining this with Lemma 6.1 proves that there exists $\varepsilon \in (0,\infty)$ such that for all $\theta \in \{\psi \in \mathbb{R}^{\mathfrak{d}}\colon \|\psi - \vartheta\| < \varepsilon\}$ it holds that $0 < \frac{\mathbf{M}(\vartheta)}{2} \le \mathbf{M}(\theta)$. Furthermore, note that the assumption that $f \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ assures that there exists $\varepsilon \in (0,\infty)$ such that for all $\theta \in \{\psi \in \mathbb{R}^{\mathfrak{d}}\colon \|\psi - \vartheta\| < \varepsilon\}$ it holds that $|f(\theta) - f(\vartheta)|^{\mathfrak{a}} \le 1$. The proof of Corollary 6.1 is thus complete. □

6.4 Generalized Kurdyka-Łojasiewicz inequalities for the training of deep ANNs

Proposition 6.2 (Generalized Łojasiewicz inequalities)**.** *Assume Setting 3.1, assume for all $i \in \{1, 2, \ldots, \ell_L\}$ that f_i is piecewise polynomial, let $\mathfrak{p}\colon [a,b]^{\ell_0} \to \mathbb{R}$ be piecewise polynomial, assume for all $E \in \mathcal{B}([a,b]^{\ell_0})$ that $\mu(E) = \int_E \mathfrak{p}(x)\,\mathrm{d}x$, and let $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ (cf. Definition 5.1). Then there exist $\varepsilon, \mathfrak{C} \in (0,\infty)$, $\mathfrak{a} \in [0,1)$ such that for all $\theta \in \{\psi \in \mathbb{R}^{\mathfrak{d}}\colon \|\psi - \vartheta\| < \varepsilon\}$, $\alpha \in [\mathfrak{a}, 1]$ it holds that*

$$|\mathcal{L}_\infty(\theta) - \mathcal{L}_\infty(\vartheta)|^\alpha \le \mathfrak{C}\|\mathcal{G}(\theta)\|. \tag{6.10}$$

Proof of Proposition 6.2. Throughout this proof for every $\varepsilon \in (0,\infty)$ let $B_\varepsilon \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy

$$B_\varepsilon = \{\theta \in \mathbb{R}^{\mathfrak{d}}\colon \|\theta - \vartheta\| < \varepsilon\} \tag{6.11}$$

and let $\mathbf{M}\colon \mathbb{R}^{\mathfrak{d}} \to [0,\infty]$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathbf{M}(\theta) = \inf\bigl(\{r \in \mathbb{R}\colon (\exists\, h \in (\mathbb{D}\mathcal{L}_\infty)(\theta)\colon r = \|h\|)\} \cup \{\infty\}\bigr) \tag{6.12}$$

(cf. Definition 3.1). Observe that Proposition 3.3 implies for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $\mathcal{G}(\theta) \in (\mathbb{D}\mathcal{L}_\infty)(\theta)$. Combining this with (6.12) shows that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\mathbf{M}(\theta) \leq \|\mathcal{G}(\theta)\|. \tag{6.13}$$

Furthermore, note that Lemma 3.1 implies that

$$\mathcal{L}_\infty \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}). \tag{6.14}$$

Therefore, we obtain for all $\varepsilon \in (0, \infty)$, $r \in [0, 1]$ that

$$\sup\nolimits_{\psi \in B_\varepsilon}\big(|\mathcal{L}_\infty(\psi) - \mathcal{L}_\infty(\vartheta)|^r\big) \leq \max\big\{1, \sup\nolimits_{\psi \in B_\varepsilon}|\mathcal{L}_\infty(\psi) - \mathcal{L}_\infty(\vartheta)|\big\} < \infty. \tag{6.15}$$

Moreover, observe that Corollary 5.1 assures that $\mathcal{L}_\infty$ is semi-algebraic. Proposition 6.1 hence proves that $\mathcal{L}_\infty$ is subanalytic. Combining this, (6.11), (6.12), (6.14), Corollary 6.1 (applied with $f \curvearrowleft \mathcal{L}_\infty$ in the notation of Corollary 6.1), and Bolte et al. [10, Theorem 3.1 and (4)] (applied with $n \curvearrowleft \mathfrak{d}$, $f \curvearrowleft (\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \mathcal{L}_\infty(\theta) \in \mathbb{R} \cup \{\infty\})$ in the notation of Bolte et al. [10, Theorem 3.1]) ensures that there exist $\varepsilon, \mathfrak{C} \in (0, \infty)$, $\mathfrak{a} \in [0, 1)$ which satisfy for all $\theta \in B_\varepsilon$ that

$$|\mathcal{L}_\infty(\theta) - \mathcal{L}_\infty(\vartheta)|^{\mathfrak{a}} \leq \mathfrak{C}\mathbf{M}(\theta). \tag{6.16}$$

Note that (6.13) and (6.16) assure for all $\theta \in B_\varepsilon$ that $|\mathcal{L}_\infty(\theta) - \mathcal{L}_\infty(\vartheta)|^{\mathfrak{a}} \leq \mathfrak{C}\|\mathcal{G}(\theta)\|$. Combining this with (6.15) demonstrates that for all $\theta \in B_\varepsilon$, $\alpha \in [\mathfrak{a}, 1]$ it holds that

$$\begin{aligned}
|\mathcal{L}_\infty(\theta) - \mathcal{L}_\infty(\vartheta)|^\alpha &\leq |\mathcal{L}_\infty(\theta) - \mathcal{L}_\infty(\vartheta)|^{\mathfrak{a}}\big[\sup\nolimits_{\psi \in B_\varepsilon}\big(|\mathcal{L}_\infty(\psi) - \mathcal{L}_\infty(\vartheta)|^{\alpha - \mathfrak{a}}\big)\big] \\
&\leq |\mathcal{L}_\infty(\theta) - \mathcal{L}_\infty(\vartheta)|^{\mathfrak{a}}\big[\max\big\{1, \sup\nolimits_{\psi \in B_\varepsilon}|\mathcal{L}_\infty(\psi) - \mathcal{L}_\infty(\vartheta)|\big\}\big] \\
&\leq \mathfrak{C}\big[\max\big\{1, \sup\nolimits_{\psi \in B_\varepsilon}|\mathcal{L}_\infty(\psi) - \mathcal{L}_\infty(\vartheta)|\big\}\big]\|\mathcal{G}(\theta)\| < \infty.
\end{aligned} \tag{6.17}$$

The proof of Proposition 6.2 is thus complete. □

7 Convergence analysis for solutions of GF differential equations

In Proposition 7.1 below we establish an abstract local convergence result for GF processes under the assumption that a Kurdyka-Łojasiewicz inequality is satisfied. The arguments used in the proof of Proposition 7.1 are essentially well-known in the scientific literature; see, e.g., Kurdyka et al. [47, Section 1], Bolte et al. [10, Theorem 4.5], Absil et al. [1, Theorem 2.2], or our previous article Eberle et al. [28] (see also [22] for a version for SDEs).

The above mentioned works [1, 10, 22, 47] assume that the objective function is C^1 or satisfies some other regularity conditions (in [10] the objective function is required to be lower-C^2 or convex). Some works, e.g. [1], also assume a certain weak decrease condition for the objective function. These assumptions are not necessary for our proof of Proposition 7.1. In fact, we do not even assume that $\mathcal{G}$ is a subgradient of the objective function $\mathcal{L}$ at every point. The only regularity we need is the chain rule $\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|\mathcal{G}(\Theta_s)\|^2 \, \mathrm{d}s$ for all $t \in [0, \infty)$. Therefore, our result is not implied by the mentioned previous works.

In Corollary 7.1 below we then prove a simplified version of Proposition 7.1. Afterwards, in Proposition 7.2 and Corollary 7.2 below we derive global convergence of every non-divergent GF trajectory. Finally, in Theorem 7.1 below we combine Corollary 7.2 with the Kurdyka-Łojasiewicz inequality for the risk function in Proposition 6.2 and the fact that the generalized gradient is a limiting subdifferential of the risk function (cf. Proposition 3.3) to establish the convergence of GF in the considered deep ANN framework and, thereby, prove Theorem 1.3 from the introduction.

7.1 Abstract local convergence results for GF processes

Proposition 7.1. *Let* $\mathfrak{d} \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathfrak{c} \in \mathbb{R}$, $\mathfrak{C}, \varepsilon \in (0, \infty)$, $\alpha \in (0,1)$, $\Theta \in C([0,\infty), \mathbb{R}^{\mathfrak{d}})$, $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, *let* $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ *be measurable, assume for all* $t \in [0,\infty)$ *that* $\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|\mathcal{G}(\Theta_s)\|^2 \, \mathrm{d}s$ *and* $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) \, \mathrm{d}s$, *and assume for all* $\theta \in \mathbb{R}^{\mathfrak{d}}$ *with* $\|\theta - \vartheta\| < \varepsilon$ *that*

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \le \mathfrak{C}\|\mathcal{G}(\theta)\|, \quad \mathfrak{c} = |\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|, \tag{7.1a}$$

$$\mathfrak{C}(1-\alpha)^{-1}\mathfrak{c}^{1-\alpha} + \|\Theta_0 - \vartheta\| < \varepsilon, \tag{7.1b}$$

and $\inf_{t \in \{s \in [0,\infty) \colon \forall\, r \in [0,s] \colon \|\Theta_r - \vartheta\| < \varepsilon\}} \mathcal{L}(\Theta_t) \ge \mathcal{L}(\vartheta)$. *Then there exists* $\psi \in \mathcal{L}^{-1}(\{\mathcal{L}(\vartheta)\})$ *such that*

(i) it holds for all $t \in [0,\infty)$ *that* $\|\Theta_t - \vartheta\| < \varepsilon$,

(ii) it holds for all $t \in [0,\infty)$ *that* $0 \le \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) \le \mathfrak{C}^2\mathfrak{c}^2(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2\mathfrak{c} + \mathfrak{c}^{2\alpha}t)^{-1}$, *and*

(iii) it holds for all $t \in [0,\infty)$ *that*

$$\begin{aligned}\|\Theta_t - \psi\| &\le \int_t^\infty \|\mathcal{G}(\Theta_s)\| \, \mathrm{d}s \le \mathfrak{C}(1-\alpha)^{-1}[\mathcal{L}(\Theta_t) - \mathcal{L}(\psi)]^{1-\alpha} \\ &\le \mathfrak{C}^{3-2\alpha}\mathfrak{c}^{2-2\alpha}(1-\alpha)^{-1}(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2\mathfrak{c} + \mathfrak{c}^{2\alpha}t)^{\alpha-1}.\end{aligned} \tag{7.2}$$

The assumption that $\inf_{t \in \{s \in [0,\infty) \colon \forall\, r \in [0,s] \colon \|\Theta_r - \vartheta\| < \varepsilon\}} \mathcal{L}(\Theta_t) \ge \mathcal{L}(\vartheta)$ means that for all $t \in [0,\infty)$ which satisfy that the trajectory of Θ remains within distance ε of ϑ until time t, it holds that $\mathcal{L}(\Theta_t) \ge \mathcal{L}(\vartheta)$. This assumption is in particular satisfied if ϑ is a local minimum of $\mathcal{L}$ with $\forall\, \theta \in \{\psi \in \mathbb{R}^{\mathfrak{d}} \colon \|\psi - \vartheta\| < \varepsilon\} \colon \mathcal{L}(\theta) \ge \mathcal{L}(\vartheta)$. But the statement of Proposition 7.1 also covers more general cases, since we only assume this lower bound for the values $\mathcal{L}(\Theta_t)$ and not for all values of $\mathcal{L}$ in a neighborhood of ϑ.

Proof of Proposition 7.1. Throughout this proof let $\mathbb{L} \colon [0,\infty) \to \mathbb{R}$ satisfy for all $t \in [0,\infty)$ that

$$\mathbb{L}(t) = \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta), \tag{7.3}$$

let $\mathbb{B} \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy

$$\mathbb{B} = \{\theta \in \mathbb{R}^{\mathfrak{d}} \colon \|\theta - \vartheta\| < \varepsilon\}, \tag{7.4}$$

let $T \in [0,\infty]$ satisfy

$$T = \inf(\{t \in [0,\infty) \colon \Theta_t \notin \mathbb{B}\} \cup \{\infty\}), \tag{7.5}$$

let $\tau \in [0, T]$ satisfy

$$\tau = \inf(\{t \in [0, T) \colon \mathbb{L}(t) = 0\} \cup \{T\}), \tag{7.6}$$

let $\ = (\ _t)_{t \in [0,\infty)} \colon [0, \infty) \to [0, \infty]$ satisfy for all $t \in [0, \infty)$ that $\ _t = \int_t^\infty \|\mathcal{G}(\Theta_s)\| \, \mathrm{d}s$, and let $\mathfrak{D} \in \mathbb{R}$ satisfy $\mathfrak{D} = \mathfrak{C}^2 \mathfrak{c}^{(2-2\alpha)}$. In the first step of our proof of items (i), (ii), and (iii) we show that for all $t \in [0, \infty)$ it holds that

$$\Theta_t \in \mathbb{B}. \tag{7.7}$$

For this we observe that (7.1), the triangle inequality, and the assumption that for all $t \in [0, \infty)$ it holds that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) \, \mathrm{d}s$ ensure that for all $t \in [0, \infty)$ we have that

$$\begin{aligned} \|\Theta_t - \vartheta\| &\leq \|\Theta_t - \Theta_0\| + \|\Theta_0 - \vartheta\| \\ &\leq \left\| \int_0^t \mathcal{G}(\Theta_s) \, \mathrm{d}s \right\| + \|\Theta_0 - \vartheta\| \\ &\leq \int_0^t \|\mathcal{G}(\Theta_s)\| \, \mathrm{d}s + \|\Theta_0 - \vartheta\| \\ &< \int_0^t \|\mathcal{G}(\Theta_s)\| \, \mathrm{d}s - \mathfrak{C}(1-\alpha)^{-1} |\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|^{1-\alpha} + \varepsilon. \end{aligned} \tag{7.8}$$

To establish (7.7), it thus sufficient to prove that $\int_0^T \|\mathcal{G}(\Theta_s)\| \, \mathrm{d}s \leq \mathfrak{C}(1-\alpha)^{-1} |\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|^{1-\alpha}$. We will accomplish this by employing an appropriate differential inequality for a fractional power of the function $\mathbb{L}$ in (7.3) (see (7.13) below for details). For this we need several technical preparations. More formally, observe that (7.3) and the assumption that for all $t \in [0, \infty)$ it holds that

$$\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|\mathcal{G}(\Theta_s)\|^2 \, \mathrm{d}s \tag{7.9}$$

imply that for almost all $t \in [0, \infty)$ it holds that $\mathbb{L}$ is differentiable at t and satisfies

$$\mathbb{L}'(t) = \tfrac{\mathrm{d}}{\mathrm{d}t}(\mathcal{L}(\Theta_t)) = -\|\mathcal{G}(\Theta_t)\|^2. \tag{7.10}$$

Moreover, note that the assumption that $\inf_{t \in \{s \in [0,\infty) \colon \forall r \in [0,s] \colon \|\Theta_r - \vartheta\| < \varepsilon\}} \mathcal{L}(\Theta_t) \geq \mathcal{L}(\vartheta)$ assures for all $t \in [0, T)$ that

$$\mathbb{L}(t) \geq 0. \tag{7.11}$$

Combining this with (7.1), (7.3), and (7.6) demonstrates for all $t \in [0, \tau)$ that

$$0 < [\mathbb{L}(t)]^\alpha = |\mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta)|^\alpha \leq \mathfrak{C} \|\mathcal{G}(\Theta_t)\|. \tag{7.12}$$

The chain rule and (7.10) hence prove that for almost all $t \in [0, \tau)$ it holds that

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}([\mathbb{L}(t)]^{1-\alpha}) &= (1-\alpha)[\mathbb{L}(t)]^{-\alpha}(-\|\mathcal{G}(\Theta_t)\|^2) \\ &\leq -(1-\alpha)\mathfrak{C}^{-1} \|\mathcal{G}(\Theta_t)\|^{-1} \|\mathcal{G}(\Theta_t)\|^2 \\ &= -\mathfrak{C}^{-1}(1-\alpha) \|\mathcal{G}(\Theta_t)\|. \end{aligned} \tag{7.13}$$

Next observe that (7.9) ensures that $[0,\infty) \ni t \mapsto \mathbb{L}(t) \in \mathbb{R}$ is absolutely continuous. This and the fact that for all $r \in (0,\infty)$ it holds that $[r,\infty) \ni y \mapsto y^{1-\alpha} \in \mathbb{R}$ is Lipschitz continuous demonstrate that for all $t \in [0,\tau)$ it holds that $[0,t] \ni s \mapsto [\mathbb{L}(s)]^{1-\alpha} \in \mathbb{R}$ is absolutely continuous. Combining this with (7.13) shows that for all $s,t \in [0,\tau)$ with $s \leq t$ it holds that

$$\int_s^t \|\mathcal{G}(\Theta_u)\| \, \mathrm{d}u \leq -\mathfrak{C}(1-\alpha)^{-1}([\mathbb{L}(t)]^{1-\alpha} - [\mathbb{L}(s)]^{1-\alpha}) \leq \mathfrak{C}(1-\alpha)^{-1}[\mathbb{L}(s)]^{1-\alpha}. \tag{7.14}$$

In the next step we note that (7.9) ensures that $[0,\infty) \ni t \mapsto \mathcal{L}(\Theta_t) \in \mathbb{R}$ is non-increasing. This and (7.3) prove that $\mathbb{L}$ is non-increasing. Combining (7.6) and (7.11) hence implies that for all $t \in [\tau, T)$ it holds that $\mathbb{L}(t) = 0$. Therefore, we obtain for all $t \in (\tau, T)$ that

$$\mathbb{L}'(t) = 0. \tag{7.15}$$

This and (7.10) assure that for almost all $t \in (\tau, T)$ it holds that

$$\mathcal{G}(\Theta_t) = 0. \tag{7.16}$$

Combining this with (7.14) demonstrates that for all $s,t \in [0,T)$ with $s \leq t$ it holds that

$$\int_s^t \|\mathcal{G}(\Theta_u)\| \, \mathrm{d}u \leq \mathfrak{C}(1-\alpha)^{-1}[\mathbb{L}(s)]^{1-\alpha}. \tag{7.17}$$

Hence, we obtain for all $t \in [0,T)$ that

$$\int_0^t \|\mathcal{G}(\Theta_u)\| \, \mathrm{d}u \leq \mathfrak{C}(1-\alpha)^{-1}[\mathbb{L}(0)]^{1-\alpha}. \tag{7.18}$$

In addition, observe that (7.1) assures that $\Theta_0 \in \mathbb{B}$. Combining this with (7.5) proves that $T > 0$. This, (7.18), and (7.1) demonstrate that

$$\int_0^T \|\mathcal{G}(\Theta_u)\| \, \mathrm{d}u \leq \mathfrak{C}(1-\alpha)^{-1}[\mathbb{L}(0)]^{1-\alpha} < \varepsilon < \infty. \tag{7.19}$$

Combining (7.5) and (7.8) hence assures that

$$T = \infty. \tag{7.20}$$

This establishes (7.7). In the next step of our proof of items (i), (ii), and (iii) we verify that $\Theta_t \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0,\infty)$, is convergent (see (7.22) below). For this observe that the assumption that for all $t \in [0,\infty)$ it holds that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) \, \mathrm{d}s$ demonstrates that for all $r,s,t \in [0,\infty)$ with $r \leq s \leq t$ it holds that

$$\|\Theta_t - \Theta_s\| = \left\| \int_s^t \mathcal{G}(\Theta_u) \, \mathrm{d}u \right\| \leq \int_s^t \|\mathcal{G}(\Theta_u)\| \, \mathrm{d}u \leq \int_r^\infty \|\mathcal{G}(\Theta_u)\| \, \mathrm{d}u = {}_r. \tag{7.21}$$

Moreover, note that (7.19) and (7.20) assure that $\infty > {}_0 \geq \limsup_{r \to \infty} {}_r = 0$. Combining this with (7.21) proves that there exist $\psi \in \mathbb{R}^{\mathfrak{d}}$ which satisfies

$$\limsup_{t \to \infty} \|\Theta_t - \psi\| = 0. \tag{7.22}$$

In the next step of our proof of items (i), (ii), and (iii) we show that $\mathcal{L}(\Theta_t)$, $t \in [0, \infty)$, converges to $\mathcal{L}(\psi)$ with convergence order 1. We accomplish this bringing a suitable differential inequality for the reciprocal of the function $\mathbb{L}$ in (7.3) into play (see (7.25) below for details). More specifically, note that (7.10), (7.20), (7.5), and (7.1) demonstrate that for almost all $t \in [0, \infty)$ it holds that

$$\mathbb{L}'(t) = -\|\mathcal{G}(\Theta_t)\|^2 \leq -\mathfrak{C}^{-2}[\mathbb{L}(t)]^{2\alpha}. \tag{7.23}$$

Hence, we obtain that $\mathbb{L}$ is non-increasing. This shows for all $t \in [0, \infty)$ that $\mathbb{L}(t) \leq \mathbb{L}(0)$. This and the fact that for all $t \in [0, \tau)$ it holds that $\mathbb{L}(t) > 0$ show that for almost all $t \in [0, \tau)$ we have that

$$\mathbb{L}'(t) \leq -\mathfrak{C}^{-2}[\mathbb{L}(t)]^{(2\alpha-2)}[\mathbb{L}(t)]^2 \leq -\mathfrak{C}^{-2}[\mathbb{L}(0)]^{(2\alpha-2)}[\mathbb{L}(t)]^2 = -\mathfrak{D}^{-1}[\mathbb{L}(t)]^2. \tag{7.24}$$

Therefore, we obtain that for almost all $t \in [0, \tau)$ it holds that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathfrak{D}}{\mathbb{L}(t)}\right) = -\left(\frac{\mathfrak{D}\,\mathbb{L}'(t)}{[\mathbb{L}(t)]^2}\right) \geq 1. \tag{7.25}$$

Moreover, observe that the fact that for all $t \in [0, \tau)$ it holds that $[0, t] \ni s \mapsto \mathbb{L}(s) \in (0, \infty)$ is absolutely continuous proves that for all $t \in [0, \tau)$ we have that $[0, t] \ni s \mapsto \mathfrak{D}[\mathbb{L}(s)]^{-1} \in (0, \infty)$ is absolutely continuous. This and (7.25) imply for all $t \in [0, \tau)$ that $\frac{\mathfrak{D}}{\mathbb{L}(t)} - \frac{\mathfrak{D}}{\mathbb{L}(0)} \geq t$. Hence, we obtain for all $t \in [0, \tau)$ that $\frac{\mathfrak{D}}{\mathbb{L}(t)} \geq \frac{\mathfrak{D}}{\mathbb{L}(0)} + t$. Therefore, we get for all $t \in [0, \tau)$ that $\mathfrak{D}\,(\frac{\mathfrak{D}}{\mathbb{L}(0)} + t)^{-1} \geq \mathbb{L}(t)$. This implies for all $t \in [0, \tau)$ that

$$\mathbb{L}(t) \leq \mathfrak{D}\,(\mathfrak{D}[\mathbb{L}(0)]^{-1} + t)^{-1} = \mathfrak{C}^2\mathfrak{c}^{2-2\alpha}(\mathfrak{C}^2\mathfrak{c}^{1-2\alpha} + t)^{-1} = \mathfrak{C}^2\mathfrak{c}^2(\mathfrak{C}^2\mathfrak{c} + \mathfrak{c}^{2\alpha}t)^{-1}. \tag{7.26}$$

The fact that for all $t \in [\tau, \infty)$ it holds that $\mathbb{L}(t) = 0$ and (7.6) therefore prove that for all $t \in [0, \infty)$ it holds that

$$0 \leq \mathbb{L}(t) \leq \mathfrak{C}^2\mathfrak{c}^2(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2\mathfrak{c} + \mathfrak{c}^{2\alpha}t)^{-1}. \tag{7.27}$$

Next note that (7.22) and the assumption that $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ assure that $\limsup_{t\to\infty}|\mathcal{L}(\Theta_t) - \mathcal{L}(\psi)| = 0$. Combining this with (7.27) demonstrates that $\mathcal{L}(\psi) = \mathcal{L}(\vartheta)$. This and (7.27) ensure for all $t \in [0, \infty)$ that

$$0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) \leq \mathfrak{C}^2\mathfrak{c}^2(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2\mathfrak{c} + \mathfrak{c}^{2\alpha}t)^{-1}. \tag{7.28}$$

In the final step of our proof of items (i), (ii), and (iii) we establish convergence rates for the real numbers $\|\Theta_t - \psi\|$, $t \in [0, \infty)$. Observe that (7.22), (7.21), and (7.17) assure for all $t \in [0, \infty)$ that

$$\|\Theta_t - \psi\| = \|\Theta_t - [\lim_{s\to\infty}\Theta_s]\| = \lim_{s\to\infty}\|\Theta_t - \Theta_s\| \leq {}_t \leq \mathfrak{C}(1-\alpha)^{-1}[\mathbb{L}(t)]^{1-\alpha}. \tag{7.29}$$

This and (7.28) ensure for all $t \in [0, \infty)$ that

$$\begin{aligned}
\|\Theta_t - \psi\| &\leq {}_t \leq \mathfrak{C}(1-\alpha)^{-1}[\mathcal{L}(\Theta_t) - \mathcal{L}(\psi)]^{1-\alpha} \\
&\leq \mathfrak{C}(1-\alpha)^{-1}\left[\mathfrak{C}^2\mathfrak{c}^2(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2\mathfrak{c} + \mathfrak{c}^{2\alpha}t)^{-1}\right]^{1-\alpha} \\
&= \mathfrak{C}^{3-2\alpha}\mathfrak{c}^{2-2\alpha}(1-\alpha)^{-1}(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2\mathfrak{c} + \mathfrak{c}^{2\alpha}t)^{\alpha-1}.
\end{aligned} \tag{7.30}$$

Combining this with (7.7) and (7.28) establishes items (i), (ii), and (iii). The proof of Proposition 7.1 is thus complete. □

Corollary 7.1. *Let* $\mathfrak{d} \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathfrak{c} \in [0,1]$, $\mathfrak{C}, \varepsilon \in (0,\infty)$, $\alpha \in (0,1)$, $\Theta \in C([0,\infty),\mathbb{R}^{\mathfrak{d}})$, $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}},\mathbb{R})$, *let* $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ *be measurable, assume for all* $t \in [0,\infty)$ *that* $\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|\mathcal{G}(\Theta_s)\|^2 \, \mathrm{d}s$ *and* $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\, \mathrm{d}s$, *and assume for all* $\theta \in \mathbb{R}^{\mathfrak{d}}$ *with* $\|\theta - \vartheta\| < \varepsilon$ *that*

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{\alpha} \leq \mathfrak{C}\|\mathcal{G}(\theta)\|, \quad \mathfrak{c} = |\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|, \tag{7.31a}$$

$$\mathfrak{C}(1-\alpha)^{-1}\mathfrak{c}^{1-\alpha} + \|\Theta_0 - \vartheta\| < \varepsilon, \tag{7.31b}$$

and $\inf_{t\in\{s\in[0,\infty)\colon \forall\, r\in[0,s]\colon \|\Theta_r - \vartheta\| < \varepsilon\}} \mathcal{L}(\Theta_t) \geq \mathcal{L}(\vartheta)$. *Then there exists* $\psi \in \mathcal{L}^{-1}(\{\mathcal{L}(\vartheta)\})$ *such that for all* $t \in [0,\infty)$ *it holds that* $\|\Theta_t - \vartheta\| < \varepsilon$, $0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) \leq (1 + \mathfrak{C}^{-2}t)^{-1}$, *and*

$$\|\Theta_t - \psi\| \leq \textstyle\int_t^\infty \|\mathcal{G}(\Theta_s)\|\, \mathrm{d}s \leq \mathfrak{C}(1-\alpha)^{-1}(1 + \mathfrak{C}^{-2}t)^{\alpha-1}. \tag{7.32}$$

Proof of Corollary 7.1. Observe that Proposition 7.1 ensures that exists $\psi \in \mathcal{L}^{-1}(\{\mathcal{L}(\vartheta)\})$ which satisfies that

(i) it holds for all $t \in [0,\infty)$ that $\|\Theta_t - \vartheta\| < \varepsilon$,

(ii) it holds for all $t \in [0,\infty)$ that $0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) \leq \mathfrak{C}^2\mathfrak{c}^2(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2\mathfrak{c} + \mathfrak{c}^{2\alpha}t)^{-1}$, and

(iii) it holds for all $t \in [0,\infty)$ that

$$\begin{aligned}\|\Theta_t - \psi\| &\leq \int_t^\infty \|\mathcal{G}(\Theta_s)\|\, \mathrm{d}s \leq \mathfrak{C}(1-\alpha)^{-1}[\mathcal{L}(\Theta_t) - \mathcal{L}(\psi)]^{1-\alpha} \\ &\leq \mathfrak{C}^{3-2\alpha}\mathfrak{c}^{2-2\alpha}(1-\alpha)^{-1}(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2\mathfrak{c} + \mathfrak{c}^{2\alpha}t)^{\alpha-1}.\end{aligned} \tag{7.33}$$

Note that item (ii) and the assumption that $\mathfrak{c} \leq 1$ imply that for all $t \in [0,\infty)$ it holds that

$$0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) \leq \mathfrak{c}^2(\mathfrak{C}^{-2}\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c} + \mathfrak{C}^{-2}\mathfrak{c}^{2\alpha}t)^{-1} \leq (1 + \mathfrak{C}^{-2}t)^{-1}. \tag{7.34}$$

This and item (iii) ensure that for all $t \in [0,\infty)$ it holds that

$$\begin{aligned}\|\Theta_t - \psi\| &\leq \int_t^\infty \|\mathcal{G}(\Theta_s)\|\, \mathrm{d}s \leq \mathfrak{C}(1-\alpha)^{-1}[\mathcal{L}(\Theta_t) - \mathcal{L}(\psi)]^{1-\alpha} \\ &\leq \mathfrak{C}(1-\alpha)^{-1}(1 + \mathfrak{C}^{-2}t)^{\alpha-1}.\end{aligned} \tag{7.35}$$

Combining this with item (i) and (7.34) establishes (7.32). The proof of Corollary 7.1 is thus complete. □

7.2 Abstract global convergence results for GF processes

We next employ Corollary 7.1 to establish under a Kurdyka-Łojasiewicz assumption the convergence of every non-divergent GF trajectory. To prove Proposition 7.2 we note that the trajectory must have a convergent subsequence with limit $\vartheta \in \mathbb{R}^{\mathfrak{d}}$. Hence, for a sufficiently large time the GF reaches a neighborhood of ϑ where the conditions of Corollary 7.1 in (7.31) are satisfied, and thus we get convergence of the entire trajectory.

Proposition 7.2. *Let $\mathfrak{d} \in \mathbb{N}$, $\Theta \in C([0,\infty),\mathbb{R}^{\mathfrak{d}})$, $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}},\mathbb{R})$, let $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ be measurable, assume that for all $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ there exist $\varepsilon, \mathfrak{C} \in (0,\infty)$, $\alpha \in (0,1)$ such that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\|\theta - \vartheta\| < \varepsilon$ it holds that $|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{\alpha} \le \mathfrak{C}\|\mathcal{G}(\theta)\|$, and assume for all $t \in [0,\infty)$ that*

$$\liminf_{s\to\infty}\|\Theta_s\| < \infty, \quad \mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \textstyle\int_0^t \|\mathcal{G}(\Theta_s)\|^2\,\mathrm{d}s, \tag{7.36a}$$

$$\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s. \tag{7.36b}$$

Then there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathfrak{C}, \tau, \beta \in (0,\infty)$ such that for all $t \in [\tau,\infty)$ it holds that

$$\|\Theta_t - \vartheta\| \le \big(1 + \mathfrak{C}^{-1}(t-\tau)\big)^{-\beta} \qquad \text{and} \qquad 0 \le \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \le \big(1 + \mathfrak{C}^{-1}(t-\tau)\big)^{-1}. \tag{7.37}$$

Proof of Proposition 7.2. Observe that (7.36) implies that $[0,\infty) \ni t \mapsto \mathcal{L}(\Theta_t) \in \mathbb{R}$ is non-increasing. Therefore, we obtain that there exists $\mathbf{m} \in [-\infty,\infty)$ which satisfies

$$\mathbf{m} = \limsup_{t\to\infty} \mathcal{L}(\Theta_t) = \liminf_{t\to\infty} \mathcal{L}(\Theta_t) = \inf_{t\in[0,\infty)} \mathcal{L}(\Theta_t). \tag{7.38}$$

Furthermore, note that that the assumption that $\liminf_{t\to\infty}\|\Theta_t\| < \infty$ ensures that there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ and $\delta = (\delta_n)_{n\in\mathbb{N}}\colon \mathbb{N} \to [0,\infty)$ which satisfy

$$\liminf_{n\to\infty} \delta_n = \infty \qquad \text{and} \qquad \limsup_{n\to\infty}\|\Theta_{\delta_n} - \vartheta\| = 0. \tag{7.39}$$

Observe that (7.38), (7.39), and the fact that $\mathcal{L}$ is continuous show that

$$\mathcal{L}(\vartheta) = \mathbf{m} \in \mathbb{R} \qquad \text{and} \qquad \forall\, t \in [0,\infty)\colon \mathcal{L}(\Theta_t) \ge \mathcal{L}(\vartheta). \tag{7.40}$$

Next let $\varepsilon, \mathfrak{C} \in (0,\infty)$, $\alpha \in (0,1)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\|\theta - \vartheta\| < \varepsilon$ that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{\alpha} \le \mathfrak{C}\|\mathcal{G}(\theta)\|. \tag{7.41}$$

Note that (7.39) and the fact that $\mathcal{L}$ is continuous demonstrate that there exist $n \in \mathbb{N}$, $\mathfrak{c} \in [0,1]$ which satisfy

$$\mathfrak{c} = |\mathcal{L}(\Theta_{\delta_n}) - \mathcal{L}(\vartheta)| \qquad \text{and} \qquad \mathfrak{C}(1-\alpha)^{-1}\mathfrak{c}^{1-\alpha} + \|\Theta_{\delta_n} - \vartheta\| < \varepsilon. \tag{7.42}$$

Next let $\Phi\colon [0,\infty) \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $t \in [0,\infty)$ that

$$\Phi_t = \Theta_{\delta_n + t}. \tag{7.43}$$

Observe that (7.36), (7.40), and (7.43) assure that for all $t \in [0,\infty)$ it holds that

$$\mathcal{L}(\Phi_t) = \mathcal{L}(\Phi_0) - \textstyle\int_0^t\|\mathcal{G}(\Phi_s)\|^2\,\mathrm{d}s, \quad \Phi_t = \Phi_0 - \int_0^t \mathcal{G}(\Phi_s)\,\mathrm{d}s, \qquad \text{and} \qquad \mathcal{L}(\Phi_t) \ge \mathcal{L}(\vartheta). \tag{7.44}$$

Combining this with (7.41), (7.42), (7.43), and Corollary 7.1 (applied with $\Theta \curvearrowleft \Phi$ in the notation of Corollary 7.1) establishes that there exists $\psi \in \mathcal{L}^{-1}(\{\mathcal{L}(\vartheta)\})$ which satisfies for all $t \in [0,\infty)$ that

$$0 \le \mathcal{L}(\Phi_t) - \mathcal{L}(\psi) \le (1 + \mathfrak{C}^{-2}t)^{-1} \qquad \text{and} \qquad \|\Phi_t - \psi\| \le \mathfrak{C}(1-\alpha)^{-1}(1 + \mathfrak{C}^{-2}t)^{\alpha-1}. \tag{7.45}$$

Note that (7.43) and (7.45) assure for all $t \in [0, \infty)$ that $0 \leq \mathcal{L}(\Theta_{\delta_n + t}) - \mathcal{L}(\psi) \leq (1 + \mathfrak{C}^{-2} t)^{-1}$ and $\|\Theta_{\delta_n+t} - \psi\| \leq \mathfrak{C}(1-\alpha)^{-1}(1 + \mathfrak{C}^{-2}t)^{\alpha-1}$. Hence, we obtain for all $\tau \in [\delta_n, \infty)$, $t \in [\tau, \infty)$ that

$$\begin{aligned} 0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) &\leq (1 + \mathfrak{C}^{-2}(t - \delta_n))^{-1} \\ &= (1 + \mathfrak{C}^{-2}(t-\tau) + \mathfrak{C}^{-2}(\tau - \delta_n))^{-1} \\ &\leq (1 + \mathfrak{C}^{-2}(t-\tau))^{-1} \end{aligned} \tag{7.46}$$

and

$$\begin{aligned} \|\Theta_t - \psi\| &\leq \mathfrak{C}(1-\alpha)^{-1}(1 + \mathfrak{C}^{-2}(t-\delta_n))^{\alpha-1} \\ &= \Big[\big[\mathfrak{C}(1-\alpha)^{-1}\big]^{\frac{1}{\alpha-1}} (1 + \mathfrak{C}^{-2}(t - \delta_n)) \Big]^{\alpha-1} \\ &= \Big[\big[\mathfrak{C}(1-\alpha)^{-1}\big]^{\frac{1}{\alpha-1}} \big[1 + \mathfrak{C}^{-2}(\tau - \delta_n)\big] + \Big[\big[\mathfrak{C}(1-\alpha)^{-1}\big]^{\frac{1}{1-\alpha}} \mathfrak{C}^2\Big]^{-1} (t-\tau) \Big]^{\alpha-1}. \end{aligned} \tag{7.47}$$

Next let $\mathscr{C}, \tau \in (0, \infty)$ satisfy

$$\mathscr{C} = \max\big\{\mathfrak{C}^2, \big[\mathfrak{C}(1-\alpha)^{-1}\big]^{\frac{1}{1-\alpha}} \mathfrak{C}^2\big\} \qquad \text{and} \qquad \tau = \delta_n + \mathfrak{C}^2 \big[\mathfrak{C}(1-\alpha)^{-1}\big]^{\frac{1}{1-\alpha}}. \tag{7.48}$$

Observe that (7.47) and (7.48) demonstrate for all $t \in [\tau, \infty)$ that

$$0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) \leq (1 + \mathfrak{C}^{-2}(t-\tau))^{-1} \leq (1 + \mathscr{C}^{-1}(t - \tau))^{-1} \tag{7.49}$$

and

$$\begin{aligned} \|\Theta_t - \psi\| &\leq \Big[\big[\mathfrak{C}(1-\alpha)^{-1}\big]^{\frac{1}{\alpha-1}} \big[1 + \mathfrak{C}^{-2}(\tau - \delta_n)\big] + \mathscr{C}^{-1}(t-\tau) \Big]^{\alpha-1} \\ &= \Big[\big[\mathfrak{C}(1-\alpha)^{-1}\big]^{\frac{1}{\alpha-1}} \big[1 + \big[\mathfrak{C}(1-\alpha)^{-1}\big]^{\frac{1}{1-\alpha}}\big] + \mathscr{C}^{-1}(t-\tau) \Big]^{\alpha-1} \\ &\leq \big[1 + \mathscr{C}^{-1}(t-\tau)\big]^{\alpha-1}. \end{aligned} \tag{7.50}$$

The proof of Proposition 7.2 is thus complete. □

The next result, Corollary 7.2, is a simplified version of Proposition 7.2 where the sufficiently large finite time $\tau \in [0, \infty)$ is incorporated in the constant $\mathscr{C}$.

Corollary 7.2. *Let $\mathfrak{d} \in \mathbb{N}$, $\Theta \in C([0,\infty), \mathbb{R}^{\mathfrak{d}})$, $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, let $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ be measurable, assume that for all $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ there exist $\varepsilon, \mathfrak{C} \in (0, \infty)$, $\alpha \in (0,1)$ such that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\|\theta - \vartheta\| < \varepsilon$ it holds that $|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \leq \mathfrak{C}\|\mathcal{G}(\theta)\|$, and assume for all $t \in [0, \infty)$ that*

$$\liminf_{s\to\infty} \|\Theta_s\| < \infty, \quad \mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \textstyle\int_0^t \|\mathcal{G}(\Theta_s)\|^2 \, ds, \tag{7.51a}$$

$$\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\, ds. \tag{7.51b}$$

Then there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathscr{C}, \beta \in (0, \infty)$ which satisfy for all $t \in [0, \infty)$ that

$$\|\Theta_t - \vartheta\| \leq \mathscr{C}(1+t)^{-\beta} \qquad \text{and} \qquad 0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq \mathscr{C}(1+t)^{-1}. \tag{7.52}$$

Proof of Corollary 7.2. Note that Proposition 7.2 demonstrates that there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathfrak{C}, \tau, \beta \in (0, \infty)$ which satisfy for all $t \in [\tau, \infty)$ that

$$\|\Theta_t - \vartheta\| \leq \big(1 + \mathfrak{C}^{-1}(t-\tau)\big)^{-\beta} \qquad \text{and} \qquad 0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq \big(1 + \mathfrak{C}^{-1}(t-\tau)\big)^{-1}. \tag{7.53}$$

In the following let $\mathscr{C} \in (0, \infty)$ satisfy

$$\mathscr{C} = \max\big\{1+\tau, (1+\tau)^{\beta}, \mathfrak{C}, \mathfrak{C}^{\beta}, (1+\tau)^{\beta}\big(\sup\nolimits_{s\in[0,\tau]}\|\Theta_s - \vartheta\|\big), (1+\tau)|\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|\big\}. \tag{7.54}$$

Observe that (7.53), (7.54), and the fact that $[0, \infty) \ni t \mapsto \mathcal{L}(\Theta_t) \in \mathbb{R}$ is non-increasing show for all $t \in [0, \tau]$ that

$$\|\Theta_t - \vartheta\| \leq \sup\nolimits_{s\in[0,\tau]}\|\Theta_s - \vartheta\| \leq \mathscr{C}(1+\tau)^{-\beta} \leq \mathscr{C}(1+t)^{-\beta} \tag{7.55}$$

and

$$0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq \mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta) \leq \mathscr{C}(1+\tau)^{-1} \leq \mathscr{C}(1+t)^{-1}. \tag{7.56}$$

Furthermore, note that (7.53) and (7.54) imply for all $t \in [\tau, \infty)$ that

$$\begin{aligned}\|\Theta_t - \vartheta\| &\leq \big(1 + \mathfrak{C}^{-1}(t-\tau)\big)^{-\beta} = \mathscr{C}\big(\mathscr{C}^{1/\beta} + \mathscr{C}^{1/\beta}\mathfrak{C}^{-1}(t-\tau)\big)^{-\beta} \\ &\leq \mathscr{C}\big(\mathscr{C}^{1/\beta} + t - \tau\big)^{-\beta} \leq \mathscr{C}(1+t)^{-\beta}.\end{aligned} \tag{7.57}$$

Moreover, observe that (7.53) and (7.54) demonstrate for all $t \in [\tau, \infty)$ that

$$\begin{aligned}0 &\leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq \mathscr{C}\big(\mathscr{C} + \mathfrak{C}^{-1}\mathscr{C}(t-\tau)\big)^{-1} \\ &\leq \mathscr{C}\big(\mathscr{C} - \tau + t\big)^{-1} \leq \mathscr{C}(1+t)^{-1}.\end{aligned} \tag{7.58}$$

The proof of Corollary 7.2 is thus complete. □

7.3 Convergence of GF processes in the training of deep ANNs

Due to the Kurdyka-Łojasiewicz inequality for the risk function from Proposition 6.2 we are now able to apply Corollary 7.2 to the risk function $\mathcal{L}_\infty$ from Setting 3.1.

Theorem 7.1. *Assume Setting 3.1, assume for all $i \in \{1, 2, \ldots, \ell_L\}$ that f_i is piecewise polynomial, let $\mathfrak{p}\colon [a,b]^{\ell_0} \to \mathbb{R}$ be piecewise polynomial, assume for all $E \in \mathcal{B}([a,b]^{\ell_0})$ that $\mu(E) = \int_E \mathfrak{p}(x)\,\mathrm{d}x$, and let $\Theta \in C([0,\infty), \mathbb{R}^{\mathfrak{d}})$ satisfy $\liminf_{t\to\infty}\|\Theta_t\| < \infty$ and $\forall\, t \in [0,\infty)\colon \Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s$ (cf. Definition 5.1). Then there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathscr{C}, \beta \in (0,\infty)$ with $0 \in (\mathbb{D}\mathcal{L}_\infty)(\vartheta)$ such that for all $t \in [0,\infty)$ it holds that*

$$\|\Theta_t - \vartheta\| \leq \mathscr{C}(1+t)^{-\beta} \qquad \text{and} \qquad 0 \leq \mathcal{L}_\infty(\Theta_t) - \mathcal{L}_\infty(\vartheta) \leq \mathscr{C}(1+t)^{-1} \tag{7.59}$$

(cf. Definition 3.1).

Proof of Theorem 7.1. Note that Proposition 3.1 shows that $\mathcal{G}$ is measurable. Furthermore, observe that [36, Lemma 3.7] ensures that for all $t \in [0, \infty)$ it holds that

$$\mathcal{L}_\infty(\Theta_t) = \mathcal{L}_\infty(\Theta_0) - \int_0^t \|\mathcal{G}(\Theta_s)\|^2 \, \mathrm{d}s. \tag{7.60}$$

Moreover, note that Lemma 3.1 assures that

$$\mathcal{L}_\infty \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}). \tag{7.61}$$

In addition, observe that Proposition 6.2 shows that for all $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ there exist $\varepsilon, \mathfrak{C} \in (0, \infty)$, $\alpha \in (0, 1)$ such that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\|\theta - \vartheta\| < \varepsilon$ it holds that $|\mathcal{L}_\infty(\theta) - \mathcal{L}_\infty(\vartheta)|^\alpha \leq \mathfrak{C}\|\mathcal{G}(\theta)\|$. Corollary 7.2, the fact that $\mathcal{G}$ is measurable, (7.60), and (7.61) therefore demonstrate that there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathscr{C}, \beta \in (0, \infty)$ which satisfy for all $t \in [0, \infty)$ that

$$\|\Theta_t - \vartheta\| \leq \mathscr{C}(1 + t)^{-\beta} \qquad \text{and} \qquad 0 \leq \mathcal{L}_\infty(\Theta_t) - \mathcal{L}_\infty(\vartheta) \leq \mathscr{C}(1 + t)^{-1}. \tag{7.62}$$

Furthermore, note that (7.60) demonstrates that $\int_0^\infty \|\mathcal{G}(\Theta_s)\|^2 \, \mathrm{d}s < \infty$. Hence, we obtain $\liminf_{s \to \infty} \|\mathcal{G}(\Theta_s)\| = 0$. This implies that there exists a strictly increasing $\tau = (\tau_n)_{n \in \mathbb{N}} \colon \mathbb{N} \to [0, \infty)$ which satisfies

$$\liminf_{n \to \infty} \tau_n = \infty \qquad \text{and} \qquad \limsup_{n \to \infty} \|\mathcal{G}(\Theta_{\tau_n})\| = 0. \tag{7.63}$$

Moreover, observe that (7.62) assures that $\limsup_{t \to \infty} \|\Theta_t - \vartheta\| = 0$. Combining this with (7.63) shows that

$$\limsup_{n \to \infty} \|\mathcal{G}(\Theta_{\tau_n})\| = \limsup_{n \to \infty} \|\Theta_{\tau_n} - \vartheta\| = 0. \tag{7.64}$$

In addition, note that Proposition 3.3 assures that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that $\mathcal{G}(\theta) \in (\mathbb{D}\mathcal{L}_\infty)(\theta)$. Therefore, we obtain for all $n \in \mathbb{N}$ that

$$\mathcal{G}(\Theta_{\tau_n}) \in (\mathbb{D}\mathcal{L}_\infty)(\Theta_{\tau_n}). \tag{7.65}$$

Combining this and (7.64) with Lemma 3.4 demonstrates that $0 \in (\mathbb{D}\mathcal{L}_\infty)(\vartheta)$. Combining this with (7.62) establishes (7.59). The proof of Theorem 7.1 is thus complete. ☐

8 Convergence analysis for GD processes

In this section we establish in Proposition 8.1 below an abstract local convergence result for GD under a Kurdyka-Łojasiewicz assumption. In the scientific literature related abstract convergence results for GD type processes under a Łojasiewicz assumption can be found, e.g., in Absil et al. [1], Attouch & Bolte [3], and Dereich & Kassing [21]. Similar arguments have recently been employed in the analysis of optimization algorithms for tensor decomposition [70], deep neural networks [20], and residual neural networks [71]. The latter two works consider the empirical risk, which is measured with respect to a finite set of training data, while we focus on the true risk defined as the expectation over the entire input distribution.

Except for [70] the above mentioned works prove convergence of GD type processes to a critical point, but do not show explicit convergence rates. On the other hand, the authors of [70] consider block coordinate descent under the assumption that the objective function is convex with respect to each block. This property is in general not satisfied for objective functions that arise in the training of DNNs with ReLU activation. The novel contribution of Proposition 8.1 is to establish a precise convergence rate with explicit constants and without such convexity assumptions.

To prove Proposition 8.1 we transfer the ideas from the continuous-time setting in Section 7 to the discrete-time setting. In addition, we require the descent statement in Lemma 8.1 below. Lemma 8.1 below is well-known, see, e.g., Lei et al. [50, Lemma 1], Attouch et al. [4, Lemma 3.1], or Karimi et al. [45]. The elementary proof is only included for completeness.

In Corollaries 8.3 and 8.4 below we establish two simplified versions of Proposition 8.1, and as a consequence we obtain in Corollary 8.5 below the convergence of GD with random initializations in an abstract setting. Afterwards, in Proposition 8.3 below we derive the convergence of GD with random initializations in the considered deep ANN framework in Setting 3.1 under the assumption that there exists a global minimum of the risk function around which suitable regularity assumptions are satisfied. Our proof of Proposition 8.3 also uses the ANN approximation result in Proposition 8.2 below which, in turn, relies on the universal approximation theorem; cf., e.g., Leshno et al. [51], Cybenko [19], Hornik [35], Lu et al. [54], and Shen et al. [66]. As a consequence of Proposition 8.3 we obtain Theorem 8.1 below and, thereby, prove Theorem 1.4 from the introduction. Finally, in Proposition 8.4 below we combine Proposition 8.3 with the existence result for global minima in Corollary 2.6 to establish the convergence of GD with random initializations in the case of shallow ANNs. As a consequence of Proposition 8.4 we derive Corollary 8.6 below and, thereby, prove Theorem 1.2 from the introduction.

8.1 One-step descent property for GD processes

Lemma 8.1. *Let $\mathfrak{d} \in \mathbb{N}$, $L \in \mathbb{R}$, let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ be open and convex, let $f \in C^1(U, \mathbb{R})$, and assume for all $x, y \in U$ that $\|(\nabla f)(x) - (\nabla f)(y)\| \leq L\|x - y\|$. Then it holds for all $x, y \in U$ that*

$$f(y) \leq f(x) + \langle (\nabla f)(x), y - x \rangle + \tfrac{L}{2}\|x - y\|^2. \tag{8.1}$$

Proof of Lemma 8.1. Observe that the fundamental theorem of calculus, the Cauchy-Schwarz inequality, and the assumption that for all $x, y \in U$ it holds that $\|(\nabla f)(x) - (\nabla f)(y)\| \leq L\|x - y\|$ assure that for all $x, y \in U$ we have that

$$\begin{aligned} f(y) - f(x) &= \big[f(x + r(y - x)) \big]_{r=0}^{r=1} \\ &= \int_0^1 \langle (\nabla f)(x + r(y - x)), y - x \rangle \, \mathrm{d}r \\ &= \langle (\nabla f)(x), y - x) \rangle + \int_0^1 \langle (\nabla f)(x + r(y - x)) - (\nabla f)(x), y - x \rangle \, \mathrm{d}r \end{aligned}$$

$$
\begin{aligned}
&\le \langle (\nabla f)(x), y-x \rangle + \int_0^1 |\langle (\nabla f)(x+r(y-x)) - (\nabla f)(x), y-x\rangle| \,\mathrm{d}r \\
&\le \langle (\nabla f)(x), y-x \rangle + \left[\int_0^1 \|(\nabla f)(x+r(y-x)) - (\nabla f)(x)\| \,\mathrm{d}r\right] \|y-x\| \\
&\le \langle (\nabla f)(x), y-x \rangle + L\|y-x\| \left[\int_0^1 \|r(y-x)\| \,\mathrm{d}r\right] \\
&= \langle (\nabla f)(x), y-x \rangle + \tfrac{L}{2}\|x-y\|^2.
\end{aligned} \tag{8.2}
$$

The proof of Lemma 8.1 is thus complete. □

Corollary 8.1. *Let $\mathfrak{d} \in \mathbb{N}$, $L, \gamma \in \mathbb{R}$, let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ be open and convex, let $f \in C^1(U, \mathbb{R})$, and assume for all $x, y \in U$ that $\|(\nabla f)(x) - (\nabla f)(y)\| \le L\|x-y\|$. Then it holds for all $x \in U$ with $x - \gamma(\nabla f)(x) \in U$ that*

$$
f(x - \gamma(\nabla f)(x)) \le f(x) + \gamma\big(\tfrac{L\gamma}{2} - 1\big)\|(\nabla f)(x)\|^2. \tag{8.3}
$$

Proof of Corollary 8.1. Note that Lemma 8.1 ensures for all $x \in U$ with $x - \gamma(\nabla f)(x) \in U$ that

$$
\begin{aligned}
f(x - \gamma(\nabla f)(x)) &\le f(x) + \langle (\nabla f)(x), -\gamma(\nabla f)(x)\rangle + \tfrac{L}{2}\|\gamma(\nabla f)(x)\|^2 \\
&= f(x) - \gamma\|(\nabla f)(x)\|^2 + \tfrac{L\gamma^2}{2}\|(\nabla f)(x)\|^2.
\end{aligned} \tag{8.4}
$$

This establishes (8.3). The proof of Corollary 8.1 is thus complete. □

Corollary 8.2. *Let $\mathfrak{d} \in \mathbb{N}$, $L \in (0, \infty)$, $\gamma \in [0, L^{-1}]$, let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ be open and convex, let $f \in C^1(U, \mathbb{R})$, and assume for all $x, y \in U$ that $\|(\nabla f)(x) - (\nabla f)(y)\| \le L\|x-y\|$. Then it holds for all $x \in U$ with $x - \gamma(\nabla f)(x) \in U$ that*

$$
f(x - \gamma(\nabla f)(x)) \le f(x) - \tfrac{\gamma}{2}\|(\nabla f)(x)\|^2 \le f(x). \tag{8.5}
$$

Proof of Corollary 8.2. Observe that Corollary 8.1, the fact that $\gamma \ge 0$, and the fact that $\frac{L\gamma}{2} - 1 \le -\frac{1}{2}$ establish (8.5). The proof of Corollary 8.2 is thus complete. □

8.2 Abstract local convergence results for GD processes

Proposition 8.1. *Let $\mathfrak{d} \in \mathbb{N}$, $\mathfrak{c} \in \mathbb{R}$, $\varepsilon, L, \mathfrak{C} \in (0, \infty)$, $\alpha \in (0,1)$, $\gamma \in (0, L^{-1}]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, let $\mathbb{B} \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy $\mathbb{B} = \{\theta \in \mathbb{R}^{\mathfrak{d}} \colon \|\theta - \vartheta\| < \varepsilon\}$, let $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy $\mathcal{L}|_{\mathbb{B}} \in C^1(\mathbb{B}, \mathbb{R})$, let $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{B}$ that $\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta)$, assume $\mathcal{G}(\vartheta) = 0$, assume for all $\theta_1, \theta_2 \in \mathbb{B}$ that $\|\mathcal{G}(\theta_1) - \mathcal{G}(\theta_2)\| \le L\|\theta_1 - \theta_2\|$, let $\Theta \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}_0$ that $\Theta_{n+1} = \Theta_n - \gamma\mathcal{G}(\Theta_n)$, and assume for all $\theta \in \mathbb{B}$ that*

$$
|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{\alpha} \le \mathfrak{C}\|\mathcal{G}(\theta)\|, \quad \mathfrak{c} = |\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|, \tag{8.6a}
$$

$$
2\mathfrak{C}(1-\alpha)^{-1}\mathfrak{c}^{1-\alpha} + \|\Theta_0 - \vartheta\| < \tfrac{\varepsilon}{\gamma L + 1}, \tag{8.6b}
$$

and $\inf_{n \in \{m \in \mathbb{N}_0 \colon \forall k \in \mathbb{N}_0 \cap [0,m] \colon \Theta_k \in \mathbb{B}\}} \mathcal{L}(\Theta_n) \ge \mathcal{L}(\vartheta)$. Then there exists $\psi \in \mathcal{L}^{-1}(\{\mathcal{L}(\vartheta)\}) \cap \mathcal{G}^{-1}(\{0\})$ such that *(i) it holds for all $n \in \mathbb{N}_0$ that $\Theta_n \in \mathbb{B}$,*

(ii) it holds for all $n \in \mathbb{N}_0$ that $0 \le \mathcal{L}(\Theta_n) - \mathcal{L}(\psi) \le 2\mathfrak{C}^2\mathfrak{c}^2(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{C}^2\mathfrak{c})^{-1}$, and

(iii) it holds for all $n \in \mathbb{N}_0$ that

$$\begin{aligned}\|\Theta_n - \psi\| \le \textstyle\sum_{k=n}^{\infty}\|\Theta_{k+1} - \Theta_k\| &\le 2\mathfrak{C}(1-\alpha)^{-1}|\mathcal{L}(\Theta_n) - \mathcal{L}(\psi)|^{1-\alpha} \\ &\le 2^{2-\alpha}\mathfrak{C}^{3-2\alpha}\mathfrak{c}^{2-2\alpha}(1-\alpha)^{-1}(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{C}^2\mathfrak{c})^{\alpha-1}. \end{aligned} \tag{8.7}$$

Observe that the assumption that $\inf_{n\in\{m\in\mathbb{N}_0 \colon \forall\, k\in\mathbb{N}_0\cap[0,m]\colon \Theta_k\in\mathbb{B}\}} \mathcal{L}(\Theta_n) \ge \mathcal{L}(\vartheta)$ is in particular satisfied if ϑ is a local minimum of $\mathcal{L}$ with $\forall\, \theta \in \mathbb{B}\colon \mathcal{L}(\theta) \ge \mathcal{L}(\vartheta)$. Hence Proposition 8.1 implies as a consequence a local convergence result of GD to a local minimum. But our assumption also covers more general cases, since we only require an estimate on the values of $\mathcal{L}(\Theta_n)$ and not for all values $\mathcal{L}(\theta)$ with $\theta \in \mathbb{B}$.

Proof of Proposition 8.1. Throughout this proof let $T \in \mathbb{N}_0 \cup \{\infty\}$ satisfy

$$T = \inf(\{n \in \mathbb{N}_0 \colon \Theta_n \notin \mathbb{B}\} \cup \{\infty\}), \tag{8.8}$$

let $\mathbb{L}\colon \mathbb{N}_0 \to \mathbb{R}$ satisfy for all $n \in \mathbb{N}_0$ that $\mathbb{L}(n) = \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta)$, and let $\tau \in \mathbb{N}_0 \cup \{\infty\}$ satisfy

$$\tau = \inf(\{n \in \mathbb{N}_0 \cap [0, T) \colon \mathbb{L}(n) = 0\} \cup \{T\}). \tag{8.9}$$

In the first step of our proof we verify that $T = \infty$, i.e., that the Gd iterates remain inside the neighborhood $\mathbb{B}$ at all times. Note that the assumption that $\mathcal{G}(\vartheta) = 0$ implies for all $\theta \in \mathbb{B}$ that

$$\gamma\|\mathcal{G}(\theta)\| = \gamma\|\mathcal{G}(\theta) - \mathcal{G}(\vartheta)\| \le \gamma L\|\theta - \vartheta\|. \tag{8.10}$$

This, the fact that $\|\Theta_0 - \vartheta\| < \varepsilon$, and the fact that

$$\begin{aligned}\|\Theta_1 - \vartheta\| &\le \|\Theta_1 - \Theta_0\| + \|\Theta_0 - \vartheta\| \\ &= \gamma\|\mathcal{G}(\Theta_0)\| + \|\Theta_0 - \vartheta\| \\ &\le (\gamma L + 1)\|\Theta_0 - \vartheta\| < \varepsilon \end{aligned} \tag{8.11}$$

ensure that $T \ge 2$. Next observe that the assumption that

$$\inf_{n\in\{m\in\mathbb{N}_0 \colon \forall\, k\in\mathbb{N}_0\cap[0,m]\colon \Theta_k\in\mathbb{B}\}} \mathcal{L}(\Theta_n) \ge \mathcal{L}(\vartheta) \tag{8.12}$$

proves for all $n \in \mathbb{N}_0 \cap [0, T)$ that $\mathbb{L}(n) \ge 0$. In addition, note that the fact that $\mathbb{B} \subseteq \mathbb{R}^{\mathfrak{d}}$ is open and convex, Corollary 8.2, and (8.6) demonstrate for all $n \in \mathbb{N}_0 \cap [0, T-1)$ that

$$\begin{aligned}\mathbb{L}(n+1) - \mathbb{L}(n) = \mathcal{L}(\Theta_{n+1}) - \mathcal{L}(\Theta_n) &\le -\tfrac{\gamma}{2}\|\mathcal{G}(\Theta_n)\|^2 \\ &= -\tfrac{1}{2}\|\mathcal{G}(\Theta_n)\|\|\gamma\mathcal{G}(\Theta_n)\| \\ &= -\tfrac{1}{2}\|\mathcal{G}(\Theta_n)\|\|\Theta_{n+1} - \Theta_n\| \\ &\le -(2\mathfrak{C})^{-1}|\mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta)|^{\alpha}\|\Theta_{n+1} - \Theta_n\| \\ &= -(2\mathfrak{C})^{-1}[\mathbb{L}(n)]^{\alpha}\|\Theta_{n+1} - \Theta_n\| \le 0. \end{aligned} \tag{8.13}$$

Therefore, we obtain that $\mathbb{N}_0 \cap [0,T) \ni n \mapsto \mathbb{L}(n) \in [0,\infty)$ is non-increasing. Combining this with (8.9) shows for all $n \in \mathbb{N}_0 \cap [\tau,T)$ that $\mathbb{L}(n) = 0$. This and (8.13) demonstrate for all $n \in \mathbb{N}_0 \cap [\tau, T-1)$ that $0 = \mathbb{L}(n+1) - \mathbb{L}(n) \leq -\frac{\gamma}{2}\|\mathcal{G}(\Theta_n)\|^2 \leq 0$. The fact that $\gamma > 0$ therefore assures for all $n \in \mathbb{N}_0 \cap [\tau, T-1)$ that $\mathcal{G}(\Theta_n) = 0$. Hence, we obtain for all $n \in \mathbb{N}_0 \cap [\tau, T)$ that

$$\Theta_n = \Theta_\tau. \tag{8.14}$$

In addition, observe that (8.9) and (8.13) ensure for all $n \in \mathbb{N}_0 \cap [0,\tau) \cap [0, T-1)$ that

$$\begin{aligned} \|\Theta_{n+1} - \Theta_n\| &\leq \frac{2\mathfrak{C}(\mathbb{L}(n) - \mathbb{L}(n+1))}{[\mathbb{L}(n)]^\alpha} = 2\mathfrak{C}\int_{\mathbb{L}(n+1)}^{\mathbb{L}(n)} [\mathbb{L}(n)]^{-\alpha}\, \mathrm{d}u \\ &\leq 2\mathfrak{C}\int_{\mathbb{L}(n+1)}^{\mathbb{L}(n)} u^{-\alpha}\, \mathrm{d}u = \frac{2\mathfrak{C}\big([\mathbb{L}(n)]^{1-\alpha} - [\mathbb{L}(n+1)]^{1-\alpha}\big)}{1-\alpha}. \end{aligned} \tag{8.15}$$

This and (8.14) show for all $n \in \mathbb{N}_0 \cap [0, T-1)$ that

$$\|\Theta_{n+1} - \Theta_n\| \leq \frac{2\mathfrak{C}\big([\mathbb{L}(n)]^{1-\alpha} - [\mathbb{L}(n+1)]^{1-\alpha}\big)}{1-\alpha}. \tag{8.16}$$

Combining this with the triangle inequality proves for all $m, n \in \mathbb{N}_0 \cap [0,T)$ with $m \leq n$ that

$$\begin{aligned} \|\Theta_n - \Theta_m\| &\leq \sum_{k=m}^{n-1} \|\Theta_{k+1} - \Theta_k\| \\ &\leq \frac{2\mathfrak{C}}{1-\alpha}\left[\sum_{k=m}^{n-1}\Big([\mathbb{L}(k)]^{1-\alpha} - [\mathbb{L}(k+1)]^{1-\alpha}\Big)\right] \\ &= \frac{2\mathfrak{C}\big([\mathbb{L}(m)]^{1-\alpha} - [\mathbb{L}(n)]^{1-\alpha}\big)}{1-\alpha} \leq \frac{2\mathfrak{C}[\mathbb{L}(m)]^{1-\alpha}}{1-\alpha}. \end{aligned} \tag{8.17}$$

This and (8.6) demonstrate for all $n \in \mathbb{N}_0 \cap [0,T)$ that

$$\|\Theta_n - \Theta_0\| \leq \frac{2\mathfrak{C}[\mathbb{L}(0)]^{1-\alpha}}{1-\alpha} = \frac{2\mathfrak{C}|\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|^{1-\alpha}}{1-\alpha} = 2\mathfrak{C}(1-\alpha)^{-1}\mathfrak{c}^{1-\alpha}. \tag{8.18}$$

Combining this with (8.10), (8.6), and the triangle inequality shows for all $n \in \mathbb{N}_0 \cap [0,T)$ that

$$\begin{aligned} \|\Theta_{n+1} - \vartheta\| &\leq \|\Theta_{n+1} - \Theta_n\| + \|\Theta_n - \vartheta\| = \gamma\|\mathcal{G}(\Theta_n)\| + \|\Theta_n - \vartheta\| \\ &\leq (\gamma L + 1)\|\Theta_n - \vartheta\| \leq (\gamma L + 1)(\|\Theta_n - \Theta_0\| + \|\Theta_0 - \vartheta\|) \\ &\leq (\gamma L + 1)(2\mathfrak{C}(1-\alpha)^{-1}\mathfrak{c}^{1-\alpha} + \|\Theta_0 - \vartheta\|) < \varepsilon. \end{aligned} \tag{8.19}$$

Hence, we obtain that

$$T = \infty. \tag{8.20}$$

Combining this with (8.6) and (8.17) proves that

$$\sum_{k=0}^{\infty}\|\Theta_{k+1}-\Theta_k\| = \lim_{n\to\infty}\left[\sum_{k=0}^{n}\|\Theta_{k+1}-\Theta_k\|\right] \leq \frac{2\mathfrak{C}[\mathbb{L}(0)]^{1-\alpha}}{1-\alpha} = \frac{2\mathfrak{C}\mathfrak{c}^{1-\alpha}}{1-\alpha} < \varepsilon < \infty. \tag{8.21}$$

Therefore, we obtain that there exists $\psi \in \mathbb{R}^{\mathfrak{d}}$ which satisfies

$$\limsup_{n\to\infty}\|\Theta_n - \psi\| = 0. \tag{8.22}$$

This establishes convergence of the GD process. We next deduce explicit convergence rates. Note that (8.19), (8.20), and (8.22) imply that $\|\psi - \vartheta\| \leq (\gamma L+1)(2\mathfrak{C}(1-\alpha)^{-1}\mathfrak{c}^{1-\alpha} + \|\Theta_0 - \vartheta\|) < \varepsilon$. Therefore, we obtain that $\psi \in \mathbb{B}$. Next observe that (8.13), (8.6), and the fact that for all $n \in \mathbb{N}_0$ it holds that $\mathbb{L}(n) \leq \mathbb{L}(0) = \mathfrak{c}$ ensure that for all $n \in \mathbb{N}_0 \cap [0,\tau)$ we have that

$$-\mathbb{L}(n) \leq \mathbb{L}(n+1) - \mathbb{L}(n) \leq -\tfrac{\gamma}{2}\|\mathcal{G}(\Theta_n)\|^2 \leq -\tfrac{\gamma}{2\mathfrak{C}^2}[\mathbb{L}(n)]^{2\alpha} \leq -\tfrac{\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}[\mathbb{L}(n)]^2. \tag{8.23}$$

This assures for all $n \in \mathbb{N}_0 \cap [0,\tau)$ that $0 < \mathbb{L}(n) \leq \frac{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}{\gamma}$. Combining this and (8.23) demonstrates for all $n \in \mathbb{N}_0 \cap [0,\tau-1)$ that

$$\begin{aligned}\frac{1}{\mathbb{L}(n)} - \frac{1}{\mathbb{L}(n+1)} &\leq \frac{1}{\mathbb{L}(n)} - \frac{1}{\mathbb{L}(n)(1-\frac{\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}\mathbb{L}(n))}\\ &= \frac{\big(1-\frac{\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}\mathbb{L}(n)\big)-1}{\mathbb{L}(n)\big(1-\frac{\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}\mathbb{L}(n)\big)} = \frac{-\frac{\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}}{\big(1-\frac{\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}\mathbb{L}(n)\big)}\\ &= -\frac{1}{\big(\frac{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}{\gamma} - \mathbb{L}(n)\big)} < -\frac{\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}.\end{aligned} \tag{8.24}$$

Therefore, we get for all $n \in \mathbb{N}_0 \cap [0,\tau)$ that

$$\begin{aligned}\frac{1}{\mathbb{L}(n)} &= \frac{1}{\mathbb{L}(0)} + \sum_{k=0}^{n-1}\left[\frac{1}{\mathbb{L}(k+1)} - \frac{1}{\mathbb{L}(k)}\right]\\ &\geq \frac{1}{\mathbb{L}(0)} + \frac{n\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}} = \frac{1}{\mathfrak{c}} + \frac{n\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}.\end{aligned} \tag{8.25}$$

Hence, we obtain for all $n \in \mathbb{N}_0 \cap [0,\tau)$ that $\mathbb{L}(n) \leq \frac{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}{n\gamma+2\mathfrak{C}^2\mathfrak{c}^{1-2\alpha}}$. Combining this with the fact that for all $n \in \mathbb{N}_0 \cap [\tau,\infty)$ it holds that $\mathbb{L}(n) = 0$ shows that for all $n \in \mathbb{N}_0$ we have that

$$\mathbb{L}(n) \leq \frac{2\mathfrak{C}^2\mathfrak{c}^2}{\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{C}^2\mathfrak{c}}. \tag{8.26}$$

This, (8.22), and the assumption that $\mathcal{L}$ is continuous prove that

$$\mathcal{L}(\psi) = \lim_{n\to\infty} \mathcal{L}(\Theta_n) = \mathcal{L}(\vartheta). \tag{8.27}$$

Combining this with (8.26) assures for all $n \in \mathbb{N}_0$ that

$$0 \le \mathcal{L}(\Theta_n) - \mathcal{L}(\psi) \le \frac{2\mathfrak{C}^2\mathfrak{c}^2}{\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{C}^2\mathfrak{c}}. \tag{8.28}$$

Furthermore, note that the fact that $\mathbb{B} \ni \theta \mapsto \mathcal{G}(\theta) \in \mathbb{R}^{\mathfrak{d}}$ is continuous, the fact that $\psi \in \mathbb{B}$, and (8.22) imply that

$$\mathcal{G}(\psi) = \lim_{n\to\infty} \mathcal{G}(\Theta_n) = \lim_{n\to\infty}(\gamma^{-1}(\Theta_n - \Theta_{n+1})) = 0. \tag{8.29}$$

Next observe that (8.26) and (8.17) ensure for all $n \in \mathbb{N}_0$ that

$$\begin{aligned} \|\Theta_n - \psi\| = \lim_{m\to\infty} \|\Theta_n - \Theta_m\| \le \sum_{k=n}^{\infty} \|\Theta_{k+1} - \Theta_k\| \le \frac{2\mathfrak{C}[\mathbb{L}(n)]^{1-\alpha}}{1-\alpha} \\ \le \frac{2^{2-\alpha}\mathfrak{C}^{3-2\alpha}\mathfrak{c}^{2-2\alpha}}{(1-\alpha)(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{C}^2\mathfrak{c})^{1-\alpha}}. \end{aligned} \tag{8.30}$$

Combining this with (8.27), (8.20), (8.29), and (8.28) establishes items (i), (ii), and (iii). The proof of Proposition 8.1 is thus complete. □

The next result, Corollary 8.3, specializes Proposition 8.1 to the case where $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ is a local minimum of $\mathcal{L}$ in the sense that for all $\theta \in \mathbb{B}$ we have that $\mathcal{L}(\theta) \ge \mathcal{L}(\vartheta)$, where $\mathbb{B}$ is a suitable neighborhood of ϑ.

Corollary 8.3. *Let $\mathfrak{d} \in \mathbb{N}$, $\mathfrak{c} \in [0,1]$, $\varepsilon, L, \mathfrak{C} \in (0,\infty)$, $\alpha \in (0,1)$, $\gamma \in (0, L^{-1}]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, let $\mathbb{B} \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy $\mathbb{B} = \{\theta \in \mathbb{R}^{\mathfrak{d}} \colon \|\theta - \vartheta\| < \varepsilon\}$, let $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy $\mathcal{L}|_{\mathbb{B}} \in C^1(\mathbb{B}, \mathbb{R})$, let $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{B}$ that $\mathcal{G}(\theta) = (\nabla\mathcal{L})(\theta)$, assume for all $\theta_1, \theta_2 \in \mathbb{B}$ that $\|\mathcal{G}(\theta_1) - \mathcal{G}(\theta_2)\| \le L\|\theta_1 - \theta_2\|$, let $\Theta = (\Theta_n)_{n\in\mathbb{N}_0} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}_0$ that $\Theta_{n+1} = \Theta_n - \gamma\mathcal{G}(\Theta_n)$, and assume for all $\theta \in \mathbb{B}$ that*

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{\alpha} \le \mathfrak{C}\|\mathcal{G}(\theta)\|, \quad \mathfrak{c} = |\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|, \tag{8.31a}$$

$$2\mathfrak{C}(1-\alpha)^{-1}\mathfrak{c}^{1-\alpha} + \|\Theta_0 - \vartheta\| < \tfrac{\varepsilon}{\gamma L+1}, \tag{8.31b}$$

and $\mathcal{L}(\theta) \ge \mathcal{L}(\vartheta)$. Then there exists $\psi \in \mathcal{L}^{-1}(\{\mathcal{L}(\vartheta)\}) \cap \mathcal{G}^{-1}(\{0\})$ such that for all $n \in \mathbb{N}_0$ it holds that $\Theta_n \in \mathbb{B}$, $0 \le \mathcal{L}(\Theta_n) - \mathcal{L}(\psi) \le 2(2 + \mathfrak{C}^{-2}\gamma n)^{-1}$, and

$$\|\Theta_n - \psi\| \le \textstyle\sum_{k=n}^{\infty} \|\Theta_{k+1} - \Theta_k\| \le 2^{2-\alpha}\mathfrak{C}(1-\alpha)^{-1}(2 + \mathfrak{C}^{-2}\gamma n)^{\alpha-1}. \tag{8.32}$$

Proof of Corollary 8.3. Note that the fact that $\mathcal{L}(\vartheta) = \inf_{\theta\in\mathbb{B}} \mathcal{L}(\theta)$ ensures that $\mathcal{G}(\vartheta) = (\nabla\mathcal{L})(\vartheta) = 0$ and $\inf_{n\in\{m\in\mathbb{N}_0 \colon \forall k\in\mathbb{N}_0\cap[0,m] \colon \Theta_k\in\mathbb{B}\}} \mathcal{L}(\Theta_n) \ge \mathcal{L}(\vartheta)$. Combining this with Proposition 8.1 implies that there exists $\psi \in \mathcal{L}^{-1}(\{\mathcal{L}(\vartheta)\}) \cap \mathcal{G}^{-1}(\{0\})$ such that

(I) it holds for all $n \in \mathbb{N}_0$ that $\Theta_n \in \mathbb{B}$,

(II) it holds for all $n \in \mathbb{N}_0$ that $0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\psi) \leq \frac{2\mathfrak{C}^2\mathfrak{c}^2}{\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{C}^2\mathfrak{c}}$, and

(III) it holds for all $n \in \mathbb{N}_0$ that

$$\begin{aligned} \|\Theta_n - \psi\| \leq \sum_{k=n}^{\infty} \|\Theta_{k+1} - \Theta_k\| &\leq \frac{2\mathfrak{C}|\mathcal{L}(\Theta_n) - \mathcal{L}(\psi)|^{1-\alpha}}{1-\alpha} \\ &\leq \frac{2^{2-\alpha}\mathfrak{C}^{3-2\alpha}\mathfrak{c}^{2-2\alpha}}{(1-\alpha)(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{C}^2\mathfrak{c})^{1-\alpha}}. \end{aligned} \tag{8.33}$$

Observe that item (II) and the assumption that $\mathfrak{c} \leq 1$ show for all $n \in \mathbb{N}_0$ that

$$\begin{aligned} 0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\psi) &\leq 2\mathfrak{c}^2\Big(\mathfrak{C}^{-2}\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^{-2}\mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{c}\Big)^{-1} \\ &\leq 2(2 + \mathfrak{C}^{-2}\gamma n)^{-1}. \end{aligned} \tag{8.34}$$

This and item (III) demonstrate for all $n \in \mathbb{N}_0$ that

$$\begin{aligned} \|\Theta_n - \psi\| \leq \sum_{k=n}^{\infty} \|\Theta_{k+1} - \Theta_k\| &\leq \frac{2\mathfrak{C}|\mathcal{L}(\Theta_n) - \mathcal{L}(\psi)|^{1-\alpha}}{1-\alpha} \\ &\leq \left[\frac{2^{2-\alpha}\mathfrak{C}}{1-\alpha}\right](2 + \mathfrak{C}^{-2}\gamma n)^{\alpha-1}. \end{aligned} \tag{8.35}$$

The proof of Corollary 8.3 is thus complete. □

8.3 Abstract global convergence results for GD processes

In Corollary 8.4 we reformulate Corollary 8.3 to show that around every local minimum point which admits a Kurdyka-Łojasiewicz inequality and a certain regularity condition there exists an open neighborhood such that the risk of every GD sequence started in this neighborhood converges with rate 1 to the risk of the local minimum.

Corollary 8.4. *Let $\mathfrak{d} \in \mathbb{N}$, $\varepsilon, L, \mathfrak{C} \in (0,\infty)$, $\alpha \in (0,1)$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, let $\mathbb{B} \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy $\mathbb{B} = \{\theta \in \mathbb{R}^{\mathfrak{d}} \colon \|\theta - \vartheta\| < \varepsilon\}$, let $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy $\mathcal{L}|_{\mathbb{B}} \in C^1(\mathbb{B}, \mathbb{R})$, let $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{B}$ that $\mathcal{G}(\theta) = (\nabla\mathcal{L})(\theta)$, assume for all $\theta_1, \theta_2 \in \mathbb{B}$ that $\|\mathcal{G}(\theta_1) - \mathcal{G}(\theta_2)\| \leq L\|\theta_1 - \theta_2\|$, for every $\theta \in \mathbb{R}^{\mathfrak{d}}$, $\gamma \in \mathbb{R}$ let $\Theta^{\gamma,\theta} = (\Theta^{\gamma,\theta}_n)_{n\in\mathbb{N}_0} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}_0$ that $\Theta^{\gamma,\theta}_0 = \theta$ and $\Theta^{\gamma,\theta}_{n+1} = \Theta^{\gamma,\theta}_n - \gamma\mathcal{G}(\Theta^{\gamma,\theta}_n)$, and assume for all $\theta \in \mathbb{B}$ that*

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{\alpha} \leq \mathfrak{C}\|\mathcal{G}(\theta)\| \qquad \textit{and} \qquad \mathcal{L}(\theta) \geq \mathcal{L}(\vartheta). \tag{8.36}$$

Then there exist $\delta, \mathscr{C} \in (0,\infty)$ such that for all $\theta \in \{\psi \in \mathbb{R}^{\mathfrak{d}} \colon \|\psi - \vartheta\| < \delta\}$, $\gamma \in (0, L^{-1}]$, $n \in \mathbb{N}_0$ it holds that $0 \leq \mathcal{L}(\Theta^{\gamma,\theta}_n) - \mathcal{L}(\vartheta) \leq \mathscr{C}(1 + \gamma n)^{-1}$.

Proof of Corollary 8.4. Note that the fact that $\mathcal{L}$ is continuous ensures that there exist $\mathfrak{c} \in [0,1]$, $\delta \in (0, \varepsilon)$ which satisfy for all $\theta \in \{\psi \in \mathbb{R}^{\mathfrak{d}} \colon \|\psi - \vartheta\| < \delta\}$, $\gamma \in (0, L^{-1}]$ that

$$\mathfrak{c} = |\mathcal{L}(\theta) - \mathcal{L}(\vartheta)| \qquad \text{and} \qquad 2\mathfrak{C}(1-\alpha)^{-1}\mathfrak{c}^{1-\alpha} + \|\theta - \vartheta\| < \tfrac{\varepsilon}{2} \leq \tfrac{\varepsilon}{\gamma L + 1}. \tag{8.37}$$

Observe that (8.37) and Corollary 8.3 (applied for every $\theta \in \{\psi \in \mathbb{R}^{\mathfrak{d}} \colon \|\psi - \vartheta\| < \delta\}$, $\gamma \in (0, L^{-1}]$ with $\varepsilon \curvearrowleft \delta$, $\gamma \curvearrowleft \gamma$, $\mathbb{B} \curvearrowleft \{\psi \in \mathbb{R}^{\mathfrak{d}} \colon \|\psi - \vartheta\| < \delta\}$, $\Theta \curvearrowleft \Theta^{\gamma,\theta}$ in the notation of Corollary 8.3) demonstrate that for all $\theta \in \{\psi \in \mathbb{R}^{\mathfrak{d}} \colon \|\psi - \vartheta\| < \delta\}$, $\gamma \in (0, L^{-1}]$ there exists $\psi \in \mathcal{L}^{-1}(\{\mathcal{L}(\vartheta)\})$ such that for all $n \in \mathbb{N}_0$ it holds that

$$0 \leq \mathcal{L}(\Theta_n^{\gamma,\theta}) - \mathcal{L}(\psi) \leq 2(2 + \mathfrak{C}^{-2}\gamma n)^{-1}. \tag{8.38}$$

Hence, we obtain for all $\theta \in \{\psi \in \mathbb{R}^{\mathfrak{d}} \colon \|\psi - \vartheta\| < \delta\}$, $\gamma \in (0, L^{-1}]$, $n \in \mathbb{N}_0$ that

$$\begin{aligned} 0 &\leq \mathcal{L}(\Theta_n^{\gamma,\theta}) - \mathcal{L}(\vartheta) \leq 2(2 + \mathfrak{C}^{-2}\gamma_n)^{-1} \\ &\leq 2(\min\{2, \mathfrak{C}^{-2}\}(1 + \gamma n))^{-1} \\ &= \max\{1, 2\mathfrak{C}^2\}(1 + \gamma n)^{-1}. \end{aligned} \tag{8.39}$$

The proof of Corollary 8.4 is thus complete. □

8.4 Abstract convergence result for GD with random initializations

The next result, Corollary 8.5, establishes convergence in probability of the GD method with multiple random initializations under a Łojasiewicz type assumption. The proof relies on Corollary 8.4 and the fact that for a sufficiently high number of initializations at least one of the GD trajectories will start in a suitable open domain of attraction with high probability.

Corollary 8.5. *Let $\mathfrak{d} \in \mathbb{N}$, $\varepsilon, L, \mathfrak{C}, \gamma \in (0, \infty)$, $\alpha \in (0,1)$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ satisfy $\gamma L \leq 1$, let $\mathbb{B} \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy $\mathbb{B} = \{\theta \in \mathbb{R}^{\mathfrak{d}} \colon \|\theta - \vartheta\| < \varepsilon\}$, let $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy $\mathcal{L}|_{\mathbb{B}} \in C^1(\mathbb{B}, \mathbb{R})$, let $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{B}$ that $\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta)$, assume for all $\theta_1, \theta_2 \in \mathbb{B}$ that $\|\mathcal{G}(\theta_1) - \mathcal{G}(\theta_2)\| \leq L\|\theta_1 - \theta_2\|$, assume for all $\theta \in \mathbb{B}$ that*

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{\alpha} \leq \mathfrak{C}\|\mathcal{G}(\theta)\| \qquad \textit{and} \qquad \mathcal{L}(\theta) \geq \mathcal{L}(\vartheta), \tag{8.40}$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $K, n \in \mathbb{N}_0$ let $\Theta_n^K \colon \Omega \to \mathbb{R}^{\mathfrak{d}}$ and $\mathbf{k}_n^K \colon \Omega \to \mathbb{N}$ be random variables, assume that Θ_0^K, $K \in \mathbb{N}$, are i.i.d., assume for all $\delta \in (0, \infty)$ that $\mathbb{P}(\|\Theta_0^1 - \vartheta\| < \delta) > 0$, and assume for all $K \in \mathbb{N}$, $n \in \mathbb{N}_0$, $\omega \in \Omega$ that

$$\Theta_{n+1}^K(\omega) = \Theta_n^K(\omega) - \gamma \mathcal{G}(\Theta_n^K(\omega)) \qquad \textit{and} \qquad \mathbf{k}_n^K(\omega) \in \arg\min\nolimits_{\kappa \in \{1,2,\dots,K\}} \mathcal{L}(\Theta_n^{\kappa}(\omega)). \tag{8.41}$$

Then

$$\liminf\nolimits_{K \to \infty} \mathbb{P}\big(\limsup\nolimits_{n \to \infty} \mathcal{L}(\Theta_n^{\mathbf{k}_n^K}) \leq \mathcal{L}(\vartheta)\big) = 1. \tag{8.42}$$

Proof of Corollary 8.5. Note that (8.41) shows for all $K \in \mathbb{N}$ that

$$\begin{aligned} &\mathbb{P}\big(\limsup_{n\to\infty} \mathcal{L}\big(\Theta_n^{\mathbf{k}_n^K}\big) \leq \mathcal{L}(\vartheta)\big) \\ &\geq \mathbb{P}(\exists\, \kappa \in \{1,2,\ldots,K\} \colon \limsup_{n\to\infty} \mathcal{L}(\Theta_n^\kappa) \leq \mathcal{L}(\vartheta)). \end{aligned} \tag{8.43}$$

Furthermore, observe that Corollary 8.4 demonstrates that there exist $\delta, \mathscr{C} \in (0,\infty)$ which satisfy for all $\kappa \in \mathbb{N}$, $\omega \in \{w \in \Omega \colon \|\Theta_0^\kappa(w) - \vartheta\| < \delta\}$, $n \in \mathbb{N}_0$ that $0 \leq \mathcal{L}(\Theta_n^\kappa(\omega)) - \mathcal{L}(\vartheta) \leq \mathscr{C}(1+\gamma n)^{-1}$. Therefore, we obtain for all $\kappa \in \mathbb{N}$, $\omega \in \{w \in \Omega \colon \|\Theta_0^\kappa(w) - \vartheta\| < \delta\}$ that

$$\limsup_{n\to\infty} \mathcal{L}(\Theta_n^\kappa(\omega)) \leq \mathcal{L}(\vartheta). \tag{8.44}$$

This shows for all $\kappa \in \mathbb{N}$ that

$$\{\omega \in \Omega \colon \|\Theta_0^\kappa(\omega) - \vartheta\| < \delta\} \subseteq \Big\{\omega \in \Omega \colon \limsup_{n\to\infty} \mathcal{L}(\Theta_n^\kappa(\omega)) \leq \mathcal{L}(\vartheta)\Big\}. \tag{8.45}$$

Hence, we obtain for all $K \in \mathbb{N}$ that

$$\begin{aligned} &\mathbb{P}\big(\exists\, \kappa \in \{1,2,\ldots,K\} \colon \|\Theta_0^\kappa - \vartheta\| < \delta\big) = \mathbb{P}\big(\cup_{\kappa=1}^K \{\|\Theta_0^\kappa - \vartheta\| < \delta\}\big) \\ &\leq \mathbb{P}\big(\cup_{\kappa=1}^K \{\limsup_{n\to\infty} \mathcal{L}(\Theta_n^\kappa) \leq \mathcal{L}(\vartheta)\}\big) \\ &= \mathbb{P}\big(\exists\, \kappa \in \{1,2,\ldots,K\} \colon \limsup_{n\to\infty} \mathcal{L}(\Theta_n^\kappa) \leq \mathcal{L}(\vartheta)\big). \end{aligned} \tag{8.46}$$

Moreover, note that the assumption that Θ_0^κ, $\kappa \in \mathbb{N}$, are i.i.d. proves that for all $K \in \mathbb{N}$ we have that

$$\begin{aligned} &\mathbb{P}\big(\exists\, \kappa \in \{1,2,\ldots,K\} \colon \|\Theta_0^\kappa - \vartheta\| < \delta\big) \\ &= 1 - \mathbb{P}\big(\forall\, \kappa \in \{1,2,\ldots,K\} \colon \|\Theta_0^\kappa - \vartheta\| \geq \delta\big) \\ &= 1 - \big[\mathbb{P}\big(\|\Theta_0^1 - \vartheta\| \geq \delta\big)\big]^K. \end{aligned} \tag{8.47}$$

The fact that $\mathbb{P}(\|\Theta_0^1 - \vartheta\| \geq \delta) = 1 - \mathbb{P}(\|\Theta_0^1 - \vartheta\| < \delta) < 1$ therefore implies that

$$\liminf_{K\to\infty} \mathbb{P}\big(\exists\, \kappa \in \{1,2,\ldots,K\} \colon \|\Theta_0^\kappa - \vartheta\| < \delta\big) = 1. \tag{8.48}$$

Combining this with (8.43) and (8.46) establishes (8.42). The proof of Corollary 8.5 is thus complete. ☐

8.5 Approximation results for deep ANNs

We next show an L^2-universal approximation result for shallow ANNs. In Lemma 8.2 the target function is not necessarily continuous and takes values in a multidimensional space $\mathbb{R}^\delta$. We establish Lemma 8.2 by employing the universal approximation theorem for $\mathbb{R}$-valued functions in Leshno et al. [51, Proposition 1 in Section 4]. The proof is only included for completeness.

Lemma 8.2. *Let $d, \delta \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in (a, \infty)$, $\varepsilon \in (0, \infty)$ and let $f = (f_1, \dots, f_\delta) \colon [a,b]^d \to \mathbb{R}^\delta$ and $\mathfrak{p} \colon [a,b]^d \to [0, \infty)$ be bounded and measurable. Then there exist $n \in \mathbb{N}$, $\mathfrak{w}_1, \mathfrak{w}_2, \dots, \mathfrak{w}_n \in \mathbb{R}^{1 \times d}$, $\mathfrak{b}_1, \mathfrak{b}_2, \dots, \mathfrak{b}_n \in \mathbb{R}$, $\mathfrak{v}_1, \mathfrak{v}_2, \dots, \mathfrak{v}_n \in \mathbb{R}^\delta$ such that*

$$\textstyle\int_{[a,b]^d} \| f(x) - \sum_{i=1}^n \mathfrak{v}_i \max\{\mathfrak{w}_i x + \mathfrak{b}_i, 0\} \|^2 \, \mathfrak{p}(x) \, \mathrm{d}x < \varepsilon. \tag{8.49}$$

Proof of Lemma 8.2. Throughout this proof let $\mu \colon \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$ satisfy for all $E \in \mathcal{B}(\mathbb{R}^d)$ that

$$\mu(E) = \textstyle\int_{[a,b]^d \cap E} \mathfrak{p}(x) \, \mathrm{d}x \tag{8.50}$$

and for every $i \in \{1, 2, \dots, \delta\}$ let $F_i \colon \mathbb{R}^d \to \mathbb{R}$ satisfy for all $x \in [a,b]^d$, $y \in \mathbb{R}^d \backslash [a,b]^d$ that

$$F_i(x) = f_i(x) \qquad \text{and} \qquad F_i(y) = 0. \tag{8.51}$$

Observe that (8.50) and the assumption that $\mathfrak{p}$ is bounded and measurable ensure that μ is a finite, absolutely continuous, and compactly supported measure. The assumption that f is bounded hence implies that for all $i \in \{1, 2, \dots, \delta\}$ it holds that $\int_{\mathbb{R}^d} |F_i(x)|^2 \, \mu(\mathrm{d}x) < \infty$. Combining this, the universal approximation theorem (cf. Leshno et al. [51, Proposition 1 in Section 4] (applied with $\sigma \curvearrowleft (\mathbb{R} \ni x \mapsto \max\{x, 0\} \in \mathbb{R})$, $\mu \curvearrowleft \mu$, $p \curvearrowleft 2$ in the notation of [51, Proposition 1 in Section 4])), (8.50), (8.51), and the fact that μ is a finite, absolutely continuous, and compactly supported measure proves that for every $i \in \{1, 2, \dots, \delta\}$ there exist $\mathbf{n}^{(i)} \in \mathbb{N}$, $\mathbf{w}_1^{(i)}, \mathbf{w}_2^{(i)}, \dots, \mathbf{w}_{\mathbf{n}^{(i)}}^{(i)} \in \mathbb{R}^{1 \times d}$, $\mathbf{b}_1^{(i)}, \mathbf{b}_2^{(i)}, \dots, \mathbf{b}_{\mathbf{n}^{(i)}}^{(i)}, \mathbf{v}_1^{(i)}, \mathbf{v}_2^{(i)}, \dots, \mathbf{v}_{\mathbf{n}^{(i)}}^{(i)} \in \mathbb{R}$ which satisfy

$$\begin{aligned} &\textstyle\int_{[a,b]^d} |f_i(x) - \sum_{k=1}^{\mathbf{n}^{(i)}} \mathbf{v}_k^{(i)} \max\{\mathbf{w}_k^{(i)} x + \mathbf{b}_k^{(i)}, 0\}|^2 \, \mathfrak{p}(x) \, \mathrm{d}x \\ &= \textstyle\int_{\mathbb{R}^d} |F_i(x) - \sum_{k=1}^{\mathbf{n}^{(i)}} \mathbf{v}_k^{(i)} \max\{\mathbf{w}_k^{(i)} x + \mathbf{b}_k^{(i)}, 0\}|^2 \, \mu(\mathrm{d}x) < \frac{\varepsilon}{\delta}. \end{aligned} \tag{8.52}$$

In the following let $e_1, e_2, \dots, e_\delta \in \mathbb{R}^\delta$ satisfy $e_1 = (1, 0, 0, \dots, 0)$, $e_2 = (0, 1, 0, \dots, 0)$, $\dots$, $e_\delta = (0, \dots, 0, 1)$, let $n \in \mathbb{N}$ satisfy $n = \sum_{i=1}^\delta \mathbf{n}^{(i)}$, and let $\mathfrak{w}_1, \mathfrak{w}_2, \dots, \mathfrak{w}_n \in \mathbb{R}^{1 \times d}$, $\mathfrak{b}_1, \mathfrak{b}_2, \dots, \mathfrak{b}_n \in \mathbb{R}$, $\mathfrak{v}_1, \mathfrak{v}_2, \dots, \mathfrak{v}_n \in \mathbb{R}^\delta$ satisfy for all $j \in \{1, 2, \dots, \delta\}$, $k \in \{1, 2, \dots, \mathbf{n}^{(j)}\}$ that

$$\mathfrak{w}_{k + \sum_{i=1}^{j-1} \mathbf{n}^{(i)}} = \mathbf{w}_k^{(j)}, \qquad \mathfrak{b}_{k + \sum_{i=1}^{j-1} \mathbf{n}^{(i)}} = \mathbf{b}_k^{(j)}, \qquad \text{and} \qquad \mathfrak{v}_{k + \sum_{i=1}^{j-1} \mathbf{n}^{(i)}} = \mathbf{v}_k^{(j)} e_j. \tag{8.53}$$

Note that (8.53) assures that for all $c_1, c_2, \dots, c_n \in \mathbb{R}$, $x \in [a,b]^d$ it holds that

$$\big\| f(x) - \textstyle\sum_{i=1}^n c_i \mathfrak{v}_i \big\|^2 = \sum_{j=1}^\delta \big| f_j(x) - \sum_{k=1}^{\mathbf{n}^{(j)}} \mathbf{v}_k^{(j)} c_{k + \sum_{i=1}^{j-1} \mathbf{n}^{(i)}} \big|^2. \tag{8.54}$$

Combining this with (8.52) and (8.53) demonstrates that

$$\begin{aligned} &\textstyle\int_{[a,b]^d} \big\| f(x) - \sum_{i=1}^n \big[\max\{\mathfrak{w}_i x + \mathfrak{b}_i, 0\}\big] \mathfrak{v}_i \big\|^2 \mathfrak{p}(x) \, \mathrm{d}x \\ &= \sum_{j=1}^\delta \int_{[a,b]^d} \big| f_j(x) - \textstyle\sum_{k=1}^{\mathbf{n}^{(j)}} \mathbf{v}_k^{(j)} \max\{\mathfrak{w}_{k + \sum_{i=1}^{j-1} \mathbf{n}^{(i)}} x + \mathfrak{b}_{k + \sum_{i=1}^{j-1} \mathbf{n}^{(i)}}, 0\} \big|^2 \mathfrak{p}(x) \, \mathrm{d}x \\ &= \sum_{j=1}^\delta \int_{[a,b]^d} \big| f_j(x) - \textstyle\sum_{k=1}^{\mathbf{n}^{(j)}} \mathbf{v}_k^{(j)} \max\{\mathbf{w}_k^{(j)} x + \mathbf{b}_k^{(j)}, 0\} \big|^2 \mathfrak{p}(x) \, \mathrm{d}x < \delta \big[\frac{\varepsilon}{\delta}\big] = \varepsilon. \end{aligned} \tag{8.55}$$

The proof of Lemma 8.2 is thus complete. ☐

As a consequence of Lemma 8.2 we show in Proposition 8.2 a universal approximation result for deep ANNs as the width increases to infinity.

Proposition 8.2. *Let $d, \delta \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in [a, \infty)$, $(\rho_{\mathfrak{a}})_{\mathfrak{a} \in \mathbb{N}} \subseteq (\mathbb{N} \cap (1, \infty))$, let $\ell^{\mathfrak{a}} = (\ell^{\mathfrak{a}}_0, \ell^{\mathfrak{a}}_1, \ldots, \ell^{\mathfrak{a}}_{\rho_{\mathfrak{a}}}) \in \{d\} \times \mathbb{N}^{\rho_{\mathfrak{a}}-1} \times \{\delta\}$, $\mathfrak{a} \in \mathbb{N}$, satisfy*

$$\liminf_{\mathfrak{a} \to \infty} \min\{\ell^{\mathfrak{a}}_1, \ell^{\mathfrak{a}}_2, \ldots, \ell^{\mathfrak{a}}_{\rho_{\mathfrak{a}}-1}\} = \infty, \tag{8.56}$$

for every $\mathfrak{a} \in \mathbb{N}$ let $\mathfrak{d}_{\mathfrak{a}} = \sum_{k=1}^{\rho_{\mathfrak{a}}} \ell^{\mathfrak{a}}_k(\ell^{\mathfrak{a}}_{k-1} + 1)$, let $f\colon [a, b]^d \to \mathbb{R}^\delta$ and $\mathfrak{p}\colon [a, b]^d \to [0, \infty)$ be bounded and measurable, for every $\mathfrak{a} \in \mathbb{N}$, $k \in \{1, 2, \ldots, \rho_{\mathfrak{a}}\}$, $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}_{\mathfrak{a}}}) \in \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}$ let $\mathfrak{w}^{k,\theta}_{\mathfrak{a}} = (\mathfrak{w}^{k,\theta}_{\mathfrak{a},i,j})_{(i,j) \in \{1,\ldots,\ell^{\mathfrak{a}}_k\} \times \{1,\ldots,\ell^{\mathfrak{a}}_{k-1}\}} \in \mathbb{R}^{\ell^{\mathfrak{a}}_k \times \ell^{\mathfrak{a}}_{k-1}}$ and $\mathfrak{b}^{k,\theta}_{\mathfrak{a}} = (\mathfrak{b}^{k,\theta}_{\mathfrak{a},1}, \ldots, \mathfrak{b}^{k,\theta}_{\mathfrak{a},\ell^{\mathfrak{a}}_k}) \in \mathbb{R}^{\ell^{\mathfrak{a}}_k}$ satisfy for all $i \in \{1, 2, \ldots, \ell^{\mathfrak{a}}_k\}$, $j \in \{1, 2, \ldots, \ell^{\mathfrak{a}}_{k-1}\}$ that

$$\mathfrak{w}^{k,\theta}_{\mathfrak{a},i,j} = \theta_{(i-1)\ell^{\mathfrak{a}}_{k-1} + j + \sum_{h=1}^{k-1} \ell^{\mathfrak{a}}_h(\ell^{\mathfrak{a}}_{h-1}+1)} \qquad \text{and} \qquad \mathfrak{b}^{k,\theta}_{\mathfrak{a},i} = \theta_{\ell^{\mathfrak{a}}_k \ell^{\mathfrak{a}}_{k-1} + i + \sum_{h=1}^{k-1} \ell^{\mathfrak{a}}_h(\ell^{\mathfrak{a}}_{h-1}+1)}, \tag{8.57}$$

let $\mathfrak{M}\colon (\cup_{n \in \mathbb{N}} \mathbb{R}^n) \to (\cup_{n \in \mathbb{N}} \mathbb{R}^n)$ satisfy for all $n \in \mathbb{N}$, $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ that $\mathfrak{M}(x) = (\max\{x_1, 0\}, \ldots, \max\{x_n, 0\})$, for every $\mathfrak{a} \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}$ let $\mathcal{N}^{k,\theta}_{\mathfrak{a}}\colon \mathbb{R}^d \to \mathbb{R}^{\ell^{\mathfrak{a}}_k}$, $k \in \mathbb{N} \cap [1, \rho_{\mathfrak{a}}]$, satisfy for all $k \in \mathbb{N} \cap [1, \rho_{\mathfrak{a}})$, $x \in \mathbb{R}^d$ that

$$\mathcal{N}^{1,\theta}_{\mathfrak{a}}(x) = \mathfrak{b}^{1,\theta}_{\mathfrak{a}} + \mathfrak{w}^{1,\theta}_{\mathfrak{a}} x \qquad \text{and} \qquad \mathcal{N}^{k+1,\theta}_{\mathfrak{a}}(x) = \mathfrak{b}^{k+1,\theta}_{\mathfrak{a}} + \mathfrak{w}^{k+1,\theta}_{\mathfrak{a}}\big(\mathfrak{M}(\mathcal{N}^{k,\theta}_{\mathfrak{a}}(x))\big), \tag{8.58}$$

and for every $\mathfrak{a} \in \mathbb{N}$ let $\mathcal{L}_{\mathfrak{a}}\colon \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}$ that $\mathcal{L}_{\mathfrak{a}}(\theta) = \int_{[a,b]^d} \|\mathcal{N}^{\rho_{\mathfrak{a}},\theta}_{\mathfrak{a}}(x) - f(x)\|^2 \mathfrak{p}(x)\, dx$. Then

$$\limsup_{\mathfrak{a} \to \infty} \inf_{\theta \in \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}} \mathcal{L}_{\mathfrak{a}}(\theta) = 0. \tag{8.59}$$

Proof of Proposition 8.2. Throughout this proof let $\varepsilon \in (0, \infty)$. Observe that Lemma 8.2 proves that there exist $n \in \mathbb{N}$, $\mathbf{w} \in \mathbb{R}^{n \times d}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{v} \in \mathbb{R}^{\delta \times n}$ which satisfy

$$\int_{[a,b]^d} \|\mathbf{v}\mathfrak{M}(\mathbf{w}x + \mathbf{b}) - f(x)\|^2 \mathfrak{p}(x)\, dx < \varepsilon. \tag{8.60}$$

Furthermore, note that (8.56) assures that there exists $\mathbf{A} \in \mathbb{N}$ which satisfies for all $\mathfrak{a} \in \mathbb{N} \cap [\mathbf{A}, \infty)$, $i \in \mathbb{N} \cap (1, \rho_{\mathfrak{a}})$ that $\ell^{\mathfrak{a}}_1 \geq n$ and $\ell^{\mathfrak{a}}_i \geq 2\delta$. Combining this with Beck et al. [6, Lemma 2.10] (applied for every $\mathfrak{a} \in \mathbb{N} \cap [\mathbf{A}, \infty)$ with $L \curvearrowleft 2$, $(l_0, l_1, l_2) \curvearrowleft (d, n, \delta)$, $d \curvearrowleft n(d+1) + \delta(n+1)$, $\mathfrak{L} \curvearrowleft \rho_{\mathfrak{a}}$, $(\mathfrak{l}_0, \mathfrak{l}_1, \ldots, \mathfrak{l}_{\mathfrak{L}}) \curvearrowleft \ell^{\mathfrak{a}}$, $\mathfrak{d} \curvearrowleft \mathfrak{d}_{\mathfrak{a}}$ in the notation of [6, Lemma 2.10]) shows for every $\mathfrak{a} \in \mathbb{N} \cap [\mathbf{A}, \infty)$ that there exists $\theta_{\mathfrak{a}} \in \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}$ which satisfies for all $x \in \mathbb{R}^d$ that

$$\mathcal{N}^{\rho_{\mathfrak{a}},\theta_{\mathfrak{a}}}_{\mathfrak{a}}(x) = \mathbf{v}\mathfrak{M}(\mathbf{w}x + \mathbf{b}). \tag{8.61}$$

Observe that (8.60) and (8.61) ensure for all $\mathfrak{a} \in \mathbb{N} \cap [\mathbf{A}, \infty)$ that

$$\inf_{\vartheta \in \mathbb{R}^{\mathfrak{d}_{\mathfrak{a}}}} \mathcal{L}_{\mathfrak{a}}(\vartheta) \leq \mathcal{L}_{\mathfrak{a}}(\theta_{\mathfrak{a}}) = \int_{[a,b]^d} \|\mathcal{N}^{\rho_{\mathfrak{a}},\theta_{\mathfrak{a}}}_{\mathfrak{a}}(x) - f(x)\|^2 \mathfrak{p}(x)\, dx < \varepsilon. \tag{8.62}$$

This completes the proof of Proposition 8.2. ☐

8.6 Convergence of GD with random initializations in the training of deep ANNs

We next combine the Kurdyka-Łojasiewicz inequality from Proposition 6.2 with the abstract convergence result for GD with random initializations from Corollary 8.5 to prove convergence in probability of GD with random initializations for deep ANNs with a fixed architecture. In Proposition 8.3 the parameter vector $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ is assumed to be a local minimum of the risk function $\mathcal{L}_\infty$ in a neighborhood of which the regularity assumptions in Corollary 8.5 are satisfied. The convergence holds for every sufficiently small positive learning rate $\gamma \in (0, \mathfrak{g}]$.

Proposition 8.3. *Assume Setting 3.1, assume for all $i \in \{1, 2, \ldots, \ell_L\}$ that f_i is piecewise polynomial, let $\mathfrak{p} \colon [a,b]^{\ell_0} \to \mathbb{R}$ be piecewise polynomial, assume for all $E \in \mathcal{B}([a,b]^{\ell_0})$ that $\mu(E) = \int_E \mathfrak{p}(x)\,\mathrm{d}x$, let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, assume $(\mathcal{L}_\infty)|_U \in C^1(U, \mathbb{R})$, let $\mathfrak{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in U$ that $\mathfrak{G}(\theta) = (\nabla \mathcal{L}_\infty)(\theta)$, assume that $\mathfrak{G}|_U$ is locally Lipschitz continuous, let $\vartheta \in U$ satisfy $\mathcal{L}_\infty(\vartheta) = \inf_{\theta \in U} \mathcal{L}_\infty(\theta)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $K, n \in \mathbb{N}_0$, $\gamma \in \mathbb{R}$ let $\Theta_n^{K,\gamma} \colon \Omega \to \mathbb{R}^{\mathfrak{d}}$ and $\mathbf{k}_n^{K,\gamma} \colon \Omega \to \mathbb{N}$ be random variables, assume for all $\gamma \in \mathbb{R}$ that $\Theta_0^{K,\gamma}$, $K \in \mathbb{N}$, are i.i.d., assume for all $\gamma, \delta \in (0,1)$ that $\mathbb{P}(\|\Theta_0^{1,\gamma} - \vartheta\| < \delta) > 0$, and assume for all $K \in \mathbb{N}$, $n \in \mathbb{N}_0$, $\gamma \in \mathbb{R}$, $\omega \in \Omega$ that*

$$\Theta_{n+1}^{K,\gamma}(\omega) = \Theta_n^{K,\gamma}(\omega) - \gamma \mathfrak{G}(\Theta_n^{K,\gamma}(\omega)), \tag{8.63a}$$

$$\mathbf{k}_n^{K,\gamma}(\omega) \in \arg\min\nolimits_{\kappa \in \{1,2,\ldots,K\}} \mathcal{L}_\infty(\Theta_n^{\kappa,\gamma}(\omega)) \tag{8.63b}$$

(cf. Definition 5.1). Then there exists $\mathfrak{g} \in (0, \infty)$ such that for all $\gamma \in (0, \mathfrak{g}]$ it holds that

$$\liminf\nolimits_{K \to \infty} \mathbb{P}\Big(\limsup\nolimits_{n \to \infty} \mathcal{L}_\infty(\Theta_n^{\mathbf{k}_n^{K,\gamma},\gamma}) \le \inf\nolimits_{\theta \in U} \mathcal{L}_\infty(\theta)\Big) = 1. \tag{8.64}$$

Proof of Proposition 8.3. Note that Corollary 3.2 assures for all open $V \subseteq \mathbb{R}^{\mathfrak{d}}$ and all $\theta \in V$ with $(\mathcal{L}_\infty)|_V \in C^1(V, \mathbb{R})$ that $\mathcal{G}(\theta) = (\nabla \mathcal{L}_\infty)(\theta)$. The assumption that $U \subseteq \mathbb{R}^{\mathfrak{d}}$ is open, the assumption that $(\mathcal{L}_\infty)|_U \in C^1(U, \mathbb{R})$, and the assumption that for all $\theta \in U$ it holds that $\mathfrak{G}(\theta) = (\nabla \mathcal{L}_\infty)(\theta)$ therefore demonstrates that for all $\theta \in U$ it holds that

$$\mathcal{G}(\theta) = (\nabla \mathcal{L}_\infty)(\theta) = \mathfrak{G}(\theta). \tag{8.65}$$

Furthermore, observe that Proposition 6.2, the assumption that $U \subseteq \mathbb{R}^{\mathfrak{d}}$ is open, and the assumption that $\mathfrak{G}|_U$ is locally Lipschitz continuous assure that there exist $\mathbb{L}, \varepsilon, \mathfrak{C} \in (0, \infty)$, $\alpha \in (0,1)$ which satisfy for all $v, w \in \{\psi \in \mathbb{R}^{\mathfrak{d}} \colon \|\psi - \vartheta\| < \varepsilon\}$ that

$$v \in U, \qquad |\mathcal{L}_\infty(v) - \mathcal{L}_\infty(\vartheta)|^\alpha \le \mathfrak{C} \|\mathcal{G}(v)\|, \tag{8.66a}$$

$$\|\mathfrak{G}(v) - \mathfrak{G}(w)\| \le \mathbb{L} \|v - w\|. \tag{8.66b}$$

Moreover, note that Lemma 3.1 shows that $\mathcal{L}_\infty \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$. Combining this, (8.65), (8.66), Corollary 8.5 (applied for every $\gamma \in (0, \mathbb{L}^{-1}] \cap (0,1)$ with $\mathfrak{d} \curvearrowleft \mathfrak{d}$, $\varepsilon \curvearrowleft \varepsilon$, $L \curvearrowleft \mathbb{L}$, $\mathfrak{C} \curvearrowleft \mathfrak{C}$, $\gamma \curvearrowleft \gamma$, $\alpha \curvearrowleft \alpha$, $\vartheta \curvearrowleft \vartheta$, $\mathcal{L} \curvearrowleft \mathcal{L}_\infty$, $\mathcal{G} \curvearrowleft \mathfrak{G}$ in the notation of Corollary 8.5), and the assumption that $\mathcal{L}_\infty(\vartheta) = \inf_{\theta \in U} \mathcal{L}_\infty(\theta)$ assures that for all $\gamma \in (0, \mathbb{L}^{-1}] \cap (0,1)$ it holds that

$$\begin{aligned}
&\liminf_{K\to\infty} \mathbb{P}\Big(\limsup_{n\to\infty} \mathcal{L}_\infty\big(\Theta_n^{\mathbf{k}_n^{K,\gamma},\gamma}\big) \le \inf_{\theta\in U} \mathcal{L}_\infty(\theta)\Big) \\
&= \liminf_{K\to\infty} \mathbb{P}\Big(\limsup_{n\to\infty} \mathcal{L}_\infty\big(\Theta_n^{\mathbf{k}_n^{K,\gamma},\gamma}\big) \le \mathcal{L}_\infty(\vartheta)\Big) = 1. \qquad (8.67)
\end{aligned}$$

The proof of Proposition 8.3 is thus complete. □

As a consequence of Proposition 8.3 and the universal approximation result from Proposition 8.2 we verify in Theorem 8.1 that the risk of the GD method with random initializations converges in probability to 0 as the number of GD steps, the number of random initializations, and the width of the ANNs increase to ∞ and as the step size of the GD method decreases to 0. In item (i) we establish convergence in probability, and as a consequence we obtain in item (ii) convergence with respect to the metric $\mathbb{E}[\min\{|X-Y|,1\}]$ on the space of random variables.

Theorem 8.1. *Let $d, \delta \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in [a,\infty)$, $(\rho_\mathfrak{a})_{\mathfrak{a}\in\mathbb{N}} \subseteq (\mathbb{N} \cap (1,\infty))$, let $\ell^\mathfrak{a} = (\ell_0^\mathfrak{a}, \ell_1^\mathfrak{a}, \ldots, \ell_{\rho_\mathfrak{a}}^\mathfrak{a}) \in \{d\} \times \mathbb{N}^{\rho_\mathfrak{a}-1} \times \{\delta\}$, $\mathfrak{a} \in \mathbb{N}$, satisfy*

$$\liminf_{\mathfrak{a}\to\infty} \min\{\ell_1^\mathfrak{a}, \ell_2^\mathfrak{a}, \ldots, \ell_{\rho_\mathfrak{a}-1}^\mathfrak{a}\} = \infty, \qquad (8.68)$$

for every $\mathfrak{a} \in \mathbb{N}$ let $\mathfrak{d}_\mathfrak{a} = \sum_{k=1}^{\rho_\mathfrak{a}} \ell_k^\mathfrak{a}(\ell_{k-1}^\mathfrak{a}+1)$, let $f = (f_1, \ldots, f_\delta) \colon [a,b]^d \to \mathbb{R}^\delta$ and $\mathfrak{p} \colon [a,b]^d \to [0,\infty)$ be functions, assume for all $i \in \{1,2,\ldots,\delta\}$ that f_i and $\mathfrak{p}$ are piecewise polynomial, for every $\mathfrak{a} \in \mathbb{N}$, $k \in \{1,2,\ldots,\rho_\mathfrak{a}\}$, $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}_\mathfrak{a}}) \in \mathbb{R}^{\mathfrak{d}_\mathfrak{a}}$ let $\mathfrak{w}_\mathfrak{a}^{k,\theta} = (\mathfrak{w}_{\mathfrak{a},i,j}^{k,\theta})_{(i,j)\in\{1,\ldots,\ell_k^\mathfrak{a}\}\times\{1,\ldots,\ell_{k-1}^\mathfrak{a}\}} \in \mathbb{R}^{\ell_k^\mathfrak{a}\times\ell_{k-1}^\mathfrak{a}}$ and $\mathfrak{b}_\mathfrak{a}^{k,\theta} = (\mathfrak{b}_{\mathfrak{a},1}^{k,\theta}, \ldots, \mathfrak{b}_{\mathfrak{a},\ell_k^\mathfrak{a}}^{k,\theta}) \in \mathbb{R}^{\ell_k^\mathfrak{a}}$ satisfy for all $i \in \{1,2,\ldots,\ell_k^\mathfrak{a}\}$, $j \in \{1,2,\ldots,\ell_{k-1}^\mathfrak{a}\}$ that

$$\mathfrak{w}_{\mathfrak{a},i,j}^{k,\theta} = \theta_{(i-1)\ell_{k-1}^\mathfrak{a}+j+\sum_{h=1}^{k-1}\ell_h^\mathfrak{a}(\ell_{h-1}^\mathfrak{a}+1)} \qquad \text{and} \qquad \mathfrak{b}_{\mathfrak{a},i}^{k,\theta} = \theta_{\ell_k^\mathfrak{a}\ell_{k-1}^\mathfrak{a}+i+\sum_{h=1}^{k-1}\ell_h^\mathfrak{a}(\ell_{h-1}^\mathfrak{a}+1)}, \qquad (8.69)$$

let $\mathfrak{M} \colon (\cup_{n\in\mathbb{N}}\mathbb{R}^n) \to (\cup_{n\in\mathbb{N}}\mathbb{R}^n)$ satisfy for all $n \in \mathbb{N}$, $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ that $\mathfrak{M}(x) = (\max\{x_1,0\}, \ldots, \max\{x_n,0\})$, for every $\mathfrak{a} \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}_\mathfrak{a}}$ let $\mathcal{N}_\mathfrak{a}^{k,\theta} \colon \mathbb{R}^d \to \mathbb{R}^{\ell_k^\mathfrak{a}}$, $k \in \mathbb{N} \cap [1,\rho_\mathfrak{a}]$, satisfy for all $k \in \mathbb{N} \cap [1,\rho_\mathfrak{a})$, $x \in \mathbb{R}^d$ that

$$\mathcal{N}_\mathfrak{a}^{1,\theta}(x) = \mathfrak{b}_\mathfrak{a}^{1,\theta} + \mathfrak{w}_\mathfrak{a}^{1,\theta} x \qquad \text{and} \qquad \mathcal{N}_\mathfrak{a}^{k+1,\theta}(x) = \mathfrak{b}_\mathfrak{a}^{k+1,\theta} + \mathfrak{w}_\mathfrak{a}^{k+1,\theta}\big(\mathfrak{M}(\mathcal{N}_\mathfrak{a}^{k,\theta}(x))\big), \qquad (8.70)$$

for every $\mathfrak{a} \in \mathbb{N}$ let $\mathcal{L}_\mathfrak{a} \colon \mathbb{R}^{\mathfrak{d}_\mathfrak{a}} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}_\mathfrak{a}}$ that $\mathcal{L}_\mathfrak{a}(\theta) = \int_{[a,b]^d} \|\mathcal{N}_\mathfrak{a}^{\rho_\mathfrak{a},\theta}(x) - f(x)\|^2\, \mathfrak{p}(x)\, \mathrm{d}x$, for every $\mathfrak{a} \in \mathbb{N}$ let $\vartheta_\mathfrak{a} \in (\mathcal{L}_\mathfrak{a})^{-1}(\{\inf_{\theta\in\mathbb{R}^{\mathfrak{d}_\mathfrak{a}}} \mathcal{L}_\mathfrak{a}(\theta)\})$, $\varepsilon_\mathfrak{a} \in (0,1)$ satisfy that $\mathcal{L}_\mathfrak{a}|_{\{\theta\in\mathbb{R}^{\mathfrak{d}_\mathfrak{a}} \colon \|\theta-\vartheta_\mathfrak{a}\|<\varepsilon_\mathfrak{a}\}}$ has a Lipschitz continuous derivative, for every $\mathfrak{a} \in \mathbb{N}$ let $\mathcal{G}_\mathfrak{a} \colon \mathbb{R}^{\mathfrak{d}_\mathfrak{a}} \to \mathbb{R}^{\mathfrak{d}_\mathfrak{a}}$ satisfy for all $\theta \in \cup_{U\subseteq\mathbb{R}^{\mathfrak{d}_\mathfrak{a}},\, U \text{ is open},\, \mathcal{L}_\mathfrak{a}|_U \in C^1(U,\mathbb{R})} U$ that $\mathcal{G}_\mathfrak{a}(\theta) = (\nabla\mathcal{L}_\mathfrak{a})(\theta)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $n, \mathfrak{a}, K \in \mathbb{N}_0$, $\gamma \in \mathbb{R}$ let $\Theta_{\mathfrak{a},n}^{K,\gamma} \colon \Omega \to \mathbb{R}^{\mathfrak{d}_\mathfrak{a}}$ and $\mathbf{k}_{\mathfrak{a},n}^{K,\gamma} \colon \Omega \to \mathbb{N}$ be random variables, assume for all $\mathfrak{a} \in \mathbb{N}$, $\gamma \in \mathbb{R}$ that $\Theta_{\mathfrak{a},0}^{K,\gamma}$, $K \in \mathbb{N}$, are i.i.d., assume for all

$\mathfrak{a} \in \mathbb{N}$, $\gamma, r \in (0,1)$, $\theta \in \mathbb{R}^{\mathfrak{d}_\mathfrak{a}}$ that $\mathbb{P}(\|\Theta^{1,\gamma}_{\mathfrak{a},0} - \theta\| < r) > 0$, and assume for all $n \in \mathbb{N}_0$, $\mathfrak{a}, K \in \mathbb{N}$, $\gamma \in \mathbb{R}$, $\omega \in \Omega$ that

$$\Theta^{K,\gamma}_{\mathfrak{a},n+1}(\omega) = \Theta^{K,\gamma}_{\mathfrak{a},n}(\omega) - \gamma \mathcal{G}_\mathfrak{a}(\Theta^{K,\gamma}_{\mathfrak{a},n}(\omega)), \tag{8.71a}$$

$$\mathbf{k}^{K,\gamma}_{\mathfrak{a},n}(\omega) \in \arg\min\nolimits_{\kappa \in \{1,2,\dots,K\}} \mathcal{L}_\mathfrak{a}(\Theta^{\kappa,\gamma}_{\mathfrak{a},n}(\omega)) \tag{8.71b}$$

(cf. Definition 5.1). Then

(i) there exist $\mathbf{A} \colon \mathbb{R} \to \mathbb{R}$ and $\mathfrak{g} \colon \mathbb{N} \to (0,\infty)$ such that

$$\inf\nolimits_{\varepsilon \in (0,\infty)} \inf\nolimits_{\mathfrak{a} \in \mathbb{N} \cap [\mathbf{A}(\varepsilon),\infty)} \inf\nolimits_{\gamma \in (0,\mathfrak{g}(\mathfrak{a})]} \liminf\nolimits_{K \to \infty} \mathbb{P}\big(\limsup\nolimits_{n \to \infty} \mathcal{L}_\mathfrak{a}\big(\Theta^{\mathbf{k}^{K,\gamma}_{\mathfrak{a},n},\gamma}_{\mathfrak{a},n}\big) \le \varepsilon\big) = 1 \tag{8.72}$$

and

(ii) it holds that

$$\limsup\nolimits_{\mathfrak{a} \to \infty} \limsup\nolimits_{\gamma \searrow 0} \limsup\nolimits_{K \to \infty} \mathbb{E}\big[\limsup\nolimits_{n \to \infty} \min\big\{\mathcal{L}_\mathfrak{a}\big(\Theta^{\mathbf{k}^{K,\gamma}_{\mathfrak{a},n},\gamma}_{\mathfrak{a},n}\big), 1\big\}\big] = 0. \tag{8.73}$$

Proof of Theorem 8.1. Throughout this proof for every $\mathfrak{a} \in \mathbb{N}$ let $U_\mathfrak{a} \subseteq \mathbb{R}^{\mathfrak{d}_\mathfrak{a}}$ satisfy

$$U_\mathfrak{a} = \{\theta \in \mathbb{R}^{\mathfrak{d}_\mathfrak{a}} \colon \|\theta - \vartheta_\mathfrak{a}\| < \varepsilon_\mathfrak{a}\}. \tag{8.74}$$

Observe the assumption that $\mathfrak{p}$ is piecewise polynomial and the assumption that for all $i \in \{1,2,\dots,\delta\}$ it holds that f_i is piecewise polynomial imply that f and $\mathfrak{p}$ are bounded and measurable. Combining this, [36, Lemma 2.4], Proposition 8.3 (applied for every $\mathfrak{a} \in \mathbb{N}$ with $a \curvearrowleft a$, $b \curvearrowleft b$, $\curvearrowleft 1/2$, $\curvearrowleft 1$, $(\mathbb{N}_0 \ni k \mapsto \ell_k \in \mathbb{N}) \curvearrowleft (\mathbb{N}_0 \ni k \mapsto \ell_{\min\{k,\rho_\mathfrak{a}\}^\mathfrak{a}} \in \mathbb{N})$, $L \curvearrowleft \rho_\mathfrak{a}$, $\mathfrak{d} \curvearrowleft \mathfrak{d}_\mathfrak{a}$, $f \curvearrowleft f$, $\mu \curvearrowleft (\mathcal{B}([a,b]^d) \ni E \mapsto \int_E \mathfrak{p}(x)\,dx \in [0,\infty])$, $U \curvearrowleft U_\mathfrak{a}$, $\mathfrak{G} \curvearrowleft \mathcal{G}_\mathfrak{a}$, $\vartheta \curvearrowleft \vartheta_\mathfrak{a}$ in the notation of Proposition 8.3), and the fact that for all $\mathfrak{a} \in \mathbb{N}$ it holds that

$$\mathcal{L}_\mathfrak{a}(\vartheta_\mathfrak{a}) = \inf\nolimits_{\theta \in \mathbb{R}^{\mathfrak{d}_\mathfrak{a}}} \mathcal{L}_\mathfrak{a}(\theta) = \inf\nolimits_{\theta \in U_\mathfrak{a}} \mathcal{L}_\mathfrak{a}(\theta) \tag{8.75}$$

proves that there there exists $\mathfrak{g} \colon \mathbb{N} \to (0,\infty)$ which satisfies for all $\mathfrak{a} \in \mathbb{N}$, $\gamma \in (0,\mathfrak{g}(\mathfrak{a})]$ that

$$\liminf\nolimits_{K \to \infty} \mathbb{P}\Big(\limsup\nolimits_{n \to \infty} \mathcal{L}_\mathfrak{a}\big(\Theta^{\mathbf{k}^{K,\gamma}_{\mathfrak{a},n},\gamma}_{\mathfrak{a},n}\big) \le \inf\nolimits_{\theta \in \mathbb{R}^{\mathfrak{d}_\mathfrak{a}}} \mathcal{L}_\mathfrak{a}(\theta)\Big) = 1. \tag{8.76}$$

Proposition 8.2 hence establishes that there exists $\mathbf{A} \colon \mathbb{R} \to \mathbb{R}$ which satisfies for all $\varepsilon \in (0,\infty)$, $\mathfrak{a} \in \mathbb{N} \cap [\mathbf{A}(\varepsilon),\infty)$ that

$$\inf\nolimits_{\theta \in \mathbb{R}^{\mathfrak{d}_\mathfrak{a}}} \mathcal{L}_\mathfrak{a}(\theta) \le \varepsilon. \tag{8.77}$$

Note that (8.76) and (8.77) assure that for all $\varepsilon \in (0,\infty)$, $\mathfrak{a} \in \mathbb{N} \cap [\mathbf{A}(\varepsilon),\infty)$, $\gamma \in (0,\mathfrak{g}(\mathfrak{a})]$ it holds that

$$\begin{aligned} &\liminf\nolimits_{K \to \infty} \mathbb{P}\Big(\limsup\nolimits_{n \to \infty} \mathcal{L}_\mathfrak{a}\big(\Theta^{\mathbf{k}^{K,\gamma}_{\mathfrak{a},n},\gamma}_{\mathfrak{a},n}\big) \le \varepsilon\Big) \\ &\ge \liminf\nolimits_{K \to \infty} \mathbb{P}\Big(\limsup\nolimits_{n \to \infty} \mathcal{L}_\mathfrak{a}\big(\Theta^{\mathbf{k}^{K,\gamma}_{\mathfrak{a},n},\gamma}_{\mathfrak{a},n}\big) \le \inf\nolimits_{\theta \in \mathbb{R}^{\mathfrak{d}_\mathfrak{a}}} \mathcal{L}_\mathfrak{a}(\theta)\Big) = 1. \end{aligned} \tag{8.78}$$

This establishes item (i). Observe that for all $\varepsilon \in (0,\infty)$ and all random variables $Z_n \colon \Omega \to [0,\infty)$, $n \in \mathbb{N}$, it holds that

$$\begin{aligned}
&\mathbb{E}[\limsup_{n\to\infty} \min\{Z_n, 1\}] \\
&\le \mathbb{E}\big[\min\{\limsup_{n\to\infty} Z_n, 1\}\big] \\
&\le \mathbb{E}\big[\min\{\limsup_{n\to\infty} Z_n, 1\}\mathbb{1}_{\{\limsup_{n\to\infty} Z_n > \varepsilon\}}\big] \\
&\quad + \mathbb{E}\big[\min\{\limsup_{n\to\infty} Z_n, 1\}\mathbb{1}_{\{\limsup_{n\to\infty} Z_n \le \varepsilon\}}\big] \\
&\le \mathbb{E}\big[\mathbb{1}_{\{\limsup_{n\to\infty} Z_n > \varepsilon\}}\big] + \mathbb{E}\big[\min\{\varepsilon, 1\}\mathbb{1}_{\{\limsup_{n\to\infty} Z_n \le \varepsilon\}}\big] \\
&\le \mathbb{P}\big(\limsup_{n\to\infty} Z_n > \varepsilon\big) + \min\{\varepsilon, 1\} \\
&\le \mathbb{P}(\limsup_{n\to\infty} Z_n > \varepsilon) + \varepsilon.
\end{aligned} \tag{8.79}$$

Furthermore, note that (8.78) assures that for all $\varepsilon \in (0,\infty)$, $\mathfrak{a} \in \mathbb{N} \cap [\mathbf{A}(\varepsilon), \infty)$, $\gamma \in (0, \mathfrak{g}(\mathfrak{a})]$ it holds that

$$\begin{aligned}
&\limsup_{K\to\infty} \mathbb{P}\Big(\limsup_{n\to\infty} \mathcal{L}_{\mathfrak{a}}\big(\Theta_{\mathfrak{a},n}^{\mathbf{k}_{\mathfrak{a},n}^{K,\gamma},\gamma}\big) > \varepsilon\Big) \\
&= \limsup_{K\to\infty} \Big[1 - \mathbb{P}\Big(\limsup_{n\to\infty} \mathcal{L}_{\mathfrak{a}}\big(\Theta_{\mathfrak{a},n}^{\mathbf{k}_{\mathfrak{a},n}^{K,\gamma},\gamma}\big) \le \varepsilon\Big)\Big] = 0.
\end{aligned} \tag{8.80}$$

Combining this with (8.79) ensures that for all $\varepsilon \in (0,\infty)$, $\mathfrak{a} \in \mathbb{N} \cap [\mathbf{A}(\varepsilon), \infty)$, $\gamma \in (0, \mathfrak{g}(\mathfrak{a})]$ it holds that

$$\limsup_{K\to\infty} \mathbb{E}\big[\limsup_{n\to\infty} \min\{\mathcal{L}_{\mathfrak{a}}\big(\Theta_{\mathfrak{a},n}^{\mathbf{k}_{\mathfrak{a},n}^{K,\gamma},\gamma}\big), 1\}\big] \le \varepsilon. \tag{8.81}$$

This establishes item (ii). The proof of Theorem 8.1 is thus complete. □

8.7 Convergence of GD with random initializations in the training of shallow ANNs

In this section we employ the general convergence results for deep ANNs from Subsection 8.6 to establish convergence of the risk of the GD method for shallow ANNs. This time the regularity assumptions can be omitted, since they follow from the existence result for regular global minima for shallow ANNs in Corollary 2.6.

Proposition 8.4. *Assume Setting 3.1, assume $L = 2$, assume $f \in C([a,b], \mathbb{R})$, assume that f is piecewise polynomial, let $\mathfrak{p} \colon [a,b] \to \mathbb{R}$ be piecewise polynomial, assume for all $E \in \mathcal{B}([a,b])$ that $\mu(E) = \int_E \mathfrak{p}(x)\,\mathrm{d}x$, let $\mathfrak{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \{\vartheta \in \mathbb{R}^{\mathfrak{d}} \colon \mathcal{L}_\infty$ is differentiable at $\vartheta\}$ that $\mathfrak{G}(\theta) = (\nabla \mathcal{L}_\infty)(\theta)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $K, n \in \mathbb{N}_0$, $\gamma \in \mathbb{R}$ let $\Theta_n^{K,\gamma} \colon \Omega \to \mathbb{R}^{\mathfrak{d}}$ and $\mathbf{k}_n^{K,\gamma} \colon \Omega \to \mathbb{N}$ be random variables, assume for all $\gamma \in \mathbb{R}$ that $\Theta_0^{K,\gamma}$, $K \in \mathbb{N}$, are i.i.d., assume for all $\gamma, \delta \in (0,1)$ that $\mathbb{P}(\|\Theta_0^{1,\gamma} - \vartheta\| < \delta) > 0$, and assume for all $K \in \mathbb{N}$, $n \in \mathbb{N}_0$, $\gamma \in \mathbb{R}$, $\omega \in \Omega$ that*

$$\Theta_{n+1}^{K,\gamma}(\omega) = \Theta_n^{K,\gamma}(\omega) - \gamma \mathfrak{G}(\Theta_n^{K,\gamma}(\omega)), \tag{8.82a}$$

$$\mathbf{k}_n^{K,\gamma}(\omega) \in \operatorname{arg\,min}_{\kappa \in \{1,2,\ldots,K\}} \mathcal{L}_\infty(\Theta_n^{\kappa,\gamma}(\omega)) \tag{8.82b}$$

(cf. Definition 5.1). Then there exists $\mathfrak{g} \in (0, \infty)$ *such that for all* $\gamma \in (0, \mathfrak{g}]$ *it holds that*

$$\liminf_{K \to \infty} \mathbb{P}\Big(\limsup_{n \to \infty} \mathcal{L}_\infty(\Theta_n^{\mathbf{k}_n^{K,\gamma},\gamma}) = \inf_{\theta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}_\infty(\theta)\Big) = 1. \tag{8.83}$$

Proof of Proposition 8.4. Observe that the assumption that $f \in C([a, b], \mathbb{R})$ and the assumption that f is piecewise polynomial imply that f is Lipschitz continuous. Furthermore, note that (3.3) and the assumption that $\sup_{r \in [1,\infty)} \sup_{x \in \mathbb{R}} |(\mathfrak{R}_r)'(x)| < \infty$ assure that for all $x \in \mathbb{R}$ it holds that $(\cup_{r \in \mathbb{N}} \{\mathfrak{R}_r\}) \subseteq C^1(\mathbb{R}, \mathbb{R})$, $\mathfrak{R}_\infty(x) = \max\{x, 0\}$, $\sup_{r \in \mathbb{N}} \sup_{y \in [-|x|,|x|]} |(\mathfrak{R}_r)'(y)| < \infty$, and

$$\limsup_{r \to \infty} (|\mathfrak{R}_r(x) - \mathfrak{R}_\infty(x)| + |(\mathfrak{R}_r)'(x) - \mathbb{1}_{(0,\infty)}(x)|) = 0. \tag{8.84}$$

Combining this, Corollary 2.6, and item (ii) in Proposition 3.1 with the fact that f is Lipschitz continuous shows that there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ and an open $U \subseteq \mathbb{R}^{\mathfrak{d}}$ which satisfy that

(i) it holds that $(\mathcal{L}_\infty)|_U \in C^1(U, \mathbb{R})$,

(ii) it holds for all $\theta \in U$ that $\mathcal{G}(\theta) = (\nabla \mathcal{L}_\infty)(\theta)$,

(iii) it holds that $\mathcal{G}|_U$ is locally Lipschitz continuous,

(iv) it holds that $\vartheta \in U$, and

(v) it holds that $\mathcal{L}_\infty(\vartheta) = \inf_{\theta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}_\infty(\theta)$.

Observe that item (iv) and item (v) ensure that

$$\mathcal{L}_\infty(\vartheta) = \inf_{\theta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}_\infty(\theta) = \inf_{\theta \in U} \mathcal{L}_\infty(\theta). \tag{8.85}$$

Moreover, note that item (i), item (ii), and the assumption that for all $\theta \in \{\vartheta \in \mathbb{R}^{3\mathfrak{h}+1} \colon \mathcal{L}_\infty$ is differentiable at $\vartheta\}$ it holds that $\mathfrak{G}(\theta) = (\nabla \mathcal{L}_\infty)(\theta)$ assure that for all $\theta \in U$ it holds that

$$\mathcal{G}(\theta) = (\nabla \mathcal{L}_\infty)(\theta) = \mathfrak{G}(\theta). \tag{8.86}$$

Therefore, we obtain that $\mathfrak{G}|_U = \mathcal{G}|_U$. This and item (iii) ensure that $\mathfrak{G}|_U$ is locally Lipschitz continuous. Combining this, item (i), (8.85), (8.86), Proposition 8.3, and the fact that $U \subseteq \mathbb{R}^{\mathfrak{d}}$ is open proves that there exists $\mathfrak{g} \in (0, \infty)$ such that for all $\gamma \in (0, \mathfrak{g}]$ it holds that

$$\begin{aligned} &\liminf_{K \to \infty} \mathbb{P}\Big(\limsup_{n \to \infty} \mathcal{L}_\infty(\Theta_n^{\mathbf{k}_n^{K,\gamma},\gamma}) = \inf_{\theta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}_\infty(\theta)\Big) \\ &= \liminf_{K \to \infty} \mathbb{P}\Big(\limsup_{n \to \infty} \mathcal{L}_\infty(\Theta_n^{\mathbf{k}_n^{K,\gamma},\gamma}) \le \inf_{\theta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}_\infty(\theta)\Big) \\ &= \liminf_{K \to \infty} \mathbb{P}\Big(\limsup_{n \to \infty} \mathcal{L}_\infty(\Theta_n^{\mathbf{k}_n^{K,\gamma},\gamma}) \le \inf_{\theta \in U} \mathcal{L}_\infty(\theta)\Big) = 1. \end{aligned} \tag{8.87}$$

The proof of Proposition 8.4 is thus complete. □

Corollary 8.6. *Let $N \in \mathbb{N}$, ${}_0, {}_1, \ldots, {}_N, a, b \in \mathbb{R}$ satisfy $a = {}_0 < {}_1 < \ldots < {}_N = b$, let $f \in C([a,b], \mathbb{R})$, let $\mathfrak{p}\colon [a,b] \to [0,\infty)$ be a function, assume for all $i \in \{1,2,\ldots,N\}$ that $f|_{({i-1}, {i})}$ and $\mathfrak{p}|_{({i-1}, {i})}$ are polynomials, for every $\mathfrak{h} \in \mathbb{N}$ let $\mathcal{L}_{\mathfrak{h}}\colon \mathbb{R}^{3\mathfrak{h}+1} \to \mathbb{R}$ satisfy for all $\theta = (\theta_1, \ldots, \theta_{3\mathfrak{h}+1}) \in \mathbb{R}^{3\mathfrak{h}+1}$ that*

$$\mathcal{L}_{\mathfrak{h}}(\theta) = \int_a^b \big(f(x) - \theta_{\mathfrak{d}} - \textstyle\sum_{j=1}^{\mathfrak{h}} \theta_{2\mathfrak{h}+j} \max\{\theta_j x + \theta_{\mathfrak{h}+j}, 0\}\big)^2 \mathfrak{p}(x)\, dx, \tag{8.88}$$

for every $\mathfrak{h} \in \mathbb{N}$ let $\mathcal{G}_{\mathfrak{h}}\colon \mathbb{R}^{3\mathfrak{h}+1} \to \mathbb{R}^{3\mathfrak{h}+1}$ satisfy for all $\theta \in \{\vartheta \in \mathbb{R}^{3\mathfrak{h}+1}\colon \mathcal{L}_{\mathfrak{h}}$ is differentiable at $\vartheta\}$ that $\mathcal{G}_{\mathfrak{h}}(\theta) = (\nabla \mathcal{L}_{\mathfrak{h}})(\theta)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $n, \mathfrak{h}, K \in \mathbb{N}_0$, $\gamma \in \mathbb{R}$ let $\Theta^{K,\gamma}_{\mathfrak{h},n}\colon \Omega \to \mathbb{R}^{3\mathfrak{h}+1}$ and $\mathbf{k}^{K,\gamma}_{\mathfrak{h},n}\colon \Omega \to \mathbb{N}$ be random variables, assume for all $\mathfrak{h} \in \mathbb{N}$, $\gamma \in \mathbb{R}$ that $\Theta^{K,\gamma}_{\mathfrak{h},0}$, $K \in \mathbb{N}$, are i.i.d., assume for all $\mathfrak{h} \in \mathbb{N}$, $\gamma, r \in (0,1)$, $\theta \in \mathbb{R}^{3\mathfrak{h}+1}$ that $\mathbb{P}(\|\Theta^{1,\gamma}_{\mathfrak{h},0} - \theta\| < r) > 0$, and assume for all $n, \mathfrak{h} \in \mathbb{N}_0$, $K \in \mathbb{N}$, $\gamma \in \mathbb{R}$, $\omega \in \Omega$ that

$$\Theta^{K,\gamma}_{\mathfrak{h},n+1}(\omega) = \Theta^{K,\gamma}_{\mathfrak{h},n}(\omega) - \gamma \mathcal{G}_{\mathfrak{h}}(\Theta^{K,\gamma}_{\mathfrak{h},n}(\omega)), \tag{8.89a}$$

$$\mathbf{k}^{K,\gamma}_{\mathfrak{h},n}(\omega) \in \arg\min\nolimits_{\kappa \in \{1,2,\ldots,K\}} \mathcal{L}_{\mathfrak{h}}(\Theta^{\kappa,\gamma}_{\mathfrak{h},n}(\omega)). \tag{8.89b}$$

Then

(i) there exist $\mathbf{H}\colon \mathbb{R} \to \mathbb{R}$ and $\mathfrak{g}\colon \mathbb{N} \to (0,\infty)$ such that

$$\inf\nolimits_{\varepsilon \in (0,\infty)} \inf\nolimits_{\mathfrak{h} \in \mathbb{N} \cap [\mathbf{H}(\varepsilon),\infty)} \inf\nolimits_{\gamma \in (0,\mathfrak{g}(\mathfrak{h})]} \liminf\nolimits_{K \to \infty} \mathbb{P}\big(\limsup\nolimits_{n \to \infty} \mathcal{L}_{\mathfrak{h}}\big(\Theta^{\mathbf{k}^{K,\gamma}_{\mathfrak{h},n},\gamma}_{\mathfrak{h},n}\big) \le \varepsilon\big) = 1 \tag{8.90}$$

and

(ii) it holds that

$$\limsup\nolimits_{\mathfrak{h} \to \infty} \limsup\nolimits_{\gamma \searrow 0} \limsup\nolimits_{K \to \infty} \mathbb{E}\big[\limsup\nolimits_{n \to \infty} \min\{\mathcal{L}_{\mathfrak{h}}\big(\Theta^{\mathbf{k}^{K,\gamma}_{\mathfrak{h},n},\gamma}_{\mathfrak{h},n}\big), 1\}\big] = 0. \tag{8.91}$$

Proof of Corollary 8.6. Observe that Corollary 2.6 demonstrates that for every $\mathfrak{h} \in \mathbb{N}$ there exist $\vartheta_{\mathfrak{h}} \in \mathbb{R}^{3\mathfrak{h}+1}$, $\mathbb{L}_{\mathfrak{h}} \in \mathbb{R}$, and an open $V_{\mathfrak{h}} \subseteq \mathbb{R}^{3\mathfrak{h}+1}$ which satisfy that

(I) it holds that $\vartheta_{\mathfrak{h}} \in V_{\mathfrak{h}}$,

(II) it holds that $\mathcal{L}_{\mathfrak{h}}(\vartheta_{\mathfrak{h}}) = \inf_{\psi \in \mathbb{R}^{3\mathfrak{h}+1}} \mathcal{L}_{\mathfrak{h}}(\psi)$,

(III) it holds that $\mathcal{L}_{\mathfrak{h}}|_{V_{\mathfrak{h}}} \in C^1(V_{\mathfrak{h}}, \mathbb{R})$, and

(IV) it holds for all $\theta_1, \theta_2 \in V_{\mathfrak{h}}$ that $\|(\nabla \mathcal{L}_{\mathfrak{h}})(\theta_1) - (\nabla \mathcal{L}_{\mathfrak{h}})(\theta_2)\| \le \mathbb{L}_{\mathfrak{h}} \|\theta_1 - \theta_2\|$.

Furthermore, note that the fact that for all $\mathfrak{h} \in \mathbb{N}$, $\theta \in \{\vartheta \in \mathbb{R}^{3\mathfrak{h}+1}\colon \mathcal{L}_{\mathfrak{h}}$ is differentiable at $\vartheta\}$ it holds that

$$\mathcal{G}_{\mathfrak{h}}(\theta) = (\nabla \mathcal{L}_{\mathfrak{h}})(\theta) \tag{8.92}$$

proves that for all $\mathfrak{h} \in \mathbb{N}$, $\theta \in \cup_{U \subseteq \mathbb{R}^{3\mathfrak{h}+1},\, U \text{ is open},\, \mathcal{L}_{\mathfrak{h}}|_U \in C^1(U,\mathbb{R})} V$ it holds that

$$\mathcal{G}_{\mathfrak{h}}(\theta) = (\nabla \mathcal{L}_{\mathfrak{h}})(\theta). \tag{8.93}$$

Combining this, item (I), item (II), item (III), item (IV), and item (i) in Theorem 8.1 (applied with $d \curvearrowleft 1$, $\delta \curvearrowleft 1$, $a \curvearrowleft a$, $b \curvearrowleft b$, $(\mathbb{N} \ni \mathfrak{a} \mapsto \rho_{\mathfrak{a}} \in \mathbb{N} \cap (1,\infty)) \curvearrowleft (\mathbb{N} \ni \mathfrak{a} \mapsto 2 \in \mathbb{N} \cap (1,\infty))$, $(\mathbb{N} \ni \mathfrak{a} \mapsto \ell^{\mathfrak{a}} \in \mathbb{N}^3) \curvearrowleft (\mathbb{N} \ni \mathfrak{a} \mapsto (d, \mathfrak{a}, \delta) \in \mathbb{N}^3)$, $(\mathbb{N} \ni \mathfrak{a} \mapsto \mathfrak{d}_{\mathfrak{a}} \in \mathbb{N}) \curvearrowleft (\mathbb{N} \ni \mathfrak{a} \mapsto (3\mathfrak{a}+1) \in \mathbb{N})$, $(\mathbb{N} \ni \mathfrak{a} \mapsto \vartheta_{\mathfrak{a}} \in (\cup_{k\in\mathbb{N}} \mathbb{R}^k)) \curvearrowleft (\mathbb{N} \ni \mathfrak{a} \mapsto \vartheta_{\mathfrak{a}} \in (\cup_{k\in\mathbb{N}} \mathbb{R}^k))$ in the notation of Theorem 8.1) establishes items (i) and (ii). The proof of Corollary 8.6 is thus complete. □

Acknowledgments

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2044-390685587, Mathematics Münster: Dynamics-Geometry-Structure. This project has been partially supported by the startup fund project of Shenzhen Research Institute of Big Data under grant No. T00120220001.

References

[1] P.-A. Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM J. Optim.*, 16(2):531–547, 2005. doi:10.1137/040605266.

[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 Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332, Long Beach, California, USA, 6 2019. PMLR. URL: http://proceedings.mlr.press/v97/arora19a.html.

[3] Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.*, 116(1-2, Ser. B):5–16, 2009. doi:10.1007/s10107-007-0133-5.

[4] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Program.*, 137(1-2, Ser. A):91–129, 2013. doi:10.1007/s10107-011-0484-9.

[5] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, pages 773–781. Curran Associates, Inc., 2013. URL: http://papers.nips.cc/paper/4900-non-strongly-convex-smooth-stochastic-approximation-with-convergence-rate-o1n.pdf.

[6] Christian Beck, Arnulf Jentzen, and Benno Kuckuck. Full error analysis for the training of deep neural networks. *Infin. Dimens. Anal. Quantum Probab. Relat. Top.*, 25(2):Paper No. 2150020, 76, 2022. doi:10.1142/S021902572150020X.

[7] Dimitri P. Bertsekas and John N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000. doi:10.1137/S1052623497331063.

[8] Edward Bierstone and Pierre D. Milman. Semianalytic and subanalytic sets. *Inst. Hautes Études Sci. Publ. Math.*, 67:5–42, 1988. URL: http://www.numdam.org/item?id=PMIHES_1988__67__5_0.

[9] Jacek Bochnak, Michel Coste, and Marie-Francoise Roy. *Real algebraic geometry*, volume 36 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3)*. Springer-Verlag, Berlin, 1998. Translated from the 1987 French original, Revised by the authors. doi:10.1007/978-3-662-03718-8.

[10] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.*, 17(4):1205–1223, 2006. doi:10.1137/050644641.

[11] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning, 2018. arXiv:1606.04838.

[12] Zhengdao Chen, Grant Rotskoff, Joan Bruna, and Eric Vanden-Eijnden. A dynamical central limit theorem for shallow neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22217–22230. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper/2020/file/fc5b3186f1cf0daece964f78259b7ba0-Paper.pdf.

[13] Patrick Cheridito, Arnulf Jentzen, Adrian Riekert, and Florian Rossmannek. A proof of convergence for gradient descent in the training of artificial neural networks for constant target functions. *J. Complexity*, 72:Paper No. 101646, 2022. doi:10.1016/j.jco.2022.101646.

[14] Patrick Cheridito, Arnulf Jentzen, and Florian Rossmannek. Landscape analysis for shallow ReLU neural networks: complete classification of critical points for affine target functions, 2021. Accepted in Journal of Nonlinear Science. arXiv:2103.10922.

[15] Lénaïc Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 2021. doi:10.1007/s10107-021-01636-z.

[16] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 3036–3046. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

[17] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.

[18] Michel Coste. *An introduction to semialgebraic geometry*. Istituti editoriali e poligrafici internazionali, Pisa, 2000.

[19] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems*, 2(4):303–314, 1989. doi:10.1007/BF02551274.

[20] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D. Lee. Stochastic subgradient method converges on tame functions. *Found. Comput. Math.*, 20(1):119–154, 2020. doi:10.1007/s10208-018-09409-5.

[21] Steffen Dereich and Sebastian Kassing. Convergence of stochastic gradient descent schemes for Lojasiewicz-landscapes, 2021. arXiv:2102.09385.

[22] Steffen Dereich and Sebastian Kassing. Cooling down stochastic differential equations: almost sure convergence, 2021. arXiv:2106.03510.

[23] Steffen Dereich and Sebastian Kassing. On minimal representations of shallow ReLU networks, 2021. arXiv:2108.05643.

[24] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the*

36th International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685, Long Beach, California, USA, 6 2019. PMLR. URL: `http://proceedings.mlr.press/v97/du19c.html`.

[25] 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, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL: `https://openreview.net/forum?id=S1eK3i09YQ`.

[26] Weinan E, Chao Ma, Stephan Wojtowytsch, and Lei Wu. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't, 2020. `arXiv:2009.10713`.

[27] Weinan E, Chao Ma, and Lei Wu. A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Sci. China Math.*, 63(7):1235–1258, 2020. `doi:10.1007/s11425-019-1628-5`.

[28] Simon Eberle, Arnulf Jentzen, Adrian Riekert, and Georg S. Weiss. Existence, uniqueness, and convergence rates for gradient flows in the training of artificial neural networks with ReLU activation, 2021. `arXiv:2108.08106`.

[29] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010. `doi:10.1090/gsm/019`.

[30] Benjamin Fehrman, Benjamin Gess, and Arnulf Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *J. Mach. Learn. Res.*, 21:Paper No. 136, 48, 2020.

[31] Davide Gallon, Arnulf Jentzen, and Felix Lindner. Blow up phenomena for gradient descent optimization methods in the training of artificial neural networks, 2022.

[32] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 797–842, Paris, France, 03–06 Jul 2015. PMLR.

[33] F. Girosi and T. Poggio. Networks and the best approximation property. *Biological Cybernetics*, 63:169–176, 1990.

[34] Lars Hörmander. *The analysis of linear partial differential operators. I*. Classics in Mathematics. Springer-Verlag, Berlin, 2003. Distribution theory and Fourier analysis, Reprint of the second (1990) edition [Springer, Berlin; MR1065993 (91m:35001a)]. `doi:10.1007/978-3-642-61497-2`.

[35] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251 – 257, 1991. URL: `https://doi.org/10.1016/0893-6080(91)90009-T`, `doi:http://www.sciencedirect.com/science/article/pii/089360809190009T`.

[36] Martin Hutzenthaler, Arnulf Jentzen, Katharina Pohl, Adrian Riekert, and Luca Scarpa. Convergence proof for stochastic gradient descent in the training of deep neural networks with ReLU activation for constant target functions. 2021. Revision requested from Electronic Research Archive. `arXiv:2112.07369`.

[37] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL: `https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf`.

[38] Arnulf Jentzen and Timo Kröger. Convergence rates for gradient descent in the training of overparameterized artificial neural networks with biases, 2021. `arXiv:2102.11840`.

[39] Arnulf Jentzen, Benno Kuckuck, Ariel Neufeld, and Philippe von Wurstemberger. Strong error analysis for stochastic gradient descent optimization algorithms. *IMA Journal of Numerical Analysis*, 41(1):455–492, 2021. `doi:10.1093/imanum/drz055`.

[40] Arnulf Jentzen and Adrian Riekert. Convergence analysis for gradient flows in the training of artificial neural networks with ReLU activation, 2021. arXiv:2107.04479.

[41] Arnulf Jentzen and Adrian Riekert. A proof of convergence for stochastic gradient descent in the training of artificial neural networks with ReLU activation for constant target functions, 2021. Accepted in Zeitschrift fuer Angewandte Mathematik und Physik (ZAMP). arXiv:2104.00277.

[42] Arnulf Jentzen and Adrian Riekert. A proof of convergence for the gradient descent optimization method with random initializations in the training of neural networks with ReLU activation for piecewise linear target functions, 2021. arXiv:2108.04620.

[43] P. Kainen, V. Kůrková, and A. Vogt. Best approximation by heaviside perceptron networks. *Neural Networks*, 13(7):695–697, 2000. URL: https://www.sciencedirect.com/science/article/pii/S0893608000000563, doi:https://doi.org/10.1016/S0893-6080(00)00056-3.

[44] Tobias Kaiser. Integration of semialgebraic functions and integrated Nash functions. *Math. Z.*, 275(1-2):349–366, 2013. doi:10.1007/s00209-012-1138-1.

[45] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition, 2020. arXiv:1608.04636.

[46] Achim Klenke. *Probability theory*. Universitext. Springer, London, second edition, 2014. A comprehensive course. doi:10.1007/978-1-4471-5361-0.

[47] Krzysztof Kurdyka, Tadeusz Mostowski, and Adam Parusiński. Proof of the gradient conjecture of R. Thom. *Ann. of Math. (2)*, 152(3):763–792, 2000. doi:10.2307/2661354.

[48] Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Math. Program.*, 176(1–2):311–337, July 2019. doi:10.1007/s10107-019-01374-3.

[49] Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1246–1257, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL: http://proceedings.mlr.press/v49/lee16.html.

[50] Y. Lei, T. Hu, G. Li, and K. Tang. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10):4394–4400, 2020. doi:10.1109/TNNLS.2019.2952219.

[51] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861 – 867, 1993. URL: https://doi.org/10.1016/S0893-6080(05)80131-5, doi:http://www.sciencedirect.com/science/article/pii/S0893608005801315.

[52] Xiao Li, Andre Milzarek, and Junwen Qiu. Convergence of random reshuffling under the Kurdyka-Łojasiewicz inequality, 2021. arXiv:2110.04926.

[53] S. Łojasiewicz. Sur les trajectoires du gradient d'une fonction analytique. In *Geometry seminars, 1982–1983 (Bologna, 1982/1983)*, pages 115–117. Univ. Stud. Bologna, Bologna, 1984.

[54] Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM J. Math. Anal.*, 53(5):5465–5506, 2021. doi:10.1137/20M134695X.

[55] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, pages 451–459. Curran Associates, Inc., 2011. URL: https://proceedings.neurips.cc/paper/2011/file/40008b9a5380fcacce3976bf7c08af5b-Paper.pdf.

[56] Yurii Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course. doi:10.1007/978-1-4419-8853-9.

[57] Peter Ochs. Unifying abstract inexact convergence theorems and block coordinate variable metric iPiano. *SIAM J. Optim.*, 29(1):541–570, 2019. doi:10.1137/17M1124085.

[58] Ioannis Panageas and Georgios Piliouras. Gradient Descent Only Converges to Minimizers: Non-Isolated Critical Points and Invariant Regions. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 2:1–2:12, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. doi:10.4230/LIPIcs.ITCS.2017.2.

[59] Ioannis Panageas, Georgios Piliouras, and Xiao Wang. First-order methods almost always avoid saddle points: the case of vanishing step-sizes, 2019. arXiv:1906.07772.

[60] Vivak Patel. Stopping criteria for, and strong convergence of, stochastic gradient descent on Bottou-Curtis-Nocedal functions, 2021. arXiv:2004.00475.

[61] Philipp Petersen, Mones Raslan, and Felix Voigtlaender. Topological Properties of the Set of Functions Generated by Neural Networks of Fixed Size. *Found. Comput. Math.*, 21(2):375–444, 2021. doi:10.1007/s10208-020-09461-0.

[62] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, page 1571–1578, Madison, WI, USA, 2012. Omnipress.

[63] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 1998. doi:10.1007/978-3-642-02431-3.

[64] Sebastian Ruder. An overview of gradient descent optimization algorithms, 2017. arXiv:1609.04747.

[65] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pages 4433–4441. PMLR, 2018.

[66] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons, 2020. arXiv:1906.05497.

[67] Masahiro Shiota. *Geometry of subanalytic and semialgebraic sets*, volume 150 of *Progress in Mathematics*. Birkhäuser Boston, Inc., Boston, MA, 1997. doi:10.1007/978-1-4612-2008-4.

[68] Grzegorz Swirszcz, Wojciech Marian Czarnecki, and Razvan Pascanu. Local minima in training of neural networks, 2017. arXiv:1611.06310.

[69] Lou van den Dries and Chris Miller. Geometric categories and o-minimal structures. *Duke Math. J.*, 84(2):497–540, 1996. doi:10.1215/S0012-7094-96-08416-1.

[70] Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM J. Imaging Sci.*, 6(3):1758–1789, 2013. doi:10.1137/120887795.

[71] Jinshan Zeng, Tim Tsz-Kit Lau, Shaobo Lin, and Yuan Yao. Global convergence of block coordinate descent in deep learning, 2018. arXiv:1803.00225.

[72] Guodong Zhang, James Martens, and Roger B Grosse. Fast convergence of natural gradient descent for over-parameterized neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8082–8093. Curran Associates, Inc., 2019. URL: http://papers.nips.cc/paper/9020-fast-convergence-of-natural-gradient-descent-for-over-parameterized-neural-networks.pdf.

[73] Yaoyu Zhang, Yuqing Li, Zhongwang Zhang, Tao Luo, and Zhi-Qin John Xu. Embedding principle: a hierarchical structure of loss landscape of deep neural networks, 2021. arXiv:2111.15527.

[74] Yaoyu Zhang, Zhongwang Zhang, Tao Luo, and Zhi-Qin John Xu. Embedding principle of loss landscape of deep neural networks, 2021. arXiv:2105.14573.

[75] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109:467–492, 2020. doi:10.1007/s10994-019-05839-6.