Electronic
Research Archive

Research article

Existence, uniqueness, and convergence rates for gradient flows in the training of artificial neural networks with ReLU activation

Simon Eberle[1,*]**, Arnulf Jentzen**[2,3,*]**, Adrian Riekert**[3] **and Georg S. Weiss**[4]

[1] Basque Center of Applied Mathematics, Spain
[2] School of Data Science and Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China
[3] Applied Mathematics: Institute for Analysis and Numerics, University of Münster, Germany
[4] Faculty of Mathematics, AG Analysis of Partial Differential Equations, University of Duisburg-Essen, Germany

* **Correspondence:** Email: seberle@bcamath.org, ajentzen@uni-muenster.de.

Abstract: The training of artificial neural networks (ANNs) with rectified linear unit (ReLU) activation via gradient descent (GD) type optimization schemes is nowadays a common industrially relevant procedure. GD type optimization schemes can be regarded as temporal discretization methods for the gradient flow (GF) differential equations associated to the considered optimization problem and, in view of this, it seems to be a natural direction of research to *first aim to develop a mathematical convergence theory for time-continuous GF differential equations* and, thereafter, to aim to extend such a time-continuous convergence theory to implementable time-discrete GD type optimization methods. In this article we establish two basic results for GF differential equations in the training of fully-connected feedforward ANNs with one hidden layer and ReLU activation. In the first main result of this article we establish in the training of such ANNs under the assumption that the probability distribution of the input data of the considered supervised learning problem is absolutely continuous with a bounded density function that every GF differential equation admits for every initial value a solution which is also unique among a suitable class of solutions. In the second main result of this article we prove in the training of such ANNs under the assumption that the target function and the density function of the probability distribution of the input data are piecewise polynomial that every non-divergent GF trajectory converges with an appropriate rate of convergence to a critical point and that the risk of the non-divergent GF trajectory converges with rate 1 to the risk of the critical point. We establish this result by proving that the considered risk function is *semialgebraic* and, consequently, satisfies the *Kurdyka-Łojasiewicz inequality*, which allows us to show convergence of every non-divergent GF trajectory.

Keywords: deep learning; artificial intelligence; optimization; gradient flow; Kurdyka-Łojasiewicz inequalities

1. Introduction

The training of artificial neural networks (ANNs) with rectified linear unit (ReLU) activation via gradient descent (GD) type optimization schemes is nowadays a common industrially relevant procedure which appears, for instance, in the context of natural language processing, face recognition, fraud detection, and game intelligence. Although there exist a large number of numerical simulations in which GD type optimization schemes are effectively used to train ANNs with ReLU activation, till this day in the scientific literature there is in general no mathematical convergence analysis which explains the success of GD type optimization schemes in the training of such ANNs.

GD type optimization schemes can be regarded as temporal discretization methods for the gradient flow (GF) differential equations associated to the considered optimization problem and, in view of this, it seems to be a natural direction of research to *first aim to develop a mathematical convergence theory for time-continuous GF differential equations* and, thereafter, to aim to extend such a time-continuous convergence theory to implementable time-discrete GD type optimization methods.

Although there is in general no theoretical analysis which explains the success of GD type optimization schemes in the training of ANNs in the literature, there are several auspicious analysis approaches as well as several promising partial error analyses regarding the training of ANNs via GD type optimization schemes and GFs, respectively, in the literature. For convex objective functions, the convergence of GF and GD processes to the global minimum in different settings has been proved, e.g., in [1–5]. For general non-convex objective functions, even under smoothness assumptions GF and GD processes can show wild oscillations and admit infinitely many limit points, cf., e.g., [6]. A standard condition which excludes this undesirable behavior is the *Kurdyka-Łojasiewicz inequality* and we point to [7–16] for convergence results for GF and GD processes under Łojasiewicz type assumptions. It is in fact one of the main contributions of this work to demonstrate that the objective functions occurring in the training of ANNs with ReLU activation satisfy an appropriate Kurdyka-Łojasiewicz inequality, provided that both the target function and the density of the probability distribution of the input data are *piecewise polynomial*. For further abstract convergence results for GF and GD processes in the non-convex setting we refer, e.g., to [17–21] and the references mentioned therein.

In the *overparametrized regime*, where the number of training parameters is much larger than the number of training data points, GF and GD processes can be shown to converge to global minima in the training of ANNs with high probability, cf., e.g., [22–28]. As the number of neurons increases to infinity, the corresponding GF processes converge (with appropriate rescaling) to a measure-valued process which is known in the scientific literature as Wasserstein GF. For results on the convergence behavior of Wasserstein GFs in the training of ANNs we point, e.g., to [29–31], [32, Section 5.1], and the references mentioned therein.

A different approach is to consider only very *special target functions* and we refer, in particular, to [33, 34] for a convergence analysis for GF and GD processes in the case of constant target functions and to [35] for a convergence analysis for GF and GD processes in the training of ANNs with piecewise linear target functions. In the case of linear target functions, a complete characterization of the non-global local minima and the saddle points of the risk function has been obtained in [36].

In this article we establish two basic results for GF differential equations in the training of *fully-connected feedforward ANNs* with one hidden layer and ReLU activation. Specifically, in the first main result of this article, see Theorem 1.1 below, we establish in the training of such ANNs under

the assumption that the probability distribution of the input data of the considered supervised learning problem is absolutely continuous with a bounded density function that every GF differential equation *possesses* for every initial value *a solution* which is also unique among a suitable class of solutions (see (1.6) in Theorem 1.1 for details). In the second main result of this article, see Theorem 1.2 below, we prove in the training of such ANNs under the assumption that the target function and the density function are piecewise polynomial (see (1.8) below for details) that every non-divergent GF trajectory *converges* with an appropriate speed of convergence (see (1.11) below) *to a critical point*.

In Theorems 1.1 and 1.2 we consider ANNs with $d \in \mathbb{N} = \{1, 2, 3, \dots\}$ neurons on the input layer (d-dimensional input), $H \in \mathbb{N}$ neurons on the hidden layer (H-dimensional hidden layer), and 1 neuron on the output layer (1-dimensional output). There are thus Hd scalar real weight parameters and H scalar real bias parameters to describe the affine linear transformation between 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 between the H-dimensional hidden layer and the 1-dimensional output layer. Altogether there are thus

$$\mathfrak{d} = Hd + H + H + 1 = Hd + 2H + 1 \tag{1.1}$$

real numbers to describe the ANNs in Theorems 1.1 and 1.2. We also refer to Figure 1 for a graphical illustration of the architecture of an example ANN with $d = 4$ neurons on the input layer and $H = 5$ neurons on the hidden layer.

The real numbers $a \in \mathbb{R}$, $b \in (a, \infty)$ in Theorems 1.1 and 1.2 are used to specify the set $[a, b]^d$ in which the input data of the considered supervised learning problem takes values in and the function $f \colon [a, b]^d \to \mathbb{R}$ in Theorem 1.1 specifies the target function of the considered supervised learning problem.

In Theorem 1.1 we assume that the target function is an element of the set $C([a, b]^d, \mathbb{R})$ of continuous functions from $[a, b]^d$ to $\mathbb{R}$ but beside this continuity hypothesis we do not impose further regularity assumptions on the target function.

The function $\mathfrak{p} \colon [a, b]^d \to [0, \infty)$ in Theorems 1.1 and 1.2 is an unnormalized density function of the probability distribution of the input data of the considered supervised learning problem and in Theorem 1.1 we impose that this unnormalized density function is bounded and measurable.

In Theorems 1.1 and 1.2 we consider ANNs with the ReLU activation function

$$\mathbb{R} \ni x \mapsto \max\{x, 0\} \in \mathbb{R}. \tag{1.2}$$

The ReLU activation function fails to be differentiable and this lack of regularity also transfers to the risk function of the considered supervised learning problem; cf. (1.5) below. We thus need to employ appropriately generalized gradients of the risk function to specify the dynamics of the GFs. As in [34, Setting 2.1 and Proposition 2.3] (cf. also [33, 37]), we accomplish this, first, by approximating the ReLU activation function through 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 and, thereafter, by specifying the generalized gradient function as the limit of the gradients of the approximated risk functions; see (1.3) and (1.5) in Theorem 1.1 and (1.9) and (1.10) in Theorem 1.2 for details.

We now present the precise statement of Theorem 1.1 and, thereafter, provide further comments regarding Theorem 1.2.
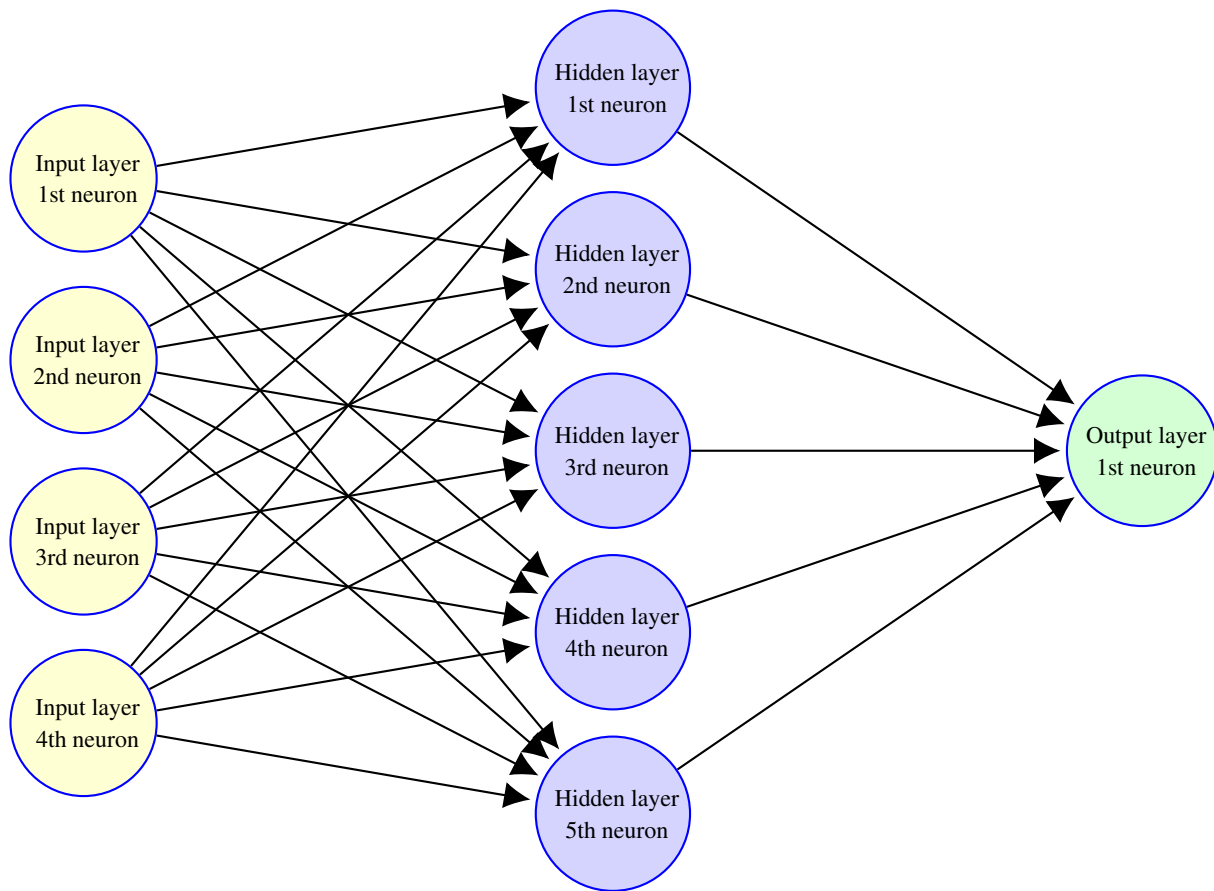
Figure 1. Graphical illustration of the architecture of an example fully-connected feedforward ANN with one hidden layer with 4 neurons on the input layer, 5 neurons on the hidden layer, and 1 neuron on the output layer corresponding to $d = 4$ and $H = 5$ in Theorems 1.1 and 1.2. In this example there are $Hd = 20$ arrows from the input layer to the hidden layer corresponding to $Hd = 20$ weight parameters to describe the affine linear transformation from the input layer to the hidden layer, there are $H = 5$ bias parameters to describe the affine linear transformation from the input layer to the hidden layer, there are $H = 5$ arrows from the hidden layer to the output layer corresponding to $H = 5$ weight parameters to describe the affine linear transformation from the hidden layer to the output layer, and there is 1 bias parameter to describe the affine linear transformation from the hidden layer to the output layer. The overall number $\mathfrak{d} \in \mathbb{N}$ of ANN parameters thus satisfies $\mathfrak{d} = Hd + H + H + 1 = Hd + 2H + 1 = 20 + 10 + 1 = 31$ (cf. (1.1), Theorems 1.1 and 1.2.

Theorem 1.1 (Existence and uniqueness of solutions of GFs in the training of ANNs). *Let $d, H, \mathfrak{d} \in \mathbb{N}$, $a \in \mathbb{R}$, $\mathscr{b} \in (a, \infty)$, $f \in C([a, \mathscr{b}]^d, \mathbb{R})$ satisfy $\mathfrak{d} = dH + 2H + 1$, let $\mathfrak{p} \colon [a, \mathscr{b}]^d \to [0, \infty)$ be bounded and measurable, let $\mathfrak{R}_r \in C(\mathbb{R}, \mathbb{R})$, $r \in \mathbb{N} \cup \{\infty\}$, satisfy for all $x \in \mathbb{R}$ that $(\cup_{r \in \mathbb{N}} \{\mathfrak{R}_r\}) \subseteq C^1(\mathbb{R}, \mathbb{R})$, $\mathfrak{R}_\infty(x) = \max\{x, 0\}$, $\sup_{r \in \mathbb{N}} \sup_{y \in [-|x|, |x|]} |(\mathfrak{R}_r)'(y)| < \infty$, and*

$$\limsup_{r \to \infty} (|\mathfrak{R}_r(x) - \mathfrak{R}_\infty(x)| + |(\mathfrak{R}_r)'(x) - \mathbf{1}_{(0,\infty)}(x)|) = 0, \tag{1.3}$$

for every $\theta = (\theta_1, \ldots, \theta_\mathfrak{d}) \in \mathbb{R}^\mathfrak{d}$ *let* $\mathbf{D}^\theta \subseteq \mathbb{N}$ *satisfy*

$$\mathbf{D}^\theta = \{i \in \{1, 2, \ldots, H\} : |\theta_{Hd+i}| + \sum_{j=1}^d |\theta_{(i-1)d+j}| = 0\}, \tag{1.4}$$

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

$$\mathcal{L}_r(\theta) = \int_{[a,b]^d} (f(x_1, \ldots, x_d)$$
$$- \theta_\mathfrak{d} - \sum_{i=1}^H \theta_{H(d+1)+i}[\mathfrak{R}_r(\theta_{Hd+i} + \sum_{j=1}^d \theta_{(i-1)d+j} x_j)])^2 \mathfrak{p}(x) \, \mathrm{d}(x_1, \ldots, x_d), \tag{1.5}$$

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

(i) *it holds that* $\mathcal{G}$ *is locally bounded and measurable and*

(ii) *there exists a unique* $\Theta \in C([0, \infty), \mathbb{R}^\mathfrak{d})$ *which satisfies for all* $t \in [0, \infty)$, $s \in [t, \infty)$ *that* $\mathbf{D}^{\Theta_t} \subseteq \mathbf{D}^{\Theta_s}$ *and*

$$\Theta_t = \theta - \int_0^t \mathcal{G}(\Theta_u) \, \mathrm{d}u. \tag{1.6}$$

Theorem 1.1 is a direct consequence of Theorem 3.3 below. In Theorem 1.2 we also assume that the target function $f : [a, b]^d \to \mathbb{R}$ is continuous but additionally assume that, roughly speaking, both the target function $f : [a, b]^d \to \mathbb{R}$ and the unnormalized density function $\mathfrak{p} : [a, b]^d \to [0, \infty)$ coincide with polynomial functions on suitable subsets of their domain of definition $[a, b]^d$. In Theorem 1.2 the $(n \times d)$-matrices $\alpha_i^k \in \mathbb{R}^{n \times d}$, $i \in \{1, 2, \ldots, n\}$, $k \in \{0, 1\}$, and the n-dimensional vectors $\beta_i^k \in \mathbb{R}^n$, $i \in \{1, 2, \ldots, n\}$, $k \in \{0, 1\}$, are used to describe these subsets and the functions $P_i^k : \mathbb{R}^d \to \mathbb{R}$, $i \in \{1, 2, \ldots, n\}$, $k \in \{0, 1\}$, constitute the polynomials with which the target function and the unnormalized density function should partially coincide. More formally, in (1.8) in Theorem 1.2 we assume that for every $x \in [a, b]^d$ we have that

$$\mathfrak{p}(x) = \sum_{i \in \{1,2,\ldots,n\}, \alpha_i^0 x + \beta_i^0 \in [0,\infty)^n} P_i^0(x) \quad \text{and} \quad f(x) = \sum_{i \in \{1,2,\ldots,n\}, \alpha_i^1 x + \beta_i^1 \in [0,\infty)^n} P_i^1(x). \tag{1.7}$$

In (1.11) in Theorem 1.2 we prove that there exists a strictly positive real number $\beta \in (0, \infty)$ such that for every GF trajectory $\Theta : [0, \infty) \to \mathbb{R}^\mathfrak{d}$ which does not diverge to infinity in the sense* that $\liminf_{t \to \infty} \|\Theta_t\| < \infty$ we have that $\Theta_t \in \mathbb{R}^\mathfrak{d}$, $t \in [0, \infty)$, converges with order β to a critical point $\vartheta \in \mathcal{G}^{-1}(\{0\}) = \{\theta \in \mathbb{R}^\mathfrak{d} : \mathcal{G}(\theta) = 0\}$ and we have that the risk $\mathcal{L}(\Theta_t) \in \mathbb{R}$, $t \in [0, \infty)$, converges with order 1 to the risk $\mathcal{L}(\vartheta)$ of the critical point ϑ. We now present the precise statement of Theorem 1.2.

Theorem 1.2 (Convergence rates for GFs trajectories in the training of ANNs). *Let* $d, H, \mathfrak{d}, n \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in (a, \infty)$, $f \in C([a, b]^d, \mathbb{R})$ *satisfy* $\mathfrak{d} = dH + 2H + 1$, *for every* $i \in \{1, 2, \ldots, n\}$, $k \in \{0, 1\}$ *let* $\alpha_i^k \in \mathbb{R}^{n \times d}$, *let* $\beta_i^k \in \mathbb{R}^n$, *and let* $P_i^k : \mathbb{R}^d \to \mathbb{R}$ *be a polynomial, let* $\mathfrak{p} : [a, b]^d \to [0, \infty)$ *satisfy for all* $k \in \{0, 1\}$, $x \in [a, b]^d$ *that*

$$kf(x) + (1-k)\mathfrak{p}(x) = \sum_{i=1}^n \left[P_i^k(x) \mathbf{1}_{[0,\infty)^n}(\alpha_i^k x + \beta_i^k) \right], \tag{1.8}$$

---

*Note that the functions $\|\cdot\| : (\cup_{n \in \mathbb{N}} \mathbb{R}^n) \to \mathbb{R}$ and $\langle \cdot, \cdot \rangle : (\cup_{n \in \mathbb{N}} (\mathbb{R}^n \times \mathbb{R}^n)) \to \mathbb{R}$ 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 $\langle x, y \rangle = \sum_{i=1}^\mathfrak{d} x_i y_i$.

let $\mathfrak{R}_r \in C(\mathbb{R}, \mathbb{R})$, $r \in \mathbb{N} \cup \{\infty\}$, satisfy for all $x \in \mathbb{R}$ that $(\cup_{r \in \mathbb{N}}\{\mathfrak{R}_r\}) \subseteq C^1(\mathbb{R}, \mathbb{R})$, $\mathfrak{R}_\infty(x) = \max\{x, 0\}$, $\sup_{r \in \mathbb{N}} \sup_{y \in [-|x|, |x|]} |(\mathfrak{R}_r)'(y)| < \infty$, and

$$\limsup_{r \to \infty} (|\mathfrak{R}_r(x) - \mathfrak{R}_\infty(x)| + |(\mathfrak{R}_r)'(x) - \mathbf{1}_{(0,\infty)}(x)|) = 0, \tag{1.9}$$

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

$$\mathcal{L}_r(\theta) = \int_{[a,b]^d} \left(f(x_1, \ldots, x_d)\right.$$
$$\left. - \theta_{\mathfrak{d}} - \sum_{i=1}^{H} \theta_{H(d+1)+i} \left[\mathfrak{R}_r\left(\theta_{Hd+i} + \sum_{j=1}^{d} \theta_{(i-1)d+j} x_j\right)\right]\right)^2 \mathfrak{p}(x) \, \mathrm{d}(x_1, \ldots, x_d), \tag{1.10}$$

let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \{\vartheta \in \mathbb{R}^{\mathfrak{d}}: ((\nabla \mathcal{L}_r)(\vartheta))_{r \in \mathbb{N}} \text{ is convergent}\}$ that $\mathcal{G}(\theta) = \lim_{r \to \infty}(\nabla \mathcal{L}_r)(\theta)$, and let $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ satisfy $\liminf_{t \to \infty}\|\Theta_t\| < \infty$ and $\forall\, t \in [0, \infty): \Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) \, \mathrm{d}s$. Then there exist $\vartheta \in \mathcal{G}^{-1}(\{0\})$, $\mathfrak{C}, \beta \in (0, \infty)$ which satisfy for all $t \in [0, \infty)$ that

$$\|\Theta_t - \vartheta\| \leq \mathfrak{C}(1 + t)^{-\beta} \qquad and \qquad |\mathcal{L}_\infty(\Theta_t) - \mathcal{L}_\infty(\vartheta)| \leq \mathfrak{C}(1 + t)^{-1}. \tag{1.11}$$

Theorem 1.2 above is an immediate consequence of Theorem 5.4 in Subsection 5.3 below. Theorem 1.2 is related to Theorem 1.1 in our previous article [37]. In particular, [37, Theorem 1.1] uses weaker assumptions than Theorem 1.2 above but Theorem 1.2 above establishes a stronger statement when compared to [37, Theorem 1.1]. Specifically, on the one hand in [37, Theorem 1.1] the target function is only assumed to be a continuous function and the unnormalized density is only assumed to be measurable and integrable while in Theorem 1.2 it is additionally assumed that both the target function and the unnormalized density are *piecewise polynomial* in the sense of (1.8) above. On the other hand [37, Theorem 1.1] only asserts that the risk of every bounded GF trajectory converges to the risk of critical point while Theorem 1.2 assures that every non-divergent GF trajectory *converges with a strictly positive rate of convergence* to a critical point (the rate of convergence is given through the strictly positive real number $\beta \in (0, \infty)$ appearing in the exponent on the left inequality in (Eq 1.11) in Theorem 1.2) and also assures that the risk of the non-divergent GF trajectory *converges with rate 1* to the risk of the critical point (the convergence rate 1 is ensured through the 1 appearing in the exponent on the right inequality in (Eq 1.11) in Theorem 1.2).

We also point out that Theorem 1.2 assumes that the GF trajectory is non-divergent in the sense that $\liminf_{t \to \infty}\|\Theta_t\| < \infty$. In general, it remains an open problem to establish sufficient conditions which ensure that the GF trajectory has this non-divergence property. In this aspect we also refer to Gallon et al. [38] for counterexamples for which it has been proved that every GF trajectory with sufficiently small initial risk does in the training of ANNs diverge to ∞ in the sense that $\liminf_{t \to \infty}\|\Theta_t\| = \infty$.

The remainder of this article is organized in the following way. In Section 2 we establish several regularity properties for the risk function of the considered supervised learning problem and its generalized gradient function. In Section 3 we employ the findings from Section 2 to establish existence and uniqueness properties for solutions of GF differential equations. In particular, in Section 3 we present the proof of Theorem 1.1 above. In Section 4 we establish under the assumption that both the target function $f: [a, b]^d \to \mathbb{R}$ and the unnormalized density function $\mathfrak{p}: [a, b]^d \to [0, \infty)$ are piecewise polynomial that the risk function is *semialgebraic* in the sense of Definition 4.3 in Section 4 (see Corollary 4.10 in Section 4 for details). In Section 5 we engage the results from Sections 2 and 4 to establish several convergence rate results for solutions of GF differential equations and, thereby, we also prove Theorem 1.2 above.

2. Properties of the risk function and its generalized gradient function

In this section we establish several regularity properties for the risk function $\mathcal{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ and its generalized gradient function $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$. In particular, in Proposition 2.12 in Subsection 2.5 below we prove for every parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}}$ in the ANN parameter space $\mathbb{R}^{\mathfrak{d}} = \mathbb{R}^{dH+2H+1}$ that the generalized gradient $\mathcal{G}(\theta)$ is a limiting subdifferential of the risk function $\mathcal{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ at θ. In Definition 2.8 in Subsection 2.5 we recall the notion of subdifferentials (which are sometimes also referred to as Fréchet subdifferentials in the scientific literature) and in Definition 2.9 in Subsection 2.5 we recall the notion of limiting subdifferentials. In the scientific literature Definitions 2.8 and 2.9 can in a slightly different presentational form, e.g., be found in Rockafellar & Wets [39, Definition 8.3] and Bolte et al. [9, Definition 2.10], respectively.

Our proof of Proposition 2.12 uses the continuously differentiability result for the risk function in Proposition 2.3 in Subsection 2.2 and the local Lipschitz continuity result for the generalized gradient function in Corollary 2.7 in Subsection 2.4. Corollary 2.7 will also be employed in Section 3 below to establish existence and uniqueness results for solutions of GF differential equations. Proposition 2.3 follows directly from [37, Proposition 2.10, Lemmas 2.11 and 2.12]. Our proof of Corollary 2.7, in turn, employs the known representation result for the generalized gradient function in Proposition 2.2 in Subsection 2.2 below and the local Lipschitz continuity result for certain parameter integrals in Corollary 2.6 in Subsection 2.4. Statements related to Proposition 2.2 can, e.g., be found in [37, Proposition 2.2], [33, Proposition 2.3], and [34, Proposition 2.3].

Our proof of Corollary 2.6 uses the elementary abstract local Lipschitz continuity result for certain parameter integrals in Lemma 2.5 in Subsection 2.4 and the local Lipschitz continuity result for active neuron regions in Lemma 2.4 in Subsection 2.3 below. Lemma 2.4 is a generalization of [35, Lemma 7], Lemma 2.5 is a slight generalization of [35, Lemma 6], and Corollary 2.6 is a generalization of [37, Lemma 2.12] and [35, Corollary 9]. The proof of Lemma 2.5 is therefore omitted.

In Setting 2.1 in Subsection 2.1 below we present the mathematical setup to describe ANNs with ReLU activation, the risk function $\mathcal{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$, and its generalized gradient function $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$. Moreover, in (2.6) in Setting 2.1 we define for a given parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}}$ the set of hidden neurons which have all input parameters equal to zero. Such neurons are sometimes called degenerate (cf. Cheridito et al. [36]) and can cause problems with the differentiability of the risk function, which is why we exclude degenerate neurons in Proposition 2.3 and Corollary 2.7 below.

2.1. Mathematical description of artificial neural networks (ANNs)

In this subsection we present in Setting 2.1 below the mathematical setup that we employ to state most of the mathematical results of this work. We also refer to Figure 2 below for a table in which we briefly list the mathematical objects introduced in Setting 2.1.

Setting 2.1. *Let* $d, H, \mathfrak{d} \in \mathbb{N}$, $\mathfrak{a} \in \mathbb{R}$, $\mathfrak{b} \in (\mathfrak{a}, \infty)$, $f \in C([\mathfrak{a}, \mathfrak{b}]^d, \mathbb{R})$ *satisfy* $\mathfrak{d} = dH + 2H + 1$, *let* $\mathfrak{w} = ((\mathfrak{w}_{i,j}^{\theta})_{(i,j) \in \{1,\dots,H\} \times \{1,\dots,d\}})_{\theta \in \mathbb{R}^{\mathfrak{d}}} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{H \times d}$, $\mathfrak{b} = ((\mathfrak{b}_1^{\theta}, \dots, \mathfrak{b}_H^{\theta}))_{\theta \in \mathbb{R}^{\mathfrak{d}}} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^H$, $\mathfrak{v} = ((\mathfrak{v}_1^{\theta}, \dots, \mathfrak{v}_H^{\theta}))_{\theta \in \mathbb{R}^{\mathfrak{d}}} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^H$, *and* $\mathfrak{c} = (\mathfrak{c}^{\theta})_{\theta \in \mathbb{R}^{\mathfrak{d}}} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ *satisfy for all* $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \dots, H\}$, $j \in \{1, 2, \dots, d\}$ *that*

$$\mathfrak{w}_{i,j}^{\theta} = \theta_{(i-1)d+j}, \qquad \mathfrak{b}_i^{\theta} = \theta_{Hd+i}, \qquad \mathfrak{v}_i^{\theta} = \theta_{H(d+1)+i}, \qquad and \qquad \mathfrak{c}^{\theta} = \theta_{\mathfrak{d}}, \tag{2.1}$$

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

$$\limsup_{r \to \infty} (|\mathfrak{R}_r(x) - \max\{x, 0\}| + |(\mathfrak{R}_r)'(x) - \mathbf{1}_{(0,\infty)}(x)|) = 0 \tag{2.2}$$

and $\sup_{r \in \mathbb{N}} \sup_{y \in [-|x|,|x|]} |(\mathfrak{R}_r)'(y)| < \infty$, *let* $\lambda \colon \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$ *be the Lebesgue–Borel measure on* $\mathbb{R}^d$, *let* $\mathfrak{p} \colon [\mathcal{a}, \mathcal{b}]^d \to [0, \infty)$ *be bounded and measurable, let* $\mathcal{N} = (\mathcal{N}^\theta)_{\theta \in \mathbb{R}^{\mathfrak{d}}} \colon \mathbb{R}^{\mathfrak{d}} \to C(\mathbb{R}^d, \mathbb{R})$ *and* $\mathcal{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ *satisfy for all* $\theta \in \mathbb{R}^{\mathfrak{d}}$, $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ *that*

$$\mathcal{N}^\theta(x) = \mathfrak{c}^\theta + \sum_{i=1}^H \mathfrak{v}_i^\theta \max\{\mathfrak{b}_i^\theta + \sum_{j=1}^d \mathfrak{w}_{i,j}^\theta x_j, 0\} \tag{2.3}$$

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

$$\mathfrak{L}_r(\theta) = \int_{[\mathcal{a}, \mathcal{b}]^d} (f(y) - \mathfrak{c}^\theta - \sum_{i=1}^H \mathfrak{v}_i^\theta [\mathfrak{R}_r(\mathfrak{b}_i^\theta + \sum_{j=1}^d \mathfrak{w}_{i,j}^\theta y_j)])^2 \mathfrak{p}(y) \, \lambda(\mathrm{d}y), \tag{2.4}$$

for every $\varepsilon \in (0, \infty)$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ *let* $B_\varepsilon(\theta) \subseteq \mathbb{R}^{\mathfrak{d}}$ *satisfy* $B_\varepsilon(\theta) = \{\vartheta \in \mathbb{R}^{\mathfrak{d}} \colon \|\theta - \vartheta\| < \varepsilon\}$, *for every* $\theta \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \dots, H\}$ *let* $I_i^\theta \subseteq \mathbb{R}^d$ *satisfy*

$$I_i^\theta = \{x = (x_1, \dots, x_d) \in [\mathcal{a}, \mathcal{b}]^d \colon \mathfrak{b}_i^\theta + \sum_{j=1}^d \mathfrak{w}_{i,j}^\theta x_d > 0\}, \tag{2.5}$$

for every $\theta \in \mathbb{R}^{\mathfrak{d}}$ *let* $\mathbf{D}^\theta \subseteq \mathbb{N}$ *satisfy*

$$\mathbf{D}^\theta = \{i \in \{1, 2, \dots, H\} \colon |\mathfrak{b}_i^\theta| + \sum_{j=1}^d |\mathfrak{w}_{i,j}^\theta| = 0\}, \tag{2.6}$$

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

Next we add some explanations regarding the mathematical framework presented in Setting 2.1 above. In Setting 2.1

- the natural number $d \in \mathbb{N}$ represents *the number of neurons on the input layer* of the considered ANNs,

- the natural number $H \in \mathbb{N}$ represents *the number of neurons on the hidden layer* of the considered ANNs, and

- the natural number $\mathfrak{d} \in \mathbb{N}$ measures *the overall number of parameters* of the considered ANNs

(cf. (1.1) and Figure 1 above). The real numbers $\mathcal{a} \in \mathbb{R}$, $\mathcal{b} \in (\mathcal{a}, \infty)$ in Setting 2.1 are employed to specify the d-dimensional set $[\mathcal{a}, \mathcal{b}]^d \subseteq \mathbb{R}^d$ in which the input data of the supervised learning problem considered in Setting 2.1 takes values in and which, thereby, also serves as *the domain of definition of the target function* of the considered supervised learning problem.

In Setting 2.1 the function $f \colon [\mathcal{a}, \mathcal{b}]^d \to \mathbb{R}$ represents *the target function* of the considered supervised learning problem. In Setting 2.1 the target function f is assumed to be an element of the set $C([\mathcal{a}, \mathcal{b}]^d, \mathbb{R})$ of continuous functions from the d-dimensional set $[\mathcal{a}, \mathcal{b}]^d$ to the reals $\mathbb{R}$ (first line in Setting 2.1).

The matrix valued function $\mathfrak{w} = ((\mathfrak{w}_{i,j}^\theta)_{(i,j) \in \{1,\dots,H\} \times \{1,\dots,d\}})_{\theta \in \mathbb{R}^{\mathfrak{d}}} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{H \times d}$ in Setting 2.1 is used to represent *the inner weight parameters* of the ANNs considered in Setting 2.1. In particular, in Setting 2.1 we have for every $\theta \in \mathbb{R}^{\mathfrak{d}}$ that the $H \times d$-matrix $\mathfrak{w}^\theta = (\mathfrak{w}_{i,j}^\theta)_{(i,j) \in \{1,\dots,H\} \times \{1,\dots,d\}} \in \mathbb{R}^{H \times d}$ represents the *weight parameter matrix* for the affine linear transformation from the d-dimensional input layer to the H-dimensional hidden layer of the ANN associated to the ANN parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}}$ (cf. (2.1), (2.3), and Figure 1).

Mathematical object	Rough description of the introduced mathematical object
d	*Dimension of the input layer* of the considered ANNs (number of neurons on the input layer of the considered ANNs)
H	*Dimension of the hidden layer* of the considered ANNs (number of neurons on the hidden layer of the considered ANNs)
$\mathfrak{d}$	*Overall number of parameters* of the considered ANNs
$[\mathscr{a}, \mathscr{b}]^d$	Subset of $\mathbb{R}^d$ specifying the domain in which the input data of the considered learning problem takes values in
$f: [\mathscr{a}, \mathscr{b}]^d \to \mathbb{R}$	*Target function* of the considered supervised learning problem
$\mathfrak{w}^\theta = (\mathfrak{w}^\theta_{i,j})_{(i,j)\in\{1,\ldots,H\}\times\{1,\ldots,d\}}$	Real $H \times d$-matrix consisting of *the inner weight parameters* of the ANN associated to the parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}}$ (cf. (2.1), (2.3), and Figure 1)
$\mathfrak{b}^\theta = (\mathfrak{b}^\theta_1, \ldots, \mathfrak{b}^\theta_H)$	Real H-dimensional vector consisting of *the inner bias parameters* of the ANN associated to the parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}}$ (cf. (2.1), (2.3), and Figure 1)
$\mathfrak{v}^\theta = (\mathfrak{v}^\theta_1, \ldots, \mathfrak{v}^\theta_H)$	Real H-dimensional vector consisting of *the outer weight parameters* of the ANN associated to the parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}}$ (cf. (2.1), (2.3), and Figure 1)
$\mathfrak{c}^\theta$	Real number representing *the outer bias parameter* of the ANN associated to the parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}}$ (cf. (2.1), (2.3), and Figure 1)
$\mathfrak{R}_r: \mathbb{R} \to \mathbb{R}$ for $r \in \mathbb{N}$	*Continuously differentiable approximations* of the ReLU activation function $\mathbb{R} \ni x \mapsto \max\{x, 0\} \in \mathbb{R}$ (cf. (2.2))
$\lambda: \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$	Lebesgue–Borel measure on $\mathbb{R}^d$
$\mathfrak{p}: [\mathscr{a}, \mathscr{b}]^d \to \mathbb{R}$	*Unnormalized density function* of the considered supervised learning problem (cf. (2.3) and (2.4))
$\mathscr{N}^\theta: \mathbb{R}^d \to \mathbb{R}$	Realization function of the ANN associated to the parameter vector $\theta \in \mathbb{R}^{\mathfrak{d}}$ (cf. (2.3))
$\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$	*Risk function* of the considered supervised learning problem (cf. (2.3))
$\mathfrak{L}_r: \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ for $r \in \mathbb{N}$	Approximated risk function specified through the continuously differentiable approximations $(\mathfrak{R}_r)_{r\in\mathbb{N}}$ (cf. (2.4))
$B_\varepsilon(\theta)$	Open ball in $\mathbb{R}^{\mathfrak{d}}$ with radius $\varepsilon \in (0, \infty)$ and center $\theta \in \mathbb{R}^{\mathfrak{d}}$
I^θ_i	Subset of the set $[\mathscr{a}, \mathscr{b}]^d$ on which the i-th neuron on the hidden layer is *active* (cf. (2.5) and Figure 1)
$\mathbf{D}^\theta$	Subset of the set $\{1, 2, \ldots, H\}$ representing the set of *degenerate neurons on the hidden layer* (cf. (2.6) and Figure 1)
$\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$	*Generalized gradient function* of the risk function $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$

Figure 2. List of the mathematical objects introduced in Setting 2.1.

The vector valued function $\mathfrak{b} = ((\mathfrak{b}_1^\theta, \ldots, \mathfrak{b}_H^\theta))_{\theta \in \mathbb{R}^\mathfrak{d}} \colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}^H$ in Setting 2.1 is used to represent *the inner bias parameters* of the ANNs considered in Setting 2.1. In particular, in Setting 2.1 we have for every $\theta \in \mathbb{R}^\mathfrak{d}$ that the d-dimensional vector $\mathfrak{b}^\theta = (\mathfrak{b}_1^\theta, \ldots, \mathfrak{b}_H^\theta) \in \mathbb{R}^H$ represents the *bias parameter vector* for the affine linear transformation from the d-dimensional input layer to the H-dimensional hidden layer of the ANN associated to the ANN parameter vector $\theta \in \mathbb{R}^\mathfrak{d}$ (cf. (2.1), (2.3), and Figure 1).

The vector valued function $\mathfrak{v} = ((\mathfrak{v}_1^\theta, \ldots, \mathfrak{v}_H^\theta))_{\theta \in \mathbb{R}^\mathfrak{d}} \colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}^H$ in Setting 2.1 is used to describe *the outer weight parameters* of the ANNs considered in Setting 2.1. In particular, in Setting 2.1 we have for every $\theta \in \mathbb{R}^\mathfrak{d}$ that the transpose of the H-dimensional vector $\mathfrak{v}^\theta = (\mathfrak{v}_1^\theta, \ldots, \mathfrak{v}_H^\theta) \in \mathbb{R}^H$ represents the *weight parameter matrix* for the affine linear transformation from the H-dimensional hidden layer to the 1-dimensional output layer of the ANN associated to the ANN parameter vector $\theta \in \mathbb{R}^\mathfrak{d}$ (cf. (2.1), (2.3), and Figure 1).

The real valued function $\mathfrak{c} = (\mathfrak{c}^\theta)_{\theta \in \mathbb{R}^\mathfrak{d}} \colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}$ in Setting 2.1 is used to represent *the outer bias parameters* of the ANNs considered in Setting 2.1. In particular, in Setting 2.1 we have for every $\theta \in \mathbb{R}^\mathfrak{d}$ that the real number $\mathfrak{c}^\theta \in \mathbb{R}$ describes the *bias parameter* for the affine linear transformation from the H-dimensional hidden layer to the 1-dimensional output layer of the ANN associated to the ANN parameter vector $\theta \in \mathbb{R}^\mathfrak{d}$ (cf. (2.1), (2.3), and Figure 1).

In Setting 2.1 we consider *ANNs with the ReLU activation function* $\mathbb{R} \ni x \mapsto \max\{x, 0\} \in \mathbb{R}$ (cf. (1.2)). The ReLU activation function fails to be differentiable and this lack of differentiability typically transfers from the activation function to *the realization functions* $\mathscr{N}^\theta \colon \mathbb{R}^d \to \mathbb{R}$, $\theta \in \mathbb{R}^\mathfrak{d}$, of the considered ANNs and *the risk function* $\mathcal{L} \colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}$ of the considered supervised learning problem, both, introduced in (2.3) in Setting 2.1. In general, there thus do not exist standard derivatives and standard gradients of the risk function and, in view of this, we need to introduce suitably generalized gradients of the risk function to specify the GF dynamics. As in [34, Setting 2.1 and Proposition 2.3] (cf. also [33, 37]), we accomplish this,

- first, by approximating the ReLU activation function 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,

- then, by using these continuously differentiable approximations of the ReLU activation function to specify approximated risk functions, and,

- finally, by specifying *the generalized gradient function* as the pointwise limit of the standard gradients of the approximated risk functions.

In Setting 2.1 the functions $\mathfrak{R}_r \colon \mathbb{R} \to \mathbb{R}$, $r \in \mathbb{N}$, serves as such appropriate *continuously differentiable approximations of the ReLU activation function* and the hypothesis in (2.2) ensures that these functions converge pointwise to the ReLU activation function and that the derivatives of these functions converge pointwise to the left derivative of the ReLU activation function (cf. also (1.3) in Theorem 1.1 and (1.9) in Theorem 1.2). These continuously differentiable approximations of the ReLU activation function are then used in (2.4) in Setting 2.1 (cf. also (1.5) in Theorem 1.1 and (1.10) in Theorem 1.2) to introduce continuously differentiable approximated risk functions $\mathfrak{L}_r \colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}$, $r \in \mathbb{N}$, which converge pointwise to the risk function $\mathcal{L} \colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}$ (cf., e.g., [37, Proposition 2.2]). Finally, the standard gradients of the approximated risk functions $\mathfrak{L}_r \colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}$, $r \in \mathbb{N}$, are then used to introduce the generalized gradient function $\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_\mathfrak{d}) \colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}^\mathfrak{d}$ in Setting 2.1. In this regard we also note that Proposi-

tion 2.2 in Subsection 2.2 below, in particular, ensures that the function $\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_\mathfrak{d})\colon \mathbb{R}^\mathfrak{d} \to \mathbb{R}^\mathfrak{d}$ in Setting 2.1 is indeed uniquely defined.

2.2. Differentiability properties of the risk function

Proposition 2.2. *Assume Setting 2.1. Then it holds for all* $\theta \in \mathbb{R}^\mathfrak{d}$, $i \in \{1, 2, \ldots, H\}$, $j \in \{1, 2, \ldots, d\}$ *that*

$$\mathcal{G}_{(i-1)d+j}(\theta) = 2\mathfrak{v}_i^\theta \int_{I_i^\theta} x_j (\mathcal{N}^\theta(x) - f(x))\mathfrak{p}(x)\,\lambda(\mathrm{d}x),$$

$$\mathcal{G}_{Hd+i}(\theta) = 2\mathfrak{v}_i^\theta \int_{I_i^\theta} (\mathcal{N}^\theta(x) - f(x))\mathfrak{p}(x)\,\lambda(\mathrm{d}x),$$

$$\mathcal{G}_{H(d+1)+i}(\theta) = 2 \int_{[\mathfrak{a},\mathfrak{b}]^d} [\max\{\mathfrak{b}_i^\theta + \textstyle\sum_{j=1}^d \mathfrak{w}_{i,j}^\theta x_j, 0\}](\mathcal{N}^\theta(x) - f(x))\mathfrak{p}(x)\,\lambda(\mathrm{d}x),$$

$$and \qquad \mathcal{G}_\mathfrak{d}(\theta) = 2 \int_{[\mathfrak{a},\mathfrak{b}]^d} (\mathcal{N}^\theta(x) - f(x))\mathfrak{p}(x)\,\lambda(\mathrm{d}x). \tag{2.7}$$

Proof of Proposition 2.2. Observe that, e.g., [37, Proposition 2.2] establishes (2.7). The proof of Proposition 2.2 is thus complete.

Proposition 2.3. *Assume Setting 2.1 and let* $U \subseteq \mathbb{R}^\mathfrak{d}$ *satisfy* $U = \{\theta \in \mathbb{R}^\mathfrak{d} \colon \mathbf{D}^\theta = \varnothing\}$. *Then*

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

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

(iii) *it holds that* $\nabla(\mathcal{L}|_U) = \mathcal{G}|_U$.

Proof of Proposition 2.3. Note that [37, Proposition 2.10, Lemmas 2.11 and 2.12] establish items (i), (ii), and (iii). The proof of Proposition 2.3 is thus complete.

2.3. Local Lipschitz continuity of active neuron regions

Lemma 2.4. *Let* $d \in \mathbb{N}$, $\mathfrak{a} \in \mathbb{R}$, $\mathfrak{b} \in (\mathfrak{a}, \infty)$, *for every* $v = (v_1, \ldots, v_{d+1}) \in \mathbb{R}^{d+1}$ *let* $I^v \subseteq [\mathfrak{a}, \mathfrak{b}]^d$ *satisfy* $I^v = \{x \in [\mathfrak{a}, \mathfrak{b}]^d \colon v_{d+1} + \sum_{i=1}^d v_i x_i > 0\}$, *for every* $n \in \mathbb{N}$ *let* $\lambda_n \colon \mathcal{B}(\mathbb{R}^n) \to [0, \infty]$ *be the Lebesgue–Borel measure on* $\mathbb{R}^n$, *let* $\mathfrak{p} \colon [\mathfrak{a}, \mathfrak{b}]^d \to [0, \infty)$ *be bounded and measurable, and let* $u \in \mathbb{R}^{d+1} \setminus \{0\}$. *Then there exist* $\varepsilon, \mathfrak{C} \in (0, \infty)$ *such that for all* $v, w \in \mathbb{R}^{d+1}$ *with* $\max\{\|u - v\|, \|u - w\|\} \leq \varepsilon$ *it holds that*

$$\int_{I^v \Delta I^w} \mathfrak{p}(x)\,\lambda_d(\mathrm{d}x) \leq \mathfrak{C}\|v - w\|. \tag{2.8}$$

Proof of Lemma 2.4. Observe that for all $v, w \in \mathbb{R}^{d+1}$ we have that

$$\int_{I^v \Delta I^w} \mathfrak{p}(x)\,\lambda_d(\mathrm{d}x) \leq (\sup_{x \in [\mathfrak{a},\mathfrak{b}]^d} \mathfrak{p}(x))\lambda_d(I^v \Delta I^w). \tag{2.9}$$

Moreover, note that the fact that for all $y \in \mathbb{R}$ it holds that $y \geq -|y|$ ensures that for all $v = (v_1, \ldots, v_{d+1}) \in \mathbb{R}^{d+1}$, $i \in \{1, 2, \ldots, d+1\}$ with $\|u - v\| < |u_i|$ it holds that

$$u_i v_i = (u_i)^2 + (v_i - u_i)u_i \geq |u_i|^2 - |u_i - v_i||u_i| \geq |u_i|^2 - \|u - v\||u_i| > 0. \tag{2.10}$$

Next observe that for all $v_1, v_2, w_1, w_2 \in \mathbb{R}$ with $\min\{|v_1|, |w_1|\} > 0$ it holds that

$$\left|\tfrac{v_2}{v_1} - \tfrac{w_2}{w_1}\right| = \tfrac{|v_2 w_1 - w_2 v_1|}{|v_1 w_1|} = \tfrac{|v_2(w_1 - v_1) + v_1(v_2 - w_2)|}{|v_1 w_1|} \leq \left[\tfrac{|v_2| + |v_1|}{|v_1 w_1|}\right][|v_1 - w_1| + |v_2 - w_2|]. \tag{2.11}$$

Combining this and (2.10) demonstrates for all $v = (v_1, \ldots, v_{d+1})$, $w = (w_1, \ldots, w_{d+1}) \in \mathbb{R}^{d+1}$, $i \in \{1, 2, \ldots, d\}$ with $\max\{\|v - u\|, \|w - u\|\} < |u_1|$ that $v_1 w_1 > 0$ and

$$\left|\tfrac{v_i}{v_1} - \tfrac{w_i}{w_1}\right| \leq \left[\tfrac{2\|v\|}{|v_1 w_1|}\right][2\|v - w\|] \leq \left[\tfrac{4\|v - u\| + 4\|u\|}{|v_1 w_1|}\right]\|v - w\|. \tag{2.12}$$

Hence, we obtain for all $v = (v_1, \ldots, v_{d+1})$, $w = (w_1, \ldots, w_{d+1}) \in \mathbb{R}^{d+1}$, $i \in \{1, 2, \ldots, d\}$ with $\max\{\|v - u\|, \|w - u\|\} \leq \tfrac{|u_1|}{2}$ and $|u_1| > 0$ that $v_1 w_1 > 0$ and

$$\left|\tfrac{v_i}{v_1} - \tfrac{w_i}{w_1}\right| \leq \tfrac{(2|u_1| + 4\|u\|)\|v - w\|}{|u_1 + (v_1 - u_1)||u_1 + (w_1 - u_1)|} \leq \tfrac{6\|u\|\|v - w\|}{(|u_1| - \|v - u\|)(|u_1| - \|w - u\|)} \leq \left[\tfrac{24\|u\|}{|u_1|^2}\right]\|v - w\|. \tag{2.13}$$

In the following we distinguish between the case $\max_{i \in \{1, 2, \ldots, d\}}|u_i| = 0$, the case $(\max_{i \in \{1, 2, \ldots, d\}}|u_i|, d) \in (0, \infty) \times [2, \infty)$, and the case $(\max_{i \in \{1, 2, \ldots, d\}}|u_i|, d) \in (0, \infty) \times \{1\}$. We first prove (2.8) in the case

$$\max_{i \in \{1, 2, \ldots, d\}}|u_i| = 0. \tag{2.14}$$

Note that (2.14) and the assumption that $u \in \mathbb{R}^{d+1}\backslash\{0\}$ imply that $|u_{d+1}| > 0$. Moreover, observe that (2.14) shows that for all $v = (v_1, \ldots, v_{d+1}) \in \mathbb{R}^{d+1}$, $x = (x_1, \ldots, x_d) \in I^u \Delta I^v$ we have that

$$\begin{aligned}
&\left|\left(\left[\sum_{i=1}^d v_i x_i\right] + v_{d+1}\right) - \left(\left[\sum_{i=1}^d u_i x_i\right] + u_{d+1}\right)\right| \\
&= \left|\left[\sum_{i=1}^d v_i x_i\right] + v_{d+1}\right| + \left|\left[\sum_{i=1}^d u_i x_i\right] + u_{d+1}\right| \geq \left|\left[\sum_{i=1}^d u_i x_i\right] + u_{d+1}\right| = |u_{d+1}|.
\end{aligned} \tag{2.15}$$

In addition, note that for all $v = (v_1, \ldots, v_{d+1}) \in \mathbb{R}^{d+1}$, $x = (x_1, \ldots, x_d) \in [\alpha, \theta]^d$ it holds that

$$\begin{aligned}
&\left|\left(\left[\sum_{i=1}^d v_i x_i\right] + v_{d+1}\right) - \left(\left[\sum_{i=1}^d u_i x_i\right] + u_{d+1}\right)\right| \leq \left[\sum_{i=1}^d |v_i - u_i||x_i|\right] + |v_{d+1} - u_{d+1}| \\
&\leq \max\{|\alpha|, |\theta|\}\left[\sum_{i=1}^d |v_i - u_i|\right] + |v_{d+1} - u_{d+1}| \leq (1 + d \max\{|\alpha|, |\theta|\})\|v - u\|.
\end{aligned} \tag{2.16}$$

This and (2.15) prove that for all $v \in \mathbb{R}^{d+1}$ with $\|u - v\| \leq \tfrac{|u_{d+1}|}{2 + d \max\{|\alpha|, |\theta|\}}$ we have that $I^u \Delta I^v = \varnothing$, i.e., $I^u = I^v$. Therefore, we get for all $v, w \in \mathbb{R}^{d+1}$ with $\max\{\|u - v\|, \|u - w\|\} \leq \tfrac{|u_{d+1}|}{2 + d \max\{|\alpha|, |\theta|\}}$ that $I^v = I^w = I^u$. Hence, we obtain for all $v, w \in \mathbb{R}^{d+1}$ with $\max\{\|u - v\|, \|u - w\|\} \leq \tfrac{|u_{d+1}|}{2 + d \max\{|\alpha|, |\theta|\}}$ that $\lambda_d(I^v \Delta I^w) = 0$. This establishes (2.8) in the case $\max_{i \in \{1, 2, \ldots, d\}}|u_i| = 0$. In the next step we prove (2.8) in the case

$$(\max_{i \in \{1, 2, \ldots, d\}}|u_i|, d) \in (0, \infty) \times [2, \infty). \tag{2.17}$$

For this we assume without loss of generality that $|u_1| > 0$. In the following let $J_x^{v,w} \subseteq \mathbb{R}$, $x \in [\alpha, \theta]^{d-1}$, $v, w \in \mathbb{R}^{d+1}$, satisfy for all $x = (x_2, \ldots, x_d) \in [\alpha, \theta]^{d-1}$, $v, w \in \mathbb{R}^{d+1}$ that $J_x^{v,w} = \{y \in [\alpha, \theta] : (y, x_2, \ldots, x_d) \in I^v\backslash I^w\}$. Next observe that Fubini's theorem and the fact that for all $v \in \mathbb{R}^{d+1}$ it holds that I^v is measurable show that for all $v, w \in \mathbb{R}^{d+1}$ we have that

$$\begin{aligned}
\lambda_d(I^v \Delta I^w) &= \int_{[\alpha, \theta]^d} \mathbf{1}_{I^v \Delta I^w}(x)\,\lambda_d(dx) = \int_{[\alpha, \theta]^d} \left(\mathbf{1}_{I^v\backslash I^w}(x) + \mathbf{1}_{I^w\backslash I^v}(x)\right)\lambda_d(dx) \\
&= \int_{[\alpha, \theta]^{d-1}} \int_{[\alpha, \theta]} \left(\mathbf{1}_{I^v\backslash I^w}(y, x_2, \ldots, x_d) + \mathbf{1}_{I^w\backslash I^v}(y, x_2, \ldots, x_d)\right)\lambda_1(dy)\,\lambda_{d-1}(d(x_2, \ldots, x_d)) \\
&= \int_{[\alpha, \theta]^{d-1}} \int_{[\alpha, \theta]} \left(\mathbf{1}_{J_x^{v,w}}(y) + \mathbf{1}_{J_x^{w,v}}(y)\right)\lambda_1(dy)\,\lambda_{d-1}(dx) \\
&= \int_{[\alpha, \theta]^{d-1}} \left(\lambda_1(J_x^{v,w}) + \lambda_1(J_x^{w,v})\right)\lambda_{d-1}(dx).
\end{aligned} \tag{2.18}$$

Furthermore, note that for all $x = (x_2, \ldots, x_d) \in [\alpha, \ell]^{d-1}$, $v = (v_1, \ldots, v_{d+1})$, $w = (w_1, \ldots, w_{d+1}) \in \mathbb{R}^{d+1}$, $\mathfrak{s} \in \{-1, 1\}$ with $\min\{\mathfrak{s}v_1, \mathfrak{s}w_1\} > 0$ it holds that

$$
\begin{aligned}
J_x^{v,w} &= \{y \in [\alpha, \ell] : (y, x_2, \ldots, x_d) \in I^v \backslash I^w\} \\
&= \left\{y \in [\alpha, \ell] : v_1 y + \left[\textstyle\sum_{i=2}^d v_i x_i\right] + v_{d+1} > 0 \geq w_1 y + \left[\textstyle\sum_{i=2}^d w_i x_i\right] + w_{d+1}\right\} \\
&= \left\{y \in [\alpha, \ell] : -\tfrac{\mathfrak{s}}{v_1}\left(\left[\textstyle\sum_{i=2}^d v_i x_i\right] + v_{d+1}\right) < \mathfrak{s}y \leq -\tfrac{\mathfrak{s}}{w_1}\left(\left[\textstyle\sum_{i=2}^d w_i x_i\right] + w_{d+1}\right)\right\}.
\end{aligned}
\tag{2.19}
$$

Hence, we obtain for all $x = (x_2, \ldots, x_d) \in [\alpha, \ell]^{d-1}$, $v = (v_1, \ldots, v_{d+1})$, $w = (w_1, \ldots, w_{d+1}) \in \mathbb{R}^{d+1}$, $\mathfrak{s} \in \{-1, 1\}$ with $\min\{\mathfrak{s}v_1, \mathfrak{s}w_1\} > 0$ that

$$
\begin{aligned}
\lambda_1(J_x^{v,w}) &\leq \left|\tfrac{\mathfrak{s}}{v_1}\left(\left[\textstyle\sum_{i=2}^d v_i x_i\right] + v_{d+1}\right) - \tfrac{\mathfrak{s}}{w_1}\left(\left[\textstyle\sum_{i=2}^d w_i x_i\right] + w_{d+1}\right)\right| \\
&\leq \left[\textstyle\sum_{i=2}^d \left|\tfrac{v_i}{v_1} - \tfrac{w_i}{w_1}\right| |x_i|\right] + \left|\tfrac{v_{d+1}}{v_1} - \tfrac{w_{d+1}}{w_1}\right| \\
&\leq \max\{|\alpha|, |\ell|\}\left[\textstyle\sum_{i=2}^d \left|\tfrac{v_i}{v_1} - \tfrac{w_i}{w_1}\right|\right] + \left|\tfrac{v_{d+1}}{v_1} - \tfrac{w_{d+1}}{w_1}\right|.
\end{aligned}
\tag{2.20}
$$

Furthermore, observe that (2.10) demonstrates for all $v = (v_1, \ldots, v_{d+1}) \in \mathbb{R}^{d+1}$ with $\|u - v\| < |u_1|$ that $u_1 v_1 > 0$. This implies that for all $v = (v_1, \ldots, v_{d+1})$, $w = (w_1, \ldots, w_{d+1}) \in \mathbb{R}^{d+1}$ with $\max\{\|u - v\|, \|u - w\|\} < |u_1|$ there exists $\mathfrak{s} \in \{-1, 1\}$ such that $\min\{\mathfrak{s}v_1, \mathfrak{s}w_1\} > 0$. Combining this and (2.13) with (2.20) proves that there exists $\mathfrak{C} \in \mathbb{R}$ such that for all $x \in [\alpha, \ell]^{d-1}$, $v, w \in \mathbb{R}^{d+1}$ with $\max\{\|u - v\|, \|u - w\|\} \leq \tfrac{|u_1|}{2}$ we have that $\lambda_1(J_x^{v,w}) + \lambda_1(J_x^{w,v}) \leq \mathfrak{C}\|v - w\|$. This, (2.18), and (2.9) establish (2.8) in the case $(\max_{i \in \{1,2,\ldots,d\}} |u_i|, d) \in (0, \infty) \times [2, \infty)$. Finally, we prove (2.8) in the case

$$
(\max_{i \in \{1,2,\ldots,d\}} |u_i|, d) \in (0, \infty) \times \{1\}.
\tag{2.21}
$$

Note that (2.21) demonstrates that $|u_1| > 0$. In addition, observe that for all $v = (v_1, v_2)$, $w = (w_1, w_2) \in \mathbb{R}^2$, $\mathfrak{s} \in \{-1, 1\}$ with $\min\{\mathfrak{s}v_1, \mathfrak{s}w_1\} > 0$ it holds that

$$
\begin{aligned}
I^v \backslash I^w &= \{y \in [\alpha, \ell] : v_1 y + v_2 > 0 \geq w_1 y + w_2\} = \left\{y \in [\alpha, \ell] : -\tfrac{\mathfrak{s}v_2}{v_1} < \mathfrak{s}y \leq -\tfrac{\mathfrak{s}w_2}{w_1}\right\} \\
&\subseteq \left\{y \in \mathbb{R} : -\tfrac{\mathfrak{s}v_2}{v_1} < \mathfrak{s}y \leq -\tfrac{\mathfrak{s}w_2}{w_1}\right\}.
\end{aligned}
\tag{2.22}
$$

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

$$
\lambda_1(I^v \backslash I^w) \leq \left|\left(-\tfrac{\mathfrak{s}v_2}{v_1}\right) - \left(-\tfrac{\mathfrak{s}w_2}{w_1}\right)\right| = \left|\tfrac{v_2}{v_1} - \tfrac{w_2}{w_1}\right|.
\tag{2.23}
$$

Furthermore, note that (2.10) ensures for all $v = (v_1, v_2) \in \mathbb{R}^2$ with $\|u - v\| < |u_1|$ that $u_1 v_1 > 0$. This proves that for all $v = (v_1, v_2)$, $w = (w_1, w_2) \in \mathbb{R}^2$ with $\max\{\|u - v\|, \|u - w\|\} < |u_1|$ there exists $\mathfrak{s} \in \{-1, 1\}$ such that $\min\{\mathfrak{s}v_1, \mathfrak{s}w_1\} > 0$. Combining this with (2.23) demonstrates for all $v = (v_1, v_2)$, $w = (w_1, w_2) \in \mathbb{R}^2$ with $\max\{\|u - v\|, \|u - w\|\} < |u_1|$ that $\min\{|v_1|, |w_1|\} > 0$ and

$$
\lambda_1(I^v \Delta I^w) = \lambda_1(I^v \backslash I^w) + \lambda_1(I^w \backslash I^v) \leq 2\left|\tfrac{v_2}{v_1} - \tfrac{w_2}{w_1}\right|.
\tag{2.24}
$$

This, (2.13), and (2.9) establish (2.8) in the case $(\max_{i \in \{1,2,\ldots,d\}} |u_i|, d) \in (0, \infty) \times \{1\}$. The proof of Lemma 2.4 is thus complete.

2.4. Local Lipschitz continuity properties for the generalized gradient function

Lemma 2.5. *Let $d, n \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in (a, \infty)$, $x \in \mathbb{R}^n$, $\mathfrak{C}, \varepsilon \in (0, \infty)$, let $\phi \colon \mathbb{R}^n \times [a, b]^d \to \mathbb{R}$ be locally bounded and measurable, assume for all $r \in (0, \infty)$ that*

$$\sup\nolimits_{y,z \in \mathbb{R}^n, \, \|y\|+\|z\| \leq r, \, y \neq z} \sup\nolimits_{s \in [a,b]^d} \frac{|\phi(y,s) - \phi(z,s)|}{\|y - z\|} < \infty, \tag{2.25}$$

let $\mu \colon \mathcal{B}([a, b]^d) \to [0, \infty)$ be a finite measure, let $I^y \in \mathcal{B}([a, b]^d)$, $y \in \mathbb{R}^n$, satisfy for all $y, z \in \{v \in \mathbb{R}^n \colon \|x - v\| \leq \varepsilon\}$ that $\mu(I^y \Delta I^z) \leq \mathfrak{C}\|y - z\|$, and let $\Phi \colon \mathbb{R}^n \to \mathbb{R}$ satisfy for all $y \in \mathbb{R}^n$ that

$$\Phi(y) = \int_{I^y} \phi(y, s)\, \mu(\mathrm{d}s). \tag{2.26}$$

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

Proof of Lemma 2.5. The proof is analogous to the proof of [35, Lemma 6]. $\quad\square$

Corollary 2.6. *Assume Setting 2.1, let $\phi \colon \mathbb{R}^{\mathfrak{d}} \times [a, b]^d \to \mathbb{R}$ be locally bounded and measurable, and assume for all $r \in (0, \infty)$ that*

$$\sup\nolimits_{\theta, \vartheta \in \mathbb{R}^{\mathfrak{d}}, \, \|\theta\|+\|\vartheta\| \leq r, \, \theta \neq \vartheta} \sup\nolimits_{x \in [a,b]^d} \frac{|\phi(\theta,x) - \phi(\vartheta,x)|}{\|\theta - \vartheta\|} < \infty. \tag{2.27}$$

Then

(i) *it holds that*

$$\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \int_{[a,b]^d} \phi(\theta, x) \mathfrak{p}(x)\, \lambda(\mathrm{d}x) \in \mathbb{R} \tag{2.28}$$

   is locally Lipschitz continuous and

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

$$\{\vartheta \in \mathbb{R}^{\mathfrak{d}} \colon i \notin \mathbf{D}^{\vartheta}\} \ni \theta \mapsto \int_{I_i^{\theta}} \phi(\theta, x) \mathfrak{p}(x)\, \lambda(\mathrm{d}x) \in \mathbb{R} \tag{2.29}$$

   is locally Lipschitz continuous.

Proof of Corollary 2.6. First observe that Lemma 2.5 (applied for every $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $n \curvearrowleft \mathfrak{d}$, $x \curvearrowleft \theta$, $\mu \curvearrowleft (\mathcal{B}([a, b]^d) \ni A \mapsto \int_A \mathfrak{p}(x)\, \lambda(\mathrm{d}x) \in [0, \infty))$, $(I^y)_{y \in \mathbb{R}^n} \curvearrowleft ([a, b]^d)_{y \in \mathbb{R}^{\mathfrak{d}}}$ in the notation of Lemma 2.5) establishes item (i). In the following let $i \in \{1, 2, \ldots, H\}$, $\theta \in \{\vartheta \in \mathbb{R}^{\mathfrak{d}} \colon i \notin \mathbf{D}^{\vartheta}\}$. Note that Lemma 2.4 shows that there exist $\varepsilon, \mathfrak{C} \in (0, \infty)$ which satisfy for all $\vartheta_1, \vartheta_2 \in \mathbb{R}^{\mathfrak{d}}$ with $\max\{\|\theta - \vartheta_1\|, \|\theta - \vartheta_2\|\} \leq \varepsilon$ that

$$\int_{I_i^{\vartheta_1} \Delta I_i^{\vartheta_2}} \mathfrak{p}(x)\, \lambda(\mathrm{d}x) \leq \mathfrak{C}\|\vartheta_1 - \vartheta_2\|. \tag{2.30}$$

Combining this with Lemma 2.5 (applied for every $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $n \curvearrowleft \mathfrak{d}$, $x \curvearrowleft \theta$, $\mu \curvearrowleft (\mathcal{B}([a, b]^d) \ni A \mapsto \int_A \mathfrak{p}(x)\, \lambda(\mathrm{d}x) \in [0, \infty))$, $(I^y)_{y \in \mathbb{R}^n} \curvearrowleft (I_i^y)_{y \in \mathbb{R}^{\mathfrak{d}}}$ in the notation of Lemma 2.5) demonstrates that there exists $\mathscr{C} \in \mathbb{R}$ such that for all $\vartheta_1, \vartheta_2 \in \mathbb{R}^{\mathfrak{d}}$ with $\max\{\|\theta - \vartheta_1\|, \|\theta - \vartheta_2\|\} \leq \varepsilon$ it holds that

$$\left| \int_{I_i^{\vartheta_1}} \phi(\vartheta_1, x) \mathfrak{p}(x)\, \lambda(\mathrm{d}x) - \int_{I_i^{\vartheta_2}} \phi(\vartheta_2, x) \mathfrak{p}(x)\, \lambda(\mathrm{d}x) \right| \leq \mathscr{C}\|\vartheta_1 - \vartheta_2\|. \tag{2.31}$$

This establishes item (ii). The proof of Corollary 2.6 is thus complete. $\quad\square$

Corollary 2.7. *Assume Setting 2.1. Then*

(i) it holds for all $k \in \mathbb{N} \cap (Hd + H, \eth]$ that

$$\mathbb{R}^{\eth} \ni \theta \mapsto \mathcal{G}_k(\theta) \in \mathbb{R} \tag{2.32}$$

is locally Lipschitz continuous,

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

$$\{\vartheta \in \mathbb{R}^{\eth} : i \notin \mathbf{D}^{\vartheta}\} \ni \theta \mapsto \mathcal{G}_{(i-1)d+j}(\theta) \in \mathbb{R} \tag{2.33}$$

is locally Lipschitz continuous, and

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

$$\{\vartheta \in \mathbb{R}^{\eth} : i \notin \mathbf{D}^{\vartheta}\} \ni \theta \mapsto \mathcal{G}_{Hd+i}(\theta) \in \mathbb{R} \tag{2.34}$$

is locally Lipschitz continuous.

Proof of Corollary 2.7. Observe that (2.7) and Corollary 2.6 establish items (i), (ii), and (iii). The proof of Corollary 2.7 is thus complete.

2.5. Subdifferentials

Definition 2.8 (Subdifferential)**.** Let $n \in \mathbb{N}$, $f \in C(\mathbb{R}^n, \mathbb{R})$, $x \in \mathbb{R}^n$. Then we denote by $\hat{\partial} f(x) \subseteq \mathbb{R}^n$ the set given by

$$\hat{\partial} f(x) = \left\{ y \in \mathbb{R}^n : \liminf_{\mathbb{R}^n \setminus \{0\} \ni h \to 0} \left(\frac{f(x+h) - f(x) - \langle y, h \rangle}{\|h\|} \right) \geq 0 \right\}. \tag{2.35}$$

Definition 2.9 (Limiting subdifferential)**.** Let $n \in \mathbb{N}$, $f \in C(\mathbb{R}^n, \mathbb{R})$, $x \in \mathbb{R}^n$. Then we denote by $\partial f(x) \subseteq \mathbb{R}^n$ the set given by

$$\partial f(x) = \bigcap_{\varepsilon \in (0, \infty)} \overline{\left[\bigcup_{y \in \{z \in \mathbb{R}^n : \|x - z\| < \varepsilon\}} \hat{\partial} f(y) \right]} \tag{2.36}$$

(cf. Definition 2.8).

Lemma 2.10. *Let $n \in \mathbb{N}$, $f \in C(\mathbb{R}^n, \mathbb{R})$, $x \in \mathbb{R}^n$. Then*

$$\partial f(x) = \{y \in \mathbb{R}^n : \exists z = (z_1, z_2) : \mathbb{N} \to \mathbb{R}^n \times \mathbb{R}^n : ([\forall k \in \mathbb{N} : z_2(k) \in \hat{\partial} f(z_1(k))],$$
$$[\limsup_{k \to \infty} (\|z_1(k) - x\| + \|z_2(k) - y\|) = 0])\} \tag{2.37}$$

(cf. Definitions 2.8 and 2.9).

Proof of Lemma 2.10. Note that (2.36) establishes (2.37). The proof of Lemma 2.10 is thus complete.

Lemma 2.11. *Let $n \in \mathbb{N}$, $f \in C(\mathbb{R}^n, \mathbb{R})$, let $U \subseteq \mathbb{R}^n$ be open, assume $f|_U \in C^1(U, \mathbb{R})$, and let $x \in U$. Then $\hat{\partial} f(x) = \partial f(x) = \{(\nabla f)(x)\}$ (cf. Definitions 2.8 and 2.9).*

Proof of Lemma 2.11. This is a direct consequence of, e.g., Rockafellar & Wets [39, Exercise 8.8]. The proof of Lemma 2.11 is thus complete.

Proposition 2.12. *Assume Setting 2.1 and let $\theta \in \mathbb{R}^{\mathfrak{d}}$. Then $\mathcal{G}(\theta) \in \partial \mathcal{L}(\theta)$ (cf. Definition 2.9).*

Proof of Proposition 2.12. Throughout this proof let $\vartheta = (\vartheta_n)_{n \in \mathbb{N}} \colon \mathbb{N} \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$, $i \in \{1, 2, \ldots, H\}$, $j \in \{1, 2, \ldots, d\}$ that $\mathfrak{w}_{i,j}^{\vartheta_n} = \mathfrak{w}_{i,j}^{\theta}$, $\mathfrak{b}_i^{\vartheta_n} = \mathfrak{b}_i^{\theta} - \frac{1}{n}\mathbf{1}_{\mathbf{D}^{\theta}}(i)$, $\mathfrak{v}_i^{\vartheta_n} = \mathfrak{v}_i^{\theta}$, and $\mathfrak{c}^{\vartheta_n} = \mathfrak{c}^{\theta}$. We prove Proposition 2.12 through an application of Lemma 2.10. Observe that for all $n \in \mathbb{N}$, $i \in \{1, 2, \ldots, H\} \backslash \mathbf{D}^{\theta}$ it holds that $\mathfrak{b}_i^{\vartheta_n} = \mathfrak{b}_i^{\theta}$. This implies for all $n \in \mathbb{N}$, $i \in \{1, 2, \ldots, H\} \backslash \mathbf{D}^{\theta}$ that

$$i \notin \mathbf{D}^{\vartheta_n}. \tag{2.38}$$

In addition, note that for all $n \in \mathbb{N}$, $i \in \mathbf{D}^{\theta}$ it holds that $\mathfrak{b}_i^{\vartheta_n} = -\frac{1}{n} < 0$. This shows for all $n \in \mathbb{N}$, $i \in \mathbf{D}^{\theta}$ that

$$i \notin \mathbf{D}^{\vartheta_n}. \tag{2.39}$$

Hence, we obtain for all $n \in \mathbb{N}$ that $\mathbf{D}^{\vartheta_n} = \varnothing$. Combining this with Proposition 2.3 and Lemma 2.11 demonstrates that for all $n \in \mathbb{N}$ it holds that $\hat{\partial}\mathcal{L}(\vartheta_n) = \{(\nabla \mathcal{L})(\vartheta_n)\} = \{\mathcal{G}(\vartheta_n)\}$ (cf. Definition 2.8). Moreover, observe that $\lim_{n \to \infty} \vartheta_n = \theta$. It thus remains to show that $\mathcal{G}(\vartheta_n)$, $n \in \mathbb{N}$, converges to $\mathcal{G}(\theta)$. Note that Corollary 2.7 ensures that for all $k \in \mathbb{N} \cap (Hd + H, \mathfrak{d}]$ it holds that

$$\lim_{n \to \infty} \mathcal{G}_k(\vartheta_n) = \mathcal{G}_k(\theta). \tag{2.40}$$

Furthermore, observe that Corollary 2.7, (2.38) and (2.39) assure that for all $i \in \{1, 2, \ldots, H\} \backslash \mathbf{D}^{\theta}$, $j \in \{1, 2, \ldots, d\}$ it holds that

$$\lim_{n \to \infty} \mathcal{G}_{(i-1)d+j}(\vartheta_n) = \mathcal{G}_{(i-1)d+j}(\theta) \qquad \text{and} \qquad \lim_{n \to \infty} \mathcal{G}_{Hd+i}(\vartheta_n) = \mathcal{G}_{Hd+i}(\theta). \tag{2.41}$$

In addition, note that for all $n \in \mathbb{N}$, $i \in \mathbf{D}^{\theta}$ we have that $I_i^{\vartheta_n} = I_i^{\theta} = \varnothing$. Hence, we obtain for all $i \in \mathbf{D}^{\theta}$, $j \in \{1, 2, \ldots, d\}$ that

$$\lim_{n \to \infty} \mathcal{G}_{(i-1)d+j}(\vartheta_n) = 0 = \mathcal{G}_{(i-1)d+j}(\theta) \qquad \text{and} \qquad \lim_{n \to \infty} \mathcal{G}_{Hd+i}(\vartheta_n) = 0 = \mathcal{G}_{Hd+i}(\theta). \tag{2.42}$$

Combining this, (2.40) and (2.41) demonstrates that $\lim_{n \to \infty} \mathcal{G}(\vartheta_n) = \mathcal{G}(\theta)$. This and Lemma 2.10 assure that $\mathcal{G}(\theta) \in \partial \mathcal{L}(\theta)$. The proof of Proposition 2.12 is thus complete.

3. Existence and uniqueness properties for solutions of gradient flows (GFs)

In this section we employ the local Lipschitz continuity result for the generalized gradient function in Corollary 2.7 from Section 2 to establish existence and uniqueness results for solutions of GF differential equations. Specifically, in Proposition 3.1 in Subsection 3.1 below we prove the existence of solutions GF differential equations, in Lemma 3.2 in Subsection 3.2 below we establish the uniqueness of solutions of GF differential equations among a suitable class of GF solutions, and in Theorem 3.3 in Subsection 3.3 below we combine Proposition 3.1 and Lemma 3.2 to establish the unique existence of solutions of GF differential equations among a suitable class of GF solutions. Theorem 1.1 in the introduction is an immediate consequence of Theorem 3.3.

Roughly speaking, we show in Theorem 3.3 the unique existence of solutions of GF differential equations among the class of GF solutions which satisfy that the set of all degenerate neurons of the GF solution at time $t \in [0, \infty)$ is non-decreasing in the time variable $t \in [0, \infty)$. In other words, in Theorem 3.3 we prove the unique existence of GF solutions with the property that once a neuron has become degenerate it will remain degenerate for subsequent times.

Our strategy of the proof of Theorem 3.3 and Proposition 3.1, respectively, can, loosely speaking, be described as follows. Corollary 2.7 above implies that the components of the generalized gradient function $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ corresponding to non-degenerate neurons are locally Lipschitz continuous so that the classical Picard-Lindelöf local existence and uniqueness theorem for ordinary differential equations can be brought into play for those components. On the other hand, if at some time $t \in [0, \infty)$ the i-th neuron is degenerate, then Proposition 2.2 above shows that the corresponding components of the generalized gradient function $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ vanish. The GF differential equation is thus satisfied if the neuron remains degenerate at all subsequent times $s \in [t, \infty)$. Using these arguments we prove in Proposition 3.1 the existence of GF solutions by induction on the number of non-degenerate neurons of the initial value.

3.1. Existence properties for solutions of GF differential equations

Proposition 3.1. *Assume Setting 2.1 and let $\theta \in \mathbb{R}^{\mathfrak{d}}$. Then there exists $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ which satisfies for all $t \in [0, \infty)$, $s \in [t, \infty)$ that*

$$\Theta_t = \theta - \int_0^t \mathcal{G}(\Theta_u) \, \mathrm{d}u \qquad and \qquad \mathbf{D}^{\Theta_t} \subseteq \mathbf{D}^{\Theta_s}. \tag{3.1}$$

Proof of Proposition 3.1. We prove the statement by induction on the quantity $H - \#(\mathbf{D}^{\theta}) \in \mathbb{N} \cap [0, H]$. Assume first that $H - \#(\mathbf{D}^{\theta}) = 0$, i.e., $\mathbf{D}^{\theta} = \{1, 2, \ldots, H\}$. Observe that this implies that $\mathfrak{w}^{\theta} = 0$ and $\mathfrak{b}^{\theta} = 0$. In the following let $\kappa \in \mathbb{R}$ satisfy

$$\kappa = \int_{[\alpha, \mathscr{6}]^d} f(x) \mathfrak{p}(x) \, \lambda(\mathrm{d}x). \tag{3.2}$$

Note that the Picard–Lindelöf Theorem shows that there exists a unique $c \in C([0, \infty), \mathbb{R})$ which satisfies for all $t \in [0, \infty)$ that

$$c(0) = \mathfrak{c}^{\theta} \qquad \text{and} \qquad c(t) = c(0) + 2\kappa t - 2 \left(\int_{[\alpha, \mathscr{6}]^d} \mathfrak{p}(x) \, \lambda(\mathrm{d}x) \right) \left(\int_0^t c(s) \, \mathrm{d}s \right). \tag{3.3}$$

Next let $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ satisfy for all $t \in [0, \infty)$, $i \in \{1, 2, \ldots, H\}$, $j \in \{1, 2, \ldots, d\}$ that

$$\mathfrak{w}_{i,j}^{\Theta_t} = \mathfrak{w}_{i,j}^{\theta} = \mathfrak{b}_i^{\Theta_t} = \mathfrak{b}_i^{\theta} = 0, \qquad \mathfrak{v}_i^{\Theta_t} = \mathfrak{v}_i^{\theta}, \qquad \text{and} \qquad \mathfrak{c}^{\Theta_t} = c(t). \tag{3.4}$$

Observe that (2.7), (3.3), and (3.4) ensure for all $t \in [0, \infty)$ that

$$
\begin{aligned}
\mathfrak{c}^{\Theta_t} &= \mathfrak{c}^{\theta} + 2\kappa t - 2\left(\int_{[\mathfrak{a},\mathfrak{b}]^d} \mathfrak{p}(x)\,\lambda(\mathrm{d}x)\right)\left(\int_0^t \mathfrak{c}^{\Theta_s}\,\mathrm{d}s\right) \\
&= \mathfrak{c}^{\theta} - 2\int_0^t\left(-\kappa + \int_{[\mathfrak{a},\mathfrak{b}]^d} \mathfrak{c}^{\Theta_s}\mathfrak{p}(x)\,\lambda(\mathrm{d}x)\right)\mathrm{d}s \\
&= \mathfrak{c}^{\theta} - 2\int_0^t\int_{[\mathfrak{a},\mathfrak{b}]^d}\left(\mathfrak{c}^{\Theta_s} + \sum_{i=1}^{H}[\mathfrak{v}_i^{\Theta_s}\max\{\mathfrak{b}_i^{\Theta_s} + \sum_{j=1}^{d}\mathfrak{w}_{i,j}^{\Theta_s}x_j,0\}] - f(x)\right)\mathfrak{p}(x)\,\lambda(\mathrm{d}x)\,\mathrm{d}s \\
&= \mathfrak{c}^{\theta} - 2\int_0^t\int_{[\mathfrak{a},\mathfrak{b}]^d}(\mathcal{N}^{\Theta_s}(x) - f(x))\mathfrak{p}(x)\,\lambda(\mathrm{d}x)\,\mathrm{d}s = \mathfrak{c}^{\theta} - \int_0^t \mathcal{G}_{\mathfrak{d}}(\Theta_s)\,\mathrm{d}s.
\end{aligned}
\tag{3.5}
$$

Next note that (3.4) and (2.7) show for all $t \in [0, \infty)$, $i \in \mathbb{N} \cap [1, \mathfrak{d})$ that $\mathbf{D}^{\Theta_t} = \{1, 2, \ldots, H\}$ and $\mathcal{G}_i(\Theta_t) = 0$. Combining this with (3.4) and (3.5) proves that Θ satisfies (3.1). This establishes the claim in the case $\#(\mathbf{D}^{\theta}) = H$.

For the induction step assume that $\#(\mathbf{D}^{\theta}) < H$ and assume that for all $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ with $\#(\mathbf{D}^{\vartheta}) > \#(\mathbf{D}^{\theta})$ there exists $\Theta \in C([0,\infty), \mathbb{R}^{\mathfrak{d}})$ which satisfies for all $t \in [0, \infty)$, $s \in [t, \infty)$ that $\Theta_t = \vartheta - \int_0^t \mathcal{G}(\Theta_u)\,\mathrm{d}u$ and $\mathbf{D}^{\Theta_t} \subseteq \mathbf{D}^{\Theta_s}$. In the following let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy

$$
U = \{\vartheta \in \mathbb{R}^{\mathfrak{d}} : \mathbf{D}^{\vartheta} \subseteq \mathbf{D}^{\theta}\}
\tag{3.6}
$$

and let $\mathfrak{G} : U \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\vartheta \in U$, $i \in \{1, 2, \ldots, \mathfrak{d}\}$ that

$$
\mathfrak{G}_i(\vartheta) = \begin{cases} 0 & : i \in \{(\ell-1)d + j : \ell \in \mathbf{D}^{\theta}, j \in \mathbb{N} \cap [1, d]\} \cup \{Hd + \ell : \ell \in \mathbf{D}^{\theta}\} \\ \mathcal{G}_i(\vartheta) & : \text{else.} \end{cases}
\tag{3.7}
$$

Observe that (3.6) assures that $U \subseteq \mathbb{R}^{\mathfrak{d}}$ is open. In addition, note that Corollary 2.7 implies that $\mathfrak{G}$ is locally Lipschitz continuous. Combining this with the Picard–Lindelöf Theorem demonstrates that there exist a unique maximal $\tau \in (0, \infty]$ and $\Psi \in C([0, \tau), U)$ which satisfy for all $t \in [0, \tau)$ that

$$
\Psi_t = \theta - \int_0^t \mathfrak{G}(\Psi_u)\,\mathrm{d}u.
\tag{3.8}
$$

Next observe that (3.7) ensures that for all $t \in [0, \tau)$, $i \in \mathbf{D}^{\theta}$, $j \in \{1, 2, \ldots, d\}$ we have that

$$
\mathfrak{w}_{i,j}^{\Psi_t} = \mathfrak{w}_{i,j}^{\theta} = \mathfrak{b}_i^{\Psi_t} = \mathfrak{b}_i^{\theta} = 0 \qquad \text{and} \qquad \mathfrak{v}_i^{\Psi_t} = \mathfrak{v}_i^{\theta}.
\tag{3.9}
$$

This, (3.7), and (2.7) demonstrate for all $t \in [0, \tau)$ that $\mathcal{G}(\Psi_t) = \mathfrak{G}(\Psi_t)$. In addition, note that (3.6) and (3.9) imply for all $t \in [0, \tau)$ that $\mathbf{D}^{\Psi_t} = \mathbf{D}^{\theta}$. Hence, if $\tau = \infty$ then Ψ satisfies (3.1). Next assume that $\tau < \infty$. Observe that the Cauchy-Schwarz inequality and [37, Lemma 3.1] prove for all $s, t \in [0, \tau)$ with $s \leq t$ that

$$
\begin{aligned}
\|\Psi_t - \Psi_s\| &\leq \int_s^t \|\mathcal{G}(\Psi_u)\|\,\mathrm{d}u \leq (t-s)^{1/2}\left[\int_s^t \|\mathcal{G}(\Psi_u)\|^2\,\mathrm{d}u\right]^{1/2} \\
&\leq (t-s)^{1/2}\left[\int_0^t \|\mathcal{G}(\Psi_u)\|^2\,\mathrm{d}u\right]^{1/2} = (t-s)^{1/2}(\mathcal{L}(\Psi_0) - \mathcal{L}(\Psi_t))^{1/2} \\
&\leq (t-s)^{1/2}(\mathcal{L}(\Psi_0))^{1/2}.
\end{aligned}
\tag{3.10}
$$

Hence, we obtain for all $(t_n)_{n\in\mathbb{N}} \subseteq [0, \tau)$ with $\liminf_{n\to\infty} t_n = \tau$ that (Ψ_{t_n}) is a Cauchy sequence. This implies that $\vartheta := \lim_{t\uparrow\tau} \Psi_t \in \mathbb{R}^{\mathfrak{d}}$ exists. Furthermore, note that the fact that τ is maximal proves that $\vartheta \notin U$. Therefore, we have that $\mathbf{D}^{\vartheta} \backslash \mathbf{D}^{\theta} \neq \emptyset$. Moreover, observe that (3.9) shows that for all $i \in \mathbf{D}^{\theta}$, $j \in \{1, 2, \ldots, d\}$ it holds that $\mathrm{w}_{i,j}^{\vartheta} = \mathfrak{b}_i^{\vartheta} = 0$ and, therefore, $i \in \mathbf{D}^{\vartheta}$. This demonstrates that $\#(\mathbf{D}^{\vartheta}) > \#(\mathbf{D}^{\theta})$. Combining this with the induction hypothesis ensures that there exists $\Phi \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ which satisfies for all $t \in [0, \infty)$, $s \in [t, \infty)$ that

$$\Phi_t = \vartheta - \int_0^t \mathcal{G}(\Phi_u)\,\mathrm{d}u \qquad \text{and} \qquad \mathbf{D}^{\Phi_t} \subseteq \mathbf{D}^{\Phi_s}. \tag{3.11}$$

In the following let $\Theta: [0, \infty) \to \mathbb{R}^{\mathfrak{d}}$ satisfy for all $t \in [0, \infty)$ that

$$\Theta_t = \begin{cases} \Psi_t & : t \in [0, \tau) \\ \Phi_{t-\tau} & : t \in [\tau, \infty). \end{cases} \tag{3.12}$$

Note that the fact that $\vartheta = \lim_{t\uparrow\tau} \Psi_t$ and the fact that $\Phi_0 = \vartheta$ imply that Θ is continuous. Furthermore, observe that the fact that $\mathcal{G}$ is locally bounded and (3.8) ensure that

$$\Theta_\tau = \vartheta = \lim_{t\uparrow\tau} \Psi_t = \lim_{t\uparrow\tau}\left[\theta - \int_0^t \mathcal{G}(\Psi_s)\,\mathrm{d}s\right] = \theta - \int_0^\tau \mathcal{G}(\Psi_s)\,\mathrm{d}s = \theta - \int_0^\tau \mathcal{G}(\Theta_s)\,\mathrm{d}s. \tag{3.13}$$

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

$$\begin{aligned} \Theta_t &= (\Theta_t - \Theta_\tau) + \Theta_\tau = (\Phi_{t-\tau} - \Phi_0) + \Theta_\tau = -\int_0^{t-\tau} \mathcal{G}(\Phi_s)\,\mathrm{d}s + \theta - \int_0^\tau \mathcal{G}(\Theta_s)\,\mathrm{d}s \\ &= -\int_t^\tau \mathcal{G}(\Theta_s) + \theta - \int_0^\tau \mathcal{G}(\Theta_s)\,\mathrm{d}s = \theta - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s. \end{aligned} \tag{3.14}$$

This shows that Θ satisfies (3.1). The proof of Proposition 3.1 is thus complete. $\quad\blacksquare$

3.2. Uniqueness properties for solutions of GF differential equations

Lemma 3.2. *Assume Setting 2.1 and let $\theta \in \mathbb{R}^{\mathfrak{d}}$, $\Theta^1, \Theta^2 \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ satisfy for all $t \in [0, \infty)$, $s \in [t, \infty)$, $k \in \{1, 2\}$ that*

$$\Theta_t^k = \theta - \int_0^t \mathcal{G}(\Theta_u^k)\,\mathrm{d}u \qquad \text{and} \qquad \mathbf{D}^{\Theta_t^k} \subseteq \mathbf{D}^{\Theta_s^k}. \tag{3.15}$$

Then it holds for all $t \in [0, \infty)$ that $\Theta_t^1 = \Theta_t^2$.

Proof of Lemma 3.2. Assume for the sake of contradiction that there exists $t \in [0, \infty)$ such that $\Theta_t^1 \neq \Theta_t^2$. By translating the variable t if necessary, we may assume without loss of generality that $\inf\{t \in [0, \infty): \Theta_t^1 \neq \Theta_t^2\} = 0$. Next note that the fact that Θ^1 and Θ^2 are continuous implies that there exists $\delta \in (0, \infty)$ which satisfies for all $t \in [0, \delta]$, $k \in \{1, 2\}$ that $\mathbf{D}^{\Theta_t^k} \subseteq \mathbf{D}^{\theta}$. Furthermore, observe that (3.15) ensures for all $t \in [0, \infty)$, $i \in \mathbf{D}^{\theta}$, $k \in \{1, 2\}$ that $i \in \mathbf{D}^{\Theta_t^k}$. Hence, we obtain for all $t \in [0, \infty)$, $i \in \mathbf{D}^{\theta}$, $j \in \{1, 2, \ldots, d\}$, $k \in \{1, 2\}$ that

$$\mathcal{G}_{(i-1)d+j}(\Theta_t^k) = \mathcal{G}_{Hd+i}(\Theta_t^k) = \mathcal{G}_{H(d+1)+i}(\Theta_t^k) = 0. \tag{3.16}$$

In addition, note that the fact that Θ^1 and Θ^2 are continuous implies that there exists a compact $K \subseteq \{\vartheta \in \mathbb{R}^{\mathfrak{d}} : \mathbf{D}^{\vartheta} \subseteq \mathbf{D}^{\theta}\}$ which satisfies for all $t \in [0, \delta]$, $k \in \{1, 2\}$ that $\Theta_t^k \in K$. Moreover, observe that Corollary 2.7 proves that for all $i \in \{1, 2, \dots, H\} \backslash \mathbf{D}^{\theta}$, $j \in \{1, 2, \dots, d\}$ it holds that $\mathcal{G}_{(i-1)d+j}, \mathcal{G}_{Hd+i}, \mathcal{G}_{H(d+1)+i}, \mathcal{G}_{\mathfrak{d}} : K \to \mathbb{R}$ are Lipschitz continuous. This and (3.16) show that there exists $L \in (0, \infty)$ such that for all $t \in [0, \delta]$ we have that

$$\|\mathcal{G}(\Theta_t^1) - \mathcal{G}(\Theta_t^2)\| \leq L \|\Theta_t^1 - \Theta_t^2\|. \tag{3.17}$$

In the following let $M \colon [0, \infty) \to [0, \infty)$ satisfy for all $t \in [0, \infty)$ that $M_t = \sup_{s \in (0, t]} \|\Theta_s^1 - \Theta_s^2\|$. Note that the fact that $\inf\{t \in [0, \infty) : \Theta_t^1 \neq \Theta_t^2\} = 0$ proves for all $t \in (0, \infty)$ that $M_t > 0$. Moreover, observe that (3.17) ensures for all $t \in (0, \delta)$ that

$$\begin{aligned}
\|\Theta_t^1 - \Theta_t^2\| &= \left\| \int_0^t \mathcal{G}(\Theta_u^1) \, du - \int_0^t \mathcal{G}(\Theta_u^2) \, du \right\| \leq \int_0^t \|\mathcal{G}(\Theta_u^1) - \mathcal{G}(\Theta_u^2)\| \, du \\
&\leq L \int_0^t \|\Theta_u^1 - \Theta_u^2\| \, du \leq L t M_t.
\end{aligned} \tag{3.18}$$

Combining this with the fact that M is non-decreasing shows for all $t \in (0, \delta)$, $s \in (0, t]$ that

$$\|\Theta_s^1 - \Theta_s^2\| \leq L s M_s \leq L t M_t. \tag{3.19}$$

This demonstrates for all $t \in (0, \min\{L^{-1}, \delta\})$ that

$$0 < M_t \leq L t M_t < M_t, \tag{3.20}$$

which is a contradiction. The proof of Lemma 3.2 is thus complete.

3.3. Existence and uniqueness properties for solutions of GF differential equations

Theorem 3.3. *Assume Setting 2.1 and let $\theta \in \mathbb{R}^{\mathfrak{d}}$. Then there exists a unique $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ which satisfies for all $t \in [0, \infty)$, $s \in [t, \infty)$ that*

$$\Theta_t = \theta - \int_0^t \mathcal{G}(\Theta_u) \, du \qquad and \qquad \mathbf{D}^{\Theta_t} \subseteq \mathbf{D}^{\Theta_s}. \tag{3.21}$$

Proof of Theorem 3.3. Proposition 3.1 establishes the existence and Lemma 3.2 establishes the uniqueness. The proof of Theorem 3.3 is thus complete.

4. Semialgebraic sets and functions

In this section we establish in Corollary 4.10 in Subsection 4.3 below that under the assumption that both the target function $f \colon [\mathfrak{a}, \mathfrak{b}]^d \to \mathbb{R}$ and the unnormalized density function $\mathfrak{p} \colon [\mathfrak{a}, \mathfrak{b}]^d \to [0, \infty)$ are piecewise polynomial in the sense of Definition 4.9 in Subsection 4.3 we have that the risk function $\mathcal{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ is a semialgebraic function in the sense of Definition 4.3 in Subsection 4.1. In Definition 4.9 we specify precisely what we mean by a piecewise polynomial function, in Definition 4.2 in Subsection 4.1 we recall the notion of a semialgebraic set, and in Definition 4.3 we recall the notion of a semialgebraic function. In the scientific literature Definitions 4.2 and 4.3 can in a slightly different

presentational form, e.g., be found in Bierstone & Milman [40, Definitions 1.1 and 1.2] and Attouch et al. [8, Definition 2.1].

Note that the risk function $\mathcal{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ is given through a parametric integral in the sense that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ we have that

$$\mathcal{L}(\theta) = \int_{[\mathfrak{a},\mathfrak{b}]^d} (f(y) - \mathcal{N}^{\theta}(y))^2 \, \mathfrak{p}(y) \, \lambda(\mathrm{d}y). \tag{4.1}$$

In general, parametric integrals of semialgebraic functions are no longer semialgebraic functions and the characterization of functions that can occur as such integrals is quite involved (cf. Kaiser [41]). This is the reason why we introduce in Definition 4.6 in Subsection 4.2 below a suitable subclass of the class of semialgebraic functions which is rich enough to contain the realization functions of ANNs with ReLU activation (cf. (4.30) in Subsection 4.2 below) and which can be shown to be closed under integration (cf. Proposition 4.8 in Subsection 4.2 below for the precise statement).

4.1. Semialgebraic sets and functions

Definition 4.1 (Set of polynomials)**.** Let $n \in \mathbb{N}_0$. Then we denote by $\mathscr{P}_n \subseteq C(\mathbb{R}^n, \mathbb{R})$ the set[†] of all polynomials from $\mathbb{R}^n$ to $\mathbb{R}$.

Definition 4.2 (Semialgebraic sets)**.** Let $n \in \mathbb{N}$ and let $A \subseteq \mathbb{R}^n$ be a set. Then we say that A is a semialgebraic set if and only if there exist $k \in \mathbb{N}$ and $(P_{i,j,\ell})_{(i,j,\ell)\in\{1,2,\dots,k\}^2\times\{0,1\}} \subseteq \mathscr{P}_n$ such that

$$A = \bigcup_{i=1}^{k} \bigcap_{j=1}^{k} \{x \in \mathbb{R}^n \colon P_{i,j,0}(x) = 0 < P_{i,j,1}(x)\} \tag{4.2}$$

(cf. Definition 4.1).

Definition 4.3 (Semialgebraic functions)**.** Let $m, n \in \mathbb{N}$ and let $f \colon \mathbb{R}^n \to \mathbb{R}^m$ be a function. Then we say that f is a semialgebraic function if and only if it holds that $\{(x, f(x)) \colon x \in \mathbb{R}^n\} \subseteq \mathbb{R}^{m+n}$ is a semialgebraic set (cf. Definition 4.2).

Lemma 4.4. *Let $n \in \mathbb{N}$ and let $f, g \colon \mathbb{R}^n \to \mathbb{R}$ be semialgebraic functions (cf. Definition 4.3). Then*

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

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

Proof of Lemma 4.4. Note that, e.g., Coste [42, Corollary 2.9] (see, e.g., also Bierstone & Milman [40, Section 1]) establishes items (i) and (ii). The proof of Lemma 4.4 is thus complete.

4.2. On the semialgebraic property of certain parametric integrals

Definition 4.5 (Set of rational functions)**.** Let $n \in \mathbb{N}$. Then we denote by $\mathscr{R}_n$ the set given by

$$\mathscr{R}_n = \left\{ R \colon \mathbb{R}^n \to \mathbb{R} \colon \left[\exists \, P, Q \in \mathscr{P}_n \colon \forall \, x \in \mathbb{R}^n \colon R(x) = \begin{cases} \frac{P(x)}{Q(x)} & \colon Q(x) \neq 0 \\ 0 & \colon Q(x) = 0 \end{cases} \right] \right\} \tag{4.3}$$

(cf. Definition 4.1).

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

Definition 4.6. Let $m \in \mathbb{N}$, $n \in \mathbb{N}_0$. Then we denote by $\mathscr{A}_{m,n}$ the $\mathbb{R}$-vector space given by

$$\mathscr{A}_{m,n} = \text{span}\Big(\Big\{f \colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R} \colon \Big[\exists\, r \in \mathbb{N},\ A_1, A_2, \ldots, A_r \in \{\{0\}, [0, \infty), (0, \infty)\},$$

$$R \in \mathscr{R}_m,\ Q \in \mathscr{P}_n,\ P = (P_{i,j})_{(i,j) \in \{1,2,\ldots,r\} \times \{0,1,\ldots,n\}} \subseteq \mathscr{P}_m \colon \forall\, \theta \in \mathbb{R}^m,\ x = (x_1, \ldots, x_n) \in \mathbb{R}^n \colon$$

$$f(\theta, x) = R(\theta)Q(x)\big[\textstyle\prod_{i=1}^{r} \mathbf{1}_{A_i}(P_{i,0}(\theta) + \sum_{j=1}^{n} P_{i,j}(\theta)x_j)\big]\Big]\Big\}\Big) \quad (4.4)$$

(cf. Definitions 4.1 and 4.5).

Lemma 4.7. *Let* $m \in \mathbb{N}$, $f \in \mathscr{A}_{m,0}$ *(cf. Definition 4.6). Then* f *is semialgebraic (cf. Definition 4.3).*

Proof of Lemma 4.7. Throughout this proof let $r \in \mathbb{N}$, $A_1, A_2, \ldots, A_r \in \{\{0\}, [0, \infty), (0, \infty)\}$, $R \in \mathscr{R}_m$, $P = (P_i)_{i \in \{1,2,\ldots,r\}} \subseteq \mathscr{P}_m$, and let $g \colon \mathbb{R}^m \to \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^m$ that

$$g(\theta) = R(\theta) \textstyle\prod_{i=1}^{r} \mathbf{1}_{A_i}(P_i(\theta)) \quad (4.5)$$

(cf. Definitions 4.1 and 4.5). Due to the fact that sums of semialgebraic functions are again semialgebraic (cf. Lemma 4.4), it suffices to show that g is semialgebraic. Furthermore, observe that for all $y \in \mathbb{R}$ it holds that $\mathbf{1}_{(0,\infty)}(y) = 1 - \mathbf{1}_{[0,\infty)}(-y)$ and $\mathbf{1}_{\{0\}}(y) = \mathbf{1}_{[0,\infty)}(y)\mathbf{1}_{[0,\infty)}(-y)$. Hence, by linearity we may assume for all $i \in \{1, 2, \ldots, r\}$ that $A_i = [0, \infty)$. Next let $Q_1, Q_2 \in \mathscr{P}_m$ satisfy for all $x \in \mathbb{R}^m$ that

$$R(x) = \begin{cases} \frac{Q_1(x)}{Q_2(x)} & : Q_2(x) \neq 0 \\ 0 & : Q_2(x) = 0. \end{cases} \quad (4.6)$$

Note that the graph of $\mathbb{R}^m \ni \theta \mapsto R(\theta) \in \mathbb{R}$ is given by

$$\{(\theta, y) \in \mathbb{R}^m \times \mathbb{R} \colon Q_2(\theta) = 0,\ y = 0\} \cup \{(\theta, y) \in \mathbb{R}^m \times \mathbb{R} \colon Q_2(\theta) \neq 0,\ Q_2(\theta)y - Q_1(\theta) = 0\}. \quad (4.7)$$

Since both of these sets are described by polynomial equations and inequalities, it follows that $\mathbb{R}^m \ni \theta \mapsto R(\theta) \in \mathbb{R}$ is semialgebraic. In addition, observe that for all $i \in \{1, 2, \ldots, r\}$ the graph of $\mathbb{R}^m \ni \theta \mapsto \mathbf{1}_{[0,\infty)}(P_i(\theta)) \in \mathbb{R}$ is given by

$$\{(\theta, y) \in \mathbb{R}^m \times \mathbb{R} \colon P_i(\theta) < 0,\ y = 0\} \cup \{(\theta, y) \in \mathbb{R}^m \times \mathbb{R} \colon P_i(\theta) \geq 0,\ y = 1\}. \quad (4.8)$$

This demonstrates for all $i \in \{1, 2, \ldots, r\}$ that $\mathbb{R}^m \ni \theta \mapsto \mathbf{1}_{[0,\infty)}(P_i(\theta)) \in \mathbb{R}$ is semialgebraic. Combining this and (4.5) with Lemma 4.4 demonstrates that g is semialgebraic. The proof of Lemma 4.7 is thus complete. $\qquad\blacksquare$

Proposition 4.8. *Let* $m, n \in \mathbb{N}$, $\mathscr{a} \in \mathbb{R}$, $\mathscr{b} \in (\mathscr{a}, \infty)$, $f \in \mathscr{A}_{m,n}$ *(cf. Definition 4.6). Then*

$$\left[\mathbb{R}^m \times \mathbb{R}^{n-1} \ni (\theta, x_1, \ldots, x_{n-1}) \mapsto \int_{\mathscr{a}}^{\mathscr{b}} f(\theta, x_1, \ldots, x_n)\, dx_n \in \mathbb{R}\right] \in \mathscr{A}_{m,n-1}. \quad (4.9)$$

Proof of Proposition 4.8. By linearity of the integral it suffices to consider a function f of the form

$$f(\theta, x) = R(\theta)Q(x) \prod_{i=1}^{r} \mathbf{1}_{A_i}(P_{i,0}(\theta) + \textstyle\sum_{j=1}^{n} P_{i,j}(\theta)x_j) \quad (4.10)$$

where $r \in \mathbb{N}$, $(P_{i,j})_{(i,j) \in \{1,2,\ldots,r\} \times \{0,1,\ldots,n\}} \subseteq \mathscr{P}_m$, $A_1, A_2, \ldots, A_r \in \{\{0\}, (0,\infty), [0,\infty)\}$, $Q \in \mathscr{P}_n$, and $R \in \mathscr{R}_m$ (cf. Definitions 4.1 and 4.5). Moreover, note that for all $y \in \mathbb{R}$ it holds that $\mathbf{1}_{(0,\infty)}(y) = 1 - \mathbf{1}_{[0,\infty)}(-y)$ and $\mathbf{1}_{\{0\}}(y) = \mathbf{1}_{[0,\infty)}(y)\mathbf{1}_{[0,\infty)}(-y)$. Hence, by linearity we may assume that $A_i = [0,\infty)$ for all $i \in \{1,2,\ldots,r\}$. Furthermore, by linearity we may assume that Q is of the form

$$Q(x_1, \ldots, x_n) = \textstyle\prod_{\ell=1}^n (x_\ell)^{i_\ell} \tag{4.11}$$

with $i_1, i_2, \ldots, i_n \in \mathbb{N}_0$. In the following let $\mathfrak{s} \colon \mathbb{R} \to \mathbb{R}$ satisfy for all $x \in \mathbb{R}$ that $\mathfrak{s}(x) = \mathbf{1}_{(0,\infty)}(x) - \mathbf{1}_{(0,\infty)}(-x)$, for every $\theta \in \mathbb{R}^m$, $k \in \{-1,0,1\}$ let $\mathcal{S}_k^\theta \subseteq \{1,2,\ldots,r\}$ satisfy $\mathcal{S}_k^\theta = \{i \in \{1,2,\ldots,r\} \colon \mathfrak{s}(P_{i,n}(\theta)) = k\}$, and for every $i \in \{1,2,\ldots,r\}$ let $Z_i \colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ satisfy for all $(\theta, x) \in \mathbb{R}^m \times \mathbb{R}^n$ that

$$Z_i(\theta, x) = -P_{i,0}(\theta) - \textstyle\sum_{j=1}^{n-1} P_{i,j}(\theta)x_j. \tag{4.12}$$

Observe that (4.10), (4.11), and (4.12) imply for all $\theta \in \mathbb{R}^m$, $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ that

$$f(\theta, x) = R(\theta)\left(\textstyle\prod_{\ell=1}^n (x_\ell)^{i_\ell}\right)\left(\textstyle\prod_{i=1}^r \mathbf{1}_{[0,\infty)}(P_{i,n}(\theta)x_n - Z_i(\theta, x))\right). \tag{4.13}$$

This shows that $f(\theta, x)$ can only be nonzero if

$$\begin{aligned}
\forall\, i \in \mathcal{S}_1^\theta \colon\ & x_n \geq \frac{Z_i(\theta, x)}{P_{i,n}(\theta)}, \\
\forall\, i \in \mathcal{S}_{-1}^\theta \colon\ & x_n \leq \frac{Z_i(\theta, x)}{P_{i,n}(\theta)}, \\
\forall\, i \in \mathcal{S}_0^\theta \colon\ & -Z_i(\theta, x) \geq 0.
\end{aligned} \tag{4.14}$$

Hence, if for given $\theta \in \mathbb{R}^m$, $(x_1, \ldots, x_{n-1}) \in \mathbb{R}^{n-1}$ there exists $x_n \in [\mathscr{a}, \mathscr{b}]$ which satisfies these conditions then (4.13) and the fact that $\int y^{i_n}\, dy = \frac{1}{i_n+1}y^{i_n+1}$ imply that

$$\begin{aligned}
& \int_{\mathscr{a}}^{\mathscr{b}} f(\theta, x_1, \ldots, x_n)\, dx_n \\
&= \frac{R(\theta)}{i_n + 1}\left(\textstyle\prod_{\ell=1}^{n-1} x_\ell^{i_\ell}\right)\left[\left(\min\left\{\mathscr{b}, \min_{j \in \mathcal{S}_{-1}^\theta} \frac{Z_j(\theta, x)}{P_{j,n}(\theta)}\right\}\right)^{i_n+1} - \left(\max\left\{\mathscr{a}, \max_{j \in \mathcal{S}_1^\theta} \frac{Z_j(\theta, x)}{P_{j,n}(\theta)}\right\}\right)^{i_n+1}\right].
\end{aligned} \tag{4.15}$$

Otherwise, we have that $\int_{\mathscr{a}}^{\mathscr{b}} f(\theta, x_1, \ldots, x_n)\, dx_n = 0$. It remains to write these expressions in the different cases as a sum of functions of the required form in Definition 4.6 by introducing suitable indicator functions. Note that there are four possible cases where the integral is nonzero:

- It holds that $\mathscr{a} < \max_{j \in \mathcal{S}_1^\theta} \frac{Z_j(\theta,x)}{P_{j,n}(\theta)} < \min_{j \in \mathcal{S}_{-1}^\theta} \frac{Z_j(\theta,x)}{P_{j,n}(\theta)} < \mathscr{b}$. In this case, we have

$$\begin{aligned}
& \int_{\mathscr{a}}^{\mathscr{b}} f(\theta, x_1, \ldots, x_n)\, dx_n \\
&= \frac{R(\theta)}{i_n + 1}\left(\textstyle\prod_{\ell=1}^{n-1} x_\ell^{i_\ell}\right)\left[\left(\min_{j \in \mathcal{S}_{-1}^\theta} \frac{Z_j(\theta, x)}{P_{j,n}(\theta)}\right)^{i_n+1} - \left(\max_{j \in \mathcal{S}_1^\theta} \frac{Z_j(\theta, x)}{P_{j,n}(\theta)}\right)^{i_n+1}\right].
\end{aligned} \tag{4.16}$$

- It holds that $\mathscr{a} < \max_{j \in \mathcal{S}_1^\theta} \frac{Z_j(\theta,x)}{P_{j,n}(\theta)} < \mathscr{b} \leq \min_{j \in \mathcal{S}_{-1}^\theta} \frac{Z_j(\theta,x)}{P_{j,n}(\theta)}$. In this case, we have

$$\int_{\mathscr{a}}^{\mathscr{b}} f(\theta, x_1, \ldots, x_n)\, dx_n = \frac{R(\theta)}{i_n + 1}\left(\textstyle\prod_{\ell=1}^{n-1} x_\ell^{i_\ell}\right)\left[\mathscr{b}^{i_n+1} - \left(\max_{j \in \mathcal{S}_1^\theta} \frac{Z_j(\theta, x)}{P_{j,n}(\theta)}\right)^{i_n+1}\right]. \tag{4.17}$$

- It holds that $\max_{j\in\mathcal{S}_1^\theta}\frac{Z_j(\theta,x)}{P_{j,n}(\theta)} \leq \alpha < \min_{j\in\mathcal{S}_{-1}^\theta}\frac{Z_j(\theta,x)}{P_{j,n}(\theta)} < \ell$. In this case, we have

$$\int_\alpha^\ell f(\theta,x_1,\ldots,x_n)\,\mathrm{d}x_n = \frac{R(\theta)}{i_n+1}\Big(\prod_{\ell=1}^{n-1}x_\ell^{i_\ell}\Big)\bigg[\Big(\min_{j\in\mathcal{S}_{-1}^\theta}\frac{Z_j(\theta,x)}{P_{j,n}(\theta)}\Big)^{i_n+1} - \alpha^{i_n+1}\bigg]. \tag{4.18}$$

- It holds that $\max_{j\in\mathcal{S}_1^\theta}\frac{Z_j(\theta,x)}{P_{j,n}(\theta)} \leq \alpha < \ell \leq \min_{j\in\mathcal{S}_{-1}^\theta}\frac{Z_j(\theta,x)}{P_{j,n}(\theta)}$. In this case, we have

$$\int_\alpha^\ell f(\theta,x_1,\ldots,x_n)\,\mathrm{d}x_n = \frac{R(\theta)}{i_n+1}\Big(\prod_{\ell=1}^{n-1}x_\ell^{i_\ell}\Big)\big[\ell^{i_n+1} - \alpha^{i_n+1}\big]. \tag{4.19}$$

Since these four cases are disjoint, by summing over all possible choices $A, B, C \subseteq \{1,2,\ldots,r\}$ of the sets $\mathcal{S}_k^\theta$, $k \in \{-1,0,1\}$, and all choices of (non-empty) subsets $\mathcal{I}, \mathcal{J}$ of $\mathcal{S}_1^\theta$, $\mathcal{S}_{-1}^\theta$ where the maximal/minimal values are achieved, we can write

$$\int_\alpha^\ell f(\theta,x_1,\ldots,x_n)\,\mathrm{d}x_n = \frac{R(\theta)}{i_n+1}\Big(\prod_{\ell=1}^{n-1}x_\ell^{i_\ell}\Big)[(I)+(II)+(III)+(IV)], \tag{4.20}$$

where $(I),(II),(III),(IV)$ denote the functions of $\theta \in \mathbb{R}^m$ and $(x_1,\ldots,x_{n-1}) \in \mathbb{R}^{n-1}$ given by

$$
\begin{aligned}
(I) = \sum_{A\dot\cup B\dot\cup C=\{1,\ldots,r\}} & \bigg[\prod_{j\in A}\mathbf{1}_{(0,\infty)}(P_{j,n}(\theta))\prod_{j\in B}\mathbf{1}_{(0,\infty)}(-P_{j,n}(\theta))\prod_{j\in C}\Big(\mathbf{1}_{\{0\}}(P_{j,n}(\theta))\mathbf{1}_{[0,\infty)}(-Z_j(\theta,x))\Big)\bigg]\\
& \sum_{\varnothing\neq\mathcal{I}\subseteq A}\sum_{\varnothing\neq\mathcal{J}\subseteq B}\bigg[\bigg[\prod_{i\in\mathcal{I}}\Big(\mathbf{1}_{(\alpha,\ell)}\Big(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}\Big)\mathbf{1}_{\{0\}}\Big(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}-\frac{Z_{\min\mathcal{I}}(\theta,x)}{P_{\min\mathcal{I},n}(\theta)}\Big)\Big)\\
& \times\prod_{j\in A\setminus\mathcal{I}}\mathbf{1}_{(0,\infty)}\Big(\frac{Z_{\min\mathcal{I}}(\theta,x)}{P_{\min\mathcal{I},n}(\theta)}-\frac{Z_j(\theta,x)}{P_{j,n}(\theta)}\Big)\prod_{i\in\mathcal{J}}\Big(\mathbf{1}_{(\alpha,\ell)}\Big(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}\Big)\mathbf{1}_{\{0\}}\Big(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}-\frac{Z_{\min\mathcal{J}}(\theta,x)}{P_{\min\mathcal{J},n}(\theta)}\Big)\Big)\\
& \times\prod_{j\in B\setminus\mathcal{J}}\mathbf{1}_{(0,\infty)}\Big(\frac{Z_j(\theta,x)}{P_{j,n}(\theta)}-\frac{Z_{\min\mathcal{J}}(\theta,x)}{P_{\min\mathcal{J},n}(\theta)}\Big)\mathbf{1}_{(0,\infty)}\Big(\frac{Z_{\min\mathcal{J}}(\theta,x)}{P_{\min\mathcal{J},n}(\theta)}-\frac{Z_{\min\mathcal{I}}(\theta,x)}{P_{\min\mathcal{I},n}(\theta)}\Big)\bigg]\\
& \times\bigg[\Big(\frac{Z_{\min\mathcal{J}}(\theta,x)}{P_{\min\mathcal{J},n}(\theta)}\Big)^{i_n+1}-\Big(\frac{Z_{\min\mathcal{I}}(\theta,x)}{P_{\min\mathcal{I},n}(\theta)}\Big)^{i_n+1}\bigg]\bigg],
\end{aligned}
\tag{4.21}
$$

$$
\begin{aligned}
(II) = \sum_{A\dot\cup B\dot\cup C=\{1,\ldots,r\}} & \bigg[\prod_{j\in A}\mathbf{1}_{(0,\infty)}(P_{j,n}(\theta))\prod_{j\in B}\mathbf{1}_{(0,\infty)}(-P_{j,n}(\theta))\prod_{j\in C}\Big(\mathbf{1}_{\{0\}}(P_{j,n}(\theta))\mathbf{1}_{[0,\infty)}(-Z_j(\theta,x))\Big)\bigg]\\
& \sum_{\varnothing\neq\mathcal{I}\subseteq A}\bigg[\bigg[\prod_{i\in\mathcal{I}}\Big(\mathbf{1}_{(\alpha,\ell)}\Big(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}\Big)\mathbf{1}_{\{0\}}\Big(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}-\frac{Z_{\min\mathcal{I}}(\theta,x)}{P_{\min\mathcal{I},n}(\theta)}\Big)\Big)\\
& \times\prod_{j\in A\setminus\mathcal{I}}\mathbf{1}_{(0,\infty)}\Big(\frac{Z_{\min\mathcal{I}}(\theta,x)}{P_{\min\mathcal{I},n}(\theta)}-\frac{Z_j(\theta,x)}{P_{j,n}(\theta)}\Big)\prod_{i\in B}\Big(\mathbf{1}_{[\ell,\infty)}\Big(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}\Big)\Big)\bigg]\\
& \times\bigg[\ell^{i_n+1}-\Big(\frac{Z_{\min\mathcal{I}}(\theta,x)}{P_{\min\mathcal{I},n}(\theta)}\Big)^{i_n+1}\bigg]\bigg],
\end{aligned}
\tag{4.22}
$$

$$(III) = \sum_{A\dot\cup B\dot\cup C=\{1,\dots,r\}} \left[\prod_{j\in A}\mathbf{1}_{(0,\infty)}(P_{j,n}(\theta)) \prod_{j\in B}\mathbf{1}_{(0,\infty)}(-P_{j,n}(\theta)) \prod_{j\in C}\left(\mathbf{1}_{\{0\}}(P_{j,n}(\theta))\mathbf{1}_{[0,\infty)}(-Z_j(\theta,x))\right)\right]$$

$$\sum_{\varnothing\neq\mathcal{J}\subseteq B}\left[\left[\prod_{i\in A}\left(\mathbf{1}_{(-\infty,\mathscr{a}]}\left(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}\right)\right)\prod_{i\in\mathcal{J}}\left(\mathbf{1}_{(\mathscr{a},\mathscr{b})}\left(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}\right)\mathbf{1}_{\{0\}}\left(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)} - \frac{Z_{\min\mathcal{J}}(\theta,x)}{P_{\min\mathcal{J},n}(\theta)}\right)\right)\right.\right. \quad (4.23)$$

$$\times \prod_{j\in B\setminus\mathcal{J}}\mathbf{1}_{(0,\infty)}\left(\frac{Z_j(\theta,x)}{P_{j,n}(\theta)} - \frac{Z_{\min\mathcal{J}}(\theta,x)}{P_{\min\mathcal{J},n}(\theta)}\right)\right]\times\left[\left(\frac{Z_{\min\mathcal{J}}(\theta,x)}{P_{\min\mathcal{J},n}(\theta)}\right)^{i_n+1} - \mathscr{a}^{i_n+1}\right]\right],$$

and

$$(IV) = \sum_{A\dot\cup B\dot\cup C=\{1,\dots,r\}}\left[\prod_{j\in A}\mathbf{1}_{(0,\infty)}(P_{j,n}(\theta))\prod_{j\in B}\mathbf{1}_{(0,\infty)}(-P_{j,n}(\theta))\prod_{j\in C}\left(\mathbf{1}_{\{0\}}(P_{j,n}(\theta))\mathbf{1}_{[0,\infty)}(-Z_j(\theta,x))\right)\right]$$

$$(4.24)$$

$$\times\left(\prod_{i\in A}\mathbf{1}_{(-\infty,\mathscr{a}]}\left(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}\right)\prod_{i\in B}\mathbf{1}_{[\mathscr{b},\infty)}\left(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}\right)\right)\left[\mathscr{b}^{i_n+1} - \mathscr{a}^{i_n+1}\right].$$

Note that the first products over all elements of A, B, C precisely describe the conditions that $\mathcal{S}_1^\theta = A$, $\mathcal{S}_{-1}^\theta = B$, $\mathcal{S}_0^\theta = C$, and $\forall j\in\mathcal{S}_0^\theta:\; -Z_j(\theta,x)\geq 0$. Furthermore, observe that, e.g., in (I) we we must have for all $i\in\mathcal{I}$, $j\in A\setminus\mathcal{I}$ that $\frac{Z_j(\theta,x)}{P_{j,n}(\theta)} < \frac{Z_{\min\mathcal{I}}(\theta,x)}{P_{\min\mathcal{I},n}(\theta)} = \frac{Z_i(\theta,x)}{P_{i,n}(\theta)}\in(\mathscr{a},\mathscr{b})$ in order to obtain a non-zero value. In other words, the maximal value of $\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}$, $i\in A$, is achieved exactly for $i\in\mathcal{I}$, and similarly the minimal value of $\frac{Z_j(\theta,x)}{P_{j,n}(\theta)}$, $j\in B$, is achieved exactly for $j\in\mathcal{J}$ (and analogously in $(II),(III)$). Moreover, note that we have for all $i\in\mathcal{I}\subseteq A$ that

$$\mathbf{1}_{(\mathscr{a},\mathscr{b})}\left(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}\right) = \mathbf{1}_{(\mathscr{a},\infty)}\left(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}\right)\mathbf{1}_{(-\infty,\mathscr{b})}\left(\frac{Z_i(\theta,x)}{P_{i,n}(\theta)}\right)$$

$$= \mathbf{1}_{(0,\infty)}(Z_i(\theta,x) - \mathscr{a}P_{i,n}(\theta))\mathbf{1}_{(0,\infty)}(\mathscr{b}P_{i,n}(\theta) - Z_i(\theta,x)). \quad (4.25)$$

Here $Z_i(\theta,x)$ is polynomial in θ and linear in $x_1,\dots,x_{n-1}$, and thus of the form required by Definition 4.6. Similarly, the other indicator functions can be brought into the correct form, taking into account the different signs of $P_{j,n}(\theta)$ for $j\in A$ and $j\in B$. Moreover, observe that the remaining terms can be written as linear combinations of rational functions in θ and polynomials in x. Hence, we obtain that the functions defined by $(I),(II),(III),(IV)$ are elements of $\mathscr{A}_{m,n-1}$. The proof of Proposition 4.8 is thus complete.

4.3. On the semialgebraic property of the risk function

Definition 4.9. Let $d\in\mathbb{N}$, let $A\subseteq\mathbb{R}^d$ be a set, and let $f\colon A\to\mathbb{R}$ be a function. Then we say that f is piecewise polynomial if and only if there exist $n\in\mathbb{N}$, $\alpha_1,\alpha_2,\dots,\alpha_n\in\mathbb{R}^{n\times d}$, $\beta_1,\beta_2,\dots,\beta_n\in\mathbb{R}^n$, $P_1,P_2,\dots,P_n\in\mathscr{P}_d$ such that for all $x\in A$ it holds that

$$f(x) = \sum_{i=1}^n\left[P_i(x)\mathbf{1}_{[0,\infty)^n}(\alpha_i x + \beta_i)\right] \quad (4.26)$$

(cf. Definition 4.1).

Corollary 4.10. *Assume Setting 2.1 and assume that f and $\mathfrak{p}$ are piecewise polynomial (cf. Definition 4.9). Then $\mathcal{L}$ is semialgebraic (cf. Definition 4.3).*

Proof of Corollary 4.10. Throughout this proof let $F \colon \mathbb{R}^d \to \mathbb{R}$ and $\mathfrak{P} \colon \mathbb{R}^d \to \mathbb{R}$ satisfy for all $x \in \mathbb{R}^d$ that

$$F(x) = \begin{cases} f(x) & : x \in [a, b]^d \\ 0 & : x \notin [a, b]^d \end{cases} \quad \text{and} \quad \mathfrak{P}(x) = \begin{cases} \mathfrak{p}(x) & : x \in [a, b]^d \\ 0 & : x \notin [a, b]^d. \end{cases} \tag{4.27}$$

Note that (4.27) and the assumption that f and $\mathfrak{p}$ are piecewise polynomial assure that

$$\left[\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^d \ni (\theta, x) \mapsto F(x) \in \mathbb{R} \right] \in \mathscr{A}_{\mathfrak{d},d} \quad \text{and} \quad \left[\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^d \ni (\theta, x) \mapsto \mathfrak{P}(x) \in \mathbb{R} \right] \in \mathscr{A}_{\mathfrak{d},d} \tag{4.28}$$

(cf. Definition 4.6). In addition, observe that the fact that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $x \in \mathbb{R}^d$ we have that

$$\begin{aligned}
\mathcal{N}^{\theta}(x) &= \mathfrak{c}^{\theta} + \sum_{i=1}^{H} \mathfrak{v}_i^{\theta} \max\left\{ \textstyle\sum_{\ell=1}^{d} \mathfrak{w}_{i,\ell}^{\theta} x_{\ell} + \mathfrak{b}_i^{\theta}, 0 \right\} \\
&= \mathfrak{c}^{\theta} + \sum_{i=1}^{H} \mathfrak{v}_i^{\theta} \left(\textstyle\sum_{\ell=1}^{d} \mathfrak{w}_{i,\ell}^{\theta} x_{\ell} + \mathfrak{b}_i^{\theta} \right) \mathbf{1}_{[0,\infty)} \left(\textstyle\sum_{\ell=1}^{d} \mathfrak{w}_{i,\ell}^{\theta} x_{\ell} + \mathfrak{b}_i^{\theta} \right)
\end{aligned} \tag{4.29}$$

demonstrates that

$$\left[\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^d \ni (\theta, x) \mapsto \mathcal{N}^{\theta}(x) \in \mathbb{R} \right] \in \mathscr{A}_{\mathfrak{d},d}. \tag{4.30}$$

Combining this with (4.28) and the fact that $\mathscr{A}_{\mathfrak{d},d}$ is an algebra proves that

$$\left[\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^d \ni (\theta, x) \mapsto (\mathcal{N}^{\theta}(x) - F(x))^2 \mathfrak{P}(x) \in \mathbb{R} \right] \in \mathscr{A}_{\mathfrak{d},d}. \tag{4.31}$$

This, Proposition 4.8, and induction demonstrate that

$$\left[\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \int_a^b \int_a^b \cdots \int_a^b (\mathcal{N}^{\theta}(x) - F(x))^2 \mathfrak{P}(x) \, \mathrm{d}x_d \cdots \mathrm{d}x_2 \, \mathrm{d}x_1 \in \mathbb{R} \right] \in \mathscr{A}_{\mathfrak{d},0}. \tag{4.32}$$

Fubini's theorem hence implies that $\mathcal{L} \in \mathscr{A}_{\mathfrak{d},0}$. Combining this and Lemma 4.7 shows that $\mathcal{L}$ is semialgebraic. The proof of Corollary 4.10 is thus complete.

5. Convergence rates for solutions of GF differential equations

In this section we employ the findings from Sections 2 and 4 to establish in Proposition 5.2 in Subsection 5.2 below, in Proposition 5.3 in Subsection 5.2, and in Theorem 5.4 in Subsection 5.3 below several convergence rate results for solutions of GF differential equations. Theorem 1.2 in the introduction is a direct consequence of Theorem 5.4. Our proof of Theorem 5.4 is based on an application of Proposition 5.3 and our proof of Proposition 5.3 uses Proposition 5.2. Our proof of Proposition 5.2, in turn, employs Proposition 5.1 in Subsection 5.1 below. In Proposition 5.1 we establish that under the assumption that the target function $f \colon [a, b]^d \to \mathbb{R}$ and the unnormalized density function $\mathfrak{p} \colon [a, b]^d \to [0, \infty)$ are piecewise polynomial (see Definition 4.9 in Subsection 4.3) we have that the risk function $\mathcal{L} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}$ satisfies an appropriately generalized *Kurdyka-Łojasiewicz inequality*.

In the proof of Proposition 5.1 the classical *Łojasiewicz inequality* for semialgebraic or subanalytic functions (cf., e.g., Bierstone & Milman [40]) is not directly applicable since the generalized gradient

function $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ is not continuous. We will employ the more general results from Bolte et al. [9] which also apply to not necessarily continuously differentiable functions.

The arguments used in the proof of Proposition 5.2 are slight adaptions of well-known arguments in the literature; see, e.g., Kurdyka et al. [12, Section 1], Bolte et al. [9, Theorem 4.5], or Absil et al. [6, Theorem 2.2]. On the one hand, in Kurdyka et al. [12, Section 1] and Absil et al. [6, Theorem 2.2] it is assumed that the object function of the considered optimization problem is analytic and in Bolte et al. [9, Theorem 4.5] it is assumed that the objective function of the considered optimization problem is convex or lower C^2 and Proposition 5.2 does not require these assumptions. On the other hand, Bolte et al. [9, Theorem 4.5] consider more general differential dynamics and the considered gradients are allowed to be more general than the specific generalized gradient function $\mathcal{G} \colon \mathbb{R}^{\mathfrak{d}} \to \mathbb{R}^{\mathfrak{d}}$ which is considered in Proposition 5.2.

5.1. Generalized Kurdyka-Łojasiewicz inequality for the risk function

Proposition 5.1 (Generalized Kurdyka-Łojasiewicz inequality). *Assume Setting 2.1, assume that $\mathfrak{p}$ and f are piecewise polynomial, and let $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ (cf. Definition 4.9). Then there exist $\varepsilon, \mathfrak{D} \in (0, \infty)$, $\alpha \in (0, 1)$ such that for all $\theta \in B_\varepsilon(\vartheta)$ it holds that*

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \le \mathfrak{D}\|\mathcal{G}(\theta)\|. \tag{5.1}$$

Proof of Proposition 5.1. Throughout this proof let $\mathbf{M} \colon \mathbb{R}^{\mathfrak{d}} \to [0, \infty]$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathbf{M}(\theta) = \inf(\{\|h\| \colon h \in \partial\mathcal{L}(\theta)\} \cup \{\infty\}). \tag{5.2}$$

Note that Proposition 2.12 implies for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathbf{M}(\theta) \le \|\mathcal{G}(\theta)\|. \tag{5.3}$$

Furthermore, observe that Corollary 4.10, the fact that semialgebraic functions are subanalytic, and Bolte et al. [9, Theorem 3.1 and Remark 3.2] ensure that there exist $\varepsilon, \mathfrak{D} \in (0, \infty)$, $\mathfrak{a} \in [0, 1)$ which satisfy for all $\theta \in B_\varepsilon(\vartheta)$ that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{\mathfrak{a}} \le \mathfrak{D}\mathbf{M}(\theta). \tag{5.4}$$

Combining this and (5.3) with the fact that $\sup_{\theta \in B_\varepsilon(\vartheta)}|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)| < \infty$ demonstrates that for all $\theta \in B_\varepsilon(\vartheta)$, $\alpha \in (\mathfrak{a}, 1)$ we have that

$$\begin{aligned}
|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha &\le |\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{\mathfrak{a}}(\sup_{\psi \in B_\varepsilon(\vartheta)}|\mathcal{L}(\psi) - \mathcal{L}(\vartheta)|^{\alpha - \mathfrak{a}}) \\
&\le (\mathfrak{D} \sup_{\psi \in B_\varepsilon(\vartheta)}|\mathcal{L}(\psi) - \mathcal{L}(\vartheta)|^{\alpha - \mathfrak{a}})\|\mathcal{G}(\theta)\|.
\end{aligned} \tag{5.5}$$

This completes the proof of Proposition 5.1. $\qquad\square$

5.2. Local convergence for solutions of GF differential equations

Proposition 5.2. *Assume Setting 2.1 and let $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\varepsilon, \mathfrak{D} \in (0, \infty)$, $\alpha \in (0, 1)$ satisfy for all $\theta \in B_\varepsilon(\vartheta)$ that*

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \le \mathfrak{D}\|\mathcal{G}(\theta)\|. \tag{5.6}$$

Then there exists $\delta \in (0, \varepsilon)$ such that for all $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ with $\Theta_0 \in B_\delta(\vartheta)$, $\forall t \in [0, \infty)$: $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) \, ds$, and $\inf_{t \in \{s \in [0, \infty) : \Theta_s \in B_\varepsilon(\vartheta)\}} \mathcal{L}(\Theta_t) \geq \mathcal{L}(\vartheta)$ there exists $\psi \in \mathcal{L}^{-1}(\{\mathcal{L}(\vartheta)\})$ such that for all $t \in [0, \infty)$ it holds that

$$\Theta_t \in B_\varepsilon(\vartheta), \qquad \int_0^\infty \|\mathcal{G}(\Theta_s)\| \, ds \leq \varepsilon, \qquad |\mathcal{L}(\Theta_t) - \mathcal{L}(\psi)| \leq (1 + \mathfrak{D}^{-2} t)^{-1}, \tag{5.7}$$

$$and \qquad \|\Theta_t - \psi\| \leq \left[1 + (\mathfrak{D}^{-1/\alpha}(1 - \alpha))^{\frac{\alpha}{1-\alpha}} t\right]^{-\min\left\{1, \frac{1-\alpha}{\alpha}\right\}}. \tag{5.8}$$

Proof of Proposition 5.2. Note that the fact that $\mathcal{L}$ is continuous implies that there exists $\delta \in (0, \varepsilon/3)$ which satisfies for all $\theta \in B_\delta(\vartheta)$ that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{1-\alpha} \leq \min\left\{\frac{\varepsilon(1 - \alpha)}{3\mathfrak{D}}, \frac{1 - \alpha}{\mathfrak{D}}, 1\right\}. \tag{5.9}$$

In the following let $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ satisfy $\forall t \in [0, \infty)$: $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) \, ds$, $\Theta_0 \in B_\delta(\vartheta)$, and

$$\inf_{t \in \{s \in [0, \infty) : \Theta_s \in B_\varepsilon(\vartheta)\}} \mathcal{L}(\Theta_t) \geq \mathcal{L}(\vartheta). \tag{5.10}$$

In the first step we show that for all $t \in [0, \infty)$ it holds that

$$\Theta_t \in B_\varepsilon(\vartheta). \tag{5.11}$$

Observe that, e.g., [37, Lemma 3.1] ensures for all $t \in [0, \infty)$ that

$$\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|\mathcal{G}(\Theta_s)\|^2 \, ds. \tag{5.12}$$

This implies that $[0, \infty) \ni t \mapsto \mathcal{L}(\Theta_t) \in [0, \infty)$ is non-increasing. Next let $L \colon [0, \infty) \to \mathbb{R}$ satisfy for all $t \in [0, \infty)$ that

$$L(t) = \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \tag{5.13}$$

and let $T \in [0, \infty]$ satisfy

$$T = \inf(\{t \in [0, \infty) : \|\Theta_t - \vartheta\| \geq \varepsilon\} \cup \{\infty\}). \tag{5.14}$$

We intend to show that $T = \infty$. Note that (5.10) assures for all $t \in [0, T)$ that $L(t) \geq 0$. Moreover, observe that (5.12) and (5.13) ensure that for almost all $t \in [0, T)$ it holds that L is differentiable at t and satisfies $L'(t) = \frac{d}{dt}(\mathcal{L}(\Theta_t)) = -\|\mathcal{G}(\Theta_t)\|^2$. In the following let $\tau \in [0, T]$ satisfy

$$\tau = \inf(\{t \in [0, T) : L(t) = 0\} \cup \{T\}). \tag{5.15}$$

Note that the fact that L is non-increasing implies that for all $s \in [\tau, T)$ it holds that $L(s) = 0$. Combining this with (5.12) demonstrates for almost all $s \in (\tau, T)$ that $\mathcal{G}(\Theta_s) = 0$. This proves for all $s \in [\tau, T)$ that $\Theta_s = \Theta_\tau$. Next observe that (5.6) ensures that for all $t \in [0, \tau)$ it holds that

$$0 < [L(t)]^\alpha = |\mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta)|^\alpha \leq \mathfrak{D}\|\mathcal{G}(\Theta_t)\|. \tag{5.16}$$

Combining this with the chain rule proves for almost all $t \in [0, \tau)$ that

$$\frac{\mathrm{d}}{\mathrm{d}t}([L(t)]^{1-\alpha}) = (1-\alpha)[L(t)]^{-\alpha}\left(-\|\mathcal{G}(\Theta_t)\|^2\right)$$
$$\leq -(1-\alpha)\mathfrak{D}^{-1}\|\mathcal{G}(\Theta_t)\|^{-1}\|\mathcal{G}(\Theta_t)\|^2 = -\mathfrak{D}^{-1}(1-\alpha)\|\mathcal{G}(\Theta_t)\|. \tag{5.17}$$

In addition, note that the fact that $[0, \infty) \ni t \mapsto L(t) \in \mathbb{R}$ is absolutely continuous and the fact that for all $r \in (0, \infty)$ it holds that $(r, \infty) \ni y \mapsto y^{1-\alpha} \in \mathbb{R}$ is Lipschitz continuous demonstrate for all $t \in [0, \tau)$ that $[0, t] \ni s \mapsto [L(s)]^{1-\alpha} \in \mathbb{R}$ is absolutely continuous. Integrating (5.17) hence shows for all $s, t \in [0, \tau)$ with $t \leq s$ that

$$\int_t^s \|\mathcal{G}(\Theta_u)\| \, \mathrm{d}u \leq -\mathfrak{D}(1-\alpha)^{-1}([L(s)]^{1-\alpha} - [L(t)]^{1-\alpha}) \leq \mathfrak{D}(1-\alpha)^{-1}[L(t)]^{1-\alpha}. \tag{5.18}$$

This and the fact that for almost all $s \in (\tau, T)$ it holds that $\mathcal{G}(\Theta_s) = 0$ ensure that for all $s, t \in [0, T)$ with $t \leq s$ we have that

$$\int_t^s \|\mathcal{G}(\Theta_u)\| \, \mathrm{d}u \leq \mathfrak{D}(1-\alpha)^{-1}[L(t)]^{1-\alpha}. \tag{5.19}$$

Combining this with (5.9) demonstrates for all $t \in [0, T)$ that

$$\|\Theta_t - \Theta_0\| = \left\|\int_0^t \mathcal{G}(\Theta_s) \, \mathrm{d}s\right\| \leq \int_0^t \|\mathcal{G}(\Theta_s)\| \, \mathrm{d}s \leq \frac{\mathfrak{D}|\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|^{1-\alpha}}{1-\alpha} \leq \min\left\{\frac{\varepsilon}{3}, 1\right\}. \tag{5.20}$$

This, the fact that $\delta < \varepsilon/3$, and the triangle inequality assure for all $t \in [0, T)$ that

$$\|\Theta_t - \vartheta\| \leq \|\Theta_t - \Theta_0\| + \|\Theta_0 - \vartheta\| \leq \frac{\varepsilon}{3} + \delta \leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \frac{2\varepsilon}{3}. \tag{5.21}$$

Combining this with (5.14) proves that $T = \infty$. This establishes (5.11).

Next observe that the fact that $T = \infty$ and (5.20) prove that

$$\int_0^\infty \|\mathcal{G}(\Theta_s)\| \, \mathrm{d}s \leq \min\left\{\frac{\varepsilon}{3}, 1\right\} \leq \varepsilon < \infty. \tag{5.22}$$

In the following let $\sigma \colon [0, \infty) \to [0, \infty)$ satisfy for all $t \in [0, \infty)$ that

$$\sigma(t) = \int_t^\infty \|\mathcal{G}(\Theta_s)\| \, \mathrm{d}s. \tag{5.23}$$

Note that (5.22) proves that $\limsup_{t\to\infty} \sigma(t) = 0$. In addition, observe that (5.22) assures that there exists $\psi \in \mathbb{R}^{\mathfrak{d}}$ such that

$$\limsup_{t\to\infty}\|\Theta_t - \psi\| = 0. \tag{5.24}$$

In the next step we combine the weak chain rule for the risk function in (5.12) with (5.11) and (5.6) to obtain that for almost all $t \in [0, \infty)$ we have that

$$L'(t) = -\|\mathcal{G}(\Theta_t)\|^2 \leq -\mathfrak{D}^{-2}[L(t)]^{2\alpha}. \tag{5.25}$$

In addition, note that the fact that L is non-increasing and (5.9) ensure that for all $t \in [0, \infty)$ it holds that $L(t) \leq L(0) \leq 1$. Therefore, we get for almost all $t \in [0, \infty)$ that

$$L'(t) \leq -\mathfrak{D}^{-2}[L(t)]^2. \tag{5.26}$$

Combining this with the fact that for all $t \in [0, \tau)$ it holds that $L(t) > 0$ establishes for almost all $t \in [0, \tau)$ that

$$\frac{d}{dt}\left(\frac{\mathfrak{D}^2}{L(t)}\right) = -\frac{\mathfrak{D}^2 L'(t)}{[L(t)]^2} \geq 1. \tag{5.27}$$

The fact that for all $t \in [0, \tau)$ it holds that $[0, t] \ni s \mapsto L(s) \in (0, \infty)$ is absolutely continuous hence demonstrates for all $t \in [0, \tau)$ that

$$\frac{\mathfrak{D}^2}{L(t)} \geq \frac{\mathfrak{D}^2}{L(0)} + t \geq \mathfrak{D}^2 + t. \tag{5.28}$$

Therefore, we infer for all $t \in [0, \tau)$ that

$$L(t) \leq \mathfrak{D}^2\left(\mathfrak{D}^2 + t\right)^{-1} = \left(1 + \mathfrak{D}^{-2}t\right)^{-1}. \tag{5.29}$$

This and the fact that for all $t \in [\tau, \infty)$ it holds that $L(t) = 0$ prove that for all $t \in [0, \infty)$ we have that

$$|\mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta)| = L(t) \leq \left(1 + \mathfrak{D}^{-2}t\right)^{-1}. \tag{5.30}$$

Furthermore, observe that (5.24) and the fact that $\mathcal{L}$ is continuous imply that $\limsup_{t \to \infty} |\mathcal{L}(\Theta_t) - \mathcal{L}(\psi)| = 0$. Hence, we obtain that $\mathcal{L}(\psi) = \mathcal{L}(\vartheta)$. This shows for all $t \in [0, \infty)$ that

$$|\mathcal{L}(\Theta_t) - \mathcal{L}(\psi)| \leq \left(1 + \mathfrak{D}^{-2}t\right)^{-1}. \tag{5.31}$$

In the next step we establish a convergence rate for the quantity $\|\Theta_t - \psi\|$, $t \in [0, \infty)$. We accomplish this by employing an upper bound for the tail length of the curve $\Theta_t \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0, \infty)$. More formally, note that (5.19), (5.11), and (5.6) demonstrate for all $t \in [0, \infty)$ that

$$\sigma(t) = \int_t^\infty \|\mathcal{G}(\Theta_u)\| \, du = \lim_{s \to \infty}\left[\int_t^s \|\mathcal{G}(\Theta_u)\| \, du\right] \tag{5.32}$$
$$\leq \mathfrak{D}(1 - \alpha)^{-1}[L(t)]^{1-\alpha} \leq \mathfrak{D}(1 - \alpha)^{-1}(\mathfrak{D}\|\mathcal{G}(\Theta_t)\|)^{\frac{1-\alpha}{\alpha}}.$$

Next observe that the fact that for all $t \in [0, \infty)$ it holds that $\sigma(t) = \int_0^\infty \|\mathcal{G}(\Theta_s)\| \, ds - \int_0^t \|\mathcal{G}(\Theta_s)\| \, ds$ shows that for almost all $t \in [0, \infty)$ we have that $\sigma'(t) = -\|\mathcal{G}(\Theta_t)\|$. This and (5.32) yield for almost all $t \in [0, \infty)$ that $\sigma(t) \leq \mathfrak{D}^{1/\alpha}(1 - \alpha)^{-1}[-\sigma'(t)]^{\frac{1-\alpha}{\alpha}}$. Therefore, we obtain for almost all $t \in [0, \infty)$ that

$$\sigma'(t) \leq -[(1 - \alpha)\mathfrak{D}^{-1/\alpha}\sigma(t)]^{\frac{\alpha}{1-\alpha}}. \tag{5.33}$$

Combining this with the fact that σ is absolutely continuous implies for all $t \in [0, \infty)$ that

$$\sigma(t) - \sigma(0) \leq -[(1 - \alpha)\mathfrak{D}^{-1/\alpha}]^{\frac{\alpha}{1-\alpha}} \int_0^t [\sigma(s)]^{\frac{\alpha}{1-\alpha}} \, ds. \tag{5.34}$$

In the following let $\beta, \mathfrak{C} \in (0, \infty)$ satisfy $\beta = \max\{1, \frac{\alpha}{1-\alpha}\}$ and $\mathfrak{C} = \left((1 - \alpha)\mathfrak{D}^{-1/\alpha}\right)^{\frac{\alpha}{1-\alpha}}$. Note that (5.34) and the fact that for all $t \in [0, \infty)$ it holds that $\sigma(t) \leq \sigma(0) \leq 1$ ensure that for all $t \in [0, \infty)$ it holds that

$$\sigma(t) \leq \sigma(0) - \mathfrak{C} \int_0^t [\sigma(s)]^\beta \, ds. \tag{5.35}$$

This, the fact that σ is non-increasing, and the fact that for all $t \in [0, \infty)$ it holds that $0 \leq \sigma(t) \leq 1$ prove that for all $t \in [0, \infty)$ we have that

$$[\sigma(t)]^\beta \leq \sigma(t) \leq \sigma(0) - \mathfrak{C}[\sigma(t)]^\beta t \leq 1 - \mathfrak{C}t[\sigma(t)]^\beta. \tag{5.36}$$

Hence, we obtain for all $t \in [0, \infty)$ that $\sigma(t) \leq (1 + \mathfrak{C}t)^{-\frac{1}{\beta}}$. Combining this with the fact that for all $t \in [0, \infty)$ it holds that

$$\begin{aligned}
\|\Theta_t - \psi\| &\leq \limsup_{s\to\infty}\|\Theta_t - \Theta_s\| = \limsup_{s\to\infty}\left\|\int_t^s \mathcal{G}(\Theta_u)\,\mathrm{d}u\right\| \leq \limsup_{s\to\infty}\left[\int_t^s \|\mathcal{G}(\Theta_u)\|\,\mathrm{d}u\right] \\
&= \int_t^\infty \|\mathcal{G}(\Theta_u)\|\,\mathrm{d}u = \sigma(t)
\end{aligned} \tag{5.37}$$

shows that for all $t \in [0, \infty)$ we have that $\|\Theta_t - \psi\| \leq (1 + \mathfrak{C}t)^{-1/\beta}$. This, (5.11), (5.22), and (5.31) establish (5.8). The proof of Proposition 5.2 is thus complete.

5.3. Global convergence for solutions of GF differential equations

Proposition 5.3. *Assume Setting 2.1, assume that $\mathfrak{p}$ and f are piecewise polynomial, and let $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ satisfy*

$$\liminf_{t\to\infty}\|\Theta_t\| < \infty \qquad and \qquad \forall t \in [0, \infty): \Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s)\,\mathrm{d}s \tag{5.38}$$

(cf. Definition 4.9). Then there exist $\vartheta \in \mathcal{G}^{-1}(\{0\})$, $\mathfrak{C}, \tau, \beta \in (0, \infty)$ which satisfy for all $t \in [\tau, \infty)$ that

$$\|\Theta_t - \vartheta\| \leq (1 + \mathfrak{C}(t - \tau))^{-\beta} \qquad and \qquad |\mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta)| \leq (1 + \mathfrak{C}(t - \tau))^{-1}. \tag{5.39}$$

Proof of Proposition 5.3. First observe that [37, Lemma 3.1] ensures that for all $t \in [0, \infty)$ it holds that

$$\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|\mathcal{G}(\Theta_s)\|^2\,\mathrm{d}s. \tag{5.40}$$

This implies that $[0, \infty) \ni t \mapsto \mathcal{L}(\Theta_t) \in [0, \infty)$ is non-increasing. Hence, we obtain that there exists $\mathbf{m} \in [0, \infty)$ which satisfies that

$$\mathbf{m} = \limsup_{t\to\infty} \mathcal{L}(\Theta_t) = \liminf_{t\to\infty} \mathcal{L}(\Theta_t) = \inf_{t\in[0,\infty)} \mathcal{L}(\Theta_t). \tag{5.41}$$

Moreover, note that the assumption that $\liminf_{t\to\infty}\|\Theta_t\| < \infty$ ensures that there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ and $\tau = (\tau_n)_{n\in\mathbb{N}}: \mathbb{N} \to [0, \infty)$ which satisfy $\liminf_{n\to\infty} \tau_n = \infty$ and

$$\limsup_{n\to\infty}\|\Theta_{\tau_n} - \vartheta\| = 0. \tag{5.42}$$

Combining this with (5.41) and the fact that $\mathcal{L}$ is continuous shows that

$$\mathcal{L}(\vartheta) = \mathbf{m} \qquad and \qquad \forall t \in [0, \infty): \mathcal{L}(\Theta_t) \geq \mathcal{L}(\vartheta). \tag{5.43}$$

Next observe that Proposition 5.1 demonstrates that there exist $\varepsilon, \mathfrak{D} \in (0, \infty)$, $\alpha \in (0, 1)$ such that for all $\theta \in B_\varepsilon(\vartheta)$ we have that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \leq \mathfrak{D}\|\mathcal{G}(\theta)\|. \tag{5.44}$$

Combining this and (5.42) with Proposition 5.2 demonstrates that there exists $\delta \in (0, \varepsilon)$ which satisfies for all $\Phi \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ with $\Phi_0 \in B_\delta(\vartheta)$, $\forall t \in [0, \infty)$: $\Phi_t = \Phi_0 - \int_0^t \mathcal{G}(\Phi_s) \, ds$, and $\inf_{t \in \{s \in [0, \infty) \colon \Phi_s \in B_\varepsilon(\vartheta)\}} \mathcal{L}(\Phi_t) \geq \mathcal{L}(\vartheta)$ that it holds for all $t \in [0, \infty)$ that

$$\Phi_t \in B_\varepsilon(\vartheta), \qquad |\mathcal{L}(\Phi_t) - \mathcal{L}(\vartheta)| \leq (1 + \mathfrak{D}^{-2} t)^{-1}, \tag{5.45}$$

$$\text{and} \qquad \|\Phi_t - \vartheta\| \leq \left[1 + (\mathfrak{D}^{-1/\alpha}(1 - \alpha))^{\frac{\alpha}{1-\alpha}} t \right]^{-\min\{1, \frac{1-\alpha}{\alpha}\}}. \tag{5.46}$$

Moreover, note that (5.42) ensures that there exists $n \in \mathbb{N}$ which satisfies $\Theta_{\tau_n} \in B_\delta(\vartheta)$. Next let $\Phi \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ satisfy for all $t \in [0, \infty)$ that

$$\Phi_t = \Theta_{t+\tau_n}. \tag{5.47}$$

Observe that (5.43) and (5.47) assure that

$$\Phi_0 \in B_\delta(\vartheta), \qquad \inf_{t \in [0, \infty)} \mathcal{L}(\Phi_t) \geq \mathcal{L}(\vartheta), \qquad \text{and} \qquad \forall t \in [0, \infty) : \Phi_t = \Phi_0 - \int_0^t \mathcal{G}(\Phi_s) \, ds. \tag{5.48}$$

Combining this with (5.46) proves for all $t \in [\tau_n, \infty)$ that

$$|\mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta)| \leq \left(1 + \mathfrak{D}^{-2}(t - \tau_n) \right)^{-1} \tag{5.49}$$

and

$$\|\Theta_t - \vartheta\| \leq \left[1 + (\mathfrak{D}^{-1/\alpha}(1 - \alpha))^{\frac{\alpha}{1-\alpha}} (t - \tau_n) \right]^{-\min\{1, \frac{1-\alpha}{\alpha}\}}. \tag{5.50}$$

Next note that [37, Corollary 2.15] shows that $\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \|\mathcal{G}(\theta)\| \in [0, \infty)$ is lower semicontinuous. The fact that $\liminf_{s \to \infty} \|\mathcal{G}(\Theta_s)\| = 0$ and the fact that $\limsup_{t \to \infty} \|\Theta_t - \vartheta\| = 0$ hence imply that $\mathcal{G}(\vartheta) = 0$. Combining this with (5.49) and (5.50) establishes (5.39). The proof of Proposition 5.3 is thus complete.

By choosing a sufficiently large $\mathscr{C} \in (0, \infty)$ we can conclude a simplified version of Proposition 5.3. This is precisely the subject of the next result, Theorem 5.4 below. Theorem 1.2 in the introduction is a direct consequence of Theorem 5.4.

Theorem 5.4. *Assume Setting 2.1, assume that $\mathfrak{p}$ and f are piecewise polynomial, and let $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ satisfy $\liminf_{t \to \infty} \|\Theta_t\| < \infty$ and $\forall t \in [0, \infty) : \Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) \, ds$ (cf. Definition 4.9). Then there exist $\vartheta \in \mathcal{G}^{-1}(\{0\})$, $\mathscr{C}, \beta \in (0, \infty)$ which satisfy for all $t \in [0, \infty)$ that*

$$\|\Theta_t - \vartheta\| \leq \mathscr{C}(1 + t)^{-\beta} \qquad \text{and} \qquad |\mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta)| \leq \mathscr{C}(1 + t)^{-1}. \tag{5.51}$$

Proof of Theorem 5.4. Observe that Proposition 5.3 assures that there exist $\vartheta \in \mathcal{G}^{-1}(\{0\})$, $\mathfrak{C}, \tau, \beta \in (0, \infty)$ which satisfy for all $t \in [\tau, \infty)$ that

$$\|\Theta_t - \vartheta\| \leq (1 + \mathfrak{C}(t - \tau))^{-\beta} \tag{5.52}$$

and

$$|\mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta)| \leq (1 + \mathfrak{C}(t - \tau))^{-1}. \tag{5.53}$$

In the following let $\mathscr{C} \in (0, \infty)$ satisfy

$$\mathscr{C} = \max\left\{\mathfrak{C}^{-1}, 1 + \tau, \mathfrak{C}^{-\beta}, (1 + \tau)^{\beta}, (1 + \tau)^{\beta}[\sup_{s \in [0,\tau]} \|\Theta_s - \vartheta\|], (1 + \tau)\mathcal{L}(\Theta_0)\right\}. \tag{5.54}$$

Note that (5.53), (5.54), and the fact that $[0, \infty) \ni t \mapsto \mathcal{L}(\Theta_t) \in [0, \infty)$ is non-increasing show for all $t \in [0, \tau]$ that

$$\|\Theta_t - \vartheta\| \le \sup_{s \in [0,\tau]} \|\Theta_s - \vartheta\| \le \mathscr{C}(1 + \tau)^{-\beta} \le \mathscr{C}(1 + t)^{-\beta} \tag{5.55}$$

and

$$|\mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta)| = \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \le \mathcal{L}(\Theta_t) \le \mathcal{L}(\Theta_0) \le \mathscr{C}(1 + \tau)^{-1} \le \mathscr{C}(1 + t)^{-1}. \tag{5.56}$$

Moreover, observe that (5.52) and (5.54) imply for all $t \in [\tau, \infty)$ that

$$\|\Theta_t - \vartheta\| \le \mathscr{C}(\mathscr{C}^{1/\beta} + \mathfrak{C}\mathscr{C}^{1/\beta}(t - \tau))^{-\beta} \le \mathscr{C}(\mathscr{C}^{1/\beta} - \tau + t)^{-\beta} \le \mathscr{C}(1 + t)^{-\beta}. \tag{5.57}$$

In addition, note that (5.53) and (5.54) demonstrate for all $t \in [\tau, \infty)$ that

$$|\mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta)| \le \mathscr{C}(\mathscr{C} + \mathfrak{C}\mathscr{C}(t - \tau))^{-1} \le \mathscr{C}(\mathscr{C} - \tau + t)^{-1} \le \mathscr{C}(1 + t)^{-1}. \tag{5.58}$$

This completes the proof of Theorem 5.4. $\qquad \square$

Acknowledgments

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. F. Bach, E. Moulines, Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$, in *Advances in Neural Information Processing Systems*, **26** (2013), 773–781. Available from: `https://proceedings.neurips.cc/paper/2013/file/7fe1f8abaad094e0b 5cb1b01d712f708-Paper.pdf`.

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

3. E. Moulines, F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning, in *Advances in Neural Information Processing Systems*, **24** (2011), 451–459. Available from: https://proceedings.neurips.cc/paper/2011/file/40008b9a5380fcacce3976bf7c08af5b-Paper.pdf.

4. Y. Nesterov, *Introductory Lectures on Convex Optimization*, 2004. https://doi.org/10.1007/978-1-4419-8853-9

5. A. Rakhlin, O. Shamir, K. Sridharan, Making gradient descent optimal for strongly convex stochastic optimization, in *Proceedings of the 29th International Conference on Machine Learning*, Madison, WI, USA, (2012), 1571–1578.

6. P. A. Absil, R. Mahony, B. Andrews, Convergence of the iterates of descent methods for analytic cost functions, *SIAM J. Optim.*, **16** (2005), 531–547. https://doi.org/10.1137/040605266

7. H. Attouch, J. Bolte, On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, *Math. Program.*, **116** (2009), 5–16. https://doi.org/10.1007/s10107-007-0133-5

8. H. Attouch, J. Bolte, B. F. Svaiter, Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods, *Math. Program.*, **137** (2013), 91–129. https://doi.org/10.1007/s10107-011-0484-9

9. J. Bolte, A. Daniilidis, A. Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems, *SIAM J. Optim.*, **17** (2007), 1205–1223. https://doi.org/10.1137/050644641

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

11. H. Karimi, J. Nutini, M. Schmidt, Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition, in *Machine Learning and Knowledge Discovery in Databases*, (2016), 795–811. https://doi.org/10.1007/978-3-319-46128-1_50

12. K. Kurdyka, T. Mostowski, A. Parusiński, Proof of the gradient conjecture of R. Thom, *Ann. Math.*, **152** (2000), 763–792. https://doi.org/10.2307/2661354

13. J. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. Jordan, B. Recht, First-order methods almost always avoid strict saddle points, *Math. Program.*, **176** (2019), 311–337. https://doi.org/10.1007/s10107-019-01374-3

14. J. D. Lee, M. Simchowitz, M. I. Jordan, B. Recht, Gradient descent only converges to minimizers, in *29th Annual Conference on Learning Theory*, **49** (2016), 1246–1257. Available from: http://proceedings.mlr.press/v49/lee16.html.

15. S. Łojasiewicz, Sur les trajectoires du gradient d'une fonction analytique, in *Geometry Seminars*, (1983), 115–117.

16. P. Ochs, Unifying abstract inexact convergence theorems and block coordinate variable metric iPiano, *SIAM J. Optim.*, **29** (2019), 541–570. https://doi.org/10.1137/17M1124085

17. D. P. Bertsekas, J. N. Tsitsiklis, Gradient convergence in gradient methods with errors, *SIAM J. Optim.*, **10** (2000), 627–642. https://doi.org/10.1137/S1052623497331063

18. B. Fehrman, B. Gess, A. Jentzen, Convergence rates for the stochastic gradient descent method for non-convex objective functions, *J. Mach. Learn. Res.*, **21** (2022), 5354–5401. Available from: https://dl.acm.org/doi/abs/10.5555/3455716.3455852.

19. Y. Lei, T. Hu, G. Li, K. Tang, Stochastic gradient descent for nonconvex learning without bounded gradient assumptions, *IEEE Trans. Neural Networks Learn. Syst.*, **31** (2019), 4394–4400. https://doi.org/10.1109/TNNLS.2019.2952219

20. V. Patel. Stopping criteria for, and strong convergence of, stochastic gradient descent on Bottou-Curtis-Nocedal functions, *Math. Program.*, **195** (2022), 693–734. https://doi.org/10.1007/s10107-021-01710-6

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

22. S. Arora, S. Du, W. Hu, Z. Li, R. Wang, Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks, in *Proceedings of the 36th International Conference on Machine Learning*, **97** (2019), 322–332. Available from: http://proceedings.mlr.press/v97/arora19a.html.

23. L. Chizat, E. Oyallon, F. Bach, On lazy training in differentiable programming, in *Advances in Neural Information Processing Systems*, **32** (2019). Available from: https://proceedings.neurips.cc/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.

24. S. S. Du, X. Zhai, B. Póczos, A. Singh, Gradient descent provably optimizes over-parameterized neural networks, in *International Conference on Learning Representations*, 2019. Available from: https://openreview.net/forum?id=S1eK3i09YQ.

25. W. E, C. Ma, L. Wu, A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics, *Sci. China Math.*, **63** (2020), 1235–1258. https://doi.org/10.1007/s11425-019-1628-5

26. A. Jacot, F. Gabriel, C. Hongler, Neural tangent kernel: convergence and generalization in neural networks, in *Advances in Neural Information Processing Systems*, **31** (2018). Available from: https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

27. A. Jentzen, T. Kröger, Convergence rates for gradient descent in the training of overparameterized artificial neural networks with biases, preprint, arXiv:2102.11840.

28. G. Zhang, J. Martens, R. Grosse, Fast convergence of natural gradient descent for over-parameterized neural networks, in *Advances in Neural Information Processing Systems*, **32** (2019), 8082–8093. Available from: https://proceedings.neurips.cc/paper/2019/file/1da546f25222c1ee710cf7e2f7a3ff0c-Paper.pdf.

29. Z. Chen, G. Rotskoff, J. Bruna, E. Vanden-Eijnden, A dynamical central limit theorem for shallow neural networks, in *Advances in Neural Information Processing Systems*, **33** (2020), 22217–22230. Available from: https://proceedings.neurips.cc/paper/2020/file/fc5b3186f1cf0daece964f78259b7ba0-Paper.pdf.

30. L. Chizat, Sparse optimization on measures with over-parameterized gradient descent, *Math. Program.*, **194** (2022), 487–532. https://doi.org/10.1007/s10107-021-01636-z

31. L. Chizat, F. Bach, On the global convergence of gradient descent for over-parameterized models using optimal transport, in *Advances in Neural Information Processing Systems*, **31** (2018), 3036–3046. Available from: https://proceedings.neurips.cc/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.

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

33. P. Cheridito, A. Jentzen, A. Riekert, F. Rossmannek, A proof of convergence for gradient descent in the training of artificial neural networks for constant target functions, *J. Complexity*, **72** (2022), 101646. https://doi.org/10.1016/j.jco.2022.101646

34. A. Jentzen, A. Riekert, A proof of convergence for stochastic gradient descent in the training of artificial neural networks with ReLU activation for constant target functions, *Z. Angew. Math. Phys.*, **73** (2022), 188. https://doi.org/10.1007/s00033-022-01716-w

35. A. Jentzen, A. Riekert, A proof of convergence for the gradient descent optimization method with random initializations in the training of neural networks with ReLU activation for piecewise linear target functions, *J. Mach. Learn. Res.*, **23** (2022), 1–50. Available from: https://www.jmlr.org/papers/volume23/21-0962/21-0962.pdf.

36. P. Cheridito, A. Jentzen, F. Rossmannek, Landscape analysis for shallow neural networks: complete classification of critical points for affine target functions, *J. Nonlinear Sci.*, **32** (2022), 64. https://doi.org/10.1007/s00332-022-09823-8

37. A. Jentzen, A. Riekert, Convergence analysis for gradient flows in the training of artificial neural networks with ReLU activation, *J. Math. Anal. Appl.*, **517** (2023), 126601. https://doi.org/10.1016/j.jmaa.2022.126601

38. D. Gallon, A. Jentzen, F. Lindner, Blow up phenomena for gradient descent optimization methods in the training of artificial neural networks, preprint, arXiv:2211.15641.

39. R. T. Rockafellar, R. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 1998. https://doi.org/10.1007/978-3-642-02431-3

40. E. Bierstone, P. D. Milman, Semianalytic and subanalytic sets, *Inst. Hautes Études Sci. Publ. Math.*, **67** (1998), 5–42. https://doi.org/10.1007/BF02699126

41. T. Kaiser, Integration of semialgebraic functions and integrated Nash functions, *Math. Z.*, **275** (2013), 349–366. https://doi.org/10.1007/s00209-012-1138-1

42. M. Coste, An introduction to semialgebraic geometry, 2000. Available from: http://blogs.mat.ucm.es/jesusr/wp-content/uploads/sites/52/2020/03/SAG.pdf.