Journal of Machine Learning Research 23 (2022) 1-50 Submitted 8/21; Revised 3/22; Published 8/22

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

Arnulf Jentzen AJENTZEN@UNI-MUENSTER.DE
School of Data Science and Shenzhen Research Institute of Big Data
The Chinese University of Hong Kong, Shenzhen, China
and
Applied Mathematics: Institute for Analysis and Numerics
University of Münster, Germany

Adrian Riekert ARIEKERT@UNI-MUENSTER.DE
Applied Mathematics: Institute for Analysis and Numerics
University of Münster, Germany

Editor: Ohad Shamir

Abstract

Gradient descent (GD) type optimization methods are the standard instrument to train artificial neural networks (ANNs) with rectified linear unit (ReLU) activation. Despite the great success of GD type optimization methods in numerical simulations for the training of ANNs with ReLU activation, it remains – even in the simplest situation of the plain vanilla GD optimization method and ANNs with one hidden layer – an open problem to prove (or disprove) the conjecture that the risk of the GD optimization method converges in the training of such ANNs to zero. In this article we establish in the situation where the probability distribution of the input data is equivalent to the continuous uniform distribution on a compact interval, where the probability distribution for the random initialization of the ANN parameters is the standard normal distribution, and where the target function under consideration is continuous and piecewise affine linear that the risk of the considered GD process converges exponentially fast to zero with a positive probability. Roughly speaking, the key ingredients in our mathematical convergence analysis are (i) to prove that suitable sets of global minima of the risk functions are *twice continuously differentiable submanifolds of the ANN parameter spaces*, (ii) to prove that the Hessians of the risk functions on these sets of global minima satisfy an appropriate *maximal rank condition*, and, thereafter, (iii) to apply the machinery in [Fehrman, B., Gess, B., Jentzen, A., Convergence rates for the stochastic gradient descent method for non-convex objective functions. J. Mach. Learn. Res. 21(136): 1–48, 2020] to establish local convergence of the GD optimization method. As a consequence, we obtain convergence of the risk to zero as the width of the ANNs, the number of independent random initializations, and the number of GD steps increase to infinity.

Keywords: Gradient descent, artificial neural networks, non-convex optimization

1. Introduction

Gradient descent (GD) type optimization methods are the standard schemes to train artificial neural networks (ANNs) with rectified linear unit (ReLU) activation; cf., e.g., (Goodfellow et al., 2016, Chapter 5). Even though GD type optimization methods seem to perform very effectively in numerical simulations, until today in general there is no mathematical convergence analysis in the literature which explains the success of GD optimization methods in the training of ANNs with ReLU activation.

There are, however, several promising mathematical analysis approaches for GD optimization methods in the scientific literature. In the case of convex objective functions, the convergence of GD type optimizations methods to the global minimum in different settings was shown, e.g., in Bach and Moulines (2013); Jentzen et al. (2021); Moulines and Bach (2011); Nesterov (2015, 2004); Rakhlin et al. (2012); Schmidt and Roux (2013).

Typically, the objective functions occurring in the training of ANNs with ReLU activation are non-convex and, instead, admit infinitely many non-global local minima and saddle points. In view of this, it becomes important to study the landscapes of the risk functions in the training of ANNs and to develop an understanding of the appearance of critical points (such as non-global local extrema and saddle points) of the risk functions. Recently, in the article Cheridito et al. (2022b) a characterization of the saddle points and non-global local minima of the risk function was obtained for the case of affine target functions. Sufficient conditions which ensure that the convergence of GD type optimization methods to saddle points can be excluded have been revealed, e.g., in Ge et al. (2015); Lee et al. (2019, 2016); Panageas and Piliouras (2017); Panageas et al. (2019).

Another promising direction of research is to study the convergence of GD type optimization methods for the training of ANNs in the so-called overparametrized regime, where the number of ANN parameters has to be sufficiently large when compared to the number of used input-output data pairs. In this situation the risks of GD type optimization methods can be shown to converge to zero with high probability; see, e.g., Arora et al. (2019); Du et al. (2019b); E et al. (2020b); Jentzen and Kröger (2021); Li and Liang (2018); Rotskoff and Vanden-Eijnden (2018); Zhang et al. (2019) for the case of ANNs with one hidden layer and see, e.g., Allen-Zhu et al. (2019a,b); Du et al. (2019a); Sankararaman et al. (2020); Zou et al. (2020) for the case of ANNs with more than one hidden layer. The results in these articles apply to the empirical risk, which is measured with respect to a finite set of input-output data pairs.

For convergence results for GD type optimization schemes without convexity but under Łojasiewicz type assumptions we point, e.g., to Absil et al. (2005); Attouch and Bolte (2009); Dereich and Kassing (2021); Karimi et al. (2020); Lei et al. (2020); Wojtowytsch (2021); Xu and Yin (2013). Further abstract convergence results for GD type optimization schemes in the non-convex setting can be found, e.g., in Akyildiz and Sabanis (2021); Bertsekas and Tsitsiklis (2000); Dereich and Müller-Gronbach (2019); Fehrman et al. (2020); Lovas et al. (2020); Patel (2021) and the references mentioned therein. In particular, the article Fehrman et al. (2020) shows convergence towards the global minimum value of some GD type optimization algorithms with random initializations, provided that the set of global minima of the objective function is locally a suitable submanifold of the parameter space and provided that the Hessian of the objective function satisfies a certain maximal rank condition at these global minima. A key contribution of this work is to demonstrate that these regularity assumptions are satisfied in the training of ANNs with one hidden layer and ReLU activation provided that the target function is piecewise affine linear.

We also refer, e.g., to Cheridito et al. (2020); Jentzen and von Wurstemberger (2020); Lu et al. (2020); Shamir (2019) for lower bounds and divergence results for GD type optimization methods. For more detailed overviews and further literature on GD type optimization schemes we point, e.g., to Bercu and Fort (2013), Bottou et al. (2018), E et al. (2020a), (Fehrman et al., 2020, Section 1.1), (Jentzen et al., 2021, Section 1), and Ruder (2017).

There are different variants of GD type optimization methods in the scientific literature, such as the plain vanilla GD optimization method, GD optimization methods with momentum, and adaptive GD optimization methods (cf., e.g., Ruder (2017)), and the plain vanilla GD optimization method with independent random initializations is maybe the GD based ANN training scheme which is most accessible for a mathematical convergence analysis. Despite the above mentioned promising mathematical analysis approaches in the literature, it remains – even in the simple situation of the plain vanilla GD optimization method with independent random initializations and ANNs with one hidden layer and ReLU activation – an open problem to prove (or disprove) the conjecture that the risk of the GD optimization method converges to the risk of the global minima of the risk function in the training of such ANNs. It is one of the key contributions of this article to establish convergence of the plain vanilla GD optimization method with multiple independent random initializations for ANNs with one hidden layer and ReLU activation in the situation where the probability distribution of the input data is equivalent to the continuous uniform distribution on a compact interval with a Lipschitz continuous density, where the probability distributions for the random initializations of the ANN parameters are standard normal distributions, and where the target function under consideration is continuous and piecewise affine linear. Specifically, we obtain convergence of the risk to zero as the width of the ANNs, the number of independent random initializations, and the number of GD steps increase to infinity. The precise formulation of this statement is given in Theorem 2 below within this introductory section.

Theorem 2 is a consequence of the first main result of this article, Theorem 1 below, which establishes a local convergence result for the plain vanilla GD optimization method with a single random initialization. Specifically, in Theorem 1 we prove in the case of ANNs with three layers (1-dimensional input layer, H-dimensional hidden layer, and 1-dimensional output layer) and in the case of a continuous and piecewise affine linear target function $f\colon [a,b] \to \mathbb{R}$ with $N \in \mathbb{N} \cap [1,H]$ grid points that for every sufficiently small learning rate γ we have that the risk of the considered GD process with learning rate γ and standard normal random initialization (see (1.2) in Theorem 1) converges exponentially to zero with a positive probability (see (1.3) in Theorem 1). We now present the precise statement of Theorem 1 in a self-contained style and, thereafter, we provide some additional explanations regarding the mathematical objects in Theorem 1.

Theorem 1 *Let $H, \mathfrak{d} \in \mathbb{N}$, $N \in \mathbb{N} \cap [1,H]$, $x_0, x_1, \ldots, x_N, a \in \mathbb{R}$, $b \in (a,\infty)$, $f \in C([a,b],\mathbb{R})$ satisfy $\mathfrak{d} = 3H+1$ and $a = x_0 < x_1 < \cdots < x_N = b$, assume for all $i \in \{1,2,\ldots,N\}$ that $f|_{[x_{i-1},x_i]}$ is affine linear, let $\mathfrak{p}\colon [a,b] \to (0,\infty)$ be Lipschitz con-*

tinuous, 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*

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

let $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space, let* $\Theta_n^\gamma\colon \Omega \to \mathbb{R}^{\mathfrak{d}}$*,* $\gamma \in \mathbb{R}$*,* $n \in \mathbb{N}_0$*, be random variables, assume for every* $\gamma \in \mathbb{R}$ *that* Θ_0^γ *is standard normally distributed, let* $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ *satisfy for all* $\theta \in \{\vartheta \in \mathbb{R}^{\mathfrak{d}}\colon \mathcal{L} \text{ is differentiable at } \vartheta\}$ *that* $\mathcal{G}(\theta) = (\nabla\mathcal{L})(\theta)$*, and assume for all* $\gamma \in \mathbb{R}$*,* $n \in \mathbb{N}_0$*,* $\omega \in \Omega$ *that*

$$\Theta_{n+1}^\gamma(\omega) = \Theta_n^\gamma(\omega) - \gamma\mathcal{G}(\Theta_n^\gamma(\omega)). \tag{1.2}$$

Then there exist $\mathfrak{c}, \mathfrak{C} \in (0,\infty)$ *such that for all* $\gamma \in (0, \mathfrak{c}]$ *it holds that*

$$\mathbb{P}\bigl(\limsup\nolimits_{n\to\infty} \mathcal{L}(\Theta_n^\gamma) = 0\bigr) \geq \mathbb{P}\bigl(\forall\, n \in \mathbb{N}_0\colon \mathcal{L}(\Theta_n^\gamma) \leq \mathfrak{C}\exp(-\mathfrak{c}\gamma n)\bigr) \geq \mathfrak{c} > 0. \tag{1.3}$$

Theorem 1 is an immediate consequence of Corollary 49 below (applied with $\rho \curvearrowleft 0$ in the notation of Corollary 49). Corollary 49, in turn, is a direct consequence of Theorem 48 in Subsection 5.2 below, which is the main result of this article.

In Theorem 1 the target function (the function which describes the relationship between the input and the output data in the considered supervised learning problem) is described through the continuous function $f \in C([a,b], \mathbb{R})$ from the compact interval $[a,b]$ to the real numbers $\mathbb{R}$. In Theorem 1 this target function $f\colon [a,b] \to \mathbb{R}$ is assumed to be piecewise affine linear in the sense that there exist $N \in \mathbb{N}$, $\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}$ with

$$a = \boldsymbol{x}_0 < \boldsymbol{x}_1 < ... < \boldsymbol{x}_N = b \tag{1.4}$$

so that for all $i \in \{1, 2, \dots, N\}$ we have that the target function $[\boldsymbol{x}_{i-1}, \boldsymbol{x}_i] \ni x \mapsto f(x) \in \mathbb{R}$ restricted to the sub-interval $[\boldsymbol{x}_{i-1}, \boldsymbol{x}_i]$ is affine linear; see above (1.1) in Theorem 1.

In Theorem 1 we also assume that the probability distribution of the input data in the considered supervised learning problem is absolutely continuous with respect to the standard uniform distribution on $[a,b]$ with a Lipschitz continuous, strictly positive density. More specifically, the Lipschitz continuous function $\mathfrak{p}\colon [a,b] \to (0,\infty)$ in Theorem 1 is assumed to be an unnormalized density of the probability distribution of the input data with respect to the Lebesgue measure restricted to $[a,b]$.

In (1.1) in Theorem 1 we consider fully connected feedforward ANNs with the ReLU activation $\mathbb{R} \ni x \mapsto \max\{x,0\} \in \mathbb{R}$ and three layers: one input layer with 1 neuron on the input layer (1-dimensional input), one hidden layer with $H \in \mathbb{N}$ neurons on the hidden layer (H-dimensional hidden layer), and one output layer with 1 neuron on the output layer (1-dimensional output). In particular, we describe in (1.1) the risk function $\mathcal{L}\colon \mathbb{R}^{3H+1} \to \mathbb{R}$ associated to the supervised learning problem considered in Theorem 1.

The risk function $\mathcal{L}\colon \mathbb{R}^{3H+1} \to \mathbb{R}$ fails to be continuously differentiable due to the lack of differentiability of the ReLU activation function $\mathbb{R} \ni x \mapsto \max\{x,0\} \in \mathbb{R}$ and, in view of this, one needs to introduce appropriate generalized gradients of the risk function which mathematically describe the behavior of GD steps in implementations in numerical simulations to mathematically formulate the GD optimization method for the training of ANNs with ReLU activation. The function $\mathcal{G}\colon \mathbb{R}^{3H+1} \to \mathbb{R}^{3H+1}$ in Theorem 1 describes a generalized gradient function of the risk function $\mathcal{L}\colon \mathbb{R}^{3H+1} \to \mathbb{R}$ in (1.1). We assume for every $\theta \in \mathbb{R}^{3H+1}$ at which $\mathcal{L}$ is differentiable that $\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta)$, while the value of $\mathcal{G}(\theta)$ at the remaining points $\theta \in \mathbb{R}^{3H+1}$ is not specified. In Proposition 11 below we identify a suitable open set $\mathfrak{V} \subseteq \mathbb{R}^{3H+1}$ with full Lebesgue measure on which $\mathcal{L}$ is continuously differentiable, and we also derive an explicit representation for $\mathcal{G}(\theta)$ for every $\theta \in \mathfrak{V}$.

For every learning rate $\gamma \in \mathbb{R}$ the random variables $\Theta_n \colon \Omega \to \mathbb{R}^{3H+1}$, $n \in \mathbb{N}_0$, in (1.2) describe the GD process with learning rate γ and standard normal initialization $\Theta_0^\gamma \colon \Omega \to \mathbb{R}^{3H+1}$. The conclusion of Theorem 1 in (1.3) demonstrates for every sufficiently small positive learning rate $\gamma \in (0, \mathfrak{c}]$ that the risk of the GD process converges exponentially fast to zero with a strictly positive probability.

Roughly speaking, we prove Theorem 1 and Theorem 48, respectively, (i) by showing that for every number $H \in \mathbb{N} \cap [N, \infty)$ of neurons on the hidden layer there exists a natural number $k \in \mathbb{N} \cap [1, \mathfrak{d})$ such that a suitable subset of the set of global minima of the risk function $\mathcal{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ in (1.1) is a twice continuously differentiable k-dimensional submanifold of the ANN parameter space $\mathbb{R}^{\mathfrak{d}} = \mathbb{R}^{3H+1}$ (cf. Lemma 21 and Corollary 29 in Section 3 below), (ii) by proving that the ranks of the Hessian matrices of the risk function on this suitable set of global minima of the risk function $\mathcal{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ in (1.1) are equal to $\mathfrak{d} - k$, and, thereafter, (iii) by applying the machinery in Fehrman et al. (2020) to establish convergence of the GD optimization method provided that the initial value is contained in a suitable open neighborhood of certain global minima.

As a consequence of Theorem 1, we obtain in Theorem 2 below that the plain vanilla GD optimization method with multiple independent random initializations converges in the training of ANNs with one hidden layer and ReLU activation (under the same assumptions on the target function and the probability distribution of the input data as in Theorem 1). We now present the precise statement of Theorem 2.

Theorem 2 *Let $N \in \mathbb{N}$, $x_0, x_1, \ldots, x_N, a \in \mathbb{R}$, $b \in (a, \infty)$, $f \in C([a,b], \mathbb{R})$ satisfy $a = x_0 < x_1 < \cdots < x_N = b$, assume for all $i \in \{1, 2, \ldots, N\}$ that $f|_{[x_{i-1}, x_i]}$ is affine linear, let $\mathfrak{p} \colon [a,b] \to (0, \infty)$ be Lipschitz continuous, let $\mathcal{L}^H \colon \mathbb{R}^{3H+1} \to \mathbb{R}$, $H \in \mathbb{N}$, satisfy for all $H \in \mathbb{N}$, $\theta = (\theta_1, \ldots, \theta_{3H+1}) \in \mathbb{R}^{3H+1}$ that*

$$\mathcal{L}^H(\theta) = \int_a^b \big(f(x) - \theta_{\mathfrak{d}} - \textstyle\sum_{j=1}^H \theta_{2H+j} [\max\{\theta_j x + \theta_{H+j}, 0\}] \big)^2 \mathfrak{p}(x) \, \mathrm{d}x, \tag{1.5}$$

let $\mathcal{G}^H \colon \mathbb{R}^{3H+1} \to \mathbb{R}^{3H+1}$, $H \in \mathbb{N}$, satisfy for all $H \in \mathbb{N}$, $\theta \in \{\vartheta \in \mathbb{R}^{3H+1} \colon \mathcal{L}^H$ is differentiable at $\vartheta\}$ that $\mathcal{G}^H(\theta) = (\nabla \mathcal{L}^H)(\theta)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\Theta_n^{H,k,\gamma} \colon \Omega \to \mathbb{R}^{3H+1}$, $H, k \in \mathbb{N}$, $\gamma \in \mathbb{R}$, $n \in \mathbb{N}_0$, and $\mathbf{k}_n^{H,k,\gamma} \colon \Omega \to \mathbb{N}$, $H, k \in \mathbb{N}$, $\gamma \in \mathbb{R}$, $n \in \mathbb{N}_0$, be random variables, assume for all $H \in \mathbb{N}$, $\gamma \in \mathbb{R}$ that $\Theta_0^{H,k,\gamma}$, $k \in \mathbb{N}$, are independent standard normal random vectors, and assume for all $H, k \in \mathbb{N}$, $\gamma \in \mathbb{R}$, $n \in \mathbb{N}_0$, $\omega \in \Omega$ that

$$\Theta_{n+1}^{H,k,\gamma}(\omega) = \Theta_n^{H,k,\gamma}(\omega) - \gamma \mathcal{G}^H(\Theta_n^{H,k,\gamma}(\omega)) \tag{1.6}$$

and

$$\mathbf{k}_n^{H,k,\gamma}(\omega) \in \arg\min\nolimits_{\ell \in \{1,2,\ldots,k\}} \mathcal{L}^H(\Theta_n^{H,\ell,\gamma}(\omega)). \tag{1.7}$$

Then there exists $\mathfrak{g} \in (0, \infty)$ such that for all $\gamma \in (0, \mathfrak{g}]$ it holds that

$$\liminf\nolimits_{H \to \infty} \liminf\nolimits_{K \to \infty} \mathbb{P}\Big(\limsup\nolimits_{n \to \infty} \mathcal{L}^H\big(\Theta_n^{H, \mathbf{k}_n^{H,K,\gamma}, \gamma}\big) = 0 \Big) = 1. \tag{1.8}$$

Theorem 2 is a direct consequence of Corollary 50 below. Corollary 50, in turn, follows from Theorem 48 (see Subsection 5.3 below for details).

In Theorem 2 we have for every number $H \in \mathbb{N}$ of neurons on the hidden layer, every natural number $k \in \mathbb{N}$, and every learning rate $\gamma \in \mathbb{R}$ that the random variables $\Theta_n^{H,k,\gamma} \colon \Omega \to \mathbb{R}^{3H+1}$, $n \in \mathbb{N}_0$, in (1.6) describe the GD process with learning rate γ. Note that the natural number $k \in \mathbb{N}$ counts the number of random initializations, while the index $n \in \mathbb{N}_0$ specifies the current gradient step. Observe that the assumption in Theorem 2 that for all $H \in \mathbb{N}$, $\gamma \in \mathbb{R}$ it holds that $\Theta_0^{H,k,\gamma} \colon \Omega \to \mathbb{R}^{3H+1}$, $k \in \mathbb{N}$, are i.i.d. random variables ensures for all $H \in \mathbb{N}$, $n \in \mathbb{N}_0$, $\gamma \in \mathbb{R}$ that the random variables $\Theta_n^{H,k,\gamma} \colon \Omega \to \mathbb{R}^{3H+1}$, $k \in \mathbb{N}$, are i.i.d. Loosely speaking, for every number $H \in \mathbb{N}$ of neurons on the hidden layer, every natural number $k \in \mathbb{N}$, every learning rate $\gamma \in \mathbb{R}$, and every number $n \in \mathbb{N}$ of GD steps we have that the random variable $\mathbf{k}_n^{H,k,\gamma} \colon \Omega \to \mathbb{N}$ in (1.7) selects an independent random initialization with the smallest risk. The conclusion of Theorem 2, equation (1.8), reveals that there exists a sufficiently small strictly positive real number $\mathfrak{g} \in (0, \infty)$ such that for every learning rate $\gamma \in (0, \mathfrak{g}]$ we have as the number $K \in \mathbb{N}$ of independent random realizations and the number $H \in \mathbb{N}$ of neurons on the hidden layer increase to infinity convergence to one of the probability that the risk of the GD optimization method with independent standard normal random initializations converges to zero.

The remainder of this article is organized as follows. In Section 2 we establish several regularity properties for the Hessian matrix of the risk function of the considered supervised learning problem. In Section 3 we employ the findings from Section 2 to establish that a suitable subset of the set of global minima of the risk function constitutes a C^∞-submanifold of the ANN parameter space $\mathbb{R}^{\mathfrak{d}} = \mathbb{R}^{3H+1}$ on which the Hessian matrix of the risk function has maximal rank. In Section 4 we engage the findings from Section 3 to establish that the risk of certain solutions of GF differential equations converges exponentially quick to zero. Finally, in Section 5 we establish that the risk of certain GD processes converges exponentially quick to zero and, thereby, we also prove Theorems 1 and 2 above.

2. Second order differentiability properties of the risk function

In this section we establish in Lemma 14 in Subsection 2.4 below an explicit representation result for the Hessian matrix of the risk function of the considered supervised learning problem. In particular, in Lemma 14 we identify a suitable open subset of the ANN parameter space with full Lebesgue measure on which the risk function is twice continuously differentiable (see (2.3) below for details). This is nontrivial due to the fact that the ReLU activation function $\mathbb{R} \ni x \mapsto \max\{x, 0\} \in \mathbb{R}$ is not everywhere differentiable. Results related to Lemma 14 have been shown in (Cheridito et al., 2022b, Lemma 3.8).

Corollary 16 in Subsection 2.4 specializes Lemma 14 to the specific situation where the ANN parameter represents a global minima of the risk function. In Lemma 15 in Subsection 2.4 we employ Lemma 14 to conclude under the assumption that the target function is locally Lipschitz continuous that the second derivative of the risk function is locally Lipschitz continuous. In Lemma 17, Lemma 18, and Corollary 19 in Subsection 2.5 below we use Lemma 14 to derive suitable upper bounds for the absolute values of the second order partial derivatives of the risk function. Lemma 15, Corollary 16, and Corollary 19 are all employed in Section 3 below.

Our proof of Lemma 14 employs the well-known Leibniz integral rule type result in Lemma 13 in Subsection 2.4, the known representation and regularity result for the first derivative of the risk function in Proposition 11 in Subsection 2.4, the elementary continuity result in Lemma 12 in Subsection 2.4, the elementary and well-known differentiability result for certain parameter integrals in Lemma 4 in Subsection 2.2 below, and the elementary continuity result for certain parameter integrals involving indicator functions in Lemma 5 in Subsection 2.2 and Corollary 9 in Subsection 2.3 below. Proposition 11 is a direct consequence of Propositions 2.2 and 2.11 in Jentzen and Riekert (2022); see also Proposition 2.4 in Cheridito et al. (2022a) for a similar result. Our proof of Lemma 15 also uses the local Lipschitz continuity results for certain parameter integrals involving indicator functions in Corollary 10 in Subsection 2.3. Our proofs of Corollaries 9 and 10, in turn, employ the elementary Lipschitz continuity result for certain parameter integrals involving indicator functions in Lemma 6 in Subsection 2.2 as well as the local Lipschitz continuity results for active neuron regions in Lemma 7 and Corollary 8 in Subsection 2.3.

2.1 Mathematical description of artificial neural networks (ANNs)

Setting 3 *Let* $H, \mathfrak{d} \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in (a, \infty)$, $f \in C([a,b], \mathbb{R})$ *satisfy* $\mathfrak{d} = 3H + 1$, *let* $\mathfrak{w} = ((\mathfrak{w}^\theta_1, \dots, \mathfrak{w}^\theta_H))_{\theta \in \mathbb{R}^\mathfrak{d}} \colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}^H$, $\mathfrak{b} = ((\mathfrak{b}^\theta_1, \dots, \mathfrak{b}^\theta_H))_{\theta \in \mathbb{R}^\mathfrak{d}} \colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}^H$, $\mathfrak{v} = ((\mathfrak{v}^\theta_1, \dots, \mathfrak{v}^\theta_H))_{\theta \in \mathbb{R}^\mathfrak{d}} \colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}^H$, $\mathfrak{c} = (\mathfrak{c}^\theta)_{\theta \in \mathbb{R}^\mathfrak{d}} \colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}$, *and* $\mathfrak{q} = ((\mathfrak{q}^\theta_1, \dots, \mathfrak{q}^\theta_H)) \colon \mathbb{R}^\mathfrak{d} \to (-\infty, \infty]^H$ *satisfy for all* $\theta = (\theta_1, \dots, \theta_\mathfrak{d}) \in \mathbb{R}^\mathfrak{d}$, $j \in \{1, 2, \dots, H\}$ *that* $\mathfrak{w}^\theta_j = \theta_j$, $\mathfrak{b}^\theta_j = \theta_{H+j}$, $\mathfrak{v}^\theta_j = \theta_{2H+j}$, $\mathfrak{c}^\theta = \theta_\mathfrak{d}$, *and*

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

let $\mathfrak{p} \colon [a,b] \to (0, \infty)$ *be Lipschitz continuous, let* $\mathfrak{R} \colon \mathbb{R} \to \mathbb{R}$, $\mathscr{N} = (\mathscr{N}^\theta)_{\theta \in \mathbb{R}^\mathfrak{d}} \colon \mathbb{R}^\mathfrak{d} \to C(\mathbb{R}, \mathbb{R})$, *and* $\mathcal{L} \colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}$ *satisfy for all* $\theta \in \mathbb{R}^\mathfrak{d}$, $x \in \mathbb{R}$ *that* $\mathfrak{R}(x) = \max\{x, 0\}$, $\mathscr{N}^\theta(x) = \mathfrak{c}^\theta + \sum_{j=1}^H \mathfrak{v}^\theta_j [\mathfrak{R}(\mathfrak{w}^\theta_j x + \mathfrak{b}^\theta_j)]$, *and*

$$\mathcal{L}(\theta) = \int_a^b (\mathscr{N}^\theta(y) - f(y))^2 \mathfrak{p}(y) \, \mathrm{d}y, \tag{2.2}$$

let $I^\theta_j \subseteq \mathbb{R}$, $\theta \in \mathbb{R}^\mathfrak{d}$, $j \in \{1, 2, \dots, H\}$, *satisfy for all* $\theta \in \mathbb{R}^\mathfrak{d}$, $j \in \{1, 2, \dots, H\}$ *that* $I^\theta_j = \{x \in [a,b] \colon \mathfrak{w}^\theta_j x + \mathfrak{b}^\theta_j > 0\}$, *let* $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_\mathfrak{d}) \colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}^\mathfrak{d}$ *satisfy for all* $\theta \in \{\vartheta \in \mathbb{R}^\mathfrak{d} \colon \mathcal{L}$ *is differentiable at* $\vartheta\}$ *that* $\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta)$, *and let* $\mathfrak{V} \subseteq \mathbb{R}^\mathfrak{d}$ *satisfy*

$$\mathfrak{V} = \big\{\theta \in \mathbb{R}^\mathfrak{d} \colon \big(\textstyle\prod_{j=1}^H \prod_{v \in \{a,b\}} (\mathfrak{w}^\theta_j v + \mathfrak{b}^\theta_j) \neq 0\big)\big\}. \tag{2.3}$$

2.2 Regularity properties for parametric integrals of Lipschitz continuous functions

Lemma 4 *Let* $n \in \mathbb{N}$, $j \in \{1, 2, \dots, n\}$, $\mathfrak{u} \in \mathbb{R}$, $\mathfrak{v} \in (\mathfrak{u}, \infty)$, *let* $\phi \colon \mathbb{R}^n \times [\mathfrak{u}, \mathfrak{v}] \to \mathbb{R}$ *be locally bounded and measurable, let* $\mu \colon \mathcal{B}([\mathfrak{u}, \mathfrak{v}]) \to [0, \infty]$ *be*[1] *a finite measure, let* $\Phi \colon \mathbb{R}^n \to \mathbb{R}$ *satisfy for all* $x \in \mathbb{R}^n$ *that*

$$\Phi(x) = \int_\mathfrak{u}^\mathfrak{v} \phi(x, s) \, \mu(\mathrm{d}s), \tag{2.4}$$

1. For all $u \in \mathbb{R}$, $v \in (u, \infty)$ we denote by $\mathcal{B}([u,v])$ the Borel sigma algebra on $[u,v]$, i.e., the smallest σ-algebra which contains all open subsets of $[u,v]$.

let $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, $\delta, c \in (0, \infty)$ *satisfy for all* $s \in [\mathfrak{u}, \mathfrak{v}]$, $h \in (-\delta, \delta)$ *that*

$$|\phi(x_1, \dots, x_{j-1}, x_j + h, x_{j+1}, \dots, x_n, s) - \phi(x, s)| \le c|h|, \tag{2.5}$$

let $E \subseteq [\mathfrak{u}, \mathfrak{v}]$ *be measurable, assume* $\mu([\mathfrak{u}, \mathfrak{v}] \backslash E) = 0$, *and assume for all* $s \in E$ *that* $\mathbb{R} \ni v \mapsto \phi(x_1, \dots, x_{j-1}, v, x_{j+1}, \dots, x_n, s) \in \mathbb{R}$ *is differentiable at* x_j. *Then*

(i) it holds that $\mathbb{R} \ni v \mapsto \Phi(x_1, \dots, x_{j-1}, v, x_{j+1}, \dots, x_n) \in \mathbb{R}$ *is differentiable at* x_j *and*

(ii) it holds that

$$\big(\tfrac{\partial}{\partial x_j} \Phi\big)(x_1, \dots, x_n) = \int_E \big(\tfrac{\partial}{\partial x_j} \phi\big)(x_1, \dots, x_n, s)\, \mu(\mathrm{d}s). \tag{2.6}$$

Proof [Proof of Lemma 4] Observe that (2.5) and the dominated convergence theorem establish items (i) and (ii). The proof of Lemma 4 is thus complete. ■

Lemma 5 *Let* $n \in \mathbb{N}$, $\mathfrak{u} \in \mathbb{R}$, $\mathfrak{v} \in (\mathfrak{u}, \infty)$, $x \in \mathbb{R}^n$, $c, \varepsilon \in (0, \infty)$, $\phi \in C(\mathbb{R}^n \times [\mathfrak{u}, \mathfrak{v}], \mathbb{R})$, *let* $\mu \colon \mathcal{B}([\mathfrak{u}, \mathfrak{v}]) \to [0, \infty]$ *be a finite measure, let* $I^y \in \mathcal{B}([\mathfrak{u}, \mathfrak{v}])$, $y \in \mathbb{R}^n$, *satisfy*[2] *for all* $y, z \in \{v \in \mathbb{R}^n \colon \|x - v\| \le \varepsilon\}$ *that* $\mu(I^y \Delta I^z) \le c\|y - z\|$, *and let* $\Phi \colon \mathbb{R}^n \to \mathbb{R}$ *satisfy for all* $y \in \mathbb{R}^n$ *that*

$$\Phi(y) = \int_{I^y} \phi(y, s)\, \mu(\mathrm{d}s). \tag{2.7}$$

Then it holds that $\{v \in \mathbb{R}^n \colon \|x - v\| \le \varepsilon\} \ni y \mapsto \Phi(y) \in \mathbb{R}$ *is continuous.*

Proof [Proof of Lemma 5] Throughout this proof let $y \in \{v \in \mathbb{R}^n \colon \|x - v\| \le \varepsilon\}$ and let $z = (z_k)_{k \in \mathbb{N}} \colon \mathbb{N} \to \{v \in \mathbb{R}^n \colon \|x - v\| \le \varepsilon\}$ satisfy $\limsup_{k \to \infty} \|z_k - y\| = 0$. Note that for all $k \in \mathbb{N}$ it holds that

$$\begin{aligned} |\Phi(y) - \Phi(z_k)| \le \int_{I^y \cap I^{z_k}} |\phi(y, s) - \phi(z_k, s)|\, \mu(\mathrm{d}s) + \int_{I^y \backslash I^{z_k}} |\phi(y, s)|\, \mu(\mathrm{d}s) \\ + \int_{I^{z_k} \backslash I^y} |\phi(z_k, s)|\, \mu(\mathrm{d}s). \end{aligned} \tag{2.8}$$

Next observe that the assumption that ϕ is continuous and the dominated convergence theorem demonstrate that

$$\limsup_{k \to \infty} \left[\int_{I^y \cap I^{z_k}} |\phi(y, s) - \phi(z_k, s)|\, \mu(\mathrm{d}s) \right] = 0. \tag{2.9}$$

Moreover, note that the fact that for all $k \in \mathbb{N}$ it holds that $\mu(I^y \Delta I^{z_k}) \le c\|y - z_k\|$ and the assumption that ϕ is continuous prove that for all $k \in \mathbb{N}$ we have that

$$\limsup_{k \to \infty} \left[\int_{I^y \backslash I^{z_k}} |\phi(y, s)|\, \mu(\mathrm{d}s) + \int_{I^{z_k} \backslash I^y} |\phi(z_k, s)|\, \mu(\mathrm{d}s) \right] = 0. \tag{2.10}$$

2. Throughout this article we denote by $\|\cdot\| \colon \big(\bigcup_{n \in \mathbb{N}} \mathbb{R}^n\big) \to \mathbb{R}$ and $\langle \cdot, \cdot \rangle \colon \big(\bigcup_{n \in \mathbb{N}} (\mathbb{R}^n \times \mathbb{R}^n)\big) \to \mathbb{R}$ 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 $\|x\| = \big[\sum_{i=1}^n |x_i|^2\big]^{1/2}$ and $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$.

Combining this with (2.8) and (2.9) establishes that $\limsup_{k\to\infty}|\Phi(y)-\Phi(z_k)|=0$. The proof of Lemma 5 is thus complete. ■

Lemma 6 *Let $n\in\mathbb{N}$, $\mathfrak{u}\in\mathbb{R}$, $\mathfrak{v}\in(\mathfrak{u},\infty)$, $x\in\mathbb{R}^n$, $c,\varepsilon\in(0,\infty)$, let $\phi\colon\mathbb{R}^n\times[\mathfrak{u},\mathfrak{v}]\to\mathbb{R}$ be locally Lipschitz continuous, let $\mu\colon\mathcal{B}([\mathfrak{u},\mathfrak{v}])\to[0,\infty]$ be a finite measure, let $I^y\in\mathcal{B}([\mathfrak{u},\mathfrak{v}])$, $y\in\mathbb{R}^n$, satisfy for all $y,z\in\{v\in\mathbb{R}^n\colon\|x-v\|\le\varepsilon\}$ that $\mu(I^y\Delta I^z)\le c\|y-z\|$, and let $\Phi\colon\mathbb{R}^n\to\mathbb{R}$ satisfy for all $y\in\mathbb{R}^n$ that*

$$\Phi(y)=\int_{I^y}\phi(y,s)\,\mu(\mathrm{d}s). \tag{2.11}$$

Then there exists $\mathfrak{C}\in\mathbb{R}$ such that for all $y,z\in\{v\in\mathbb{R}^n\colon\|x-v\|\le\varepsilon\}$ it holds that $|\Phi(y)-\Phi(z)|\le\mathfrak{C}\|y-z\|$.

Proof [Proof of Lemma 6] Observe that the assumption that ϕ is locally Lipschitz continuous ensures that there exists $\mathfrak{C}\in\mathbb{R}$ which satisfies for all $y,z\in\{v\in\mathbb{R}^n\colon\|x-v\|\le\varepsilon\}$, $s\in[\mathfrak{u},\mathfrak{v}]$ with $y\neq z$ that

$$\tfrac{|\phi(y,s)-\phi(z,s)|}{\|y-z\|}+|\phi(y,s)|+|\phi(z,s)|\le\mathfrak{C}. \tag{2.12}$$

Furthermore, note that (2.11) ensures for all $y,z\in\mathbb{R}^n$ that

$$|\Phi(y)-\Phi(z)|\le\int_{I^y\cap I^z}|\phi(y,s)-\phi(z,s)|\,\mu(\mathrm{d}s)+\int_{I^y\setminus I^z}|\phi(y,s)|\,\mu(\mathrm{d}s)+\int_{I^z\setminus I^y}|\phi(z,s)|\,\mu(\mathrm{d}s). \tag{2.13}$$

In addition, observe that (2.12) shows for all $y,z\in\{v\in\mathbb{R}^n\colon\|x-v\|\le\varepsilon\}$ that

$$\int_{I^y\cap I^z}|\phi(y,s)-\phi(z,s)|\,\mu(\mathrm{d}s)\le\mathfrak{C}\|y-z\|\mu([\mathfrak{u},\mathfrak{v}]). \tag{2.14}$$

Moreover, note that (2.12) and the assumption that for all $y,z\in\{v\in\mathbb{R}^n\colon\|x-v\|\le\varepsilon\}$ it holds that $\mu(I^y\Delta I^z)\le c\|y-z\|$ prove that for all $y,z\in\{v\in\mathbb{R}^n\colon\|x-v\|\le\varepsilon\}$ we have that

$$\int_{I^y\setminus I^z}|\phi(y,s)|\,\mu(\mathrm{d}s)+\int_{I^z\setminus I^y}|\phi(z,s)|\,\mu(\mathrm{d}s)\le c\mathfrak{C}\|y-z\|. \tag{2.15}$$

Combining this with (2.13) and (2.14) establishes for all $y,z\in\{v\in\mathbb{R}^n\colon\|x-v\|\le\varepsilon\}$ that

$$|\Phi(y)-\Phi(z)|\le\mathfrak{C}(c+\mu([\mathfrak{u},\mathfrak{v}]))\|y-z\|. \tag{2.16}$$

The proof of Lemma 6 is thus complete. ■

2.3 Local Lipschitz continuity for active neuron regions

Lemma 7 *Let $a \in \mathbb{R}$, $b \in (a, \infty)$, $u = (u_1, u_2) \in \mathbb{R}^2 \backslash \{0\}$, let $\mathfrak{p} \colon [a, b] \to \mathbb{R}$ be bounded and measurable, and let $I^v \subseteq \mathbb{R}$, $v \in \mathbb{R}^2$, satisfy for all $v = (v_1, v_2) \in \mathbb{R}^2$ that $I^v = \{x \in [a, b] \colon v_1 x + v_2 > 0\}$. Then there exist $c, \varepsilon \in (0, \infty)$ such that for all $v, w \in \mathbb{R}^2$ with $\max\{\|u - v\|, \|u - w\|\} \leq \varepsilon$ it holds that*

$$\left| \int_{I^v \Delta I^w} \mathfrak{p}(x) \, \mathrm{d}x \right| \leq c \|v - w\|. \tag{2.17}$$

Proof [Proof of Lemma 7] Throughout this proof let $M \in \mathbb{R}$ satisfy $M = \sup_{x \in [a,b]} |\mathfrak{p}(x)|$. In the following we distinguish between the case $u_1 = 0$ and the case $u_1 \neq 0$.

We first prove (2.17) in the case

$$u_1 = 0. \tag{2.18}$$

Observe that (2.18) and the assumption that $u = (u_1, u_2) \in \mathbb{R}^2 \backslash \{0\}$ imply that $u_2 \neq 0$. Moreover, note that (2.18) shows for all $v = (v_1, v_2) \in \mathbb{R}^2$, $x \in I^u \Delta I^v$ that

$$|(u_1 x + u_2) - (v_1 x + v_2)| = |u_1 x + u_2| + |v_1 x + v_2| \geq |u_1 x + u_2| = |u_2|. \tag{2.19}$$

In addition, observe that for all $v = (v_1, v_2) \in \mathbb{R}^2$, $x \in [a, b]$ we have that

$$|(u_1 x + u_2) - (v_1 x + v_2)| \leq |u_1 - v_1||x| + |u_2 - v_2| \leq (1 + \max\{|a|, |b|\}) \|u - v\|. \tag{2.20}$$

Combining this with (2.19) demonstrates for all $v \in \mathbb{R}^2$ with $\|u - v\| < \frac{|u_2|}{1 + \max\{|a|, |b|\}}$ that $I^u \Delta I^v = \varnothing$ and, therefore, $I^u = I^v$. Hence, we obtain for all $v, w \in \mathbb{R}^2$ with $\max\{\|u - v\|, \|u - w\|\} \leq \frac{|u_2|}{2 + \max\{|a|, |b|\}}$ that $I^v = I^u = I^w$ and, therefore, $\int_{I^v \Delta I^w} \mathfrak{p}(x) \, \mathrm{d}x = 0$. This establishes (2.17) in the case $u_1 = 0$.

In the next step we prove (2.17) in the case $u_1 \neq 0$. Note that for all $v = (v_1, v_2)$, $w = (w_1, w_2) \in \mathbb{R}^2$, $\mathfrak{s} \in \{-1, 1\}$ with $\min\{\mathfrak{s} v_1, \mathfrak{s} w_1\} > 0$ it holds that

$$\begin{aligned} I^v \backslash I^w &= \{y \in [a, b] \colon v_1 y + v_2 > 0 \geq w_1 y + w_2\} = \left\{ y \in [a, b] \colon -\tfrac{\mathfrak{s} v_2}{v_1} < \mathfrak{s} y \leq -\tfrac{\mathfrak{s} w_2}{w_1} \right\} \\ &\subseteq \left\{ y \in \mathbb{R} \colon -\tfrac{\mathfrak{s} v_2}{v_1} < \mathfrak{s} y \leq -\tfrac{\mathfrak{s} w_2}{w_1} \right\}. \end{aligned} \tag{2.21}$$

Hence, we obtain for all $v = (v_1, v_2)$, $w = (w_1, w_2) \in \mathbb{R}^2$, $\mathfrak{s} \in \{-1, 1\}$ with $\min\{\mathfrak{s} v_1, \mathfrak{s} w_1\} > 0$ that

$$\int_{I^v \backslash I^w} 1 \, \mathrm{d}x \leq \left| \left(-\tfrac{\mathfrak{s} w_2}{w_1} \right) - \left(-\tfrac{\mathfrak{s} v_2}{v_1} \right) \right| = \left| \tfrac{v_2}{v_1} - \tfrac{w_2}{w_1} \right|. \tag{2.22}$$

Furthermore, observe that the fact that for all $y \in \mathbb{R}$ it holds that $y \geq -|y|$ implies that for all $v = (v_1, v_2) \in \mathbb{R}^2$ with $\|u - v\| < |u_1|$ we have that

$$u_1 v_1 = (u_1)^2 + (v_1 - u_1) u_1 \geq |u_1|^2 - |u_1 - v_1||u_1| \geq |u_1|^2 - \|u - v\||u_1| > 0. \tag{2.23}$$

This ensures that for all $v = (v_1, v_2)$, $w = (w_1, w_2) \in \mathbb{R}^2$ with $\max\{\|u - v\|, \|u - w\|\} < |u_1|$ there exists $\mathfrak{s} \in \{-1, 1\}$ such that $\min\{\mathfrak{s} v_1, \mathfrak{s} w_1\} > 0$. Combining this with (2.22) demonstrates for all $v = (v_1, v_2)$, $w = (w_1, w_2) \in \mathbb{R}^2$ with $\max\{\|u - v\|, \|u - w\|\} \leq \frac{|u_1|}{2}$ that

$$\begin{aligned}\left|\int_{I^v \Delta I^w} \mathfrak{p}(x)\,\mathrm{d}x\right| &\leq M\left[\int_{I^v \Delta I^w} 1\,\mathrm{d}x\right] \leq 2M\left|\tfrac{v_2}{v_1} - \tfrac{w_2}{w_1}\right| = 2M\left|\frac{v_2(w_1 - v_1) - v_1(w_2 - v_2)}{v_1 w_1}\right| \\ &\leq 2M\left[\left|\frac{v_2(w_1 - v_1)}{v_1 w_1}\right| + \left|\frac{v_1(w_2 - v_2)}{v_1 w_1}\right|\right] \leq 2M\left[\frac{|v_2|\|v - w\|}{|v_1 w_1|} + \frac{|v_1|\|v - w\|}{|v_1 w_1|}\right] \\ &\leq \frac{4M\|v\|\|v - w\|}{|v_1 w_1|} \leq \left[\frac{16M\|v\|}{|u_1|^2}\right]\|v - w\| \leq \left[\frac{32M\|u\|}{|u_1|^2}\right]\|v - w\|.\end{aligned} \tag{2.24}$$

This establishes (2.17) in the case $u_1 \neq 0$. The proof of Lemma 7 is thus complete. ■

Corollary 8 *Assume Setting 3 and let $\theta \in \mathfrak{V}$. Then there exist $c, \varepsilon \in (0, \infty)$ such that for all $\vartheta_1, \vartheta_2 \in \mathbb{R}^{\mathfrak{d}}$ with $\max\{\|\vartheta_1 - \theta\|, \|\vartheta_2 - \theta\|\} \leq \varepsilon$ it holds that*

$$\int_{\cup_{i,j=1}^H((I_i^{\vartheta_1} \cap I_j^{\vartheta_1}) \Delta (I_i^{\vartheta_2} \cap I_j^{\vartheta_2}))} \mathfrak{p}(x)\,\mathrm{d}x \leq \int_{\cup_{i=1}^H (I_i^{\vartheta_1} \Delta I_i^{\vartheta_2})} \mathfrak{p}(x)\,\mathrm{d}x \leq c\|\vartheta_1 - \vartheta_2\|. \tag{2.25}$$

Proof [Proof of Corollary 8] Note that (2.3) ensures that $\min_{k \in \{1,2,\ldots,H\}}(|\mathfrak{w}_k^\theta| + |\mathfrak{b}_k^\theta|) > 0$. Combining this with Lemma 7 shows that there exist $c, \varepsilon \in (0, \infty)$ such that for all $k \in \{1, 2, \ldots, H\}$, $\vartheta_1, \vartheta_2 \in \mathbb{R}^{\mathfrak{d}}$ with $\max\{\|\theta - \vartheta_1\|, \|\theta - \vartheta_2\|\} \leq \varepsilon$ we have that

$$\int_{I_k^{\vartheta_1} \Delta I_k^{\vartheta_2}} \mathfrak{p}(x)\,\mathrm{d}x \leq c\|\vartheta_1 - \vartheta_2\|. \tag{2.26}$$

Next observe that the fact that for all sets $A, \mathbb{A}, B, \mathbb{B}$ it holds that

$$(A \cap \mathbb{A}) \backslash (B \cap \mathbb{B}) \subseteq (A \backslash B) \cup (\mathbb{A} \backslash \mathbb{B}) \subseteq (A \Delta B) \cup (\mathbb{A} \Delta \mathbb{B}) \tag{2.27}$$

implies that for all sets $A, \mathbb{A}, B, \mathbb{B}$ we have that

$$(A \cap \mathbb{A}) \Delta (B \cap \mathbb{B}) \subseteq (A \Delta B) \cup (\mathbb{A} \Delta \mathbb{B}). \tag{2.28}$$

Hence, we obtain for all $\vartheta_1, \vartheta_2 \in \mathbb{R}^{\mathfrak{d}}$, $i, j \in \{1, 2, \ldots, H\}$ that $(I_i^{\vartheta_1} \cap I_j^{\vartheta_1}) \Delta (I_i^{\vartheta_2} \cap I_j^{\vartheta_2}) \subseteq (I_i^{\vartheta_1} \Delta I_i^{\vartheta_2}) \cup (I_j^{\vartheta_2} \Delta I_j^{\vartheta_2})$. Combining this with (2.26) proves for all $\vartheta_1, \vartheta_2 \in \mathbb{R}^{\mathfrak{d}}$ with $\max\{\|\theta - \vartheta_1\|, \|\theta - \vartheta_2\|\} \leq \varepsilon$ that

$$\begin{aligned}\int_{\cup_{i,j=1}^H((I_i^{\vartheta_1} \cap I_j^{\vartheta_1}) \Delta (I_i^{\vartheta_2} \cap I_j^{\vartheta_2}))} \mathfrak{p}(x)\,\mathrm{d}x &\leq \int_{\cup_{k=1}^H (I_k^{\vartheta_1} \Delta I_k^{\vartheta_2})} \mathfrak{p}(x)\,\mathrm{d}x \\ &\leq \sum_{k=1}^H \left[\int_{I_k^{\vartheta_1} \Delta I_k^{\vartheta_2}} \mathfrak{p}(x)\,\mathrm{d}x\right] \leq cH\|\vartheta_1 - \vartheta_2\|.\end{aligned} \tag{2.29}$$

The proof of Corollary 8 is thus complete. ■

Corollary 9 *Assume Setting 3 and let* $i, j \in \{1, 2, \ldots, H\}$, $\phi \in C(\mathbb{R}^{\mathfrak{d}} \times [a, b], \mathbb{R})$. *Then*

(i) it holds that

$$\mathfrak{V} \ni \theta \mapsto \int_{I_i^\theta} \phi(\theta, x)\mathfrak{p}(x) \, \mathrm{d}x \in \mathbb{R} \tag{2.30}$$

is continuous and

(ii) it holds that

$$\mathfrak{V} \ni \theta \mapsto \int_{I_i^\theta \cap I_j^\theta} \phi(\theta, x)\mathfrak{p}(x) \, \mathrm{d}x \in \mathbb{R} \tag{2.31}$$

is continuous.

Proof [Proof of Corollary 9] Throughout this proof let $\theta \in \mathfrak{V}$. Note that Corollary 8 and Lemma 5 (applied with $n \curvearrowleft \mathfrak{d}$, $\mathfrak{u} \curvearrowleft a$, $\mathfrak{v} \curvearrowleft b$, $x \curvearrowleft \theta$, $\mu \curvearrowleft (\mathcal{B}([a, b]) \ni A \mapsto \int_A \mathfrak{p}(x) \, \mathrm{d}x \in [0, \infty])$ in the notation of Lemma 5) assure that there exists $\varepsilon \in (0, \infty)$ such that

$$\{\psi \in \mathbb{R}^{\mathfrak{d}} \colon \|\theta - \psi\| \leq \varepsilon\} \ni \vartheta \mapsto \int_{I_i^\vartheta} \phi(\vartheta, x)\mathfrak{p}(x) \, \mathrm{d}x \in \mathbb{R} \tag{2.32}$$

and

$$\{\psi \in \mathbb{R}^{\mathfrak{d}} \colon \|\theta - \psi\| \leq \varepsilon\} \ni \vartheta \mapsto \int_{I_i^\vartheta \cap I_j^\vartheta} \phi(\vartheta, x)\mathfrak{p}(x) \, \mathrm{d}x \in \mathbb{R} \tag{2.33}$$

are continuous. This shows items (i) and (ii). The proof of Corollary 9 is thus complete. ■

Corollary 10 *Assume Setting 3, let* $i, j \in \{1, 2, \ldots, H\}$, *and let* $\phi \colon \mathbb{R}^{\mathfrak{d}} \times [a, b] \to \mathbb{R}$ *be locally Lipschitz continuous. Then*

(i) it holds that

$$\mathfrak{V} \ni \theta \mapsto \int_{I_i^\theta} \phi(\theta, x)\mathfrak{p}(x) \, \mathrm{d}x \in \mathbb{R} \tag{2.34}$$

is locally Lipschitz continuous and

(ii) it holds that

$$\mathfrak{V} \ni \theta \mapsto \int_{I_i^\theta \cap I_j^\theta} \phi(\theta, x)\mathfrak{p}(x) \, \mathrm{d}x \in \mathbb{R} \tag{2.35}$$

is locally Lipschitz continuous.

Proof [Proof of Corollary 10] Throughout this proof let $\theta \in \mathfrak{V}$. Observe that Corollary 8 and Lemma 6 (applied with $n \curvearrowleft \mathfrak{d}$, $\mathfrak{u} \curvearrowleft a$, $\mathfrak{v} \curvearrowleft b$, $x \curvearrowleft \theta$, $\mu \curvearrowleft (\mathcal{B}([a, b]) \ni A \mapsto \int_A \mathfrak{p}(x) \, \mathrm{d}x \in [0, \infty])$ in the notation of Lemma 6) demonstrate that there exist $\varepsilon, \mathfrak{C} \in (0, \infty)$ such that for all $\vartheta_1, \vartheta_2 \in \mathbb{R}^{\mathfrak{d}}$ with $\max\{\|\theta - \vartheta_1\|, \|\theta - \vartheta_2\|\} \leq \varepsilon$ it holds that

$$\left| \int_{I_i^{\vartheta_1}} \phi(\vartheta_1, x)\mathfrak{p}(x) \, \mathrm{d}x - \int_{I_i^{\vartheta_2}} \phi(\vartheta_2, x)\mathfrak{p}(x) \, \mathrm{d}x \right| \leq \mathfrak{C}\|\vartheta_1 - \vartheta_2\| \tag{2.36}$$

and

$$\left| \int_{I_i^{\vartheta_1} \cap I_j^{\vartheta_1}} \phi(\vartheta_1, x)\mathfrak{p}(x)\,\mathrm{d}x - \int_{I_i^{\vartheta_2} \cap I_j^{\vartheta_2}} \phi(\vartheta_2, x)\mathfrak{p}(x)\,\mathrm{d}x \right| \le \mathfrak{C}\|\vartheta_1 - \vartheta_2\|. \tag{2.37}$$

This establishes items (i) and (ii). The proof of Corollary 10 is thus complete. ■

2.4 Explicit representations for the Hessian matrix of the risk function

Proposition 11 *Assume Setting 3 and let $\theta \in \mathfrak{V}$. Then*

(i) it holds that $\mathcal{L}$ is differentiable at θ and

(ii) it holds for all $i \in \{1, 2, \dots, H\}$ that

$$\begin{aligned} \mathcal{G}_i(\theta) &= \left(\tfrac{\partial}{\partial \theta_i}\mathcal{L}\right)(\theta) = 2\mathfrak{v}_i^\theta \int_{I_i^\theta} x(\mathscr{N}^\theta(x) - f(x))\mathfrak{p}(x)\,\mathrm{d}x, \\ \mathcal{G}_{H+i}(\theta) &= \left(\tfrac{\partial}{\partial \theta_{H+i}}\mathcal{L}\right)(\theta) = 2\mathfrak{v}_i^\theta \int_{I_i^\theta} (\mathscr{N}^\theta(x) - f(x))\mathfrak{p}(x)\,\mathrm{d}x, \\ \mathcal{G}_{2H+i}(\theta) &= \left(\tfrac{\partial}{\partial \theta_{2H+i}}\mathcal{L}\right)(\theta) = 2\int_a^b \left[\mathfrak{R}(\mathfrak{w}_i^\theta x + \mathfrak{b}_i^\theta)\right](\mathscr{N}^\theta(x) - f(x))\mathfrak{p}(x)\,\mathrm{d}x, \\ \textit{and} \qquad \mathcal{G}_{\mathfrak{d}}(\theta) &= \left(\tfrac{\partial}{\partial \theta_{\mathfrak{d}}}\mathcal{L}\right)(\theta) = 2\int_a^b (\mathscr{N}^\theta(x) - f(x))\mathfrak{p}(x)\,\mathrm{d}x. \end{aligned} \tag{2.38}$$

Proof [Proof of Proposition 11] Note that the assumption that $\theta \in \mathfrak{V}$ implies that for all $i \in \{1, 2, \dots, H\}$ it holds that $|\mathfrak{w}_i^\theta| + |\mathfrak{b}_i^\theta| > 0$. Hence, we obtain that

$$\mathcal{L}(\theta)\left(\textstyle\sum_{i=1}^H |\mathfrak{v}_i^\theta| \mathbb{1}_{\{0\}}\left(|\mathfrak{w}_i^\theta| + |\mathfrak{b}_i^\theta|\right)\right) = 0. \tag{2.39}$$

Combining this with (Jentzen and Riekert, 2022, Proposition 2.2 and Proposition 2.11) establishes items (i) and (ii). The proof of Proposition 11 is thus complete. ■

Lemma 12 *Assume Setting 3, let $i \in \{1, 2, \dots, H\}$, $r, s \in \mathbb{N}_0$, let $\psi\colon \mathbb{R} \to \mathbb{R}$ satisfy for all $x \in \mathbb{R}\backslash\{0\}$ that $\psi(x) = x^{-1}$, and let $\mathbf{c}\colon (-\infty, \infty] \to \mathbb{R}$ satisfy for all $x \in (-\infty, \infty]$ that $\mathbf{c}(x) = \max\{\min\{x, b\}, a\}$. Then*

(i) it holds for all continuous $\phi\colon \mathfrak{V} \times [a, b] \to \mathbb{R}$ that

$$\mathfrak{V} \ni \theta \mapsto \left[\psi([\mathfrak{w}_i^\theta]^r |\mathfrak{w}_i^\theta|^s)\right]\left[\phi(\theta, \mathbf{c}(\mathfrak{q}_i^\theta))\right]\mathbb{1}_{[a,b]}(\mathfrak{q}_i^\theta) \in \mathbb{R} \tag{2.40}$$

is continuous and

(ii) it holds for all locally Lipschitz continuous $\phi\colon \mathfrak{V} \times [a, b] \to \mathbb{R}$ that

$$\mathfrak{V} \ni \theta \mapsto \left[\psi([\mathfrak{w}_i^\theta]^r |\mathfrak{w}_i^\theta|^s)\right]\left[\phi(\theta, \mathbf{c}(\mathfrak{q}_i^\theta))\right]\mathbb{1}_{[a,b]}(\mathfrak{q}_i^\theta) \in \mathbb{R} \tag{2.41}$$

is locally Lipschitz continuous.

Proof [Proof of Lemma 12] Observe that (2.3) shows for all $\theta \in \mathfrak{V}$ that $|\mathfrak{w}_i^\theta| + |\mathfrak{b}_i^\theta| > 0$. Hence, we obtain for all $\theta \in \mathfrak{V}$ with $\mathfrak{w}_i^\theta = 0$ that $\mathfrak{b}_i^\theta \neq 0$. This implies that for all $\theta \in \mathfrak{V}$ with $\mathfrak{w}_i^\theta = 0$ there exists $\varepsilon \in (0, \infty)$ such that for all $\vartheta \in \{\psi \in \mathbb{R}^{\mathfrak{d}} \colon \|\psi - \theta\| < \varepsilon\}$ we have that $\mathfrak{q}_i^\vartheta \notin [a, b]$. Combining this with (2.1) and the fact that for all $\theta \in \mathfrak{V}$ it holds that $\mathfrak{q}_i^\theta \notin \{a, b\}$ establishes items (i) and (ii). The proof of Lemma 12 is thus complete. ■

Lemma 13 *Let $a \in \mathbb{R}$, $b \in (a, \infty)$, let $U \subseteq \mathbb{R}$ be open, let $\phi = (\phi_x(t))_{(x,t) \in [a,b] \times U} \in C([a, b] \times U, \mathbb{R})$ satisfy for all $x \in [a, b]$ that $\phi_x \in C^1(U, \mathbb{R})$, assume that $[a, b] \times U \ni (x, t) \mapsto (\phi_x)'(t) \in \mathbb{R}$ is continuous, let $\psi_0, \psi_1 \in C^1(U, [a, b])$, and let $\Phi \colon U \to \mathbb{R}$ satisfy for all $t \in U$ that*

$$\Phi(t) = \int_{\psi_0(t)}^{\psi_1(t)} \phi_x(t) \, \mathrm{d}x. \tag{2.42}$$

Then

(i) it holds that $\Phi \in C^1(U, \mathbb{R})$ and

(ii) it holds for all $t \in U$ that

$$\Phi'(t) = \big[\phi_{\psi_1(t)}(t)\big]\big[(\psi_1)'(t)\big] - \big[\phi_{\psi_0(t)}(t)\big]\big[(\psi_0)'(t)\big] + \int_{\psi_0(t)}^{\psi_1(t)} (\phi_x)'(t) \, \mathrm{d}x. \tag{2.43}$$

Proof [Proof of Lemma 13] Throughout this proof let $\Psi \colon [a, b] \times U \to \mathbb{R}$ satisfy for all $x \in [a, b]$, $t \in U$ that

$$\Psi(x, t) = \int_a^x \phi_y(t) \, \mathrm{d}y. \tag{2.44}$$

Note that (2.42) and (2.44) imply for all $t \in U$ that

$$\Phi(t) = \int_a^{\psi_1(t)} \phi_x(t) \, \mathrm{d}x - \int_a^{\psi_0(t)} \phi_x(t) \, \mathrm{d}x = \Psi(\psi_1(t), t) - \Psi(\psi_0(t), t). \tag{2.45}$$

Next observe that the fundamental theorem of calculus ensures for all $x \in [a, b]$, $t \in U$ that $\frac{\partial}{\partial x} \Psi(x, t) = \phi_x(t)$. In addition, note that Lemma 4 assures for all $x \in [a, b]$, $t \in U$ that $\frac{\partial}{\partial t} \Psi(x, t) = \int_a^x (\phi_y)'(t) \, \mathrm{d}y$. Furthermore, observe that the assumption that $[a, b] \times U \ni (x, t) \mapsto \phi_x(t) \in \mathbb{R}$ is continuous, the assumption that $[a, b] \times U \ni (x, t) \mapsto (\phi_x)'(t) \in \mathbb{R}$ is continuous, and the dominated convergence theorem demonstrate that $[a, b] \times U \ni (x, t) \mapsto \frac{\partial}{\partial x} \Psi(x, t) \in \mathbb{R}$ and $[a, b] \times U \ni (x, t) \mapsto \frac{\partial}{\partial t} \Psi(x, t) \in \mathbb{R}$ are continuous. Hence, we obtain that $\Psi \in C^1([a, b] \times U, \mathbb{R})$. Combining this with (2.45) and the chain rule shows for all $t \in U$ that $\Phi \in C^1(U, \mathbb{R})$ and

$$\begin{aligned}
\Phi'(t) &= (\psi_1)'(t) \big(\tfrac{\partial}{\partial x} \Psi\big)(\psi_1(t), t) + \big(\tfrac{\partial}{\partial t} \Psi\big)(\psi_1(t), t) \\
&\quad - (\psi_0)'(t) \big(\tfrac{\partial}{\partial x} \Psi\big)(\psi_0(t), t) - \big(\tfrac{\partial}{\partial t} \Psi\big)(\psi_0(t), t) \\
&= \big[(\psi_1)'(t)\big]\big[\phi_{\psi_1(t)}(t)\big] + \int_a^{\psi_1(t)} (\phi_x)'(t) \, \mathrm{d}x - \big[(\psi_0)'(t)\big]\big[\phi_{\psi_0(t)}(t)\big] - \int_a^{\psi_0(t)} (\phi_x)'(t) \, \mathrm{d}x \\
&= \big[(\psi_1)'(t)\big]\big[\phi_{\psi_1(t)}(t)\big] - \big[(\psi_0)'(t)\big]\big[\phi_{\psi_0(t)}(t)\big] + \int_{\psi_0(t)}^{\psi_1(t)} (\phi_x)'(t) \, \mathrm{d}x.
\end{aligned} \tag{2.46}$$

The proof of Lemma 13 is thus complete. ∎

Lemma 14 *Assume Setting 3, let $\psi\colon \mathbb{R} \to \mathbb{R}$ satisfy for all $x \in \mathbb{R}\backslash\{0\}$ that $\psi(x) = x^{-1}$, and let $\mathbf{c}\colon (-\infty, \infty] \to \mathbb{R}$ satisfy for all $x \in (-\infty, \infty]$ that $\mathbf{c}(x) = \max\{\min\{x, b\}, a\}$. Then*

(i) it holds that $\mathfrak{V} \subseteq \mathbb{R}^{\mathfrak{d}}$ is open,

(ii) it holds that $\mathcal{L}|_{\mathfrak{V}} \in C^2(\mathfrak{V}, \mathbb{R})$, and

(iii) it holds for all $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathfrak{V}$, $i, j \in \{1, 2, \ldots, H\}$ that

$$\big(\tfrac{\partial^2}{\partial \theta_j \partial \theta_{\mathfrak{d}}} \mathcal{L}\big)(\theta) = 2\mathfrak{v}_j^\theta \textstyle\int_{I_j^\theta} x \mathfrak{p}(x)\,\mathrm{d}x, \tag{2.47}$$

$$\big(\tfrac{\partial^2}{\partial \theta_{H+j} \partial \theta_{\mathfrak{d}}} \mathcal{L}\big)(\theta) = 2\mathfrak{v}_j^\theta \textstyle\int_{I_j^\theta} \mathfrak{p}(x)\,\mathrm{d}x, \tag{2.48}$$

$$\big(\tfrac{\partial^2}{\partial \theta_{2H+j} \partial \theta_{\mathfrak{d}}} \mathcal{L}\big)(\theta) = 2 \textstyle\int_a^b \big[\mathfrak{R}(\mathfrak{w}_j^\theta x + \mathfrak{b}_j^\theta)\big] \mathfrak{p}(x)\,\mathrm{d}x, \tag{2.49}$$

$$\big(\tfrac{\partial^2}{\partial \theta_{\mathfrak{d}}^2} \mathcal{L}\big)(\theta) = 2 \textstyle\int_a^b \mathfrak{p}(x)\,\mathrm{d}x, \tag{2.50}$$

$$\begin{split}\big(\tfrac{\partial^2}{\partial \theta_j \partial \theta_{2H+i}} \mathcal{L}\big)(\theta) &= 2\mathfrak{v}_j^\theta \textstyle\int_{I_j^\theta} x \big[\mathfrak{R}(\mathfrak{w}_i^\theta x + \mathfrak{b}_i^\theta)\big] \mathfrak{p}(x)\,\mathrm{d}x \\ &\quad + 2\mathbb{1}_{\{i\}}(j) \textstyle\int_{I_i^\theta} x(\mathscr{N}^\theta(x) - f(x))\mathfrak{p}(x)\,\mathrm{d}x,\end{split} \tag{2.51}$$

$$\begin{split}\big(\tfrac{\partial^2}{\partial \theta_{H+j} \partial \theta_{2H+i}} \mathcal{L}\big)(\theta) &= 2\mathfrak{v}_j^\theta \textstyle\int_{I_j^\theta} \big[\mathfrak{R}(\mathfrak{w}_i^\theta x + \mathfrak{b}_i^\theta)\big] \mathfrak{p}(x)\,\mathrm{d}x \\ &\quad + 2\mathbb{1}_{\{i\}}(j) \textstyle\int_{I_i^\theta} (\mathscr{N}^\theta(x) - f(x))\mathfrak{p}(x)\,\mathrm{d}x,\end{split} \tag{2.52}$$

$$\big(\tfrac{\partial^2}{\partial \theta_{2H+j} \partial \theta_{2H+i}} \mathcal{L}\big)(\theta) = 2 \textstyle\int_a^b \big[\mathfrak{R}(\mathfrak{w}_i^\theta x + \mathfrak{b}_i^\theta)\big]\big[\mathfrak{R}(\mathfrak{w}_j^\theta x + \mathfrak{b}_j^\theta)\big] \mathfrak{p}(x)\,\mathrm{d}x, \tag{2.53}$$

$$\begin{split}\big(\tfrac{\partial^2}{\partial \theta_j \partial \theta_i} \mathcal{L}\big)(\theta) &= 2\mathfrak{v}_i^\theta \mathfrak{v}_j^\theta \textstyle\int_{I_i^\theta \cap I_j^\theta} x^2 \mathfrak{p}(x)\,\mathrm{d}x \\ &\quad - 2\mathfrak{v}_i^\theta \mathfrak{b}_i^\theta \mathbb{1}_{\{i\}}(j) \mathbb{1}_{[a,b]}(\mathfrak{q}_i^\theta)[\psi(\mathfrak{w}_i^\theta |\mathfrak{w}_i^\theta|)][\mathbf{c}(\mathfrak{q}_i^\theta)](\mathscr{N}^\theta(\mathbf{c}(\mathfrak{q}_i^\theta)) - f(\mathbf{c}(\mathfrak{q}_i^\theta)))\mathfrak{p}(\mathbf{c}(\mathfrak{q}_i^\theta)),\end{split} \tag{2.54}$$

$$\begin{split}\big(\tfrac{\partial^2}{\partial \theta_j \partial \theta_{H+i}} \mathcal{L}\big)(\theta) &= 2\mathfrak{v}_i^\theta \mathfrak{v}_j^\theta \textstyle\int_{I_i^\theta \cap I_j^\theta} x \mathfrak{p}(x)\,\mathrm{d}x \\ &\quad + 2\mathfrak{v}_i^\theta \mathbb{1}_{\{i\}}(j) \mathbb{1}_{[a,b]}(\mathfrak{q}_i^\theta)[\psi(|\mathfrak{w}_i^\theta|)][\mathbf{c}(\mathfrak{q}_i^\theta)](\mathscr{N}^\theta(\mathbf{c}(\mathfrak{q}_i^\theta)) - f(\mathbf{c}(\mathfrak{q}_i^\theta)))\mathfrak{p}(\mathbf{c}(\mathfrak{q}_i^\theta)),\end{split} \tag{2.55}$$

and

$$\begin{split}\big(\tfrac{\partial^2}{\partial \theta_{H+j} \partial \theta_{H+i}} \mathcal{L}\big)(\theta) &= 2\mathfrak{v}_i^\theta \mathfrak{v}_j^\theta \textstyle\int_{I_i^\theta \cap I_j^\theta} \mathfrak{p}(x)\,\mathrm{d}x \\ &\quad + 2\mathfrak{v}_i^\theta \mathbb{1}_{\{i\}}(j) \mathbb{1}_{[a,b]}(\mathfrak{q}_i^\theta)[\psi(|\mathfrak{w}_i^\theta|)](\mathscr{N}^\theta(\mathbf{c}(\mathfrak{q}_i^\theta)) - f(\mathbf{c}(\mathfrak{q}_i^\theta)))\mathfrak{p}(\mathbf{c}(\mathfrak{q}_i^\theta)).\end{split} \tag{2.56}$$

Proof [Proof of Lemma 14] Note that (2.3) establishes item (i). Next observe that Proposition 11 ensures that $\mathfrak{V} \ni \theta \mapsto \mathcal{L}(\theta) \in \mathbb{R}$ is differentiable and satisfies $\nabla(\mathcal{L}|_{\mathfrak{V}}) = \mathcal{G}|_{\mathfrak{V}}$. In addition, note that (2.38) and Corollary 9 prove that $\nabla(\mathcal{L}|_{\mathfrak{V}})$ is continuous. Hence, we obtain that $\mathcal{L}|_{\mathfrak{V}} \in C^1(\mathfrak{V}, \mathbb{R})$. Combining this with (2.38), Lemma 4, and the product rule establishes (2.47) to (2.53).

In the next step we prove (2.54) to (2.56) and for this let $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathfrak{V}$, $i, j \in \{1, 2, \ldots, H\}$. In our proof of (2.54) to (2.56) we distinguish between the case $(i \neq j)$, the case $((i = j) \wedge (\max\{\mathfrak{w}^\theta_i a + \mathfrak{b}^\theta_i, \mathfrak{w}^\theta_i b + \mathfrak{b}^\theta_i\} < 0))$, the case $((i = j) \wedge (\min\{\mathfrak{w}^\theta_i a + \mathfrak{b}^\theta_i, \mathfrak{w}^\theta_i b + \mathfrak{b}^\theta_i\} > 0))$, the case $((i = j) \wedge (\mathfrak{w}^\theta_i a + \mathfrak{b}^\theta_i < 0 < \mathfrak{w}^\theta_i b + \mathfrak{b}^\theta_i))$, and the case $((i = j) \wedge (\mathfrak{w}^\theta_i a + \mathfrak{b}^\theta_i > 0 > \mathfrak{w}^\theta_i b + \mathfrak{b}^\theta_i))$. We first establish (2.54) to (2.56) in the case $(i \neq j)$. Observe that for all $k \in \{0, 1\}$ and almost all $x \in [a, b]$ it holds that

$$\tfrac{\partial}{\partial \theta_{kH+j}} \mathscr{N}^\theta(x) = \tfrac{\partial}{\partial \theta_{kH+j}} \big(\theta_{2H+j}[\mathfrak{R}(\theta_j x + \theta_{H+j})]\big) = \mathfrak{v}^\theta_j x^{1-k} \mathbb{1}_{I^\theta_j}(x). \tag{2.57}$$

Combining this with (2.38) and Lemma 4 (applied for every $k, \ell \in \{0, 1\}$ with $n \curvearrowleft \mathfrak{d}$, $j \curvearrowleft kH + j$, $\phi \curvearrowleft (\mathbb{R}^{\mathfrak{d}} \times [a, b] \ni (\vartheta, x) \mapsto x^{1-\ell}(\mathscr{N}^\vartheta(x) - f(x))\mathfrak{p}(x)\mathbb{1}_{I^\vartheta_i}(x) \in \mathbb{R})$ in the notation of Lemma 4) demonstrates for all $k, \ell \in \{0, 1\}$ that

$$\begin{split}&\big(\tfrac{\partial^2}{\partial \theta_{kH+j} \partial \theta_{\ell H+i}} \mathcal{L}\big)(\theta) = \big(\tfrac{\partial}{\partial \theta_{kH+j}} \mathcal{G}_{\ell H+i}\big)(\theta) \\ &= \tfrac{\partial}{\partial \theta_{kH+j}} \left(2\mathfrak{v}^\theta_i \int_a^b x^{1-\ell}(\mathscr{N}^\theta(x) - f(x))\mathfrak{p}(x)\mathbb{1}_{I^\theta_i}(x)\,\mathrm{d}x \right) = 2\mathfrak{v}^\theta_i \mathfrak{v}^\theta_j \int_{I^\theta_i \cap I^\theta_j} x^{2-k-\ell}\mathfrak{p}(x)\,\mathrm{d}x.\end{split} \tag{2.58}$$

This establishes (2.54) to (2.56) in the case $(i \neq j)$.

We next prove (2.54) to (2.56) in the case

$$(i = j) \wedge (\max\{\mathfrak{w}^\theta_i a + \mathfrak{b}^\theta_i, \mathfrak{w}^\theta_i b + \mathfrak{b}^\theta_i\} < 0). \tag{2.59}$$

Note that (2.59) implies that there exists $\delta \in (0, \infty)$ such that for all $h \in \mathbb{R}^{\mathfrak{d}}$ with $\|h\| < \delta$ it holds that $\mathfrak{q}^{\theta+h}_i \notin [a, b]$ and $I^{\theta+h}_i = \varnothing$. Combining this with (2.38) ensures that $\big(\tfrac{\partial^2}{\partial \theta_i^2} \mathcal{L}\big)(\theta) = \big(\tfrac{\partial^2}{\partial \theta_i \partial \theta_{H+i}} \mathcal{L}\big)(\theta) = \big(\tfrac{\partial^2}{\partial \theta_{H+i}^2} \mathcal{L}\big)(\theta) = 0$, as desired.

In the next step we prove (2.54) to (2.56) in the case

$$(i = j) \wedge (\min\{\mathfrak{w}^\theta_i a + \mathfrak{b}^\theta_i, \mathfrak{w}^\theta_i b + \mathfrak{b}^\theta_i\} > 0). \tag{2.60}$$

Observe that (2.60) implies that there exists $\delta \in (0, \infty)$ such that for all $h \in \mathbb{R}^{\mathfrak{d}}$ with $\|h\| < \delta$ it holds that $\mathfrak{q}^{\theta+h}_i \notin [a, b]$ and $I^{\theta+h}_i = [a, b]$. Combining (2.38) and Lemma 4 hence shows that $\big(\tfrac{\partial^2}{\partial \theta_i^2} \mathcal{L}\big)(\theta) = 2(\mathfrak{v}^\theta_i)^2 \int_a^b x^2 \mathfrak{p}(x)\,\mathrm{d}x$, $\big(\tfrac{\partial^2}{\partial \theta_i \partial \theta_{H+i}} \mathcal{L}\big)(\theta) = 2(\mathfrak{v}^\theta_i)^2 \int_a^b x\mathfrak{p}(x)\,\mathrm{d}x$, and $\big(\tfrac{\partial^2}{\partial \theta_{H+i}^2} \mathcal{L}\big)(\theta) = 2(\mathfrak{v}^\theta_i)^2 \int_a^b \mathfrak{p}(x)\,\mathrm{d}x$, as claimed.

In the remaining cases we employ Lemma 13 since the interval I^θ_i depends on $\mathfrak{w}^\theta_i$ and $\mathfrak{b}^\theta_i$ in these cases. We first consider the case

$$(i = j) \wedge (\mathfrak{w}^\theta_i a + \mathfrak{b}^\theta_i < 0 < \mathfrak{w}^\theta_i b + \mathfrak{b}^\theta_i). \tag{2.61}$$

Note that (2.61) ensures that there exists an open neighborhood $U \subseteq \mathbb{R}^{\mathfrak{d}}$ of θ which satisfies for all $\vartheta \in U$ that $\mathfrak{w}^\vartheta_i > 0$, $\mathfrak{q}^\vartheta_i \in (a, b)$, and $I^\vartheta_i = (\mathfrak{q}^\vartheta_i, b]$. Furthermore, observe that

$U \ni \vartheta \mapsto \mathfrak{q}_i^\vartheta = -\frac{\mathfrak{b}_i^\vartheta}{\mathfrak{w}_i^\vartheta} \in \mathbb{R}$ is continuously differentiable and satisfies $\frac{\partial}{\partial \theta_i} \mathfrak{q}_i^\theta = \frac{\mathfrak{b}_i^\theta}{(\mathfrak{w}_i^\theta)^2} = -\frac{\mathfrak{q}_i^\theta}{\mathfrak{w}_i^\theta}$ and $\frac{\partial}{\partial \theta_{H+i}} \mathfrak{q}_i^\theta = -\frac{1}{\mathfrak{w}_i^\theta}$. Combining Lemma 13 and (2.38) hence shows that

$$\begin{aligned}
\big(\tfrac{\partial^2}{\partial \theta_i^2} \mathcal{L}\big)(\theta) &= 2(\mathfrak{v}_i^\theta)^2 \textstyle\int_{I_i^\theta} x^2 \mathfrak{p}(x)\, \mathrm{d}x - \Big[\tfrac{2\mathfrak{v}_i^\theta \mathfrak{b}_i^\theta}{(\mathfrak{w}_i^\theta)^2}\Big] \mathfrak{q}_i^\theta (\mathscr{N}^\theta(\mathfrak{q}_i^\theta) - f(\mathfrak{q}_i^\theta)) \mathfrak{p}(\mathfrak{q}_i^\theta), \\
\big(\tfrac{\partial^2}{\partial \theta_i \partial \theta_{H+i}} \mathcal{L}\big)(\theta) &= 2(\mathfrak{v}_i^\theta)^2 \textstyle\int_{I_i^\theta} x \mathfrak{p}(x)\, \mathrm{d}x + \Big[\tfrac{2\mathfrak{v}_i^\theta}{\mathfrak{w}_i^\theta}\Big] \mathfrak{q}_i^\theta (\mathscr{N}^\theta(\mathfrak{q}_i^\theta) - f(\mathfrak{q}_i^\theta)) \mathfrak{p}(\mathfrak{q}_i^\theta), \\
\text{and} \quad \big(\tfrac{\partial^2}{\partial \theta_{H+i}^2} \mathcal{L}\big)(\theta) &= 2(\mathfrak{v}_i^\theta)^2 \textstyle\int_{I_i^\theta} \mathfrak{p}(x)\, \mathrm{d}x + \Big[\tfrac{2\mathfrak{v}_i^\theta}{\mathfrak{w}_i^\theta}\Big] (\mathscr{N}^\theta(\mathfrak{q}_i^\theta) - f(\mathfrak{q}_i^\theta)) \mathfrak{p}(\mathfrak{q}_i^\theta).
\end{aligned} \tag{2.62}$$

This establishes (2.54) to (2.56) in the case $((i = j) \wedge (\mathfrak{w}_i^\theta a + \mathfrak{b}_i^\theta < 0 < \mathfrak{w}_i^\theta b + \mathfrak{b}_i^\theta))$. It remains to consider the case

$$(i = j) \wedge (\mathfrak{w}_i^\theta a + \mathfrak{b}_i^\theta > 0 > \mathfrak{w}_i^\theta b + \mathfrak{b}_i^\theta) \tag{2.63}$$

Note that (2.63) assures that $\mathfrak{w}_i^\theta < 0$, $\mathfrak{q}_i^\theta \in (a, b)$, and $I_i^\theta = [a, \mathfrak{q}_i^\theta)$. Combining Lemma 13 and (2.38) therefore demonstrates that

$$\begin{aligned}
\big(\tfrac{\partial^2}{\partial \theta_i^2} \mathcal{L}\big)(\theta) &= 2(\mathfrak{v}_i^\theta)^2 \textstyle\int_{I_i^\theta} x^2 \mathfrak{p}(x)\, \mathrm{d}x + \Big[\tfrac{2\mathfrak{v}_i^\theta \mathfrak{b}_i^\theta}{(\mathfrak{w}_i^\theta)^2}\Big] \mathfrak{q}_i^\theta (\mathscr{N}^\theta(\mathfrak{q}_i^\theta) - f(\mathfrak{q}_i^\theta)) \mathfrak{p}(\mathfrak{q}_i^\theta), \\
\big(\tfrac{\partial^2}{\partial \theta_i \partial \theta_{H+i}} \mathcal{L}\big)(\theta) &= 2(\mathfrak{v}_i^\theta)^2 \textstyle\int_{I_i^\theta} x \mathfrak{p}(x)\, \mathrm{d}x - \Big[\tfrac{2\mathfrak{v}_i^\theta}{\mathfrak{w}_i^\theta}\Big] \mathfrak{q}_i^\theta (\mathscr{N}^\theta(\mathfrak{q}_i^\theta) - f(\mathfrak{q}_i^\theta)) \mathfrak{p}(\mathfrak{q}_i^\theta), \\
\text{and} \quad \big(\tfrac{\partial^2}{\partial \theta_{H+i}^2} \mathcal{L}\big)(\theta) &= 2(\mathfrak{v}_i^\theta)^2 \textstyle\int_{I_i^\theta} \mathfrak{p}(x)\, \mathrm{d}x - \Big[\tfrac{2\mathfrak{v}_i^\theta}{\mathfrak{w}_i^\theta}\Big] (\mathscr{N}^\theta(\mathfrak{q}_i^\theta) - f(\mathfrak{q}_i^\theta)) \mathfrak{p}(\mathfrak{q}_i^\theta).
\end{aligned} \tag{2.64}$$

This establishes (2.54) to (2.56) in the case $((i = j) \wedge (\mathfrak{w}_i^\theta a + \mathfrak{b}_i^\theta > 0 > \mathfrak{w}_i^\theta b + \mathfrak{b}_i^\theta))$.

Finally, observe that Corollary 9 and item (i) in Lemma 12 imply that the partial derivatives in (2.47) to (2.56) are continuous on $\mathfrak{V}$. The proof of Lemma 14 is thus complete. ■

Lemma 15 *Assume Setting 3 and assume that f is Lipschitz continuous. Then*

(i) it holds that $\mathfrak{V} \subseteq \mathbb{R}^{\mathfrak{d}}$ is open,

(ii) it holds that $\mathcal{L}|_{\mathfrak{V}} \in C^2(\mathfrak{V}, \mathbb{R})$, and

(iii) it holds that $\mathfrak{V} \ni \theta \mapsto (\operatorname{Hess} \mathcal{L})(\theta) \in \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}}$ is locally Lipschitz continuous.

Proof [Proof of Lemma 15] Note that Lemma 14 establishes items (i) and (ii). Moreover, observe that Lemma 14, Corollary 10, item (ii) in Lemma 12, the assumption that f is Lipschitz continuous, and the assumption that $\mathfrak{p}$ is Lipschitz continuous establish item (iii). The proof of Lemma 15 is thus complete. ■

Corollary 16 *Assume Setting 3, let $\theta \in \mathfrak{V}$, $i, j \in \{1, 2, \ldots, H\}$, and assume for all $x \in [a, b]$ that $\mathscr{N}^\theta(x) = f(x)$. Then*

$$\begin{aligned} \big(\tfrac{\partial^2}{\partial \theta_i \partial \theta_j} \mathcal{L}\big)(\theta) &= 2\mathfrak{v}_i^\theta \mathfrak{v}_j^\theta \int_{I_i^\theta \cap I_j^\theta} x^2 \mathfrak{p}(x)\, \mathrm{d}x, \\ \big(\tfrac{\partial^2}{\partial \theta_i \partial \theta_{H+j}} \mathcal{L}\big)(\theta) &= 2\mathfrak{v}_i^\theta \mathfrak{v}_j^\theta \int_{I_i^\theta \cap I_j^\theta} x \mathfrak{p}(x)\, \mathrm{d}x, \\ \text{and} \qquad \big(\tfrac{\partial^2}{\partial \theta_{H+i} \partial \theta_{H+j}} \mathcal{L}\big)(\theta) &= 2\mathfrak{v}_i^\theta \mathfrak{v}_j^\theta \int_{I_i^\theta \cap I_j^\theta} \mathfrak{p}(x)\, \mathrm{d}x. \end{aligned} \tag{2.65}$$

Proof [Proof of Corollary 16] Note that the assumption that for all $x \in [a, b]$ it holds that $\mathscr{N}^\theta(x) = f(x)$ and Lemma 14 establish (2.65). The proof of Corollary 16 is thus complete. ■

2.5 Upper bounds for the entries of the Hessian matrix of the risk function

Lemma 17 *Assume Setting 3, let $\mathfrak{D} \in [1, \infty)$, $A \in \mathbb{R}$ satisfy $A = \max\{1, |a|, |b|, b-a\}$, and let $\theta \in \mathfrak{V}$ satisfy $\max_{i \in \{1,2,\ldots,\mathfrak{d}\}} |\theta_i| \le \mathfrak{D}$ and $\min_{j \in \{1,2,\ldots,H\}} \big((\mathfrak{w}_j^\theta - \frac{1}{2}) \mathbb{1}_{[a,b]}(\mathfrak{q}_j^\theta)\big) \ge 0$. Then*

$$\begin{aligned} &\max\nolimits_{i,j \in \{1,2,\ldots,\mathfrak{d}\}} \big| \big(\tfrac{\partial^2}{\partial \theta_i \partial \theta_j} \mathcal{L}\big)(\theta) \big| \\ &\le \big(8A^3 \mathfrak{D}^2 + 8A^2 \mathfrak{D}^2 \big[\sup\nolimits_{x \in [a,b]} |\mathscr{N}^\theta(x) - f(x)|\big]\big) \big(\sup\nolimits_{x \in [a,b]} \mathfrak{p}(x)\big). \end{aligned} \tag{2.66}$$

Proof [Proof of Lemma 17] Throughout this proof let $\psi\colon \mathbb{R} \to \mathbb{R}$ satisfy for all $x \in \mathbb{R} \backslash \{0\}$ that $\psi(x) = x^{-1}$ and let $\mathbf{c}\colon (-\infty, \infty] \to \mathbb{R}$ satisfy for all $x \in (-\infty, \infty]$ that $\mathbf{c}(x) = \max\{\min\{x, b\}, a\}$. Observe that Lemma 14 implies for all $i, j \in \{1, 2, \ldots, H\}$ that

$$\big| \big(\tfrac{\partial^2}{\partial \theta_{\mathfrak{d}}^2} \mathcal{L}\big)(\theta) \big| = 2 \big| \textstyle\int_a^b \mathfrak{p}(x)\, \mathrm{d}x \big| \le 2A \big(\sup\nolimits_{x \in [a,b]} \mathfrak{p}(x)\big), \tag{2.67}$$

$$\begin{aligned} \big| \big(\tfrac{\partial^2}{\partial \theta_{2H+j} \partial \theta_{\mathfrak{d}}} \mathcal{L}\big)(\theta) \big| &= 2 \big| \textstyle\int_a^b [\mathfrak{R}(\mathfrak{w}_j^\theta x + \mathfrak{b}_j^\theta)] \mathfrak{p}(x)\, \mathrm{d}x \big| \le 2 \textstyle\int_a^b |\mathfrak{R}(\mathfrak{w}_j^\theta x + \mathfrak{b}_j^\theta)| \mathfrak{p}(x)\, \mathrm{d}x \\ &\le 2A(|\mathfrak{w}_j^\theta| + |\mathfrak{b}_j^\theta|) \textstyle\int_a^b \mathfrak{p}(x)\, \mathrm{d}x \le 4A^2 \mathfrak{D} \big(\sup\nolimits_{x \in [a,b]} \mathfrak{p}(x)\big), \end{aligned} \tag{2.68}$$

$$\begin{aligned} &\big| \big(\tfrac{\partial^2}{\partial \theta_{2H+i} \partial \theta_{2H+j}} \mathcal{L}\big)(\theta) \big| = 2 \big| \textstyle\int_a^b [\mathfrak{R}(\mathfrak{w}_i^\theta x + \mathfrak{b}_i^\theta)][\mathfrak{R}(\mathfrak{w}_j^\theta x + \mathfrak{b}_j^\theta)] \mathfrak{p}(x)\, \mathrm{d}x \big| \\ &\le 2 \textstyle\int_a^b |\mathfrak{R}(\mathfrak{w}_i^\theta x + \mathfrak{b}_i^\theta) \mathfrak{R}(\mathfrak{w}_j^\theta x + \mathfrak{b}_j^\theta)| \mathfrak{p}(x)\, \mathrm{d}x \\ &\le 2A^2(|\mathfrak{w}_i^\theta| + |\mathfrak{b}_i^\theta|)(|\mathfrak{w}_j^\theta| + |\mathfrak{b}_j^\theta|) \textstyle\int_a^b \mathfrak{p}(x)\, \mathrm{d}x \le 8A^3 \mathfrak{D}^2 \big(\sup\nolimits_{x \in [a,b]} \mathfrak{p}(x)\big), \end{aligned} \tag{2.69}$$

$$\big| \big(\tfrac{\partial^2}{\partial \theta_{\mathfrak{d}} \partial \theta_j} \mathcal{L}\big)(\theta) \big| = 2 |\mathfrak{v}_j^\theta| \big| \textstyle\int_{I_j^\theta} x \mathfrak{p}(x)\, \mathrm{d}x \big| \le 2A^2 \mathfrak{D} \big(\sup\nolimits_{x \in [a,b]} \mathfrak{p}(x)\big), \tag{2.70}$$

$$\big| \big(\tfrac{\partial^2}{\partial \theta_{\mathfrak{d}} \partial \theta_{H+j}} \mathcal{L}\big)(\theta) \big| = 2 |\mathfrak{v}_j^\theta| \textstyle\int_{I_j^\theta} \mathfrak{p}(x)\, \mathrm{d}x \le 2A \mathfrak{D} \big(\sup\nolimits_{x \in [a,b]} \mathfrak{p}(x)\big), \tag{2.71}$$

$$\begin{aligned}\left|\left(\tfrac{\partial^2}{\partial\theta_{2H+i}\partial\theta_j}\mathcal{L}\right)(\theta)\right| &\le 2|\mathfrak{v}_j^\theta|\left|\textstyle\int_{I_j^\theta} x[\mathfrak{R}(\mathfrak{w}_i^\theta x+\mathfrak{b}_i^\theta)]\mathfrak{p}(x)\,\mathrm{d}x\right| + 2\left|\textstyle\int_{I_i^\theta} x(\mathscr{N}^\theta(x)-f(x))\mathfrak{p}(x)\,\mathrm{d}x\right| \\ &\le \left(4A^3\mathfrak{D}^2+2A^2\left[\sup\nolimits_{x\in[a,b]}|\mathscr{N}^\theta(x)-f(x)|\right]\right)\left(\sup\nolimits_{x\in[a,b]}\mathfrak{p}(x)\right),\end{aligned} \tag{2.72}$$

and

$$\begin{aligned}\left|\left(\tfrac{\partial^2}{\partial\theta_{2H+i}\partial\theta_{H+j}}\mathcal{L}\right)(\theta)\right| &\le 2|\mathfrak{v}_j^\theta|\left|\textstyle\int_{I_j^\theta} [\mathfrak{R}(\mathfrak{w}_i^\theta x+\mathfrak{b}_i^\theta)]\mathfrak{p}(x)\,\mathrm{d}x\right| + 2\left|\textstyle\int_{I_i^\theta} (\mathscr{N}^\theta(x)-f(x))\mathfrak{p}(x)\,\mathrm{d}x\right| \\ &\le \left(4A^2\mathfrak{D}^2+2A\left[\sup\nolimits_{x\in[a,b]}|\mathscr{N}^\theta(x)-f(x)|\right]\right)\left(\sup\nolimits_{x\in[a,b]}\mathfrak{p}(x)\right).\end{aligned} \tag{2.73}$$

In addition, note that Lemma 14 and the fact that for all $i\in\{1,2,\ldots,H\}$ with $\mathfrak{q}_i^\theta\in[a,b]$ it holds that $\mathfrak{w}_i^\theta\ge\frac{1}{2}$ show that for all $i,j\in\{1,2,\ldots,H\}$ we have that

$$\begin{aligned}\left|\left(\tfrac{\partial^2}{\partial\theta_i\partial\theta_j}\mathcal{L}\right)(\theta)\right| &\le 2|\mathfrak{v}_i^\theta\mathfrak{v}_j^\theta|\left|\textstyle\int_{I_i^\theta\cap I_j^\theta} x^2\mathfrak{p}(x)\,\mathrm{d}x\right| \\ &\quad + \mathbb{1}_{[a,b]}(\mathfrak{q}_i^\theta)\left|2\mathfrak{v}_i^\theta\mathfrak{b}_i^\theta[\psi(|\mathfrak{w}_i^\theta|^2)][\mathbf{c}(\mathfrak{q}_i^\theta)](\mathscr{N}^\theta(\mathbf{c}(\mathfrak{q}_i^\theta))-f(\mathbf{c}(\mathfrak{q}_i^\theta)))\mathfrak{p}(\mathbf{c}(\mathfrak{q}_i^\theta))\right| \\ &\le \left(2A^3\mathfrak{D}^2+8A\mathfrak{D}^2\left[\sup\nolimits_{x\in[a,b]}|\mathscr{N}^\theta(x)-f(x)|\right]\right)\left(\sup\nolimits_{x\in[a,b]}\mathfrak{p}(x)\right),\end{aligned} \tag{2.74}$$

$$\begin{aligned}\left|\left(\tfrac{\partial^2}{\partial\theta_i\partial\theta_{H+j}}\mathcal{L}\right)(\theta)\right| &\le 2|\mathfrak{v}_i^\theta\mathfrak{v}_j^\theta|\left|\textstyle\int_{I_i^\theta\cap I_j^\theta} x\mathfrak{p}(x)\,\mathrm{d}x\right| \\ &\quad + \mathbb{1}_{[a,b]}(\mathfrak{q}_i^\theta)\left|2\mathfrak{v}_i^\theta[\psi(\mathfrak{w}_i^\theta)][\mathbf{c}(\mathfrak{q}_i^\theta)](\mathscr{N}^\theta(\mathbf{c}(\mathfrak{q}_i^\theta))-f(\mathbf{c}(\mathfrak{q}_i^\theta)))\mathfrak{p}(\mathbf{c}(\mathfrak{q}_i^\theta))\right| \\ &\le \left(2A^2\mathfrak{D}^2+4A\mathfrak{D}\left[\sup\nolimits_{x\in[a,b]}|\mathscr{N}^\theta(x)-f(x)|\right]\right)\left(\sup\nolimits_{x\in[a,b]}\mathfrak{p}(x)\right),\end{aligned} \tag{2.75}$$

and

$$\begin{aligned}\left|\left(\tfrac{\partial^2}{\partial\theta_{H+i}\partial\theta_{H+j}}\mathcal{L}\right)(\theta)\right| &\le 2|\mathfrak{v}_i^\theta\mathfrak{v}_j^\theta|\left|\textstyle\int_{I_i^\theta\cap I_j^\theta} \mathfrak{p}(x)\,\mathrm{d}x\right| \\ &\quad + \mathbb{1}_{[a,b]}(\mathfrak{q}_i^\theta)\left|2\mathfrak{v}_i^\theta[\psi(\mathfrak{w}_i^\theta)](\mathscr{N}^\theta(\mathbf{c}(\mathfrak{q}_i^\theta))-f(\mathbf{c}(\mathfrak{q}_i^\theta)))\mathfrak{p}(\mathbf{c}(\mathfrak{q}_i^\theta))\right| \\ &\le \left(2A\mathfrak{D}^2+4\mathfrak{D}\left[\sup\nolimits_{x\in[a,b]}|\mathscr{N}^\theta(x)-f(x)|\right]\right)\left(\sup\nolimits_{x\in[a,b]}\mathfrak{p}(x)\right).\end{aligned} \tag{2.76}$$

Combining this with the fact that $\{A,\mathfrak{D}\}\subseteq[1,\infty)$ establishes (2.66). The proof of Lemma 17 is thus complete. ∎

Lemma 18 *Assume Setting 3 and let $\theta\in\mathbb{R}^{\mathfrak{d}}$, $A\in\mathbb{R}$ satisfy $A=\max\{1,|a|,|b|\}$. Then*

$$\begin{aligned}\sup\nolimits_{x\in[a,b]}|\mathscr{N}^\theta(x)| &\le |\mathfrak{c}^\theta| + A\left[\textstyle\sum_{i=1}^H|\mathfrak{v}_i^\theta|(|\mathfrak{w}_i^\theta|+|\mathfrak{b}_i^\theta|)\right] \\ &\le \left[\max\nolimits_{i\in\{1,2,\ldots,\mathfrak{d}\}}|\theta_i|\right] + 2AH\left[\max\nolimits_{i\in\{1,2,\ldots,\mathfrak{d}\}}|\theta_i|^2\right].\end{aligned} \tag{2.77}$$

Proof [Proof of Lemma 18] Observe that for all $i\in\{1,2,\ldots,H\}$, $x\in[a,b]$ it holds that

$$|\mathfrak{v}_i^\theta\mathfrak{R}(\mathfrak{w}_i^\theta x+\mathfrak{b}_i^\theta)| \le |\mathfrak{v}_i^\theta|(|\mathfrak{w}_i^\theta x|+|\mathfrak{b}_i^\theta|) \le |\mathfrak{v}_i^\theta|(|\mathfrak{w}_i^\theta|+|\mathfrak{b}_i^\theta|)A. \tag{2.78}$$

This and the triangle inequality demonstrate for all $x \in [a,b]$ that

$$\begin{aligned} |\mathscr{N}^\theta(x)| &\leq |\mathfrak{c}^\theta| + \textstyle\sum_{i=1}^H |\mathfrak{v}_i^\theta \mathfrak{R}(\mathfrak{w}_i^\theta x + \mathfrak{b}_i^\theta)| \leq |\mathfrak{c}^\theta| + A\Big[\textstyle\sum_{i=1}^H |\mathfrak{v}_i^\theta|(|\mathfrak{w}_i^\theta| + |\mathfrak{b}_i^\theta|)\Big] \\ &\leq \big[\max_{i\in\{1,2,\dots,\mathfrak{d}\}}|\theta_i|\big] + 2AH\big[\max_{i\in\{1,2,\dots,\mathfrak{d}\}}|\theta_i|^2\big]. \end{aligned} \tag{2.79}$$

The proof of Lemma 18 is thus complete. ∎

Corollary 19 *Assume Setting 3, let $\mathfrak{D} \in [1,\infty)$, $A \in \mathbb{R}$ satisfy $A = \max\{1, |a|, |b|, b-a\}$, and let $\theta \in \mathfrak{V}$ satisfy $\max_{i\in\{1,2,\dots,\mathfrak{d}\}}|\theta_i| \leq \mathfrak{D}$ and $\min_{j\in\{1,2,\dots,H\}}\big((\mathfrak{w}_j^\theta - \frac{1}{2})\mathbb{1}_{[a,b]}(\mathfrak{q}_j^\theta)\big) \geq 0$. Then*

$$\begin{aligned} &\max_{i,j\in\{1,2,\dots,\mathfrak{d}\}}\big|\big(\tfrac{\partial^2}{\partial\theta_i\partial\theta_j}\mathcal{L}\big)(\theta)\big| \\ &\leq \big[8A^3\mathfrak{D}^2 + 8A^2\mathfrak{D}^2\big(\mathfrak{D} + 2AH\mathfrak{D}^2 + \sup_{x\in[a,b]}|f(x)|\big)\big]\big(\sup_{x\in[a,b]}\mathfrak{p}(x)\big) \\ &= \big[8A^3\mathfrak{D}^2 + 8A^2\mathfrak{D}^3 + 16A^3H\mathfrak{D}^4 + 8A^2\mathfrak{D}^2\big(\sup_{x\in[a,b]}|f(x)|\big)\big]\big(\sup_{x\in[a,b]}\mathfrak{p}(x)\big). \end{aligned} \tag{2.80}$$

Proof [Proof of Corollary 19] Note that Lemma 18 and the triangle inequality prove that for all $x \in [a,b]$ it holds that

$$|\mathscr{N}^\theta(x) - f(x)| \leq \mathfrak{D} + 2AH\mathfrak{D}^2 + |f(x)| \leq \mathfrak{D} + 2AH\mathfrak{D}^2 + \sup_{y\in[a,b]}|f(y)|. \tag{2.81}$$

This and Lemma 17 establish (2.80). The proof of Corollary 19 is thus complete. ∎

3. Regularity properties for the set of global minima of the risk function

In this section we establish in Corollary 29 in Subsection 3.3 below under the assumption that the target function is piecewise affine linear that there exists a natural number $k \in \{1, 2, \dots, \mathfrak{d}\}$ such that a suitable subset of the set of global minima of the considered risk function constitutes a k-dimensional C^∞-submanifold of the ANN parameter space on which the Hessian matrix of the risk function has the maximal rank $\mathfrak{d}-k$. Related results regarding C^∞-manifolds of ANN parameter vectors with identical realization functions have recently been established in Dereich and Kassing (2022).

Our proof of Corollary 29 employs Proposition 26 in Subsection 3.3 as well as the elementary and well-known eigenvalue estimate in Lemma 28 in Subsection 3.3. In the scientific literature Lemma 28 is, e.g., proved in (Golub and Van Loan, 2013, Section 2.3.2). In Proposition 26 we establish under the assumption that the target function is piecewise affine linear with varying slopes in consecutive sub-intervals that a suitable subset of the set of global minima of the risk function represents an $(H+1)$-dimensional C^∞-submanifold of the ANN parameter space on which the Hessian matrix of the risk function has the maximal rank $\mathfrak{d} - (H+1) = (3H+1) - (H+1) = 2H$ where $H \in \mathbb{N}$ represents the number of neurons on the hidden layer (see Setting 3 for details).

Our proof of Proposition 26 uses Lemma 21 in Subsection 3.1, Proposition 23 in Subsection 3.2, and the elementary and well-known properties for tangent spaces of submanifolds in Lemma 25 in Subsection 3.3. The notion of tangent spaces is recalled in Definition 24 in Subsection 3.3. Only for the sake of completeness we include in this section the detailed proof for Lemma 25. Our proof of Proposition 23, in turn, is based on an application of the auxiliary result in Lemma 22 in Subsection 3.2 and in Lemma 22 and Proposition 23 we show that certain matrices involving appropriate sub-integrals of the unnormalized density function have a strictly positive determinant.

In Lemma 21 in Subsection 3.1 we verify that a suitable subset of the ANN parameter space is a non-empty $(H+1)$-dimensional C^∞-submanifold of the ANN parameter space $\mathbb{R}^{\mathfrak{d}}$. Our proof of Lemma 21 is based on an application of the regular level set theorem which we recall in Proposition 20 below. In the scientific literature Proposition 20 is sometimes also referred to as submersion level set theorem, regular value theorem, or preimage theorem. Proposition 20 is, e.g., proved as Theorem 9.9 in Tu (2011).

3.1 Submanifolds of the ANN parameter space

Proposition 20 *Let $\mathfrak{d}, n \in \mathbb{N}$, let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $g \in C^\infty(U, \mathbb{R}^n)$, and assume for all $x \in g^{-1}(\{0\})$ that $\operatorname{rank}(g'(x)) = n$. Then it holds that $g^{-1}(\{0\}) \subseteq U$ is a $(\mathfrak{d}-n)$-dimensional C^∞-submanifold of $\mathbb{R}^{\mathfrak{d}}$.*

In Proposition 20 above, we have for every $x \in U$ that the differential $g'(x)$ is a linear map from $\mathbb{R}^{\mathfrak{d}}$ to $\mathbb{R}^n$, which can be identified with a matrix in $\mathbb{R}^{n \times \mathfrak{d}}$, the Jacobian.

Lemma 21 *Assume Setting 3, let $\mathscr{x}_0, \mathscr{x}_1, \dots, \mathscr{x}_H, \alpha_1, \alpha_2, \dots, \alpha_H, \mathfrak{D}, \mathscr{y} \in \mathbb{R}$ satisfy $a = \mathscr{x}_0 < \mathscr{x}_1 < \cdots < \mathscr{x}_H = b$ and*

$$\mathfrak{D} \geq 1 + |\mathscr{y}| + (1 + 2\max\nolimits_{j \in \{1,2,\dots,H\}} |\alpha_j|)(1 + |a| + |b|), \tag{3.1}$$

and let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ be given by

$$\begin{aligned} \mathcal{M} = \big\{\theta \in (-\mathfrak{D}, \mathfrak{D})^{\mathfrak{d}} \colon \big(\big[\min\{\mathfrak{w}_1^\theta a + \mathfrak{b}_1^\theta, \mathfrak{w}_1^\theta b + \mathfrak{b}_1^\theta, \mathfrak{v}_1^\theta\} > 0\big], \, [\mathfrak{v}_1^\theta(\mathfrak{w}_1^\theta a + \mathfrak{b}_1^\theta) + \mathfrak{c}^\theta = \mathscr{y}], \\ [\mathfrak{w}_1^\theta \mathfrak{v}_1^\theta = \alpha_1], \, \big[\forall\, j \in \mathbb{N} \cap (1, H] \colon \mathfrak{w}_j^\theta > 1/2, \, \mathfrak{q}_j^\theta = \mathscr{x}_{j-1}, \, \mathfrak{w}_j^\theta \mathfrak{v}_j^\theta = \alpha_j - \alpha_{j-1}\big]\big)\big\}. \end{aligned} \tag{3.2}$$

Then

(i) it holds that $\mathcal{M} \neq \varnothing$ and

(ii) it holds that $\mathcal{M}$ is a $(H+1)$-dimensional C^∞-submanifold of $\mathbb{R}^{\mathfrak{d}}$.

Note that in (3.2) and subsequent equations, we have that $\mathbb{N} \cap (1, H] = \{1, 2, \dots, H\} \backslash \{1\} = \{i \in \mathbb{N} \colon 2 \leq i \leq H\}$.

Proof [Proof of Lemma 21] Throughout this proof let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy

$$U = \big\{\theta \in (-\mathfrak{D}, \mathfrak{D})^{\mathfrak{d}} \colon \big(\big[\min\{\mathfrak{w}_1^\theta a + \mathfrak{b}_1^\theta, \mathfrak{w}_1^\theta b + \mathfrak{b}_1^\theta, \mathfrak{v}_1^\theta\} > 0\big], \, \big[\forall\, j \in \mathbb{N} \cap (1, H] \colon \mathfrak{w}_j^\theta > 1/2\big]\big)\big\}, \tag{3.3}$$

let $g = (g_1, \dots, g_{2H}) \colon U \to \mathbb{R}^{2H}$ satisfy for all $\theta \in U$, $j \in \{1, 2, \dots, H\}$ that

$$g_j(\theta) = \begin{cases} \mathfrak{w}_1^\theta \mathfrak{v}_1^\theta - \alpha_1 & : j = 1 \\ \mathfrak{w}_j^\theta \mathfrak{v}_j^\theta - (\alpha_j - \alpha_{j-1}) & : j > 1 \end{cases} \tag{3.4}$$

and

$$g_{H+j}(\theta) = \begin{cases} \mathfrak{v}_1^\theta(\mathfrak{w}_1^\theta a + \mathfrak{b}_1^\theta) + \mathfrak{c}^\theta - \mathscr{y} & : j = 1 \\ \mathfrak{q}_j^\theta - \mathscr{x}_{j-1} & : j > 1, \end{cases} \tag{3.5}$$

and let $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ satisfy

$$\big([\mathfrak{w}_1^\vartheta = \alpha_1],\ [\forall\, i \in \mathbb{N} \cap (1, H] \colon \mathfrak{w}_i^\vartheta = 1],\ [\mathfrak{b}_1^\vartheta = |\alpha_1|(|a|+|b|)+1],\ [\forall\, i \in \mathbb{N} \cap (1, H] \colon \mathfrak{b}_i^\vartheta = -\mathscr{x}_{i-1}],\\ [\mathfrak{v}_1^\vartheta = 1],\ [\forall\, i \in \mathbb{N} \cap (1, H] \colon \mathfrak{v}_i^\vartheta = \alpha_i - \alpha_{i-1}],\ [\mathfrak{c}^\vartheta = \mathscr{y} - \mathfrak{v}_1^\vartheta(\mathfrak{w}_1^\vartheta a + \mathfrak{b}_1^\vartheta)]\big). \tag{3.6}$$

Observe that (3.6) ensures that $\mathfrak{v}_1^\vartheta > 0$, $\mathfrak{w}_1^\vartheta \mathfrak{v}_1^\vartheta = \alpha_1$, and $\mathfrak{v}_1^\vartheta(\mathfrak{w}_1^\vartheta a + \mathfrak{b}_1^\vartheta) + \mathfrak{c}^\vartheta = \mathscr{y}$. Moreover, note that $\min\{\mathfrak{w}_1^\vartheta a + \mathfrak{b}_1^\vartheta, \mathfrak{w}_1^\vartheta b + \mathfrak{b}_1^\vartheta\} = \min\{\alpha_1 a, \alpha_1 b\} + |\alpha_1|(|a|+|b|)+1 \geq 1 > 0$. In addition, observe that for all $j \in \mathbb{N} \cap (1, H]$ we have that $\mathfrak{w}_j^\vartheta = 1 > 1/2$, $\mathfrak{q}_j^\vartheta = -\mathfrak{b}_j^\vartheta / \mathfrak{w}_j^\vartheta = \mathscr{x}_{j-1}$, and $\mathfrak{w}_j^\vartheta \mathfrak{v}_j^\vartheta = \alpha_j - \alpha_{j-1}$. Furthermore, note that for all $i \in \mathbb{N} \cap (1, H]$ it holds that $|\mathfrak{w}_i^\vartheta| = 1 < \mathfrak{D}$, $|\mathfrak{v}_i^\vartheta| \leq 2 \max_{j \in \{1,2,\ldots,H\}} |\alpha_j| < \mathfrak{D}$, and $|\mathfrak{b}_i^\vartheta| \leq 1 + |a| + |b| < \mathfrak{D}$. Moreover, observe that $|\mathfrak{w}_1^\vartheta| = |\alpha_1| < \mathfrak{D}$, $|\mathfrak{b}_1^\vartheta| \leq (1 + \max_{j \in \{1,2,\ldots,H\}} |\alpha_j|)(1 + |a| + |b|) < \mathfrak{D}$, $|\mathfrak{v}_1^\vartheta| = 1 < \mathfrak{D}$, and

$$\begin{aligned} |\mathfrak{c}^\vartheta| &\leq |\mathscr{y}| + |\mathfrak{v}_1^\vartheta \mathfrak{w}_1^\vartheta a| + |\mathfrak{v}_1^\vartheta \mathfrak{b}_1^\vartheta| = |\mathscr{y}| + |\alpha_1||a| + |\alpha_1|(|a|+|b|) + 1 \\ &\leq |\mathscr{y}| + \big(1 + 2 \max\nolimits_{j \in \{1,2,\ldots,H\}} |\alpha_j|\big)(1 + |a| + |b|) < \mathfrak{D}. \end{aligned} \tag{3.7}$$

This implies that $\vartheta \in (-\mathfrak{D}, \mathfrak{D})^{\mathfrak{d}}$. Hence, we obtain that $\vartheta \in \mathcal{M}$. This establishes item (i). In the next step we prove item (ii) through an application of the regular value theorem in Proposition 20. Note that (3.3) assures that $U \subseteq \mathbb{R}^{\mathfrak{d}}$ is open. In addition, observe that the fact that for all $\theta \in U$, $j \in \mathbb{N} \cap (1, H]$ it holds that $\mathfrak{w}_j^\theta > 0$ ensures that $g \in C^\infty(U, \mathbb{R}^{2H})$. Moreover, note that

$$g^{-1}(\{0\}) = \big\{\theta \in U \colon \big([\mathfrak{w}_1^\theta \mathfrak{v}_1^\theta = \alpha_1],\ [\mathfrak{v}_1^\theta(\mathfrak{w}_1^\theta a + \mathfrak{b}_1^\theta) + \mathfrak{c}^\theta = \mathscr{y}],\\ \big[\forall\, j \in \mathbb{N} \cap (1, H] \colon \mathfrak{q}_j^\theta = \mathscr{x}_{j-1},\ \mathfrak{w}_j^\theta \mathfrak{v}_j^\theta = \alpha_j - \alpha_{j-1}\big]\big)\big\}. \tag{3.8}$$

This implies that

$$g^{-1}(\{0\}) = \big\{\theta \in (-\mathfrak{D}, \mathfrak{D})^{\mathfrak{d}} \colon \big(\big[\min\{\mathfrak{w}_1^\theta a + \mathfrak{b}_1^\theta, \mathfrak{w}_1^\theta b + \mathfrak{b}_1^\theta, \mathfrak{v}_1^\theta\} > 0\big],\\ [\forall\, j \in \mathbb{N} \cap (1, H] \colon \mathfrak{w}_j^\theta > 1/2],\ [\mathfrak{w}_1^\theta \mathfrak{v}_1^\theta = \alpha_1],\ [\mathfrak{v}_1^\theta(\mathfrak{w}_1^\theta a + \mathfrak{b}_1^\theta) + \mathfrak{c}^\theta = \mathscr{y}],\\ \big[\forall\, j \in \mathbb{N} \cap (1, H] \colon \mathfrak{q}_j^\theta = \mathscr{x}_{j-1},\ \mathfrak{w}_j^\theta \mathfrak{v}_j^\theta = \alpha_j - \alpha_{j-1}\big]\big)\big\} = \mathcal{M}. \tag{3.9}$$

Next observe that (3.4), (3.5), and the fact that for all $\theta \in U$, $j \in \mathbb{N} \cap [1, H]$ it holds that $\mathfrak{w}_j^\theta = \theta_j$, $\mathfrak{b}_j^\theta = \theta_{H+j}$, and $\mathfrak{v}_j^\theta = \theta_{2H+j}$ ensure that for all $\theta \in U$, $j \in \mathbb{N} \cap (1, H]$, $\ell \in \mathbb{N} \cap [1, 2H]$ we have that

$$\big(\tfrac{\partial}{\partial \theta_{2H+j}} g_\ell\big)(\theta) = \begin{cases} \mathfrak{w}_j^\theta \neq 0 & : \ell = j \\ 0 & : \ell \neq j \end{cases} \tag{3.10}$$

and

$$\big(\tfrac{\partial}{\partial \theta_{H+j}} g_\ell\big)(\theta) = \begin{cases} -(\mathfrak{w}_j^\theta)^{-1} \neq 0 & : \ell = H + j \\ 0 & : \ell \neq H + j. \end{cases} \tag{3.11}$$

In addition, note that (3.4) and (3.5) show for all $\theta \in U$, $\ell \in \mathbb{N} \cap [1, 2H]$ that

$$\left(\tfrac{\partial}{\partial \theta_1} g_\ell\right)(\theta) = \begin{cases} \mathfrak{v}_1^\theta \neq 0 & : \ell = 1 \\ \mathfrak{v}_1^\theta a & : \ell = H+1 \\ 0 & : \ell \notin \{1, H+1\} \end{cases} \tag{3.12}$$

and

$$\left(\tfrac{\partial}{\partial \theta_{H+1}} g_\ell\right)(\theta) = \begin{cases} \mathfrak{v}_1^\theta \neq 0 & : \ell = H+1 \\ 0 & : \ell \neq H+1. \end{cases} \tag{3.13}$$

This demonstrates that for all $\theta \in U$ it holds that the $((2H) \times (2H))$-matrix with entries $\left(\frac{\partial}{\partial \theta_i} g_\ell\right)(\theta) \in \mathbb{R}$, $(i, \ell) \in (\{1\} \cup \{H + j \colon j \in \mathbb{N} \cap [1, H]\} \cup \{2H + j \colon j \in \mathbb{N} \cap (1, H]\}) \times \{1, 2, \ldots, 2H\}$, is invertible. Hence, we obtain for all $\theta \in U$ that $\operatorname{rank}(g'(\theta)) = 2H$. Combining this with Proposition 20 establishes item (ii). The proof of Lemma 21 is thus complete. ■

3.2 Determinants of submatrices of the Hessian matrix of the risk function

Lemma 22 *Let $a \in \mathbb{R}$, $b \in (a, \infty)$, let $\mathfrak{p} \colon [a, b] \to (0, \infty)$ be bounded and measurable, let $\mathcal{Q}_N \subseteq \mathbb{R}^{N+1}$, $N \in \mathbb{N}$, satisfy for all $N \in \mathbb{N}$ that $\mathcal{Q}_N = \{\mathscr{x} = (\mathscr{x}_1, \ldots, \mathscr{x}_{N+1}) \in \mathbb{R}^{N+1} \colon a \leq \mathscr{x}_1 < \mathscr{x}_2 < \cdots < \mathscr{x}_{N+1} \leq b\}$, and let $A^{N,\mathscr{x}} = (A^{N,\mathscr{x}}_{i,j})_{(i,j) \in \{1,2,\ldots,2N\}^2} \in \mathbb{R}^{(2N) \times (2N)}$, $\mathscr{x} \in \mathcal{Q}_N$, $N \in \mathbb{N}$, satisfy for all $N \in \mathbb{N}$, $\mathscr{x} = (\mathscr{x}_1, \ldots, \mathscr{x}_{N+1}) \in \mathcal{Q}_N$, $i, j \in \{1, 2, \ldots, N\}$ that*

$$A^{N,\mathscr{x}}_{i,j} = \textstyle\int_{\mathscr{x}_{\max\{i,j\}}}^{\mathscr{x}_{N+1}} x^2 \mathfrak{p}(x)\,\mathrm{d}x, \qquad A^{N,\mathscr{x}}_{N+i,j} = A^{N,\mathscr{x}}_{i,N+j} = \textstyle\int_{\mathscr{x}_{\max\{i,j\}}}^{\mathscr{x}_{N+1}} x \mathfrak{p}(x)\,\mathrm{d}x,$$
$$\text{and} \qquad A^{N,\mathscr{x}}_{N+i,N+j} = \textstyle\int_{\mathscr{x}_{\max\{i,j\}}}^{\mathscr{x}_{N+1}} \mathfrak{p}(x)\,\mathrm{d}x. \tag{3.14}$$

Then it holds for all $N \in \mathbb{N}$, $\mathscr{x} \in \mathcal{Q}_N$ that

$$\det(A^{N,\mathscr{x}}) = \prod_{i=1}^{N} \left(\left[\textstyle\int_{\mathscr{x}_i}^{\mathscr{x}_{i+1}} x^2 \mathfrak{p}(x)\,\mathrm{d}x\right] \left[\textstyle\int_{\mathscr{x}_i}^{\mathscr{x}_{i+1}} \mathfrak{p}(x)\,\mathrm{d}x\right] - \left[\textstyle\int_{\mathscr{x}_i}^{\mathscr{x}_{i+1}} x \mathfrak{p}(x)\,\mathrm{d}x\right]^2 \right) > 0. \tag{3.15}$$

Proof [Proof of Lemma 22] Throughout this proof let $E^{N,\mathscr{x}}_i \in \mathbb{R}$, $i \in \{1, 2, \ldots, N\}$, $\mathscr{x} \in \mathcal{Q}_N$, $N \in \mathbb{N}$, satisfy for all $N \in \mathbb{N}$, $\mathscr{x} \in \mathcal{Q}_N$, $i \in \{1, 2, \ldots, N\}$ that

$$E^{N,\mathscr{x}}_i = \left[\textstyle\int_{\mathscr{x}_i}^{\mathscr{x}_{i+1}} x^2 \mathfrak{p}(x)\,\mathrm{d}x\right] \left[\textstyle\int_{\mathscr{x}_i}^{\mathscr{x}_{i+1}} \mathfrak{p}(x)\,\mathrm{d}x\right] - \left[\textstyle\int_{\mathscr{x}_i}^{\mathscr{x}_{i+1}} x \mathfrak{p}(x)\,\mathrm{d}x\right]^2. \tag{3.16}$$

Observe that the Cauchy-Schwarz inequality and the fact that for all $x \in [a, b]$ it holds that $\mathfrak{p}(x) > 0$ ensure that for all $N \in \mathbb{N}$, $\mathscr{x} \in \mathcal{Q}_N$, $i \in \{1, 2, \ldots, N\}$ we have that

$$\begin{aligned} \left| \textstyle\int_{\mathscr{x}_i}^{\mathscr{x}_{i+1}} x \mathfrak{p}(x)\,\mathrm{d}x \right| &= \left| \textstyle\int_{\mathscr{x}_i}^{\mathscr{x}_{i+1}} \left[x \sqrt{\mathfrak{p}(x)}\right] \left[\sqrt{\mathfrak{p}(x)}\right] \mathrm{d}x \right| \\ &< \left[\textstyle\int_{\mathscr{x}_i}^{\mathscr{x}_{i+1}} x^2 \mathfrak{p}(x)\,\mathrm{d}x\right]^{1/2} \left[\textstyle\int_{\mathscr{x}_i}^{\mathscr{x}_{i+1}} \mathfrak{p}(x)\,\mathrm{d}x\right]^{1/2}. \end{aligned} \tag{3.17}$$

Hence, we obtain for all $N \in \mathbb{N}$, $\mathscr{x} \in \mathcal{Q}_N$, $i \in \{1, 2, \ldots, N\}$ that $E^{N,\mathscr{x}}_i > 0$. Next we claim that for all $N \in \mathbb{N}$, $x \in \mathcal{Q}_N$ it holds that

$$\det(A^{N,\mathscr{x}}) = \textstyle\prod_{i=1}^{N} E^{N,\mathscr{x}}_i > 0. \tag{3.18}$$

We now prove (3.18) by induction on $N \in \mathbb{N}$. For the base case $N = 1$ note that for all $x = (x_1, x_2) \in \mathcal{Q}_1$ it holds that

$$\det(A^{1,x}) = \det\begin{pmatrix} \int_{x_1}^{x_2} x^2 \mathfrak{p}(x)\,\mathrm{d}x & \int_{x_1}^{x_2} x\mathfrak{p}(x)\,\mathrm{d}x \\ \int_{x_1}^{x_2} x\mathfrak{p}(x)\,\mathrm{d}x & \int_{x_1}^{x_2} \mathfrak{p}(x)\,\mathrm{d}x \end{pmatrix} = E_1^{1,x} > 0. \tag{3.19}$$

This establishes (3.18) in the base case $N = 1$. For the induction step let $N \in \mathbb{N} \cap [2, \infty)$ and assume for all $x \in \mathcal{Q}_{N-1}$ that

$$\det(A^{N-1,x}) = \textstyle\prod_{i=1}^{N-1} E_i^{N-1,x} > 0. \tag{3.20}$$

Next let $x = (x_1, \ldots, x_{N+1}) \in \mathcal{Q}_N$ and let $B = (B_{i,j})_{(i,j)\in\{1,2,\ldots,2N\}^2} \in \mathbb{R}^{(2N)\times(2N)}$ satisfy for all $i, j \in \{1, 2, \ldots, 2N\}$ that

$$B_{i,j} = \begin{cases} A_{i,j}^{N,x} & : i \notin \{1, N+1\} \\ A_{1,j}^{N,x} - A_{2,j}^{N,x} & : i = 1 \\ A_{N+1,j}^{N,x} - A_{N+2,j}^{N,x} & : i = N+1. \end{cases} \tag{3.21}$$

Observe that B is the matrix that is obtained from $A^{N,x}$ by subtracting the 2nd row from the 1st row and the $(N+2)$-th row from the $(N+1)$-th row. In particular, note that (3.21) implies that $\det(B) = \det(A^{N,x})$. Next observe that the fact that for all $j \in \mathbb{N} \cap (1, N]$ it holds that $A_{1,j}^{N,x} = A_{2,j}^{N,x}$, $A_{1,N+j}^{N,x} = A_{2,N+j}^{N,x}$, $A_{N+1,j}^{N,x} = A_{N+2,j}^{N,x}$, and $A_{N+1,N+j}^{N,x} = A_{N+2,N+j}^{N,x}$ demonstrates that for all $i, j \in \mathbb{N} \cap (1, N]$ we have that

$$\begin{aligned} B_{1,1} &= A_{1,1}^{N,x} - A_{2,1}^{N,x} = \textstyle\int_{x_1}^{x_{N+1}} x^2\mathfrak{p}(x)\,\mathrm{d}x - \int_{x_2}^{x_{N+1}} x^2\mathfrak{p}(x)\,\mathrm{d}x = \int_{x_1}^{x_2} x^2\mathfrak{p}(x)\,\mathrm{d}x, \\ B_{N+1,1} &= B_{1,N+1} = \textstyle\int_{x_1}^{x_{N+1}} x\mathfrak{p}(x)\,\mathrm{d}x - \int_{x_2}^{x_{N+1}} x\mathfrak{p}(x)\,\mathrm{d}x = \int_{x_1}^{x_2} x\mathfrak{p}(x)\,\mathrm{d}x, \\ B_{N+1,N+1} &= \textstyle\int_{x_1}^{x_{N+1}} \mathfrak{p}(x)\,\mathrm{d}x - \int_{x_2}^{x_{N+1}} \mathfrak{p}(x)\,\mathrm{d}x = \int_{x_1}^{x_2} \mathfrak{p}(x)\,\mathrm{d}x, \\ B_{1,j} &= B_{N+1,j} = B_{1,N+j} = B_{N+1,N+j} = 0, \quad B_{i,j} = A_{i,j}^{N,x}, \quad B_{N+i,j} = A_{N+i,j}^{N,x}, \\ B_{i,N+j} &= A_{i,N+j}^{N,x}, \qquad \text{and} \qquad B_{N+i,N+j} = A_{N+i,N+j}^{N,x}. \end{aligned} \tag{3.22}$$

Hence, we obtain that

$$\begin{aligned} \det(B) &= (B_{1,1}B_{N+1,N+1} - B_{N+1,1}B_{1,N+1}) \det\big((B_{i,j})_{(i,j)\in(\{1,2,\ldots,2N\}\setminus\{1,N+1\})^2}\big) \\ &= E_1^{N,x} \det\big((B_{i,j})_{(i,j)\in(\{1,2,\ldots,2N\}\setminus\{1,N+1\})^2}\big). \end{aligned} \tag{3.23}$$

In addition, note that (3.20) proves that

$$\begin{aligned} \det\big((B_{i,j})_{(i,j)\in(\{1,\ldots,2N\}\setminus\{1,N+1\})^2}\big) &= \det(A^{N-1,(x_2,x_3,\ldots,x_{N+1})}) \\ &= \textstyle\prod_{i=1}^{N-1} E_i^{N-1,(x_2,x_3,\ldots,x_{N+1})} = \prod_{i=2}^{N} E_i^{N,x} > 0. \end{aligned} \tag{3.24}$$

Hence, we obtain that $\det(A^{N,x}) = \det(B) = \prod_{i=1}^{N} E_i^{N,x}$. Induction thus proves (3.18). Furthermore, observe that (3.18) establishes (3.15). The proof of Lemma 22 is thus complete. ∎

Proposition 23 *Let $N \in \mathbb{N}$, $v_1, v_2, \ldots, v_N \in \mathbb{R}\backslash\{0\}$, $x_0, x_1, \ldots, x_N \in \mathbb{R}$ satisfy $x_0 < x_1 < \cdots < x_N$, let $I_j \subseteq \mathbb{R}$, $j \in \{1, 2, \ldots, N\}$, satisfy for all $j \in \{1, 2, \ldots, N\}$ that $I_j = [x_{j-1}, x_N]$, let $\mathfrak{p}\colon [x_0, x_N] \to (0, \infty)$ be bounded and measurable, and let $A = (A_{i,j})_{(i,j)\in\{1,2,\ldots,2N\}^2} \in \mathbb{R}^{(2N)\times(2N)}$ satisfy for all $i, j \in \{1, 2, \ldots, N\}$ that*

$$A_{i,j} = 2v_i v_j \textstyle\int_{I_i \cap I_j} x^2 \mathfrak{p}(x)\,\mathrm{d}x, \qquad A_{N+i,j} = A_{i,N+j} = 2v_i v_j \textstyle\int_{I_i\cap I_j} x\mathfrak{p}(x)\,\mathrm{d}x, \qquad \textit{and} \qquad A_{N+i,N+j} = 2v_i v_j \textstyle\int_{I_i \cap I_j} \mathfrak{p}(x)\,\mathrm{d}x. \tag{3.25}$$

Then $\det(A) > 0$.

Proof [Proof of Proposition 23] Throughout this proof let $B = (B_{i,j})_{(i,j)\in\{1,2,\ldots,2N\}^2} \in \mathbb{R}^{(2N)\times(2N)}$ satisfy for all $i, j \in \{1, 2, \ldots, N\}$ that $B_{i,j} = \int_{I_i\cap I_j} x^2\mathfrak{p}(x)\,\mathrm{d}x$, $B_{N+i,j} = B_{i,N+j} = \int_{I_i\cap I_j} x\mathfrak{p}(x)\,\mathrm{d}x$, and $B_{N+i,N+j} = \int_{I_i\cap I_j} \mathfrak{p}(x)\,\mathrm{d}x$. Note that for all $i, j \in \{1, 2, \ldots, N\}$ it holds that

$$B_{i,j} = \textstyle\int_{x_{\max\{i-1,j-1\}}}^{x_N} x^2\mathfrak{p}(x)\,\mathrm{d}x, \qquad B_{N+i,j} = B_{i,N+j} = \textstyle\int_{x_{\max\{i-1,j-1\}}}^{x_N} x\mathfrak{p}(x)\,\mathrm{d}x, \qquad \text{and} \qquad B_{N+i,N+j} = \textstyle\int_{x_{\max\{i-1,j-1\}}}^{x_N} \mathfrak{p}(x)\,\mathrm{d}x. \tag{3.26}$$

Furthermore, observe that (3.25) and the fact that the determinant is linear in each row and each column show that

$$\det(A) = 4^N \left(\textstyle\prod_{i=1}^N |v_i|^4\right)\det(B). \tag{3.27}$$

In addition, note that (3.26) and Lemma 22 (applied with $a \curvearrowleft x_0$, $b \curvearrowleft x_N$, $\mathfrak{p} \curvearrowleft \mathfrak{p}$, $N \curvearrowleft N$, $x \curvearrowleft (x_0, x_1, \ldots, x_N)$ in the notation of Lemma 22) demonstrate that $\det(B) > 0$. Combining this with (3.27) ensures that $\det(A) > 0$. The proof of Proposition 23 is thus complete. ■

3.3 Regularity properties for the set of global minima of the risk function

Definition 24 (Tangent space) *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ be a set, and let $x \in \mathcal{M}$. Then we denote by $\mathcal{T}^x_{\mathcal{M}} \subseteq \mathbb{R}^{\mathfrak{d}}$ the set given by*

$$\mathcal{T}^x_{\mathcal{M}} = \left\{v \in \mathbb{R}^{\mathfrak{d}}\colon \left[\exists\, \gamma \in C^1(\mathbb{R}, \mathbb{R}^{\mathfrak{d}})\colon \left([\gamma(\mathbb{R}) \subseteq \mathcal{M}], [\gamma(0) = x], [\gamma'(0) = v]\right)\right]\right\}. \tag{3.28}$$

Lemma 25 *Let $\mathfrak{d}, k \in \mathbb{N}$, let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $f \in C^2(U, \mathbb{R})$ have locally Lipschitz continuous derivatives, let $\mathcal{M} \subseteq U$ satisfy $\mathcal{M} = \{x \in U\colon f(x) = \inf_{y\in U} f(y)\}$, assume that $\mathcal{M}$ is a k-dimensional C^2-submanifold of $\mathbb{R}^{\mathfrak{d}}$, and let $x \in \mathcal{M}$. Then*

(i) it holds for all $v \in \mathcal{T}^x_{\mathcal{M}}$ that $\big((\operatorname{Hess} f)(x)\big)v = 0$,

(ii) it holds that $\operatorname{rank}((\operatorname{Hess} f)(x)) \leq \mathfrak{d} - k$, and

(iii) it holds for all $v \in (\mathcal{T}^x_{\mathcal{M}})^\perp$ that $\big((\operatorname{Hess} f)(x)\big)v \in (\mathcal{T}^x_{\mathcal{M}})^\perp$

(cf. Definition 24).

Proof [Proof of Lemma 25] Observe that the assumption that $\mathcal{M} = \{y \in U : f(y) = \inf_{z \in U} f(z)\}$ ensures for all $y \in \mathcal{M}$ that $(\nabla f)(y) = 0$. This implies for all $\gamma \in C^1(\mathbb{R}, \mathbb{R}^{\mathfrak{d}})$, $t \in \mathbb{R}$ with $\gamma(\mathbb{R}) \subseteq \mathcal{M}$ that $(\nabla f)(\gamma(t)) = 0$. Hence, we obtain for all $\gamma \in C^1(\mathbb{R}, \mathbb{R}^{\mathfrak{d}})$, $t \in \mathbb{R}$ with $\gamma(\mathbb{R}) \subseteq \mathcal{M}$ that

$$0 = \tfrac{\mathrm{d}}{\mathrm{d}t}\big((\nabla f)(\gamma(t))\big) = \big((\operatorname{Hess} f)(\gamma(t))\big)\gamma'(t). \tag{3.29}$$

This shows for all $\gamma \in C^1(\mathbb{R}, \mathbb{R}^{\mathfrak{d}})$ with $\gamma(\mathbb{R}) \subseteq \mathcal{M}$ and $\gamma(0) = x$ that $((\operatorname{Hess} f)(x))\gamma'(0) = 0$. This establishes item (i).

Next note that the assumption that $\mathcal{M}$ is a k-dimensional C^2-submanifold of $\mathbb{R}^{\mathfrak{d}}$ proves that $\dim(\mathcal{T}^x_{\mathcal{M}}) = k$. Combining this with item (i) establishes item (ii).

Moreover, observe that item (i) and the fact that $(\operatorname{Hess} f)(x)$ is symmetric demonstrate for all $v \in \mathcal{T}^x_{\mathcal{M}}$, $w \in (\mathcal{T}^x_{\mathcal{M}})^{\perp}$ that

$$\big\langle v, \big((\operatorname{Hess} f)(x)\big)w\big\rangle = \big\langle\big((\operatorname{Hess} f)(x)\big)v, w\big\rangle = \langle 0, w\rangle = 0. \tag{3.30}$$

This establishes item (iii). The proof of Lemma 25 is thus complete. ■

Proposition 26 *Assume Setting 3, let $x_0, x_1, \ldots, x_H, \alpha_1, \alpha_2, \ldots, \alpha_H \in \mathbb{R}$ satisfy $a = x_0 < x_1 < \cdots < x_H = b$, assume for all $i \in \{1, 2, \ldots, H\}$, $x \in [x_{i-1}, x_i]$ that $f(x) = f(x_{i-1}) + \alpha_i(x - x_{i-1})$, assume $\prod_{i=1}^{H-1}(\alpha_{i+1} - \alpha_i) \neq 0$, and let $\mathfrak{D} \in \mathbb{R}$ satisfy*

$$\mathfrak{D} = 1 + |f(a)| + (1 + 2\max\nolimits_{j \in \{1,2,\ldots,H\}}|\alpha_j|)(1 + |a| + |b|). \tag{3.31}$$

Then there exists an open $U \subseteq (-\mathfrak{D}, \mathfrak{D})^{\mathfrak{d}}$ such that

(i) it holds that $U \subseteq \mathfrak{V}$,

(ii) it holds that $\mathcal{L}|_U \in C^2(U, \mathbb{R})$,

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

(iv) it holds for all $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in U$ that

$$\max\nolimits_{i,j \in \{1,2,\ldots,\mathfrak{d}\}}\big|\big(\tfrac{\partial^2}{\partial\theta_i\partial\theta_j}\mathcal{L}\big)(\theta)\big| \leq \big(24\mathfrak{D}^5 + 16H\mathfrak{D}^7\big)\big(\sup\nolimits_{x \in [a,b]}\mathfrak{p}(x)\big), \tag{3.32}$$

(v) it holds that $\{\vartheta \in U : \mathcal{L}(\vartheta) = 0\} \neq \varnothing$,

(vi) it holds that $\{\vartheta \in U : \mathcal{L}(\vartheta) = 0\}$ is a $(H+1)$-dimensional C^∞-submanifold of $\mathbb{R}^{\mathfrak{d}}$, and

(vii) it holds for all $\theta \in \{\vartheta \in U : \mathcal{L}(\vartheta) = 0\}$ that $\operatorname{rank}((\operatorname{Hess}\mathcal{L})(\theta)) = 2H = \mathfrak{d} - (H+1)$.

Proof [Proof of Proposition 26] Throughout this proof let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy

$$\begin{split} U = \big\{\theta \in (-\mathfrak{D}, \mathfrak{D})^{\mathfrak{d}} \colon \big(\big[\min\{\mathfrak{w}^\theta_1 a + \mathfrak{b}^\theta_1, \mathfrak{w}^\theta_1 b + \mathfrak{b}^\theta_1, \mathfrak{v}^\theta_1\} > 0\big], \big[\forall\, j \in \mathbb{N} \cap (1, H] \colon \mathfrak{w}^\theta_j > 1/2\big], \\ \big[\forall\, j \in \mathbb{N} \cap (1, H] \colon \mathfrak{q}^\theta_j \in (a, b)\big], \big[\forall\, j \in \mathbb{N} \cap (1, H) \colon \mathfrak{q}^\theta_j < \mathfrak{q}^\theta_{j+1}\big]\big)\big\} \end{split} \tag{3.33}$$

and let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ be given by

$$\mathcal{M} = \big\{\theta \in (-\mathfrak{D},\mathfrak{D})^{\mathfrak{d}} \colon \big(\big[\min\{\mathfrak{w}^{\theta}_1 a + \mathfrak{b}^{\theta}_1, \mathfrak{w}^{\theta}_1 b + \mathfrak{b}^{\theta}_1, \mathfrak{v}^{\theta}_1\} > 0\big],\, [\mathfrak{v}^{\theta}_1(\mathfrak{w}^{\theta}_1 a + \mathfrak{b}^{\theta}_1) + \mathfrak{c}^{\theta} = f(a)], \\ [\mathfrak{w}^{\theta}_1\mathfrak{v}^{\theta}_1 = \alpha_1],\, \big[\forall\, j \in \mathbb{N} \cap (1,H] \colon \mathfrak{w}^{\theta}_j > 1/2,\, \mathfrak{q}^{\theta}_j = x_{j-1},\, \mathfrak{w}^{\theta}_j\mathfrak{v}^{\theta}_j = \alpha_j - \alpha_{j-1}\big]\big)\big\}. \tag{3.34}$$

Note that (3.33) ensures that U is open. Furthermore, observe that (2.3) and (3.33) assure that $U \subseteq \mathfrak{V}$. This proves item (i). In addition, note that item (i), Lemma 15, and the fact that U is open establish items (ii) and (iii).

Next observe that Corollary 19, the fact that for all $\theta \in U$, $j \in \{1,2,\ldots,H\}$ with $\mathfrak{q}^{\theta}_j \in [a,b]$ it holds that $\mathfrak{w}^{\theta}_j > \frac{1}{2}$, and the fact that $\mathfrak{D} \geq \max\{|a|,|b|,b-a,\sup_{x\in[a,b]}|f(x)|,1\} \geq 1$ prove that for all $\theta \in U \subseteq (-\mathfrak{D},\mathfrak{D})^{\mathfrak{d}}$ we have that

$$\begin{aligned} \max_{i,j\in\{1,2,\ldots,\mathfrak{d}\}}\big|\big(\tfrac{\partial^2}{\partial\theta_i\partial\theta_j}\mathcal{L}\big)(\theta)\big| &\leq \big(16\mathfrak{D}^5 + 16H\mathfrak{D}^7 + 8\mathfrak{D}^4\big(\sup\nolimits_{x\in[a,b]}|f(x)|\big)\big)\big(\sup\nolimits_{x\in[a,b]}\mathfrak{p}(x)\big) \\ &\leq \big(24\mathfrak{D}^5 + 16H\mathfrak{D}^7\big)\big(\sup\nolimits_{x\in[a,b]}\mathfrak{p}(x)\big). \end{aligned} \tag{3.35}$$

This establishes item (iv).

Next note that (3.34) and Lemma 21 imply that $\mathcal{M}$ is a non-empty $(H+1)$-dimensional C^{∞}-submanifold of $\mathbb{R}^{\mathfrak{d}}$. Furthermore, observe that (3.33), (3.34), and the fact that $a < x_1 < x_2 < \cdots < x_H = b$ show that $\mathcal{M} \subseteq U$. In the next step we intend to prove that for all $\theta \in \mathcal{M}$ it holds that $\mathcal{L}(\theta) = 0$. Note that (3.33) and the fact that for all $\theta \in U$, $x \in [a,b]$ it holds that

$$\mathfrak{w}^{\theta}_1 x + \mathfrak{b}^{\theta}_1 = \Big[\tfrac{b-x}{b-a}\Big](\mathfrak{w}^{\theta}_1 a + \mathfrak{b}^{\theta}_1) + \Big[\tfrac{x-a}{b-a}\Big](\mathfrak{w}^{\theta}_1 b + \mathfrak{b}^{\theta}_1) > 0 \tag{3.36}$$

ensure that for all $\theta \in U$, $x \in [a,b]$ we have that

$$\begin{aligned} \mathscr{N}^{\theta}(x) &= \mathfrak{c}^{\theta} + \mathfrak{v}^{\theta}_1\max\{\mathfrak{w}^{\theta}_1 x + \mathfrak{b}^{\theta}_1, 0\} + \textstyle\sum_{j=2}^{H}\mathfrak{v}^{\theta}_j\max\{\mathfrak{w}^{\theta}_j x + \mathfrak{b}^{\theta}_j, 0\} \\ &= \mathfrak{c}^{\theta} + \mathfrak{v}^{\theta}_1(\mathfrak{w}^{\theta}_1 x + \mathfrak{b}^{\theta}_1) + \textstyle\sum_{j=2}^{H}\mathfrak{v}^{\theta}_j\mathfrak{w}^{\theta}_j\max\{x - \mathfrak{q}^{\theta}_j, 0\}. \end{aligned} \tag{3.37}$$

Combining this with (3.34) demonstrates that for all $\theta \in \mathcal{M}$, $x \in [a,b]$ we have that

$$\begin{aligned} \mathscr{N}^{\theta}(x) &= \mathfrak{v}^{\theta}_1\mathfrak{w}^{\theta}_1 x + \mathfrak{v}^{\theta}_1\mathfrak{b}^{\theta}_1 + \mathfrak{c}^{\theta} + \textstyle\sum_{j=2}^{H}\mathfrak{v}^{\theta}_j\mathfrak{w}^{\theta}_j\max\{x - x_{j-1}, 0\} \\ &= \mathfrak{v}^{\theta}_1\mathfrak{w}^{\theta}_1 x + f(a) - \mathfrak{v}^{\theta}_1\mathfrak{w}^{\theta}_1 a + \textstyle\sum_{j=2}^{H}\mathfrak{v}^{\theta}_j\mathfrak{w}^{\theta}_j\max\{x - x_{j-1}, 0\} \\ &= f(a) + \alpha_1(x-a) + \textstyle\sum_{j=2}^{H}(\alpha_j - \alpha_{j-1})\max\{x - x_{j-1}, 0\}. \end{aligned} \tag{3.38}$$

In addition, observe that the assumption that for all $i \in \{1,2,\ldots H\}$, $x \in [x_{i-1}, x_i]$ it holds that $f(x) = f(x_{i-1}) + \alpha_i(x - x_{i-1})$ proves that for all $j \in \{0,1,\ldots,H-1\}$, $x \in [x_j, x_{j+1}]$

we have that

$$
\begin{aligned}
f(x) &= f(x_0) + \big[\textstyle\sum_{k=1}^{j}[f(x_k) - f(x_{k-1})]\big] + [f(x) - f(x_j)] \\
&= f(a) + \big[\textstyle\sum_{k=1}^{j} \alpha_k(x_k - x_{k-1})\big] + \alpha_{j+1}(x - x_j) \\
&= f(a) + \alpha_{j+1}x + \big[\textstyle\sum_{k=1}^{j} \alpha_k(x_k - x_{k-1})\big] - \alpha_{j+1}x_j \\
&= f(a) + \alpha_{j+1}x + \big[\textstyle\sum_{k=1}^{j} \alpha_k x_k\big] - \big[\textstyle\sum_{k=1}^{j} \alpha_k x_{k-1}\big] - \alpha_{j+1}x_j \\
&= f(a) + \alpha_{j+1}x - \big(\big[\textstyle\sum_{k=1}^{j+1} \alpha_k x_{k-1}\big] - \big[\textstyle\sum_{k=1}^{j} \alpha_k x_k\big]\big) \\
&= f(a) + \alpha_{j+1}x - \big(\alpha_1 x_0 + \big[\textstyle\sum_{k=2}^{j+1} \alpha_k x_{k-1}\big] - \big[\textstyle\sum_{k=2}^{j+1} \alpha_{k-1} x_{k-1}\big]\big) \\
&= f(a) + \big(\alpha_1 x + \big[\textstyle\sum_{k=2}^{j+1}(\alpha_k - \alpha_{k-1})x\big]\big) - \big(\alpha_1 x_0 + \big[\textstyle\sum_{k=2}^{j+1}(\alpha_k - \alpha_{k-1})x_{k-1}\big]\big) \\
&= f(a) + \alpha_1(x - a) + \textstyle\sum_{k=2}^{j+1}(\alpha_k - \alpha_{k-1})(x - x_{k-1}) \\
&= f(a) + \alpha_1(x - a) + \textstyle\sum_{k=2}^{H}(\alpha_k - \alpha_{k-1})\max\{x - x_{k-1}, 0\}.
\end{aligned}
\tag{3.39}
$$

This implies that for all $x \in [a, b]$ we have that

$$
f(x) = f(a) + \alpha_1(x - a) + \textstyle\sum_{j=2}^{H}(\alpha_j - \alpha_{j-1})\max\{x - x_{j-1}, 0\}. \tag{3.40}
$$

Combining this with (3.38) demonstrates that for all $\theta \in \mathcal{M}$, $x \in [a, b]$ it holds that $\mathscr{N}^\theta(x) = f(x)$. Hence, we obtain for all $\theta \in \mathcal{M}$ that $\mathcal{L}(\theta) = 0$. Next we intend to prove that for all $\theta \in U$ with $\mathcal{L}(\theta) = 0$ it holds that $\theta \in \mathcal{M}$. Note that (2.2) and the fact that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that $[a, b] \ni x \mapsto \mathscr{N}^\theta(x) - f(x) \in \mathbb{R}$ is continuous show that for all $\theta \in \{\vartheta \in U \colon \mathcal{L}(\vartheta) = 0\} \subseteq \mathbb{R}^{\mathfrak{d}}$, $x \in [a, b]$ we have that

$$
\mathscr{N}^\theta(x) = f(x). \tag{3.41}
$$

Combining this with (3.33), (3.34), (3.37), and the fact that $\mathcal{M} \subseteq U$ demonstrates for all $\theta \in \{\vartheta \in U \colon \mathcal{L}(\vartheta) = 0\}$, $x \in [x_0, x_1 + \min\{0, (\mathfrak{q}^\theta_{\min\{2,H\}} - x_1)\mathbb{1}_{(1,\infty)}(H)\}]$ that

$$
f(a)+\alpha_1(x-a) = f(x) = \mathscr{N}^\theta(x) = \mathfrak{v}^\theta_1(\mathfrak{w}^\theta_1 x+\mathfrak{b}^\theta_1)+\mathfrak{c}^\theta = \mathfrak{v}^\theta_1\mathfrak{w}^\theta_1(x-a)+\mathfrak{v}^\theta_1(\mathfrak{w}^\theta_1 a+\mathfrak{b}^\theta_1)+\mathfrak{c}^\theta. \tag{3.42}
$$

The fact that for all $\theta \in U$ it holds that $x_1 + \min\{0, (\mathfrak{q}^\theta_{\min\{2,H\}} - x_1)\mathbb{1}_{(1,\infty)}(H)\} > x_0$ hence ensures that for all $\theta \in \{\vartheta \in U \colon \mathcal{L}(\vartheta) = 0\}$ we have that

$$
\mathfrak{w}^\theta_1\mathfrak{v}^\theta_1 = \alpha_1 \qquad \text{and} \qquad \mathfrak{v}^\theta_1(\mathfrak{w}^\theta_1 a + \mathfrak{b}^\theta_1) + \mathfrak{c}^\theta = f(a). \tag{3.43}
$$

Next observe that the fact that for all $\theta \in U$ it holds that $(a, b)\backslash\{\mathfrak{q}^\theta_1, \mathfrak{q}^\theta_2, \ldots, \mathfrak{q}^\theta_H\}$ is an open set shows that there exists $\varepsilon = (\varepsilon_{\theta,x})_{(\theta,x)\in U\times\mathbb{R}} \colon U \times \mathbb{R} \to (0, \infty)$ which satisfies for all $\theta \in U$, $x \in (a, b)\backslash\{\mathfrak{q}^\theta_1, \mathfrak{q}^\theta_2, \ldots, \mathfrak{q}^\theta_H\}$ that $(x - \varepsilon_{\theta,x}, x + \varepsilon_{\theta,x}) \subseteq (a, b)\backslash\{\mathfrak{q}^\theta_1, \mathfrak{q}^\theta_2, \ldots, \mathfrak{q}^\theta_H\}$. Combining this with (3.33) and (3.37) demonstrates for all $\theta \in U$, $x \in (a, b)\backslash\{\mathfrak{q}^\theta_1, \mathfrak{q}^\theta_2, \ldots, \mathfrak{q}^\theta_H\}$ that $(x - \varepsilon_{\theta,x}, x + \varepsilon_{\theta,x}) \ni y \mapsto \mathscr{N}^\theta(y) \in \mathbb{R}$ is affine linear. This, (3.40), (3.41), and the fact that for all $i \in \mathbb{N} \cap [1, H)$ it holds that $\alpha_{i+1} \neq \alpha_i$ prove that for all $\theta \in \{\vartheta \in U \colon \mathcal{L}(\vartheta) = 0\}$, $i \in \mathbb{N} \cap [1, H)$ we have that $x_i \in \{\mathfrak{q}^\theta_1, \mathfrak{q}^\theta_2, \ldots, \mathfrak{q}^\theta_H\}$. Combining this with the fact that for all $\theta \in U$ it holds that $\mathfrak{q}^\theta_1 \notin [a, b]$, the fact that for all $\theta \in U$, $j \in \mathbb{N} \cap (1, H]$ it holds that $\mathfrak{q}^\theta_j \in (a, b)$, the fact that for all $\theta \in U$, $j \in \mathbb{N} \cap (1, H)$ it holds that $\mathfrak{q}^\theta_j < \mathfrak{q}^\theta_{j+1}$, and the fact that $a < x_1 < x_2 < \cdots < x_H = b$ shows that for all $\theta \in \{\vartheta \in U \colon \mathcal{L}(\vartheta) = 0\}$,

$j \in \mathbb{N} \cap (1, H]$ we have that $\mathfrak{q}^\theta_j = x_{j-1}$. This, (3.36), (3.40), (3.41), and (3.43) assure that for all $\theta \in \{\vartheta \in U \colon \mathcal{L}(\vartheta) = 0\}$, $x \in [a, b]$ it holds that

$$\begin{aligned}
& f(a) + \alpha_1(x-a) + \textstyle\sum_{j=2}^H (\alpha_j - \alpha_{j-1}) \max\{x - x_{j-1}, 0\} = f(x) \\
& = \mathscr{N}^\theta(x) = \mathfrak{c}^\theta + \textstyle\sum_{j=1}^H \mathfrak{v}^\theta_j \max\{\mathfrak{w}^\theta_j x + \mathfrak{b}^\theta_j, 0\} \\
& = \mathfrak{c}^\theta + \mathfrak{v}^\theta_1 \max\{\mathfrak{w}^\theta_1 x + \mathfrak{b}^\theta_1, 0\} + \textstyle\sum_{j=2}^H \mathfrak{v}^\theta_j \mathfrak{w}^\theta_j \max\{x + (\mathfrak{w}^\theta_j)^{-1} \mathfrak{b}^\theta_j, 0\} \\
& = \mathfrak{c}^\theta + \mathfrak{v}^\theta_1 (\mathfrak{w}^\theta_1 x + \mathfrak{b}^\theta_1) + \textstyle\sum_{j=2}^H \mathfrak{v}^\theta_j \mathfrak{w}^\theta_j \max\{x - \mathfrak{q}^\theta_j, 0\} \\
& = \mathfrak{c}^\theta + \mathfrak{v}^\theta_1 \mathfrak{w}^\theta_1 (x-a) + \mathfrak{v}^\theta_1 \mathfrak{w}^\theta_1 a + \mathfrak{v}^\theta_1 \mathfrak{b}^\theta_1 + \textstyle\sum_{j=2}^H \mathfrak{v}^\theta_j \mathfrak{w}^\theta_j \max\{x - x_{j-1}, 0\} \\
& = (\mathfrak{c}^\theta + \mathfrak{v}^\theta_1 \mathfrak{w}^\theta_1 a + \mathfrak{v}^\theta_1 \mathfrak{b}^\theta_1) + \alpha_1(x-a) + \textstyle\sum_{j=2}^H \mathfrak{v}^\theta_j \mathfrak{w}^\theta_j \max\{x - x_{j-1}, 0\} \\
& = f(a) + \alpha_1(x-a) + \textstyle\sum_{j=2}^H \mathfrak{v}^\theta_j \mathfrak{w}^\theta_j \max\{x - x_{j-1}, 0\}.
\end{aligned} \tag{3.44}$$

Hence, we obtain for all $\theta \in \{\vartheta \in U \colon \mathcal{L}(\vartheta) = 0\}$, $j \in \mathbb{N} \cap (1, H]$ that $\mathfrak{v}^\theta_j \mathfrak{w}^\theta_j = \alpha_j - \alpha_{j-1}$. Combining this with (3.43) proves for all $\theta \in \{\vartheta \in U \colon \mathcal{L}(\vartheta) = 0\}$ that $\theta \in \mathcal{M}$. Hence, we obtain that $\mathcal{M} = \{\vartheta \in U \colon \mathcal{L}(\vartheta) = 0\}$. This and the fact that $\mathcal{M}$ is a non-empty $(H+1)$-dimensional C^∞-submanifold of $\mathbb{R}^{\mathfrak{d}}$ establish items (v) and (vi).

In the next step note that (3.36) ensures for all $\theta \in \mathcal{M}$ that $I^\theta_1 = [a, b]$. In addition, observe that (3.34) shows for all $\theta \in \mathcal{M}$, $j \in \mathbb{N} \cap (1, H]$ that $I^\theta_j = (x_{j-1}, b]$. Furthermore, note that (3.34) and the fact that for all $j \in \mathbb{N} \cap (1, H]$ it holds that $\alpha_j - \alpha_{j-1} \neq 0$ demonstrate that for all $\theta \in \mathcal{M}$, $i \in \mathbb{N} \cap [1, H]$ it holds that $\mathfrak{v}^\theta_i \neq 0$. This, Corollary 16, and Proposition 23 assure for all $\theta \in \mathcal{M}$ that $\det\big(\big((\tfrac{\partial^2}{\partial \theta_i \partial \theta_j} \mathcal{L})(\theta)\big)_{(i,j) \in \{1,2,\ldots,2H\}^2}\big) \neq 0$. Hence, we obtain for all $\theta \in \mathcal{M}$ that

$$\operatorname{rank}((\operatorname{Hess} \mathcal{L})(\theta)) \geq 2H. \tag{3.45}$$

Moreover, observe that the fact that $\mathcal{M} = \{\vartheta \in U \colon \mathcal{L}(\vartheta) = 0\}$ is a $(H+1)$-dimensional C^∞-submanifold of $\mathbb{R}^{\mathfrak{d}}$ and Lemma 25 imply that for all $\theta \in \mathcal{M}$ we have that $\operatorname{rank}((\operatorname{Hess} \mathcal{L})(\theta)) \leq \mathfrak{d} - (H+1) = 2H$. This and (3.45) establish item (vii). The proof of Proposition 26 is thus complete. ■

Definition 27 *Let $n \in \mathbb{N}$ and let $A \in \mathbb{R}^{n \times n} \backslash \{0\}$ be symmetric. Then we denote by $\sigma(A) \in (0, \infty)$ the real number given by*

$$\sigma(A) = \min\{\ell \in (0, \infty) \colon [\exists\, \lambda \in \{-\ell, \ell\}, v \in \mathbb{R}^n \backslash \{0\} \colon Av = \lambda v]\} \tag{3.46}$$

and we denote by $\Lambda(A) \in (0, \infty)$ the real number given by

$$\Lambda(A) = \max\{\ell \in (0, \infty) \colon [\exists\, \lambda \in \{-\ell, \ell\}, v \in \mathbb{R}^n \backslash \{0\} \colon Av = \lambda v]\} \tag{3.47}$$

Note that Definition 27 above ensures that $\sigma(A)$ is the smallest absolute value of a nonzero eigenvalue of A and $\Lambda(A)$ is the largest absolute value of a nonzero eigenvalue of A.

Lemma 28 *Let $n \in \mathbb{N}$ and let $A = (a_{i,j})_{(i,j) \in \{1,2,\ldots,n\}^2} \in \mathbb{R}^{n \times n} \backslash \{0\}$ be symmetric. Then $\Lambda(A) \leq \big[\sum_{i,j=1}^n |a_{i,j}|^2\big]^{1/2}$ (cf. Definition 27).*

Corollary 29 *Assume Setting 3, let $N \in \mathbb{N} \cap [1, H]$, $\mathscr{x}_0, \mathscr{x}_1, \ldots, \mathscr{x}_N, \alpha_1, \alpha_2, \ldots, \alpha_N \in \mathbb{R}$ satisfy $a = \mathscr{x}_0 < \mathscr{x}_1 < \cdots < \mathscr{x}_N = b$, assume for all $i \in \{1, 2, \ldots, N\}$, $x \in [\mathscr{x}_{i-1}, \mathscr{x}_i]$ that $f(x) = f(\mathscr{x}_{i-1}) + \alpha_i(x - \mathscr{x}_{i-1})$, and let $\mathfrak{D} \in \mathbb{R}$ satisfy*

$$\mathfrak{D} = 1 + |f(a)| + (1 + 2\max\nolimits_{j \in \{1,2,\ldots,N\}} |\alpha_j|)(1 + |a| + |b|). \tag{3.48}$$

Then there exist $k \in \mathbb{N} \cap [1, \mathfrak{d})$ and an open $U \subseteq (-\mathfrak{D}, \mathfrak{D})^{\mathfrak{d}}$ such that

(i) it holds that $U \subseteq \mathfrak{V}$,

(ii) it holds that $\mathcal{L}|_U \in C^2(U, \mathbb{R})$,

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

(iv) it holds for all $\theta \in U$ that

$$\Lambda((\operatorname{Hess} \mathcal{L})(\theta)) \leq (3N + 1)\big(24\mathfrak{D}^5 + 16N\mathfrak{D}^7\big)\big(\sup\nolimits_{x \in [a,b]} \mathfrak{p}(x)\big), \tag{3.49}$$

(v) it holds that $\{\vartheta \in U \colon \mathcal{L}(\vartheta) = 0\} \neq \varnothing$,

(vi) it holds that $\{\vartheta \in U \colon \mathcal{L}(\vartheta) = 0\}$ is a k-dimensional C^∞-submanifold of $\mathbb{R}^{\mathfrak{d}}$,

(vii) it holds for all $\theta \in \{\vartheta \in U \colon \mathcal{L}(\vartheta) = 0\}$ that $\operatorname{rank}((\operatorname{Hess} \mathcal{L})(\theta)) = \mathfrak{d} - k$, and

(viii) it holds that $k = \mathfrak{d} - 2[N - \#\{i \in \mathbb{N} \cap [1, N) \colon \alpha_i = \alpha_{i+1}\}]$

(cf. Definition 27).

Proof [Proof of Corollary 29] Throughout this proof we assume without loss of generality that $\prod_{i=1}^{N-1}(\alpha_{i+1} - \alpha_i) \neq 0$. (Otherwise, we can remove the points $\mathscr{x}_i \in (a, b)$ which satisfy $\alpha_{i+1} = \alpha_i$. This decreases the number N of breakpoints and changes neither the target function f nor the parameter space $\mathbb{R}^{\mathfrak{d}}$. It also does not change the number k in item (viii).)

In the following let $P \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{3N+1}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $P(\theta) = (\mathfrak{w}^\theta_1, \ldots, \mathfrak{w}^\theta_N, \mathfrak{b}^\theta_1, \ldots, \mathfrak{b}^\theta_N, \mathfrak{v}^\theta_1, \ldots, \mathfrak{v}^\theta_N, \mathfrak{c}^\theta)$, and let $\mathscr{L} \colon \mathbb{R}^{3N+1} \to \mathbb{R}$ satisfy for all $\theta = (\theta_1, \ldots, \theta_{3N+1}) \in \mathbb{R}^{3N+1}$ that

$$\mathscr{L}(\theta) = \textstyle\int_a^b \big(f(x) - \theta_{3N+1} - \sum_{j=1}^N \theta_{2N+j}[\mathfrak{R}(\theta_j x + \theta_{N+j})]\big)^2 \mathfrak{p}(x)\, \mathrm{d}x. \tag{3.50}$$

Observe that Proposition 26 (applied with $H \curvearrowleft N$, $\mathcal{L} \curvearrowleft \mathscr{L}$ in the notation of Proposition 26) demonstrates that there exists an open $V \subseteq (-\mathfrak{D}, \mathfrak{D})^{3N+1}$ which satisfies that

(I) it holds that

$$V \subseteq \big\{\theta = (\theta_1, \ldots, \theta_{3N+1}) \in \mathbb{R}^{3N+1} \colon \big(\textstyle\prod_{j=1}^N \prod_{v \in \{a,b\}} (\theta_j v + \theta_{N+j}) \neq 0\big)\big\}, \tag{3.51}$$

(II) it holds that $\mathscr{L}|_V \in C^2(V, \mathbb{R})$,

(III) it holds for all $\theta = (\theta_1, \ldots, \theta_{3N+1}) \in V$ that

$$\max\nolimits_{i,j \in \{1,2,\ldots,3N+1\}} \big|\big(\tfrac{\partial^2}{\partial\theta_i \partial\theta_j} \mathscr{L}\big)(\theta)\big| \leq \big(24\mathfrak{D}^5 + 16N\mathfrak{D}^7\big)\big(\sup\nolimits_{x \in [a,b]} \mathfrak{p}(x)\big), \tag{3.52}$$

(IV) it holds that $\{\vartheta \in V : \mathscr{L}(\vartheta) = 0\} \neq \varnothing$,

(V) it holds that $\{\vartheta \in V : \mathscr{L}(\vartheta) = 0\}$ is an $(N+1)$-dimensional C^∞-submanifold of $\mathbb{R}^{3N+1}$, and

(VI) it holds for all $\theta \in \{\vartheta \in V : \mathscr{L}(\vartheta) = 0\}$ that $\operatorname{rank}((\operatorname{Hess} \mathscr{L})(\theta)) = 2N = (3N+1) - (N+1)$.

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

$$U = \bigl\{\theta \in (-\mathfrak{D}, \mathfrak{D})^{\mathfrak{d}} \cap (P^{-1}(V)) \colon \bigl(\forall\, j \in \mathbb{N} \cap (N, H] \colon \max\bigl\{\mathfrak{w}^\theta_j a + \mathfrak{b}^\theta_j, \mathfrak{w}^\theta_j b + \mathfrak{b}^\theta_j\bigr\} < 0\bigr)\bigr\}. \tag{3.53}$$

Note that (3.53) assures that $U \subseteq \mathbb{R}^{\mathfrak{d}}$ is open. In addition, observe that (2.3), (3.53), and item (I) imply that $U \subseteq \mathfrak{V}$. This establishes item (i). Next note that item (i) and Lemma 15 prove items (ii) and (iii). Furthermore, observe that for all $\theta \in U$, $x \in [a, b]$, $i \in \mathbb{N} \cap (N, H]$ it holds that $\mathfrak{R}(\mathfrak{w}^\theta_i x + \mathfrak{b}^\theta_i) = 0$. Therefore, we obtain for all $\theta \in U$, $x \in [a, b]$ that

$$\mathscr{N}^\theta(x) = \mathfrak{c}^\theta + \textstyle\sum_{j=1}^H \mathfrak{v}^\theta_j \bigl[\mathfrak{R}(\mathfrak{w}^\theta_j x + \mathfrak{b}^\theta_j)\bigr] = \mathfrak{c}^\theta + \sum_{j=1}^N \mathfrak{v}^\theta_j \bigl[\mathfrak{R}(\mathfrak{w}^\theta_j x + \mathfrak{b}^\theta_j)\bigr]. \tag{3.54}$$

This implies for all $\theta \in U$ that

$$\mathcal{L}(\theta) = \mathscr{L}(P(\theta)). \tag{3.55}$$

Combining this with (3.52) ensures for all $\theta \in U$, $i, j \in \mathbb{N} \cap ((0, N] \cup (H, H+N] \cup (2H, 2H+N] \cup \{3H+1\})$ that

$$\bigl|\bigl(\tfrac{\partial^2}{\partial\theta_i \partial\theta_j} \mathcal{L}\bigr)(\theta)\bigr| \leq \bigl(24\mathfrak{D}^5 + 16N\mathfrak{D}^7\bigr)\bigl(\sup\nolimits_{x \in [a,b]} \mathfrak{p}(x)\bigr). \tag{3.56}$$

Moreover, note that (3.55) shows that for all $\theta \in U$, $i \in \{1, 2, \ldots, \mathfrak{d}\} \backslash ((0, N] \cup (H, H+N] \cup (2H, 2H+N] \cup \{3H+1\})$, $j \in \{1, 2, \ldots, \mathfrak{d}\}$ we have that

$$\bigl(\tfrac{\partial^2}{\partial\theta_i \partial\theta_j} \mathcal{L}\bigr)(\theta) = 0. \tag{3.57}$$

Combining this with Lemma 28 and (3.56) assures for all $\theta \in U$ that

$$\Lambda((\operatorname{Hess} \mathcal{L})(\theta)) \leq \sqrt{\textstyle\sum_{i,j=1}^H \bigl|\bigl(\tfrac{\partial^2}{\partial\theta_i \partial\theta_j} \mathscr{L}\bigr)(\theta)\bigr|^2} \leq (3N+1)\bigl(24\mathfrak{D}^5 + 16N\mathfrak{D}^7\bigr)\bigl(\sup\nolimits_{x \in [a,b]} \mathfrak{p}(x)\bigr). \tag{3.58}$$

This establishes item (iv). Furthermore, observe that items (IV) and (V), (3.53), and (3.55) establish items (v), (vi) and (viii). In addition, note that (3.55), (3.57), and item (VI) demonstrate for all $\theta \in \{\vartheta \in U : \mathcal{L}(\vartheta) = 0\}$ that $\operatorname{rank}((\operatorname{Hess} \mathcal{L})(\theta)) = 2N$. Combining this with item (viii) establishes item (vii). The proof of Corollary 29 is thus complete. ■

4. Local convergence to the set of global minima for gradient flow (GF)

In this section we employ Corollary 29 from Section 3 to establish in Proposition 44 in Subsection 4.3 below and Corollary 45 in Subsection 4.4 below that the risk of certain solutions of GF differential equations converges under the assumption that the target function is piecewise constant exponentially quick to zero. Our proof of Proposition 44 employs the abstract local convergence result for GF trajectories in Proposition 43 in Subsection 4.2. Proposition 43 and its proof are strongly inspired by (Fehrman et al., 2020, Proposition 16). Our proofs of Propositions 43 and 44 also use the several well-known concepts and results from differential geometry which we recall in Subsection 4.1 below.

In particular, Lemma 33 is a direct consequence of, e.g., (Fehrman et al., 2020, Proposition 7), Lemma 35 is proved as, e.g., (Fehrman et al., 2020, Lemma 10), Lemma 36 is proved as, e.g., (Fehrman et al., 2020, Lemma 11), Definition 37 is a slight reformulation of, e.g., (Fehrman et al., 2020, Definition 12), Proposition 39 is a slight extension of, e.g., (Fehrman et al., 2020, Proposition 13), Proposition 41 is a reformulation of (Fehrman et al., 2020, Lemma 15), and Lemma 42 is a slight generalization of (Fehrman et al., 2020, Lemma 14).

4.1 Differential geometric preliminaries

Definition 30 *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy $\mathcal{M} \neq \varnothing$. Then we denote by $\mathscr{d}_{\mathcal{M}} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ the function which satisfies for all $x \in \mathbb{R}^{\mathfrak{d}}$ that $\mathscr{d}_{\mathcal{M}}(x) = \inf_{y \in \mathcal{M}} \|x - y\|$.*

Definition 31 *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy $\mathcal{M} \neq \varnothing$. Then we denote by $\mathscr{P}_{\mathcal{M}} \subseteq \mathbb{R}^{\mathfrak{d}}$ the set given by*

$$\mathscr{P}_{\mathcal{M}} = \big\{x \in \mathbb{R}^{\mathfrak{d}} \colon (\exists_1 \, y \in \mathcal{M} \colon \|x - y\| = \mathscr{d}_{\mathcal{M}}(x))\big\} \tag{4.1}$$

and we denote by $\mathscr{p}_{\mathcal{M}} \colon \mathscr{P}_{\mathcal{M}} \to \mathbb{R}^{\mathfrak{d}}$ the function which satisfies for all $x \in \mathscr{P}_{\mathcal{M}}$ that $\mathscr{p}_{\mathcal{M}}(x) \in \mathcal{M}$ and

$$\|x - \mathscr{p}_{\mathcal{M}}(x)\| = \mathscr{d}_{\mathcal{M}}(x) \tag{4.2}$$

(cf. Definition 30).

Definition 32 *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy $\mathcal{M} \neq \varnothing$. Then we denote by $\mathbf{P}_{\mathcal{M}} \subseteq \mathbb{R}^{\mathfrak{d}}$ the set given by*

$$\mathbf{P}_{\mathcal{M}} = \bigcup\nolimits_{\substack{U \subseteq \mathbb{R}^{\mathfrak{d}} \text{ is open, } U \subseteq \mathscr{P}_{\mathcal{M}}, \\ \text{and } \mathscr{p}_{\mathcal{M}}|_U \in C^1(U, \mathbb{R}^{\mathfrak{d}})}} U \tag{4.3}$$

(cf. Definition 31).

Lemma 33 *Let $\mathfrak{d}, k \in \mathbb{N}$, let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ be a k-dimensional C^2-submanifold of $\mathbb{R}^{\mathfrak{d}}$, and let $x \in \mathcal{M}$. Then there exists an open $V \subseteq \mathbb{R}^{\mathfrak{d}}$ such that*

(i) it holds that $x \in V \subseteq \mathscr{P}_{\mathcal{M}}$ and

(ii) it holds that $\mathscr{p}_{\mathcal{M}}|_V \in C^1(V, \mathbb{R}^{\mathfrak{d}})$.

(cf. Definitions 30 and 31).

Proposition 34 *Let $\mathfrak{d}, k \in \mathbb{N}$ and let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ be a non-empty k-dimensional C^2-submanifold of $\mathbb{R}^{\mathfrak{d}}$. Then $\mathcal{M} \subseteq \mathbf{P}_{\mathcal{M}}$ (cf. Definition 32).*

Proof [Proof of Proposition 34] Observe that Lemma 33 assures that $\mathcal{M} \subseteq \mathbf{P}_{\mathcal{M}}$. The proof of Proposition 34 is thus complete. ∎

Lemma 35 *Let $\mathfrak{d}, k \in \mathbb{N}$, let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ be a non-empty k-dimensional C^2-submanifold of $\mathbb{R}^{\mathfrak{d}}$, and let $x \in \mathbf{P}_{\mathcal{M}}$ (cf. Definition 32). Then $x - \mathscr{p}_{\mathcal{M}}(x) \in (\mathcal{T}_{\mathcal{M}}^{\mathscr{p}_{\mathcal{M}}(x)})^{\perp}$ (cf. Definitions 24 and 31).*

Lemma 36 *Let $\mathfrak{d}, k \in \mathbb{N}$ and let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ be a non-empty k-dimensional C^2-submanifold of $\mathbb{R}^{\mathfrak{d}}$. Then*

(i) it holds that $\mathbf{P}_{\mathcal{M}} \backslash \mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ is open,

(ii) it holds that $\mathbf{P}_{\mathcal{M}} \backslash \mathcal{M} \ni y \mapsto \mathscr{d}_{\mathcal{M}}(y) \in \mathbb{R}$ is continuously differentiable, and

(iii) it holds for all $y \in \mathbf{P}_{\mathcal{M}} \backslash \mathcal{M}$ that

$$(\nabla \mathscr{d}_{\mathcal{M}})(y) = \tfrac{y - \mathscr{p}_{\mathcal{M}}(y)}{\|y - \mathscr{p}_{\mathcal{M}}(y)\|} \tag{4.4}$$

(cf. Definitions 30 to 32).

Definition 37 *Let $\mathfrak{d}, k \in \mathbb{N}$, let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ be a k-dimensional C^2-submanifold of $\mathbb{R}^{\mathfrak{d}}$, and let $x \in \mathcal{M}$, $r, s \in (0, \infty)$. Then we denote by $V_{\mathcal{M},x}^{r,s} \subseteq \mathbb{R}^{\mathfrak{d}}$ the set given by*

$$V_{\mathcal{M},x}^{r,s} = \Big\{ y \in \mathbb{R}^{\mathfrak{d}} \colon \exists\, \mathfrak{m} \in \mathcal{M} \colon \exists\, v \in (\mathcal{T}_{\mathcal{M}}^{\mathfrak{m}})^{\perp} \colon \big[(\|\mathfrak{m} - x\| \leq r),\, (\|v\| < s),\, (y = \mathfrak{m} + v) \big] \Big\} \tag{4.5}$$

(cf. Definition 24).

Lemma 38 *Let $\mathfrak{d}, k \in \mathbb{N}$, let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ be a k-dimensional C^2-submanifold of $\mathbb{R}^{\mathfrak{d}}$, and let $x \in \mathcal{M}$, $r, s \in (0, \infty)$. Then*

(i) it holds that

$$V_{\mathcal{M},x}^{r,s} = \Big\{ y \in \mathbb{R}^{\mathfrak{d}} \colon \exists\, \mathfrak{m} \in \mathcal{M} \colon \big[(\|\mathfrak{m} - x\| \leq r),\, (\|y - \mathfrak{m}\| < s),\, (y - \mathfrak{m} \in (\mathcal{T}_{\mathcal{M}}^{\mathfrak{m}})^{\perp}) \big] \Big\}, \tag{4.6}$$

(ii) it holds that

$$V_{\mathcal{M},x}^{r,s} \supseteq \big\{ y \in \mathbf{P}_{\mathcal{M}} \colon \big[(\|x - \mathscr{p}_{\mathcal{M}}(y)\| \leq r),\, (\|y - \mathscr{p}_{\mathcal{M}}(y)\| < s) \big] \big\}, \tag{4.7}$$

and

(iii) it holds that $x \in (V_{\mathcal{M},x}^{r,s})^{\circ}$ (cf. Definitions 24, 31, 32 and 37).

Proof [Proof of Lemma 38] Note that (4.5) establishes item (i). Next observe that (4.5) and Lemma 35 establish item (ii). Furthermore, note that item (ii) implies that

$$V_{\mathcal{M},x}^{r,s} \supseteq \big\{ y \in \mathbf{P}_{\mathcal{M}} \colon \big[(\|x - \mathscr{p}_{\mathcal{M}}(y)\| < r),\, (\|y - \mathscr{p}_{\mathcal{M}}(y)\| < s) \big] \big\}. \tag{4.8}$$

Furthermore, observe that the fact that $\mathbf{P}_{\mathcal{M}} \ni y \mapsto \mathscr{p}(y) \in \mathbb{R}^{\mathfrak{d}}$ is continuous shows that $\{y \in \mathbf{P}_{\mathcal{M}} \colon [(\|x - \mathscr{p}_{\mathcal{M}}(y)\| < r),\, (\|y - \mathscr{p}_{\mathcal{M}}(y)\| < s)]\} \subseteq \mathbb{R}^{\mathfrak{d}}$ is open. Combining this with (4.8) and the fact that $x \in \{y \in \mathbf{P}_{\mathcal{M}} \colon [(\|x - \mathscr{p}_{\mathcal{M}}(y)\| < r),\, (\|y - \mathscr{p}_{\mathcal{M}}(y)\| < s)]\}$ establishes item (iii). The proof of Lemma 38 is thus complete. ■

Proposition 39 *Let $\mathfrak{d}, k \in \mathbb{N}$, let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ be a k-dimensional C^2-submanifold of $\mathbb{R}^{\mathfrak{d}}$, let $U \subseteq \mathbf{P}_{\mathcal{M}}$ be open, and let $x \in \mathcal{M} \cap U$ (cf. Definition 32). Then there exist $R, S \in (0, \infty)$ such that*

(i) it holds for all $r \in (0, R]$, $s \in (0, S]$ that $\overline{V^{r,s}_{\mathcal{M},x}} \subseteq U$,

(ii) it holds for all $r \in (0, R]$, $s \in (0, S]$ that

$$V^{r,s}_{\mathcal{M},x} = \big\{ y \in \mathbb{R}^{\mathfrak{d}} \colon \mathscr{d}_{\mathcal{M}}(y) = \mathscr{d}_{\{\mathfrak{m} \in \mathcal{M} \colon \|x - \mathfrak{m}\| \leq r\}}(y) < s \big\}, \tag{4.9}$$

(iii) it holds for all $r \in (0, R]$, $s \in (0, S]$, $\mathfrak{m} \in \mathcal{M}$, $v \in (\mathcal{T}^{\mathfrak{m}}_{\mathcal{M}})^{\perp}$ with $\|\mathfrak{m} - x\| \leq r$ and $\|v\| < s$ that $\mathfrak{m} + v \in V^{r,s}_{\mathcal{M},x}$ and $\mathscr{p}_{\mathcal{M}}(\mathfrak{m} + v) = \mathfrak{m}$, and

(iv) it holds for all $r \in (0, R]$, $s \in (0, S]$ that

$$V^{r,s}_{\mathcal{M},x} = \big\{ y \in \mathbf{P}_{\mathcal{M}} \colon \big[(\|x - \mathscr{p}_{\mathcal{M}}(y)\| \leq r), (\|y - \mathscr{p}_{\mathcal{M}}(y)\| < s) \big] \big\} \tag{4.10}$$

(cf. Definitions 24, 30, 31 and 37).

Proof [Proof of Proposition 39] Note that (Fehrman et al., 2020, Proposition 13) establishes items (i) to (iii). In addition, observe that items (ii) and (iii) and (4.5) establish item (iv). The proof of Proposition 39 is thus complete. ■

Setting 40 *Let $\mathfrak{d} \in \mathbb{N}$, $k \in \mathbb{N} \cap (0, \mathfrak{d})$, let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $f \in C^2(U, \mathbb{R})$ have locally Lipschitz continuous derivatives, let $\mathcal{M} \subseteq U$ satisfy $\mathcal{M} = \{x \in U \colon f(x) = \inf_{y \in U} f(y)\}$, and assume that $\mathcal{M}$ is a k-dimensional C^2-submanifold of $\mathbb{R}^{\mathfrak{d}}$.*

Proposition 41 *Assume Setting 40 and let $x \in \mathcal{M}$ satisfy $\operatorname{rank}((\operatorname{Hess} f)(x)) = \mathfrak{d} - k$. Then*

(i) it holds for all $v \in ((\mathcal{T}^{x}_{\mathcal{M}})^{\perp}) \backslash \{0\}$ that $\langle ((\operatorname{Hess} f)(x))v, v \rangle \geq [\sigma((\operatorname{Hess} f)(x))] \|v\|^2 > 0$ and

(ii) it holds for all $v \in ((\mathcal{T}^{x}_{\mathcal{M}})^{\perp}) \backslash \{0\}$, $r \in [0, (\Lambda((\operatorname{Hess} f)(x)))^{-1}]$ that $\|v - r((\operatorname{Hess} f)(x))v\| \leq [1 - r\sigma((\operatorname{Hess} f)(x))] \|v\|$.

(cf. Definitions 24 and 27).

Proof [Proof of Proposition 41] Throughout this proof let $\{v_1, v_2, \ldots, v_{\mathfrak{d}-k}\} \subseteq ((\mathcal{T}^{x}_{\mathcal{M}})^{\perp}) \backslash \{0\}$ be an orthogonal basis of $(\mathcal{T}^{x}_{\mathcal{M}})^{\perp}$ with respect to which $(\operatorname{Hess} f)(x)$ is diagonal and let $\lambda_1, \lambda_2, \ldots, \lambda_{\mathfrak{d}-k} \in \mathbb{R}$ satisfy for all $i \in \{1, 2, \ldots, \mathfrak{d}-k\}$ that $((\operatorname{Hess} f)(x))v_i = \lambda_i v_i$. Note that the fact that x is a local minimum of f shows for all $i \in \{1, 2, \ldots, \mathfrak{d}-k\}$ that $\lambda_i \geq 0$. This and the assumption that $\operatorname{rank}((\operatorname{Hess} f)(x)) = \mathfrak{d} - k$ imply for all $i \in \{1, 2, \ldots, \mathfrak{d}-k\}$ that $\lambda_i > 0$. Hence, we obtain for all $i \in \{1, 2, \ldots, \mathfrak{d}-k\}$ that $\lambda_i \in [\sigma((\operatorname{Hess} f)(x)), \Lambda((\operatorname{Hess} f)(x))]$. Next let $\mathbf{v} \in ((\mathcal{T}^{x}_{\mathcal{M}})^{\perp}) \backslash \{0\}$ and let $u_1, u_2, \ldots, u_{\mathfrak{d}-k} \in \mathbb{R}$ satisfy $\mathbf{v} = \sum_{i=1}^{\mathfrak{d}-k} u_i v_i$. Observe that

$$\begin{aligned} \big\langle \big((\operatorname{Hess} f)(x)\big)\mathbf{v}, \mathbf{v} \big\rangle &= \textstyle\sum_{i=1}^{\mathfrak{d}-k} \big(\lambda_i |u_i|^2 \|v_i\|^2\big) \geq \big[\sigma((\operatorname{Hess} f)(x))\big] \big[\textstyle\sum_{i=1}^{\mathfrak{d}-k} |u_i|^2 \|v_i\|^2\big] \\ &= \big[\sigma((\operatorname{Hess} f)(x))\big] \|\mathbf{v}\|^2 > 0. \end{aligned} \tag{4.11}$$

This establishes item (i). Furthermore, note that the fact that for all $i \in \{1, 2, \dots, \mathfrak{d} - k\}$ it holds that $\lambda_i \in [\sigma((\operatorname{Hess} f)(x)), \Lambda((\operatorname{Hess} f)(x))]$ ensures that for all $r \in [0, (\Lambda(\operatorname{Hess} f)(x))^{-1}]$ we have that

$$\begin{aligned} \big\|\mathbf{v} - r\big((\operatorname{Hess} f)(x)\big)\mathbf{v}\big\|^2 &= \textstyle\sum_{i=1}^{\mathfrak{d}-k} \big(|u_i|^2 \|v_i\|^2 (1 - r\lambda_i)^2\big) \\ &\le \textstyle\sum_{i=1}^{\mathfrak{d}-k} \big(|u_i|^2 \|v_i\|^2 \big(1 - r\big[\sigma((\operatorname{Hess} f)(x))\big]\big)^2\big) \\ &= \big(1 - r\big[\sigma((\operatorname{Hess} f)(x))\big]\big)^2 \|\mathbf{v}\|^2. \end{aligned} \tag{4.12}$$

This establishes item (ii). The proof of Proposition 41 is thus complete. ■

Lemma 42 *Assume Setting 40 and let $x \in \mathcal{M}$. Then there exist $c, r, s \in (0, \infty)$ such that for all $y \in V^{r,s}_{\mathcal{M},x}$ it holds that $\overline{V^{r,s}_{\mathcal{M},x}} \subseteq (\mathbf{P}_{\mathcal{M}} \cap U)$ and*

$$\big\|(\nabla f)(y) - \big((\operatorname{Hess} f)(\mathscr{P}_{\mathcal{M}}(y))\big)(y - \mathscr{P}_{\mathcal{M}}(y))\big\| \le c(\mathscr{d}_{\mathcal{M}}(y))^2 \tag{4.13}$$

(cf. Definitions 30 to 32 and 37).

Proof [Proof of Lemma 42] Observe that Proposition 39 ensures that there exist $r, s \in (0, \infty)$ which satisfy $\overline{V^{r,s}_{\mathcal{M},x}} \subseteq U$, which satisfy

$$V^{r,s}_{\mathcal{M},x} = \big\{y \in \mathbf{P}_{\mathcal{M}} \colon \big[(\|x - \mathscr{P}_{\mathcal{M}}(y)\| \le r), (\|y - \mathscr{P}_{\mathcal{M}}(y)\| < s)\big]\big\}, \tag{4.14}$$

and which satisfy for all $\mathfrak{m} \in \mathcal{M}$, $v \in (\mathcal{T}^{\mathfrak{m}}_{\mathcal{M}})^{\perp}$ with $\|\mathfrak{m} - x\| \le r$ and $\|v\| < s$ that $\mathfrak{m} + v \in V^{r,s}_{\mathcal{M},x}$ and

$$\mathscr{P}_{\mathcal{M}}(\mathfrak{m} + v) = \mathfrak{m} \tag{4.15}$$

(cf. Definition 24). Note that (4.14), (4.15), and Lemma 35 imply for all $y \in V^{r,s}_{\mathcal{M},x}$, $t \in [0, 1]$ that $\mathscr{P}_{\mathcal{M}}(y) + t(y - \mathscr{P}_{\mathcal{M}}(y)) \in V^{r,s}_{\mathcal{M},x}$. In addition, observe that the fact that $\overline{V^{r,s}_{\mathcal{M},x}}$ is compact and the assumption that $U \ni y \mapsto (\operatorname{Hess} f)(y) \in \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}}$ is locally Lipschitz continuous prove that there exists $c \in (0, \infty)$ which satisfies for all $y, z \in \overline{V^{r,s}_{\mathcal{M},x}}$, $v \in \mathbb{R}^{\mathfrak{d}}$ that $\|((\operatorname{Hess} f)(y) - (\operatorname{Hess} f)(z))v\| \le c\|y - z\|\|v\|$. Furthermore, note that the fact that for all $y \in V^{r,s}_{\mathcal{M},x}$ it holds that $(\nabla f)(\mathscr{P}_{\mathcal{M}}(y)) = 0$ and the assumption that f is twice continuously differentiable demonstrate that for all $y \in V^{r,s}_{\mathcal{M},x}$ it holds that

$$\begin{aligned} (\nabla f)(y) &= \int_0^1 \big((\operatorname{Hess} f)(\mathscr{P}_{\mathcal{M}}(y) + t(y - \mathscr{P}_{\mathcal{M}}(y)))\big)(y - \mathscr{P}_{\mathcal{M}}(y))\,\mathrm{d}t \\ &= \big((\operatorname{Hess} f)(\mathscr{P}_{\mathcal{M}}(y))\big)(y - \mathscr{P}_{\mathcal{M}}(y)) \\ &\quad + \int_0^1 \Big((\operatorname{Hess} f)\big(\mathscr{P}_{\mathcal{M}}(y) + t(y - \mathscr{P}_{\mathcal{M}}(y))\big) - (\operatorname{Hess} f)(\mathscr{P}_{\mathcal{M}}(y))\Big)(y - \mathscr{P}_{\mathcal{M}}(y))\,\mathrm{d}t. \end{aligned} \tag{4.16}$$

Combining this with the fact that for all $y \in V^{r,s}_{\mathcal{M},x}$, $t \in [0, 1]$ it holds that

$$\big\|\big((\operatorname{Hess} f)(\mathscr{P}_{\mathcal{M}}(y) + t(y - \mathscr{P}_{\mathcal{M}}(y)) - (\operatorname{Hess} f)(\mathscr{P}_{\mathcal{M}}(y))\big)(y - \mathscr{P}_{\mathcal{M}}(y))\big\| \le ct\|y - \mathscr{P}_{\mathcal{M}}(y)\|^2 \tag{4.17}$$

implies that for all $y \in V^{r,s}_{\mathcal{M},x}$ we have that

$$\|(\nabla f)(y) - ((\operatorname{Hess} f)(\mathscr{p}_{\mathcal{M}}(y)))(y - \mathscr{p}_{\mathcal{M}}(y))\| \leq c\|y - \mathscr{p}_{\mathcal{M}}(y)\|^2 \Big[\textstyle\int_0^1 t \, \mathrm{d}t\Big] = \tfrac{c}{2}(\mathscr{d}_{\mathcal{M}}(y))^2. \tag{4.18}$$

The proof of Lemma 42 is thus complete. ■

4.2 Abstract convergence result for GF to a submanifold of global minima

Proposition 43 *Assume Setting 40, assume for all $x \in \mathcal{M}$ that $\operatorname{rank}((\operatorname{Hess} f)(x)) = \mathfrak{d} - k$, let $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ be locally bounded and measurable, assume for all $x \in U$ that $\mathcal{G}(x) = (\nabla f)(x)$, let $\Theta^\theta \in C([0,\infty), \mathbb{R}^{\mathfrak{d}})$, $\theta \in \mathbb{R}^{\mathfrak{d}}$, satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0,\infty)$ that $\Theta^\theta_t = \theta - \int_0^t \mathcal{G}(\Theta^\theta_s) \, \mathrm{d}s$, and let $x \in \mathcal{M}$. Then there exist $r, s \in (0,\infty)$ such that*

(i) it holds for all $\theta \in V^{r/2,s}_{\mathcal{M},x}$, $t \in [0,\infty)$ that $\Theta^\theta_t \in V^{r,s}_{\mathcal{M},x}$,

(ii) it holds that $\inf_{y \in \mathcal{M} \cap V^{r,s}_{\mathcal{M},x}} [\sigma((\operatorname{Hess} f)(y))] > 0$, and

(iii) it holds for all $\theta \in V^{r/2,s}_{\mathcal{M},x}$, $t \in [0,\infty)$ that

$$\mathscr{d}_{\mathcal{M}}(\Theta^\theta_t) \leq \exp\!\big(-\tfrac{t}{2}\big[\inf\nolimits_{y \in \mathcal{M} \cap V^{r,s}_{\mathcal{M},x}} \big[\sigma((\operatorname{Hess} f)(y))\big]\big]\big) \mathscr{d}_{\mathcal{M}}(\theta) \tag{4.19}$$

(cf. Definitions 27, 30 and 37).

Proof [Proof of Proposition 43] Observe that Proposition 39 and Lemma 42 prove that there exist $r, \varepsilon, \mathfrak{c} \in (0,\infty)$ which satisfy $\overline{V^{r,\varepsilon}_{\mathcal{M},x}} \subseteq U$, which satisfy

$$V^{r,s}_{\mathcal{M},x} = \big\{y \in \mathbf{P}_{\mathcal{M}} \colon \big[(\|x - \mathscr{p}_{\mathcal{M}}(y)\| \leq r), (\|y - \mathscr{p}_{\mathcal{M}}(y)\| < s)\big]\big\}, \tag{4.20}$$

and which satisfy for all $y \in \overline{V^{r,\varepsilon}_{\mathcal{M},x}}$ that

$$\|(\nabla f)(y) - (\operatorname{Hess} f)(\mathscr{p}_{\mathcal{M}}(y))(y - \mathscr{p}_{\mathcal{M}}(y))\| \leq \mathfrak{c}(\mathscr{d}_{\mathcal{M}}(y))^2 \tag{4.21}$$

(cf. Definition 32). In the following let $\kappa \in \mathbb{R}$ satisfy $\kappa = \frac{1}{2} \inf_{y \in \mathcal{M} \cap V^{r,\varepsilon}_{\mathcal{M},x}} \big[\sigma((\operatorname{Hess} f)(y))\big]$. Note that the fact that $\operatorname{Hess} f$ is locally Lipschitz continuous and the fact that the eigenvalues are continuous functions of a matrix (cf., e.g., (Kato, 1995, Section 2.5.1)) prove that $\kappa > 0$. Next observe that the fact that $\overline{V^{r,\varepsilon}_{\mathcal{M},x}}$ is compact, the fact that for all $y \in \mathbf{P}_{\mathcal{M}}$ it holds that $(\nabla f)(\mathscr{p}_{\mathcal{M}}(y)) = 0$, the fact that $\mathbf{P}_{\mathcal{M}} \ni y \mapsto \mathscr{p}_{\mathcal{M}}(y) \in \mathbb{R}^{\mathfrak{d}}$ is continuously differentiable, and the assumption that $f \in C^2(U, \mathbb{R})$ prove that there exists $c \in (0,\infty)$ which satisfies for all $y \in \overline{V^{r,\varepsilon}_{\mathcal{M},x}}$ that

$$\|(\mathscr{p}_{\mathcal{M}})'(y)[(\nabla f)(y)]\| = \|(\mathscr{p}_{\mathcal{M}})'(y)[(\nabla f)(y) - (\nabla f)(\mathscr{p}_{\mathcal{M}}(y))]\| \leq c\|y - \mathscr{p}_{\mathcal{M}}(y)\| = c\mathscr{d}_{\mathcal{M}}(y) \tag{4.22}$$

(cf. Definition 31). In the following let $s \in (0,\infty)$ satisfy

$$s = \min\Big\{\frac{\kappa}{\mathfrak{c}}, \frac{\kappa r}{2c}, \varepsilon\Big\}, \tag{4.23}$$

let $\theta \in V^{r/2,s}_{\mathcal{M},x}$, and let $\tau \in (0,\infty]$ satisfy $\tau = \inf(\{t \in [0,\infty) \colon \Theta^\theta_t \notin V^{r,s}_{\mathcal{M},x}\} \cup \{\infty\})$. Note that the assumption that for all $y \in U$ it holds that $\mathcal{G}(y) = (\nabla f)(y)$ and the fact that $U \ni y \mapsto (\nabla f)(y) \in \mathbb{R}^{\mathfrak{d}}$ is continuous assure that $[0,\tau) \ni t \mapsto \Theta^\theta_t \in \mathbb{R}^{\mathfrak{d}}$ is continuously differentiable and that for all $t \in [0,\tau)$ it holds that $\frac{\mathrm{d}}{\mathrm{d}t}\Theta^\theta_t = -(\nabla f)(\Theta^\theta_t)$. This, Lemma 36, and the chain rule show for all $t \in [0,\tau)$ that

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathscr{d}_{\mathcal{M}}(\Theta^\theta_t) = -\Big\langle (\nabla f)(\Theta^\theta_t), (\nabla \mathscr{d}_{\mathcal{M}})(\Theta^\theta_t) \Big\rangle = -\Big\langle (\nabla f)(\Theta^\theta_t), \frac{\Theta^\theta_t - \mathscr{p}_{\mathcal{M}}(\Theta^\theta_t)}{\|\Theta^\theta_t - \mathscr{p}_{\mathcal{M}}(\Theta^\theta_t)\|} \Big\rangle \tag{4.24}$$

. Next observe that (4.21), (4.23), (4.24), and Proposition 41 demonstrate for all $t \in [0,\tau)$ that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \mathscr{d}_{\mathcal{M}}(\Theta^\theta_t) &= -\Big\langle (\operatorname{Hess} f)(\mathscr{p}_{\mathcal{M}}(\Theta^\theta_t))(\Theta^\theta_t - \mathscr{p}_{\mathcal{M}}(\Theta^\theta_t)), \frac{\Theta^\theta_t - \mathscr{p}_{\mathcal{M}}(\Theta^\theta_t)}{\|\Theta^\theta_t - \mathscr{p}_{\mathcal{M}}(\Theta^\theta_t)\|} \Big\rangle \\
&\quad - \Big\langle (\nabla f)(\Theta^\theta_t) - (\operatorname{Hess} f)(\mathscr{p}_{\mathcal{M}}(\Theta^\theta_t))(\Theta^\theta_t - \mathscr{p}_{\mathcal{M}}(\Theta^\theta_t)), \frac{\Theta^\theta_t - \mathscr{p}_{\mathcal{M}}(\Theta^\theta_t)}{\|\Theta^\theta_t - \mathscr{p}_{\mathcal{M}}(\Theta^\theta_t)\|} \Big\rangle \\
&\le -2\kappa \|\Theta^\theta_t - \mathscr{p}_{\mathcal{M}}(\Theta^\theta_t)\| + \mathfrak{c}(\mathscr{d}_{\mathcal{M}}(\Theta^\theta_t))^2 \\
&= -2\kappa \mathscr{d}_{\mathcal{M}}(\Theta^\theta_t) + \mathfrak{c}(\mathscr{d}_{\mathcal{M}}(\Theta^\theta_t))^2 \le -\kappa \mathscr{d}_{\mathcal{M}}(\Theta^\theta_t).
\end{aligned} \tag{4.25}$$

Hence, we obtain for all $t \in [0,\tau)$ that

$$\mathscr{d}_{\mathcal{M}}(\Theta^\theta_t) \le e^{-\kappa t} \mathscr{d}_{\mathcal{M}}(\Theta^\theta_0) = e^{-\kappa t} \mathscr{d}_{\mathcal{M}}(\theta). \tag{4.26}$$

It remains to prove that $\tau = \infty$. To this end, note that the chain rule and Lemma 33 imply for all $t \in [0,\tau)$ that

$$\tfrac{\mathrm{d}}{\mathrm{d}t} \mathscr{p}_{\mathcal{M}}(\Theta^\theta_t) = -(D\mathscr{p}_{\mathcal{M}})(\Theta^\theta_t)(\nabla f(\Theta^\theta_t)). \tag{4.27}$$

Combining this, (4.22), and (4.26) ensures for all $t \in [0,\tau)$ that

$$\Big\| \tfrac{\mathrm{d}}{\mathrm{d}t} \mathscr{p}_{\mathcal{M}}(\Theta^\theta_t) \Big\| \le c \mathscr{d}_{\mathcal{M}}(\Theta^\theta_t) \le c e^{-\kappa t} \mathscr{d}_{\mathcal{M}}(\theta) \le c s e^{-\kappa t}. \tag{4.28}$$

This and (4.23) show for all $t \in [0,\tau)$ that

$$\|\mathscr{p}_{\mathcal{M}}(\Theta^\theta_t) - \mathscr{p}_{\mathcal{M}}(\theta)\| \le cs \int_0^t e^{-\kappa u} \,\mathrm{d}u \le \frac{\kappa r}{2} \int_0^\infty e^{-\kappa u} \,\mathrm{d}u = \frac{r}{2}. \tag{4.29}$$

Furthermore, observe that the assumption that $\theta \in V^{r/2,s}_{\mathcal{M},x}$ assures that there exists $\delta \in (0,\infty)$ which satisfies that $\theta \in V^{r/2-\delta,s}_{\mathcal{M},x}$. Combining this with (4.29) establishes for all $t \in [0,\tau)$ that $\Theta^\theta_t \in V^{r-\delta,s}_{\mathcal{M},x}$. Consequently, we must have that $\tau = \infty$. The proof of Proposition 43 is thus complete. ∎

4.3 Convergence rates for GF in the training of ANNs

Proposition 44 *Assume Setting 3, let $N \in \mathbb{N} \cap [1, H]$, $x_0, x_1, \dots, x_N, \alpha_1, \alpha_2, \dots, \alpha_N \in \mathbb{R}$ satisfy $a = x_0 < x_1 < \cdots < x_N = b$, assume for all $i \in \{1, 2, \dots, N\}$, $x \in [x_{i-1}, x_i]$ that $f(x) = f(x_{i-1}) + \alpha_i(x - x_{i-1})$, assume that $\mathcal{G}$ is locally bounded and measurable, and let $\Theta^\theta \in C([0,\infty), \mathbb{R}^{\mathfrak{d}})$, $\theta \in \mathbb{R}^{\mathfrak{d}}$, satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0, \infty)$ that*

$$\Theta_t^\theta = \theta - \int_0^t \mathcal{G}(\Theta_s^\theta)\, \mathrm{d}s. \tag{4.30}$$

Then there exist $\mathfrak{c}, \mathfrak{C} \in (0, \infty)$ and a non-empty open $U \subseteq \mathbb{R}^{\mathfrak{d}}$ such that for all $\theta \in U$, $t \in [0, \infty)$ it holds that $\mathcal{L}(\Theta_t^\theta) \le \mathfrak{C} e^{-\mathfrak{c}t}$.

Proof [Proof of Proposition 44] Throughout this proof let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy $\mathcal{M} = \{\theta \in \mathbb{R}^{\mathfrak{d}} \colon \mathcal{L}(\theta) = 0\}$. Note that Corollary 29 proves that there exist $k \in \mathbb{N} \cap [1, \mathfrak{d})$ and an open $U \subseteq \mathbb{R}^{\mathfrak{d}}$ which satisfy $U \subseteq \mathfrak{V}$, which satisfy that $\mathcal{L}|_U$ is twice continuously differentiable, which satisfy that $(\operatorname{Hess} \mathcal{L})|_U$ is locally Lipschitz continuous, which satisfy that $\mathcal{M} \cap U$ is a non-empty k-dimensional C^2-submanifold of $\mathbb{R}^{\mathfrak{d}}$, and which satisfy for all $\theta \in \mathcal{M} \cap U$ that $\operatorname{rank}((\operatorname{Hess} \mathcal{L})(\theta)) = \mathfrak{d} - k$. Combining this, Proposition 11, Lemma 38, and Proposition 39 with Proposition 43 ensures that there exist $\mathfrak{m} \in \mathcal{M} \cap U$, $\mathfrak{c} \in (0, \infty)$, $V, \mathcal{V} \in \{A \subseteq U \colon A \text{ is compact}\}$ which satisfy that

(i) it holds that $\mathfrak{m} \in V^\circ \subseteq V \subseteq \mathcal{V}$,

(ii) it holds for all $\theta \in \mathcal{V}$ that $d_{\mathcal{M} \cap U}(\theta) = d_{\mathcal{M} \cap U \cap \mathcal{V}}$,

(iii) it holds for all $\theta \in V$, $t \in [0, \infty)$ that $\Theta_t^\theta \in \mathcal{V}$, and

(iv) it holds for all $t \in [0, \infty)$ that $d_{\mathcal{M} \cap U}(\Theta_t^\theta) \le e^{-\mathfrak{c}t} d_{\mathcal{M} \cap U}(\theta)$

(cf. Definition 30). Furthermore, observe that the fact that $\mathcal{L}|_U$ is twice continuously differentiable proves that there exists $\mathfrak{C} \in (0, \infty)$ which satisfies for all $\theta, \vartheta \in \mathcal{V}$ that $|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)| \le \mathfrak{C}\|\theta - \vartheta\|$. This assures that for all $\theta \in V^\circ$, $t \in [0, \infty)$ we have that

$$\begin{aligned} \mathcal{L}(\Theta_t^\theta) &= \inf\nolimits_{\vartheta \in \mathcal{M} \cap U \cap \mathcal{V}} |\mathcal{L}(\Theta_t^\theta) - \mathcal{L}(\vartheta)| \le \mathfrak{C}\big[\inf\nolimits_{\vartheta \in \mathcal{M} \cap U \cap \mathcal{V}} \|\Theta_t^\theta - \vartheta\|\big] \\ &= \mathfrak{C}\big[d_{\mathcal{M} \cap U}(\Theta_t^\theta)\big] \le \mathfrak{C} e^{-\mathfrak{c}t} d_{\mathcal{M} \cap U}(\theta). \end{aligned} \tag{4.31}$$

The proof of Proposition 44 is thus complete. ∎

4.4 Convergence rates for GF with random initializations in the training of ANNs

Corollary 45 *Assume Setting 3, let $N \in \mathbb{N} \cap [1, H]$, $x_0, x_1, \dots, x_N, \alpha_1, \alpha_2, \dots, \alpha_N \in \mathbb{R}$ satisfy $a = x_0 < x_1 < \cdots < x_N = b$, assume for all $i \in \{1, 2, \dots, N\}$, $x \in [x_{i-1}, x_i]$ that $f(x) = f(x_{i-1}) + \alpha_i(x - x_{i-1})$, assume that $\mathcal{G}$ is locally bounded and measurable, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\Theta \colon [0, \infty) \times \Omega \to \mathbb{R}^{\mathfrak{d}}$ be a stochastic process with* *continuous sample paths, assume that Θ_0 is standard normally distributed, and assume for all $t \in [0, \infty)$, $\omega \in \Omega$ that*

$$\Theta_t(\omega) = \Theta_0(\omega) - \int_0^t \mathcal{G}(\Theta_s(\omega))\,\mathrm{d}s. \tag{4.32}$$

Then there exist $\mathfrak{c}, \mathfrak{C} \in (0, \infty)$ such that $\mathbb{P}(\forall\, t \in [0, \infty) \colon \mathcal{L}(\Theta_t) \le \mathfrak{C}e^{-\mathfrak{c}t}) > 0$.

Proof [Proof of Corollary 45] Note that Proposition 44 ensures that there exist $\mathfrak{c}, \mathfrak{C} \in (0, \infty)$ and a non-empty open $U \subseteq \mathbb{R}^{\mathfrak{d}}$ which satisfy for all $t \in [0, \infty)$, $\omega \in \Omega$ with $\Theta_0(\omega) \in U$ that $\mathcal{L}(\Theta_t(\omega)) \le \mathfrak{C}e^{-\mathfrak{c}t}$. Observe that the fact that U is a non-empty open set and the assumption that Θ_0 is standard normally distributed imply that $\mathbb{P}(\Theta_0 \in U) > 0$. This completes the proof of Corollary 45. ■

5. Local convergence to the set of global minima for gradient descent (GD)

In this section we employ Corollary 29 from Section 3 to establish in Theorem 48 in Subsection 5.2, Corollary 49 in Subsection 5.3, and Corollary 50 in Subsection 5.3 under the assumption that the target function is piecewise affine linear that the risk of certain GD processes converges to zero. Our proofs of Corollaries 49 and 50 are based on an application of Theorem 48 and our proof of Theorem 48 uses the abstract local convergence result for GD processes in Proposition 47 in Subsection 5.1 below. Proposition 47 and its proof are strongly inspired by (Fehrman et al., 2020, Proposition 17). Our proof of Proposition 47 employs the elementary uniform estimate for certain exponential sums in Lemma 46 in Subsection 5.1. For completeness we include in this section also a detailed proof for Lemma 46.

5.1 Abstract convergence result for GD to a submanifold of global minima

Lemma 46 *Let $\rho \in [0, 1)$, $c, \mathfrak{g} \in (0, \infty)$. Then there exists $\mathfrak{C} \in \mathbb{R}$ such that for all $\gamma \in (0, \mathfrak{g}]$ it holds that*

$$\sum_{k=1}^{\infty} \gamma k^{-\rho} \exp\bigl(-c\gamma(k-1)^{1-\rho}\bigr) \le \mathfrak{C}. \tag{5.1}$$

Proof [Proof of Lemma 46] First note that for all $\gamma \in (0, \mathfrak{g}]$ it holds that

$$\begin{aligned}
\sum_{k=1}^{\infty} \gamma k^{-\rho} \exp\bigl(-c\gamma(k-1)^{1-\rho}\bigr) &\le \gamma + \sum_{k=2}^{\infty} \gamma (k-1)^{-\rho} \exp\bigl(-c\gamma(k-1)^{1-\rho}\bigr) \\
&\le \mathfrak{g} + \sum_{n=1}^{\infty} \gamma n^{-\rho} \exp\bigl(-c\gamma n^{1-\rho}\bigr) \\
&\le 2\mathfrak{g} + \sum_{n=2}^{\infty} \gamma n^{-\rho} \exp\bigl(-c\gamma n^{1-\rho}\bigr).
\end{aligned} \tag{5.2}$$

Next observe that the fact that for all $\gamma \in (0,\infty)$ it holds that $[1,\infty) \ni x \mapsto x^{-\rho}\exp(-c\gamma x^{1-\rho}) \in \mathbb{R}$ is continuous and non-increasing assures that for all $\gamma \in (0,\mathfrak{g}]$ we have that

$$\begin{aligned}\sum_{n=2}^{\infty} \gamma n^{-\rho} \exp\bigl(-c\gamma n^{1-\rho}\bigr) &\leq \sum_{n=2}^{\infty}\biggl[\int_{n-1}^{n} \gamma x^{-\rho}\exp\bigl(-c\gamma x^{1-\rho}\bigr)\,\mathrm{d}x\biggr] \\ &= \int_1^\infty \gamma x^{-\rho}\exp\bigl(-c\gamma x^{1-\rho}\bigr)\,\mathrm{d}x.\end{aligned} \tag{5.3}$$

Moreover, note that the integral transformation theorem proves for all $\gamma \in (0,\mathfrak{g}]$ that

$$\begin{aligned}&\int_1^\infty \gamma x^{-\rho}\exp\bigl(-c\gamma x^{1-\rho}\bigr)\,\mathrm{d}x = \int_{\gamma^{1/(1-\rho)}}^{\infty} \gamma^{1+\frac{\rho}{1-\rho}} x^{-\rho}\exp\bigl(-cx^{1-\rho}\bigr)\gamma^{-\frac{1}{1-\rho}}\,\mathrm{d}x \\ &\leq \int_0^\infty x^{-\rho}\exp\bigl(-cx^{1-\rho}\bigr)\,\mathrm{d}x \leq \int_0^1 x^{-\rho}\,\mathrm{d}x + \int_1^\infty \exp\bigl(-cx^{1-\rho}\bigr)\,\mathrm{d}x \\ &= \frac{1}{1-\rho} + \int_1^\infty \exp\bigl(-cx^{1-\rho}\bigr)\,\mathrm{d}x.\end{aligned} \tag{5.4}$$

Furthermore, observe that the assumption that $c \in (0,\infty)$ and the assumption that $\rho \in [0,1)$ ensure that $\int_1^\infty \exp\bigl(-cx^{1-\rho}\bigr)\,\mathrm{d}x < \infty$. Combining this, (5.2), (5.3), and (5.4) establishes for all $\gamma \in (0,\mathfrak{g}]$ that

$$\sum_{k=1}^{\infty} \gamma k^{-\rho}\exp\bigl(-c\gamma(k-1)^{1-\rho}\bigr) \leq 2\mathfrak{g} + \frac{1}{1-\rho} + \int_1^\infty \exp\bigl(-cx^{1-\rho}\bigr)\,\mathrm{d}x < \infty. \tag{5.5}$$

The proof of Lemma 46 is thus complete. ∎

Proposition 47 *Assume Setting 40, assume for all $x \in \mathcal{M}$ that* $\operatorname{rank}((\operatorname{Hess} f)(x)) = \mathfrak{d} - n$, *let $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $x \in U$ that $\mathcal{G}(x) = (\nabla f)(x)$, let $x \in \mathcal{M}$, $\rho \in [0,1)$, and let $\Theta^{\theta,\gamma}\colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$, $\gamma \in \mathbb{R}$, satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $\gamma \in \mathbb{R}$, $n \in \mathbb{N}$ that $\Theta_0^{\theta,\gamma} = \theta$ and*

$$\Theta_n^{\theta,\gamma} = \Theta_{n-1}^{\theta,\gamma} - \tfrac{\gamma}{n^\rho}\mathcal{G}(\Theta_{n-1}^{\theta,\gamma}). \tag{5.6}$$

Then there exist $r, s \in (0,\infty)$ such that

(i) it holds for all $\theta \in V_{\mathcal{M},x}^{r/2,s}$, $\gamma \in (0,\min\{[\sup_{y\in\mathcal{M}\cap V_{\mathcal{M},x}^{r,s}}\Lambda((\operatorname{Hess} f)(y))]^{-1},1\}]$, $n \in \mathbb{N}_0$ that $\Theta_n^{\theta,\gamma} \in V_{\mathcal{M},x}^{r,s}$,

(ii) it holds that $\inf_{y\in\mathcal{M}\cap V_{\mathcal{M},x}^{r,s}}\bigl[\sigma((\operatorname{Hess} f)(y))\bigr] > 0$, and

(iii) it holds for all $\theta \in V_{\mathcal{M},x}^{r/2,s}$, $\gamma \in (0,\min\{[\sup_{y\in\mathcal{M}\cap V_{\mathcal{M},x}^{r,s}}\Lambda((\operatorname{Hess} f)(y))]^{-1},1\}]$, $n \in \mathbb{N}_0$ that

$$d_{\mathcal{M}}(\Theta_n^{\theta,\gamma}) \leq \exp\Bigl(-\tfrac{\gamma}{2(1-\rho)}\Bigl[\inf_{y\in\mathcal{M}\cap V_{\mathcal{M},x}^{r,s}}\bigl[\sigma((\operatorname{Hess} f)(y))\bigr]\Bigr]n^{1-\rho}\Bigr)\, d_{\mathcal{M}}(\theta) \tag{5.7}$$

(cf. Definitions 27 and 37).

Proof [Proof of Proposition 47] Note that Proposition 39 and Lemma 42 prove that there exist $r, \varepsilon, \mathfrak{c} \in (0,\infty)$ which satisfy $\overline{V^{r,\varepsilon}_{\mathcal{M},x}} \subseteq U$, which satisfy

$$V^{r,s}_{\mathcal{M},x} = \big\{ y \in \mathbf{P}_{\mathcal{M}} \colon \big[(\|x - \mathscr{P}_{\mathcal{M}}(y)\| \leq r), \, (\|y - \mathscr{P}_{\mathcal{M}}(y)\| < s) \big] \big\}, \tag{5.8}$$

and which satisfy for all $y \in \overline{V^{r,\varepsilon}_{\mathcal{M},x}}$ that

$$\|(\nabla f)(y) - (\operatorname{Hess} f)(\mathscr{P}_{\mathcal{M}}(y))(y - \mathscr{P}_{\mathcal{M}}(y))\| \leq \mathfrak{c}(\mathscr{d}_{\mathcal{M}}(y))^2 \tag{5.9}$$

(cf. Definition 32). In the following let $\kappa \in \mathbb{R}$ satisfy $\kappa = \inf_{y \in \mathcal{M} \cap V^{r,\varepsilon}_{\mathcal{M},x}} \big[\sigma((\operatorname{Hess} f)(y)) \big]$. Observe that the fact that $U \ni y \mapsto (\operatorname{Hess} f)(y) \in \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}}$ is locally Lipschitz continuous and the fact that the eigenvalues are continuous functions of a matrix (cf., e.g., (Kato, 1995, Section 2.5.1)) prove that $\kappa > 0$. Next note that the fact that $\overline{V^{r,\varepsilon}_{\mathcal{M},x}}$ is compact and the fact that $U \ni y \mapsto (\nabla f)(y) \in \mathbb{R}^{\mathfrak{d}}$ is continuously differentiable demonstrate that there exists $c \in (0,\infty)$ which satisfies for all $y \in \overline{V^{r,\varepsilon}_{\mathcal{M},x}}$ that

$$\|(\nabla f)(y)\| = \|(\nabla f)(y) - (\nabla f)(\mathscr{P}_{\mathcal{M}}(y))\| \leq c\|y - \mathscr{P}_{\mathcal{M}}(y)\| = c\,\mathscr{d}_{\mathcal{M}}(y) \tag{5.10}$$

(cf. Definition 31). In the following let $\mathfrak{C} \in (0,\infty)$ satisfy for all $\gamma \in (0,1]$ that

$$\sum_{k=1}^{\infty} \gamma k^{-\rho} \exp\Big(-\tfrac{\kappa\gamma}{2(1-\rho)} (k-1)^{1-\rho} \Big) \leq \mathfrak{C} \tag{5.11}$$

(cf. Lemma 46), let $s \in (0,\infty)$ satisfy

$$s = \min\bigg\{ \frac{\kappa}{2\mathfrak{c}}, \frac{r}{2(2 + c\mathfrak{C})}, \varepsilon \bigg\}, \tag{5.12}$$

let $\theta \in V^{r/2,s}_{\mathcal{M},x}$ and $\gamma \in (0, \min\{[\sup_{y \in \mathcal{M} \cap V^{r,s}_{\mathcal{M},x}} \Lambda((\operatorname{Hess} f)(y))]^{-1}, 1\}]$ be arbitrary, and let $\tau \in \mathbb{N} \cup \{\infty\}$ satisfy $\tau = \inf\{n \in \mathbb{N}_0 \colon \Theta^{\theta,\gamma}_n \notin V^{r,s}_{\mathcal{M},x}\}$. Observe that the fact that for all $n \in \mathbb{N} \cap (0,\tau]$ it holds that $\Theta^{\theta,\gamma}_n \in V^{r,s}_{\mathcal{M},x}$ proves that for all $n \in \mathbb{N} \cap (0,\tau]$ we have that

$$\begin{aligned} \mathscr{d}_{\mathcal{M}}(\Theta^{\theta,\gamma}_n) &\leq \|\Theta^{\theta,\gamma}_n - \mathscr{P}_{\mathcal{M}}(\Theta^{\theta,\gamma}_{n-1})\| \\ &= \Big\| \Theta^{\theta,\gamma}_{n-1} - \mathscr{P}_{\mathcal{M}}(\Theta^{\theta,\gamma}_{n-1}) - \tfrac{\gamma}{n^\rho} (\nabla f)(\Theta^{\theta,\gamma}_{n-1}) \Big\| \\ &\leq \Big\| \Theta^{\theta,\gamma}_{n-1} - \mathscr{P}_{\mathcal{M}}(\Theta^{\theta,\gamma}_{n-1}) - \tfrac{\gamma}{n^\rho} (\operatorname{Hess} f)(\mathscr{P}_{\mathcal{M}}(\Theta^{\theta,\gamma}_{n-1}))(\Theta^{\theta,\gamma}_{n-1} - \mathscr{P}_{\mathcal{M}}(\Theta^{\theta,\gamma}_{n-1})) \Big\| \\ &\quad + \tfrac{\gamma}{n^\rho} \Big\| \big((\operatorname{Hess} f)(\mathscr{P}_{\mathcal{M}}(\Theta^{\theta,\gamma}_{n-1})) \big) (\Theta^{\theta,\gamma}_{n-1} - \mathscr{P}_{\mathcal{M}}(\Theta^{\theta,\gamma}_{n-1})) - (\nabla f)(\Theta^{\theta,\gamma}_{n-1}) \Big\|. \end{aligned} \tag{5.13}$$

Combining this, Proposition 41, and (5.9) demonstrates for all $n \in \mathbb{N} \cap (0,\tau]$ that

$$\mathscr{d}_{\mathcal{M}}(\Theta^{\theta,\gamma}_n) \leq \big(1 - \tfrac{\kappa\gamma}{n^\rho}\big) \mathscr{d}_{\mathcal{M}}(\Theta^{\theta,\gamma}_{n-1}) + \tfrac{\mathfrak{c}\gamma}{n^\rho} (\mathscr{d}_{\mathcal{M}}(\Theta^{\theta,\gamma}_{n-1}))^2. \tag{5.14}$$

This, the fact that for all $n \in \mathbb{N} \cap (0,\tau]$ it holds that $\mathscr{d}_{\mathcal{M}}(\Theta^{\theta,\gamma}_{n-1}) \leq s \leq \frac{\kappa}{2\mathfrak{c}}$, and (5.12) imply that for all $n \in \mathbb{N} \cap (0,\tau]$ it holds that

$$\mathscr{d}_{\mathcal{M}}(\Theta^{\theta,\gamma}_n) \leq \big(1 - \tfrac{\kappa\gamma}{2n^\rho}\big) \mathscr{d}_{\mathcal{M}}(\Theta^{\theta,\gamma}_{n-1}). \tag{5.15}$$

By induction, we therefore obtain for all $n \in \mathbb{N} \cap (0, \tau]$ that

$$d_{\mathcal{M}}(\Theta_n^{\theta,\gamma}) \le \left[\textstyle\prod_{k=1}^n \left(1 - \frac{\kappa\gamma}{2k^\rho}\right)\right] d_{\mathcal{M}}(\theta). \tag{5.16}$$

Next note that the assumption that $\gamma \le [\sup_{y \in \mathcal{M} \cap V_{\mathcal{M},x}^{r,s}} \Lambda((\operatorname{Hess} f)(y))]^{-1} \le \kappa^{-1}$ shows for all $k \in \mathbb{N}$ that $\frac{\kappa\gamma}{2k^\rho} \in (0,1)$. This and the fact that for all $u \in (0,1)$ it holds that $\ln(1-u) \le -u$ prove that for all $n \in \mathbb{N}$ we have that

$$\ln\left[\textstyle\prod_{k=1}^n \left(1 - \frac{\kappa\gamma}{2k^\rho}\right)\right] = \textstyle\sum_{k=1}^n \ln\left(1 - \frac{\kappa\gamma}{2k^\rho}\right) \le -\frac{\kappa\gamma}{2} \sum_{k=1}^n k^{-\rho} \le -\frac{\kappa\gamma}{2} \int_0^n u^{-\rho}\, \mathrm{d}u = \frac{\kappa\gamma}{2(1-\rho)} n^{1-\rho}. \tag{5.17}$$

Combining this with (5.16) demonstrates for all $n \in \mathbb{N} \cap (0, \tau]$ that

$$d_{\mathcal{M}}(\Theta_n^{\theta,\gamma}) \le \exp\!\left(-\tfrac{\kappa\gamma}{2(1-\rho)} n^{1-\rho}\right) d_{\mathcal{M}}(\theta). \tag{5.18}$$

It only remains to show that $\tau = \infty$. Observe that (5.10) assures for all $n \in \mathbb{N} \cap (0, \tau]$ that

$$\|\Theta_n^{\theta,\gamma} - \Theta_{n-1}^{\theta,\gamma}\| = \tfrac{\gamma}{n^\rho} \|(\nabla f)(\Theta_{n-1}^{\theta,\gamma})\| \le \tfrac{c\gamma}{n^\rho} d_{\mathcal{M}}(\Theta_{n-1}^{\theta,\gamma}) \tag{5.19}$$

This, (5.18), the fact that $\gamma \le 1$, (5.11), and the triangle inequality establish for all $n \in \mathbb{N} \cap (0, \tau]$ that

$$\begin{aligned} \|\Theta_n^{\theta,\gamma} - \theta\| &\le \sum_{k=1}^n c\gamma k^{-\rho} \exp\!\left(-\tfrac{\kappa\gamma}{2(1-\rho)} (k-1)^{1-\rho}\right) d_{\mathcal{M}}(\theta) \\ &\le cs \sum_{k=1}^\infty \gamma k^{-\rho} \exp\!\left(-\tfrac{\kappa\gamma}{2(1-\rho)} (k-1)^{1-\rho}\right) \le cs\mathfrak{C}. \end{aligned} \tag{5.20}$$

Combining this with (5.18), (5.12), and the triangle inequality proves for all $n \in \mathbb{N} \cap (0, \tau]$ that

$$\begin{aligned} \|p_{\mathcal{M}}(\Theta_n^{\theta,\gamma}) - p_{\mathcal{M}}(\theta)\| &\le d_{\mathcal{M}}(\Theta_n^{\theta,\gamma}) + \|\Theta_n^{\theta,\gamma} - \theta\| + d_{\mathcal{M}}(\theta) \\ &\le s(2 + c\mathfrak{C}) \le \tfrac{r}{2}. \end{aligned} \tag{5.21}$$

Furthermore, note that the assumption that $\theta \in V_{\mathcal{M},x}^{r/2,s}$ assures that there exists $\delta \in (0, \infty)$ which satisfies that $\theta \in V_{\mathcal{M},x}^{r/2-\delta,s}$. Hence, we obtain for all $n \in \mathbb{N} \cap (0, \tau]$ that $\Theta_n^{\theta,\gamma} \in V_{\mathcal{M},x}^{r-\delta,s}$. This implies that $\tau = \infty$. The proof of Proposition 47 is thus complete. ■

5.2 Convergence rates for GD in the training of ANNs

Theorem 48 *Assume Setting 3, let $N \in \mathbb{N} \cap [1, H]$, $\rho \in [0,1)$, $x_0, x_1, \ldots, x_N, \alpha_1, \alpha_2, \ldots, \alpha_N \in \mathbb{R}$ satisfy $a = x_0 < x_1 < \cdots < x_N = b$, assume for all $i \in \{1, 2, \ldots, N\}$, $x \in [x_{i-1}, x_i]$ that $f(x) = f(x_{i-1}) + \alpha_i(x - x_{i-1})$, let $\mathfrak{D} \in \mathbb{R}$ satisfy*

$$\mathfrak{D} = 1 + |f(a)| + (1 + 2\max\nolimits_{j \in \{1,2,\ldots,H\}} |\alpha_j|)(|a| + |b| + 1), \tag{5.22}$$

and let $\Theta^{\theta,\gamma} \colon \mathbb{N}_0 \to \mathbb{R}^{\mathfrak{d}}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$, $\gamma \in \mathbb{R}$, satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $\gamma \in \mathbb{R}$, $n \in \mathbb{N}$ that $\Theta_0^{\theta,\gamma} = \theta$ and

$$\Theta_n^{\theta,\gamma} = \Theta_{n-1}^{\theta,\gamma} - \tfrac{\gamma}{n^\rho} \mathcal{G}(\Theta_{n-1}^{\theta,\gamma}). \tag{5.23}$$

Then there exist $\mathfrak{c}, \mathfrak{C} \in (0,\infty)$ *and a non-empty open* $U \subseteq (-\mathfrak{D},\mathfrak{D})^{\mathfrak{d}}$ *such that for all* $\theta \in U$, $\gamma \in (0, ((3N+1)(24\mathfrak{D}^5 + 16N\mathfrak{D}^7)(\sup_{x\in[a,b]} \mathfrak{p}(x)))^{-1}]$, $n \in \mathbb{N}_0$ *it holds that* $\mathcal{L}(\Theta_n^{\theta,\gamma}) \leq \mathfrak{C}\exp(-\mathfrak{c}\gamma n^{1-\rho})$.

Proof [Proof of Theorem 48] Throughout this proof let $\mathcal{M} \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy $\mathcal{M} = \{\theta \in \mathbb{R}^{\mathfrak{d}} \colon \mathcal{L}(\theta) = 0\}$. Observe that Corollary 29 proves that there exist $k \in \mathbb{N} \cap [1,\mathfrak{d})$ and an open $U \subseteq (-\mathfrak{D},\mathfrak{D})^{\mathfrak{d}}$ which satisfy $U \subseteq \mathfrak{V}$, which satisfy that $\mathcal{L}|_U$ is twice continuously differentiable, which satisfy for all $\theta \in U$ that $\Lambda((\operatorname{Hess}\mathcal{L})(\theta)) \leq (3N+1)(24\mathfrak{D}^5 + 16N\mathfrak{D}^7)(\sup_{x\in[a,b]} \mathfrak{p}(x))$, which satisfy that $(\operatorname{Hess}\mathcal{L})|_U$ is locally Lipschitz continuous, which satisfy that $\mathcal{M}\cap U$ is a non-empty k-dimensional C^2-submanifold of $\mathbb{R}^{\mathfrak{d}}$, and which satisfy for all $\theta \in \mathcal{M}\cap U$ that $\operatorname{rank}((\operatorname{Hess}\mathcal{L})(\theta)) = \mathfrak{d} - k$. Combining this, Proposition 11, Lemma 38, and Proposition 39 with Proposition 47 shows that there exist $\mathfrak{m} \in \mathcal{M}\cap U$, $\mathfrak{c} \in (0,\infty)$, $V, \mathcal{V} \in \{A \subseteq U \colon A \text{ is compact}\}$ such that

(i) it holds that $\mathfrak{m} \in V^\circ \subseteq V \subseteq \mathcal{V}$,

(ii) it holds for all $\theta \in \mathcal{V}$ that $d_{\mathcal{M}\cap U}(\theta) = d_{\mathcal{M}\cap U\cap\mathcal{V}}(\theta)$, and

(iii) it holds for all $\theta \in V$, $\gamma \in (0, \min\{(\sup_{\vartheta\in\mathcal{M}\cap V_2} \Lambda((\operatorname{Hess} f)(\vartheta)))^{-1}, 1\}]$, $n \in \mathbb{N}_0$ that $\Theta_n^{\theta,\gamma} \in \mathcal{V}$ and $d_{\mathcal{M}\cap U}(\Theta_n^{\theta,\gamma}) \leq \exp(-\mathfrak{c}\gamma n^{1-\rho}) d_{\mathcal{M}\cap U}(\theta)$

(cf. Definition 30). In addition, note that

$$\begin{aligned}\sup_{\vartheta\in\mathcal{M}\cap\mathcal{V}} \Lambda((\operatorname{Hess} f)(\vartheta)) &\leq \sup_{\vartheta\in U} \Lambda((\operatorname{Hess} f)(\vartheta)) \\ &\leq (3N+1)\big(24\mathfrak{D}^5 + 16N\mathfrak{D}^7\big)\big(\sup_{x\in[a,b]} \mathfrak{p}(x)\big).\end{aligned} \tag{5.24}$$

Furthermore, observe that the fact that $\mathcal{L}|_U$ is twice continuously differentiable implies that there exists $\mathfrak{C} \in (0,\infty)$ which satisfies for all $\theta, \vartheta \in \mathcal{V}$ that $|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)| \leq \mathfrak{C}\|\theta - \vartheta\|$. This ensures that for all $\theta \in V^\circ$, $\gamma \in (0, ((3N+1)(16\mathfrak{D}^5 + 8N\mathfrak{D}^7)(\sup_{x\in[a,b]} \mathfrak{p}(x)))^{-1}]$, $n \in \mathbb{N}_0$ we have that

$$\begin{aligned}\mathcal{L}(\Theta_n^{\theta,\gamma}) &= \inf_{\vartheta\in\mathcal{M}\cap U\cap\mathcal{V}} |\mathcal{L}(\Theta_n^{\theta,\gamma}) - \mathcal{L}(\vartheta)| \leq \mathfrak{C}\big[\inf_{\vartheta\in\mathcal{M}\cap U\cap\mathcal{V}} \|\Theta_n^{\theta,\gamma} - \vartheta\|\big] \\ &= \mathfrak{C}\big[d_{\mathcal{M}\cap U}(\Theta_n^{\theta,\gamma})\big] \leq \mathfrak{C}\exp(-\mathfrak{c}\gamma n^{1-\rho}) d_{\mathcal{M}\cap U}(\theta).\end{aligned} \tag{5.25}$$

The proof of Theorem 48 is thus complete. ■

5.3 Convergence results for GD with random initializations in the training of ANNs

Corollary 49 *Assume Setting 3, let* $N \in \mathbb{N}\cap[1,H]$, $\rho \in [0,1)$, $x_0, x_1, \ldots, x_N, \alpha_1, \alpha_2, \ldots, \alpha_N \in \mathbb{R}$ *satisfy* $a = x_0 < x_1 < \cdots < x_N = b$, *assume for all* $i \in \{1,2,\ldots,N\}$, $x \in [x_{i-1}, x_i]$ *that* $f(x) = f(x_{i-1}) + \alpha_i(x - x_{i-1})$, *let* $\mathfrak{D} \in \mathbb{R}$ *satisfy*

$$\mathfrak{D} = 1 + |f(a)| + (1 + 2\max_{j\in\{1,2,\ldots,H\}} |\alpha_j|)(|a| + |b| + 1), \tag{5.26}$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space, let* $\Theta_n^\gamma \colon \Omega \to \mathbb{R}^{\mathfrak{d}}$, $\gamma \in \mathbb{R}$, $n \in \mathbb{N}_0$, *be random variables, assume for all* $\gamma \in \mathbb{R}$ *that* Θ_0^γ *is standard normally distributed, and assume for all* $\gamma \in \mathbb{R}$, $n \in \mathbb{N}$, $\omega \in \Omega$ *that*

$$\Theta_n^\gamma(\omega) = \Theta_{n-1}^\gamma(\omega) - \gamma n^{-\rho}\mathcal{G}(\Theta_{n-1}^\gamma(\omega)). \tag{5.27}$$

Then there exist $\mathfrak{c}, \mathfrak{C} \in (0, \infty)$ *such that for all* $\gamma \in (0, ((3N+1)(24\mathfrak{D}^5+16N\mathfrak{D}^7)(\sup_{x\in[a,b]} \mathfrak{p}(x)))^{-1}]$ *it holds that* $\mathbb{P}(\forall\, n \in \mathbb{N}_0 \colon \mathcal{L}(\Theta_n^\gamma) \le \mathfrak{C} \exp(-\mathfrak{c}\gamma n^{1-\rho})) \ge \mathfrak{c}$.

Proof [Proof of Corollary 49] Note that Theorem 48 ensures that there exist $\mathfrak{c}, \mathfrak{C} \in (0, \infty)$ and a non-empty open $U \subseteq \mathbb{R}^{\mathfrak{d}}$ such that for all $\gamma \in (0, ((3N+1)(24\mathfrak{D}^5+16N\mathfrak{D}^7)(\sup_{x\in[a,b]} \mathfrak{p}(x)))^{-1}]$, $\omega \in \Omega$, $n \in \mathbb{N}_0$ with $\Theta_0^\gamma(\omega) \in U$ it holds that

$$\mathcal{L}(\Theta_n^\gamma(\omega)) \le \mathfrak{C} \exp(-\mathfrak{c}\gamma n^{1-\rho}). \tag{5.28}$$

Observe that the fact that U is a non-empty open set and the assumption that for all $\gamma \in \mathbb{R}$ it holds that Θ_0^γ is standard normally distributed imply that there exists $\delta \in (0, \infty)$ such that for all $\gamma \in \mathbb{R}$ we have that $\mathbb{P}(\Theta_0^\gamma \in U) \ge \delta$. This completes the proof of Corollary 49.
∎

Corollary 50 *Assume Setting 3, let* $N \in \mathbb{N} \cap [1, H]$, $x_0, x_1, \ldots, x_N, \alpha_1, \alpha_2, \ldots, \alpha_N \in \mathbb{R}$ *satisfy* $a = x_0 < x_1 < \cdots < x_N = b$, *assume for all* $i \in \{1, 2, \ldots, N\}$, $x \in [x_{i-1}, x_i]$ *that* $f(x) = f(x_{i-1}) + \alpha_i(x - x_{i-1})$, *let* $\mathfrak{D} \in \mathbb{R}$ *satisfy*

$$\mathfrak{D} = 1 + |f(a)| + (1 + 2\max\nolimits_{j\in\{1,2,\ldots,H\}}|\alpha_j|)(|a| + |b| + 1), \tag{5.29}$$

let $\Theta_n^{k,\gamma} \colon \Omega \to \mathbb{R}^{\mathfrak{d}}$, $k, n \in \mathbb{N}_0$, $\gamma \in \mathbb{R}$, *and* $\mathbf{k}_n^{k,\gamma} \colon \Omega \to \mathbb{N}$, $k, n \in \mathbb{N}_0$, $\gamma \in \mathbb{R}$, *be random variables, assume for all* $\gamma \in \mathbb{R}$ *that* $\Theta_0^{k,\gamma}$, $k \in \mathbb{N}$, *are independent standard normal random variables, and assume for all* $k \in \mathbb{N}$, $\gamma \in \mathbb{R}$, $n \in \mathbb{N}_0$, $\omega \in \Omega$ *that*

$$\Theta_{n+1}^{k,\gamma}(\omega) = \Theta_n^{k,\gamma}(\omega) - \gamma\mathcal{G}(\Theta_n^{k,\gamma}(\omega)) \tag{5.30}$$

and

$$\mathbf{k}_n^{k,\gamma}(\omega) \in \arg\min\nolimits_{\ell\in\{1,2,\ldots,k\}} \mathcal{L}(\Theta_n^{\ell,\gamma}(\omega)). \tag{5.31}$$

Then it holds for all $\gamma \in (0, ((3N + 1)(24\mathfrak{D}^5 + 16N\mathfrak{D}^7)(\sup_{x\in[a,b]} \mathfrak{p}(x)))^{-1}]$ *that*

$$\liminf\nolimits_{K\to\infty} \mathbb{P}\Big(\limsup\nolimits_{n\to\infty} \mathcal{L}\big(\Theta_n^{\mathbf{k}_n^{K,\gamma},\gamma}\big) = 0\Big) = 1. \tag{5.32}$$

Proof [Proof of Corollary 50] Throughout this proof let $\mathfrak{g} \in \mathbb{R}$ satisfy $\mathfrak{g} = ((3N+1)(24\mathfrak{D}^5 + 16N\mathfrak{D}^7)(\sup_{x\in[a,b]} \mathfrak{p}(x)))^{-1}$. Note that Theorem 48 assures that there exist $\mathfrak{c}, \mathfrak{C} \in (0, \infty)$ and a non-empty open $U \subseteq (-\mathfrak{D}, \mathfrak{D})^{\mathfrak{d}}$ such that for all $\gamma \in (0, \mathfrak{g}]$, $k \in \mathbb{N}$, $\omega \in \Omega$, $n \in \mathbb{N}_0$ with $\Theta_0^{k,\gamma}(\omega) \in U$ it holds that $\mathcal{L}(\Theta_n^{k,\gamma}(\omega)) \le \mathfrak{C}\exp(-\mathfrak{c}\gamma n)$. Hence, we obtain for all $\gamma \in (0, \mathfrak{g}]$, $k \in \mathbb{N}$, $\omega \in \Omega$ with $\Theta_0^{k,\gamma}(\omega) \in U$ that $\limsup_{n\to\infty} \mathcal{L}(\Theta_n^{k,\gamma}(\omega)) = 0$. Next observe that (5.31) ensures for all $K \in \mathbb{N}$, $\gamma \in (0, \mathfrak{g}]$ that

$$\mathbb{P}\Big(\limsup\nolimits_{n\to\infty} \mathcal{L}\big(\Theta_n^{\mathbf{k}_n^{K,\gamma},\gamma}\big) = 0\Big) \ge \mathbb{P}\Big(\exists\, k \in \{1, 2, \ldots, K\} \colon \big[\limsup\nolimits_{n\to\infty} \mathcal{L}(\Theta_n^{k,\gamma}) = 0\big]\Big). \tag{5.33}$$

Furthermore, note that the fact that for all $\gamma \in (0, \mathfrak{g}]$, $k \in \mathbb{N}$, $\omega \in \Omega$ with $\Theta_0^{k,\gamma}(\omega) \in U$ it holds that $\limsup_{n\to\infty} \mathcal{L}(\Theta_n^{k,\gamma}(\omega)) = 0$ shows that for all $K \in \mathbb{N}$, $\gamma \in (0, \mathfrak{g}]$ we have that

$$\mathbb{P}\Big(\exists\, k \in \{1, 2, \ldots, K\} \colon \big[\limsup\nolimits_{n\to\infty} \mathcal{L}(\Theta_n^{k,\gamma}) = 0\big]\Big) \ge \mathbb{P}\Big(\exists\, k \in \{1, 2, \ldots, K\} \colon \Theta_0^{k,\gamma} \in U\Big). \tag{5.34}$$

In addition, observe that the fact that for all $\gamma \in \mathbb{R}$ it holds that $\Theta_0^{k,\gamma}$, $k \in \mathbb{N}$, are i.i.d. implies that for all $K \in \mathbb{N}$, $\gamma \in (0, \mathfrak{g}]$ we have that

$$\begin{aligned}\mathbb{P}\Big(\exists\, k \in \{1,2,\dots,K\} \colon \Theta_0^{k,\gamma} \in U\Big) &= 1 - \mathbb{P}\Big(\forall\, k \in \{1,2,\dots,K\} \colon \Theta_0^{k,\gamma} \in (\mathbb{R}^{\mathfrak{d}} \backslash U)\Big) \\ &= 1 - \big[\mathbb{P}\big(\Theta_0^{1,\gamma} \in (\mathbb{R}^{\mathfrak{d}} \backslash U)\big)\big]^K.\end{aligned} \tag{5.35}$$

Moreover, note that the fact that U is non-empty and open and the fact that for all $\gamma \in \mathbb{R}$ it holds that $\Theta_0^{1,\gamma}$ is standard normally distributed prove that for all $\gamma \in \mathbb{R}$ we have that $\mathbb{P}\big(\Theta_0^{1,\gamma} \in (\mathbb{R}^{\mathfrak{d}} \backslash U)\big) < 1$. This and (5.35) demonstrate for all $\gamma \in (0, \mathfrak{g}]$ that

$$\liminf_{K \to \infty} \mathbb{P}\Big(\exists\, k \in \{1,2,\dots,K\} \colon \Theta_0^{k,\gamma} \in U\Big) = 1. \tag{5.36}$$

Combining this with (5.33) and (5.34) shows for all $\gamma \in (0, \mathfrak{g}]$ that

$$\liminf_{K \to \infty} \mathbb{P}\Big(\limsup_{n \to \infty} \mathcal{L}\big(\Theta_n^{\mathbf{k}_n^{K,\gamma},\gamma}\big) = 0\Big) = 1. \tag{5.37}$$

The proof of Corollary 50 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

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.

Ö. D. Akyildiz and S. Sabanis. Nonasymptotic analysis of Stochastic Gradient Hamiltonian Monte Carlo under local conditions for nonconvex optimization. *arXiv preprint arXiv:2002.05465*, 2021.

Z. Allen-Zhu, Y. Li, and Y. Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. 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, pages 6158–6169. Curran Associates, Inc., 2019a. URL `https://proceedings.neurips.cc/paper/2019/file/62dad6e273d32235ae02b7d321578ee8-Paper.pdf`.

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of* *the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 09–15 Jun 2019b. URL `http://proceedings.mlr.press/v97/allen-zhu19a.html`.

S. Arora, S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In K. Chaudhuri and R. 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`.

H. Attouch and J. 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.

F. Bach and E. 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`.

B. Bercu and J.-C. Fort. *Generic Stochastic Gradient Methods*, pages 1–8. American Cancer Society, 2013. URL `https://doi.org/10.1002/9780470400531.eorms1068`.

D. P. Bertsekas and J. N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000. doi: 10.1137/S1052623497331063.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *arXiv preprint arXiv:1606.04838*, 2018.

P. Cheridito, A. Jentzen, and F. Rossmannek. Non-convergence of stochastic gradient descent in the training of deep neural networks. *Journal of Complexity*, page 101540, 2020. doi: 10.1016/j.jco.2020.101540.

P. Cheridito, A. Jentzen, A. Riekert, and F. 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, 26, 2022a. doi: 10.1016/j.jco.2022.101646.

P. Cheridito, A. Jentzen, and F. Rossmannek. Landscape Analysis for Shallow Neural Networks: Complete Classification of Critical Points for Affine Target Functions. *J. Nonlinear Sci.*, 32(5):Paper No. 64, 2022b. doi: 10.1007/s00332-022-09823-8.

S. Dereich and S. Kassing. Convergence of stochastic gradient descent schemes for Lojasiewicz-landscapes. *arXiv preprint arXiv:2102.09385*, 2021.

S. Dereich and S. Kassing. On minimal representations of shallow ReLU networks. *Neural Networks*, 148:121–128, 2022. doi: https://doi.org/10.1016/j.neunet.2022.01.006.

S. Dereich and T. Müller-Gronbach. General multilevel adaptations for stochastic approximation algorithms of Robbins-Monro and Polyak-Ruppert type. *Numer. Math.*, 142(2): 279–328, 2019. doi: 10.1007/s00211-019-01024-y.

S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In K. Chaudhuri and R. 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 2019a. PMLR. URL `http://proceedings.mlr.press/v97/du19c.html`.

S. S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b. URL `https://openreview.net/forum?id=S1eK3i09YQ`.

W. E, C. Ma, S. Wojtowytsch, and L. Wu. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't. *arXiv preprint arXiv:2009.10713*, 2020a.

W. E, C. Ma, and L. 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, 2020b. doi: 10.1007/s11425-019-1628-5.

B. Fehrman, B. Gess, and A. Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *J. Mach. Learn. Res.*, 21:Paper No. 136, 48, 2020.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In P. Grünwald, E. Hazan, and S. 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.

G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2016.

A. Jentzen and T. Kröger. Convergence rates for gradient descent in the training of over-parameterized artificial neural networks with biases. *arXiv preprint arXiv:2102.11840*, 2021.

A. Jentzen and A. Riekert. Convergence analysis for gradient flows in the training of artificial neural networks with ReLU activation. *Journal of Mathematical Analysis and Applications*, page 126601, 2022. doi: https://doi.org/10.1016/j.jmaa.2022.126601.

A. Jentzen and P. von Wurstemberger. Lower error bounds for the stochastic gradient descent optimization algorithm: Sharp convergence rates for slowly and fast decaying learning rates. *Journal of Complexity*, 57:101438, 2020. doi: 10.1016/j.jco.2019.101438.

A. Jentzen, B. Kuckuck, A. Neufeld, and P. von Wurstemberger. Strong error analysis for stochastic gradient descent optimization algorithms. *IMA J. Numer. Anal.*, 41(1): 455–492, 2021. doi: 10.1093/imanum/drz055.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. *arXiv preprint arXiv:1608.04636*, 2020.

T. Kato. *Perturbation theory for linear operators*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition.

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In V. Feldman, A. Rakhlin, and O. 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`.

J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. 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.

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.

Y. Li and Y. Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. 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 8157–8166. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/54fe976ba170c19ebae453679b362263-Paper.pdf`.

A. Lovas, I. Lytras, M. Rásonyi, and S. Sabanis. Taming neural networks with TUSLA: Non-convex learning via adaptive stochastic gradient Langevin algorithms. *arXiv preprint arXiv:2006.14514*, 2020.

L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis. Dying ReLU and initialization: Theory and numerical examples. *Communications in Computational Physics*, 28(5):1671–1706, 2020. doi: 10.4208/cicp.OA-2020-0165.

E. Moulines and F. 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`.

Y. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. doi: 10.1007/978-1-4419-8853-9. A basic course.

Y. Nesterov. Universal gradient methods for convex optimization problems. *Math. Program.*, 152(1-2, Ser. A):381–404, 2015. doi: 10.1007/s10107-014-0790-0.

I. Panageas and G. Piliouras. Gradient Descent Only Converges to Minimizers: Non-Isolated Critical Points and Invariant Regions. In C. 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.

I. Panageas, G. Piliouras, and X. Wang. First-order methods almost always avoid saddle points: the case of vanishing step-sizes. *arXiv preprint arXiv:1906.07772*, 2019.

V. Patel. Stopping criteria for, and strong convergence of, stochastic gradient descent on Bottou-Curtis-Nocedal functions. *arXiv preprint arXiv:2004.00475*, 2021.

A. Rakhlin, O. Shamir, and K. 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.

G. Rotskoff and E. Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of 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/196f5641aa9dc87067da4ff90fd81e7b-Paper.pdf`.

S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2017.

K. A. Sankararaman, S. De, Z. Xu, W. R. Huang, and T. Goldstein. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. *arXiv preprint arXiv:1904.06963*, 2020.

M. Schmidt and N. L. Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

O. Shamir. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. In A. Beygelzimer and D. Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2691–2713, Phoenix, USA, 6 2019. PMLR. URL `http://proceedings.mlr.press/v99/shamir19a.html`.

L. W. Tu. *An introduction to manifolds*. Universitext. Springer, New York, second edition, 2011. doi: 10.1007/978-1-4419-7400-6.

S. Wojtowytsch. Stochastic gradient descent with noise of machine learning type. Part I: Discrete time analysis. *arXiv preprint arXiv:2105.01650*, 2021.

Y. Xu and W. 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.

G. Zhang, J. Martens, and R. 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.

D. Zou, Y. Cao, D. Zhou, and Q. Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109:467–492, 2020. doi: 10.1007/s10994-019-05839-6.