Convergence analysis for gradient flows in the training of artificial neural networks with ReLU activation

Arnulf Jentzen[1,2] and Adrian Riekert[3]

[1] Applied Mathematics: Institute for Analysis and Numerics, University of Münster, Germany, e-mail: ajentzen@uni-muenster.de

[2] School of Data Science and Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China, e-mail: ajentzen@cuhk.edu.cn

[3] Applied Mathematics: Institute for Analysis and Numerics, University of Münster, Germany, e-mail: ariekert@uni-muenster.de

February 14, 2022

Abstract

Gradient descent (GD) type optimization schemes are the standard methods to train artificial neural networks (ANNs) with rectified linear unit (ReLU) activation. Such schemes can be considered as discretizations of gradient flows (GFs) associated to the training of ANNs with ReLU activation and most of the key difficulties in the mathematical convergence analysis of GD type optimization schemes in the training of ANNs with ReLU activation seem to be already present in the dynamics of the corresponding GF differential equations. It is the key subject of this work to analyze such GF differential equations in the training of ANNs with ReLU activation and three layers (one input layer, one hidden layer, and one output layer). In particular, in this article we prove in the case where the target function is possibly multi-dimensional and continuous and in the case where the probability distribution of the input data is absolutely continuous with respect to the Lebesgue measure that the risk of every bounded GF trajectory converges to the risk of a critical point. In addition, in this article we show in the case of a 1-dimensional affine linear target function and in the case where the probability distribution of the input data coincides with the standard uniform distribution that the risk of every bounded GF trajectory converges to zero if the initial risk is sufficiently small. Finally, in the special situation where there is only one neuron on the hidden layer (1-dimensional hidden layer) we strengthen the above named result for affine linear target functions by proving that that the risk of every (not necessarily bounded) GF trajectory converges to zero if the initial risk is sufficiently small.

Contents

1 Introduction **2**

2 Properties of the risk function and its gradient **5**
2.1 Mathematical description of artificial neural networks (ANNs) 6
2.2 An upper bound for the norm of the gradient of the risk function 7
2.3 Continuous dependence of active neuron regions on ANN parameters 8
2.4 Differentiability of the risk function 10
2.5 Lower semicontinuity of the norm of the gradient of the risk function 14

3 Convergence of the risk of gradient flows (GFs) in the training of ANNs **16**
3.1 Convergence of the risk of GFs to the risk of a critical point 17
3.2 Convergence of the risk of GFs to the minimal risk 17

3.3 Risks of critical points for affine linear target functions 18
3.4 Convergence of the risk of GFs to the minimal risk for affine linear target functions 19

4 A priori estimates for GFs in the training of ANNs 19
4.1 Lyapunov type functions for GFs 20
4.2 A priori estimates for GFs with large risk 21
4.3 Invariant quantities for GFs 22

5 Properties of ANN parametrizations with small risk and one hidden neuron 22
5.1 Mean square approximations through constant functions 23
5.2 Mathematical description of ANNs with one hidden neuron 24
5.3 Properties of ANNs with small risk and one hidden neuron 25

6 Convergence of the risk of GFs in the training of ANNs with one hidden neuron 28
6.1 A priori estimates for GFs 29
6.2 Properties of ANN parameters for convergent sequences of ANN realizations . . . 31
6.3 Convergence of the risk of GFs to zero for affine linear target functions 32
6.4 Uniform convergence of realizations of GFs for affine linear target functions . . . 35

1 Introduction

Gradient descent (GD) type optimization schemes are the standard tools in the training of feedforward fully connected artificial neural networks (ANNs) with rectified linear unit (ReLU) activation. Such GD type optimization schemes can be considered as temporal discretization methods for the associated gradient flow (GF) differential equations and most of the key difficulties which arise in the mathematical convergence analysis of GD type optimization schemes in the training of ANNs with ReLU activation already arise in the mathematical convergence analysis of the corresponding GFs. It is the key subject of this article to analyze such GFs arising in the training of ANNs with ReLU activation and, in particular, to prove that the risk of every bounded GF trajectory converges in the training of ANNs with ReLU activation to the risk of a critical point. We are particularly interested in the mathematical convergence analysis of GF trajectories instead of time discrete GD optimization schemes since, on the one hand, most of the key difficulties which arise in the mathematical analysis of GD type optimization schemes in the training of ANNs with ReLU activation already arise in the mathematical analysis of the corresponding GFs and since, on the other hand, the consideration of such GF trajectories allows us to focus on precisely such key difficulties.

In the scientific literature there are several quite promising approaches regarding the mathematical convergence analysis for GD type optimization schemes and GFs, respectively. For instance, we point to [11, 13, 15, 17] for results on the convergence of GF in the training of ANNs in the overparametrized regime, where the number of neurons has to be sufficiently large when compared to the number of used input-output data pairs. Another promising idea is to view the neurons of an ANN as interacting particles and consider the limit of the associated empirical measures as the number of neurons increases to infinity. The limiting process of the corresponding GFs is known in the scientific literature as Wasserstein gradient flow; cf., e.g., [5, 9, 10], the overview article [14], and the references mentioned therein. Most convergence results for the Wasserstein gradient flow require smoothness assumptions on the considered risk function, which are not satisfied for ANNs with ReLU activation. To overcome this issue, a different parametrization for ReLU networks has been proposed in [10, Section 4.2]. In [2, 8] GF processes have been considered in the context of training deep linear neural networks, in which the employed activation function is the identity. The behavior of the realization functions of

ANNs with one hidden layer and ReLU activation under the GF dynamics has been investigated in more detail in [19, 23]. Another recent idea is to consider only very special target functions and we refer, in particular, to [6, 18] for convergence results for GF and GD processes in the case of constant target functions. In the more general case of affine linear target functions, the critical points of the risk function were characterized in [7] and parts of the analysis in this article exploit this characterization. For further abstract convergence results on GF processes we point, e.g., to [1, 4, 16, 22] and the references mentioned therein.

It is the key topic of this article to provide some first basics steps regarding the mathematical convergence analysis of GFs arising in the training of ANNs with ReLU activation. Specifically, in one of main results of this article, see item (iv) in Theorem 1.1 in this introductory section, we prove that the risk of every bounded GF trajectory converges in the training of ANNs with one hidden layer and ReLU activation to the risk of a critical point. In Theorem 1.1 below we study fully connected feedfoward ANNs with a d-dimensional input layer (with $d \in \mathbb{N} = \{1, 2, 3, ...\}$ neurons on the input layer), with an H-dimensional hidden layer (with $H \in \mathbb{N}$ neurons on the hidden layer), and with a 1-dimensional output layer (with one neuron on the output layer). There are thus Hd scalar real weight parameters and H scalar real bias parameters to describe the affine linear transformation in between the d-dimensional input layer and the H-dimensional hidden layer and there are thus H scalar real weight parameters and 1 scalar real bias parameter to describe the affine linear transformation in between the H-dimensional hidden layer and the 1-dimensional output layer. Overall the ANNs in Theorem 1.1 thus consist of precisely $\mathfrak{d} = dH + 2H + 1$ scalar real ANN parameters.

In Theorem 1.1 we study fully connected feedfoward ANNs with the ReLU activation function $\mathbb{R} \ni x \mapsto \max\{x, 0\} \in \mathbb{R}$ (which is also referred to as rectifier function) as the activation function. The ReLU activation function $\mathbb{R} \ni x \mapsto \max\{x, 0\} \in \mathbb{R}$ fails to be differentiable and can thus not be used to specify gradients in GD type optimization schemes and GFs, respectively. A common procedure to overcome this issue (cf. [18] and [6]) is to approximate the ReLU activation function $\mathbb{R} \ni x \mapsto \max\{x, 0\} \in \mathbb{R}$ through appropriate continuously differentiable functions which converge pointwise to the ReLU activation function and whose derivatives converge pointwise to the *left derivative* of the ReLU activation function. In Theorem 1.1 the function $\mathfrak{R}_\infty \colon \mathbb{R} \to \mathbb{R}$ specifies the ReLU activation function and the functions $\mathfrak{R}_r \colon \mathbb{R} \to \mathbb{R}$, $r \in \mathbb{N}$, serve as such continuously differentiable approximations of the ReLU activation function; see (1) in Theorem 1.1.

The finite measure $\mu \colon [a, b]^d \to [0, \infty]$ in Theorem 1.1 specifies up to a normalization constant the probability distribution of the input data of the supervised learning problem considered in Theorem 1.1. In Theorem 1.1 we assume that the measure $\mu \colon [a, b]^d \to [0, \infty]$ is absolutely continuous with respect to the Lebesgue measure. The functions $\mathcal{L}_r \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$, $r \in \mathbb{N} \cup \{\infty\}$, in Theorem 1.1 describe the risk functions associated to the considered ANNs in the sense that for all $r \in \mathbb{N} \cup \{\infty\}$ we have that $\mathcal{L}_r \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ is the risk function associated to the target function $f \colon [a, b]^d \to \mathbb{R}$ and the fully connected feedforward ANNs with the activation function $\mathfrak{R}_r \colon \mathbb{R} \to \mathbb{R}$; see (2) in Theorem 1.1 for details.

The function $\|\cdot\| \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ in Theorem 1.1 is nothing else but the standard norm on the ANN parameter space $\mathbb{R}^{\mathfrak{d}} = \mathbb{R}^{dH+2H+1}$. The function $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ in Theorem 1.1 specifies the generalized gradients of the risk function $\mathcal{L}_\infty \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ using the continuously differentiable approximations $\mathfrak{R}_r \colon \mathbb{R} \to \mathbb{R}$, $r \in \mathbb{N}$.

Item (i) in Theorem 1.1 asserts that the generalized gradient function $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ is locally bounded and measurable. This statement is provided to ensure that for every continuous function $\Theta = (\Theta_t)_{t \in [0,\infty)} \colon [0, \infty) \to \mathbb{R}^{\mathfrak{d}}$ and every $t \in [0, \infty)$ we have that the Lebesgue integral $\int_0^t \mathcal{G}(\Theta_s) \, \mathrm{d}s$ makes sense (cf. items (iv) and (v) in Theorem 1.1).

Item (ii) in Theorem 1.1 reveals that the generalized gradient function $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ is *lower semicontinuous*. In the case of ANNs with smooth activation functions it follows directly from Lebesgue's theorem of dominated convergence that the gradient function of the risk function

is continuous. In the case of ANNs with ReLU activation, however, the generalized gradient function $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ fails to be continuous but in item (ii) in Theorem 1.1 we prove that this generalized gradient function is instead lower semicontinuous.

Item (iii) in Theorem 1.1 connects the generalized gradient function $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ with standard gradients of the risk function $\mathcal{L}_\infty\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ by demonstrating that there exists an open set $U \subseteq \mathbb{R}^{\mathfrak{d}}$ with full Lebesgue measure such that $\mathcal{L}_\infty$ restricted to U is continuously differentiable with $\mathcal{G}|_U\colon U \to \mathbb{R}^{\mathfrak{d}}$ being the gradient of $(\mathcal{L}_\infty)|_U\colon U \to \mathbb{R}$.

Item (iv) in Theorem 1.1 establishes that the risk of every bounded GF trajectory converges in the training of the considered ANNs to the risk of a critical point. Item (v) in Theorem 1.1 reveals that the risk of every bounded GF trajectory with sufficiently small initial risk converges in the training of the considered ANNs to the risk of the global minima of $\mathcal{L}_\infty\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$. We now present the precise statement of Theorem 1.1.

Theorem 1.1. *Let $d, H, \mathfrak{d} \in \mathbb{N}$, $\mathscr{a} \in \mathbb{R}$, $\mathscr{b} \in (\mathscr{a}, \infty)$, $f \in C([\mathscr{a},\mathscr{b}]^d, \mathbb{R})$ satisfy $\mathfrak{d} = dH + 2H + 1$, let $\mathfrak{R}_r \in C(\mathbb{R},\mathbb{R})$, $r \in \mathbb{N} \cup \{\infty\}$, satisfy for all $x \in \mathbb{R}$ that $(\bigcup_{r\in\mathbb{N}}\{\mathfrak{R}_r\}) \subseteq C^1(\mathbb{R},\mathbb{R})$, $\mathfrak{R}_\infty(x) = \max\{x, 0\}$, $\sup_{r\in\mathbb{N}} \sup_{y\in[-|x|,|x|]}(|\mathfrak{R}_r(y)| + |(\mathfrak{R}_r)'(y)|) < \infty$, and*

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

let $\mu\colon \mathcal{B}([\mathscr{a},\mathscr{b}]^d) \to [0,\infty]$ be a finite measure, let $\mathcal{L}_r\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$, $r \in \mathbb{N} \cup \{\infty\}$, satisfy for all $r \in \mathbb{N} \cup \{\infty\}$, $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ that

$$\begin{aligned}\mathcal{L}_r(\theta) = \int_{[\mathscr{a},\mathscr{b}]^d} \big(&f(x_1,\ldots,x_d) \\ &- \theta_{\mathfrak{d}} - \textstyle\sum_{i=1}^{H} \theta_{H(d+1)+i}\big[\mathfrak{R}_r(\theta_{Hd+i} + \sum_{j=1}^{d} \theta_{(i-1)d+j} x_j)\big]\big)^2 \mu(\mathrm{d}(x_1,\ldots,x_d)),\end{aligned} \tag{2}$$

let $\|\cdot\|\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $x = (x_1, \ldots, x_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ that $\|x\| = [\sum_{i=1}^{\mathfrak{d}} |x_i|^2]^{1/2}$, let $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \{\vartheta \in \mathbb{R}^{\mathfrak{d}}\colon ((\nabla\mathcal{L}_r)(\vartheta))_{r\in\mathbb{N}}$ is convergent$\}$ that $\mathcal{G}(\theta) = \lim_{r\to\infty}(\nabla\mathcal{L}_r)(\theta)$, and assume that μ is absolutely continuous with respect to the Lebesgue measure on $[\mathscr{a},\mathscr{b}]^d$. Then

(i) it holds that $\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \mathcal{G}(\theta) \in \mathbb{R}^{\mathfrak{d}}$ is locally bounded and measurable,

(ii) it holds that $\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \|\mathcal{G}(\theta)\| \in \mathbb{R}$ is lower semicontinuous,

(iii) there exists an open $U \subseteq \mathbb{R}^{\mathfrak{d}}$ which satisfies $\int_{\mathbb{R}^{\mathfrak{d}}\setminus U} 1\,\mathrm{d}x = 0$, $(\mathcal{L}_\infty)|_U \in C^1(U,\mathbb{R})$, and $\nabla((\mathcal{L}_\infty)|_U) = \mathcal{G}|_U$,

(iv) it holds for all $\Theta \in C([0,\infty),\mathbb{R}^{\mathfrak{d}})$ with $\sup_{t\in[0,\infty)}\|\Theta_t\| < \infty$ and $\forall\, t \in [0,\infty)\colon \Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s$ that there exists $\vartheta \in \mathcal{G}^{-1}(\{0\})$ such that $\limsup_{t\to\infty} \mathcal{L}_\infty(\Theta_t) = \mathcal{L}_\infty(\vartheta)$, and

(v) it holds for all $\Theta \in C([0,\infty),\mathbb{R}^{\mathfrak{d}})$ with $\sup_{t\in[0,\infty)}\|\Theta_t\| < \infty$, $\forall\, t \in [0,\infty)\colon \Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s$, and $\forall\, \theta \in \mathcal{G}^{-1}(\{0\}) \cap (\mathcal{L}_\infty)^{-1}((\inf_{\vartheta\in\mathbb{R}^{\mathfrak{d}}} \mathcal{L}_\infty(\vartheta), \infty))\colon \mathcal{L}_\infty(\Theta_0) < \mathcal{L}_\infty(\theta)$ that

$$\limsup_{t\to\infty} \mathcal{L}_\infty(\Theta_t) = \inf_{\vartheta\in\mathbb{R}^{\mathfrak{d}}} \mathcal{L}_\infty(\vartheta). \tag{3}$$

Item (i) in Theorem 1.1 is a direct consequence of Corollary 2.4 below, item (ii) in Theorem 1.1 is a direct consequence of Corollary 2.16 below, item (iii) in Theorem 1.1 is a direct consequence of Corollary 2.17 below, item (iv) in Theorem 1.1 is a direct consequence of Theorem 3.2 below, and item (v) in Theorem 1.1 is a direct consequence of Corollary 3.3 below.

In Theorem 1.2 below we specialise the setup in Theorem 1.1 to the specific situation where there the input is 1-dimensional (where there is only one neuron on the input layer), where the measure $\mu\colon \mathcal{B}([\mathscr{a},\mathscr{b}]) \to [0,\infty]$ coincides with the Lebesgue–Borel measure, and where the

target function $f\colon [\mathscr{a},\mathscr{b}] \to \mathbb{R}$ is affine linear in the sense that there exist $\alpha, \beta \in \mathbb{R}$ such that for all $x \in [\mathscr{a},\mathscr{b}]$ it holds that

$$f(x) = \alpha x + \beta \tag{4}$$

to establish that the risk of every (bounded) GF trajectory with sufficiently small initial risk converges to zero. Specifically, in the specific situation of (4) we prove in Theorem 1.2 that for every continuous GF trajectory $\Theta\colon [0,\infty) \to \mathbb{R}^{3H+1}$ with

$$\sup_{t\in[0,\infty)}((H-1)\|\Theta_t\|) < \infty \tag{5}$$

and

$$\mathcal{L}_\infty(\Theta_0) < \frac{\alpha^2(\mathscr{b}-\mathscr{a})^3}{12(2\lfloor H/2\rfloor+1)^4} \tag{6}$$

we have that $\limsup_{t\to\infty} \mathcal{L}_\infty(\Theta_t) = 0$. In this specific situation of a 1-dimensional input $d = 1$ (in this specific situation where there is only one neuron on the input layer) we observe that the ANN parameter space $\mathbb{R}^{\mathfrak{d}}$ simplifies to $\mathbb{R}^{\mathfrak{d}} = \mathbb{R}^{dH+2H+1} = \mathbb{R}^{3H+1}$. Moreover, we note that in Theorem 1.2 below and in (5) above, respectively, we assume in the case where the number $H \in \mathbb{N}$ of neurons on the hidden layer is strictly bigger than 1 (in the case where $H > 1$) that the GF trajectory is bounded. We now present the precise statement of Theorem 1.2.

Theorem 1.2. *Let $H, \mathfrak{d} \in \mathbb{N}$, $\alpha, \beta, \mathscr{a} \in \mathbb{R}$, $\mathscr{b} \in (\mathscr{a},\infty)$ satisfy $\mathfrak{d} = 3H+1$, let $\mathfrak{R}_r \in C(\mathbb{R},\mathbb{R})$, $r \in \mathbb{N}\cup\{\infty\}$, satisfy for all $x \in \mathbb{R}$ that $(\bigcup_{r\in\mathbb{N}}\{\mathfrak{R}_r\}) \subseteq C^1(\mathbb{R},\mathbb{R})$, $\mathfrak{R}_\infty(x) = \max\{x,0\}$, $\sup_{r\in\mathbb{N}}\sup_{y\in[-|x|,|x|]}(|\mathfrak{R}_r(y)| + |(\mathfrak{R}_r)'(y)|) < \infty$, and*

$$\limsup_{r\to\infty}\big(|\mathfrak{R}_r(x) - \mathfrak{R}_\infty(x)| + |(\mathfrak{R}_r)'(x) - \mathbb{1}_{(0,\infty)}(x)|\big) = 0, \tag{7}$$

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

$$\mathcal{L}_r(\theta) = \int_{\mathscr{a}}^{\mathscr{b}} \big(\alpha x + \beta - \theta_{\mathfrak{d}} - \textstyle\sum_{i=1}^{H} \theta_{2H+i}\big[\mathfrak{R}_r(\theta_{H+i} + \theta_i x)\big]\big)^2 \,\mathrm{d}x, \tag{8}$$

let $\|\cdot\|\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $x = (x_1,\ldots,x_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ that $\|x\| = [\sum_{i=1}^{n}|x_i|^2]^{1/2}$, let $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \{\vartheta \in \mathbb{R}^{\mathfrak{d}}\colon ((\nabla\mathcal{L}_r)(\vartheta))_{r\in\mathbb{N}}$ is convergent$\}$ that $\mathcal{G}(\theta) = \lim_{r\to\infty}(\nabla\mathcal{L}_r)(\theta)$, and let $\Theta \in C([0,\infty),\mathbb{R}^{\mathfrak{d}})$ satisfy $\sup_{t\in[0,\infty)}((H-1)\|\Theta_t\|) < \infty$, $\forall\, t \in [0,\infty)\colon \Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s$, and $\mathcal{L}_\infty(\Theta_0) < \frac{\alpha^2(\mathscr{b}-\mathscr{a})^3}{12(2\lfloor H/2\rfloor+1)^4}$. Then $\limsup_{t\to\infty}\mathcal{L}_\infty(\Theta_t) = 0$.

Theorem 1.2 is a direct consequence of Corollary 3.5 (in the case $H > 1$) and Corollary 6.8 (in the case $H = 1$) below. The remainder of this article is organized as follows. In Section 2 below we establish certain regularity properties for the generalized gradient function $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ in Theorem 1.1 above. In Section 3 below we employ the regularity properties for the generalized gradient function $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ from Section 2 to prove items (iv) and (v) in Theorem 1.1 and to prove Theorem 1.2 under the more restrictive assumption that $\sup_{t\in[0,\infty)}\|\Theta_t\| < \infty$; cf. (5) above. In Section 4 below we establish suitable a priori bounds for GF trajectories. In Sections 5 and 6 we employ the a priori bounds from Section 4 to prove Theorem 1.2 under the more general assumption that $\sup_{t\in[0,\infty)}((H-1)\|\Theta_t\|) < \infty$; cf. (5) above.

2 Properties of the risk function and its gradient

In this section we establish several regularity properties for the risk function associated to the considered supervised learning problem; see (2) above. In particular, in Proposition 2.11 in Subsection 2.4 below we provide in (40) a sufficient condition to ensure that the risk function is differentiable and in Corollary 2.16 in Subsection 2.5 below we prove that the standard norm of the generalized gradient function $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ associated to the risk function is lower

semicontinuous. In the scientific literature results similar to Proposition 2.11 can, e.g., be found in Cheridito et al. [7]. In particular, in the case of only one neuron on the input layer (in the case of a 1-dimensional input) results similar to Proposition 2.11 have been shown in [7, Lemma 3.4 and Lemma 3.7].

Our proof of Proposition 2.11 employs the known representation result for the generalized gradient function $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ in Proposition 2.2 in Subsection 2.2 below, the well known local Lipschitz continuity result for the risk function in Lemma 2.9 in Subsection 2.4, the elementary Lipschitz type estimate for certain affine linear functions in Lemma 2.10 in Subsection 2.4, and the fact that appropriate active neuron regions depend continuously on the ANN parameters which we establish in Corollary 2.8 in Subsection 2.3 below. Our proof of Corollary 2.16 employs the fact that the absolute value of every component of the generalized gradient function $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ is lower semicontinuous which we establish in Corollary 2.15 in Subsection 2.5. Our proof of Corollary 2.15 uses the regularity results for the absolute values of the components of the generalized gradient function $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ in Lemma 2.12, Lemma 2.13, and Lemma 2.14 in Subsection 2.5. Our proof of Corollary 2.8 uses the appropriate continuity result for active neuron regions in Lemma 2.5 and the well-known results on absolutely continuous measures in Lemma 2.6 and Corollary 2.7. In the scientific literature Lemma 2.6 can, e.g., be found in Rudin [21, Theorem 6.11].

In Setting 2.1 in Subsection 2.1 below we present the mathematical framework which we frequently employ in Sections 2–4 to formulate ANNs with one hidden layer and ReLU activation and the corresponding risk functions (see (11) and (12) in Setting 2.1), in the elementary regularity result in Lemma 2.3 in Subsection 2.2 we establish an elementary a priori bound for the norm of the generalized gradient function $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$, and in the elementary regularity result in Corollary 2.4 in Subsection 2.2 we demonstrate that the generalized gradient function $\mathcal{G}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ is locally bounded and measurable. Lemma 2.3 is used in the proof of Corollary 2.4 in Subsection 2.2 and Corollary 2.4 is employed in Section 3 and in item (i) in Theorem 1.1. Only for completeness we include in this section detailed proofs for Proposition 2.2, Lemma 2.3, Corollary 2.4, Lemma 2.6, Corollary 2.7, and Lemma 2.9.

2.1 Mathematical description of artificial neural networks (ANNs)

Setting 2.1. *Let $d, H, \mathfrak{d} \in \mathbb{N}$, $\mathscr{a} \in \mathbb{R}$, $\mathscr{b} \in (\mathscr{a}, \infty)$, $f \in C([\mathscr{a},\mathscr{b}]^d, \mathbb{R})$ satisfy $\mathfrak{d} = dH + 2H + 1$, let $\mathfrak{w} = ((\mathfrak{w}^\theta_{i,j})_{(i,j)\in\{1,\ldots,H\}\times\{1,\ldots,d\}})_{\theta\in\mathbb{R}^{\mathfrak{d}}}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{H\times d}$, $\mathfrak{b} = ((\mathfrak{b}^\theta_1, \ldots, \mathfrak{b}^\theta_H))_{\theta\in\mathbb{R}^{\mathfrak{d}}}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^H$, $\mathfrak{v} = ((\mathfrak{v}^\theta_1, \ldots, \mathfrak{v}^\theta_H))_{\theta\in\mathbb{R}^{\mathfrak{d}}}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^H$, and $\mathfrak{c} = (\mathfrak{c}^\theta)_{\theta\in\mathbb{R}^{\mathfrak{d}}}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \ldots, H\}$, $j \in \{1, 2, \ldots, d\}$ that*

$$\mathfrak{w}^\theta_{i,j} = \theta_{(i-1)d+j}, \qquad \mathfrak{b}^\theta_i = \theta_{Hd+i}, \qquad \mathfrak{v}^\theta_i = \theta_{H(d+1)+i}, \qquad \text{and} \qquad \mathfrak{c}^\theta = \theta_{\mathfrak{d}}, \tag{9}$$

let $\mathfrak{R}_r \in C^1(\mathbb{R}, \mathbb{R})$, $r \in \mathbb{N}$, satisfy for all $x \in \mathbb{R}$ that

$$\limsup_{r\to\infty}\big(|\mathfrak{R}_r(x) - \max\{x, 0\}| + |(\mathfrak{R}_r)'(x) - \mathbb{1}_{(0,\infty)}(x)|\big) = 0 \tag{10}$$

and $\sup_{r\in\mathbb{N}}\sup_{y\in[-|x|,|x|]}|(\mathfrak{R}_r)'(y)| < \infty$, let $\mu\colon \mathcal{B}([\mathscr{a},\mathscr{b}]^d) \to [0,\infty]$ be a finite measure, let $\mathscr{N} = (\mathscr{N}^\theta)_{\theta\in\mathbb{R}^{\mathfrak{d}}}\colon \mathbb{R}^{\mathfrak{d}} \to C(\mathbb{R}^d, \mathbb{R})$ and $\mathcal{L}\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ that

$$\mathscr{N}^\theta(x) = \mathfrak{c}^\theta + \textstyle\sum_{i=1}^H \mathfrak{v}^\theta_i \max\{\mathfrak{b}^\theta_i + \sum_{j=1}^d \mathfrak{w}^\theta_{i,j} x_j, 0\} \tag{11}$$

and $\mathcal{L}(\theta) = \int_{[\mathscr{a},\mathscr{b}]^d}(f(y) - \mathscr{N}^\theta(y))^2\,\mu(\mathrm{d}y)$, let $\mathfrak{L}_r\colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$, $r \in \mathbb{N}$, satisfy for all $r \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathfrak{L}_r(\theta) = \int_{[\mathscr{a},\mathscr{b}]^d}\big(f(y_1, \ldots, y_2) - \mathfrak{c}^\theta - \textstyle\sum_{i=1}^H \mathfrak{v}^\theta_i\big[\mathfrak{R}_r(\mathfrak{b}^\theta_i + \sum_{j=1}^d \mathfrak{w}^\theta_{i,j} y_j)\big]\big)^2\,\mu(\mathrm{d}(y_1, \ldots, y_d)), \tag{12}$$

let $\lambda\colon \mathcal{B}([\mathscr{a},\mathscr{b}]^d) \to [0,\infty]$ be the Lebesgue–Borel measure on $[\mathscr{a},\mathscr{b}]^d$, let $\|\cdot\|\colon \big(\bigcup_{n\in\mathbb{N}}\mathbb{R}^n\big) \to \mathbb{R}$ and $\langle\cdot,\cdot\rangle\colon \big(\bigcup_{n\in\mathbb{N}}(\mathbb{R}^n \times \mathbb{R}^n)\big) \to \mathbb{R}$ satisfy for all $n \in \mathbb{N}$, $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$

that $\|x\| = [\sum_{i=1}^{n} |x_i|^2]^{1/2}$ *and* $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$, *let* $I_i^\theta \subseteq \mathbb{R}^d$, $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \ldots, H\}$, *satisfy for all* $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \ldots, H\}$ *that*

$$I_i^\theta = \big\{ x = (x_1, \ldots, x_d) \in [\mathscr{a}, \mathscr{b}]^d \colon \mathfrak{b}_i^\theta + \textstyle\sum_{j=1}^{d} \mathfrak{w}_{i,j}^\theta x_j > 0 \big\}, \tag{13}$$

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

2.2 An upper bound for the norm of the gradient of the risk function

Proposition 2.2. *Assume Setting 2.1 and let* $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \ldots, H\}$, $j \in \{1, 2, \ldots, d\}$. *Then*

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

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

(iii) it holds that $\limsup_{r \to \infty} \|(\nabla \mathfrak{L}_r)(\theta) - \mathcal{G}(\theta)\| = 0$, *and*

(iv) it holds that

$$\begin{aligned}
\mathcal{G}_{(i-1)d+j}(\theta) &= 2\mathfrak{v}_i^\theta \int_{I_i^\theta} x_j (\mathscr{N}^\theta(x) - f(x)) \, \mu(\mathrm{d}x), \\
\mathcal{G}_{Hd+i}(\theta) &= 2\mathfrak{v}_i^\theta \int_{I_i^\theta} (\mathscr{N}^\theta(x) - f(x)) \, \mu(\mathrm{d}x), \\
\mathcal{G}_{H(d+1)+i}(\theta) &= 2 \int_{[\mathscr{a}, \mathscr{b}]^d} \big[\max\big\{ \mathfrak{b}_i^\theta + \textstyle\sum_{k=1}^{d} \mathfrak{w}_{i,k}^\theta x_k, 0 \big\} \big] (\mathscr{N}^\theta(x) - f(x)) \, \mu(\mathrm{d}x), \\
\text{and} \qquad \mathcal{G}_{\mathfrak{d}}(\theta) &= 2 \int_{[\mathscr{a}, \mathscr{b}]^d} (\mathscr{N}^\theta(x) - f(x)) \, \mu(\mathrm{d}x).
\end{aligned} \tag{14}$$

Proof of Proposition 2.2. Throughout this proof we assume without loss of generality that $\mu([\mathscr{a}, \mathscr{b}]^d) > 0$. Observe that [18, Proposition 2.3] (applied with $a \curvearrowleft \mathscr{a}$, $b \curvearrowleft \mathscr{b}$, $\mu \curvearrowleft (\mathcal{B}([\mathscr{a}, \mathscr{b}]^d) \ni A \mapsto \mu(A)[\mu([\mathscr{a}, \mathscr{b}]^d)]^{-1} \in [0, 1])$ in the notation of [18, Proposition 2.3]) establishes items (i), (ii), (iii), and (iv). The proof of Proposition 2.2 is thus complete. □

Lemma 2.3. *Assume Setting 2.1 and let* $\mathbf{a} \in \mathbb{R}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ *satisfy* $\mathbf{a} = \max\{|\mathscr{a}|, |\mathscr{b}|, 1\}$. *Then*

$$\|\mathcal{G}(\theta)\|^2 \leq 4\mathcal{L}(\theta) \big(\mathbf{a}^2 (d+1) \|\theta\|^2 + 1 \big) \mu([\mathscr{a}, \mathscr{b}]^d). \tag{15}$$

Proof of Lemma 2.3. Throughout this proof assume without loss of generality that $\mu([\mathscr{a}, \mathscr{b}]^d) > 0$. Note that Proposition 2.2, [18, Proposition 2.3] (applied with $a \curvearrowleft \mathscr{a}$, $b \curvearrowleft \mathscr{b}$, $\mu \curvearrowleft (\mathcal{B}([\mathscr{a}, \mathscr{b}]^d) \ni A \mapsto \mu(A)[\mu([\mathscr{a}, \mathscr{b}]^d)]^{-1} \in [0, 1])$ in the notation of [18, Proposition 2.3]), and [18, Lemma 2.5] (applied with $a \curvearrowleft \mathscr{a}$, $b \curvearrowleft \mathscr{b}$, $\mu \curvearrowleft (\mathcal{B}([\mathscr{a}, \mathscr{b}]^d) \ni A \mapsto \mu(A)[\mu([\mathscr{a}, \mathscr{b}]^d)]^{-1} \in [0, 1])$ in the notation of [18, Lemma 2.5]) establish (15). The proof of Lemma 2.3 is thus complete. □

Corollary 2.4. *Assume Setting 2.1. Then it holds that* $\mathcal{G}$ *is locally bounded and measurable.*

Proof of Corollary 2.4. Observe that item (ii) in Proposition 2.2 ensures that for all $r \in \mathbb{N}$ it holds that $\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto (\nabla \mathfrak{L}_r)(\theta) \in \mathbb{R}^{\mathfrak{d}}$ is measurable. Combining this with item (iii) in Proposition 2.2 demonstrates that $\mathcal{G}$ is measurable. Moreover, note that Lemma 2.9 and Lemma 2.3 assure that $\mathcal{G}$ is locally bounded. This completes the proof of Corollary 2.4. □

2.3 Continuous dependence of active neuron regions on ANN parameters

Lemma 2.5. *Let $d \in \mathbb{N}$, $\mathscr{a} \in \mathbb{R}$, $\mathscr{b} \in (\mathscr{a}, \infty)$, let $I^u \subseteq [\mathscr{a}, \mathscr{b}]^d$, $u \in \mathbb{R}^{d+1}$, satisfy for all $u = (u_1, \ldots, u_{d+1}) \in \mathbb{R}^{d+1}$ that $I^u = \{x = (x_1, \ldots, x_d) \in [\mathscr{a}, \mathscr{b}]^d \colon u_{d+1} + \sum_{i=1}^d u_i x_i > 0\}$, for every $n \in \mathbb{N}$ let $\lambda_n \colon \mathcal{B}(\mathbb{R}^n) \to [0, \infty]$ be the Lebesgue–Borel measure on $\mathbb{R}^n$, and let $v \in \mathbb{R}^{d+1} \backslash \{0\}$. Then*

$$\limsup_{\mathbb{R}^{d+1} \ni u \to v} \lambda_d(I^u \Delta I^v) = 0. \tag{16}$$

Proof of Lemma 2.5. Throughout this proof let $\|\cdot\| \colon \left(\bigcup_{n \in \mathbb{N}} \mathbb{R}^n\right) \to \mathbb{R}$ satisfy for all $n \in \mathbb{N}$, $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ that $\|x\| = [\sum_{i=1}^n |x_i|^2]^{1/2}$. Observe that the fact that for all $y \in \mathbb{R}$ it holds that $y \geq -|y|$ ensures that for all $u = (u_1, \ldots, u_{d+1}) \in \mathbb{R}^{d+1}$, $i \in \{1, 2, \ldots, d+1\}$ with $\|u - v\| < |v_i|$ it holds that

$$u_i v_i = (v_i)^2 + (u_i - v_i) v_i \geq |v_i|^2 - |u_i - v_i||v_i| \geq |v_i|^2 - \|u - v\||v_i| > 0. \tag{17}$$

In the following we distinguish between the case $\max_{i \in \{1,2,\ldots,d\}} |v_i| = 0$, the case $(\max_{i \in \{1,2,\ldots,d\}} |v_i|, d) \in (0, \infty) \times [2, \infty)$, and the case $(\max_{i \in \{1,2,\ldots,d\}} |v_i|, d) \in (0, \infty) \times \{1\}$. We first prove (16) in the case

$$\max_{i \in \{1,2,\ldots,d\}} |v_i| = 0. \tag{18}$$

Note that (18) and the assumption that $v \in \mathbb{R}^{d+1} \backslash \{0\}$ imply that $v_{d+1} \neq 0$. Moreover, observe that (18) shows that for all $u = (u_1, \ldots, u_{d+1}) \in \mathbb{R}^{d+1}$, $x \in I^u \Delta I^v$ we have that

$$\begin{aligned} &\left|\left(\left[\textstyle\sum_{i=1}^d u_i x_i\right] + u_{d+1}\right) - \left(\left[\textstyle\sum_{i=1}^d v_i x_i\right] + v_{d+1}\right)\right| \\ &= \left|\left[\textstyle\sum_{i=1}^d u_i x_i\right] + u_{d+1}\right| + \left|\left[\textstyle\sum_{i=1}^d v_i x_i\right] + v_{d+1}\right| \geq \left|\left[\textstyle\sum_{i=1}^d v_i x_i\right] + v_{d+1}\right| = |v_{d+1}|. \end{aligned} \tag{19}$$

In addition, note that for all $u = (u_1, \ldots, u_{d+1}) \in \mathbb{R}^{d+1}$, $x \in [\mathscr{a}, \mathscr{b}]^d$ it holds that

$$\begin{aligned} &\left|\left(\left[\textstyle\sum_{i=1}^d u_i x_i\right] + u_{d+1}\right) - \left(\left[\textstyle\sum_{i=1}^d v_i x_i\right] + v_{d+1}\right)\right| \leq \left[\textstyle\sum_{i=1}^d |u_i - v_i||x_i|\right] + |u_{d+1} - v_{d+1}| \\ &\leq \max\{|\mathscr{a}|, |\mathscr{b}|\}\left[\textstyle\sum_{i=1}^d |u_i - v_i|\right] + |u_{d+1} - v_{d+1}| \leq (1 + d \max\{|\mathscr{a}, \mathscr{b}|\})\|u - v\|. \end{aligned} \tag{20}$$

Combining this with (19) shows that for all $u \in \mathbb{R}^{d+1}$ with $\|u - v\| < \frac{|v_{d+1}|}{1 + d \max\{|\mathscr{a}, \mathscr{b}|\}}$ it holds that $I^u \Delta I^v = \varnothing$. Hence, we obtain that $\limsup_{\mathbb{R}^{d+1} \ni u \to v} \lambda_d(I^u \Delta I^v) = 0$. This establishes (16) in the case $\max_{i \in \{1,2,\ldots,d\}} |v_i| = 0$. In the next step we prove (16) in the case

$$(\max_{i \in \{1,2,\ldots,d\}} |v_i|, d) \in (0, \infty) \times [2, \infty). \tag{21}$$

For this we assume without loss of generality that $v_1 \neq 0$. In the following let $J_x^{u,w} \subseteq \mathbb{R}$, $x \in [\mathscr{a}, \mathscr{b}]^{d-1}$, $u, w \in \mathbb{R}^{d+1}$, satisfy for all $x = (x_2, \ldots, x_d) \in [\mathscr{a}, \mathscr{b}]^{d-1}$, $u, w \in \mathbb{R}^{d+1}$ that $J_x^{u,w} = \{y \in [\mathscr{a}, \mathscr{b}] \colon (y, x_2, \ldots, x_d) \in I^u \backslash I^w\}$. Next observe that Fubini's theorem and the fact that for all $u \in \mathbb{R}^{d+1}$ it holds that I^u is measurable show that for all $u \in \mathbb{R}^{d+1}$ we have that

$$\begin{aligned} \lambda_d(I^u \Delta I^v) &= \int_{[\mathscr{a}, \mathscr{b}]^d} \mathbb{1}_{I^u \Delta I^v}(x) \, \lambda_d(\mathrm{d}x) = \int_{[\mathscr{a}, \mathscr{b}]^d} \left(\mathbb{1}_{I^u \backslash I^v}(x) + \mathbb{1}_{I^v \backslash I^u}(x)\right) \lambda_d(\mathrm{d}x) \\ &= \int_{[\mathscr{a}, \mathscr{b}]^{d-1}} \int_{[\mathscr{a}, \mathscr{b}]} \left(\mathbb{1}_{I^u \backslash I^v}(y, x_2, \ldots, x_d) + \mathbb{1}_{I^v \backslash I^u}(y, x_2, \ldots, x_d)\right) \lambda_1(\mathrm{d}y) \, \lambda_{d-1}(\mathrm{d}(x_2, \ldots, x_d)) \\ &= \int_{[\mathscr{a}, \mathscr{b}]^{d-1}} \int_{[\mathscr{a}, \mathscr{b}]} \left(\mathbb{1}_{J_x^{u,v}}(y) + \mathbb{1}_{J_x^{v,u}}(y)\right) \lambda_1(\mathrm{d}y) \, \lambda_{d-1}(\mathrm{d}x) \\ &= \int_{[\mathscr{a}, \mathscr{b}]^{d-1}} (\lambda_1(J_x^{u,v}) + \lambda_1(J_x^{v,u})) \, \lambda_{d-1}(\mathrm{d}x). \end{aligned} \tag{22}$$

Moreover, note that for all $x = (x_2, \ldots, x_d) \in [\mathscr{a}, \mathscr{b}]^{d-1}$, $u = (u_1, \ldots, u_{d+1})$, $w = (w_1, \ldots, w_{d+1}) \in \mathbb{R}^{d+1}$, $\mathfrak{s} \in \{-1, 1\}$ with $\min\{\mathfrak{s}u_1, \mathfrak{s}w_1\} > 0$ it holds that

$$\begin{aligned} J_x^{u,w} &= \{y \in [\mathscr{a}, \mathscr{b}] \colon (y, x_2, \ldots, x_d) \in I^u \backslash I^w\} \\ &= \Big\{ y \in [\mathscr{a}, \mathscr{b}] \colon u_1 y + \big[\textstyle\sum_{i=2}^{d} u_i x_i\big] + u_{d+1} > 0 \geq w_1 y + \big[\sum_{i=2}^{d} w_i x_i\big] + w_{d+1} \Big\} \\ &= \Big\{ y \in [\mathscr{a}, \mathscr{b}] \colon -\tfrac{\mathfrak{s}}{u_1}\big(\big[\textstyle\sum_{i=2}^{d} u_i x_i\big] + u_{d+1}\big) < \mathfrak{s} y \leq -\tfrac{\mathfrak{s}}{w_1}\big(\big[\sum_{i=2}^{d} w_i x_i\big] + w_{d+1}\big) \Big\}. \end{aligned} \tag{23}$$

Hence, we obtain for all $x = (x_2, \ldots, x_d) \in [\mathscr{a}, \mathscr{b}]^{d-1}$, $u = (u_1, \ldots, u_{d+1})$, $w = (w_1, \ldots, w_{d+1}) \in \mathbb{R}^{d+1}$, $\mathfrak{s} \in \{-1, 1\}$ with $\min\{\mathfrak{s}u_1, \mathfrak{s}w_1\} > 0$ that

$$\begin{aligned} \lambda_1(J_x^{u,w}) &\leq \Big| \tfrac{\mathfrak{s}}{u_1}\big(\big[\textstyle\sum_{i=2}^{d} u_i x_i\big] + u_{d+1}\big) - \tfrac{\mathfrak{s}}{w_1}\big(\big[\sum_{i=2}^{d} w_i x_i\big] + w_{d+1}\big) \Big| \\ &\leq \Big[\textstyle\sum_{i=2}^{d} \big| \frac{u_i}{u_1} - \frac{w_i}{w_1} \big| |x_i| \Big] + \Big| \frac{u_{d+1}}{u_1} - \frac{w_{d+1}}{w_1} \Big| \\ &\leq \max\{|\mathscr{a}|, |\mathscr{b}|\} \Big[\textstyle\sum_{i=2}^{d} \big| \frac{u_i}{u_1} - \frac{w_i}{w_1} \big| \Big] + \Big| \frac{u_{d+1}}{u_1} - \frac{w_{d+1}}{w_1} \Big|. \end{aligned} \tag{24}$$

Furthermore, observe that (17) demonstrates for all $u = (u_1, \ldots, u_{d+1}) \in \mathbb{R}^{d+1}$ with $\|u - v\| < \frac{|v_1|}{2}$ that $u_1 v_1 > 0$. This implies that for all $u = (u_1, \ldots, u_{d+1}) \in \mathbb{R}^{d+1}$ with $\|u - v\| < \frac{|v_1|}{2}$ there exists $\mathfrak{s} \in \{-1, 1\}$ such that $\min\{\mathfrak{s}u_1, \mathfrak{s}v_1\} > 0$. Combining this with (24) proves that there exists $\mathfrak{C} \in \mathbb{R}$ such that for all $x \in [\mathscr{a}, \mathscr{b}]^{d-1}$, $u \in \mathbb{R}^{d+1}$ with $\|u - v\| < \frac{|v_1|}{2}$ we have that $\lambda_1(J_x^{u,v}) + \lambda_1(J_x^{v,u}) \leq \mathfrak{C}\|u - v\|$. This and (22) establish (16) in the case $(\max_{i \in \{1,2,\ldots,d\}} |v_i|, d) \in (0, \infty) \times [2, \infty)$. Finally, we prove (16) in the case

$$(\max_{i \in \{1,2,\ldots,d\}} |v_i|, d) \in (0, \infty) \times \{1\}. \tag{25}$$

Note that (25) assures that $|v_1| > 0$. In addition, observe that for all $u = (u_1, u_2)$, $w = (w_1, w_2) \in \mathbb{R}^2$, $\mathfrak{s} \in \{-1, 1\}$ with $\min\{\mathfrak{s}u_1, \mathfrak{s}w_1\} > 0$ it holds that

$$\begin{aligned} I^w \backslash I^u &= \{y \in [\mathscr{a}, \mathscr{b}] \colon w_1 y + w_2 > 0 \geq u_1 y + u_2\} = \Big\{ y \in [\mathscr{a}, \mathscr{b}] \colon -\tfrac{\mathfrak{s}w_2}{w_1} < \mathfrak{s} y \leq -\tfrac{\mathfrak{s}u_2}{u_1} \Big\} \\ &\subseteq \Big\{ y \in \mathbb{R} \colon -\tfrac{\mathfrak{s}w_2}{w_1} < \mathfrak{s} y \leq -\tfrac{\mathfrak{s}u_2}{u_1} \Big\}. \end{aligned} \tag{26}$$

Hence, we obtain for all $u = (u_1, u_2)$, $w = (w_1, w_2) \in \mathbb{R}^2$, $\mathfrak{s} \in \{-1, 1\}$ with $\min\{\mathfrak{s}u_1, \mathfrak{s}w_1\} > 0$ that

$$\lambda_1(I^w \backslash I^u) \leq \Big| \Big(-\tfrac{\mathfrak{s}u_2}{u_1}\Big) - \Big(-\tfrac{\mathfrak{s}w_2}{w_1}\Big) \Big| = \Big| \tfrac{u_2}{u_1} - \tfrac{w_2}{w_1} \Big|. \tag{27}$$

Furthermore, note that (17) ensures for all $u = (u_1, u_2) \in \mathbb{R}^2$ with $\|u - v\| < |v_1|$ that $u_1 v_1 > 0$. This proves that for all $u = (u_1, u_2) \in \mathbb{R}^2$ with $\|u - v\| < |v_1|$ there exists $\mathfrak{s} \in \{-1, 1\}$ such that $\min\{\mathfrak{s}u_1, \mathfrak{s}v_1\} > 0$. Combining this with (27) demonstrates for all $u = (u_1, u_2) \in \mathbb{R}^2$ with $\|u - v\| < |v_1|$ that

$$\lambda_1(I^u \Delta I^v) = \lambda_1(I^u \backslash I^v) + \lambda_1(I^v \backslash I^u) \leq 2 \Big| \tfrac{u_2}{u_1} - \tfrac{v_2}{v_1} \Big|. \tag{28}$$

Hence, we obtain that

$$\limsup_{\mathbb{R}^2 \ni u \to v} \lambda_1(I^v \Delta I^u) = 0. \tag{29}$$

This establishes (16) in the case $(\max_{i \in \{1,2,\ldots,d\}} |v_i|, d) \in (0, \infty) \times \{1\}$. The proof of Lemma 2.5 is thus complete. □

Lemma 2.6. *Let $(E, \mathcal{E})$ be a measurable space, let $\mu \colon \mathcal{E} \to [0, \infty]$ and $\nu \colon \mathcal{E} \to [0, \infty]$ be measures, assume $\mu \ll \nu$ and $\mu(E) < \infty$, and let $\varepsilon \in (0, \infty)$. Then there exists $\delta \in (0, \infty)$ such that for all $A \in \mathcal{E}$ with $\nu(A) < \delta$ it holds that $\mu(A) < \varepsilon$.*

Proof of Lemma 2.6. Throughout this proof assume for the sake of contradiction that there exists $A = (A_n)_{n\in\mathbb{N}} \colon \mathbb{N} \to \mathcal{E}$ which satisfies for all $n \in \mathbb{N}$ that $\nu(A_n) < 2^{-n}$ and $\mu(A_n) \geq \varepsilon$ and let $B_n \in \mathcal{E}$, $n \in \mathbb{N}$, and $C \in \mathcal{E}$ satisfy for all $n \in \mathbb{N}$ that $B_n = \bigcup_{k=n}^{\infty} A_k$ and $C = \bigcap_{k=1}^{\infty} B_k$. Observe that the fact that for all $n \in \mathbb{N}$ it holds that $\nu(A_n) < 2^{-n}$ ensures that for all $n \in \mathbb{N}$ we have that

$$\nu(B_n) = \nu\big(\textstyle\bigcup_{k=n}^{\infty} A_k\big) \leq \sum_{k=n}^{\infty} \nu(A_k) \leq \sum_{k=n}^{\infty} 2^{-k} = 2^{-n}\big(\textstyle\sum_{k=0}^{\infty} 2^{-k}\big) = 2^{1-n}. \tag{30}$$

This implies that

$$\nu(C) = \nu\big(\textstyle\bigcap_{k=1}^{\infty} B_k\big) \leq \inf_{k\in\mathbb{N}} \nu(B_k) \leq \inf_{k\in\mathbb{N}}(2^{1-k}) = 0. \tag{31}$$

The assumption that $\mu \ll \nu$ hence shows that

$$\mu(C) = 0. \tag{32}$$

Moreover, note that the fact that for all $n \in \mathbb{N}$ it holds that $\mu(A_n) \geq \varepsilon$ proves that for all $n \in \mathbb{N}$ we have that $\mu(B_n) = \mu\big(\bigcup_{k=n}^{\infty} A_k\big) \geq \varepsilon$. Combining this and (32) with the fact that for all $n \in \mathbb{N}$ it holds that $B_n \supseteq B_{n+1}$ and the fact that $\mu(B_1) \leq \mu(E) < \infty$ demonstrates that

$$0 = \mu(C) = \mu\big(\textstyle\bigcap_{k=1}^{\infty} B_k\big) = \lim_{k\to\infty} \mu(B_k) \geq \varepsilon > 0. \tag{33}$$

This is a contradiction. The proof of Lemma 2.6 is thus complete. □

Corollary 2.7. *Let $(E, \mathcal{E})$ be a measurable space, let $\mu \colon \mathcal{E} \to [0,\infty]$ and $\nu \colon \mathcal{E} \to [0,\infty]$ be measures, assume $\mu \ll \nu$ and $\mu(E) < \infty$, and let $A_n \in \mathcal{E}$, $n \in \mathbb{N}$, satisfy $\limsup_{n\to\infty} \nu(A_n) = 0$. Then $\limsup_{n\to\infty} \mu(A_n) = 0$.*

Proof of Corollary 2.7. Throughout this proof let $\varepsilon \in (0,\infty)$. Observe that Lemma 2.6 proves that there exists $\delta \in (0,\infty)$ such that for all $B \in \mathcal{E}$ with $\nu(B) < \delta$ it holds that $\mu(B) < \varepsilon$. Furthermore, note that the assumption that $\limsup_{n\to\infty} \nu(A_n) = 0$ ensures that there exists $N \in \mathbb{N}$ such that for all $n \in \mathbb{N} \cap [N,\infty)$ it holds that $\nu(A_n) < \delta$. Hence, we obtain for all $n \in \mathbb{N} \cap [N,\infty)$ that $\mu(A_n) < \varepsilon$. The proof of Corollary 2.7 is thus complete. □

Corollary 2.8. *Assume Setting 2.1, let $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \ldots, H\}$ satisfy $|\mathfrak{b}_i^{\theta}| + \sum_{j=1}^{d} |\mathfrak{w}_{i,j}^{\theta}| > 0$, and assume $\mu \ll \lambda$. Then $\limsup_{\mathbb{R}^{\mathfrak{d}} \ni \vartheta \to \theta} \mu(I_i^{\theta} \Delta I_i^{\vartheta}) = 0$.*

Proof of Corollary 2.8. Throughout this proof let $\vartheta = (\vartheta_n)_{n\in\mathbb{N}} \colon \mathbb{N} \to \mathbb{R}^{\mathfrak{d}}$ satisfy $\limsup_{n\to\infty} \|\vartheta_n - \theta\| = 0$. Observe that Lemma 2.5 and the assumption that $|\mathfrak{b}_i^{\theta}| + \sum_{j=1}^{d} |\mathfrak{w}_{i,j}^{\theta}| > 0$ establish that $\limsup_{n\to\infty} \lambda(I_i^{\theta} \Delta I_i^{\vartheta_n}) = 0$. Combining this, the assumption that $\mu \ll \lambda$, the fact that $\mu([a,b]^d) < \infty$, and Corollary 2.7 implies that $\limsup_{n\to\infty} \mu(I_i^{\theta} \Delta I_i^{\vartheta_n}) = 0$. The proof of Corollary 2.8 is thus complete. □

2.4 Differentiability of the risk function

Lemma 2.9. *Let $d, H, \mathfrak{d} \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in (a,\infty)$, $f \in C([a,b]^d, \mathbb{R})$ satisfy $\mathfrak{d} = dH + 2H + 1$, let $\mathcal{N} = (\mathcal{N}^{\theta})_{\theta\in\mathbb{R}^{\mathfrak{d}}} \colon \mathbb{R}^{\mathfrak{d}} \to C(\mathbb{R}^d, \mathbb{R})$ satisfy for all $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ that*

$$\mathcal{N}^{\theta}(x) = \theta_{\mathfrak{d}} + \textstyle\sum_{i=1}^{H} \theta_{H(d+1)+i} \max\{\theta_{Hd+i} + \sum_{j=1}^{d} \theta_{(i-1)d+j} x_j, 0\}, \tag{34}$$

let $\mu \colon \mathcal{B}([a,b]^d) \to [0,\infty]$ be a finite measure, let $\|\cdot\| \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ and $\mathcal{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $\theta = (\theta_1, \ldots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ that $\|\theta\| = [\sum_{i=1}^{\mathfrak{d}} |\theta_i|^2]^{1/2}$ and $\mathcal{L}(\theta) = \int_{[a,b]^d} (\mathcal{N}^{\theta}(x) - f(x))^2 \, \mu(\mathrm{d}x)$, and let $K \subseteq \mathbb{R}^{\mathfrak{d}}$ be compact. Then there exists $\mathscr{L} \in \mathbb{R}$ such that for all $\theta, \vartheta \in K$ it holds that

$$\big(\sup_{x\in[a,b]^d} |\mathcal{N}^{\theta}(x) - \mathcal{N}^{\vartheta}(x)|\big) + |\mathcal{L}(\theta) - \mathcal{L}(\vartheta)| \leq \mathscr{L} \|\theta - \vartheta\|. \tag{35}$$

Proof of Lemma 2.9. Throughout this proof we distinguish between the case $\mu([\mathscr{a},\mathscr{b}]^d) = 0$ and the case $\mu([\mathscr{a},\mathscr{b}]^d) > 0$. We first prove (35) in the case

$$\mu([\mathscr{a},\mathscr{b}]^d) = 0. \tag{36}$$

Note that (36) ensures that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that $\mathcal{L}(\theta) = 0$. Furthermore, observe that [18, Lemma 2.4] (applied with $a \curvearrowleft \mathscr{a}$, $b \curvearrowleft \mathscr{b}$, $\mu \curvearrowleft (\mathcal{B}([\mathscr{a},\mathscr{b}]^d) \ni A \mapsto \mathbb{1}_A(a, a, \ldots, a) \in [0,1])$ in the notation of [18, Lemma 2.4]) proves that there exists $\mathscr{L} \in \mathbb{R}$ such that for all $\theta, \vartheta \in K$ it holds that $(\sup_{x\in[\mathscr{a},\mathscr{b}]^d} |\mathcal{N}^\theta(x) - \mathcal{N}^\vartheta(x)|) \leq \mathscr{L}\|\theta - \vartheta\|$. This establishes (35) in the case $\mu([\mathscr{a},\mathscr{b}]^d) = 0$. In the next step we prove (35) in the case $\mu([\mathscr{a},\mathscr{b}]^d) > 0$. Note that [18, Lemma 2.4] (applied with $a \curvearrowleft \mathscr{a}$, $b \curvearrowleft \mathscr{b}$, $\mu \curvearrowleft (\mathcal{B}([\mathscr{a},\mathscr{b}]^d) \ni A \mapsto \mu(A)[\mu([\mathscr{a},\mathscr{b}]^d)]^{-1} \in [0,1])$ in the notation of [18, Lemma 2.4]) establishes (35) in the case $\mu([\mathscr{a},\mathscr{b}]^d) > 0$. The proof of Lemma 2.9 is thus complete. □

Lemma 2.10. *Let $d \in \mathbb{N}$, $w_1, w_2 \in \mathbb{R}^d$, $b_1, b_2, \mathscr{a} \in \mathbb{R}$, $\mathscr{b} \in (\mathscr{a}, \infty)$, let $\|\cdot\| \colon \mathbb{R}^d \to \mathbb{R}$ and $\langle \cdot, \cdot \rangle \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ satisfy for all $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d) \in \mathbb{R}^d$ that $\|x\| = [\sum_{i=1}^d |x_i|^2]^{1/2}$ and $\langle x, y \rangle = \sum_{i=1}^d x_i y_i$, let $I_k \subseteq [\mathscr{a},\mathscr{b}]^d$, $k \in \{1,2\}$, satisfy for all $k \in \{1,2\}$ that $I_k = \{x \in [\mathscr{a},\mathscr{b}]^d \colon \langle w_k, x\rangle + b_k > 0\}$, and let $x \in I_1 \Delta I_2$. Then*

$$\max_{k\in\{1,2\}} |\langle w_k, x\rangle + b_k| \leq |\langle w_1 - w_2, x\rangle + b_1 - b_2| \leq \max\{|\mathscr{a}|, |\mathscr{b}|\}\sqrt{d}\|w_1 - w_2\| + |b_1 - b_2|. \tag{37}$$

Proof of Lemma 2.10. Throughout this proof assume without loss of generality that $x \in I_1 \backslash I_2$. Observe that the fact that $\langle w_2, x\rangle + b_2 \leq 0 < \langle w_1, x\rangle + b_1$ demonstrates that

$$\langle w_2 - w_1, x\rangle + b_2 - b_1 < \langle w_2, x\rangle + b_2 \leq 0 < \langle w_1, x\rangle + b_1 \leq \langle w_1 - w_2, x\rangle + b_1 - b_2. \tag{38}$$

Hence, we obtain that $\max_{k\in\{1,2\}} |\langle w_k, x\rangle + b_k| \leq |\langle w_1 - w_2, x\rangle + b_1 - b_2|$. Furthermore, note that the Cauchy-Schwarz inequality and the fact that $x \in [\mathscr{a},\mathscr{b}]^d$ assure that

$$|\langle w_1 - w_2, x\rangle + b_1 - b_2| \leq \|x\|\|w_1 - w_2\| + |b_1 - b_2| \leq \max\{|\mathscr{a}|, |\mathscr{b}|\}\sqrt{d}\|w_2 - w_1\| + |b_2 - b_1|. \tag{39}$$

This completes the proof of Lemma 2.10. □

Proposition 2.11. *Assume Setting 2.1, assume $\mu \ll \lambda$, and let $\theta \in \mathbb{R}^{\mathfrak{d}}$ satisfy*

$$\mathcal{L}(\theta)\big(\textstyle\sum_{i=1}^H |\mathfrak{v}_i^\theta| \mathbb{1}_{\{0\}}\big(|\mathfrak{b}_i^\theta| + \sum_{j=1}^d |\mathfrak{w}_{i,j}^\theta|\big)\big) = 0. \tag{40}$$

Then

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

(ii) it holds that $(\nabla\mathcal{L})(\theta) = \mathcal{G}(\theta)$.

Proof of Proposition 2.11. Throughout this proof let $M \in \mathbb{R}$ satisfy

$$M = \inf\big\{m \in \mathbb{R} \colon \mu\big(\{x \in [\mathscr{a},\mathscr{b}]^d \colon |\mathcal{N}^\theta(x) - f(x)| > m\}\big) = 0\big\} \tag{41}$$

and let $\mathfrak{C} \in \mathbb{R}$ satisfy

$$\mathfrak{C} = 1 + d\max\{|\mathscr{a}|, |\mathscr{b}|\}. \tag{42}$$

We will prove items (i) and (ii) by showing that

$$\limsup_{\mathbb{R}^{\mathfrak{d}}\backslash\{0\}\ni h\to 0} \big[\|h\|^{-1}|\mathcal{L}(\theta + h) - \mathcal{L}(\theta) - \langle \mathcal{G}(\theta), h\rangle|\big] = 0. \tag{43}$$

Observe that Proposition 2.2 ensures that for all $h \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned}\langle \mathcal{G}(\theta), h\rangle &= 2\left[\sum_{i=1}^{H}\int_{I_i^\theta}\left(\mathfrak{b}_i^h + \textstyle\sum_{j=1}^{d}\mathfrak{w}_{i,j}^h x_j\right)\mathfrak{v}_i^\theta(\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x)\right]\\ &\quad + 2\left[\sum_{i=1}^{H}\mathfrak{v}_i^h\int_{[\mathfrak{a},\mathfrak{b}]^d}\max\left\{\mathfrak{b}_i^\theta + \textstyle\sum_{j=1}^{d}\mathfrak{w}_{i,j}^\theta x_j, 0\right\}(\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x)\right]\\ &\quad + 2\mathfrak{c}^h\int_{[\mathfrak{a},\mathfrak{b}]^d}(\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x).\end{aligned} \tag{44}$$

Combining this and the fact that for all $x, y, \mathfrak{z} \in \mathbb{R}$ it holds that

$$\begin{aligned}(x-\mathfrak{z})^2 - (y-\mathfrak{z})^2 &= (x-y)(x+y-2\mathfrak{z}) = (x-y)((x-y)+2(y-\mathfrak{z}))\\ &= (x-y)^2 + 2(x-y)(y-\mathfrak{z})\end{aligned} \tag{45}$$

demonstrates that for all $h \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned}&\mathcal{L}(\theta+h) - \mathcal{L}(\theta) - \langle \mathcal{G}(\theta), h\rangle\\ &= \int_{[\mathfrak{a},\mathfrak{b}]^d}(\mathcal{N}^{\theta+h}(x) - \mathcal{N}^\theta(x))^2\,\mu(\mathrm{d}x)\\ &\quad + 2\int_{[\mathfrak{a},\mathfrak{b}]^d}(\mathcal{N}^{\theta+h}(x) - \mathcal{N}^\theta(x))(\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x) - \langle \mathcal{G}(\theta), h\rangle\\ &= \int_{[\mathfrak{a},\mathfrak{b}]^d}(\mathcal{N}^{\theta+h}(x) - \mathcal{N}^\theta(x))^2\,\mu(\mathrm{d}x)\\ &\quad + 2\int_{[\mathfrak{a},\mathfrak{b}]^d}\Big(\mathfrak{c}^h + \textstyle\sum_{i=1}^{H}\big[(\mathfrak{v}_i^\theta + \mathfrak{v}_i^h)\max\{\mathfrak{b}_i^\theta + \mathfrak{b}_i^h + \sum_{j=1}^{d}(\mathfrak{w}_{i,j}^\theta + \mathfrak{w}_{i,j}^h)x_j, 0\}\\ &\qquad - \mathfrak{v}_i^\theta\max\{\mathfrak{b}_i^\theta + \textstyle\sum_{j=1}^{d}\mathfrak{w}_{i,j}^\theta x_j, 0\}\big]\Big)(\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x)\\ &\quad - 2\left[\sum_{i=1}^{H}\int_{I_i^\theta}\left(\mathfrak{b}_i^h + \textstyle\sum_{j=1}^{d}\mathfrak{w}_{i,j}^h x_j\right)\mathfrak{v}_i^\theta(\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x)\right]\\ &\quad - 2\left[\sum_{i=1}^{H}\mathfrak{v}_i^h\int_{[\mathfrak{a},\mathfrak{b}]^d}\max\left\{\mathfrak{b}_i^\theta + \textstyle\sum_{j=1}^{d}\mathfrak{w}_{i,j}^\theta x_j, 0\right\}(\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x)\right]\\ &\quad - 2\mathfrak{c}^h\int_{[\mathfrak{a},\mathfrak{b}]^d}(\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x).\end{aligned} \tag{46}$$

This shows for all $h \in \mathbb{R}^{\mathfrak{d}}$ that

$$\begin{aligned}&\mathcal{L}(\theta+h) - \mathcal{L}(\theta) - \langle \mathcal{G}(\theta), h\rangle = \int_{[\mathfrak{a},\mathfrak{b}]^d}(\mathcal{N}^{\theta+h}(x) - \mathcal{N}^\theta(x))^2\,\mu(\mathrm{d}x)\\ &+ 2\left[\sum_{i=1}^{H}\int_{[\mathfrak{a},\mathfrak{b}]^d}(\mathfrak{v}_i^\theta + \mathfrak{v}_i^h)\big(\mathfrak{b}_i^\theta + \mathfrak{b}_i^h + \textstyle\sum_{j=1}^{d}(\mathfrak{w}_{i,j}^\theta + \mathfrak{w}_{i,j}^h)x_j\big)(\mathcal{N}^\theta(x) - f(x))\mathbb{1}_{I_i^{\theta+h}}(x)\,\mu(\mathrm{d}x)\right]\\ &- 2\left[\sum_{i=1}^{H}\int_{[\mathfrak{a},\mathfrak{b}]^d}\mathfrak{v}_i^\theta\big(\mathfrak{b}_i^\theta + \textstyle\sum_{j=1}^{d}\mathfrak{w}_{i,j}^\theta x_j\big)(\mathcal{N}^\theta(x) - f(x))\mathbb{1}_{I_i^\theta}(x)\,\mu(\mathrm{d}x)\right]\\ &- 2\left[\sum_{i=1}^{H}\int_{[\mathfrak{a},\mathfrak{b}]^d}\mathfrak{v}_i^\theta\big(\mathfrak{b}_i^h + \textstyle\sum_{j=1}^{d}\mathfrak{w}_{i,j}^h x_j\big)(\mathcal{N}^\theta(x) - f(x))\mathbb{1}_{I_i^\theta}(x)\,\mu(\mathrm{d}x)\right]\\ &- 2\left[\sum_{i=1}^{H}\int_{[\mathfrak{a},\mathfrak{b}]^d}\mathfrak{v}_i^h\big(\mathfrak{b}_i^\theta + \textstyle\sum_{j=1}^{d}\mathfrak{w}_{i,j}^\theta x_j\big)(\mathcal{N}^\theta(x) - f(x))\mathbb{1}_{I_i^\theta}(x)\,\mu(\mathrm{d}x)\right].\end{aligned} \tag{47}$$

Hence, we obtain for all $h \in \mathbb{R}^{\mathfrak{d}}$ that

$$\begin{aligned}
&\mathcal{L}(\theta+h) - \mathcal{L}(\theta) - \langle \mathcal{G}(\theta), h\rangle = \int_{[\mathscr{a},\mathscr{b}]^d} (\mathcal{N}^{\theta+h}(x) - \mathcal{N}^{\theta}(x))^2 \, \mu(\mathrm{d}x) \\
&+ 2\left[\sum_{i=1}^{H} \int_{[\mathscr{a},\mathscr{b}]^d} \mathfrak{v}_i^h \big(\mathfrak{b}_i^h + \textstyle\sum_{j=1}^{d} \mathfrak{w}_{i,j}^h x_j\big)(\mathcal{N}^{\theta}(x) - f(x)) \mathbb{1}_{I_i^{\theta+h}}(x) \, \mu(\mathrm{d}x)\right] \\
&+ 2\left[\sum_{i=1}^{H} \int_{[\mathscr{a},\mathscr{b}]^d} \mathfrak{v}_i^\theta \big(\mathfrak{b}_i^h + \textstyle\sum_{j=1}^{d} \mathfrak{w}_{i,j}^h x_j\big)(\mathcal{N}^{\theta}(x) - f(x))(\mathbb{1}_{I_i^{\theta+h}}(x) - \mathbb{1}_{I_i^{\theta}}(x)) \, \mu(\mathrm{d}x)\right] \\
&+ 2\left[\sum_{i=1}^{H} \int_{[\mathscr{a},\mathscr{b}]^d} (\mathfrak{v}_i^\theta + \mathfrak{v}_i^h)\big(\mathfrak{b}_i^\theta + \textstyle\sum_{j=1}^{d} \mathfrak{w}_{i,j}^\theta x_j\big)(\mathcal{N}^{\theta}(x) - f(x))(\mathbb{1}_{I_i^{\theta+h}}(x) - \mathbb{1}_{I_i^{\theta}}(x)) \, \mu(\mathrm{d}x)\right].
\end{aligned} \tag{48}$$

Combining this with the triangle inequality and (41) proves that for all $h \in \mathbb{R}^{\mathfrak{d}} \backslash \{0\}$ we have that

$$\begin{aligned}
\frac{|\mathcal{L}(\theta+h) - \mathcal{L}(\theta) - \langle \mathcal{G}(\theta), h\rangle|}{\|h\|} &\leq \|h\|^{-1} \int_{[\mathscr{a},\mathscr{b}]^d} (\mathcal{N}^{\theta+h}(x) - \mathcal{N}^{\theta}(x))^2 \, \mu(\mathrm{d}x) \\
&+ 2M\|h\|^{-1}\left[\sum_{i=1}^{H} \int_{[\mathscr{a},\mathscr{b}]^d} \big|\mathfrak{v}_i^h \big(\mathfrak{b}_i^h + \textstyle\sum_{j=1}^{d} \mathfrak{w}_{i,j}^h x_j\big)\big| \mathbb{1}_{I_i^{\theta+h}}(x) \, \mu(\mathrm{d}x)\right] \\
&+ 2M\left[\sum_{i=1}^{H} |\mathfrak{v}_i^\theta| \int_{[\mathscr{a},\mathscr{b}]^d} \|h\|^{-1} \big|\mathfrak{b}_i^h + \textstyle\sum_{j=1}^{d} \mathfrak{w}_{i,j}^h x_j\big| \mathbb{1}_{I_i^\theta \Delta I_i^{\theta+h}}(x) \, \mu(\mathrm{d}x)\right] \\
&+ 2M\left[\sum_{i=1}^{H} |\mathfrak{v}_i^\theta + \mathfrak{v}_i^h| \int_{[\mathscr{a},\mathscr{b}]^d} \|h\|^{-1} \big|\mathfrak{b}_i^\theta + \textstyle\sum_{j=1}^{d} \mathfrak{w}_{i,j}^\theta x_j\big| \mathbb{1}_{I_i^\theta \Delta I_i^{\theta+h}}(x) \, \mu(\mathrm{d}x)\right].
\end{aligned} \tag{49}$$

Next note that Lemma 2.9 ensures that there exists $\mathscr{L} \in \mathbb{R}$ such that for all $x \in [\mathscr{a},\mathscr{b}]^d$, $h \in \mathbb{R}^{\mathfrak{d}}$ with $\|h\| \leq 1$ it holds that

$$|\mathcal{N}^{\theta+h}(x) - \mathcal{N}^{\theta}(x)| \leq \mathscr{L}\|h\|. \tag{50}$$

Furthermore, observe that Lemma 2.10 (applied for every $i \in \{1,2,\ldots,H\}$, $h \in \mathbb{R}^{\mathfrak{d}}$, $x \in I_i^{\theta+h} \Delta I_i^\theta$ with $d \curvearrowleft d$, $w_1 \curvearrowleft (\mathfrak{w}_{i,1}^{\theta+h}, \ldots, \mathfrak{w}_{i,d}^{\theta+h})$, $w_2 \curvearrowleft (\mathfrak{w}_{i,1}^{\theta}, \ldots, \mathfrak{w}_{i,d}^{\theta})$, $b_1 \curvearrowleft \mathfrak{b}_i^{\theta+h}$, $b_2 \curvearrowleft \mathfrak{b}_i^\theta$, $\mathscr{a} \curvearrowleft \mathscr{a}$, $\mathscr{b} \curvearrowleft \mathscr{b}$, $I_1 \curvearrowleft I_i^{\theta+h}$, $I_2 \curvearrowleft I_i^\theta$, $x \curvearrowleft x$ in the notation of Lemma 2.10) and (42) show that for all $i \in \{1,2,\ldots,H\}$, $h \in \mathbb{R}^{\mathfrak{d}}$, $x \in I_i^{\theta+h} \Delta I_i^\theta$ it holds that

$$\begin{aligned}
\big|\mathfrak{b}_i^\theta + \textstyle\sum_{j=1}^{d} \mathfrak{w}_{i,j}^\theta x_j\big| \leq \big|\mathfrak{b}_i^h + \textstyle\sum_{j=1}^{d} \mathfrak{w}_{i,j}^h x_j\big| &\leq |\mathfrak{b}_i^h| + \max\{|\mathscr{a}|, |\mathscr{b}|\}\big[\textstyle\sum_{j=1}^{d} |\mathfrak{w}_{i,j}^h|\big] \\
&\leq \|h\| + d\max\{|\mathscr{a}|, |\mathscr{b}|\}\|h\| = \mathfrak{C}\|h\|.
\end{aligned} \tag{51}$$

Moreover, note that (42) implies that for all $i \in \{1,2,\ldots,H\}$, $h \in \mathbb{R}^{\mathfrak{d}} \backslash \{0\}$, $x \in [\mathscr{a},\mathscr{b}]^d$ it holds that

$$\begin{aligned}
\|h\|^{-1}\big|\mathfrak{v}_i^h \big(\mathfrak{b}_i^h + \textstyle\sum_{j=1}^{d} \mathfrak{w}_{i,j}^h x_j\big)\big| &\leq |\mathfrak{b}_i^h| + \max\{|\mathscr{a}|, |\mathscr{b}|\}\big[\textstyle\sum_{j=1}^{d} |\mathfrak{w}_{i,j}^h|\big] \\
&\leq \|h\| + d\max\{|\mathscr{a}|, |\mathscr{b}|\}\|h\| = \mathfrak{C}\|h\|.
\end{aligned} \tag{52}$$

This, (49), (50), (51), and the triangle inequality demonstrate that for all $h \in \mathbb{R}^{\mathfrak{d}} \backslash \{0\}$ with

$\|h\| \leq 1$ it holds that

$$\begin{aligned}
&\frac{|\mathcal{L}(\theta+h)-\mathcal{L}(\theta)-\langle \mathcal{G}(\theta),h\rangle|}{\|h\|} \\
&\leq \mathscr{L}^2\|h\|\big[\mu([\mathscr{a},\mathscr{b}]^d)\big] + 2M\left[\sum_{i=1}^{H}\int_{[\mathscr{a},\mathscr{b}]^d}\|h\|^{-1}\big|\mathfrak{v}_i^h\big(\mathfrak{b}_i^h+\textstyle\sum_{j=1}^{d}\mathfrak{w}_{i,j}^h x_j\big)\big|\,\mu(\mathrm{d}x)\right] \\
&\quad + 2\mathfrak{C}M\left[\sum_{i=1}^{H}(|\mathfrak{v}_i^\theta| + |\mathfrak{v}_i^\theta+\mathfrak{v}_i^h|)\big[\mu(I_i^{\theta+h}\Delta I_i^\theta)\big]\right] \\
&\leq \mathscr{L}^2\|h\|\big[\mu([\mathscr{a},\mathscr{b}]^d)\big] + 2MH\mathfrak{C}\|h\|\big[\mu([\mathscr{a},\mathscr{b}]^d)\big] + 2\mathfrak{C}M\left[\sum_{i=1}^{H}(2|\mathfrak{v}_i^\theta| + |\mathfrak{v}_i^h|)\big[\mu(I_i^{\theta+h}\Delta I_i^\theta)\big]\right] \\
&\leq (\mathscr{L}^2+2MH\mathfrak{C})\|h\|\big[\mu([\mathscr{a},\mathscr{b}]^d)\big] + 2\mathfrak{C}M\left[\sum_{i=1}^{H}\Big(2|\mathfrak{v}_i^\theta|\big[\mu(I_i^{\theta+h}\Delta I_i^\theta)\big] + \|h\|\big[\mu([\mathscr{a},\mathscr{b}]^d)\big]\Big)\right] \\
&= (\mathscr{L}^2+4MH\mathfrak{C})\|h\|\big[\mu([\mathscr{a},\mathscr{b}]^d)\big] + 4\mathfrak{C}M\left[\sum_{i=1}^{H}|\mathfrak{v}_i^\theta|\big[\mu(I_i^{\theta+h}\Delta I_i^\theta)\big]\right].
\end{aligned} \tag{53}$$

Hence, we obtain that

$$\limsup_{\mathbb{R}^{\mathfrak{d}}\setminus\{0\}\ni h\to 0}\left[\frac{|\mathcal{L}(\theta+h)-\mathcal{L}(\theta)-\langle \mathcal{G}(\theta),h\rangle|}{\|h\|}\right] \leq 4\mathfrak{C}M\left[\sum_{i=1}^{H}|\mathfrak{v}_i^\theta|\left(\limsup_{\mathbb{R}^{\mathfrak{d}}\setminus\{0\}\ni h\to 0}\mu(I_i^{\theta+h}\Delta I_i^\theta)\right)\right]. \tag{54}$$

In the following we distinguish between the case $\mathcal{L}(\theta)=0$ and the case $\mathcal{L}(\theta)>0$. We first prove (43) in the case

$$\mathcal{L}(\theta)=0. \tag{55}$$

Observe that (55) implies that for μ-almost every $x\in[\mathscr{a},\mathscr{b}]^d$ it holds that $\mathcal{N}^\theta(x)=f(x)$. This and (42) show that $M=0$. Combining this with (54) establishes (43) in the case $\mathcal{L}(\theta)=0$. In the next step we prove (43) in the case

$$\mathcal{L}(\theta)>0. \tag{56}$$

Note that (40) and (56) ensure that for all $i\in\{1,2,\ldots,H\}$ with $|\mathfrak{v}_i^\theta|>0$ it holds that $|\mathfrak{b}_i^\theta|+\sum_{j=1}^{d}|\mathfrak{w}_{i,j}^\theta|>0$. Corollary 2.8 hence proves that for all $i\in\{1,2,\ldots,H\}$ with $|\mathfrak{v}_i^\theta|>0$ we have that $\limsup_{\mathbb{R}^{\mathfrak{d}}\ni h\to 0}\mu(I_i^{\theta+h}\Delta I_i^\theta)=0$. Combining this with (54) establishes (43) in the case $\mathcal{L}(\theta)>0$. The proof of Proposition 2.11 is thus complete. □

2.5 Lower semicontinuity of the norm of the gradient of the risk function

Lemma 2.12. *Assume Setting 2.1 and let $j\in\mathbb{N}\cap(H(d+1),\mathfrak{d}]$. Then it holds that $\mathbb{R}^{\mathfrak{d}}\ni\theta\mapsto\mathcal{G}_j(\theta)\in\mathbb{R}$ is continuous.*

Proof of Lemma 2.12. Throughout this proof let $\vartheta\in\mathbb{R}^{\mathfrak{d}}$ and let $\theta=(\theta_n)_{n\in\mathbb{N}}\colon\mathbb{N}\to\mathbb{R}^{\mathfrak{d}}$ satisfy $\limsup_{n\to\infty}\|\theta_n-\vartheta\|=0$. Observe that Lemma 2.9 and the fact that $\mathbb{R}\ni x\mapsto\max\{x,0\}\in\mathbb{R}$ is continuous prove that for all $i\in\{1,2,\ldots,H\}$, $x=(x_1,\ldots,x_d)\in[\mathscr{a},\mathscr{b}]^d$ it holds that

$$\begin{aligned}
&\lim_{n\to\infty}\big(\big[\max\big\{\mathfrak{b}_i^{\theta_n}+\textstyle\sum_{k=1}^{d}\mathfrak{w}_{i,k}^{\theta_n}x_k,0\big\}\big](\mathcal{N}^{\theta_n}(x)-f(x))\big) \\
&=\big[\max\big\{\mathfrak{b}_i^{\vartheta}+\textstyle\sum_{k=1}^{d}\mathfrak{w}_{i,k}^{\vartheta}x_k,0\big\}\big](\mathcal{N}^{\vartheta}(x)-f(x))
\end{aligned} \tag{57}$$

and

$$\lim_{n\to\infty}(\mathcal{N}^{\theta_n}(x)-f(x))=\mathcal{N}^{\vartheta}(x)-f(x). \tag{58}$$

Combining (14) and Lebesgue's dominated convergence theorem therefore establishes that $\limsup_{n\to\infty}|\mathcal{G}_j(\theta_n)-\mathcal{G}_j(\vartheta)|=0$. The proof of Lemma 2.12 is thus complete. □

Lemma 2.13. *Assume Setting 2.1, assume $\mu \ll \lambda$, and let $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \ldots, H\}$ satisfy $|\mathfrak{b}_i^\theta| + \sum_{j=1}^{d} |\mathfrak{w}_{i,j}^\theta| > 0$. Then*

(i) it holds for all $j \in \{1, 2, \ldots, d\}$ that $\mathbb{R}^{\mathfrak{d}} \ni \vartheta \mapsto \mathcal{G}_{(i-1)d+j}(\vartheta) \in \mathbb{R}$ is continuous at θ and

(ii) it holds that $\mathbb{R}^{\mathfrak{d}} \ni \vartheta \mapsto \mathcal{G}_{Hd+i}(\vartheta) \in \mathbb{R}$ is continuous at θ.

Proof of Lemma 2.13. Throughout this proof let $j \in \{1, 2, \ldots, d\}$. Note that (14) implies that for all $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $v \in \{0, 1\}$ we have that

$$\begin{aligned}
&\left| \left[\int_{I_i^\vartheta} (x_j)^v (\mathcal{N}^\vartheta(x) - f(x)) \, \mu(\mathrm{d}x) \right] - \left[\int_{I_i^\theta} (x_j)^v (\mathcal{N}^\theta(x) - f(x)) \, \mu(\mathrm{d}x) \right] \right| \\
&\leq \left| \int_{[a,b]^d} (x_j)^v (\mathcal{N}^\vartheta(x) - \mathcal{N}^\theta(x)) \mathbb{1}_{I_i^\vartheta}(x) \, \mu(\mathrm{d}x) \right| \\
&\quad + \left| \int_{[a,b]^d} (x_j)^v (\mathcal{N}^\theta(x) - f(x)) (\mathbb{1}_{I_i^\theta}(x) - \mathbb{1}_{I_i^\vartheta}(x)) \, \mu(\mathrm{d}x) \right| \\
&\leq \left[\sup_{x \in [a,b]^d} |(x_j)^v (\mathcal{N}^\vartheta(x) - \mathcal{N}^\theta(x))| \right] \mu([a,b]^d) \\
&\quad + \left[\sup_{x \in [a,b]^d} |(x_j)^v (\mathcal{N}^\theta(x) - f(x))| \right] \mu(I_i^\theta \Delta I_i^\vartheta).
\end{aligned} \tag{59}$$

Next observe that Lemma 2.9 establishes that for all $v \in \{0, 1\}$ it holds that

$$\limsup_{\mathbb{R}^{\mathfrak{d}} \ni \vartheta \to \theta} \left(\sup_{x \in [a,b]^d} |(x_j)^v (\mathcal{N}^\vartheta(x) - \mathcal{N}^\theta(x))| \right) = 0. \tag{60}$$

Moreover, note that the assumption that $\mu \ll \lambda$, the assumption that $|\mathfrak{b}_i^\theta| + \sum_{k=1}^{d} |\mathfrak{w}_{i,k}^\theta| > 0$, and Corollary 2.8 imply that $\limsup_{\mathbb{R}^{\mathfrak{d}} \ni \vartheta \to \theta} \mu(I_i^\theta \Delta I_i^\vartheta) = 0$. Combining this with (59) and (60) shows that for all $v \in \{0, 1\}$ it holds that

$$\mathbb{R}^{\mathfrak{d}} \ni \vartheta \mapsto \int_{I_i^\vartheta} (x_j)^v (\mathcal{N}^\vartheta(x) - f(x)) \, \mu(\mathrm{d}x) \in \mathbb{R} \tag{61}$$

is continuous at θ. This and (14) establish that $\mathcal{G}_{(i-1)d+j}$ and $\mathcal{G}_{Hd+i}$ are continuous at θ. The proof of Lemma 2.13 is thus complete. □

Lemma 2.14. *Assume Setting 2.1 and let $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \ldots, H\}$ satisfy $|\mathfrak{b}_i^\theta| + \sum_{j=1}^{d} |\mathfrak{w}_{i,j}^\theta| = 0$. Then*

(i) it holds for all $j \in \{1, 2, \ldots, d\}$ that $\mathbb{R}^{\mathfrak{d}} \ni \vartheta \mapsto |\mathcal{G}_{(i-1)d+j}(\vartheta)| \in \mathbb{R}$ is lower semicontinuous at θ and

(ii) it holds that $\mathbb{R}^{\mathfrak{d}} \ni \vartheta \mapsto |\mathcal{G}_{Hd+i}(\vartheta)| \in \mathbb{R}$ is lower semicontinuous at θ.

Proof of Lemma 2.14. Observe that the assumption that $|\mathfrak{b}_i^\theta| + \sum_{j=1}^{d} |\mathfrak{w}_{i,j}^\theta| = 0$ proves that $I_i^\theta = \varnothing$. Combining this with (14) shows that for all $j \in \{1, 2, \ldots, d\}$ it holds that $\mathcal{G}_{(i-1)d+j}(\theta) = \mathcal{G}_{Hd+i}(\theta) = 0$. Therefore, we obtain for all $j \in \{1, 2, \ldots, d\}$ and all $\vartheta = (\vartheta_n)_{n \in \mathbb{N}} \colon \mathbb{N} \to \mathbb{R}^{\mathfrak{d}}$ with $\limsup_{n \to \infty} \|\vartheta_n - \theta\| = 0$ that

$$|\mathcal{G}_{(i-1)d+j}(\theta)| = 0 \leq \liminf_{n \to \infty} |\mathcal{G}_{(i-1)d+j}(\vartheta_n)| \tag{62}$$

and

$$|\mathcal{G}_{Hd+i}(\theta)| = 0 \leq \liminf_{n \to \infty} |\mathcal{G}_{Hd+i}(\vartheta_n)|. \tag{63}$$

Hence, we have for all $j \in \{1, 2, \ldots, d\}$ that $\mathbb{R}^{\mathfrak{d}} \ni \vartheta \mapsto |\mathcal{G}_{(i-1)d+j}(\vartheta)| \in \mathbb{R}$ and $\mathbb{R}^{\mathfrak{d}} \ni \vartheta \mapsto |\mathcal{G}_{Hd+i}(\vartheta)| \in \mathbb{R}$ are lower semicontinuous at θ. The proof of Lemma 2.14 is thus complete. □

Corollary 2.15. *Assume Setting 2.1, assume $\mu \ll \lambda$, and let $k \in \{1, 2, \ldots, \mathfrak{d}\}$. Then it holds that $\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto |\mathcal{G}_k(\theta)| \in \mathbb{R}$ is lower semicontinuous.*

Proof of Corollary 2.15. Note that Lemma 2.12 assures that for all $k \in \mathbb{N} \cap (H(d+1), \mathfrak{d}]$ it holds that $\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto |\mathcal{G}_k(\theta)| \in \mathbb{R}$ is lower semicontinuous. Moreover, observe that Lemmas 2.13 and 2.14 prove that for all $k \in \mathbb{N} \cap [1, H(d+1)]$ it holds that $\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto |\mathcal{G}_k(\theta)| \in \mathbb{R}$ is lower semicontinuous. The proof of Corollary 2.15 is thus complete. □

Corollary 2.16. *Assume Setting 2.1 and assume $\mu \ll \lambda$. Then it holds that $\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \|\mathcal{G}(\theta)\| \in \mathbb{R}$ is lower semicontinuous.*

Proof of Corollary 2.16. Throughout this proof let $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ and let $\theta = (\theta_n)_{n \in \mathbb{N}} \colon \mathbb{N} \to \mathbb{R}^{\mathfrak{d}}$ satisfy $\limsup_{n \to \infty} \|\theta_n - \vartheta\| = 0$. Note that Corollary 2.15 and the fact that for all $v = (v_{k,n})_{(k,n) \in \{1,2\} \times \mathbb{N}} \colon \{1, 2\} \times \mathbb{N} \to [0, \infty)$ it holds that

$$\liminf_{n \to \infty} (v_{1,n} + v_{2,n}) \geq (\liminf_{n \to \infty} v_{1,n}) + (\liminf_{n \to \infty} v_{2,n}) \tag{64}$$

ensure that

$$\begin{aligned} \liminf_{n \to \infty} \|\mathcal{G}(\theta_n)\|^2 &= \liminf_{n \to \infty} \left[\textstyle\sum_{j=1}^{\mathfrak{d}} |\mathcal{G}_j(\theta_n)|^2 \right] \geq \textstyle\sum_{j=1}^{\mathfrak{d}} \left[\liminf_{n \to \infty} |\mathcal{G}_j(\theta_n)|^2 \right] \\ &\geq \textstyle\sum_{j=1}^{\mathfrak{d}} |\mathcal{G}_j(\vartheta)|^2 = \|\mathcal{G}(\vartheta)\|^2. \end{aligned} \tag{65}$$

Hence, we obtain that $\|\mathcal{G}(\vartheta)\| \leq \liminf_{n \to \infty} \|\mathcal{G}(\theta_n)\|$. The proof of Corollary 2.16 is thus complete. □

Corollary 2.17. *Assume Setting 2.1 and assume $\mu \ll \lambda$. Then there exists an open $U \subseteq \mathbb{R}^{\mathfrak{d}}$ such that $\int_{\mathbb{R}^{\mathfrak{d}} \backslash U} 1 \, \mathrm{d}x = 0$, $\mathcal{L}|_U \in C^1(U, \mathbb{R})$, and $\nabla(\mathcal{L}|_U) = \mathcal{G}|_U$.*

Proof of Corollary 2.17. Throughout this proof let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy

$$U = \left\{ \theta \in \mathbb{R}^{\mathfrak{d}} \colon \left[\forall \, i \in \{1, 2, \ldots, H\} \colon \left(|\mathfrak{b}_i^\theta| + \textstyle\sum_{j=1}^{d} |\mathfrak{w}_{i,j}^\theta| > 0 \right) \right] \right\}. \tag{66}$$

Observe that (66) ensures that $U \subseteq \mathbb{R}^{\mathfrak{d}}$ is open. Moreover, note that the fact that $\mathbb{R}^{\mathfrak{d}} \backslash U \subseteq \left(\bigcup_{i=1}^{H} \{\theta \in \mathbb{R}^{\mathfrak{d}} \colon \mathfrak{b}_i^\theta = 0\} \right)$ assures that $\int_{\mathbb{R}^{\mathfrak{d}} \backslash U} 1 \, \mathrm{d}x = 0$. Furthermore, observe that Proposition 2.11 demonstrates that for all $\theta \in U$ it holds that $\mathcal{L}$ is differentiable at θ with $(\nabla \mathcal{L})(\theta) = \mathcal{G}(\theta)$. In addition, note that Lemma 2.12 and Lemma 2.13 prove that for all $\theta \in U$, $i \in \{1, 2, \ldots, \mathfrak{d}\}$ it holds that $\mathbb{R}^{\mathfrak{d}} \ni \vartheta \mapsto \mathcal{G}_i(\vartheta) \in \mathbb{R}$ is continuous at θ. Hence, we obtain that $\mathcal{L}|_U \in C^1(U, \mathbb{R})$. This completes the proof of Corollary 2.17. □

3 Convergence of the risk of gradient flows (GFs) in the training of ANNs

In this section we establish in Theorem 3.2 in Subsection 3.1 below, in Corollary 3.3 in Subsection 3.2 below, and in Corollary 3.5 in Subsection 3.4 below convergence results for the risk of GFs. In particular, in Theorem 3.2 we establish that the risk of every bounded GF trajectory converges to the risk of a critical point. Our proof of Theorem 3.2 employs the fundamental theorem of calculus type result for the risk of GFs in Lemma 3.1 in Subsection 3.1 and the fundamental fact that the standard norm of the generalized gradient function $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ is lower semicontinuous, which we established in Corollary 2.16 above. The proof of Lemma 3.1 is entirely analogous to the proof of [6, Lemma 3.5]. In Corollary 3.3 we establish that the risk of every bounded GF trajectory with sufficiently small initial risk converges to the risk of the global minima of the risk function. In Corollary 3.5 we employ the characterization result for criticial points for affine linear target functions in Cheridito et al. [7] to specialize Corollary 3.3 to the situation of affine linear target functions.

3.1 Convergence of the risk of GFs to the risk of a critical point

Lemma 3.1. *Assume Setting 2.1, let $T \in (0,\infty)$, and let $\Theta \in C([0,T],\mathbb{R}^{\mathfrak{d}})$ satisfy for all $t \in [0,T]$ that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s$ (cf. Corollary 2.4). Then it holds for all $t \in [0,T]$ that $\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|\mathcal{G}(\Theta_s)\|^2\,\mathrm{d}s$.*

Proof of Lemma 3.1. The proof of Lemma 3.1 is entirely analogous to the proof of [6, Lemma 3.5]. □

Theorem 3.2. *Assume Setting 2.1, assume $\mu \ll \lambda$, and let $\Theta \in C([0,\infty),\mathbb{R}^{\mathfrak{d}})$ satisfy for all $t \in [0,\infty)$ that $\sup_{s\in[0,\infty)}\|\Theta_s\| < \infty$ and*

$$\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s \tag{67}$$

(cf. Corollary 2.4). Then there exists $\vartheta \in \mathcal{G}^{-1}(\{0\})$ such that $\limsup_{t\to\infty}\mathcal{L}(\Theta_t) = \mathcal{L}(\vartheta)$.

Proof of Theorem 3.2. Observe that Lemma 3.1 implies that $\int_0^\infty \|\mathcal{G}(\Theta_s)\|^2\,\mathrm{d}s < \infty$. Hence, we have that $\liminf_{t\to\infty}\|\mathcal{G}(\Theta_t)\| = 0$. This proves that there exists $\tau = (\tau_n)_{n\in\mathbb{N}}\colon \mathbb{N} \to [0,\infty)$ which satisfies $\liminf_{n\to\infty}\tau_n = \infty$ and

$$\limsup_{n\to\infty}\|\mathcal{G}(\Theta_{\tau_n})\| = 0. \tag{68}$$

Note that the fact that $\sup_{n\in\mathbb{N}}\|\Theta_{\tau_n}\| \leq \sup_{t\in[0,\infty)}\|\Theta_t\| < \infty$ ensures that there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ and a strictly increasing $n\colon \mathbb{N} \to \mathbb{N}$ which satisfies

$$\limsup_{k\to\infty}\|\Theta_{\tau_{n(k)}} - \vartheta\| = 0. \tag{69}$$

Observe that (68), (69), and Corollary 2.16 demonstrate that

$$\|\mathcal{G}(\vartheta)\| \leq \liminf_{k\to\infty}\|\mathcal{G}(\Theta_{\tau_{n(k)}})\| = 0. \tag{70}$$

Furthermore, note that Lemma 3.1 assures that $[0,\infty) \ni t \mapsto \mathcal{L}(\Theta_t) \in \mathbb{R}$ is non-increasing. Combining this and (69) with Lemma 2.9 proves that $\limsup_{t\to\infty}\mathcal{L}(\Theta_t) = \lim_{k\to\infty}\mathcal{L}(\Theta_{\tau_{n(k)}}) = \mathcal{L}(\vartheta)$. The proof of Theorem 3.2 is thus complete. □

3.2 Convergence of the risk of GFs to the minimal risk

Corollary 3.3. *Assume Setting 2.1, assume $\mu \ll \lambda$, let $\mathbf{m} \in \mathbb{R}$ satisfy $\mathbf{m} = \inf_{\theta\in\mathbb{R}^{\mathfrak{d}}}\mathcal{L}(\theta)$, and let $\Theta \in C([0,\infty),\mathbb{R}^{\mathfrak{d}})$ satisfy $\sup_{t\in[0,\infty)}\|\Theta_t\| < \infty$, $\forall\, t \in [0,\infty)\colon \Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s$, and $\forall\,\theta \in \mathcal{G}^{-1}(\{0\}) \cap \mathcal{L}^{-1}((\mathbf{m},\infty))\colon \inf_{t\in[0,\infty)}\mathcal{L}(\Theta_t) < \mathcal{L}(\theta)$ (cf. Corollary 2.4). Then*

$$\limsup_{t\to\infty}\mathcal{L}(\Theta_t) = \mathbf{m}. \tag{71}$$

Proof of Corollary 3.3. Observe that Theorem 3.2 assures that there exists $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ which satisfies $\mathcal{G}(\vartheta) = 0$ and $\limsup_{t\to\infty}\mathcal{L}(\Theta_t) = \mathcal{L}(\vartheta)$. In the following we prove (71) by contradiction. We thus assume that

$$\mathcal{L}(\vartheta) > \mathbf{m}. \tag{72}$$

Note that (72) and the assumption that $\forall\,\theta \in \mathcal{G}^{-1}(\{0\}) \cap \mathcal{L}^{-1}((\mathbf{m},\infty))\colon \inf_{t\in[0,\infty)}\mathcal{L}(\Theta_t) < \mathcal{L}(\theta)$ imply that

$$\inf_{t\in[0,\infty)}\mathcal{L}(\Theta_t) < \mathcal{L}(\vartheta). \tag{73}$$

Moreover, observe that Lemma 3.1 proves that $[0,\infty) \ni t \mapsto \mathcal{L}(\Theta_t) \in \mathbb{R}$ is non-increasing. Combining this with (73) shows that

$$\mathcal{L}(\vartheta) = \limsup_{t\to\infty}\mathcal{L}(\Theta_t) = \inf_{t\in[0,\infty)}\mathcal{L}(\Theta_t) < \mathcal{L}(\vartheta). \tag{74}$$

This is a contradiction. The proof of Corollary 3.3 is thus complete. □

3.3 Risks of critical points for affine linear target functions

Proposition 3.4. *Assume Setting 2.1, assume $d = 1$, and let $\alpha, \beta \in \mathbb{R}$, $\rho \in (0, \infty)$ satisfy for all $E \in \mathcal{B}([\mathfrak{a}, \mathfrak{b}])$, $x \in [\mathfrak{a}, \mathfrak{b}]$ that $\mu(E) = \rho\lambda_1(E)$ and $f(x) = \alpha x + \beta$. Then*

(i) there exists $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ such that $\mathcal{L}(\vartheta) = \inf_{\theta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(\theta) = 0$ and

(ii) it holds for all $\theta \in \mathcal{G}^{-1}(\{0\}) \cap \mathcal{L}^{-1}((0, \infty))$ that $\mathcal{L}(\theta) \geq \frac{\rho\alpha^2(\mathfrak{b}-\mathfrak{a})^3}{12(2\lfloor H/2 \rfloor + 1)^4}$.

Proof of Proposition 3.4. Note that the assumption that $d = 1$ implies that $\mathfrak{d} = 3H + 1$. Let $\psi \in \mathbb{R}^{3H+1}$ satisfy $\mathfrak{w}_{1,1}^{\psi} = 1$, $\mathfrak{b}_1^{\psi} = -\mathfrak{a}$, $\mathfrak{v}_1^{\psi} = \alpha$, $\mathfrak{c}_1^{\psi} = \beta + \alpha\mathfrak{a}$, and $\forall\, i \in \mathbb{N} \cap (1, H]$: $\mathfrak{w}_{i,1}^{\psi} = \mathfrak{b}_i^{\psi} = \mathfrak{v}_i^{\psi} = 0$. Observe that for all $x \in [\mathfrak{a}, \mathfrak{b}]$ we have that

$$\mathcal{N}^{\psi}(x) = \alpha \max\{x - \mathfrak{a}, 0\} + \beta + \alpha\mathfrak{a} = \alpha(x - \mathfrak{a}) + \alpha\mathfrak{a} + \beta = \alpha x + \beta = f(x). \tag{75}$$

This shows that $\mathcal{L}(\psi) = 0$. Combining this with the fact that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that $\mathcal{L}(\theta) \geq 0$ establishes item (i). We now prove item (ii). For this assume in the following without loss of generality that $\alpha \neq 0$ and let $\mathfrak{G} = (\mathfrak{G}_1, \ldots, \mathfrak{G}_{\mathfrak{d}}) \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \ldots, H\}$ that

$$\begin{aligned}
\mathfrak{G}_i(\theta) &= 2\mathfrak{v}_i^{\theta} \int_{\mathfrak{a}}^{\mathfrak{b}} x(\mathcal{N}^{\theta}(x) - f(x)) \mathbb{1}_{[0,\infty)}(\mathfrak{w}_{i,1}^{\theta} x + \mathfrak{b}_i^{\theta}) \, \mathrm{d}x, \\
\mathfrak{G}_{H+i}(\theta) &= 2\mathfrak{v}_i^{\theta} \int_{\mathfrak{a}}^{\mathfrak{b}} (\mathcal{N}^{\theta}(x) - f(x)) \mathbb{1}_{[0,\infty)}(\mathfrak{w}_{i,1}^{\theta} x + \mathfrak{b}_i^{\theta}) \, \mathrm{d}x, \\
\mathfrak{G}_{2H+i}(\theta) &= 2 \int_{\mathfrak{a}}^{\mathfrak{b}} \big[\max\{\mathfrak{w}_{i,1}^{\theta} x + \mathfrak{b}_i^{\theta}, 0\}\big] (\mathcal{N}^{\theta}(x) - f(x)) \, \mathrm{d}x, \\
\text{and} \qquad \mathfrak{G}_{\mathfrak{d}}(\theta) &= 2 \int_{\mathfrak{a}}^{\mathfrak{b}} (\mathcal{N}^{\theta}(x) - f(x)) \, \mathrm{d}x
\end{aligned} \tag{76}$$

(cf., e.g., [7, Lemma 3.5]). Note that (14) and the assumption that for all $E \in \mathcal{B}([\mathfrak{a}, \mathfrak{b}])$ it holds that $\mu(E) = \rho\lambda_1(E)$ show that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \ldots, H\}$ it holds that $\mathfrak{G}_{2H+i}(\theta) = \rho^{-1}\mathcal{G}_{2H+i}(\theta)$ and $\mathfrak{G}_{\mathfrak{d}}(\theta) = \rho^{-1}\mathcal{G}_{\mathfrak{d}}(\theta)$. In the following let $\theta \in \mathbb{R}^{3H+1}$ satisfy $\mathcal{L}(\theta) > 0 = \|\mathcal{G}(\theta)\|$ and let $\vartheta \in \mathbb{R}^{3H+1}$ satisfy for all $i \in \{1, 2, \ldots, H\}$ that

$$\mathfrak{w}_{i,1}^{\vartheta} = \mathfrak{w}_{i,1}^{\theta}, \qquad \mathfrak{b}_i^{\vartheta} = \mathfrak{b}_i^{\theta}, \qquad \mathfrak{v}_i^{\vartheta} = \mathfrak{v}_i^{\theta} \mathbb{1}_{(0,\infty)}(|\mathfrak{w}_{i,1}^{\theta}| + |\mathfrak{b}_i^{\theta}|), \qquad \text{and} \qquad \mathfrak{c}^{\vartheta} = \mathfrak{c}^{\theta}. \tag{77}$$

Observe that (77) ensures that

$$\mathcal{N}^{\vartheta} = \mathcal{N}^{\theta}, \qquad \mathcal{L}(\vartheta) = \mathcal{L}(\theta), \qquad \text{and} \qquad \mathcal{G}(\vartheta) = \mathcal{G}(\theta) = 0. \tag{78}$$

Furthermore, note that the fact that for all $i \in \{1, 2, \ldots, H\}$ it holds that $\mathfrak{G}_{2H+i}(\vartheta) = \rho^{-1}\mathcal{G}_{2H+i}(\vartheta)$ and $\mathfrak{G}_{\mathfrak{d}}(\vartheta) = \rho^{-1}\mathcal{G}_{\mathfrak{d}}(\vartheta)$ assures that for all $i \in \{1, 2, \ldots, H\}$ it holds that

$$\mathfrak{G}_{2H+i}(\vartheta) = \mathfrak{G}_{\mathfrak{d}}(\vartheta) = 0. \tag{79}$$

Next observe that the fact that for all $i \in \{1, 2, \ldots, H\}$ with $|\mathfrak{w}_{i,1}^{\vartheta}| + |\mathfrak{b}_i^{\vartheta}| = 0$ it holds that $\mathfrak{v}_i^{\vartheta} = 0$ implies that for all $i \in \{1, 2, \ldots, H\}$ with $|\mathfrak{w}_{i,1}^{\vartheta}| + |\mathfrak{b}_i^{\vartheta}| = 0$ we have that

$$\mathfrak{G}_i(\vartheta) = \mathfrak{G}_{H+i}(\vartheta) = 0. \tag{80}$$

In addition, note that for all $i \in \{j \in \{1, 2, \ldots, H\} \colon |\mathfrak{w}_{j,1}^{\vartheta}| + |\mathfrak{b}_j^{\vartheta}| > 0\}$ and almost all $x \in [\mathfrak{a}, \mathfrak{b}]$ it holds that $\mathbb{1}_{[0,\infty)}(\mathfrak{w}_{i,1}^{\vartheta} x + \mathfrak{b}_i^{\vartheta}) = \mathbb{1}_{(0,\infty)}(\mathfrak{w}_{i,1}^{\vartheta} x + \mathfrak{b}_i^{\vartheta}) = \mathbb{1}_{I_i^{\vartheta}}(x)$. This shows that for all $i \in \{1, 2, \ldots, H\}$ with $|\mathfrak{w}_{i,1}^{\vartheta}| + |\mathfrak{b}_i^{\vartheta}| > 0$ it holds that

$$\mathfrak{G}_i(\vartheta) = \rho^{-1}\mathcal{G}_i(\vartheta) = 0 \qquad \text{and} \qquad \mathfrak{G}_{H+i}(\vartheta) = \rho^{-1}\mathcal{G}_{H+i}(\vartheta) = 0. \tag{81}$$

Combining (79)–(81) demonstrates that $\mathfrak{G}(\vartheta) = 0$. Cheridito et al. [7, Corollary 2.7] hence proves that there exists $n \in \{0, 2, 4, \ldots\} \cap (0, H]$ which satisfies

$$\mathcal{L}(\theta) = \mathcal{L}(\vartheta) = \rho \int_{\mathscr{a}}^{\mathscr{b}} (\mathcal{N}^{\vartheta}(x) - (\alpha x + \beta))^2 \, \mathrm{d}x = \frac{\rho \alpha^2 (\mathscr{b} - \mathscr{a})^3}{12(n+1)^4}. \tag{82}$$

Observe that the fact that $\frac{n}{2} \in \mathbb{Z}$ and the fact that $n \leq H$ assure that $n \leq 2\lfloor H/2 \rfloor$. Combining this with (82) shows that

$$\mathcal{L}(\theta) = \frac{\rho \alpha^2 (\mathscr{b} - \mathscr{a})^3}{12(n+1)^4} \geq \frac{\rho \alpha^2 (\mathscr{b} - \mathscr{a})^3}{12(2\lfloor H/2 \rfloor + 1)^4}. \tag{83}$$

This establishes item (ii). The proof of Proposition 3.4 is thus complete. □

3.4 Convergence of the risk of GFs to the minimal risk for affine linear target functions

Corollary 3.5. *Assume Setting 2.1, assume $d = 1$, let $\alpha, \beta \in \mathbb{R}$, $\rho \in (0, \infty)$ satisfy for all $E \in \mathcal{B}([\mathscr{a}, \mathscr{b}])$, $x \in [\mathscr{a}, \mathscr{b}]$ that $\mu(E) = \rho \lambda_1(E)$ and $f(x) = \alpha x + \beta$, and let $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ satisfy $\sup_{t \in [0,\infty)} \|\Theta_t\| < \infty$, $\forall \, t \in [0, \infty) \colon \Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) \, \mathrm{d}s$, and $\inf_{t \in [0,\infty)} \mathcal{L}(\Theta_t) < \frac{\rho \alpha^2 (\mathscr{b} - \mathscr{a})^3}{12(2\lfloor H/2 \rfloor + 1)^4}$ (cf. Corollary 2.4). Then*

$$\limsup_{t \to \infty} \mathcal{L}(\Theta_t) = 0. \tag{84}$$

Proof of Corollary 3.5. Note that item (i) in Proposition 3.4 implies that $\inf_{\theta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(\theta) = 0$. Moreover, observe that item (ii) in Proposition 3.4 demonstrates that for all $\theta \in \mathcal{G}^{-1}(\{0\}) \cap \mathcal{L}^{-1}((0, \infty))$ we have that

$$\mathcal{L}(\theta) \geq \frac{\rho \alpha^2 (\mathscr{b} - \mathscr{a})^3}{12(2\lfloor H/2 \rfloor + 1)^4} > \inf_{t \in [0,\infty)} \mathcal{L}(\Theta_t). \tag{85}$$

Combining this and Corollary 3.3 (applied with $\mathbf{m} \curvearrowleft 0$ in the notation of Corollary 3.3) establishes that $\limsup_{t \to \infty} \mathcal{L}(\Theta_t) = 0$. The proof of Corollary 3.5 is thus complete. □

4 A priori estimates for GFs in the training of ANNs

In this section we establish in Proposition 4.1 in Subsection 4.1 below, in Corollary 4.2 in Subsection 4.1, in Corollary 4.3 in Subsection 4.2 below, and in Proposition 4.4 in Subsection 4.3 several general a priori estimates for GF trajectories. In particular, Corollary 4.2 demonstrates that the limit value of the risk of every GF trajectory is bounded by the squared L^2-error $\inf_{\xi \in \mathbb{R}}[\int_{[\mathscr{a}, \mathscr{b}]^d} (f(x) - \xi)^2 \, \mu(\mathrm{d}x)]$ of constant approximations of the target function $f \colon [\mathscr{a}, \mathscr{b}]^d \to \mathbb{R}$. Our proof of Corollary 4.2 is based on an application of the a priori estimate in Proposition 4.1. Corollary 4.3, in particular, proves that the norm of every GF trajectory is bounded until the first time where the risk is smaller than $\inf_{\xi \in \mathbb{R}}[\int_{[\mathscr{a}, \mathscr{b}]^d} (f(x) - \xi)^2 \, \mu(\mathrm{d}x)]$. Our proof of Corollary 4.3 also employs an application of Proposition 4.1. A result similar to Proposition 4.1 has been obtained in [6, Lemma 3.2] in the special situation where the measure μ is the Lebesgue–Borel measure on $[0, 1]$ and where the target function f is a constant function, and our proof of Proposition 4.1 uses similar ideas as the proof of [6, Lemma 3.2].

In Proposition 4.4 we identify appropriate invariant quantities for the GF dynamics. In the scientific literature Proposition 4.4 has already been asserted and proved in Williams et al. [23, Lemma 3] in the case where the measure μ is chosen in a way so that the function $\mathcal{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ describes the empirical risk and where the input is 1-dimensional (where $d = 1$). Moreover, a result similar to Proposition 4.4 has also been established in Du et al. [12, Theorem 2.1] in the situation of deep ANNs without biases.

4.1 Lyapunov type functions for GFs

Proposition 4.1. *Assume Setting 2.1, let $\xi \in \mathbb{R}$, let $V \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $V(\theta) = \|\theta\|^2 + |\mathfrak{c}^\theta - 2\xi|^2$, and let $\Theta \in C([0,\infty), \mathbb{R}^{\mathfrak{d}})$ satisfy for all $t \in [0,\infty)$ that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s$ (cf. Corollary 2.4). Then it holds for all $t \in [0,\infty)$ that*

$$\begin{aligned} V(\Theta_t) &= V(\Theta_0) - 8\int_0^t \mathcal{L}(\Theta_s)\,\mathrm{d}s - 8\int_0^t \left[\int_{[a,b]^d} (f(x)-\xi)(\mathcal{N}^{\Theta_s}(x) - f(x))\,\mu(\mathrm{d}x)\right]\mathrm{d}s \\ &\leq V(\Theta_0) + 4\int_0^t \left[\int_{[a,b]^d} (f(x)-\xi)^2\,\mu(\mathrm{d}x) - \mathcal{L}(\Theta_s)\right]\mathrm{d}s. \end{aligned} \tag{86}$$

Proof of Proposition 4.1. Note that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} &(\nabla V)(\theta) \\ &= 2\big(\mathfrak{w}^\theta_{1,1}, \ldots, \mathfrak{w}^\theta_{1,d}, \mathfrak{w}^\theta_{2,1}, \ldots, \mathfrak{w}^\theta_{2,d}, \ldots, \mathfrak{w}^\theta_{H,1}, \ldots, \mathfrak{w}^\theta_{H,d}, \mathfrak{b}^\theta_1, \ldots, \mathfrak{b}^\theta_H, \mathfrak{v}^\theta_1, \ldots, \mathfrak{v}^\theta_H, 2\mathfrak{c}^\theta - 2\xi\big). \end{aligned} \tag{87}$$

This and (14) imply that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} &\langle (\nabla V)(\theta), \mathcal{G}(\theta)\rangle \\ &= 4\left[\sum_{i=1}^H \mathfrak{v}^\theta_i \int_{[a,b]^d} \big(\mathfrak{b}^\theta_i + \textstyle\sum_{j=1}^d \mathfrak{w}^\theta_{i,j} x_j\big)(\mathcal{N}^\theta(x) - f(x))\mathbb{1}_{(0,\infty)}\big(\mathfrak{b}^\theta_i + \sum_{j=1}^d \mathfrak{w}^\theta_{i,j} x_j\big)\,\mu(\mathrm{d}x)\right] \\ &+ 4\left[\sum_{i=1}^H \mathfrak{v}^\theta_i \int_{[a,b]^d} \big[\max\big\{\mathfrak{b}^\theta_i + \textstyle\sum_{j=1}^d \mathfrak{w}^\theta_{i,j} x_j, 0\big\}\big](\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x)\right] \\ &+ 8(\mathfrak{c}^\theta - \xi)\left[\int_{[a,b]^d} (\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x)\right]. \end{aligned} \tag{88}$$

Hence, we obtain for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\begin{aligned} &\langle (\nabla V)(\theta), \mathcal{G}(\theta)\rangle \\ &= 8\left[\int_{[a,b]^d} \Big(\textstyle\sum_{i=1}^H \mathfrak{v}^\theta_i \big[\max\big\{\mathfrak{b}^\theta_i + \sum_{j=1}^d \mathfrak{w}^\theta_{i,j} x_j, 0\big\}\big]\Big)(\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x)\right] \\ &+ 8(\mathfrak{c}^\theta - \xi)\left[\int_{[a,b]^d} (\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x)\right] \\ &= 8\int_{[a,b]^d} (\mathcal{N}^\theta(x) - \xi)(\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x) \\ &= 8\int_{[a,b]^d} (\mathcal{N}^\theta(x) - f(x))^2\,\mu(\mathrm{d}x) + 8\int_{[a,b]^d} (f(x)-\xi)(\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x) \\ &= 8\mathcal{L}(\theta) + 8\int_{[a,b]^d} (f(x)-\xi)(\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x). \end{aligned} \tag{89}$$

Next observe that the Cauchy-Schwarz inequality implies that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} &\int_{[a,b]^d} (f(x)-\xi)(\mathcal{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x) \\ &\geq -\left[\int_{[a,b]^d} (f(x)-\xi)^2\,\mu(\mathrm{d}x)\right]^{1/2}\left[\int_{[a,b]^d} (\mathcal{N}^\theta(x) - f(x))^2\,\mu(\mathrm{d}x)\right]^{1/2} \\ &= -\left[\int_{[a,b]^d} (f(x)-\xi)^2\,\mu(\mathrm{d}x)\right]^{1/2}\sqrt{\mathcal{L}(\theta)}. \end{aligned} \tag{90}$$

Combining this with the fact that for all $a, b \in \mathbb{R}$ it holds that $ab \leq \frac{a^2+b^2}{2}$ demonstrates that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ we have that

$$\int_{[a,b]^d} (f(x) - \xi)(\mathcal{N}^\theta(x) - f(x))\, \mu(\mathrm{d}x) \geq -\frac{1}{2}\left[\int_{[a,b]^d} (f(x) - \xi)^2\, \mu(\mathrm{d}x)\right] - \frac{\mathcal{L}(\theta)}{2}. \tag{91}$$

This, (89), the fact that $V \in C^\infty(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, and, e.g., [6, Lemma 3.1] show for all $t \in [0, \infty)$ that

$$\begin{aligned} V(\Theta_t) - V(\Theta_0) &= -\int_0^t \langle (\nabla V)(\Theta_s), \mathcal{G}(\Theta_s)\rangle \, \mathrm{d}s \\ &= -8\int_0^t \mathcal{L}(\Theta_s)\, \mathrm{d}s - 8\int_0^t \left[\int_{[a,b]^d} (f(x) - \xi)(\mathcal{N}^{\Theta_s}(x) - f(x))\, \mu(\mathrm{d}x)\right] \mathrm{d}s \\ &\leq -8\int_0^t \mathcal{L}(\Theta_s)\, \mathrm{d}s + 4\int_0^t \left[\int_{[a,b]^d} (f(x) - \xi)^2\, \mu(\mathrm{d}x) + \mathcal{L}(\Theta_s)\right] \mathrm{d}s \\ &= 4\int_0^t \left[\int_{[a,b]^d} (f(x) - \xi)^2\, \mu(\mathrm{d}x) - \mathcal{L}(\Theta_s)\right] \mathrm{d}s. \end{aligned} \tag{92}$$

The proof of Proposition 4.1 is thus complete. □

Corollary 4.2. *Assume Setting 2.1 and let $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ satisfy for all $t \in [0, \infty)$ that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\, \mathrm{d}s$ (cf. Corollary 2.4). Then*

$$\limsup_{t\to\infty} \mathcal{L}(\Theta_t) \leq \inf_{\xi\in\mathbb{R}} \left[\int_{[a,b]^d} (f(x) - \xi)^2\, \mu(\mathrm{d}x)\right]. \tag{93}$$

Proof of Corollary 4.2. Throughout this proof let $\mathbf{m}, \xi, \nu \in \mathbb{R}$ satisfy $\mathbf{m} = \limsup_{t\to\infty} \mathcal{L}(\Theta_t)$ and $\nu = \int_{[a,b]^d} (f(x) - \xi)^2\, \mu(\mathrm{d}x)$. Note that Lemma 3.1 implies that $[0, \infty) \ni t \mapsto \mathcal{L}(\Theta_t) \in \mathbb{R}$ is non-increasing. This assures that $\inf_{t\in[0,\infty)} \mathcal{L}(\Theta_t) = \mathbf{m}$. Proposition 4.1 hence demonstrates that for all $t \in [0, \infty)$ it holds that

$$\begin{aligned} 0 \leq V(\Theta_t) &\leq V(\Theta_0) + 4\int_0^t (\nu - \mathcal{L}(\Theta_s))\, \mathrm{d}s \\ &\leq V(\Theta_0) + 4\int_0^t (\nu - \mathbf{m})\, \mathrm{d}s = V(\Theta_0) - 4t(\mathbf{m} - \nu). \end{aligned} \tag{94}$$

Therefore, we obtain for all $t \in (0, \infty)$ that $\mathbf{m} - \nu \leq \frac{V(\Theta_0)}{4t}$. This shows that

$$\mathbf{m} \leq \limsup_{t\to\infty} \left[\tfrac{V(\Theta_0)}{4t} + \nu\right] = \nu. \tag{95}$$

The proof of Corollary 4.2 is thus complete. □

4.2 A priori estimates for GFs with large risk

Corollary 4.3. *Assume Setting 2.1, let $\nu, \xi \in \mathbb{R}$ satisfy $\nu = \int_{[a,b]^d} (f(x) - \xi)^2\, \mu(\mathrm{d}x)$, and let $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ satisfy for all $t \in [0, \infty)$ that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\, \mathrm{d}s$ (cf. Corollary 2.4). Then*

$$\sup_{t\in[0,\infty),\, \mathcal{L}(\Theta_t) \geq \nu \mathbb{1}_{(0,\infty)}(t)} \|\Theta_t\| \leq 3\|\Theta_0\|^2 + 8|\xi|^2 < \infty. \tag{96}$$

Proof of Corollary 4.3. Throughout this proof let $V \colon \mathbb{R}^{\mathfrak{d}} \to [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $V(\Theta) = \|\theta\|^2 + |\mathfrak{c}^\theta - 2\xi|^2$ and let $t \in (0, \infty)$ satisfy $\mathcal{L}(\Theta_t) \geq \nu$. Observe that Lemma 3.1 implies

that $[0,\infty) \ni s \mapsto \mathcal{L}(\Theta_s) \in \mathbb{R}$ is non-increasing. This shows that for all $s \in [0,t]$ it holds that $\mathcal{L}(\Theta_s) \geq \mathcal{L}(\Theta_t) \geq \nu$. Combining this with Proposition 4.1 demonstrates that

$$\|\Theta_t\| \leq V(\Theta_t) \leq V(\Theta_0) + 4\int_0^t (\nu - \mathcal{L}(\Theta_s))\,\mathrm{d}s \leq V(\Theta_0). \tag{97}$$

Furthermore, note that the fact that for all $x, y \in \mathbb{R}$ it holds that $(x+y)^2 \leq 2(x^2+y^2)$ ensures that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$V(\theta) = \|\theta\|^2 + |\mathfrak{c}^\theta - 2\xi|^2 \leq \|\theta\|^2 + 2(|\mathfrak{c}^\theta|^2 + |2\xi|^2) \leq 3\|\theta\|^2 + 8|\xi|^2. \tag{98}$$

Combining this with (97) proves that $\|\Theta_t\| \leq 3\|\Theta_0\|^2 + 8|\xi|^2 < \infty$. This completes the proof of Corollary 4.3. □

4.3 Invariant quantities for GFs

Proposition 4.4. *Assume Setting 2.1, let $W_i \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$, $i \in \{1,2,\ldots,H\}$, satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1,2,\ldots,H\}$ that $W_i(\theta) = \big[\sum_{j=1}^d (\mathfrak{w}_{i,j}^\theta)^2\big] + (\mathfrak{b}_i^\theta)^2 - (\mathfrak{v}_i^\theta)^2$, and let $\Theta \in C([0,\infty), \mathbb{R}^{\mathfrak{d}})$ satisfy for all $t \in [0,\infty)$ that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s$ (cf. Corollary 2.4). Then*

(i) it holds for all $t \in [0,\infty)$, $i \in \{1,2,\ldots,H\}$ that $W_i(\Theta_t) = W_i(\Theta_0)$ and

(ii) it holds for all $t \in [0,\infty)$ that $\sum_{i=1}^H W_i(\Theta_t) = \sum_{i=1}^H W_i(\Theta_0)$.

Proof of Proposition 4.4. Observe that the assumption that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1,2,\ldots,H\}$ it holds that $W_i(\theta) = \big[\sum_{j=1}^d (\mathfrak{w}_{i,j}^\theta)^2\big] + (\mathfrak{b}_i^\theta)^2 - (\mathfrak{v}_i^\theta)^2$ and (14) demonstrate that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1,2,\ldots,H\}$ we have that

$$\begin{aligned}
&\langle (\nabla W_i)(\theta), \mathcal{G}(\theta)\rangle \\
&= 4\mathfrak{v}_i^\theta \int_{[\mathscr{a},\mathscr{b}]^d} \big(\mathfrak{b}_i^\theta + \textstyle\sum_{j=1}^d \mathfrak{w}_{i,j}^\theta x_j\big)(\mathscr{N}^\theta(x) - f(x))\mathbb{1}_{(0,\infty)}\big(\mathfrak{b}_i^\theta + \sum_{j=1}^d \mathfrak{w}_{i,j}^\theta x_j\big)\,\mu(\mathrm{d}x) \\
&- 4\mathfrak{v}_i^\theta \int_{[\mathscr{a},\mathscr{b}]^d} \big[\max\{\mathfrak{b}_i^\theta + \textstyle\sum_{j=1}^d \mathfrak{w}_{i,j}^\theta x_j, 0\}\big](\mathscr{N}^\theta(x) - f(x))\,\mu(\mathrm{d}x) = 0.
\end{aligned} \tag{99}$$

This, the fact that for all $i \in \{1,2,\ldots,H\}$ it holds that $W_i \in C^\infty(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, and, e.g., [6, Lemma 3.1] show for all $i \in \{1,2,\ldots,H\}$, $t \in [0,\infty)$ that

$$W_i(\Theta_t) = W_i(\Theta_0) - \int_0^t \langle (\nabla W_i)(\Theta_s), \mathcal{G}(\Theta_s)\rangle\,\mathrm{d}s = W_i(\Theta_0). \tag{100}$$

This proves item (i). Next note that item (i) establishes item (ii). The proof of Proposition 4.4 is thus complete. □

5 Properties of ANN parametrizations with small risk and one hidden neuron

In Theorem 6.7 in Section 6 below we establish in the case where the measure μ (see Setting 2.1) is up to a constant the Lebesgue–Borel measure on $[\mathscr{a},\mathscr{b}]$, where the hidden layer consists of only one neuron (where $H = 1$), and where the target function $f \colon [\mathscr{a},\mathscr{b}] \to \mathbb{R}$ is affine linear that the risk of every not necessarily bounded GF trajectory converges to zero. Our proof of Theorem 6.7 employs, among other things, the a priori bounds for GF trajectories with sufficiently small initial risk in Lemma 6.2 in Subsection 6.1 below, the well known mean square approximation results in Lemma 5.1 and Corollary 5.2 in Subsection 5.1 below, the lower bound for the product of the slope of the target function and its ANN approximations in Corollary 5.6

in Subsection 5.3 below, and appropriate lower bounds for the transformation between the input and hidden layer of the considered ANN in Lemma 5.7 in Subsection 5.3.

In Lemma 5.1 in Subsection 5.1 we recall the elementary fact that the mean value of a given continuous function on a compact real interval is the best constant mean square approximation of the considered continuous function. Corollary 5.2 in Subsection 5.1 specializes Lemma 5.1 to the case where the considered continuous function is affine linear. Lemma 5.1 follows, e.g., from [3, Lemma 2.1] and only for completeness we include in this section detailed proofs for Lemma 5.1 and Corollary 5.2.

In Setting 5.3 in Subsection 5.2 below we specialize Setting 2.1 from Subsection 2.1 above and present the mathematical framework which we frequently employ in Sections 5 and 6 to formulate ANNs with ReLU activation, one hidden layer, one neuron on the input layer (corresponding to the case $d = 1$ in Setting 2.1), and one neuron on the hidden layer (corresponding to the case $H = 1$ in Setting 2.1) and the corresponding risk functions (see (106) in Setting 5.3).

In Subsection 5.3 we study realizations of ANNs whose risk is strictly smaller than the risk which can be achieved by the best constant approximation (cf. Lemma 5.1). Our proof of the a priori bound result for GF trajectories with sufficiently small initial risk in Lemma 6.2 in Subsection 6.1 employs Lemma 3.1 from Subsection 3.1, Lemma 5.1 and Corollary 5.2 from Subsection 5.1 and Lemma 5.4, Corollary 5.6, and Proposition 5.8 from Subsection 5.3. The elementary result in Lemma 5.4 in Subsection 5.3 shows that for every ANN with parameter vector $\theta = (\theta_1, \ldots, \theta_4) \in \mathbb{R}^4$ we have that the realization associated to θ is Lipschitz continuous with the Lipschitz constant $|\theta_1\theta_3|$.

Corollary 5.6 in Subsection 5.3 demonstrates in the case where there exist $\alpha, \beta \in \mathbb{R}$ such that the target function satisfies for all $x \in [\mathfrak{a}, \mathfrak{b}]$ that $f(x) = \alpha x + \beta$ that for every ANN whose risk is strictly smaller than the risk which can be achieved by the best constant approximation (cf. Lemma 5.1) with parameter vector $\theta = (\theta_1, \ldots, \theta_4) \in \mathbb{R}^4$ we have that the slope α of the target function and the slope $\theta_1\theta_3$ of the realization of the ANN must have the same sign in the sense that $\alpha\theta_1\theta_3 > 0$. Our proof of Corollary 5.6 employs an application of Lemma 5.5 in Subsection 5.3. Lemma 5.5, in turn, establishes the statement of Corollary 5.6 in the special case where the slope α of the target function is assumed to be strictly positive in the sense that $\alpha > 0$.

Lemma 5.7 in Subsection 5.3 establishes that for every ANN whose risk is strictly smaller than the risk which can be achieved by the best constant approximation (cf. Lemma 5.1) with parameter vector $\theta = (\theta_1, \ldots, \theta_4) \in \mathbb{R}^4$ we have that the hidden neuron of this ANN cannot be inactive and we must have that $\max\{\theta_1\mathfrak{a} + \theta_2, \theta_1\mathfrak{b} + \theta_2\} > 0$. This simply follows from the fact that if the neuron was inactive in the sense that $\max\{\theta_1\mathfrak{a} + \theta_2, \theta_1\mathfrak{b} + \theta_2\} \leq 0$, then the realization function associated to θ would be constant which would result in a larger risk.

Finally, Proposition 5.8 in Subsection 5.3, the main result of Section 5, loosely speaking, reveals that for every ANN whose risk is strictly smaller than the risk which can be achieved by the best constant approximation (cf. Lemma 5.1) with parameter vector $\theta = (\theta_1, \ldots, \theta_4) \in \mathbb{R}^4$ we have that the slope of the realization of the ANN θ is uniformly bounded from below and from above.

5.1 Mean square approximations through constant functions

Lemma 5.1. *Let $\xi, \mathfrak{a} \in \mathbb{R}$, $\mathfrak{b} \in (\mathfrak{a}, \infty)$, $f \in C([\mathfrak{a}, \mathfrak{b}], \mathbb{R})$. Then*

$$\int_{\mathfrak{a}}^{\mathfrak{b}} (f(x) - \xi)^2 \, \mathrm{d}x \geq \int_{\mathfrak{a}}^{\mathfrak{b}} \Big(f(x) - \tfrac{1}{\mathfrak{b}-\mathfrak{a}} \big[\textstyle\int_{\mathfrak{a}}^{\mathfrak{b}} f(y) \, \mathrm{d}y \big] \Big)^2 \mathrm{d}x. \tag{101}$$

Proof of Lemma 5.1. Throughout this proof let $\mu \in \mathbb{R}$ satisfy $\mu = (\mathfrak{b}-\mathfrak{a})^{-1} \int_{\mathfrak{a}}^{\mathfrak{b}} f(y) \, \mathrm{d}y$. Observe that for all $u \in \mathbb{R}$ it holds that

$$\int_{\mathfrak{a}}^{\mathfrak{b}} (f(x) - u)^2 \, \mathrm{d}x = \int_{\mathfrak{a}}^{\mathfrak{b}} (f(x))^2 \, \mathrm{d}x - 2u\mu(\mathfrak{b} - \mathfrak{a}) + u^2(\mathfrak{b} - \mathfrak{a}). \tag{102}$$

Hence, we obtain that

$$\begin{aligned}
&\int_{\mathscr{a}}^{\mathscr{b}} (f(x)-\xi)^2\,\mathrm{d}x - \int_{\mathscr{a}}^{\mathscr{b}} \left(f(x) - \tfrac{1}{\mathscr{b}-\mathscr{a}}\left[\textstyle\int_{\mathscr{a}}^{\mathscr{b}} f(y)\,\mathrm{d}y\right]\right)^2 \mathrm{d}x \\
&= \int_{\mathscr{a}}^{\mathscr{b}} (f(x)-\xi)^2\,\mathrm{d}x - \int_{\mathscr{a}}^{\mathscr{b}} (f(x)-\mu)^2\,\mathrm{d}x \\
&= -2\xi\mu(\mathscr{b}-\mathscr{a}) + \xi^2(\mathscr{b}-\mathscr{a}) + 2\mu^2(\mathscr{b}-\mathscr{a}) - \mu^2(\mathscr{b}-\mathscr{a}) \\
&= (\mathscr{b}-\mathscr{a})(\xi^2 - 2\xi\mu + \mu^2) = (\mathscr{b}-\mathscr{a})(\xi-\mu)^2 \geq 0.
\end{aligned} \tag{103}$$

This completes the proof of Lemma 5.1. □

Corollary 5.2. *Let $\xi, \alpha, \beta, \mathscr{a} \in \mathbb{R}$, $\mathscr{b} \in (\mathscr{a}, \infty)$. Then*

$$\int_{\mathscr{a}}^{\mathscr{b}} (\alpha x + \beta - \xi)^2\,\mathrm{d}x \geq \int_{\mathscr{a}}^{\mathscr{b}} \left((\alpha x+\beta) - (\alpha[\tfrac{\mathscr{a}+\mathscr{b}}{2}] + \beta)\right)^2 \mathrm{d}x = \frac{\alpha^2(\mathscr{b}-\mathscr{a})^3}{12}. \tag{104}$$

Proof of Corollary 5.2. Note that $\int_{\mathscr{a}}^{\mathscr{b}} (\alpha x+\beta)\,\mathrm{d}x = \frac{\alpha(\mathscr{b}^2-\mathscr{a}^2)}{2} + \beta(\mathscr{b}-\mathscr{a}) = (\mathscr{b}-\mathscr{a})(\alpha[\frac{\mathscr{b}+\mathscr{a}}{2}]+\beta)$. Lemma 5.1 hence shows that

$$\begin{aligned}
\int_{\mathscr{a}}^{\mathscr{b}} (\alpha x+\beta-\xi)^2\,\mathrm{d}x &\geq \int_{\mathscr{a}}^{\mathscr{b}} \left((\alpha x+\beta) - (\alpha[\tfrac{\mathscr{b}+\mathscr{a}}{2}]+\beta)\right)^2 \mathrm{d}x = \int_{\mathscr{a}}^{\mathscr{b}} \alpha^2 (x - [\tfrac{\mathscr{b}+\mathscr{a}}{2}])^2\,\mathrm{d}x \\
&= \left[\tfrac{\alpha^2}{3}\left(x - [\tfrac{\mathscr{b}+\mathscr{a}}{2}]\right)^3\right]_{x=\mathscr{a}}^{x=\mathscr{b}} = \frac{\alpha^2}{3}\left[\left(\frac{\mathscr{b}-\mathscr{a}}{2}\right)^3 - \left(\frac{\mathscr{a}-\mathscr{b}}{2}\right)^3\right] \\
&= \frac{\alpha^2}{24}\left[(\mathscr{b}-\mathscr{a})^3 - (\mathscr{a}-\mathscr{b})^3\right] = \frac{\alpha^2(\mathscr{b}-\mathscr{a})^3}{12}.
\end{aligned} \tag{105}$$

The proof of Corollary 5.2 is thus complete. □

5.2 Mathematical description of ANNs with one hidden neuron

Setting 5.3. *Let $\mathscr{a} \in \mathbb{R}$, $\mathscr{b} \in (\mathscr{a}, \infty)$, $\rho \in (0, \infty)$, $f \in C([\mathscr{a}, \mathscr{b}], \mathbb{R})$, $\mathfrak{w}, \mathfrak{b}, \mathfrak{v}, \mathfrak{c} \in C(\mathbb{R}^4, \mathbb{R})$ satisfy for all $\theta = (\theta_1, \ldots, \theta_4) \in \mathbb{R}^4$ that $\mathfrak{w}^\theta = \theta_1$, $\mathfrak{b}^\theta = \theta_2$, $\mathfrak{v}^\theta = \theta_3$, and $\mathfrak{c}^\theta = \theta_4$, let $\mathscr{N} = (\mathscr{N}^\theta)_{\theta\in\mathbb{R}^4} \colon \mathbb{R}^4 \to C(\mathbb{R}, \mathbb{R})$ and $\mathcal{L} \colon \mathbb{R}^4 \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^4$, $x \in \mathbb{R}$ that*

$$\mathscr{N}^\theta(x) = \mathfrak{v}^\theta \max\{\mathfrak{w}^\theta x + \mathfrak{b}^\theta, 0\} + \mathfrak{c}^\theta \tag{106}$$

and $\mathcal{L}(\theta) = \rho \int_{\mathscr{a}}^{\mathscr{b}} (\mathscr{N}^\theta(y) - f(y))^2\,\mathrm{d}y$, let $\mathfrak{R}_r \in C^1(\mathbb{R}, \mathbb{R})$, $r \in \mathbb{N}$, satisfy for all $x \in \mathbb{R}$ that

$$\limsup_{r\to\infty}\left(|\mathfrak{R}_r(x) - \max\{x, 0\}| + |(\mathfrak{R}_r)'(x) - \mathbb{1}_{(0,\infty)}(x)|\right) = 0 \tag{107}$$

and $\sup_{r\in\mathbb{N}} \sup_{y\in[-|x|,|x|]}(|\mathfrak{R}_r(y)| + |(\mathfrak{R}_r)'(y)|) < \infty$, let $\mathfrak{L}_r \colon \mathbb{R}^4 \to \mathbb{R}$, $r \in \mathbb{N}$, satisfy for all $r \in \mathbb{N}$, $\theta \in \mathbb{R}^4$ that

$$\mathfrak{L}_r(\theta) = \rho \int_{\mathscr{a}}^{\mathscr{b}} \left(\mathfrak{v}^\theta\left[\mathfrak{R}_r(\mathfrak{w}^\theta x + \mathfrak{b}^\theta)\right] + \mathfrak{c}^\theta - f(x)\right)^2 \mathrm{d}x, \tag{108}$$

let $\|\cdot\| \colon \left(\bigcup_{n\in\mathbb{N}} \mathbb{R}^n\right) \to \mathbb{R}$ and $\langle\cdot,\cdot\rangle \colon \left(\bigcup_{n\in\mathbb{N}} (\mathbb{R}^n \times \mathbb{R}^n)\right) \to \mathbb{R}$ satisfy for all $n \in \mathbb{N}$, $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ that $\|x\| = [\sum_{i=1}^n |x_i|^2]^{1/2}$ and $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$, let $\lambda \colon \mathcal{B}(\mathbb{R}) \to [0, \infty]$ be the Lebesgue–Borel measure on $\mathbb{R}$, let $I^\theta \subseteq \mathbb{R}$, $\theta \in \mathbb{R}^4$, satisfy for all $\theta \in \mathbb{R}^4$ that $I^\theta = \{x \in [\mathscr{a}, \mathscr{b}] \colon \mathfrak{w}^\theta x + \mathfrak{b}^\theta > 0\}$, and let $\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_4) \colon \mathbb{R}^4 \to \mathbb{R}^4$ satisfy for all $\theta \in \{\vartheta \in \mathbb{R}^4 \colon ((\nabla\mathfrak{L}_r)(\vartheta))_{r\in\mathbb{N}}$ is convergent$\}$ that $\mathcal{G}(\theta) = \lim_{r\to\infty}(\nabla\mathfrak{L}_r)(\theta)$.

5.3 Properties of ANNs with small risk and one hidden neuron

Lemma 5.4. *Assume Setting 5.3 and let $\theta \in \mathbb{R}^4$. Then it holds for all $x, y \in \mathbb{R}$ that*

$$|\mathcal{N}^\theta(x) - \mathcal{N}^\theta(y)| \leq |\mathfrak{w}^\theta \mathfrak{v}^\theta||x - y|. \tag{109}$$

Proof of Lemma 5.4. Observe that (106) ensures that for all $x, y \in \mathbb{R}$ it holds that

$$\begin{aligned} |\mathcal{N}^\theta(x) - \mathcal{N}^\theta(y)| &= |\mathfrak{v}^\theta \max\{\mathfrak{w}^\theta x + \mathfrak{b}^\theta, 0\} - \mathfrak{v}^\theta \max\{\mathfrak{w}^\theta y + \mathfrak{b}^\theta, 0\}| \\ &= |\mathfrak{v}^\theta||\max\{\mathfrak{w}^\theta x + \mathfrak{b}^\theta, 0\} - \max\{\mathfrak{w}^\theta y + \mathfrak{b}^\theta, 0\}| \\ &\leq |\mathfrak{v}^\theta||(\mathfrak{w}^\theta x + \mathfrak{b}^\theta) - (\mathfrak{w}^\theta y + \mathfrak{b}^\theta)| = |\mathfrak{w}^\theta \mathfrak{v}^\theta||x - y|. \end{aligned} \tag{110}$$

The proof of Lemma 5.4 is thus complete. □

Lemma 5.5. *Assume Setting 5.3, let $\alpha \in (0, \infty)$, $\beta \in \mathbb{R}$ satisfy for all $x \in [\mathfrak{a}, \mathfrak{b}]$ that $f(x) = \alpha x + \beta$, and let $\theta \in \mathbb{R}^4$ satisfy $\mathcal{L}(\theta) < \frac{\rho \alpha^2 (\mathfrak{b} - \mathfrak{a})^3}{12}$. Then*

$$\mathfrak{w}^\theta \mathfrak{v}^\theta > 0. \tag{111}$$

Proof of Lemma 5.5. We prove (111) by contradiction. We thus assume that

$$\mathfrak{w}^\theta \mathfrak{v}^\theta \leq 0. \tag{112}$$

Note that (112) ensures that for all $x, y \in [\mathfrak{a}, \mathfrak{b}]$ with $x \leq y$ it holds that

$$\mathcal{N}^\theta(x) \geq \mathcal{N}^\theta(y). \tag{113}$$

In the following we distinguish between the case $\mathcal{N}^\theta(\mathfrak{b}) \geq f(\mathfrak{b})$, the case $\mathcal{N}^\theta(\mathfrak{a}) \leq f(\mathfrak{a})$, and the case $\min\{f(\mathfrak{b}) - \mathcal{N}^\theta(\mathfrak{b}), \mathcal{N}^\theta(\mathfrak{a}) - f(\mathfrak{a})\} > 0$. We first establish the contradiction in the case

$$\mathcal{N}^\theta(\mathfrak{b}) \geq f(\mathfrak{b}). \tag{114}$$

Observe that (113) and (114) imply for all $x \in [\mathfrak{a}, \mathfrak{b}]$ that $\mathcal{N}^\theta(x) \geq \mathcal{N}^\theta(\mathfrak{b}) \geq f(\mathfrak{b}) \geq f(x)$. Combining this with Corollary 5.2 proves that $\frac{\rho \alpha^2 (\mathfrak{b} - \mathfrak{a})^3}{12} > \mathcal{L}(\theta) \geq \rho \int_{\mathfrak{a}}^{\mathfrak{b}} (f(\mathfrak{b}) - f(x))^2 \, \mathrm{d}x \geq \frac{\rho \alpha^2 (\mathfrak{b} - \mathfrak{a})^3}{12}$, which is a contradiction. In the next step we establish the contradiction in the case

$$\mathcal{N}^\theta(\mathfrak{a}) \leq f(\mathfrak{a}). \tag{115}$$

Note that (113) and (115) show for all $x \in [\mathfrak{a}, \mathfrak{b}]$ that $\mathcal{N}^\theta(x) \leq \mathcal{N}^\theta(\mathfrak{a}) \leq f(\mathfrak{a}) \leq f(x)$. This and Corollary 5.2 imply that $\frac{\rho \alpha^2 (\mathfrak{b} - \mathfrak{a})^3}{12} > \mathcal{L}(\theta) \geq \rho \int_{\mathfrak{a}}^{\mathfrak{b}} (f(\mathfrak{a}) - f(x))^2 \, \mathrm{d}x \geq \frac{\rho \alpha^2 (\mathfrak{b} - \mathfrak{a})^3}{12}$, which is a contradiction. Finally, we establish the contradiction in the case

$$\min\{f(\mathfrak{b}) - \mathcal{N}^\theta(\mathfrak{b}), \mathcal{N}^\theta(\mathfrak{a}) - f(\mathfrak{a})\} > 0. \tag{116}$$

Observe that (116) and intermediate value theorem assure that there exists $u \in [\mathfrak{a}, \mathfrak{b}]$ such that $\mathcal{N}^\theta(u) = f(u)$. This and (113) prove that $\forall \, x \in [\mathfrak{a}, u]\colon \mathcal{N}^\theta(x) \geq \mathcal{N}^\theta(u) = f(u) \geq f(x)$ and $\forall \, x \in [u, \mathfrak{b}]\colon \mathcal{N}^\theta(x) \leq \mathcal{N}^\theta(u) = f(u) \leq f(x)$. Combining this with Corollary 5.2 demonstrates that

$$\begin{aligned} \frac{\rho \alpha^2 (\mathfrak{b} - \mathfrak{a})^3}{12} &> \mathcal{L}(\theta) = \rho \int_{\mathfrak{a}}^{u} (\mathcal{N}^\theta(x) - f(x))^2 \, \mathrm{d}x + \rho \int_{u}^{\mathfrak{b}} (\mathcal{N}^\theta(x) - f(x))^2 \, \mathrm{d}x \\ &\geq \rho \int_{\mathfrak{a}}^{u} (f(u) - f(x))^2 \, \mathrm{d}x + \rho \int_{u}^{\mathfrak{b}} (f(x) - f(u))^2 \, \mathrm{d}x \\ &= \rho \int_{\mathfrak{a}}^{\mathfrak{b}} (f(x) - f(u))^2 \, \mathrm{d}x \geq \frac{\rho \alpha^2 (\mathfrak{b} - \mathfrak{a})^3}{12}. \end{aligned} \tag{117}$$

This is a contradiction. The proof of Lemma 5.5 is thus complete. □

Corollary 5.6. *Assume Setting 5.3, let $\alpha, \beta \in \mathbb{R}$ satisfy for all $x \in [\mathscr{a}, \mathscr{b}]$ that $f(x) = \alpha x + \beta$, and let $\theta \in \mathbb{R}^4$ satisfy $\mathcal{L}(\theta) < \frac{\rho\alpha^2(\mathscr{b}-\mathscr{a})^3}{12}$. Then*

$$\alpha \mathfrak{w}^\theta \mathfrak{v}^\theta > 0. \tag{118}$$

Proof of Corollary 5.6. Note that the assumption that $\mathcal{L}(\theta) < \frac{\rho\alpha^2(\mathscr{b}-\mathscr{a})^3}{12}$ assures that $\alpha \neq 0$. In the following we distinguish between the case $\alpha > 0$ and the case $\alpha < 0$. First observe that Lemma 5.5 establishes (118) in the case $\alpha > 0$. In the next step we prove (118) in the case $\alpha < 0$. Note that

$$\begin{aligned} \frac{\rho\alpha^2(\mathscr{b}-\mathscr{a})^3}{12} > \mathcal{L}(\theta) &= \rho \int_{\mathscr{a}}^{\mathscr{b}} \left(\mathfrak{v}^\theta \max\{\mathfrak{w}^\theta x + \mathfrak{b}^\theta, 0\} + \mathfrak{c}^\theta - (\alpha x + \beta)\right)^2 \mathrm{d}x \\ &= \rho \int_{\mathscr{a}}^{\mathscr{b}} \left((-\mathfrak{v}^\theta) \max\{\mathfrak{w}^\theta x + \mathfrak{b}^\theta, 0\} + (-\mathfrak{c}^\theta) - (-\alpha x - \beta)\right)^2 \mathrm{d}x. \end{aligned} \tag{119}$$

Combining this, the fact that $-\alpha > 0$, and Lemma 5.5 (applied with $\theta \curvearrowleft (\mathfrak{w}^\theta, \mathfrak{b}^\theta, -\mathfrak{v}^\theta, -\mathfrak{c}^\theta)$, $\alpha \curvearrowleft -\alpha$, $\beta \curvearrowleft -\beta$ in the notation of Lemma 5.5) demonstrates that $\alpha\mathfrak{w}^\theta\mathfrak{v}^\theta = (-\alpha)\mathfrak{w}^\theta(-\mathfrak{v}^\theta) > 0$. This establishes (118) in the case $\alpha < 0$. The proof of Corollary 5.6 is thus complete. □

Lemma 5.7. *Assume Setting 5.3, let $m \in \mathbb{R}$ satisfy $m = \rho \int_{\mathscr{a}}^{\mathscr{b}} (f(x) - (\mathscr{b}-\mathscr{a})^{-1} \int_{\mathscr{a}}^{\mathscr{b}} f(y)\,\mathrm{d}y)^2\,\mathrm{d}x$, and let $\theta \in \mathbb{R}^4$ satisfy $\mathcal{L}(\theta) < m$. Then $\max\{\mathfrak{w}^\theta \mathscr{a} + \mathfrak{b}^\theta, \mathfrak{w}^\theta \mathscr{b} + \mathfrak{b}^\theta\} > 0$.*

Proof of Lemma 5.7. We prove Lemma 5.7 by contradiction. We thus assume that

$$\max\{\mathfrak{w}^\theta \mathscr{a} + \mathfrak{b}^\theta, \mathfrak{w}^\theta \mathscr{b} + \mathfrak{b}^\theta\} \leq 0. \tag{120}$$

Observe that (120) ensures that for all $x \in [\mathscr{a}, \mathscr{b}]$ we have that

$$\mathfrak{w}^\theta x + \mathfrak{b}^\theta = \left[\tfrac{\mathscr{b}-x}{\mathscr{b}-\mathscr{a}}\right](\mathfrak{w}^\theta \mathscr{a} + \mathfrak{b}^\theta) + \left[\tfrac{x-\mathscr{a}}{\mathscr{b}-\mathscr{a}}\right](\mathfrak{w}^\theta \mathscr{b} + \mathfrak{b}^\theta) \leq 0. \tag{121}$$

This implies for all $x \in [\mathscr{a}, \mathscr{b}]$ that $\max\{\mathfrak{w}^\theta x + \mathfrak{b}^\theta, 0\} = 0$. Therefore, we obtain for all $x \in [\mathscr{a}, \mathscr{b}]$ that $\mathscr{N}^\theta(x) = \mathfrak{c}^\theta$. Combining this with Lemma 5.1 proves that $\mathcal{L}(\theta) \geq m$. This is a contradiction. The proof of Lemma 5.7 is thus complete. □

Proposition 5.8. *Assume Setting 5.3 and let $m \in \mathbb{R}$, $\varepsilon \in (0, \infty)$ satisfy $m = \rho \int_{\mathscr{a}}^{\mathscr{b}} (f(x) - (\mathscr{b}-\mathscr{a})^{-1} \int_{\mathscr{a}}^{\mathscr{b}} f(y)\,\mathrm{d}y)^2\,\mathrm{d}x$. Then there exists $\mathfrak{C} \in (0, \infty)$ such that for all $\theta \in \{\vartheta \in \mathbb{R}^4 \colon \mathcal{L}(\vartheta) \leq m - \varepsilon\}$ it holds that $\mathfrak{C}^{-1} \leq |\mathfrak{w}^\theta \mathfrak{v}^\theta| \leq \mathfrak{C}$.*

Proof of Proposition 5.8. Throughout this proof assume without loss of generality that $\varepsilon \leq m$, assume without loss of generality that $\{\vartheta \in \mathbb{R}^4 \colon \mathcal{L}(\vartheta) \leq m - \varepsilon\} \neq \varnothing$, and let $M \in \mathbb{R}$ satisfy $M = \max\{1, \sup_{x \in [\mathscr{a}, \mathscr{b}]} |f(x)|\}$. We first prove that there exists $\mathfrak{C} \in (0, \infty)$ such that for all $\theta \in \{\vartheta \in \mathbb{R}^4 \colon \mathcal{L}(\vartheta) \leq m - \varepsilon\}$ it holds that

$$\mathfrak{C}^{-1} \leq |\mathfrak{w}^\theta \mathfrak{v}^\theta|. \tag{122}$$

Note that Lemma 5.4 implies for all $\theta \in \mathbb{R}^4$, $x \in [\mathscr{a}, \mathscr{b}]$ that $|\mathscr{N}^\theta(x) - \mathscr{N}^\theta(\mathscr{a})| \leq |\mathfrak{w}^\theta \mathfrak{v}^\theta||x - \mathscr{a}| \leq |\mathfrak{w}^\theta \mathfrak{v}^\theta|(\mathscr{b}-\mathscr{a})$. Combining this, Lemma 5.1, and Minkowski's inequality establishes for all $\theta \in \mathbb{R}^4$ that

$$\begin{aligned} \sqrt{\mathcal{L}(\theta)} &= \left[\rho \int_{\mathscr{a}}^{\mathscr{b}} (\mathscr{N}^\theta(x) - f(x))^2\,\mathrm{d}x\right]^{1/2} \\ &\geq \left[\rho \int_{\mathscr{a}}^{\mathscr{b}} (\mathscr{N}^\theta(\mathscr{a}) - f(x))^2\,\mathrm{d}x\right]^{1/2} - \left[\rho \int_{\mathscr{a}}^{\mathscr{b}} (\mathscr{N}^\theta(x) - \mathscr{N}^\theta(\mathscr{a}))^2\,\mathrm{d}x\right]^{1/2} \\ &\geq \inf_{\xi \in \mathbb{R}} \left[\rho \int_{\mathscr{a}}^{\mathscr{b}} (f(x) - \xi)^2\,\mathrm{d}x\right]^{1/2} - \left[\rho \int_{\mathscr{a}}^{\mathscr{b}} |\mathfrak{w}^\theta \mathfrak{v}^\theta|^2 |\mathscr{b} - \mathscr{a}|^2\,\mathrm{d}x\right]^{1/2} \\ &= \sqrt{m} - |\mathfrak{w}^\theta \mathfrak{v}^\theta| \sqrt{\rho(\mathscr{b}-\mathscr{a})^3}. \end{aligned} \tag{123}$$

This implies for all $\theta \in \mathbb{R}^4$ that

$$|\mathfrak{w}^\theta \mathfrak{v}^\theta| \geq \tfrac{\sqrt{m}-\sqrt{\mathcal{L}(\theta)}}{\sqrt{\rho(\mathscr{b}-\mathscr{a})^3}}. \tag{124}$$

Hence, we obtain for all $\theta \in \{\vartheta \in \mathbb{R}^4 \colon \mathcal{L}(\vartheta) \leq m - \varepsilon\}$ that

$$|\mathfrak{w}^\theta \mathfrak{v}^\theta| \geq \tfrac{\sqrt{m}-\sqrt{\mathcal{L}(\theta)}}{\sqrt{\rho(\mathscr{b}-\mathscr{a})^3}} \geq \tfrac{\sqrt{m}-\sqrt{m-\varepsilon}}{\sqrt{\rho(\mathscr{b}-\mathscr{a})^3}} > 0. \tag{125}$$

This establishes (122). In the next step we verify that there exists $\mathfrak{C} \in (0, \infty)$ such that for all $\theta \in \{\vartheta \in \mathbb{R}^4 \colon \mathcal{L}(\vartheta) \leq m - \varepsilon\}$ it holds that

$$|\mathfrak{w}^\theta \mathfrak{v}^\theta| \leq \mathfrak{C}. \tag{126}$$

We prove (126) by contradiction. In the following we thus assume that

$$\sup_{\theta \in \{\vartheta \in \mathbb{R}^4 \colon \mathcal{L}(\vartheta) \leq m-\varepsilon\}} |\mathfrak{w}^\theta \mathfrak{v}^\theta| = \infty. \tag{127}$$

Observe that (127) ensures that there exist $\theta_n \in \{\vartheta \in \mathbb{R}^4 \colon \mathcal{L}(\vartheta) \leq m - \varepsilon\}$, $n \in \mathbb{N}$, which satisfy for all $n \in \mathbb{N}$ that

$$|\mathfrak{w}^{\theta_n} \mathfrak{v}^{\theta_n}| \geq 2(n+1)^2 M > 0. \tag{128}$$

Roughly speaking, we next establish that for all sufficiently large n it holds that the function $[\mathscr{a}, \mathscr{b}] \ni x \mapsto \mathcal{N}^{\theta_n}(x) \in \mathbb{R}$ is almost constant in the sense that $\limsup_{n\to\infty} \lambda(I^{\theta_n}) = 0$ and, thereafter, we use this to prove (126). Note that (106) ensures that for all $n \in \mathbb{N}$, $x \in I^{\theta_n}$ it holds that $\mathcal{N}^{\theta_n}(x) = \mathfrak{w}^{\theta_n} \mathfrak{v}^{\theta_n} x + (\mathfrak{b}^{\theta_n} \mathfrak{v}^{\theta_n} + \mathfrak{c}^{\theta_n})$. Combining this with (128) and the fact that for all $\alpha, \beta, c \in \mathbb{R}$ with $\alpha \neq 0$ it holds that

$$\begin{aligned} &\lambda(\{x \in [\mathscr{a}, \mathscr{b}] \colon |\alpha x + \beta| \leq |c|\}) \leq \lambda(\{x \in \mathbb{R} \colon |\alpha x + \beta| \leq |c|\}) \\ &= \lambda\big(\{x \in \mathbb{R} \colon |x + \tfrac{\beta}{\alpha}| \leq \tfrac{|c|}{|\alpha|}\}\big) = \lambda\big([-\tfrac{\beta}{\alpha} - \tfrac{|c|}{|\alpha|}, -\tfrac{\beta}{\alpha} + \tfrac{|c|}{|\alpha|}]\big) = \tfrac{2|c|}{|\alpha|} \end{aligned} \tag{129}$$

implies that for all $n \in \mathbb{N}$ we have that

$$\lambda\Big(\big\{x \in I^{\theta_n} \colon |\mathcal{N}^{\theta_n}(x)| \leq (n+1)M\big\}\Big) \leq \min\Big\{\lambda(I^{\theta_n}), \tfrac{2(n+1)M}{|\mathfrak{w}^{\theta_n} \mathfrak{v}^{\theta_n}|}\Big\} \leq \min\big\{\lambda(I^{\theta_n}), \tfrac{1}{n+1}\big\}. \tag{130}$$

Hence, we obtain for all $n \in \mathbb{N}$ that

$$\begin{aligned} \lambda\Big(\big\{x \in I^{\theta_n} \colon |\mathcal{N}^{\theta_n}(x)| > (n+1)M\big\}\Big) &= \lambda(I^{\theta_n}) - \lambda\Big(\big\{x \in I^{\theta_n} \colon |\mathcal{N}^{\theta_n}(x)| \leq (n+1)M\big\}\Big) \\ &\geq \lambda(I^{\theta_n}) - \min\big\{\lambda(I^{\theta_n}), \tfrac{1}{n+1}\big\} \\ &= \max\big\{0, \lambda(I^{\theta_n}) - \tfrac{1}{n+1}\big\}. \end{aligned} \tag{131}$$

Furthermore, observe that for all $x \in I^{\theta_n}$ with $|\mathcal{N}^{\theta_n}(x)| > (n+1)M$ it holds that

$$|\mathcal{N}^{\theta_n}(x) - f(x)| \geq |\mathcal{N}^{\theta_n}(x)| - |f(x)| \geq |\mathcal{N}^{\theta_n}(x)| - M > (n+1)M - M = nM. \tag{132}$$

Combining this with (128) and (131) implies that for all $n \in \mathbb{N}$ it holds that

$$m > m - \varepsilon \geq \mathcal{L}(\theta_n) \geq n^2 M^2 \max\big\{0, \lambda(I^{\theta_n}) - \tfrac{1}{n+1}\big\}. \tag{133}$$

Hence, we obtain that

$$\begin{aligned} 0 &\leq \limsup_{n\to\infty} [\lambda(I^{\theta_n})] = \limsup_{n\to\infty} \big[\lambda(I^{\theta_n}) - \tfrac{1}{n+1}\big] \leq \limsup_{n\to\infty} \big[\max\big\{0, \lambda(I^{\theta_n}) - \tfrac{1}{n+1}\big\}\big] \\ &\leq \limsup_{n\to\infty} \big[\tfrac{m}{n^2 M^2}\big] = 0. \end{aligned} \tag{134}$$

Next note that (106) ensures that for all $n \in \mathbb{N}$, $x \in [\mathfrak{a}, \mathfrak{b}] \backslash I^{\theta_n}$ it holds that $\mathcal{N}^{\theta_n}(x) = \mathfrak{c}^{\theta_n}$. This implies for all $n \in \mathbb{N}$ that

$$\mathcal{L}(\theta_n) \geq \rho \int_{[\mathfrak{a},\mathfrak{b}] \backslash I^{\theta_n}} (f(x) - \mathcal{N}^{\theta_n}(x))^2 \, \mathrm{d}x \geq \inf_{\xi \in \mathbb{R}} \left[\rho \int_{[\mathfrak{a},\mathfrak{b}] \backslash I^{\theta_n}} (f(x) - \xi)^2 \, \mathrm{d}x \right]. \tag{135}$$

Furthermore, observe that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} [\mathfrak{a}, \mathfrak{b}] \backslash I^{\theta_n} &= \{x \in [\mathfrak{a}, \mathfrak{b}] \colon \mathfrak{w}^{\theta_n} x + \mathfrak{b}^{\theta_n} \leq 0\} = \{x \in [\mathfrak{a}, \mathfrak{b}] \colon \mathfrak{w}^{\theta_n} x \leq -\mathfrak{b}^{\theta_n}\} \\ &= \begin{cases} [\mathfrak{a}, \mathfrak{b}] & \colon \mathfrak{b}^{\theta_n} \leq \mathfrak{w}^{\theta_n} = 0 \\ \varnothing & \colon \mathfrak{b}^{\theta_n} > \mathfrak{w}^{\theta_n} = 0 \\ [\mathfrak{a}, \mathfrak{b}] \cap (-\infty, -\frac{\mathfrak{b}^{\theta_n}}{\mathfrak{w}^{\theta_n}}] & \colon \mathfrak{w}^{\theta_n} > 0 \\ [\mathfrak{a}, \mathfrak{b}] \cap [-\frac{\mathfrak{b}^{\theta_n}}{\mathfrak{w}^{\theta_n}}, \infty) & \colon \mathfrak{w}^{\theta_n} < 0. \end{cases} \end{aligned} \tag{136}$$

Lemma 5.1 hence proves that for all $n \in \mathbb{N}$ it holds that

$$\inf_{\xi \in \mathbb{R}} \left[\rho \int_{[\mathfrak{a},\mathfrak{b}] \setminus I^{\theta_n}} (f(x) - \xi)^2 \, \mathrm{d}x \right] = \inf_{\xi \in [-M, M]} \left[\rho \int_{[\mathfrak{a},\mathfrak{b}] \setminus I^{\theta_n}} (f(x) - \xi)^2 \, \mathrm{d}x \right]. \tag{137}$$

This, (135), (136), and Lemma 5.1 demonstrate for all $n \in \mathbb{N}$ that

$$\begin{aligned} \mathcal{L}(\theta_n) &\geq \inf_{\xi \in [-M, M]} \left[\rho \int_{[\mathfrak{a},\mathfrak{b}] \setminus I^{\theta_n}} (f(x) - \xi)^2 \, \mathrm{d}x \right] \\ &= \inf_{\xi \in [-M, M]} \left[\rho \int_{[\mathfrak{a},\mathfrak{b}]} (f(x) - \xi)^2 \, \mathrm{d}x - \rho \int_{I^{\theta_n}} (f(x) - \xi)^2 \, \mathrm{d}x \right] \\ &\geq \inf_{\xi \in [-M, M]} \left[\rho \int_{[\mathfrak{a},\mathfrak{b}]} (f(x) - \xi)^2 \, \mathrm{d}x - \rho \int_{I^{\theta_n}} (|f(x)| + |\xi|)^2 \, \mathrm{d}x \right] \\ &\geq \inf_{\xi \in [-M, M]} \left[\rho \int_{[\mathfrak{a},\mathfrak{b}]} (f(x) - \xi)^2 \, \mathrm{d}x - \rho \int_{I^{\theta_n}} (2M)^2 \, \mathrm{d}x \right] \\ &\geq \left[\inf_{\xi \in [-M, M]} \rho \int_{\mathfrak{a}}^{\mathfrak{b}} (f(x) - \xi)^2 \, \mathrm{d}x \right] - 4\rho M^2 \lambda(I^{\theta_n}) \\ &= m - 4\rho M^2 \lambda(I^{\theta_n}). \end{aligned} \tag{138}$$

Combining this with (133) and (134) shows that

$$m > m - \varepsilon \geq \liminf_{n \to \infty} \mathcal{L}(\theta_n) \geq m. \tag{139}$$

This is a contradiction. The proof of Proposition 5.8 is thus complete. □

6 Convergence of the risk of GFs in the training of ANNs with one hidden neuron

The main result of this section, Theorem 6.7 in Subsection 6.3 below, demonstrates in the special situation where the measure μ (see Setting 2.1) is up to a constant the Lebesgue–Borel measure on $[\mathfrak{a}, \mathfrak{b}]$, where the hidden layer consists of only one neuron (where $H = 1$), and where the target function $f \colon [\mathfrak{a}, \mathfrak{b}] \to \mathbb{R}$ is affine linear that the risk of every not necessarily bounded GF trajectory converges to zero. Our proof of Theorem 6.7 employs some of the results in Sections 3 and 5, the a priori bounds for GF trajectories with sufficiently small initial risk

in Lemma 6.2 in Subsection 6.1 below, the convergence properties of ANNs with uniformly convergent realization functions in Lemma 6.4 in Subsection 6.2, and the well-known fact for integral equations in Lemma 6.6 in Subsection 6.3. Only for completeness we include in this section a detailed proof for Lemma 6.6.

In our proof of Theorem 6.7 we first employ Lemma 3.1 in Subsection 3.1 to obtain that $[0,\infty) \ni t \mapsto \mathcal{G}(\Theta_t) \in \mathbb{R}^4$ is L^2-integrable. This allows us to extract a subsequence along which the standard norm of the generalized gradient converges to zero. In the next step Lemma 6.2 enables us to conclude that the realization functions of the corresponding ANNs are uniformly equicontinuous. This, in turn, allows us to bring the Arzela-Ascoli theorem into play to obtain that along some sub-subsequence the realization functions converge uniformly on $[\mathscr{a},\mathscr{b}]$. It then remains to prove that the limit of these uniformly convergent ANN realization functions coincides with the affine linear target function. We verify this by employing Lemma 6.4 in combination with a careful analysis of the gradient given by (158).

As a consequence of Theorem 6.7, we prove in Corollary 6.9 in the special situation where the measure μ (see Setting 2.1) is up to a constant the Lebesgue–Borel measure on $[\mathscr{a},\mathscr{b}]$, where the hidden layer consists of only one neuron (where $H = 1$), and where the target function $f\colon [\mathscr{a},\mathscr{b}] \to \mathbb{R}$ is affine linear that the realization functions of the GF trajectory converge to the target function not only in L^2-sense (Theorem 6.7) but even uniformly in the set of all continuous functions $C([\mathscr{a},\mathscr{b}],\mathbb{R})$ from $[\mathscr{a},\mathscr{b}]$ to $\mathbb{R}$.

Our formulations of the statements in Lemma 6.2, Corollary 6.3, Theorem 6.7, and Corollary 6.9 also exploit the elementary regularity result in Lemma 6.1 in Subsection 6.1. Lemma 6.1 clarifies in the framework of Setting 5.3 that the generalized gradient function $\mathcal{G}\colon \mathbb{R}^4 \to \mathbb{R}^4$ is locally bounded and measurable and, thereby, in particular ensures for every continuous function $\Theta\colon [0,\infty) \to \mathbb{R}^4$ and every $t \in [0,\infty)$ that the Lebesgue integral $\int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s$ is well-defined. Lemma 6.1 is an immediate consequence of the more general result in Corollary 2.4 from Subsection 2.2 above.

6.1 A priori estimates for GFs

Lemma 6.1. *Assume Setting 5.3. Then it holds that $\mathcal{G}$ is locally bounded and measurable.*

Proof of Lemma 6.1. Note that Corollary 2.4 establishes that $\mathcal{G}$ is locally bounded and measurable. The proof of Lemma 6.1 is thus complete. □

Lemma 6.2. *Assume Setting 5.3, let $\Theta \in C([0,\infty),\mathbb{R}^4)$ satisfy for all $t \in [0,\infty)$ that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s$, let $m \in \mathbb{R}$ satisfy $m = \rho\int_{\mathscr{a}}^{\mathscr{b}}(f(x) - (\mathscr{b}-\mathscr{a})^{-1}\int_{\mathscr{a}}^{\mathscr{b}} f(y)\,\mathrm{d}y)^2\,\mathrm{d}x$, and assume $\mathcal{L}(\Theta_0) < m$ (cf. Lemma 6.1). Then*

(i) it holds that $\sup_{t\in[0,\infty)}|\mathfrak{w}^{\Theta_t}\mathfrak{v}^{\Theta_t}| < \infty$,

(ii) it holds that

$$\sup_{t\in[0,\infty)}|\mathfrak{w}^{\Theta_t}| \leq \big[\sup_{t\in[0,\infty)} \max\{1, |\mathfrak{w}^{\Theta_0}|^2 + |\mathfrak{b}^{\Theta_0}|^2 - |\mathfrak{v}^{\Theta_0}|^2 + |\mathfrak{w}^{\Theta_t}\mathfrak{v}^{\Theta_t}|^2\}\big]^{1/2} < \infty, \tag{140}$$

(iii) it holds for all $t \in [0,\infty)$ that

$$\sup_{x\in[\mathscr{a},\mathscr{b}]}|\mathcal{N}^{\Theta_t}(x)| \leq 2\big[\sup_{x\in[\mathscr{a},\mathscr{b}]}|f(x)|\big] + (\mathscr{b}-\mathscr{a})|\mathfrak{w}^{\Theta_t}\mathfrak{v}^{\Theta_t}| < \infty, \tag{141}$$

and

(iv) it holds for all $\alpha,\beta \in \mathbb{R}$ with $\forall\, x \in [\mathscr{a},\mathscr{b}]\colon f(x) = \alpha x + \beta$ that $\inf_{t\in[0,\infty)} \alpha\mathfrak{w}^{\Theta_t}\mathfrak{v}^{\Theta_t} > 0$.

Proof of Lemma 6.2. Throughout this proof let $w = (w_t)_{t\in[0,\infty)}$, $b = (b_t)_{t\in[0,\infty)}$, $v = (v_t)_{t\in[0,\infty)}$, $c = (c_t)_{t\in[0,\infty)} \in C([0,\infty),\mathbb{R})$ satisfy for all $t \in [0,\infty)$ that

$$w_t = \mathfrak{w}^{\Theta_t}, \qquad b_t = \mathfrak{b}^{\Theta_t}, \qquad v_t = \mathfrak{v}^{\Theta_t}, \qquad \text{and} \qquad c_t = \mathfrak{c}^{\Theta_t}, \tag{142}$$

let $M \in \mathbb{R}$ satisfy $M = \sup_{x\in[\mathscr{a},\mathscr{b}]}|f(x)|$, let $A \in \mathbb{R}$ satisfy $A = |w_0|^2 + |b_0|^2 - |v_0|^2$, and let $\mathfrak{C} \in [0,\infty]$ satisfy $\mathfrak{C} = \sup_{t\in[0,\infty)}|w_t v_t|$. Observe that Lemma 3.1 demonstrates for all $t \in [0,\infty)$ that $\mathcal{L}(\Theta_t) \leq \mathcal{L}(\Theta_0) < m$. Corollary 5.2, Corollary 5.6, and Proposition 5.8 hence establish items (i) and (iv).

In the next step we prove item (ii). Note that Proposition 4.4 implies for all $t \in [0,\infty)$ that

$$|w_t|^2 - |v_t|^2 \leq |w_t|^2 + |b_t|^2 - |v_t|^2 = |w_0|^2 + |b_0|^2 - |v_0|^2 = A. \tag{143}$$

Combining this with the fact that $\sup_{t\in[0,\infty)}|w_t v_t| = \mathfrak{C} < \infty$ ensures for all $t \in [0,\infty)$ with $|w_t| \geq 1$ that

$$|w_t|^2 \leq A + |v_t|^2 \leq A + \tfrac{\mathfrak{C}^2}{|w_t|^2} \leq A + \mathfrak{C}^2. \tag{144}$$

Hence, we obtain for all $t \in [0,\infty)$ that $|w_t|^2 \leq \max\{A + \mathfrak{C}^2, 1\} < \infty$. This establishes item (ii).

Finally, we prove item (iii). Observe that Lemma 5.1 implies that $m \leq \rho\int_{\mathscr{a}}^{\mathscr{b}}(f(y))^2\,\mathrm{d}y \leq \rho(\mathscr{b}-\mathscr{a})M^2$. Combining this with Lemma 3.1 assures that for all $t \in [0,\infty)$ we have that

$$\rho\int_{\mathscr{a}}^{\mathscr{b}}(\mathscr{N}^{\Theta_t}(y) - f(y))^2\,\mathrm{d}y = \mathcal{L}(\Theta_t) \leq \mathcal{L}(\Theta_0) \leq m \leq \rho(\mathscr{b}-\mathscr{a})M^2. \tag{145}$$

This shows that there exists $x = (x_t)_{t\in[0,\infty)}\colon [0,\infty) \to [\mathscr{a},\mathscr{b}]$ which satisfies for all $t \in [0,\infty)$ that

$$|\mathscr{N}^{\Theta_t}(x_t) - f(x_t)| \leq M. \tag{146}$$

In addition, note that Lemma 5.4 ensures that for all $t \in [0,\infty)$, $x, y \in [\mathscr{a},\mathscr{b}]$ it holds that $|\mathscr{N}^{\Theta_t}(x) - \mathscr{N}^{\Theta_t}(y)| \leq |w_t v_t||x-y|$. Hence, we obtain for all $t \in [0,\infty)$, $y \in [\mathscr{a},\mathscr{b}]$ that

$$\begin{aligned}|\mathscr{N}^{\Theta_t}(y)| &\leq |\mathscr{N}^{\Theta_t}(x_t)| + |\mathscr{N}^{\Theta_t}(y) - \mathscr{N}^{\Theta_t}(x_t)| \\ &\leq |f(x_t)| + |\mathscr{N}^{\Theta_t}(x_t) - f(x_t)| + |w_t v_t||y - x_t| \\ &\leq M + M + |w_t v_t|(\mathscr{b}-\mathscr{a}) = 2M + |w_t v_t|(\mathscr{b}-\mathscr{a}).\end{aligned} \tag{147}$$

This establishes item (iii). The proof of Lemma 6.2 is thus complete. □

Corollary 6.3. *Assume Setting 5.3 and let $\Theta \in C([0,\infty),\mathbb{R}^4)$ satisfy for all $t \in [0,\infty)$ that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s$ (cf. Lemma 6.1). Then*

(i) it holds that $\sup_{t\in[0,\infty)}|\mathfrak{w}^{\Theta_t}\mathfrak{v}^{\Theta_t}| < \infty$ and

(ii) it holds that $\sup_{t\in[0,\infty)}|\mathfrak{w}^{\Theta_t}| < \infty$.

Proof of Corollary 6.3. Throughout this proof let $m \in \mathbb{R}$ satisfy

$$m = \rho\int_{\mathscr{a}}^{\mathscr{b}}\big(f(x) - (\mathscr{b}-\mathscr{a})^{-1}\int_{\mathscr{a}}^{\mathscr{b}} f(y)\,\mathrm{d}y\big)^2\,\mathrm{d}x. \tag{148}$$

In the following we distinguish between the case $\inf_{t\in[0,\infty)}\mathcal{L}(\Theta_t) \geq m$ and the case $\inf_{t\in[0,\infty)}\mathcal{L}(\Theta_t) < m$. We first establish items (i) and (ii) in the case

$$\inf_{t\in[0,\infty)}\mathcal{L}(\Theta_t) \geq m. \tag{149}$$

Observe that (149) and Corollary 4.3 show that

$$\sup_{t\in[0,\infty)}\|\Theta_t\| \leq 3\|\Theta_0\|^2 + 8\big|(\mathscr{b}-\mathscr{a})^{-1}\int_{\mathscr{a}}^{\mathscr{b}} f(y)\,\mathrm{d}y\big|^2 < \infty. \tag{150}$$

This establishes items (i) and (ii) in the case $\inf_{t\in[0,\infty)} \mathcal{L}(\Theta_t) \geq m$. In the next step we prove items (i) and (ii) in the case

$$\inf_{t\in[0,\infty)} \mathcal{L}(\Theta_t) < m. \tag{151}$$

Note that (151) assures that there exists $T \in [0,\infty)$ which satisfies that $\mathcal{L}(\Theta_T) < m$. Observe that the fact that that $\Theta\colon [0,\infty) \to \mathbb{R}^4$ is continuous implies that $\sup_{t\in[0,T]}|\mathfrak{w}^{\Theta_t}\mathfrak{v}^{\Theta_t}| < \infty$ and $\sup_{t\in[0,T]}|\mathfrak{w}^{\Theta_t}| < \infty$. Next let $\mathit{\Theta} \in C([0,\infty),\mathbb{R}^4)$ satisfy for all $t \in [0,\infty)$ that $\mathit{\Theta}_t = \Theta_{T+t}$. Note that the integral transformation theorem ensures for all $t \in [0,\infty)$ that $\mathcal{L}(\mathit{\Theta}_0) = \mathcal{L}(\Theta_T) < m$ and

$$\begin{aligned}\mathit{\Theta}_t = \Theta_{T+t} &= \Theta_0 - \int_0^{T+t} \mathcal{G}(\Theta_s)\,\mathrm{d}s = \left[\Theta_0 - \int_0^T \mathcal{G}(\Theta_s)\,\mathrm{d}s\right] - \int_T^{T+t} \mathcal{G}(\Theta_s)\,\mathrm{d}s \\ &= \Theta_T - \int_0^t \mathcal{G}(\Theta_{T+s})\,\mathrm{d}s = \mathit{\Theta}_0 - \int_0^t \mathcal{G}(\mathit{\Theta}_s)\,\mathrm{d}s.\end{aligned} \tag{152}$$

Lemma 6.2 hence proves that $\sup_{t\in[T,\infty)}|\mathfrak{w}^{\Theta_t}\mathfrak{v}^{\Theta_t}| = \sup_{t\in[0,\infty)}|\mathfrak{w}^{\mathit{\Theta}_t}\mathfrak{v}^{\mathit{\Theta}_t}| < \infty$ and $\sup_{t\in[T,\infty)}|\mathfrak{w}^{\Theta_t}| = \sup_{t\in[0,\infty)}|\mathfrak{w}^{\mathit{\Theta}_t}| < \infty$. This establishes items (i) and (ii) in the case $\inf_{t\in[0,\infty)} \mathcal{L}(\Theta_t) < m$. The proof of Corollary 6.3 is thus complete. □

6.2 Properties of ANN parameters for convergent sequences of ANN realizations

Lemma 6.4. *Assume Setting 5.3, let $(\theta_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^4$, $h \in C([\mathfrak{a},\mathfrak{b}],\mathbb{R})$ satisfy*

$$\limsup_{n\to\infty} \sup_{x\in[\mathfrak{a},\mathfrak{b}]} |\mathcal{N}^{\theta_n}(x) - h(x)| = 0, \tag{153}$$

and assume that h is not constant. Then

(i) there exists $\vartheta \in \mathbb{R}^4$ which satisfies $\mathcal{N}^{\vartheta}|_{[\mathfrak{a},\mathfrak{b}]} = h$,

(ii) it holds that $\limsup_{n\to\infty}|\mathfrak{w}^{\theta_n}\mathfrak{v}^{\theta_n} - \mathfrak{w}^{\vartheta}\mathfrak{v}^{\vartheta}| = 0$, and

(iii) it holds that $\limsup_{n\to\infty} \lambda(I^{\theta_n}\Delta I^{\vartheta}) = 0$.

Proof of Lemma 6.4. Observe that [20, Theorem 3.8] ensures that there exists $\vartheta \in \mathbb{R}^4$ which satisfies $\mathcal{N}^{\vartheta}|_{[\mathfrak{a},\mathfrak{b}]} = h$. This establishes item (i).

In the next step we prove that $\limsup_{n\to\infty} \lambda(I^{\vartheta}\backslash I^{\theta_n}) = 0$. Note that the assumption that h is not constant implies that $\lambda(I^{\vartheta}) > 0$ and $\mathfrak{w}^{\vartheta}\mathfrak{v}^{\vartheta} \neq 0$. Moreover, observe that (106) ensures that for all $n \in \mathbb{N}$, $x \in [\mathfrak{a},\mathfrak{b}]\backslash I^{\theta_n}$ it holds that $\mathcal{N}^{\theta_n}(x) = \mathfrak{c}^{\theta_n}$. This, the fact that for all $x \in I^{\vartheta}$ it holds that $\mathcal{N}^{\vartheta}(x) = \mathfrak{w}^{\vartheta}\mathfrak{v}^{\vartheta}x + \mathfrak{v}^{\vartheta}\mathfrak{b}^{\vartheta} + \mathfrak{c}^{\vartheta}$, and Corollary 5.2 imply that for all $n \in \mathbb{N}$ we have that

$$\int_{\mathfrak{a}}^{\mathfrak{b}} |\mathcal{N}^{\theta_n}(x) - \mathcal{N}^{\vartheta}(x)|^2\,\mathrm{d}x \geq \int_{I^{\vartheta}\backslash I^{\theta_n}} (h(x) - \mathfrak{c}^{\theta_n})^2\,\mathrm{d}x \geq \frac{|\mathfrak{w}^{\vartheta}\mathfrak{v}^{\vartheta}|^2(\lambda(I^{\vartheta}\backslash I^{\theta_n}))^3}{12}. \tag{154}$$

Furthermore, note that (153) assures that

$$\limsup_{n\to\infty}\left[\int_{\mathfrak{a}}^{\mathfrak{b}} |\mathcal{N}^{\theta_n}(x) - \mathcal{N}^{\vartheta}(x)|^2\,\mathrm{d}x\right] = 0. \tag{155}$$

This, (154), and the fact that $\mathfrak{w}^{\vartheta}\mathfrak{v}^{\vartheta} \neq 0$ demonstrate that $\limsup_{n\to\infty} \lambda(I^{\vartheta}\backslash I^{\theta_n}) = 0$. Hence, we have that $\limsup_{n\to\infty}|\lambda(I^{\vartheta} \cap I^{\theta_n}) - \lambda(I^{\vartheta})| = 0$. Next observe that (106) shows that for all

$n \in \mathbb{N}$, $x \in I^{\vartheta} \cap I^{\theta_n}$ it holds that $\mathcal{N}^{\theta_n}(x) - \mathcal{N}^{\vartheta}(x) = (\mathfrak{w}^{\theta_n}\mathfrak{v}^{\theta_n} - \mathfrak{w}^{\vartheta}\mathfrak{v}^{\vartheta})x + (\mathfrak{v}^{\theta_n}\mathfrak{b}^{\theta_n} + \mathfrak{c}^{\theta_n} - \mathfrak{v}^{\vartheta}\mathfrak{b}^{\vartheta} - \mathfrak{c}^{\vartheta})$. Combining this and Corollary 5.2 proves for all $n \in \mathbb{N}$ that

$$\begin{aligned} \int_{\mathscr{a}}^{\mathscr{b}} |\mathcal{N}^{\theta_n}(x) - \mathcal{N}^{\vartheta}(x)|^2 \, \mathrm{d}x &\geq \int_{I^{\vartheta} \cap I^{\theta_n}} |\mathcal{N}^{\theta_n}(x) - \mathcal{N}^{\vartheta}(x)|^2 \, \mathrm{d}x \\ &\geq \frac{|\mathfrak{w}^{\theta_n}\mathfrak{v}^{\theta_n} - \mathfrak{w}^{\vartheta}\mathfrak{v}^{\vartheta}|^2 (\lambda(I^{\vartheta} \cap I^{\theta_n}))^3}{12}. \end{aligned} \tag{156}$$

This, (155), and the fact that $\liminf_{n\to\infty} \lambda(I^{\vartheta} \cap I^{\theta_n}) = \lambda(I^{\vartheta}) > 0$ ensure that $\limsup_{n\to\infty} |\mathfrak{w}^{\theta_n}\mathfrak{v}^{\theta_n} - \mathfrak{w}^{\vartheta}\mathfrak{v}^{\vartheta}| = 0$, which establishes item (ii).

It remains to prove that $\limsup_{n\to\infty} \lambda(I^{\theta_n} \backslash I^{\vartheta}) = 0$. Note that (106) implies that for all $x \in [\mathscr{a}, \mathscr{b}] \backslash I^{\vartheta}$ it holds that $\mathcal{N}^{\vartheta}(x) = \mathfrak{c}^{\vartheta}$. This, the fact that for all $n \in \mathbb{N}$, $x \in I^{\theta_n}$ we have that $\mathcal{N}^{\theta_n}(x) = \mathfrak{w}^{\theta_n}\mathfrak{v}^{\theta_n} x + \mathfrak{v}^{\theta_n}\mathfrak{b}^{\theta_n} + \mathfrak{c}^{\theta_n}$, and Corollary 5.2 show that for all $n \in \mathbb{N}$ it holds that

$$\int_{\mathscr{a}}^{\mathscr{b}} |\mathcal{N}^{\theta_n}(x) - \mathcal{N}^{\vartheta}(x)|^2 \, \mathrm{d}x \geq \int_{I^{\theta_n} \backslash I^{\vartheta}} |\mathcal{N}^{\theta_n}(x) - \mathfrak{c}^{\vartheta}|^2 \, \mathrm{d}x \geq \frac{|\mathfrak{w}^{\theta_n}\mathfrak{v}^{\theta_n}|^2 (\lambda(I^{\theta_n} \backslash I^{\vartheta}))^3}{12}. \tag{157}$$

Combining this and (155) with the fact that $\liminf_{n\to\infty} \mathfrak{w}^{\theta_n}\mathfrak{v}^{\theta_n} = \mathfrak{w}^{\vartheta}\mathfrak{v}^{\vartheta} \neq 0$ demonstrates that $\limsup_{n\to\infty} \lambda(I^{\theta_n} \backslash I^{\vartheta}) = 0$. This proves item (iii). The proof of Lemma 6.4 is thus complete. □

6.3 Convergence of the risk of GFs to zero for affine linear target functions

Proposition 6.5. *Assume Setting 5.3 and let $\theta \in \mathbb{R}^{\mathfrak{d}}$. Then*

$$\begin{aligned} \mathcal{G}_1(\theta) &= 2\rho\mathfrak{v}^{\theta} \int_{I^{\theta}} x(\mathcal{N}^{\theta}(x) - f(x)) \, \mathrm{d}x, \\ \mathcal{G}_2(\theta) &= 2\rho\mathfrak{v}^{\theta} \int_{I^{\theta}} (\mathcal{N}^{\theta}(x) - f(x)) \, \mathrm{d}x, \\ \mathcal{G}_3(\theta) &= 2\rho \int_{\mathscr{a}}^{\mathscr{b}} \big[\max\{\mathfrak{w}^{\theta}x + \mathfrak{b}^{\theta}, 0\}\big](\mathcal{N}^{\theta}(x) - f(x)) \, \mathrm{d}x, \\ \text{and} \qquad \mathcal{G}_4(\theta) &= 2\rho \int_{\mathscr{a}}^{\mathscr{b}} (\mathcal{N}^{\theta}(x) - f(x)) \, \mathrm{d}x. \end{aligned} \tag{158}$$

Proof of Proposition 6.5. Observe that Proposition 2.2 establishes (158). The proof of Proposition 6.5 is thus complete. □

Lemma 6.6. *Let $\mathscr{a} \in \mathbb{R}$, $\mathscr{b} \in (\mathscr{a}, \infty)$, $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{R}$ satisfy*

$$\int_{\mathscr{a}}^{\mathscr{b}} x((\alpha_1 x + \beta_1) - (\alpha_2 x + \beta_2)) \, \mathrm{d}x = \int_{\mathscr{a}}^{\mathscr{b}} ((\alpha_1 x + \beta_1) - (\alpha_2 x + \beta_2)) \, \mathrm{d}x = 0. \tag{159}$$

Then $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$.

Proof of Lemma 6.6. Note that (159) assures that

$$\begin{aligned} 0 &= (\alpha_1 - \alpha_2) \left[\int_{\mathscr{a}}^{\mathscr{b}} x((\alpha_1 x + \beta_1) - (\alpha_2 x + \beta_2)) \, \mathrm{d}x \right] \\ &\quad + (\beta_1 - \beta_2) \left[\int_{\mathscr{a}}^{\mathscr{b}} ((\alpha_1 x + \beta_1) - (\alpha_2 x + \beta_2)) \, \mathrm{d}x \right] \\ &= \int_{\mathscr{a}}^{\mathscr{b}} ((\alpha_1 - \alpha_2)x + (\beta_1 - \beta_2))((\alpha_1 x + \beta_1) - (\alpha_2 x + \beta_2)) \, \mathrm{d}x \\ &= \int_{\mathscr{a}}^{\mathscr{b}} ((\alpha_1 - \alpha_2)x + (\beta_1 - \beta_2))^2 \, \mathrm{d}x. \end{aligned} \tag{160}$$

This and the fact that for all $x \in [\mathscr{a}, \mathscr{b}]$ it holds that $((\alpha_1 - \alpha_2)x + (\beta_1 - \beta_2))^2 \geq 0$ show that for all $x \in [\mathscr{a}, \mathscr{b}]$ it holds that $((\alpha_1 - \alpha_2)x + (\beta_1 - \beta_2)) = 0$. Hence, we obtain that $\alpha_1 - \alpha_2 = \beta_1 - \beta_2 = 0$. The proof of Lemma 6.6 is thus complete. □

Theorem 6.7. *Assume Setting 5.3, let $\alpha, \beta \in \mathbb{R}$ satisfy for all $x \in [\mathscr{a}, \mathscr{b}]$ that $f(x) = \alpha x + \beta$, and let $\Theta \in C([0, \infty), \mathbb{R}^4)$ satisfy for all $t \in [0, \infty)$ that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\, \mathrm{d}s$ and $\mathcal{L}(\Theta_0) < \frac{\rho\alpha^2(\mathscr{b}-\mathscr{a})^3}{12}$ (cf. Lemma 6.1). Then $\limsup_{t\to\infty} \mathcal{L}(\Theta_t) = 0$.*

Proof of Theorem 6.7. Throughout this proof let $w = (w_t)_{t\in[0,\infty)}$, $b = (b_t)_{t\in[0,\infty)}$, $v = (v_t)_{t\in[0,\infty)}$, $c = (c_t)_{t\in[0,\infty)} \in C([0, \infty), \mathbb{R})$ satisfy for all $t \in [0, \infty)$ that

$$w_t = \mathfrak{w}^{\Theta_t}, \qquad b_t = \mathfrak{b}^{\Theta_t}, \qquad v_t = \mathfrak{v}^{\Theta_t}, \qquad \text{and} \qquad c_t = \mathfrak{c}^{\Theta_t} \tag{161}$$

and let $\mathcal{I}_t \subseteq [\mathscr{a}, \mathscr{b}]$, $t \in [0, \infty)$, satisfy for all $t \in [0, \infty)$ that $\mathcal{I}_t = I^{\Theta_t}$. Observe that Lemma 3.1 implies that $[0, \infty) \ni t \mapsto \mathcal{L}(\Theta_t) \in \mathbb{R}$ is non-increasing. Hence, we obtain that

$$\limsup_{t\to\infty} \mathcal{L}(\Theta_t) = \liminf_{t\to\infty} \mathcal{L}(\Theta_t) = \inf_{t\in[0,\infty)} \mathcal{L}(\Theta_t). \tag{162}$$

Next note that Lemma 3.1 proves that $\int_0^\infty \|\mathcal{G}(\Theta_s)\|^2\, \mathrm{d}s < \infty$. This demonstrates that $\liminf_{t\to\infty} \|\mathcal{G}(\Theta_t)\| = 0$. Therefore, we obtain that there exist $\tau_n \in [0, \infty)$, $n \in \mathbb{N}$, which satisfy $\liminf_{n\to\infty} \tau_n = \infty$ and $\limsup_{n\to\infty} \|\mathcal{G}(\Theta_{\tau_n})\| = 0$. Observe that Lemma 6.2 implies that

$$\sup_{n\in\mathbb{N}} |w_{\tau_n} v_{\tau_n}| < \infty \qquad \text{and} \qquad \sup_{n\in\mathbb{N}} \sup_{x\in[\mathscr{a},\mathscr{b}]} |\mathcal{N}^{\Theta_{\tau_n}}(x)| < \infty. \tag{163}$$

Combining this and Lemma 5.4 proves that there exists $\mathfrak{C} \in \mathbb{R}$ such that for all $x, y \in [\mathscr{a}, \mathscr{b}]$, $n \in \mathbb{N}$ it holds that $|\mathcal{N}^{\Theta_{\tau_n}}(x) - \mathcal{N}^{\Theta_{\tau_n}}(y)| \leq \mathfrak{C}|x - y|$ and $|\mathcal{N}^{\Theta_{\tau_n}}(x)| \leq \mathfrak{C}$. The Arzela-Ascoli theorem hence shows that there exist $h \in C([\mathscr{a}, \mathscr{b}], \mathbb{R})$ and a strictly increasing $k\colon \mathbb{N} \to \mathbb{N}$ which satisfy

$$\limsup_{n\to\infty} \sup_{x\in[\mathscr{a},\mathscr{b}]} |\mathcal{N}^{\Theta_{\tau_{k(n)}}}(x) - h(x)| = 0. \tag{164}$$

Combining this with (162) and the assumption that $\mathcal{L}(\Theta_0) < \frac{\rho\alpha^2(\mathscr{b}-\mathscr{a})^3}{12}$ implies that

$$\rho \int_{\mathscr{a}}^{\mathscr{b}} (f(x) - h(x))^2\, \mathrm{d}x = \limsup_{n\to\infty} \mathcal{L}(\Theta_{\tau_{k(n)}}) = \inf_{t\in[0,\infty)} \mathcal{L}(\Theta_t) < \frac{\rho\alpha^2(\mathscr{b} - \mathscr{a})^3}{12}. \tag{165}$$

This and Corollary 5.2 assure that h is not constant. Lemma 6.4 hence ensures that there exists $\vartheta \in \mathbb{R}^4$ which satisfies $\mathcal{N}^\vartheta|_{[\mathscr{a},\mathscr{b}]} = h$. Combining this and (165) with Corollary 5.2, Corollary 5.6, and Lemma 5.7 demonstrates that $\alpha\mathfrak{w}^\vartheta\mathfrak{v}^\vartheta > 0$ and $I^\vartheta \neq \varnothing$. In addition, note that (158) and (164) show that

$$\begin{aligned} 0 &= \frac{1}{2\rho}\Big[\lim_{n\to\infty} \mathcal{G}_4(\Theta_{\tau_{k(n)}})\Big] = \lim_{n\to\infty}\left[\int_{\mathscr{a}}^{\mathscr{b}} (\mathcal{N}^{\Theta_{\tau_{k(n)}}}(x) - (\alpha x + \beta))\, \mathrm{d}x\right] \\ &= \int_{\mathscr{a}}^{\mathscr{b}} (\mathcal{N}^\vartheta(x) - (\alpha x + \beta))\, \mathrm{d}x. \end{aligned} \tag{166}$$

Furthermore, observe that (164) and Lemma 6.4 prove that $\limsup_{n\to\infty} \lambda(\mathcal{I}_{\tau_{k(n)}} \Delta I^\vartheta) = 0$. Combining this and the fact that $\limsup_{n\to\infty} \sup_{x\in[\mathscr{a},\mathscr{b}]} |\mathcal{N}^{\Theta_{\tau_{k(n)}}}(x) - \mathcal{N}^\vartheta(x)| = 0$ demonstrates that

$$\limsup_{n\to\infty} \left| \int_{\mathcal{I}_{\tau_{k(n)}}} x(\mathcal{N}^{\Theta_{\tau_{k(n)}}}(x) - (\alpha x + \beta))\, \mathrm{d}x - \int_{I^\vartheta} x(\mathcal{N}^\vartheta(x) - (\alpha x + \beta))\, \mathrm{d}x \right| = 0 \tag{167}$$

and

$$\limsup_{n\to\infty} \left| \int_{\mathcal{I}_{\tau_{k(n)}}} (\mathcal{N}^{\Theta_{\tau_{k(n)}}}(x) - (\alpha x + \beta))\, \mathrm{d}x - \int_{I^\vartheta} (\mathcal{N}^\vartheta(x) - (\alpha x + \beta))\, \mathrm{d}x \right| = 0. \tag{168}$$

Moreover, note that the fact that $\limsup_{n\to\infty}\|\mathcal{G}(\Theta_{\tau_{k(n)}})\| = 0$ and (158) imply that

$$\begin{aligned}&\limsup_{n\to\infty}\left|v_{\tau_{k(n)}}\int_{\mathcal{I}_{\tau_{k(n)}}} x(\mathcal{N}^{\Theta_{\tau_{k(n)}}}(x)-(\alpha x+\beta))\,\mathrm{d}x\right|\\ &=\limsup_{n\to\infty}\left|v_{\tau_{k(n)}}\int_{\mathcal{I}_{\tau_{k(n)}}} (\mathcal{N}^{\Theta_{\tau_{k(n)}}}(x)-(\alpha x+\beta))\,\mathrm{d}x\right| = 0.\end{aligned} \tag{169}$$

In the next step we show that

$$\left|\int_{I^\vartheta} x(\mathcal{N}^\vartheta(x)-(\alpha x+\beta))\,\mathrm{d}x\right| = \left|\int_{I^\vartheta} (\mathcal{N}^\vartheta(x)-(\alpha x+\beta))\,\mathrm{d}x\right| = 0. \tag{170}$$

We prove (170) by contradiction. We thus assume that

$$\left|\int_{I^\vartheta} x(\mathcal{N}^\vartheta(x)-(\alpha x+\beta))\,\mathrm{d}x\right| + \left|\int_{I^\vartheta} (\mathcal{N}^\vartheta(x)-(\alpha x+\beta))\,\mathrm{d}x\right| > 0. \tag{171}$$

Observe that (167)–(169) and (171) prove that $\limsup_{n\to\infty}|v_{\tau_{k(n)}}| = 0$. In addition, note that Lemma 6.4 assures that $\lim_{n\to\infty}(w_{\tau_{k(n)}}v_{\tau_{k(n)}}) = \mathfrak{w}^\vartheta\mathfrak{v}^\vartheta \neq 0$. Combining this with item (ii) in Lemma 6.2 demonstrates that $\infty = \liminf_{n\to\infty}|w_{\tau_{k(n)}}| < \infty$. This contradiction establishes (170). Next observe that for all $x \in I^\vartheta$ it holds that $\mathcal{N}^\vartheta(x) = \mathfrak{w}^\vartheta\mathfrak{v}^\vartheta x + \mathfrak{v}^\vartheta\mathfrak{b}^\vartheta + \mathfrak{c}^\vartheta$. Combining this, (170), and Lemma 6.6 ensures that for all $x \in I^\vartheta$ it holds that

$$\mathcal{N}^\vartheta(x) = \alpha x + \beta. \tag{172}$$

Note that for all $q \in (\mathscr{a},\mathscr{b})$ with $I^\vartheta = (q,\mathscr{b}]$ it holds that $\forall\, x \in [\mathscr{a},q]\colon \mathcal{N}^\vartheta(x) = \mathcal{N}^\vartheta(q) = \alpha q+\beta$. This, (166), and (170) imply that for all $q \in (\mathscr{a},\mathscr{b})$ with $I^\vartheta = (q,\mathscr{b}]$ we have that

$$\begin{aligned}0 &= \int_{\mathscr{a}}^{\mathscr{b}} (\mathcal{N}^\vartheta(x)-(\alpha x+\beta))\,\mathrm{d}x = \int_{\mathscr{a}}^{q} (\mathcal{N}^\vartheta(x)-(\alpha x+\beta))\,\mathrm{d}x\\ &= \int_{\mathscr{a}}^{q} (\alpha q-\alpha x)\,\mathrm{d}x = \alpha\int_{\mathscr{a}}^{q}(q-x)\,\mathrm{d}x = \frac{\alpha(q-\mathscr{a})^2}{2} \neq 0.\end{aligned} \tag{173}$$

Furthermore, observe that for all $q \in (\mathscr{a},\mathscr{b})$ with $I^\vartheta = [\mathscr{a},q)$ we have that $\forall\, x \in [q,\mathscr{b}]\colon \mathcal{N}^\vartheta(x) = \mathcal{N}^\vartheta(q) = \alpha q+\beta$. This, (166), and (170) ensure that for all $q \in (\mathscr{a},\mathscr{b})$ with $I^\vartheta = [\mathscr{a},q)$ it holds that

$$\begin{aligned}0 &= \int_{\mathscr{a}}^{\mathscr{b}} (\mathcal{N}^\vartheta(x)-(\alpha x+\beta))\,\mathrm{d}x = \int_{q}^{\mathscr{b}} (\mathcal{N}^\vartheta(x)-(\alpha x+\beta))\,\mathrm{d}x\\ &= \int_{q}^{\mathscr{b}} (\alpha q-\alpha x)\,\mathrm{d}x = \alpha\int_{q}^{\mathscr{b}}(q-x)\,\mathrm{d}x = -\frac{\alpha(\mathscr{b}-q)^2}{2} \neq 0.\end{aligned} \tag{174}$$

Combining this, (173), and the fact that $\lambda(I^\vartheta) > 0$ shows that $I^\vartheta \in \{[\mathscr{a},\mathscr{b}],(\mathscr{a},\mathscr{b}],[\mathscr{a},\mathscr{b})\}$. This implies that $(\mathscr{a},\mathscr{b}) \subseteq I^\vartheta$. Combining this with (172) assures that for all $x \in (\mathscr{a},\mathscr{b})$ we have that $\mathcal{N}^\vartheta(x) = \alpha x+\beta = f(x)$. Hence, we obtain that

$$\int_{\mathscr{a}}^{\mathscr{b}} (f(x)-h(x))^2\,\mathrm{d}x = \int_{\mathscr{a}}^{\mathscr{b}} (f(x)-\mathcal{N}^\vartheta(x))^2\,\mathrm{d}x = 0. \tag{175}$$

This, (162), and (165) imply that $\lim_{t\to\infty}\mathcal{L}(\Theta_t) = \mathcal{L}(\vartheta) = 0$. The proof of Theorem 6.7 is thus complete. □

Corollary 6.8. *Let $\alpha, \beta, \mathscr{a} \in \mathbb{R}$, $\mathscr{b} \in (\mathscr{a}, \infty)$, let $\mathfrak{R}_r \in C(\mathbb{R}, \mathbb{R})$, $r \in \mathbb{N} \cup \{\infty\}$, satisfy for all $x \in \mathbb{R}$ that $(\bigcup_{r \in \mathbb{N}} \{\mathfrak{R}_r\}) \subseteq C^1(\mathbb{R}, \mathbb{R})$, $\mathfrak{R}_\infty(x) = \max\{x, 0\}$, $\sup_{r \in \mathbb{N}} \sup_{y \in [-|x|, |x|]} (|\mathfrak{R}_r(y)| + |(\mathfrak{R}_r)'(y)|) < \infty$, and*

$$\limsup_{r \to \infty} \big(|\mathfrak{R}_r(x) - \mathfrak{R}_\infty(x)| + |(\mathfrak{R}_r)'(x) - \mathbb{1}_{(0,\infty)}(x)| \big) = 0, \tag{176}$$

let $\mathcal{L}_r \colon \mathbb{R}^4 \to \mathbb{R}$, $r \in \mathbb{N} \cup \{\infty\}$, satisfy for all $r \in \mathbb{N} \cup \{\infty\}$, $\theta = (\theta_1, \ldots, \theta_4) \in \mathbb{R}^4$ that

$$\mathcal{L}_r(\theta) = \int_{\mathscr{a}}^{\mathscr{b}} \big(\alpha x + \beta - \theta_4 - \theta_3 \mathfrak{R}_r(\theta_2 + \theta_1 x) \big)^2 \, \mathrm{d}x, \tag{177}$$

let $\mathcal{G} \colon \mathbb{R}^4 \to \mathbb{R}^4$ satisfy for all $\theta \in \{\vartheta \in \mathbb{R}^4 \colon ((\nabla \mathcal{L}_r)(\vartheta))_{r \in \mathbb{N}} \text{ is convergent}\}$ that $\mathcal{G}(\theta) = \lim_{r \to \infty} (\nabla \mathcal{L}_r)(\theta)$, let $\Theta \in C([0, \infty), \mathbb{R}^4)$ satisfy for all $t \in [0, \infty)$ that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) \, \mathrm{d}s$, and assume $\mathcal{L}_\infty(\Theta_0) < \frac{\alpha^2 (\mathscr{b} - \mathscr{a})^3}{12}$. Then $\limsup_{t \to \infty} \mathcal{L}_\infty(\Theta_t) = 0$.

Proof of Corollary 6.8. Note that Theorem 6.7 (applied with $\rho \curvearrowleft 1$ in the notation of Theorem 6.7) shows that $\limsup_{t \to \infty} \mathcal{L}_\infty(\Theta_t) = 0$. The proof of Corollary 6.8 is thus complete. □

6.4 Uniform convergence of realizations of GFs for affine linear target functions

Corollary 6.9. *Assume Setting 5.3, let $\alpha, \beta \in \mathbb{R}$ satisfy for all $x \in [\mathscr{a}, \mathscr{b}]$ that $f(x) = \alpha x + \beta$, and let $\Theta \in C([0, \infty), \mathbb{R}^4)$ satisfy for all $t \in [0, \infty)$ that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) \, \mathrm{d}s$ and $\mathcal{L}(\Theta_0) < \frac{\rho \alpha^2 (\mathscr{b} - \mathscr{a})^3}{12}$ (cf. Lemma 6.1). Then*

$$\limsup_{t \to \infty} \big(\sup_{x \in [\mathscr{a}, \mathscr{b}]} |\mathcal{N}^{\Theta_t}(x) - (\alpha x + \beta)| \big) = 0. \tag{178}$$

Proof of Corollary 6.9. Observe that Lemma 6.2 assures that there exists $\mathfrak{C} \in (0, \infty)$ such that for all $t \in [0, \infty)$ it holds that $|\mathfrak{w}^{\Theta_t} \mathfrak{v}^{\Theta_t}| \leq \mathfrak{C}$. We now prove (178) by contradiction. In the following we thus assume that

$$\limsup_{t \to \infty} \big(\sup_{x \in [\mathscr{a}, \mathscr{b}]} |\mathcal{N}^{\Theta_t}(x) - (\alpha x + \beta)| \big) > 0. \tag{179}$$

Note that (179) assures that there exist $\varepsilon \in (0, \infty)$ and $\tau_n \in [0, \infty)$, $n \in \mathbb{N}$, which satisfy $\liminf_{t \to \infty} \tau_n = \infty$ and

$$\inf_{n \in \mathbb{N}} \big(\sup_{x \in [\mathscr{a}, \mathscr{b}]} |\mathcal{N}^{\Theta_{\tau_n}}(x) - (\alpha x + \beta)| \big) > \varepsilon. \tag{180}$$

Observe that (180) shows that there exist $x_n \in [\mathscr{a}, \mathscr{b}]$, $n \in \mathbb{N}$, which satisfy for all $n \in \mathbb{N}$ that $|\mathcal{N}^{\Theta_{\tau_n}}(x_n) - (\alpha x_n + \beta)| \geq \varepsilon$. Moreover, note that Lemma 5.4 proves that for all $n \in \mathbb{N}$, $y, z \in [\mathscr{a}, \mathscr{b}]$ it holds that

$$\begin{aligned} |[\mathcal{N}^{\Theta_{\tau_n}}(y) - (\alpha y + \beta)] - [\mathcal{N}^{\Theta_{\tau_n}}(z) - (\alpha z + \beta)]| &\leq |\mathcal{N}^{\Theta_{\tau_n}}(y) - \mathcal{N}^{\Theta_{\tau_n}}(z)| + |\alpha| |y - z| \\ &\leq (\mathfrak{C} + |\alpha|) |y - z|. \end{aligned} \tag{181}$$

Next let $\delta \in (0, \infty)$ satisfy $\delta = \frac{\varepsilon}{2(\mathfrak{C} + |\alpha|)}$. Observe that (181) ensures that for all $n \in \mathbb{N}$, $y \in [x_n - \delta, x_n + \delta] \cap [\mathscr{a}, \mathscr{b}]$ it holds that

$$\begin{aligned} &|\mathcal{N}^{\Theta_{\tau_n}}(y) - (\alpha y + \beta)| \\ &\geq |\mathcal{N}^{\Theta_{\tau_n}}(x_n) - (\alpha x_n + \beta)| - |[\mathcal{N}^{\Theta_{\tau_n}}(x_n) - (\alpha x_n + \beta)] - [\mathcal{N}^{\Theta_{\tau_n}}(y) - (\alpha y + \beta)]| \\ &\geq \varepsilon - (\mathfrak{C} + |\alpha|) |x_n - y| \geq \varepsilon - (\mathfrak{C} + |\alpha|) \delta = \varepsilon - \tfrac{\varepsilon}{2} = \tfrac{\varepsilon}{2}. \end{aligned} \tag{182}$$

Furthermore, note that for all $n \in \mathbb{N}$ we have that $\lambda([x_n - \delta, x_n + \delta] \cap [\mathscr{a}, \mathscr{b}]) \geq \min\{\delta, \mathscr{b} - \mathscr{a}\}$. This demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\mathcal{L}(\Theta_{\tau_n}) \geq \rho \int_{[x_n - \delta, x_n + \delta] \cap [\mathscr{a}, \mathscr{b}]} |\mathcal{N}^{\Theta_{\tau_n}}(y) - (\alpha y + \beta)|^2 \, \mathrm{d}y \geq \frac{\rho \varepsilon^2 \min\{\delta, \mathscr{b} - \mathscr{a}\}}{4}. \tag{183}$$

Combining this with Theorem 6.7 shows that

$$0 = \limsup_{t \to \infty} \mathcal{L}(\Theta_t) \geq \limsup_{n \to \infty} \mathcal{L}(\Theta_{\tau_n}) \geq \tfrac{\rho \varepsilon^2 \min\{\delta, \mathscr{b} - \mathscr{a}\}}{4} > 0. \tag{184}$$

This is a contradiction. The proof of Corollary 6.9 is thus complete. □

Acknowledgements

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.

References

[1] P.-A. Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM J. Optim.*, 16(2):531–547, 2005. doi:10.1137/040605266.

[2] Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 02 2021. iaaa039. doi:10.1093/imaiai/iaaa039.

[3] Christian Beck, Sebastian Becker, Philipp Grohs, Nor Jaafari, and Arnulf Jentzen. Solving stochastic differential equations and Kolmogorov equations by means of deep learning, 2018. Accepted in Journal of Scientific Computing. arXiv:1806.00421.

[4] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.*, 17(4):1205–1223, 2006. doi:10.1137/050644641.

[5] Zhengdao Chen, Grant Rotskoff, Joan Bruna, and Eric Vanden-Eijnden. A dynamical central limit theorem for shallow neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22217–22230. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper/2020/file/fc5b3186f1cf0daece964f78259b7ba0-Paper.pdf.

[6] Patrick Cheridito, Arnulf Jentzen, Adrian Riekert, and Florian Rossmannek. A proof of convergence for gradient descent in the training of artificial neural networks for constant target functions, 2021. arXiv:2102.09924.

[7] Patrick Cheridito, Arnulf Jentzen, and Florian Rossmannek. Landscape analysis for shallow relu neural networks: complete classification of critical points for affine target functions, 2021. arXiv:2103.10922.

[8] Yacine Chitour, Zhenyu Liao, and Romain Couillet. A geometric approach of gradient descent algorithms in neural networks, 2019. arXiv:1811.03568.

[9] Lénaïc Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 2021. doi:10.1007/s10107-021-01636-z.

[10] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 3036–3046. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

[11] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.

[12] Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. 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/fe131d7f5a6b38b23cc967316c13dae2-Paper.pdf.

[13] Simon S. Du, Xiyu Zhai, Barnabás Poczós, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks, 2018. arXiv:1810.02054.

[14] Weinan E, Chao Ma, Stephan Wojtowytsch, and Lei Wu. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't, 2020. arXiv:2009.10713.

[15] Weinan E, Chao Ma, and Lei Wu. A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Sci. China Math.*, 63(7):1235–1258, 2020. doi:10.1007/s11425-019-1628-5.

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

[17] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

[18] Arnulf Jentzen and Adrian Riekert. A proof of convergence for stochastic gradient descent in the training of artificial neural networks with ReLU activation for constant target functions, 2021. arXiv:2104.00277.

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

[20] Philipp Petersen, Mones Raslan, and Felix Voigtlaender. Topological Properties of the Set of Functions Generated by Neural Networks of Fixed Size. *Found. Comput. Math.*, 21(2):375–444, 2021. doi:10.1007/s10208-020-09461-0.

[21] Walter Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition, 1987.

[22] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bull. Math. Sci.*, 7(1):87–154, 2017. doi:10.1007/s13373-017-0101-1.

[23] Francis Williams, Matthew Trager, Daniele Panozzo, Claudio Silva, Denis Zorin, and Joan Bruna. Gradient dynamics of shallow univariate ReLU networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/1f6419b1cbe79c71410cb320fc094775-Paper.pdf.