

Bounded subgradient trajectories in semialgebraic optimization

Xiaopeng Li

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2025

© 2025

Xiaopeng Li

All Rights Reserved

Abstract

Bounded subgradient trajectories in semialgebraic optimization

Xiaopeng Li

Solving modern data science problems relies on optimization algorithms that often succeed in high-dimensional, large-scale settings. First-order methods, in particular, are widely used due to their low per-iteration cost and scalability. However, despite their empirical success, theoretical understanding of their behavior in such settings remains limited. Classical optimization theory is typically built on assumptions about the objective function, such as convexity, smoothness, and coercivity, which are rarely satisfied in practice. In addition, common assumptions on algorithms, such as the boundedness of iterates or the existence of limit points, are often difficult to verify. To address these challenges, we develop a framework that replaces these assumptions with easily checkable conditions, using tools from dynamical systems, semialgebraic geometry, and variational analysis.

A key result of this thesis is that a broad class of optimization problems, including phase retrieval, matrix sensing, and neural networks, have bounded gradient flows. This property is central to multiple aspects of optimization. For landscape analysis, we develop practical tools for certifying the absence of bad local minima at infinity. For first-order algorithms, we analyze momentum methods and the proximal random reshuffling algorithm, proving global convergence of iterates and establishing improved convergence rates.

Table of Contents

Acknowledgments	v
Dedication	vi
Introduction	1
Chapter 1: Background	4
1.1 Clarke subdifferential	4
1.2 Semialgebraic functions	5
1.3 Subgradient trajectories	6
Chapter 2: Typical models with bounded subgradient trajectories	12
2.1 Convex model	12
2.2 Nonconvex coercive model	13
2.3 Nonconvex noncoercive model	14
2.3.1 Phase Retrieval	14
2.3.2 Asymmetric matrix sensing	15
2.3.3 Nonsmooth matrix factorization	17
2.3.4 Nonnegative matrix factorization	20
2.3.5 Linear neural network	25

2.3.6	One dimensional deep sigmoid neural network	28
Chapter 3:	Landscape at infinity	32
3.1	Setwise local minimum	37
3.2	Proof of Theorem 3.1	42
3.3	Applications	47
3.3.1	Deep linear neural network	47
3.3.2	One dimensional deep sigmoid neural network	48
3.3.3	Matrix sensing	49
3.3.4	Nonsmooth matrix factorization	49
Chapter 4:	Convergence of momentum methods	51
4.1	Convergence results	54
4.2	Proof of the length formula	67
4.3	Proof of Lemma 4.2	79
4.4	Proof of Theorem 4.3	83
Chapter 5:	Proximal random reshuffling algorithm	89
5.1	Literature review	93
5.1.1	Nonsmooth component functions	94
5.1.2	Locally smooth component functions	95
5.2	Main results	98
5.2.1	Definitions	98
5.2.2	Assumptions	101

5.2.3	Theorems	103
5.2.4	Examples	108
5.3	Proofs of main results	111
5.3.1	Tracking lemma	112
5.3.2	Reachability of (ϵ, δ) -near approximate stationarity	121
5.3.3	Convergence to (ϵ, δ) -near approximate stationarity	122
5.3.4	Convergence to $(\epsilon, 0)$ -near approximate stationarity	123
5.3.5	Convergence to $(0, 0)$ -near approximate stationarity	124
5.4	Proof of intermediate results	139
5.4.1	Proof of Proposition 5.2	139
5.4.2	Proof of Theorem 5.4	141
5.4.3	Projection formula	149
5.4.4	Continuous length formula	152
5.4.5	Uniform boundedness of subgradient trajectories	156
	References	161

List of Figures

3.1	Gradient method initialized uniformly at random in $[-1, 1]^4$ with constant step size 0.01 sometimes gets stuck at a spurious local minimum at infinity (3 among 10 trials in the experiment).	34
3.2	Function devoid of spurious valleys containing a spurious local minimum at infinity.	35
3.3	Local minimum at infinity of $f(w_1, w_2) = \frac{1}{2}[(w_2\sigma(w_1) - 1)^2 + (w_2\sigma(-w_1) + 3)^2]$.	38
3.4	An example of function with infinitely many critical values	45
5.1	Iterates of the gradient descent with different constant step sizes applied to $f(x, y) = x^2y^2 - x - y$	90
5.2	Iterate norms and function values of the subgradient method with constant step sizes $\alpha_k = \alpha$ applied to ℓ_1 matrix completion instance $f(x_1, x_2, y_1, y_2) := 1 - x_1y_1 + 1 - x_2y_1 + 1 - x_2y_2 $	91
5.3	Iterate norms and function values of the subgradient method with non-summable diminishing step sizes $\alpha_k = \alpha/(k + 1)^\beta$ applied to robust matrix completion instance.	91
5.4	Iterate norms and function values of the subgradient method with summable diminishing step sizes $\alpha_k = \alpha/(k + 1)^\beta$ applied to robust matrix completion instance.	92
5.1	Visualization of function evolution in the proof of Proposition 5.1.	154

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Cédric Jozs. This thesis would not have been possible without his unwavering support and guidance. I have greatly benefited from his innovative ideas in continuous optimization, his rigorous approach to mathematical proofs, and his strategic mentorship. His personal strength, especially in the face of a prolonged and difficult health battle, has been a source of deep inspiration to me. I wish him continued strength and recovery.

I am also thankful to Donald Goldfarb, Henry Lam, Mert Gürbüzbalaban, and Tianyi Lin for serving on my dissertation committee and for their thoughtful feedback, encouragement, and support throughout the process.

I would like to thank my colleagues and collaborators at Columbia IEOR for the stimulating discussions, fruitful collaborations, and shared experiences in coursework, sports, or travel. In particular, I am grateful to Lexiao Lai for his deep technical insight and engaging research discussions, as well as for his generous support with both academic logistics and daily life matters.

I am also grateful to the administrative staff at Columbia IEOR, especially Lizbeth Morales (former), Winsor Yang, and Shi Yee Lee, whose professionalism greatly assisted my progress through the program.

Lastly, I would like to thank those who supported me professionally during my academic development and job search, especially Jim Dai, Tom Luo, and David Yao.

Dedication

To my parents, Jun Xiao and Weiping Li, for their unwavering love, patience, and sacrifices throughout every stage of my life. Their constant support has sustained me in countless ways, from teaching me how to learn and live well, to encouraging me to persevere through every challenge. This work would not have been possible without them.

To my former partner, Liang Ou, for her steady encouragement, support, and belief in me during the most challenging moments of this journey. Her constant affirmation gave me the courage to be myself, and her companionship inspired me to pursue what I truly value and enjoy. I remain deeply grateful for her presence during this formative chapter of my life.

This work is dedicated to you.

Introduction

The past decade has witnessed tremendous advances in data science, enabling remarkable progress in fields from classical computer vision tasks such as recommendation systems [1, 2, 3] and face recognition [4, 5, 6] to more recent advancements in autonomous driving [7, 8] and large language models [9, 10]. The success of these models relies heavily on the landscape of their loss functions and the optimization algorithms used to train them. In the era of high-dimensional large-scale data, first-order algorithms, namely the optimization algorithms that require only first-order derivatives, becomes more and more prevailing due to its low per-iteration costs [11, 12]. However, their applications often lack guarantees, raising concerns about their robustness and reliability. Practitioners often need multiple executions to fine-tune parameter settings, leading to significant energy consumption and increased computational costs [13, 14].

A central challenge in establishing theoretical guarantees for the applications discussed above is that classical optimization theory often rests on assumptions such as convexity, smoothness, or coercivity of the objective function [15, 16, 17], which are frequently violated in modern data science problems [18, 19]. As illustrated in Example 3.1, gradient descent can fail even on a simple matrix completion task due to the absence of these regularity properties. In addition, standard assumptions on the algorithms, such as the boundedness of iterates or the existence of limit points [20, 21, 22, 23], may not provide a satisfactory explanation for their empirical success. For example, Figure 5.1 shows that even when the iterates of gradient descent remain bounded, it may fail to converge. Figure 5.3 demonstrates that the subgradient method with a standard diminishing step size can generate unbounded iterates in the context of ℓ_1 matrix completion (unconstrained version

of Example 5.2). Lastly, Figure 5.4 underscores the difficulty of empirically verifying boundedness in practice.

In this thesis, we aim to provide theoretical guarantees for minimizing a broad class of functions using first-order methods. To address the challenges posed by the absence of a global Lipschitz constant and the lack of standard boundedness assumptions on algorithmic iterates, we adopt a continuous time dynamics perspective inspired by [24, 25, 26, 27, 22, 28, 29, 30, 31]. Specifically, we utilize subgradient trajectories (Definition 1.6) to analyze the function landscape and the behavior of first-order algorithms. We introduce the notion of bounded subgradient trajectories (Definition 1.7) as a key condition for characterizing the global dynamics of some first-order algorithms. Away from critical points, algorithmic iterates are shown to closely track a bounded subgradient trajectory, approaching arbitrarily near a critical point. Once the iterates enter a neighborhood of a critical point, they either remain within the neighborhood and converge to a stationary point, or they exit the region with a sufficiently large decrease in the objective value. Since such escapes can occur only finitely many times, the iterates ultimately stabilize and converge.

The results in this thesis rely on a key structural assumption that the objective function is semialgebraic (Definition 1.4). This class is sufficiently expressive to model a wide range of problems arising in data science [32, 33, 28], while remaining structured enough to permit the use of powerful tools from real algebraic geometry and variational analysis. In contrast, general functions may exhibit highly oscillatory behavior, which rarely occurs in data science applications [34, 35]. A review of the relevant properties of semialgebraic functions is provided in Section 1.2. To accommodate functions involving components such as exponentials or logarithms, one may consider the broader class of functions that are definable in an o-minimal structure over the real field [36, 37]. Since all the results in this thesis stated for semialgebraic functions extend naturally to definable functions, we focus primarily on the semialgebraic setting for clarity and simplicity.

This thesis is organized as follows. Chapter 1 provides the necessary background for the subsequent chapters, introducing the concept of Clarke subdifferential, semialgebraic functions, and subgradient trajectories, along with their fundamental properties used throughout the thesis. In

Chapter 2, we identify a collection of representative models from data science that exhibit bounded subgradient trajectories; these results serve as the foundation for the landscape and convergence analyses in the later chapters. Chapter 3 introduces the notion of spurious local minima at infinity and presents a practical tool (Theorem 3.1) to certify the absence of such undesirable behavior by leveraging trajectory boundedness. In Chapter 4, we investigate momentum methods for locally smooth semialgebraic functions with bounded subgradient trajectories. Using a specialized length formula (Lemma 4.1), we establish results on local convergence (Theorem 4.1), global convergence (Theorem 4.2), and saddle point avoidance (Theorem 4.3). Finally, Chapter 5 extends this framework to nonsmooth constrained settings and stochastic algorithms. We analyze the proximal random reshuffling algorithm for composite finite-sum problems and show that, under various regularity assumptions on the objective, the algorithm can recover approximate stationary points with quantifiable accuracy (Theorems 5.1 to 5.4).

Chapter 1: Background

We first list some basic notations and concepts. Let $\mathbb{N} := \{0, 1, 2, \dots\}$, $\mathbb{N}^* := \{1, 2, \dots\}$, $\mathbb{R} := (-\infty, \infty)$, $\mathbb{R}_+ := [0, \infty)$, $\mathbb{R}_{++} := (0, \infty)$, and $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$. Let $\|\cdot\|$ be the induced norm of an inner product $\langle \cdot, \cdot \rangle$ on \mathbb{R}^n . Given $S \subset \mathbb{R}^n$, let $\overset{\circ}{S}$ and \overline{S} denote the interior and closure of S in \mathbb{R}^n respectively. Let $B(a, r)$ and $\overset{\circ}{B}(a, r)$ respectively denote the closed and open balls of center $a \in \mathbb{R}^n$ and radius $r \geq 0$. Given $x \in \mathbb{R}^n$, consider the distance of x to S defined by $d(x, S) := \inf\{\|x - y\| : y \in S\}$. Given a set-valued mapping $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ and $y \in \mathbb{R}^m$, let $F^{-1}(y) := \{x \in \mathbb{R}^n : F(x) \ni y\}$. Given $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, the domain, graph, and epigraph are respectively given by $\text{dom } f := \{x \in \mathbb{R}^n : f(x) < \infty\}$, $\text{graph } f := \{(x, t) \in \mathbb{R}^n \times \mathbb{R} : \Phi(x) = t\}$, and $\text{epi } f := \{(x, t) \in \mathbb{R}^{n+1} : f(x) \leq t\}$. A function $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is convex (respectively lower semicontinuous) if $\text{epi } f$ is convex (respectively closed).

1.1 Clarke subdifferential

In this section, we will review some concepts and results on generalized derivative in the sense of Clarke [38, p. 336], since we would like to also consider nonsmooth functions. We begin by recalling the definition of several normal cones [38, Definition 6.3].

Definition 1.1. Given $S \subset \mathbb{R}^n$ and $\bar{x} \in S$, the regular normal, normal, and convexified normal cones are given respectively by

$$\begin{aligned} \widehat{N}_S(\bar{x}) &:= \left\{ v \in \mathbb{R}^n : \limsup_{\substack{x \rightarrow \bar{x}, x \neq \bar{x} \\ S}} \frac{\langle v, x - \bar{x} \rangle}{\|x - \bar{x}\|} \leq 0 \right\}, \\ N_S(\bar{x}) &:= \left\{ v \in \mathbb{R}^n : \exists (x_k, v_k) \rightarrow (\bar{x}, v), (x_k, v_k) \in S \times \widehat{N}_S(x_k) \right\}, \\ \overline{N}_S(\bar{x}) &:= \overline{\text{conv}} N_S(\bar{x}), \end{aligned}$$

where $x \xrightarrow[S]{} \bar{x}$ means $x \in S$ converges to \bar{x} .

Definition 1.2. Given $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, the Clarke subdifferential is the set-valued mapping $\partial\Phi : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined by

$$\partial f(\bar{x}) := \begin{cases} \{v \in \mathbb{R}^n : (v, -1) \in \overline{N}_{\text{epi } f}((\bar{x}, f(\bar{x}))\} & \text{if } |f(\bar{x})| < \infty \\ \emptyset & \text{else.} \end{cases}$$

We say $x \in \text{dom}(f)$ is a Clarke critical point if $0 \in \partial f(x)$ and $v \in \mathbb{R}$ is a Clarke critical value if $f(x) = v$ for some $x \in \text{dom}(f)$ such that $0 \in \partial f(x)$.

Definition 1.3. A function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is *locally Lipschitz* if for all $a \in \mathbb{R}^n$, there exist positive constants r and L such that

$$\forall x, y \in B(a, r), \quad \|f(x) - f(y)\| \leq L\|x - y\|.$$

Notice that for a locally Lipschitz function, by [39, Theorem 3.2], the derivative exists almost everywhere. It is also well known that for any locally Lipschitz function f and any $x \in \mathbb{R}^n$, the Clarke subdifferential $\partial f(x)$ is a nonempty, convex, and compact set [40, Proposition 2.1.2(a)].

1.2 Semialgebraic functions

Definition 1.4. [41, 42] A subset S of \mathbb{R}^n is *semialgebraic* if it is a finite union of sets of the form $\{x \in \mathbb{R}^n : p_i(x) = 0, i = 1, \dots, k; p_i(x) > 0, i = k + 1, \dots, m\}$ where p_1, \dots, p_m are polynomials defined from \mathbb{R}^n to \mathbb{R} . A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is semialgebraic if its graph, that is to say $\{(x, t) \in \mathbb{R}^{n+1} : f(x) = t\}$, is a semialgebraic set.

Next we recall several useful properties of semialgebraic functions. They will be frequently used in later chapters.

Lemma 1.1 (semialgebraic Morse-Sard theorem [43]). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be locally Lipschitz and semialgebraic. Then f has finitely many critical values.*

Theorem 1.1 (Kurdyka-Łojasiewicz inequality [44, 43]). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be locally Lipschitz and semialgebraic. Let X be a bounded subset of \mathbb{R}^n and $v \in \mathbb{R}$ be a critical value of f in \overline{X} . There exists $\rho > 0$ and a strictly increasing continuous semialgebraic function $\psi : [0, \rho) \rightarrow [0, \infty)$ which belongs to $C^1((0, \rho))$ with $\psi(0) = 0$ such that*

$$\forall x \in X, \quad |f(x) - v| \in (0, \rho) \implies d(0, \partial(\psi \circ |f - v|)(x)) \geq 1. \quad (1.1)$$

Proposition 1.1 (Uniform Kurdyka-Łojasiewicz inequality [45, Proposition 5]). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be locally Lipschitz and semialgebraic. Let X be a bounded subset of \mathbb{R}^n and V be the set of critical values of f in \overline{X} if it is non-empty, otherwise $V := \{0\}$. There exists a concave semialgebraic diffeomorphism $\psi : [0, \infty) \rightarrow [0, \infty)$ such that*

$$\forall x \in X \setminus (\partial \tilde{f})^{-1}(0), \quad d(0, \partial(\psi \circ \tilde{f})(x)) \geq 1, \quad (1.2)$$

where $\tilde{f}(x) := d(f(x), V)$ for all $x \in \mathbb{R}^n$.

1.3 Subgradient trajectories

In this section, we will introduce some basic concepts and fundamental properties related to subgradient trajectories.

Definition 1.5. [46, Definition 1 p. 12] Given two real numbers $a < b$, a function $x : [a, b] \rightarrow \mathbb{R}^n$ is *absolutely continuous* if for all $\epsilon > 0$, there exists $\delta > 0$ such that, for any finite collection of disjoint subintervals $[a_1, b_1], \dots, [a_m, b_m]$ of $[a, b]$ such that $\sum_{i=1}^m (b_i - a_i) \leq \delta$, we have $\sum_{i=1}^m \|x(b_i) - x(a_i)\| \leq \epsilon$.

By virtue of [47, Theorem 20.8], $x : [a, b] \rightarrow \mathbb{R}^n$ is absolutely continuous if and only if it is differentiable almost everywhere on (a, b) , its derivative x' is Lebesgue integrable, and $x(t) - x(a) = \int_a^t x'(\tau) d\tau$ for all $t \in [a, b]$. Given a non-compact interval I of \mathbb{R} , $x : I \rightarrow \mathbb{R}^n$ is absolutely continuous if it is absolutely continuous on all compact subintervals of I .

Definition 1.6. An absolutely continuous function $x : [0, \infty) \rightarrow \mathbb{R}^n$ is called a *subgradient trajectory* of $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ starting at $x_0 \in \text{dom}(\partial f)$ if it satisfies the following differential inclusion with initial condition:

$$x'(t) \in -\partial f(x(t)), \quad \text{for almost every } t \geq 0, \quad x(0) = x_0, \quad (1.3)$$

where “almost every” means all elements except for those in a set of zero measure.

However, a subgradient trajectory may not always exist for arbitrary f , even if f is a smooth function. Let $f(x) = -\frac{1}{3}x^3$ and $x_0 = 1$, then it is easy to see $x(t) = \frac{1}{1-t}$ is the unique solution for $t \in [0, 1)$ and it cannot be extended to an absolutely continuous function on $[0, \infty)$ due to the singularity at $t = 1$. In this case, one would seek a family of functions including many loss functions arising in applications that guarantee the existence of a subgradient trajectory.

We say a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is *bounded below* if $\inf_{\mathbb{R}^n} f = c > -\infty$. It was shown in [48, Theorem 3.2] that a primal lower nice function bounded below by a linear function suffices. However, in general it is not easy to check whether those nonconvex functions in statistical learning problems are primal lower nice. For easily checkable conditions, the following result generalized from [49, Proposition 2.3] for differentiable functions tells us that a locally Lipschitz function bounded below also suffices.

Proposition 1.2. *If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is locally Lipschitz and bounded below, then there exists a subgradient trajectory of f starting at arbitrary $x_0 \in \mathbb{R}^n$.*

Proof. For a fixed real number $\tau > 0$, define a sequence x_k^τ recurrently by letting $x_0^\tau := x_0$ and

$$x_{k+1}^\tau \in \arg \min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{\|x - x_k^\tau\|^2}{2\tau} \right\}, \quad \forall k \in \mathbb{N}.$$

A solution exists because f is bounded below and the objective function is coercive. Any solution satisfies

$$v_{k+1}^\tau := \frac{x_{k+1}^\tau - x_k^\tau}{\tau} \in -\partial f(x_{k+1}^\tau), \quad \forall k \in \mathbb{N}.$$

Define two functions $x^\tau, \tilde{x}^\tau : \mathbb{R}_+ \rightarrow \mathbb{R}^n$ where $\mathbb{R}_+ := [0, \infty)$ by

$$x^\tau(t) := x_{k+1}^\tau, \quad \tilde{x}^\tau(t) := x_k^\tau + (t - k\tau)v_{k+1}^\tau, \quad \forall t \in (k\tau, (k+1)\tau]$$

for all $k \in \mathbb{N}$, with initial condition $x^\tau(0) = \tilde{x}^\tau(0) = x_0$. Note that \tilde{x}^τ is absolutely continuous because it is piecewise affine. On the contrary, x^τ is not continuous. Also, define $v^\tau : \mathbb{R}_+ \rightarrow \mathbb{R}^n$ by

$$v^\tau(t) := v_{k+1}^\tau, \quad \forall t \in (k\tau, (k+1)\tau], \quad \forall k \in \mathbb{N},$$

and choose $v^\tau(0) \in -\partial f(x_0)$. Since $(\tilde{x}^\tau)' = v^\tau$ on $(k\tau, (k+1)\tau)$ for all $k \in \mathbb{N}$, and $v^\tau(t) \in -\partial f(x^\tau(t))$ for all $t \geq 0$, we conclude that $(\tilde{x}^\tau)'(t) \in -\partial f(x^\tau(t))$ for almost every $t \in \mathbb{R}_+$. By optimality of x_{k+1}^τ , we have

$$f(x_{k+1}^\tau) + \frac{\|x_{k+1}^\tau - x_k^\tau\|^2}{2\tau} \leq f(x_k^\tau), \quad \forall k \in \mathbb{N}.$$

For any $l \in \mathbb{N}$, we have

$$\sum_{k=0}^l \frac{\|x_{k+1}^\tau - x_k^\tau\|^2}{2\tau} \leq f(x_0^\tau) - f(x_{l+1}^\tau) \leq f(x_0) - \inf_{\mathbb{R}^n} f =: C < \infty$$

since f is bounded below. Observe that

$$\sum_{k=0}^l \frac{\|x_{k+1}^\tau - x_k^\tau\|^2}{2\tau} = \sum_{k=0}^l \frac{\tau}{2} \|v_{k+1}^\tau\|^2 = \frac{1}{2} \sum_{k=0}^l \int_{k\tau}^{(k+1)\tau} \|(\tilde{x}^\tau)'(t)\|^2 dt.$$

Fix $T \geq 0$ from now on. From the above, we have

$$\int_0^T \|(\tilde{x}^\tau)'(t)\|^2 dt \leq \sum_{k=0}^{\lfloor T/\tau \rfloor} \int_{k\tau}^{(k+1)\tau} \|(\tilde{x}^\tau)'(t)\|^2 dt \leq 2C. \quad (1.4)$$

Since \tilde{x}^τ is absolutely continuous, for any $s, t \in [0, T]$ we have

$$\|\tilde{x}^\tau(t) - \tilde{x}^\tau(s)\| = \left\| \int_s^t (\tilde{x}^\tau)'(u) du \right\| \quad (1.5a)$$

$$\leq \left(\int_0^T \|(\tilde{x}^\tau)'(t)\|^2 dt \right)^{1/2} |t - s|^{1/2} \leq \sqrt{2C} |t - s|^{1/2} \quad (1.5b)$$

where we use the Cauchy-Schwarz inequality. Now one can see $(\tilde{x}^\tau)_{\tau>0}$ is a family of uniformly bounded and equicontinuous functions on the compact interval $[0, T]$. Therefore, by Arzelà-Ascoli theorem [50, Theorem 7.25], there exists a sequence of positive reals $(\tau_k)_{k \in \mathbb{N}}$ such that $\tau_k \rightarrow 0$ and $\tilde{x}^{\tau_k} \rightarrow x^*$ uniformly on $[0, T]$ as $k \rightarrow \infty$. For all $k \in \mathbb{N}$ and $t \in (k\tau, (k+1)\tau]$, we have $\tilde{x}^\tau((k+1)\tau) = x_k^\tau + \tau v_{k+1}^\tau = x_{k+1}^\tau = x^\tau(t)$. Thus $\|\tilde{x}^\tau(t) - x^\tau(t)\| = \|\tilde{x}^\tau(t) - \tilde{x}^\tau((k+1)\tau)\| \leq \sqrt{2C}\tau^{1/2}$ for all $t \in [0, T]$ where the inequality is due to (1.5) (take $s := (k+1)\tau$). Combined with the fact that $\tilde{x}^{\tau_k} \rightarrow x^*$ uniformly on $[0, T]$, one can see that $x^{\tau_k} \rightarrow x^*$ uniformly on $[0, T]$. Since (1.4) implies that $((\tilde{x}^{\tau_k})')_{k \in \mathbb{N}}$ is a bounded sequence in $L^2([0, T], \mathbb{R}^n)$, there exists a subsequence $(\tau_{k_j})_{j \in \mathbb{N}}$ such that $(\tilde{x}^{\tau_{k_j}})' \rightarrow v^*$ weakly in $L^1([0, T], \mathbb{R}^n)$ as $j \rightarrow \infty$ by [51, Corollary 14 p. 413]. Since $\tilde{x}^{\tau_{k_j}}$ is absolutely continuous, for all $t \in [0, T]$, we have

$$\tilde{x}^{\tau_{k_j}}(t) - \tilde{x}^{\tau_{k_j}}(0) = \int_0^t (\tilde{x}^{\tau_{k_j}})'(u) du.$$

Take $j \rightarrow \infty$ on both sides, we have

$$x^*(t) - x^*(0) = \int_0^t v^*(u) du,$$

where the convergence of the integral relies on the fact that the constant functions equal to the canonical basis of \mathbb{R}^n lie in $L^\infty([0, T], \mathbb{R}^n)$. Thus, x^* is absolutely continuous and $(x^*)'(t) = v^*(t)$ for almost every $t \in [0, T]$. Recall that for all $k \in \mathbb{N}$, it holds for almost every $t \in [0, T]$ that

$$(\tilde{x}^{\tau_k})'(t) = v^{\tau_k}(t) \in -\partial f(x^{\tau_k}(t)).$$

Since f is locally Lipschitz, the set-valued function $-\partial f$ is upper semicontinuous [40, 2.1.5 Proposition (d) p. 29] with nonempty compact values [40, 2.1.2 Proposition (a) p. 27], hence proper upper hemicontinuous [46, Proposition 1 p. 60]. In addition, $x^{\tau_k} \rightarrow x^*$ uniformly on $[0, T]$ and $(\tilde{x}^{\tau_k})' \rightarrow (x^*)'$ weakly in $L^1([0, T], \mathbb{R}^n)$. Therefore, $(x^*)'(t) \in -\partial f(x^*(t))$ for almost all $t \in [0, T]$ by [46, Theorem 1 p. 60]¹. The initial condition also holds since $\tilde{x}^\tau(0) = x_0$ for all $\tau > 0$.

We have proved that for any initial point x_0 , there exists $x^* : [0, T] \rightarrow \mathbb{R}^n$ such that $(x^*)'(t) = -\partial f(x^*(t))$ holds for almost every $t \in [0, T]$ with any $T > 0$. Since T is independent of x_0 , by setting $T = 1$, there exists a sequence of absolutely continuous functions $(x_k)_{k \in \mathbb{N}}$ such that

$$x'_k(t) \in -\partial f(x_k(t)), \quad \text{for a.e. } t \in [0, 1], \quad x_k(0) = x_{k-1}(1),$$

for all $k \in \mathbb{N}$ where $x_{-1}(0) = x_0$. Therefore, the desired function $x : [0, \infty) \rightarrow \mathbb{R}^n$ can be defined in a piecewise fashion by

$$x(t) := x_k(t - k), \quad t \in [k, k + 1), \quad \forall k \in \mathbb{N}.$$

By construction, x is absolutely continuous on any compact interval $[a, b] \subset [0, \infty)$. □

We remark here that with Proposition 1.2, one can recover Ekeland's variational principle [52, Corollary 2.3] [53, Corollary] for locally Lipschitz lower bounded functions with a chain rule (see [54, Theorem 3.1] for an extension to lower semi-continuous lower bounded functions). Indeed, Proposition 1.2 implies that for all $\epsilon > 0$, there exists $(x, s) \in \text{graph } \partial f$ such that $f(x) \leq \inf f + \epsilon$ and $\|s\| \leq \epsilon$.² Note that Proposition 1.2 only guarantees the existence of a solution to (1.3) for all $t \geq 0$, but the solution $x(t)$ could go to infinity as $t \rightarrow \infty$. This motivates the following definition.

Definition 1.7. A function $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ has *bounded subgradient trajectories* if for any $x_0 \in \text{dom}(f)$, there exists a constant $r > 0$, such that for any subgradient trajectory x of f starting at x_0 ,

¹In the theorem we take $F := -\partial f$, $X = Y := \mathbb{R}^n$, and $I := [0, T]$.

²This follows from the formula $f(x(t)) - \inf f \geq \int_t^\infty d(0, \partial f(x(\tau)))^2 d\tau$ where $d(x, X) := \inf_{y \in X} \|x - y\|$ (see [22, Lemma 5.2] and [55, Proposition 4.10]).

we have $\|x(t)\| \leq r$ for all $t \geq 0$.

Finally, notice that when f is continuously differentiable, by [40, Proposition 2.2.4], (1.3) reduces to the classical Cauchy problem of differential equation

$$x'(t) = -\nabla f(x(t)), \quad \text{for all } t \geq 0, \quad x(0) = x_0.$$

and subgradient trajectory reduces to gradient trajectory by imposing x to be continuously differentiable. Recall the descent property of gradient trajectories [56, Proposition 17.1.1], i.e., $f \circ x$ is a decreasing function for any gradient trajectory x of f . We want this nice property to hold even in a more general case. We adopt the notion of chain rule in [22, Definition 5.1]. Note that functions admitting a chain rule are also referred to as path differentiable [57, Definition 3].

Definition 1.8. Let $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be locally Lipschitz over $\text{dom}(f)$. We say f admits a chain rule if for any absolutely continuous function $x : [0, \infty) \rightarrow \mathbb{R}^n$, we have

$$(f \circ x)'(t) = \langle v, x'(t) \rangle, \quad \forall v \in \partial f(x(t)),$$

for almost every $t \in [0, \infty)$.

Thus, for any locally Lipschitz function that admits a chain rule, by [22, Lemma 5.2], the function value is always decreasing in time along the subgradient trajectory. A detailed discussion on what class of functions admits a chain rule can be found in [57]. Note that general Lipschitz functions are far from admitting a chain rule since they generically have a maximal Clarke subdifferential [34, 58, 59].

Chapter 2: Typical models with bounded subgradient trajectories

In this chapter, we focus on the main condition of this thesis: the boundedness of subgradient trajectories. We formalize this notion and show that it holds in a wide range of models commonly encountered in data science. The results on bounded subgradient trajectories form a foundational tool for the convergence and landscape analyses developed in the subsequent chapters. To help with exposition, we proceed in order of increasing complexity, starting with convex models, followed by nonconvex coercive problems, and concluding with nonconvex noncoercive settings.

2.1 Convex model

Theorem 2.1. *For a convex proper l.s.c. function $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, suppose $x^* \in \text{dom}(f)$ is a minimizer, then f has bounded subgradient trajectories.*

Proof. By [56, Theorem 17.2.2], for any initial point $x_0 \in \text{dom}(f)$, there exists a unique subgradient trajectory $x : \mathbb{R}_+ \rightarrow \mathbb{R}^n$. It is easy to see $t \mapsto \|x(t) - x^*\|^2$ is absolutely continuous and the chain rule can be applied so that for a.e. $t \in \mathbb{R}_+$,

$$\frac{d}{dt} \|x(t) - x^*\|^2 = 2\langle x'(t), x(t) - x^* \rangle = -2\langle g_t, x(t) - x^* \rangle \leq -2(f(x(t)) - f(x^*)) \leq 0,$$

where $g_t \in \partial f(x(t))$. Thus, $\|x(t) - x^*\| \leq \|x_0 - x^*\|$ for all $t \in \mathbb{R}_+$, which implies $\|x(t)\| \leq \|x_0 - x^*\| + \|x^*\|$. Therefore, f has bounded subgradient trajectories. \square

Example 2.1. One of the most fundamental and widely used convex models in data science is linear regression. A general form of linear regression can be written as a special case of a linear

feedforward neural network without hidden layers [60], given by

$$f(W) := \sum_{i=1}^m \|Wx_i - y_i\|^2, \quad (2.1)$$

where $W \in \mathbb{R}^{n \times r}$ is the parameter matrix, $x_i \in \mathbb{R}^r$ are the input vectors, and $y_i \in \mathbb{R}^n$ are the target outputs for $i = 1, \dots, m$. The function f is convex, as it is a sum of convex quadratic functions. More importantly, f always admits a minimizer for any given dataset $\{(x_i, y_i)\}_{i=1}^m$, since the associated optimality condition is the normal equation, whose solution exists even if the original system $Wx_i = y_i$ is inconsistent. Consequently, Theorem 2.1 is applicable in this setting.

2.2 Nonconvex coercive model

Recall that a function $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is said to be *coercive* if $f(x) \rightarrow \infty$ as $\|x\| \rightarrow \infty$. Coercive functions arise naturally in data science, particularly when regularization is used to control model complexity. Intuitively, the coercive shape of the function prevents subgradient trajectories from escaping to infinity, as long as the function value decreases along the trajectory.

Proposition 2.1. *For a locally Lipschitz continuous function $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, if it is coercive and satisfies chain rule, then f has bounded subgradient trajectories.*

Proof. The existence of subgradient trajectories is ensured by Proposition 1.2. For any absolutely continuous x , $f \circ x$ is also absolutely continuous and $(f \circ x)'(t) = \langle v, x'(t) \rangle$, for all $v \in \partial f(x(t))$. Since $x'(t) \in -\partial f(x(t))$, by taking $v = -x'(t)$, we obtain $(f \circ x)'(t) = -\|x'(t)\|^2 < 0$. Thus, $f \circ x$ is decreasing and $f(x(t)) \leq f(x_0)$ for all $t \in \mathbb{R}_+$. By coercivity of f , x is bounded and f has bounded subgradient trajectories. \square

Example 2.2. Coercive functions frequently appear in data science models through regularization. One classical example is a ReLU neural network [19] with ℓ_2 -regularization:

$$\mathcal{L}(W) := \sum_{i=1}^m \|W_L \sigma(W_{L-1} \cdots \sigma(W_1 x_i) \cdots) - y_i\|^2 + \lambda \sum_{\ell=1}^L \|W_\ell\|_F^2, \quad (2.2)$$

where $\sigma(z) = \max\{0, z\}$ is the ReLU activation and $\lambda > 0$. The regularization renders the function coercive, ensuring the boundedness of gradient or subgradient trajectories during training.

2.3 Nonconvex noncoercive model

2.3.1 Phase Retrieval

The problem of solving systems of quadratic equations of the form $y_i = \langle a_i, x \rangle^2$, $1 \leq i \leq m$, has applications in numerous contexts. One of the most classical applications is the so-called phase retrieval problem. This problem has attracted high interest due to its broad applications in X-ray crystallography [61], microscopy [62], astronomy [63] and optical imaging [64]. Here we consider a slightly more general formulation

$$\min_x f(x) := \sum_{i=1}^m (\langle A_i x, x \rangle - y_i)^2. \quad (2.3)$$

Proposition 2.2. *Let $A_i \in \mathbb{R}^{n \times n}$ be symmetric positive semidefinite and $y_i \in \mathbb{R}$ for all $i = 1, \dots, m$. Then (2.3) has bounded subgradient trajectories.*

Proof. Since A_i is real symmetric and positive semidefinite for all $i = 1, \dots, m$, by orthogonal decomposition, we can write $A_i = \sum_{j=1}^r \lambda_{ij} v_{ij} v_{ij}^T$, where $(v_{ij})_{j=1}^r$ are orthonormal, for some $1 \leq r \leq n$ and $\lambda_{ij} > 0$ for all $j = 1, \dots, r$. Define

$$V := \text{Span}\{v_{ij} : i = 1, \dots, m, j = 1, \dots, r\}.$$

Notice that

$$\nabla f(x) = 4 \sum_{i=1}^m (\langle A_i x, x \rangle - b_i) A_i x = 4 \sum_{i=1}^m \sum_{j=1}^r \lambda_{ij} \langle v_{ij}, x \rangle (\langle A_i x, x \rangle - b_i) v_{ij} \in V.$$

Therefore, $\nabla f(x) \in V$ for all $x \in \mathbb{R}^{2n}$. Denote V^\perp as the orthogonal complement of the subspace V , then for any given initial point x_0 , the subgradient trajectory $x(\cdot)$ of f can be decomposed as $x(t) =$

$x_V(t) + x_{V^\perp}(t)$, where $x_V(t) \in V$ and $x_{V^\perp}(t) \in V^\perp$ for all $t \geq 0$. Note that $x'(t) = -\nabla f(x(t)) \in V$, thus $x_{V^\perp}(t) \equiv x_{V^\perp}(0)$ and we write $x(t) = x_V(t) + x_{V^\perp}(0)$ for all $t \geq 0$. Since $f(x)$ is a decreasing function over $t \geq 0$,

$$\begin{aligned} \sum_{j=1}^r \lambda_{ij} \langle v_{ij}, x(t) \rangle^2 &= |\langle A_i x(t), x(t) \rangle| \leq \sqrt{2(\langle A_i x(t), x(t) \rangle - b_i)^2 + 2b_i^2} \\ &\leq \sqrt{2f(x(t)) + 2b_i^2} \leq \sqrt{2f(x_0) + 2b_i^2}. \end{aligned}$$

Recall that $\lambda_{ij} > 0$ for all $j = 1, \dots, r$, hence $\langle v_{ij}, x(t) \rangle$ is bounded over $t \geq 0$ and so is $\langle v_{ij}, x_V(t) \rangle$. As $\text{Span}\{v_{ij} : i = 1, \dots, m, j = 1, \dots, r\} = V$, we can extract a basis of vectors v_{ij} to form a basis of V , and denote this basis as $\{u_\ell : \ell = 1, \dots, d\}$. Then one can write $x_V(t) = \sum_{\ell=1}^d \zeta_\ell(t) u_\ell$. Notice that for each $\ell = 1, \dots, d$, there must exist (i_ℓ, j_ℓ) such that $v_{i_\ell j_\ell} = u_\ell$. Thus,

$$\|x_V(t)\|^2 = \sum_{\ell=1}^d \zeta_\ell(t)^2 = \sum_{\ell=1}^d \langle u_\ell, x_V(t) \rangle^2 = \sum_{\ell=1}^d \langle v_{i_\ell j_\ell}, x_V(t) \rangle^2$$

is bounded over $t \geq 0$. Finally, $x(t) = x_V(t) + x_{V^\perp}(0)$ is bounded over $t \geq 0$. \square

2.3.2 Asymmetric matrix sensing

Matrix sensing is a widely used model in computer vision and statistics; see for instance [65, 66]. Given $r \geq 1$, the goal is to recover an unknown target matrix $M \in \mathbb{R}^{n_1 \times n_2}$ of rank less than or equal to r from a set of linear measurements $b_i = \langle A_i, M \rangle_F$, where $A_i \in \mathbb{R}^{n_1 \times n_2}$ for $i = 1, \dots, m$ are sensing matrices and $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. In order to do so, we minimize the mean square loss

$$f(X, Y) := \frac{1}{2m} \sum_{i=1}^m (\langle A_i, XY^T \rangle_F - b_i)^2. \quad (2.4)$$

where $X \in \mathbb{R}^{n_1 \times r}$ and $Y \in \mathbb{R}^{n_2 \times r}$.

A sufficient condition is to require the sensing matrices to be *lower bounded*, i.e., there exists

a constant $c > 0$ such that for any matrix $\tilde{M} \in \mathbb{R}^{n_1 \times n_2}$ with $\text{rank}(\tilde{M}) \leq r$,

$$\frac{1}{m} \sum_{i=1}^m \langle A_i, \tilde{M} \rangle_F^2 \geq c \|\tilde{M}\|_F^2.$$

A special case of lower bounded sensing matrices is to take each A_i be a matrix unit $E_{j,k}$, i.e., a matrix with only one nonzero entry at j -th row and k -th column with value 1. For example, we can let $m = n_1 n_2$, and $A_1 = E_{1,1}, A_2 = E_{1,2}, \dots, A_{n_1 n_2} = E_{n_1, n_2}$. In this case, $\sum_{i=1}^m \langle A_i, \tilde{M} \rangle_F^2 = \|\tilde{M}\|_F^2$, so the above condition holds. In fact, under such setting, the objective function is equivalent to simple matrix factorization $f(X, Y) = \|XY^T - M\|_F^2$.

Proposition 2.3. *Matrix sensing with loss function (2.4) and lower bounded sensing matrices has bounded gradient trajectories.*

Proof. Since f is locally Lipschitz and lower bounded, by Proposition 1.2 there exists a gradient trajectory for any initial point. The gradient trajectories of f satisfy the initial value problem

$$\begin{aligned} \dot{X} &= -\frac{1}{m} \sum_{i=1}^m (\langle A_i, XY^T \rangle_F - b_i) A_i Y, \\ \dot{Y} &= -\frac{1}{m} \sum_{i=1}^m (\langle A_i, XY^T \rangle_F - b_i) A_i^T X, \\ X(0) &= X_0, \quad Y(0) = Y_0. \end{aligned}$$

Notice that $\dot{X}^T X = Y^T \dot{Y}$ and $X^T \dot{X} = \dot{Y}^T Y$, so

$$\frac{d}{dt} (X^T X - Y^T Y) = \dot{X}^T X + X^T \dot{X} - \dot{Y}^T Y - Y^T \dot{Y} = 0.$$

This implies that $X^T X - Y^T Y = C$ where $C \in \mathbb{R}^{r \times r}$ is a constant. Since the function value is decreasing along gradient trajectories [22, Lemma 5.2], there exists a constant c_1 such that $f(X(t), Y(t)) \leq c_1$ for all $t \geq 0$. Combined with the assumption that sensing matrices are lower

bounded, there exist constants c and c_2 such that

$$\begin{aligned} c\|XY^T\|_F^2 &\leq \frac{1}{m} \sum_{i=1}^m \langle A_i, XY^T \rangle_F^2 \leq \frac{1}{m} \sum_{i=1}^m [2(\langle A_i, XY^T \rangle_F - b_i)^2 + 2b_i^2] \\ &= 2f(X, Y) + \frac{2}{m} \sum_{i=1}^m b_i^2 \leq 2c_1 + \frac{2}{m} \sum_{i=1}^m b_i^2 =: c_2. \end{aligned}$$

We have $\|XY^T\|_F^2 \leq c_3 := c_2/c$. Notice that

$$\|X^T X\|_F^2 + \|Y^T Y\|_F^2 = \|X^T X - Y^T Y\|_F^2 + 2\|XY^T\|_F^2 \leq \|C\|_F^2 + 2c_3.$$

Define the constant $c_4 := 2c_3 + \|C\|_F^2$. By the Cauchy-Schwarz inequality,

$$\|X\|_F^4 + \|Y\|_F^4 \leq \text{rank}(X)\|X^T X\|_F^2 + \text{rank}(Y)\|Y^T Y\|_F^2 \leq (n_1 + n_2 + r)c_4.$$

Thus, X and Y are bounded. □

2.3.3 Nonsmooth matrix factorization

In this subsection, we consider the application of Theorem 3.1 in a nonsmooth setting, namely, the nonsmooth matrix factorization problem. We consider minimizing the loss function

$$f(X, Y) := \|XY^T - M\|_1, \tag{2.5}$$

where $X \in \mathbb{R}^{m \times r}$, $Y \in \mathbb{R}^{n \times r}$ are decision variables and $M \in \mathbb{R}^{m \times n}$ is the given data matrix. Here $\|A\|_1 := \sum_{i=1}^m \sum_{j=1}^n |A_{ij}|$ for any $A \in \mathbb{R}^{m \times n}$. In robust principal component analysis (PCA) problem with sparse noise, (2.5) is usually used as a surrogate function for the original ℓ_0 -norm formulation; see [67, 68].

To verify (2.5) has bounded subgradient trajectories, we discover that the auto-balancing property in [69, Theorem 2.2] also holds for nonsmooth matrix factorization. The result can be summarized in the following proposition.

Proposition 2.4. *Nonsmooth matrix factorization with loss function (2.5) has bounded subgradient trajectories.*

Proof. Since f is locally Lipschitz and lower bounded, by Proposition 1.2 there exists a subgradient trajectory for any initial point. Let $(X_0, Y_0) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$. Consider an absolutely continuous function $Z : [0, \infty) \rightarrow \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$ such that

$$Z'(t) \in -\partial f(Z(t)), \quad \text{for almost every } t \geq 0, \quad \text{and } Z(0) = (X_0, Y_0).$$

By [40, Theorem 2.3.10],

$$\partial f(X, Y) = \left\{ \left(\begin{array}{c} \Lambda Y \\ \Lambda^T X \end{array} \right) \middle| \Lambda \in \text{sign}(XY^T - M) \right\}$$

where sign is an element-wise operation mapping each entry of a matrix to a real number in $[-1, 1]$ such that

$$\text{sign}(x) := \begin{cases} -1 & \text{if } x < 0, \\ [-1, 1] & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

Hence, with $Z = (X, Y)$, for almost every $t \geq 0$ we have

$$X'(t) = -\Lambda(t)Y(t), \quad Y'(t) = -\Lambda(t)^T X(t), \quad (2.6a)$$

$$\Lambda(t) \in \text{sign}(X(t)Y(t)^T - M). \quad (2.6b)$$

Consider $\phi : [0, \infty) \rightarrow \mathbb{R}$ defined by $\phi(t) := X(t)^T X(t) - Y(t)^T Y(t)$. By taking derivative, we have

$$\phi'(t) = X'(t)^T X(t) + X(t)^T X'(t) - Y'(t)^T Y(t) - Y(t)^T Y'(t). \quad (2.7)$$

Combining (2.6a) and (2.7), we have

$$\begin{aligned}\phi'(t) &= -Y(t)^T \Lambda(t)^T X(t) - X(t)^T \Lambda(t) Y(t) \\ &\quad + X(t)^T \Lambda(t) Y(t) + Y(t)^T \Lambda(t)^T X(t) = 0.\end{aligned}$$

Hence the continuous function ϕ is constant on $[0, \infty)$. Also, we have

$$\begin{aligned}\|X^T X - Y^T Y\|_F^2 &= \|X^T X\|_F^2 + \|Y^T Y\|_F^2 - 2\langle X^T X, Y^T Y \rangle_F \\ &= \|X^T X\|_F^2 + \|Y^T Y\|_F^2 - 2\|XY^T\|_F^2 \\ &\geq \|X^T X\|_2^2 + \|Y^T Y\|_2^2 - 2\|XY^T\|_F^2 \\ &= \|X\|_2^4 + \|Y\|_2^4 - 2\|XY^T\|_F^2 \\ &\geq \|X\|_2^4 + \|Y\|_2^4 - 2mn\|XY^T\|_1^2 \\ &\geq \|X\|_2^4 + \|Y\|_2^4 - 2mn(\|XY^T - M\|_1 + \|M\|_1)^2.\end{aligned}$$

Here $\|\cdot\|_2$ denotes the spectral norm. Therefore, for all $t \geq 0$, we have

$$\begin{aligned}\|X(t)\|_2^4 + \|Y(t)\|_2^4 &\leq \|X(t)^T X(t) - Y(t)^T Y(t)\|_F^2 \\ &\quad + 2mn(\|X(t)Y(t)^T - M\|_1 + \|M\|_1)^2 \\ &\leq \|X_0^T X_0 - Y_0^T Y_0\|_F^2 \\ &\quad + 2mn(\|X_0 Y_0^T - M\|_1 + \|M\|_1)^2.\end{aligned}$$

□

2.3.4 Nonnegative matrix factorization

Let $A \in \mathbb{R}^{m \times n}$ and $p \geq 1$, we define the p -norm of A by

$$\|A\|_p := \left(\sum_{i=1}^m \sum_{j=1}^n |A_{ij}|^p \right)^{1/p}.$$

Let $M \in \mathbb{R}^{m \times n}$, in nonnegative ℓ_p matrix factorization, we aim to minimize

$$\begin{aligned} f : \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r} &\longrightarrow \mathbb{R} \\ (X, Y) &\longmapsto \|XY^T - M\|_p^p, \end{aligned}$$

subject to $(X, Y) \in \mathbb{R}_+^{m \times r} \times \mathbb{R}_+^{n \times r} =: C$. Without the nonnegativity constraints, ℓ_p matrix factorization was studied in [70], and was shown to be robust against outliers when $p < 2$. Note that when $p = 2$, the above example reduces to the problem of nonnegative matrix factorization (NMF) [71, 72, 18, 73].

Next we verify the boundedness of the subgradient trajectories. We begin with some notations. Let $A, B \in \mathbb{R}^{m \times n}$, we denote by $A \odot B \in \mathbb{R}^{m \times n}$ their Hadamard product, whose (i, j) -entry is given by $(A \odot B)_{ij} := A_{ij}B_{ij}$. Let $p \geq 0$, we denote by $|A|^{\circ p} \in \mathbb{R}^{m \times n}$ the matrix obtained by taking absolute value and then raising to p th power for each element in A , namely, $(|A|^{\circ p})_{ij} := |A_{ij}|^p$. We use the convention that $0^0 = 1$. Let sign denote the element-wise operation that maps each entry of a matrix to a subset of $[-1, 1]$ such that

$$\text{sign}(t) := \begin{cases} -1 & \text{if } t < 0, \\ [-1, 1] & \text{if } t = 0, \\ 1 & \text{if } t > 0. \end{cases}$$

By [40, 2.3.10 Theorem (Chain Rule II)], we have

$$\partial f(X, Y) = \left\{ \left(\begin{array}{c} \left(\Lambda \odot |XY^T - M|^{\circ(p-1)} \right) Y \\ \left(\Lambda \odot |XY^T - M|^{\circ(p-1)} \right)^T X \end{array} \right) : \Lambda \in \text{sign}(XY^T - M) \right\}.$$

We next study the solutions to (2.8), which is an equivalent characterization of the subgradient trajectories of Φ by [74, Theorem 2.3(b)].

Lemma 2.1. *Given $X_0 \in \mathbb{R}_+^{m \times r}$, $Y_0 \in \mathbb{R}_+^{n \times r}$, and $M \in \mathbb{R}^{m \times n}$, there exist $c_1, \dots, c_r \in \mathbb{R}$ such that any solution $(X, Y, \Lambda) : \mathbb{R}_+ \rightarrow \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r} \times \mathbb{R}^{m \times n}$ to*

$$\begin{cases} X' = P_{T_{\mathbb{R}_+^{m \times r}}(X)} \left(- \left(\Lambda \odot |XY^T - M|^{\circ(p-1)} \right) Y \right) \\ Y' = P_{T_{\mathbb{R}_+^{n \times r}}(Y)} \left(- \left(\Lambda \odot |XY^T - M|^{\circ(p-1)} \right)^T X \right) \\ \Lambda \in \text{sign}(XY^T - M), X(0) = X_0, Y(0) = Y_0 \end{cases} \quad (2.8)$$

satisfies that

$$\sum_{i=1}^m X_{ik}(t)^2 - \sum_{j=1}^n Y_{jk}(t)^2 = c_k, \quad \forall t \in \mathbb{R}_+, \quad \forall k \in \llbracket 1, r \rrbracket.$$

Proof. For all $t \in \mathbb{R}_+$, let

$$\begin{aligned} L(t) &:= X'(t)^T X(t) + X(t)^T X'(t), \quad R(t) := Y'(t)^T Y(t) + Y^T(t) Y'(t) \\ &\text{and } E(t) := -\Lambda(t) \odot |X(t)Y(t)^T - M|^{\circ(p-1)}. \end{aligned}$$

For $k \in \llbracket 1, r \rrbracket$ and $t \in \mathbb{R}_+$, define the following index sets

$$I_k^X(t) := \left\{ i \in \llbracket 1, m \rrbracket : X'_{ik}(t) \neq \sum_{j=1}^n E_{ij}(t) Y_{jk}(t) \right\}$$

and

$$I_k^Y(t) := \left\{ j \in \llbracket 1, n \rrbracket : Y'_{jk}(t) \neq \sum_{i=1}^m E_{ij}(t) X_{ik}(t) \right\}.$$

Consider the k -th diagonal element of $L(t)$ for $k \in \llbracket 1, r \rrbracket$, we have that

$$L_{kk}(t) = 2 \sum_{i=1}^m X_{ik}(t) X'_{ik}(t) = 2 \sum_{i \in I_k^X(t)} X_{ik}(t) X'_{ik}(t) + 2 \sum_{i \notin I_k^X(t)} X_{ik}(t) X'_{ik}(t).$$

Notice that if $i \in I_k^X(t)$, then $X'_{ik}(t) = 0$. Thus

$$L_{kk}(t) = 2 \sum_{i \notin I_k^X(t)} X_{ik}(t) \sum_{j=1}^n E_{ij}(t) Y_{jk}(t) = 2 \sum_{i \notin I_k^X(t)} \sum_{j=1}^n E_{ij}(t) X_{ik}(t) Y_{jk}(t).$$

Furthermore, notice that if $j \in I_k^Y(t)$, then $Y_{jk}(t) = 0$ and

$$L_{kk}(t) = 2 \sum_{i \notin I_k^X(t)} \sum_{j \notin I_k^Y(t)} E_{ij}(t) X_{ik}(t) Y_{jk}(t).$$

Similarly, consider the k -th diagonal element of $R(t)$ for $k \in \llbracket 1, r \rrbracket$, one has

$$R_{kk}(t) = 2 \sum_{j=1}^n Y_{jk}(t) Y'_{jk}(t) = 2 \sum_{j \in I_k^Y(t)} Y_{jk}(t) Y'_{jk}(t) + 2 \sum_{j \notin I_k^Y(t)} Y_{jk}(t) Y'_{jk}(t).$$

Notice that if $j \in I_k^Y(t)$, then $Y'_{jk}(t) = 0$. Thus,

$$R_{kk}(t) = 2 \sum_{j \notin I_k^Y(t)} Y_{jk}(t) \sum_{i=1}^m E_{ij}(t) X_{ik}(t) = 2 \sum_{j \notin I_k^Y(t)} \sum_{i=1}^m E_{ij}(t) X_{ik}(t) Y_{jk}(t).$$

Moreover, if $i \in I_k^X(t)$, then $X_{ik}(t) = 0$, thus

$$R_{kk}(t) = 2 \sum_{j \notin I_k^Y(t)} \sum_{i \notin I_k^X(t)} E_{ij}(t) X_{ik}(t) Y_{jk}(t) = 2 \sum_{i \notin I_k^X(t)} \sum_{j \notin I_k^Y(t)} E_{ij}(t) X_{ik}(t) Y_{jk}(t) = L_{kk}(t)$$

by exchanging the order of two finite sums. Finally, for all $k \in \llbracket 1, r \rrbracket$ and $t \in \mathbb{R}_+$,

$$\frac{d}{dt} \left(\sum_{i=1}^m X_{ik}(t)^2 - \sum_{j=1}^n Y_{jk}(t)^2 \right) = L_{kk}(t) - R_{kk}(t) = 0.$$

Therefore, for all $k \in \llbracket 1, r \rrbracket$ and $t \in \mathbb{R}_+$,

$$\sum_{i=1}^m X_{ik}(t)^2 - \sum_{j=1}^n Y_{jk}(t)^2 = c_k := \sum_{i=1}^m X_{ik}(0)^2 - \sum_{j=1}^n Y_{jk}(0)^2,$$

where c_k 's are constants independent of t . □

Using lemma 2.1, we next prove that the products of the entries of X and Y in (2.8) remain bounded throughout time.

Lemma 2.2. *Given $X_0 \in \mathbb{R}_+^{m \times r}$, $Y_0 \in \mathbb{R}_+^{n \times r}$, and $M \in \mathbb{R}^{m \times n}$, there exists $d > 0$ such that any solution $(X, Y) : \mathbb{R}_+ \rightarrow \mathbb{R}_+^{m \times r} \times \mathbb{R}_+^{n \times r}$ to (2.8) satisfies that for every $i \in \llbracket 1, m \rrbracket$ and $j \in \llbracket 1, n \rrbracket$,*

$$|X_{ik}(t)Y_{jk}(t)| \leq d, \quad \forall t \in \mathbb{R}_+, \quad \forall k \in \llbracket 1, r \rrbracket.$$

Proof. Note that any solution to (2.8) is a subgradient trajectory of $f + \delta_{\mathbb{R}_+^{m \times r} \times \mathbb{R}_+^{n \times r}}$ [74, Theorem 2.3(b)]. By [55, Corollary 5.4] and [22, Lemma 6.3], we have that $t \mapsto f(X(t), Y(t))$ is a decreasing function over \mathbb{R}_+ . Thus, we have that for all $t \in \mathbb{R}_+$,

$$\begin{aligned} \|X(t)Y(t)^T\|_p &\leq \|X(t)Y(t)^T - M\|_p + \|M\|_p \\ &\leq \|X_0Y_0^T - M\|_p + \|M\|_p =: \hat{d} \end{aligned}$$

where $\hat{d} > 0$ is a constant. By the equivalence of norms, there is a constant $\hat{c}_p > 0$ such that

$$\|X(t)Y(t)^T\|_1 \leq \hat{c}_p \|X(t)Y(t)^T\|_p \leq \hat{c}_p \hat{d}$$

This implies that for all $t \in \mathbb{R}_+$,

$$\left| \sum_{k=1}^r X_{ik}(t)Y_{jk}(t) \right| = |[X(t)Y(t)^T]_{ij}| \leq \|X(t)Y(t)^T\|_1 \leq \hat{c}_p \hat{d}, \quad \forall i \in \llbracket 1, m \rrbracket, j \in \llbracket 1, n \rrbracket.$$

Notice that $X \in \mathbb{R}_+^{m \times r}$ and $Y \in \mathbb{R}_+^{n \times r}$, Thus, we have for all $t \in \mathbb{R}_+$,

$$\sum_{k=1}^r |X_{ik}(t)Y_{jk}(t)| = \left| \sum_{k=1}^r X_{ik}(t)Y_{jk}(t) \right| \leq \hat{c}_p \hat{d},$$

and the desired result follows immediately by setting $d := \hat{c}_p \hat{d}$. \square

Finally, by combining Lemma 2.1 and Lemma 2.2, we can obtain the desired result that ℓ_p -nonnegative matrix factorization has bounded subgradient trajectories.

Proposition 2.5. *Let $p \geq 1$ and let $f : \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r} \rightarrow \mathbb{R}$ be defined by $f(X, Y) := \frac{1}{p} \|XY^T - M\|_p^p$. Let $\Phi := f + \delta_C$ where $C := \mathbb{R}_+^{m \times r} \times \mathbb{R}_+^{n \times r}$. Then for any $(X_0, Y_0) \in \text{dom } \Phi$, there exists a unique subgradient trajectory of Φ initialized at (X_0, Y_0) , and this subgradient trajectory is bounded.*

Proof. For any $(X_0, Y_0) \in \text{dom } \Phi$, existence of a subgradient trajectory initialized at (X_0, Y_0) is a result of [74, Theorem 3.1]. As Φ is primal lower nice [75, Definition 1.1] at every point in $\text{dom } \Phi$, such a subgradient trajectory must be unique [48, Theorem 2.9]. We next show that this subgradient trajectory is also bounded. Recall that every subgradient trajectory is a solution to (2.8) by [74, Theorem 2.3(b)]. Thus, it suffices to show that every solution to (2.8) is bounded. From lemmas 2.1 and 2.2, there exist constants c_k 's for any $k \in \llbracket 1, r \rrbracket$ and $d > 0$ such that for any solution $(X(\cdot), Y(\cdot))$ to (2.8), we have

$$\begin{aligned} \left(\sum_{i=1}^m X_{ik}(t)^2 + \sum_{j=1}^n Y_{jk}(t)^2 \right)^2 &= \left(\sum_{i=1}^m X_{ik}(t)^2 - \sum_{j=1}^n Y_{jk}(t)^2 \right)^2 + 4 \sum_{i=1}^m X_{ik}(t)^2 \sum_{j=1}^n Y_{jk}(t)^2 \\ &\leq c_k^2 + 4mnd^2. \end{aligned}$$

for any $t \in \mathbb{R}_+$. Thus, for all $k \in \llbracket 1, r \rrbracket$,

$$\sum_{i=1}^m X_{ik}(t)^2 + \sum_{j=1}^n Y_{jk}(t)^2 \leq \sqrt{4mnd^2 + c_k^2}.$$

Summing both sides up over k yields

$$\|X(t)\|_2^2 + \|Y(t)\|_2^2 \leq \sum_{k=1}^r \sqrt{4mnd^2 + c_k^2}.$$

Hence, every solution to (2.8) is bounded. \square

2.3.5 Linear neural network

Consider minimizing the loss function of linear neural network without bias term

$$f(W_1, \dots, W_L) := \frac{1}{2} \|W_L \cdots W_1 X - Y\|_F^2, \quad (2.11)$$

where $X \in \mathbb{R}^{d_0 \times d_x}$, $Y \in \mathbb{R}^{d_L \times d_x}$, and $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$ for $i = 1, \dots, L$. Here $\|\cdot\|_F$ denotes the Frobenius norm.

Proposition 2.6. *Linear neural network with loss function (2.11) has bounded gradient trajectories.*

An existing proof of Proposition 2.6 under additional assumptions on network structure, initialization, input data, or target data can be found, for instance, in [76, 77, 78]. To the best of our knowledge, the closest result to Proposition 2.6 is [76, Theorem 3.2], which shows that gradient trajectories are bounded if XX^T is of full rank. In the proof of Proposition 2.6, we show that this rank assumption on X can be removed and hence Proposition 2.6 applies to any linear neural network.

Proof of Proposition 2.6. Since f is locally Lipschitz and lower bounded, by Proposition 1.2 there exists a gradient trajectory for any initial point. By [76, Lemma 2.1], the gradient trajectories of f satisfy the initial value problem

$$\dot{W}_i = -(W_L \cdots W_{i+1})^T (W_L \cdots W_1 X - Y) (W_{i-1} \cdots W_1 X)^T, \quad (2.12a)$$

$$W_i(0) = W_i^0, \quad W_i^0 \in \mathbb{R}^{d_i \times d_{i-1}} \text{ is a given constant matrix,} \quad (2.12b)$$

for all $i = 1, \dots, L$. Note that if $i = L$, (2.12a) reduces to

$$\dot{W}_L = -(W_L \cdots W_1 X - Y)(W_{L-1} \cdots W_1 X)^T,$$

and if $i = 1$, (2.12a) reduces to

$$\dot{W}_1 = -(W_L \cdots W_2)^T (W_L \cdots W_1 X - Y) X^T.$$

Note that [76, Theorem 3.2] proved the boundedness of gradient trajectories of f when XX^T is invertible. Thus, we only need to show we can always reduce the boundedness of gradient trajectories of f for general X to the boundedness of gradient trajectories of another function g in the same form as f but with invertible XX^T . Let $X = U\Sigma V^T$ be a singular value decomposition, where $U \in \mathbb{R}^{d_0 \times d_0}$ and $V \in \mathbb{R}^{d_x \times d_x}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{d_0 \times d_x}$ is a rectangular matrix satisfying

$$\Sigma = \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix}, \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_r) \succ 0,$$

where $r \leq \min\{d_0, d_x\}$. Eliminating X in (2.12a), it reduces to

$$\begin{aligned} \dot{W}_i &= -(W_L \cdots W_{i+1})^T (W_L \cdots W_1 U \Sigma V^T - Y) (W_{i-1} \cdots W_1 U \Sigma V^T)^T \\ &= -(W_L \cdots W_{i+1})^T (W_L \cdots W_1 U \Sigma - YV) (W_{i-1} \cdots W_1 U \Sigma)^T. \end{aligned}$$

Define $Z := YV \in \mathbb{R}^{d_L \times d_x}$, and (2.12) reduces to

$$\dot{W}_i = -(W_L \cdots W_{i+1})^T (W_L \cdots W_1 U \Sigma - Z) (W_{i-1} \cdots W_1 U \Sigma)^T, \quad (2.13a)$$

$$W_i(0) = W_i^0, \quad \forall i = 1, \dots, L. \quad (2.13b)$$

Denote $\bar{W}_1 := W_1 U \in \mathbb{R}^{d_1 \times d_0}$ and $\bar{W}_1^0 := W_1^0 U \in \mathbb{R}^{d_1 \times d_0}$. To keep the notation consistent, also let

$\bar{W}_i := W_i$ and $\bar{W}_i^0 := W_i^0$ for $i = 2, \dots, L$. Thus, (2.13) reduces to

$$\dot{\bar{W}}_i = -(\bar{W}_L \cdots \bar{W}_{i+1})^T (\bar{W}_L \cdots \bar{W}_1 \Sigma - Z) (\bar{W}_{i-1} \cdots \bar{W}_1 \Sigma)^T, \quad (2.14a)$$

$$\bar{W}_i(0) = \bar{W}_i^0, \quad \forall i = 1, \dots, L. \quad (2.14b)$$

Partition the matrices \bar{W}_1 , \bar{W}_1^0 , and Z into two column blocks:

$$\bar{W}_1 = \begin{bmatrix} \bar{W}_{11} & \bar{W}_{12} \end{bmatrix}, \quad \bar{W}_1^0 = \begin{bmatrix} \bar{W}_{11}^0 & \bar{W}_{12}^0 \end{bmatrix}, \quad Z = \begin{bmatrix} Z_1 & Z_2 \end{bmatrix},$$

where \bar{W}_{11} , \bar{W}_{11}^0 , and Z_1 consist of the first r columns of \bar{W}_1 , \bar{W}_1^0 and Z respectively. Thus, when $i = 1$, (2.14) can be reduced into

$$\dot{\bar{W}}_{11} = -(\bar{W}_L \cdots \bar{W}_2)^T (\bar{W}_L \cdots \bar{W}_2 \bar{W}_{11} \Lambda - Z_1) \Lambda^T, \quad \dot{\bar{W}}_{12} = 0,$$

$$\bar{W}_{11}(0) = \bar{W}_{11}^0, \quad \bar{W}_{12}(0) = \bar{W}_{12}^0.$$

When $i = 2, \dots, L$, (2.14) can be reduced into

$$\dot{\bar{W}}_i = -(\bar{W}_L \cdots \bar{W}_{i+1})^T (\bar{W}_L \cdots \bar{W}_2 \bar{W}_{11} \Lambda - Z_1) (\bar{W}_{i-1} \cdots \bar{W}_2 \bar{W}_{11} \Lambda)^T,$$

$$\bar{W}_i(0) = \bar{W}_i^0.$$

It indicates that $\bar{W}_{12}(t) = \bar{W}_{12}^0$ for all $t \geq 0$. Denote $\tilde{W}_1 := \bar{W}_{11}$ and $\tilde{W}_1^0 := \bar{W}_{11}^0$. To keep the notation consistent, also let $\tilde{W}_i := \bar{W}_i$ and $\tilde{W}_i^0 := \bar{W}_i^0$ for $i = 2, \dots, L$. Therefore, (2.14) reduces to

$$\dot{\tilde{W}}_i = -(\tilde{W}_L \cdots \tilde{W}_{i+1})^T (\tilde{W}_L \cdots \tilde{W}_1 \Lambda - Z_1) (\tilde{W}_{i-1} \cdots \tilde{W}_1 \Lambda)^T, \quad (2.15a)$$

$$\tilde{W}_i(0) = \tilde{W}_i^0, \quad \forall i = 1, \dots, L. \quad (2.15b)$$

Define the new function g as

$$g(\tilde{W}_1, \dots, \tilde{W}_L) := \frac{1}{2} \|\tilde{W}_L \cdots \tilde{W}_1 \Lambda - Z_1\|_F^2.$$

Notice that the gradient trajectories of g satisfy (2.15). To prove f has bounded gradient trajectories, it is equivalent to prove g has bounded gradient trajectories, because $\|W_1\|_F = \|W_1 U\|_F = \|\bar{W}_1\|_F$ and $\|\bar{W}_1(t)\|_F^2 = \|\tilde{W}_1(t)\|_F^2 + \|\bar{W}_{12}^0\|_F^2$ for all $t \geq 0$. Since $\Lambda \Lambda^T$ is invertible, by [76, Theorem 3.2], g has bounded gradient trajectories, and so does f . \square

2.3.6 One dimensional deep sigmoid neural network

Though famous for its benign theoretical properties, linear neural network is rarely used in practice because of its low representation power. We want to take a step further in the case of non-linear deep neural network. In this subsection, we focus on neural network with sigmoid activation function in one dimensional case.

Consider minimizing the following loss function of sigmoid neural network

$$f(w_1, \dots, w_L) := \frac{1}{2} (w_L \sigma(w_{L-1} \cdots \sigma(w_1 x)) - y)^2, \quad (2.16)$$

where $\sigma(z) := (1 + e^{-z})^{-1}$ is the sigmoid function and $w_i, x, y \in \mathbb{R}$ for all $i = 1, \dots, L$.

Notice that the techniques in the proof of Proposition 2.6 cannot be adapted to this case because the auto-balancing property in [69, Theorem 2.1] does not hold. Surprisingly, it is still true that (2.16) has bounded gradient trajectories.

Proposition 2.7. *One dimensional sigmoid neural network with loss function (2.16) has bounded gradient trajectories.*

Proof. Since f is locally Lipschitz and lower bounded, by Proposition 1.2 there exists a gradient trajectory for any initial point. For simplicity, define p_i for $i = 0, \dots, L-2$ recursively by $p_0 := x$,

$p_1 := \sigma(w_1x)$ and $p_{i+1} := \sigma(w_{i+1}p_i)$. The gradient trajectories of f satisfy

$$\dot{w}_L = -(w_L\sigma(w_{L-1}p_{L-2}) - y)\sigma(w_{L-1}p_{L-2}), \quad (2.17a)$$

$$\dot{w}_i = \frac{p_{i-1}}{1 + e^{w_i p_{i-1}}} \dot{w}_{i+1} w_{i+1}, \quad i = L - 1, \dots, 1. \quad (2.17b)$$

We will prove each w_i is bounded inductively from the last layer to the first layer. The relation between the last two layers w_L and w_{L-1} , and the relation between the first two layers can be regarded as the base cases.

We claim that there exists a time T such that \dot{w}_i and w_i does not change sign for all $t \geq T$ and for all i . To verify this, first notice that the claim is true for the last layer, i.e., \dot{w}_L and w_L will not change sign for all $t \geq T$. Suppose \dot{w}_L changes sign, by continuity and mean value theorem, there exists $t^* > 0$ such that $\dot{w}_L(t^*) = 0$. However, $\dot{w}_L(t^*) = 0$ implies $\dot{w}_i(t^*) = 0$ for all i , meaning that a critical point is achieved and the gradient trajectory is stopped for all $t \geq t^*$. In this case, all w_i 's are trivially bounded. Thus, we assume the trajectory will never stop at a finite time. In this case, either $\dot{w}_L(t) > 0$ or $\dot{w}_L(t) < 0$ for all $t \geq 0$. Since w_L is monotonic, it either keeps the sign unchanged or changes the sign only once. Thus, there exists $T_L > 0$ such that w_L does not change sign on $[T_L, \infty)$. Notice that for all $i \geq 2$, $p_{i-1}(t) \in (0, 1)$ for all $t \geq 0$. Since $\dot{w}_L w_L$ does not change sign on $[T_L, \infty)$, (2.17b) implies that \dot{w}_{L-1} does not change sign on $[T_L, \infty)$ either. Therefore, we conclude that w_{L-1} is monotonic. Similarly, there exists $T_{L-1} > T_L$ such that \dot{w}_{L-1} and w_{L-1} does not change sign on $[T_{L-1}, \infty)$. Recursively using the above argument, we can show the claim is true for all $i \geq 2$ on $[T_2, \infty)$. For $i = 1$, although $p_0 = x$ may not be in $(0, 1)$, since x is a constant, the fact that \dot{w}_2 and w_2 do not change sign still implies that \dot{w}_1 does not change sign and hence there exists $T_1 > T_2$ such that w_1 does not change sign on $[T_1, \infty)$. Therefore, the claim holds for $i = 1, \dots, L$ by choosing $T = T_1$.

By the claim proved in the last paragraph, for $i = 1, \dots, L$, either $\dot{w}_i w_i$ is nonnegative or $\dot{w}_i w_i$ is negative on $[T, \infty)$. Now we are going to prove each w_i is bounded. The first step is to prove the last two layers w_L and w_{L-1} are bounded. Consider the case where $\dot{w}_L w_L$ is nonnegative on

$[T, \infty)$. (2.17b) implies that $\dot{w}_{L-1} \geq 0$ and w_{L-1} is increasing over $[T, \infty)$, so there exists a constant c_{L-1} such that $w_{L-1}(t) \geq c_{L-1}$ for all $t \geq 0$. Since $p_{L-2} \in (0, 1)$, we have $\sigma(w_{L-1}p_{L-2}) \geq \sigma(-|c_{L-1}|) > 0$. Again, by [22, Lemma 5.2], $\frac{d}{dt}f(w_1, \dots, w_L) \leq 0$ and $f(w_1, \dots, w_L) \leq C$ for some constant C on $[0, \infty)$. Thus, it is easy to see $|w_L|\sigma(w_{L-1}p_{L-2}) \leq C_1$ for some constant C_1 on $[0, \infty)$. Since $\sigma(w_{L-1}p_{L-2}) \in [\sigma(-|c_{L-1}|), 1)$, we conclude $|w_L|$ is bounded. Suppose w_{L-1} is unbounded. Since it is increasing and does not change sign, $w_{L-1}(t) > 0$ for all $t \geq T$ and $w_{L-1}(t) \rightarrow \infty$ as $t \rightarrow \infty$. By (2.17b),

$$\dot{w}_{L-1} = \frac{p_{L-2}}{1 + e^{w_L p_{L-2}}} \dot{w}_L w_L \leq \dot{w}_L w_L, \quad (2.18)$$

because $p_{L-2} \in (0, 1)$ and $1 + e^{w_L p_{L-2}} > 1$. By (2.18), $w_{L-1} - \frac{1}{2}w_L^2$ is a decreasing function on $[T, \infty)$. Hence, $w_{L-1} - \frac{1}{2}w_L^2 \leq C_2$ for some constant C_2 . Notice that w_L is bounded but $w_{L-1}(t) \rightarrow \infty$ as $t \rightarrow \infty$, so a contradiction occurs. Therefore, w_{L-1} is bounded.

Now we consider the case where $\dot{w}_L w_L$ is negative on $[T, \infty)$. In this case, (2.17b) implies $\dot{w}_{L-1} \leq 0$, so w_{L-1} is decreasing on $[T, \infty)$ and there exists a constant d_{L-1} such that $w_{L-1} \leq d_{L-1}$. Since $p_{L-2}/(1 + e^{w_L p_{L-2}}) \in (0, 1)$ and $\dot{w}_L w_L \leq 0$ on $[T, \infty)$, we have $\dot{w}_{L-1} \geq \dot{w}_L w_L$. This shows $w_{L-1} - \frac{1}{2}w_L^2$ is increasing on $[T, \infty)$, and hence $w_{L-1} \geq \tilde{d}_{L-1}$ for some constant \tilde{d}_{L-1} . Therefore, $w_{L-1} \in [\tilde{d}_{L-1}, d_{L-1}]$ is bounded. By exactly the same argument as in the case when $\dot{w}_L w_L$ is nonnegative, we know $\sigma(w_{L-1}p_{L-2}) \in [\sigma(-|\tilde{d}_{L-1}|), 1)$ and w_L is bounded by using the boundedness of objective function f .

Up to now, we have proved boundedness for the last two layers w_L and w_{L-1} . For $i = 2, \dots, L-2$, by discussing two cases $\dot{w}_{i+1}w_{i+1} \geq 0$ and $\dot{w}_{i+1}w_{i+1} \leq 0$, together with the boundedness of w_{i+1} , we can prove that w_i is bounded by exactly the same argument as we did in the last two paragraphs. The induction starts with proving w_{L-2} is bounded and ends with proving w_2 is bounded. Once we prove w_2 is bounded, consider the relation between w_1 and w_2 ,

$$(1 + e^{w_1 x})\dot{w}_1 = x\dot{w}_2 w_2.$$

If $x = 0$, then $\dot{w}_1 = 0$ implies w_1 is a constant over $[0, \infty)$, so it must be bounded. Suppose $x \neq 0$, by taking integration with respect to t and multiplying x on both sides, we have

$$w_1 x + e^{w_1 x} = \frac{x^2}{2} w_2^2 + C_3.$$

Let $z = w_1 x$, then $z + e^z \rightarrow \pm\infty$ as $z \rightarrow \pm\infty$. Thus, the boundedness of w_2 implies the boundedness of $z = w_1 x$. Since $x \neq 0$ is a constant, w_1 is bounded. Therefore, we proved that w_i is bounded for all $i = 1, \dots, L$. □

Chapter 3: Landscape at infinity

Having explored a variety of examples that exhibit bounded subgradient trajectories in Chapter 2, we now turn to the question of how such property can be leveraged to analyze the optimization landscape and the global behavior of first-order algorithms. As a first step, this chapter focuses on the simplest setting, studying the landscape of the objective function itself. The results presented in this chapter are based on the following article:

C. Jozs and X. Li, “Certifying the absence of spurious local minima at infinity,” *SIAM Journal on Optimization*, vol. 33, pp. 1416–1439, 3 2023

The idea that the absence of spurious local minima alone does not guarantee the success of first-order methods was first expressed in the context of binary classification in the mid-nineties. It was shown that gradient trajectories are bounded if the objective function satisfies several technical conditions tailored to the problem at hand [80, Theorems 3.6-3.8]. This property was referred to as having no attractors at infinity. More recently, it was proved that adding an exponential neuron to a wide class of neural networks eliminates all spurious local minima [81], but it was soon realized that this procedure simply sends them to infinity [82]. These results suggest that besides spurious local minima, a certain notion of spurious local minima at infinity also affects the convergence of first-order methods to global optima. However, the current optimization literature lacks a precise definition of local minima at infinity, and, accordingly, there is little theoretical understanding of them. Worse still, classical tools for landscape analysis, such as the gradient and the Hessian, cannot detect spurious local minima at infinity even in simple scenarios (recall Example 3.1), let alone handle nonsmooth functions without a gradient.

Example 3.1. Consider an instance of matrix completion problem, i.e., minimize

$$f(x_1, x_2, y_1, y_2) := (x_1 y_1 - 1)^2 + (x_2 y_1 - 1)^2 + (x_2 y_2 - 1)^2.$$

By solving $\nabla f(x_1, x_2, y_1, y_2) = 0$, the set of critical points of f can be decomposed into four connected components:

$$C_1 = \{(x_1, x_2, y_1, y_2) = (t, t, 1/t, 1/t) \mid t \in \mathbb{R} \setminus \{0\}\},$$

$$C_2 = \{(x_1, x_2, y_1, y_2) = (t, 0, 1/t, -1/t) \mid t \in \mathbb{R} \setminus \{0\}\},$$

$$C_3 = \{(x_1, x_2, y_1, y_2) = (t, -t, 0, -1/t) \mid t \in \mathbb{R} \setminus \{0\}\},$$

$$C_4 = \{(x_1, x_2, y_1, y_2) = (0, 0, 0, 0)\}.$$

The critical values are $f(C_1) = \{0\}$, $f(C_2) = f(C_3) = \{2\}$, and $f(C_4) = \{3\}$. Furthermore, C_1 is the set of global minima, and by computing the Hessian $\nabla^2 f$, we find that it has positive and negative eigenvalues at all points in C_2 , C_3 and C_4 . Therefore, f has no spurious local minima and all saddle points are strict [83, Definition 2]. One would expect first-order methods like gradient descent to converge to a global minimum for almost all initial points [83, Theorem 11]. However, the numerical experiments in Figure 3.1 show otherwise. This is because the function is not coercive.

Two newly proposed concepts related to spurious local minima at infinity are setwise local minima [84] and spurious valleys [85]. Setwise local minima [84, Definition 2.5] generalize the notion of local minima from points to compact sets. It was recently established that the uniform limit (on all compact subsets) of a sequence of continuous functions which are devoid of spurious setwise local minima is itself devoid of spurious strict setwise local minima [84, Proposition 2.7]. However, due to the boundedness assumption, setwise local minima cannot be directly used to study spurious local minima at infinity. Spurious valleys [85, Definition 1] do have the potential to handle spurious local minima at infinity but they fail to detect them when there are flat regions, such

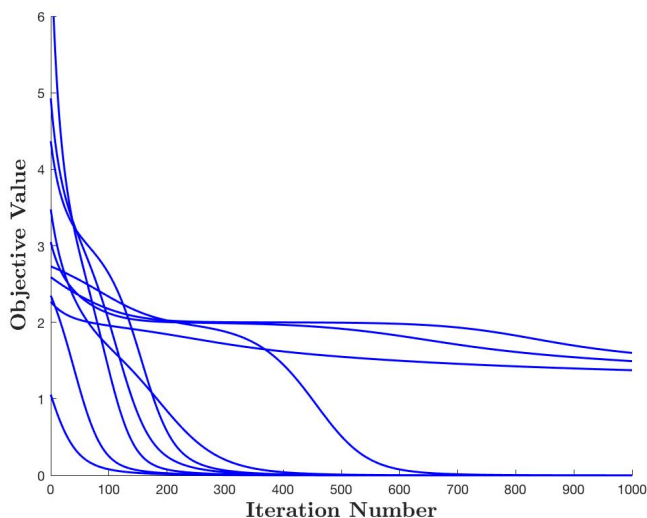


Figure 3.1: Gradient method initialized uniformly at random in $[-1, 1]^4$ with constant step size 0.01 sometimes gets stuck at a spurious local minimum at infinity (3 among 10 trials in the experiment).

as in the ReLU network with one-hidden layer $(x_1, x_2) \mapsto (x_2 \max\{x_1, 0\} - 1)^2$ (see Figure 3.2). Spurious valleys also rely on the notion of path-connectedness, which is actually not necessary for defining spurious local minimum at infinity. In this chapter, we extend the concept of setwise local minima by relaxing the boundedness assumption. This enables us to define spurious local minima at infinity as unbounded setwise local minima over which the infimum of the objective function is greater than the global infimum. It also allows us to handle classical spurious local minima and flat regions in a unified way.

An existing strategy to analyze the landscape of non-coercive functions is to construct a strictly decreasing path to a global minimum from any initial point. Such a path was shown to exist in half-rectified neural network [86]. This strategy is used to prove the existence of spurious local minima in neural networks with almost all nonlinear activations [87]. It also explains the phase transition from the existence of sub-optimal basins in narrow networks to their disappearance in wide networks [88]. Finally, it is used to prove the absence of spurious valleys for over-parametrized one-hidden layer neural network [85]. However, such a strategy needs to be tailored to each application since one needs to select a particular path for each specific loss function. In this chapter, we

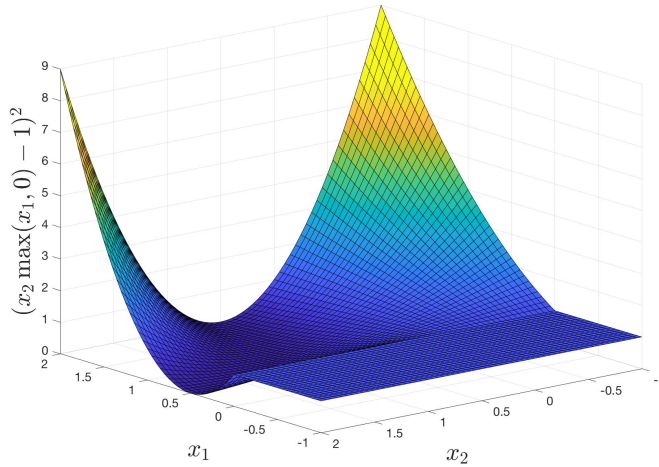


Figure 3.2: Function devoid of spurious valleys containing a spurious local minimum at infinity.

instead develop a theory allowing one to use a common decreasing path - subgradient trajectory - to analyze the landscape in various different contexts. We can then rule out spurious setwise local minima (and thus those at infinity) for a general class of functions. Our main result is as follows.

Theorem 3.1. *Suppose a locally Lipschitz function is bounded below, admits a chain rule, has finitely many critical values, and has bounded subgradient trajectories. Then it has no spurious local minima if and only if it has no spurious setwise local minima.*

The above statement is meant to help readers get a first taste of our main result in this chapter. For the terminology in the theorem, see Definition 1.3 (locally Lipschitz), Definition 1.8 (chain rule), Definition 1.7 (bounded subgradient trajectories), and Definition 3.2 (setwise local minimum). A discussion on the role of its assumptions is also given following the proof of Theorem 3.1 in Section 3.2. Let us mention already that two of its assumptions, namely those regarding the chain rule and critical values, automatically hold for functions definable in an o-minimal expansion of the real field [36] (by [57, Proposition 2 (iv)] and the definable Morse-Sard theorem [43, Corollary 9 (ii)]). This includes semi-algebraic, globally subanalytic, and log-exp functions, and importantly, many applications of interest nowadays [57, Section 4.1]. The locally Lipschitz and lower bounded assumptions usually come for free in applications, so that in practice the sole assumption that one needs to check for is that subgradient trajectories are bounded. Theorem 3.1

thus serves as a handy device to conclude that there are no spurious setwise local minima for a family of functions that are widely used in machine learning, especially in deep neural networks and matrix sensing. We summarize the problems that we are going to consider in the following corollary.

Corollary 3.1. *The following problems have no spurious local minima at infinity:*

1. *deep linear neural network*

$$\inf_{W_1, \dots, W_L} \|W_L \cdots W_1 X - Y\|_F^2;$$

2. *one dimensional deep neural network with sigmoid activation function σ*

$$\inf_{w_1, \dots, w_L} (w_L \sigma(w_{L-1} \cdots \sigma(w_1 x)) - y)^2;$$

3. *matrix recovery with restricted isometry property (RIP)*

$$\inf_{X, Y} \frac{1}{2m} \sum_{i=1}^m (\langle A_i, XY^T \rangle_F - b_i)^2;$$

4. *nonsmooth matrix factorization where $\text{rank}(M) = 1$ and $M_{ij} \neq 0$*

$$\inf_{x, y} \sum_{i=1}^m \sum_{j=1}^n |x_i y_j - M_{ij}|.$$

Again, the statement above aims at giving readers some feeling on what type of functions we are considering. More rigorous descriptions of the applications will be given in Section 3.3.

This chapter is organized as follows. Section 3.1 contains background material on setwise local minima. Section 3.2 contains the proof of our main result, namely Theorem 3.1. Finally, Section 3.3 contains applications of our main result as delineated in Corollary 3.1.

3.1 Setwise local minimum

In this section, we present the formal definitions and some useful properties of setwise local minimum and local minimum at infinity. We first review the classical definition of local and global minima.

Definition 3.1. A point $x \in \mathbb{R}^n$ is a *local minimum* (respectively, *global minimum*) of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ if $f(x) \leq f(y)$ for all $y \in \mathring{B}(x, \epsilon)$ for some $\epsilon > 0$ (respectively, $y \in \mathbb{R}^n$). A local minimum is *spurious* if it is not a global minimum.

From Definition 3.1, one can see the definition of a local minimum only considers the landscape of a function at any finite point. To discuss the function landscape at infinity, we generalize the notion of setwise local minimum first proposed in [84, Definition 2.5].

Definition 3.2 (Setwise local minimum). A nonempty closed subset $S \subset \mathbb{R}^n$ is a *setwise local minimum* of a continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ if there exists an open set $U \subset \mathbb{R}^n$ such that $S \subset U$ and $f(x) \leq f(y)$ for all $x \in S, y \in U \setminus S$.

It is easy to see that a local minimum is a setwise local minimum by taking S to be a singleton. We also define a *strict setwise local minimum* by replacing $f(x) \leq f(y)$ with $f(x) < f(y)$ in Definition 3.2.

Definition 3.3 (Setwise global minimum). A subset S of \mathbb{R}^n is a *setwise global minimum* of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ if S is a setwise local minimum of f and $\inf_S f = \inf_{\mathbb{R}^n} f$.

Note that $\inf_S f$ is a shorthand for $\inf_{x \in S} f(x)$, and similar for $\sup_S f$ and $\max_S f$. Setwise local minima include setwise global minima as a special case, and we say a setwise local minimum is *spurious* if it is not a setwise global minimum. Note that Definition 3.2 is not exactly the same as [84, Definition 2.5] because we do not require a setwise local minimum to be a compact set. In other words, a setwise local minimum can be either bounded or unbounded, and we say a (spurious) setwise local minimum is a (spurious) *local minimum at infinity* if it is unbounded. For example,

consider the loss function of a one-hidden layer neural network with sigmoid activation σ and two data points $(1, 1)$ and $(-1, -3)$ in Figure 3.3. One can see that S is a setwise local minimum (in particular, a local minimum at infinity) and U is the corresponding open set in Definition 3.2. Finally, observe that \mathbb{R}^n is a strict setwise local minimum at infinity of any function.

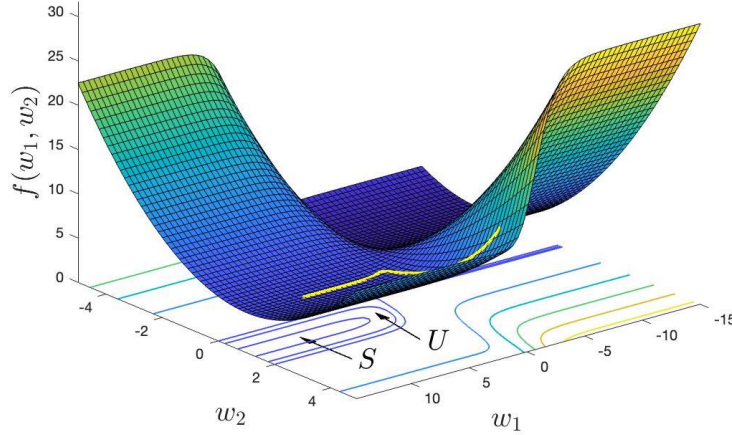


Figure 3.3: Local minimum at infinity of $f(w_1, w_2) = \frac{1}{2}[(w_2\sigma(w_1) - 1)^2 + (w_2\sigma(-w_1) + 3)^2]$.

Now we introduce one of the most useful properties of setwise local minima in Lemma 3.1. This property is intuitive and will be used in different scenarios throughout this chapter. Let \bar{S} , S° , and $\partial S := \bar{S} \setminus S^\circ$ respectively denote the closure, interior, and boundary of a subset S of \mathbb{R}^n .

Lemma 3.1. *If S is a setwise local minimum of a continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, then $f(x) = \sup_S f$ for all $x \in \partial S$.*

Proof. Let $S \subset \mathbb{R}^n$ be a setwise local minimum of a continuous function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Let $U \supset S$ be an open set such that $f(x) \leq f(y)$ for all $x \in S$ and $y \in U \setminus S$. Note that S is closed, so its boundary is defined by $\partial S := S \setminus S^\circ$. Let $z \in \partial S$ and consider any real number $\epsilon > 0$. Since $f(z) + (-\epsilon, \epsilon)$ is a neighborhood of $f(z)$, by continuity of f , there exists a neighborhood $N(z)$ of z such that $f(N(z)) \subset f(z) + (-\epsilon, \epsilon)$. Since U is a neighborhood of z , $N'(z) := U \cap N(z)$ is also a neighborhood of z with $f(N'(z)) \subset f(z) + (-\epsilon, \epsilon)$. The set $N'(z) \cap S$ is nonempty because $z \in S$ and the set $N'(z) \setminus S$ is nonempty because $z \in \partial S$. For any $x \in N'(z) \cap S$ and $y \in N'(z) \setminus S$, it

follows that

$$\inf_{U \setminus S} f - \epsilon \leq f(y) - \epsilon < f(z) < f(x) + \epsilon \leq \sup_S f + \epsilon \leq \inf_{U \setminus S} f + \epsilon.$$

The last inequality follows from the definition of setwise local minima. As $\epsilon > 0$ was arbitrary, we deduce that

$$\inf_{U \setminus S} f = f(z) = \sup_S f.$$

Thus, f is a constant on the boundary of S and f attains its maximum over S on the boundary of S . □

It is worth relating our notion of setwise local minimum to the concept of *valley* proposed in [85, Definition 1]. A *valley* of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is defined as a path-connected component¹ of a sublevel set of f . These two definitions are distinct in general. The interval $[-1, 1]$ is a setwise local minimum of $f : \mathbb{R} \rightarrow \mathbb{R}$ defined by $f(x) := 0$ for all $x \in \mathbb{R}$ yet it is not a valley. Conversely, $X := \{(x_1, x_2) \in (\mathbb{R} \setminus \{0\}) \times \mathbb{R} \mid x_2 = \sin(1/x_1)\}$ is a valley of $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ where f is defined as the distance between x and X , yet it is not a setwise local minimum since it is not closed. (The sublevel set of f corresponding to the value zero is composed of two path-connected components, namely X and $\{0\} \times [-1, 1]$, whose union is \bar{X} .) Under some mild conditions, their relation can be summarized in Proposition 3.1.

Proposition 3.1. *For a continuous function from \mathbb{R}^n to \mathbb{R} ,*

- (a) *a path-connected component of a strict setwise local minimum (respectively, setwise local minimum) is a valley (respectively, subset of a valley);*
- (b) *a connected component² of a sublevel set which has finitely many connected components is a strict setwise local minimum.*

¹A subset S of \mathbb{R}^n is path-connected if for all $x, y \in S$, there exists a continuous function $\gamma : [0, 1] \rightarrow S$ such that $\gamma(0) = x$ and $\gamma(1) = y$. A maximal path-connected set is called a path-connected component. Path-connected components can be viewed as equivalence classes over a set.

²A subset S of \mathbb{R}^n is disconnected if there exist nonempty disjoint open (in S) sets A and B such that $S = A \cup B$. It is connected if it is not disconnected. A maximal connected set is called a connected component.

Proof. (a) Let S be a setwise local minimum. By Lemma 3.1, we know that $c := \sup_S f = f(z)$ for all $z \in \partial S$. Take a path-connected component C of S . Then $C \subset [f \leq c] := \{x \in \mathbb{R}^n \mid f(x) \leq c\}$. Since C is path-connected, there exists a path-connected component V of $[f \leq c]$ such that $C \subset V$. By definition, V is a valley. This shows that a path-connected component of a setwise local minimum is a subset of a valley.

If in addition, S is a strict setwise local minimum, then we distinguish two cases. If $S = \mathbb{R}^n$, then the path-connected component C of S is equal to \mathbb{R}^n and is therefore a valley. Otherwise, it suffices to show that $V \subset S$. Indeed, V is then a path-connected subset of S containing the path-connected component C of S , so that by maximality, $V = C$. Therefore C is valley.

Consider an open set $U \supset S$ such that $f(x) > f(y)$ for all $x \in U \setminus S$ and $y \in S$. In order to show that $V \subset S$, it suffices to show that $V \cap (U \setminus S) = \emptyset$ and $V \cap U^c = \emptyset$ because if so, then $V = (V \cap S) \cup (V \cap (U \setminus S)) \cup (V \cap U^c) = V \cap S$. Since $f(w) \leq c$ for all $w \in V$ and $f(w) > c$ for all $w \in U \setminus S$ (the supremum function value $f(y) = c$ can be attained by some $y \in \partial S \subset S$), we know that $V \cap (U \setminus S) = \emptyset$. Thus, $V = (V \cap S) \cup (V \cap U^c)$. Note that $V \cap S$ is nonempty and closed because $C \subset V \cap S$ and V and S are both closed. Since $V \cap U^c$ is also closed and V is connected, $V \cap U^c$ must be empty.

(b) Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a continuous function and $a \in \mathbb{R}$ be a nonempty sublevel set of f . By continuity of f , $[f \leq a]$ is closed in \mathbb{R}^n . Suppose $[f \leq a]$ has finitely many connected components C_1, \dots, C_k . Denote \overline{B} as the closure of any set $B \subset \mathbb{R}^n$. Since C_i 's are connected, by [89, Theorem 23.4], $\overline{C_i}$'s are also connected. Since $C_i \subset [f \leq a]$, $\overline{C_i} \subset \overline{[f \leq a]} = [f \leq a]$. By [89, Theorem 25.1], $\overline{C_i}$ has no intersection with any other C_j for $j \neq i$. Together with the fact that $[f \leq a] = \bigcup_{i=1}^k C_i$, we have $\overline{C_i} \subset C_i$, and hence $\overline{C_i} = C_i$. Thus, each C_i is closed in \mathbb{R}^n .

For any fixed i , denote $C_{-i} := [f \leq a] \setminus C_i$, then $C_{-i} = \bigcup_{j=1, j \neq i}^k C_j$ is a closed set disjoint with C_i . By [89, Theorem 32.2], there exist disjoint open sets $D, E \subset \mathbb{R}^n$ such that

$C_i \subset D$ and $C_{-i} \subset E$. Take $U = D$ in Definition 3.2, then $f(x) \leq a$ for all $x \in C_i$ because $C_i \subset [f \leq a]$. Furthermore, $f(y) > a$ for all $y \in U \setminus C_i$ because $(U \setminus C_i) \cap [f \leq a] = \emptyset$. This verifies that C_i is a strict setwise local minimum of f .

□

Remark 3.1. The assumption on finiteness of connected components is necessary, or else counterexample may occur when the function is oscillatory. For example,

$$f(x) := \begin{cases} 0 & \text{if } x \leq 0, \\ x^2 \sin \frac{1}{x} & \text{if } x > 0. \end{cases}$$

The function f is continuous on \mathbb{R} , but the sublevel set $\{x \in \mathbb{R} \mid f(x) \leq 0\}$ has infinitely many connected components. Take a connected component $C_1 = (-\infty, 0]$ (also path-connected, thus a valley), and it is not a setwise local minimum because for any open set U containing C_1 , there exists some $x_0 \in U$ such that $f(x_0) < 0$.

Finally, we discuss the case of coercive functions.

Proposition 3.2. *If a continuous function from \mathbb{R}^n to \mathbb{R} is coercive, then it has no spurious local minima at infinity.*

Proof. Let S be a spurious setwise local minimum at infinity. Since $\inf_S f > \inf_{\mathbb{R}^n} f$, it must be that $S \neq \mathbb{R}^n$. By Definition 3.2, there exists $y \in S^c$ such that $S \subset \{x \in \mathbb{R}^n \mid f(x) \leq f(y)\}$. Since f is coercive, its sublevel sets are bounded and hence S is bounded. S is thus not a spurious local minimum at infinity. □

In many statistical learning problems, the loss functions without regularizer are usually not coercive, so spurious local minima at infinity may exist. Therefore, it is important to develop some device to check whether spurious local minima exist or not so that optimization algorithms can be designed to avoid getting trapped in them.

3.2 Proof of Theorem 3.1

This section contains the proof of the main result, i.e., Theorem 3.1. After the proof, we will explain the necessity of the assumptions in Theorem 3.1 by raising some counterexamples. For emphasis, we summarize all assumptions in Theorem 3.1 below.

Assumption 3.1. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a function such that it

- (a) is bounded below, namely, $\inf_{\mathbb{R}^n} f > -\infty$;
- (b) is locally Lipschitz continuous on \mathbb{R}^n ; see Definition 1.3;
- (c) admits a chain rule; see Definition 1.8;
- (d) has finitely many critical values; see Section 1.1;
- (e) has bounded subgradient trajectories; see Definition 1.7.

Proof of Theorem 3.1. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a function satisfying Assumption 3.1. If f has no spurious setwise local minima, then f has no spurious local minima. We next prove the converse. Let $S \subset \mathbb{R}^n$ be a setwise local minimum of f . We seek to show that S is a setwise global minimum of f . If $S^\circ = \emptyset$, then by Lemma 3.1 $f(x) = \sup_S f$ for all $x \in \partial S = \bar{S} \setminus S^\circ = S$ since S is closed by Definition 3.2. Thus f is constant on S . By definition of setwise local minima (Definition 3.2), there exists an open subset U of \mathbb{R}^n containing S such that the constant value of f on S is less than or equal to $f(y)$ for all $y \in U \setminus S$. Thus every point in S is a local minimum of f . Since every local minimum of f is a global minimum, $\inf_S f = \inf_{\mathbb{R}^n} f$ and S is a setwise global minimum according to Definition 3.3. The rest of the proof deals with the case when $S^\circ \neq \emptyset$. Let C be the set of all critical points of f in S and consider the following optimization problem:

$$\inf_{x \in C} f(x). \tag{3.1}$$

We claim that the set of (global) solutions to (3.1) is nonempty, and that any solution is a local

minimum of f belonging to the setwise local minimum S . We first show that the feasible set of (3.1) is nonempty.

Since $S^\circ \neq \emptyset$, let $x_0 \in S^\circ$. If $x_0 \in C$, then the feasible set C is nonempty. We thus assume that $x_0 \notin C$. Since f is locally Lipschitz and bounded below, by Proposition 1.2 there exists a subgradient trajectory $x : [0, \infty) \rightarrow \mathbb{R}^n$ starting at x_0 . We next show that $x([0, \infty)) \subset S$. We reason by contradiction and assume that $S^c \cap x([0, \infty)) \neq \emptyset$, where S^c is the complement of S in \mathbb{R}^n . Then S° and S^c are disjoint open subsets of \mathbb{R}^n such that $S^\circ \cap x([0, \infty)) \neq \emptyset$ (the intersection contains x_0), $S^c \cap x([0, \infty)) \neq \emptyset$, and $x([0, \infty)) = (x([0, \infty)) \cap S^\circ) \cup (x([0, \infty)) \cap S^c) \subset \mathbb{R}^n \setminus \partial S$ (since³ $f(x(t)) < f(x(0)) = f(x_0) \leq f(x)$ for all $t > 0$ and $x \in \partial S$, where the last inequality follows from Lemma 3.1). Thus the connected set $x([0, \infty))$ is the union of two relatively open disjoint nonempty sets, which is a contradiction.

Since f has bounded subgradient trajectories and $x(\cdot)$ is an arbitrary subgradient trajectory starting at x_0 , by Definition 1.7 and without loss of generality there exists $r > 0$ such that $\|x(t)\| \leq r$ for all $t \geq 0$. We next show that there exists a critical point of f in $\mathring{B}(0, r) \cap S$. Suppose that there exist two constants $T, \epsilon > 0$ for which $\|x'(t)\| \geq \epsilon$ for all $t \geq T$ such that $x'(t) \in -\partial f(x(t))$. By [22, Lemma 5.2], we have $(f \circ x)'(t) = -\|x'(t)\|^2 \leq -\epsilon^2$ for almost every $t \geq T$. By integrating, we get $f(x(t)) - f(x(T)) \leq -\epsilon^2 t$ and thus $f(x(t))$ converges to $-\infty$ as $t \rightarrow \infty$. This is impossible since $x(t) \in \mathring{B}(0, r)$ and f is continuous. Hence there exists a time sequence $t_k \rightarrow \infty$ such that $\|x'(t_k)\| \rightarrow 0$ as $k \rightarrow \infty$ and $x'(t_k) \in -\partial f(x(t_k))$ for all $k \in \mathbb{N} := \{0, 1, 2, \dots\}$. By the Bolzano–Weierstrass theorem, there exists a subsequence $x(t_{k_j})$ of $x(t_k)$ such that $x(t_{k_j}) \rightarrow \tilde{x} \in \mathbb{R}^n$ as $j \rightarrow \infty$. Since $x'(t_{k_j}) \in -\partial f(x(t_{k_j}))$, by [40, 2.1.5 Proposition (b) p. 29] we have $0 \in -\partial f(\tilde{x})$. Finally, since $x([0, \infty)) \subset S$ and S is closed, we have $\tilde{x} \in C$. We obtain that $C \neq \emptyset$ as desired.

Since f has finitely many critical values and $C \neq \emptyset$, the set of solutions to (3.1) is nonempty. Let $x^* \in C$ be a solution, that is to say $f(x^*) = \min_C f$. Recall that C is a subset of the setwise local minimum S . If x^* is a local minimum of f , then it is a global minimum of f in S since every local

³If there exists $t > 0$ such that $f(x(t)) = f(x(0))$, then $f(x(t)) - f(x(0)) = \int_0^t \|x'(s)\|^2 ds = 0$ and $x'(s) = 0$ for almost every $s \in (0, t)$. Since $x'(s) \in -\partial f(x(s))$ for almost every $s > 0$, by [40, 2.1.5 Proposition (b) p. 29] we have $0 \in \partial f(x(0))$.

minimum of f is a global minimum. Thus $\inf_S f = \inf_{\mathbb{R}^n} f$ and S is a setwise global minimum. For the remainder of the proof, we consider the case where x^* is not a local minimum and show that this leads to a contradiction. We first show that there exists $s_0 \in S^\circ$ such that $f(s_0) < f(x^*)$. This is clearly true if $x^* \in S^\circ$ since one can then find a ball centered at x^* inside S° . If $x^* \in S \setminus S^\circ = \partial S$, then we reason by contradiction and assume that $f(x) \geq f(x^*)$ for all $x \in S^\circ$. By Lemma 3.1, we have $f(x) = f(x^*) = \sup_S f \geq f(y) \geq f(x^*)$ for all $(x, y) \in (S \setminus S^\circ) \times S^\circ$. Hence $f(x^*) = f(x)$ for all $x \in S$. Since S is a setwise local minimum, there exists an open set U such that $f(x) \geq f(x^*)$ holds for all $x \in U \setminus S$. Thus $f(x) \geq f(x^*)$ for all $x \in U$ and x^* is a local minimum. This yields a contradiction. Hence let $s_0 \in S^\circ$ be such that $f(s_0) < f(x^*)$. The nonempty closed set $S' := S \cap [f \leq (f(s_0) + f(x^*))/2]$ is a setwise local minimum of f where $[f \leq \alpha] := \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}$. Indeed, for all $x \in S'$ and $y \in U \setminus S' = (U \setminus S) \cup (U \setminus [f \leq (f(s_0) + f(x^*))/2])$, we have⁴ $f(x) \leq (f(s_0) + f(x^*))/2 \leq f(y)$. Since $s_0 \in S^\circ$ and $f(s_0) < (f(s_0) + f(x^*))/2$, we have $s_0 \in S^\circ \cap [f \leq (f(s_0) + f(x^*))/2]^\circ = (S \cap [f \leq (f(s_0) + f(x^*))/2])^\circ = (S')^\circ$. Hence the setwise local minimum S' has nonempty interior. Also, $S' \subset S$ and $\sup_{S'} f < f(x^*) = \min_C f$ where we remind the reader that C is the set of critical points in S . Thus S' is devoid of critical points. However, by the previous paragraph, setwise local minima of f with nonempty interior must contain a critical point. This yields a contradiction. \square

Remark 3.2 (Finitely many critical values). This assumption is not intuitive and we explain why it is necessary by the following example. Define $f : \mathbb{R} \rightarrow \mathbb{R}$ as

$$f(x) := \begin{cases} (x+4)^2 - 8 & \text{if } x \leq -2; \\ -x^2 & \text{if } x \in [-2, 0]; \\ -2^{-k}(x-2k)^{2k+1} - 3(1-2^{-k}) & \text{if } x \in [2k, 2k+1], k \in \mathbb{N}; \\ 2^{-k}(x-2k)^{2k+1} - 3(1-2^{-k}) & \text{if } x \in [2k-1, 2k], k \in \mathbb{N}^*. \end{cases}$$

To be more intuitive, we give the plot of f on $[-7, 7]$ in Figure 3.4.

⁴Indeed, for any sets A, B , and C it holds that $A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$.

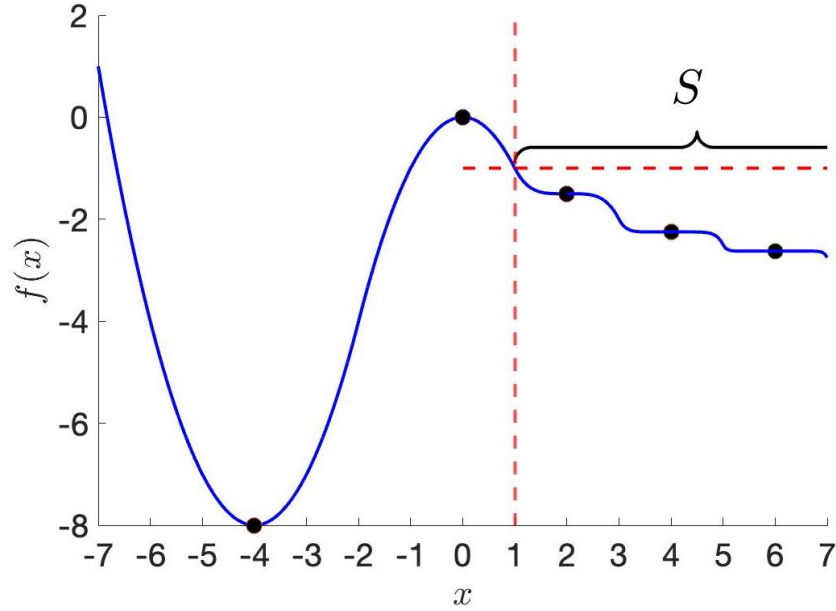


Figure 3.4: An example of function with infinitely many critical values

By standard calculus, one can see f is continuously differentiable, $f(x) \rightarrow -3$ as $x \rightarrow \infty$, and $f(x) \geq -8$ over \mathbb{R} . Furthermore, $\{-4\} \cup \{2k\}_{k \in \mathbb{N}}$ are all critical points of f , with critical values $\{-8\} \cup \{-3(1 - 2^{-k})\}_{k \in \mathbb{N}}$ respectively. Finally, the subgradient trajectory of f starting at $x_0 < 0$ will converge to the critical point $x = -4$; the one starting at $x_0 = 0$ will stay at the critical point $x = 0$; and the one starting at $x_0 > 0$ such that $2k < x_0 \leq 2k + 2$ will converge to $x = 2k + 2$, for all $k \in \mathbb{N}$. This shows f has bounded subgradient trajectories. Thus, f satisfies all conditions in Assumption 3.1 except the finiteness of critical values. It is also easy to see f has no spurious local minima because all of its critical points are either global minimum ($x = -4$), or local maximum ($x = 0$), or saddle points. However, for any $a > 0$, the set $[a, \infty)$ is a spurious local minima at infinity. This shows that Theorem 3.1 may not hold for functions with infinitely many critical values.

Remark 3.3 (Bounded subgradient trajectories). This is the main assumption of Theorem 3.1. From the proof of Theorem 3.1, one can easily see that instead of requiring all subgradient trajectories being bounded, only one bounded subgradient trajectory is needed for each initial point. However, this relaxed one is necessary because without it, one could easily think of a smooth func-

tion without any spurious local minimum, yet has spurious local minimum at infinity. This is the case of the function in Figure 3.3 in which the yellow curve corresponds to an unbounded gradient trajectory. In order to prove the necessity of the boundedness assumption, it suffices to consider the univariate function f defined in [84, Figure 4(a)] defined by

$$f(x) := \frac{x^2(1+x^2)}{1+x^4}, \quad f'(x) = -\frac{2x(x^4 - 2x^2 - 1)}{(x^4 + 1)^2}.$$

By solving $f'(x) = 0$, we know that f has three critical points, among which $x = 0$ is the global minimum and $x = \pm(\sqrt{2} - 1)^{-1/2}$ are two global maxima. Thus, f is bounded below, continuously differentiable (hence locally Lipschitz and admits a chain rule), has finitely many critical values, and has no spurious local minima. Since f is strictly decreasing for all $x \geq (\sqrt{2} - 1)^{-1/2} \approx 1.55$ and $f(x) \rightarrow 1$ as $x \rightarrow \infty$, one can easily see $[2, \infty)$ is a spurious local minimum at infinity. This shows that Theorem 3.1 does not hold and the reason is that f does not have bounded subgradient trajectories. To see this explicitly, consider the Cauchy problem

$$\dot{x} = \frac{2x(x^4 - 2x^2 - 1)}{(x^4 + 1)^2}, \quad x(0) = 2.$$

By using separation of variables, the unique solution $x(t)$ is given by

$$\begin{aligned} c + 2t &= \frac{1}{4}x^4 + x^2 + (2 + \sqrt{2}) \log(x^2 - \sqrt{2} - 1) \\ &\quad + (2 - \sqrt{2}) \log(x^2 + \sqrt{2} - 1) - \log x =: g(x), \end{aligned}$$

where c is a constant determined by $x(0) = 2$. It is easy to see that x is strictly increasing so $x(t) \geq 2$ for all $t \in [0, \infty)$. Note that g is continuous on $[2, \infty)$, so if x is bounded, then $g \circ x$ is bounded. This contradicts the fact that $g(x(t)) = 2t + c \rightarrow \infty$ as $t \rightarrow \infty$, and thus f has an unbounded subgradient trajectory.

3.3 Applications

In this section, we use Theorem 3.1 to analyze the landscape of some widely used loss functions in unconstrained optimization. To be more specific, we will consider deep linear neural network, one dimensional deep sigmoid neural network, matrix sensing, and nonsmooth matrix factorization in the following four subsections respectively.

3.3.1 Deep linear neural network

As a prototypical example in deep learning, the landscape of deep linear neural network has been widely studied; see for example [90, 91, 85]. Recall the objective function f of deep linear neural network defined in (2.11). It was recently established that f has no spurious valleys [85, Theorem 11], however this fact alone does not imply the absence of spurious local minima at infinity (recall Figure 3.2). Together with the fact that f has no spurious local minima [92, Corollary 1] and that f is semi-algebraic, it can be deduced that f has no spurious setwise local minima (and thus no spurious local minima at infinity).

The proof of the absence of spurious valleys [85, Theorem 11] is tailored to the problem at hand. Using linear algebra, it argues that from any initial point one can construct a piecewise linear path to a global minimum along which the objective function is non-increasing. The proof spans multiple pages and requires several technical lemmas. The proof that we propose is shorter and follows a general principle, namely Theorem 3.1, that applies to various problems as the next subsections will show. The first four assumptions of Theorem 3.1 are easy to verify: f is nonnegative, hence bounded below; f is continuously differentiable, hence locally Lipschitz and admits a chain rule; f is semi-algebraic, by [42, Corollary 1.1], it has finitely many critical values. Since f has bounded gradient trajectories Proposition 2.6, we verified that f satisfies Assumption 3.1. Thus, (2.11) has no spurious setwise local minima if and only if it has no spurious local minima. Since a local minimum at infinity is an unbounded setwise local minimum, and f has no spurious local minima, we conclude that f has no spurious local minima at infinity. This

proves the first result in Corollary 3.1.

3.3.2 One dimensional deep sigmoid neural network

Landscape analysis of one or two-hidden layer sigmoid neural network can be found, for instance, in [85, 87, 88]. However, none of the results above can be easily generalized to arbitrary many layers.

We want to apply Theorem 3.1 to conclude that (2.16) has no spurious setwise local minimum, and hence no local minima at infinity. Again, the first three assumptions in Assumption 3.1 are easy to verify: f is nonnegative, hence bounded below; f is continuously differentiable, hence locally Lipschitz, and admits a chain rule. Note that f is not semi-algebraic, but it is definable in the real exponential field [93] [57, Section 6.2], so by Morse–Sard theorem for definable functions [43, Corollary 9(ii)], it has finitely many critical values.

With Proposition 2.7, we can conclude that f has no spurious setwise local minimum if and only if it has no spurious local minima. However, from the gradient of f , we can easily see that any critical point of it will be a global minimum, so f has neither spurious local minimum nor spurious setwise local minimum. This verifies the second result in Corollary 3.1.

Unfortunately, unlike linear neural networks, the result in Proposition 2.7 is not true in general even in one-hidden layer case, if more than one data point is given; see Example 3.2. However, it is still an open question whether the gradient trajectories will be bounded in the over-parameterized case (in which case there exists at least one achievable global minimum).

Example 3.2. Consider the following function

$$f(w_1, w_2) := \frac{1}{2}[(w_2\sigma(w_1) - 1)^2 + (w_2\sigma(-w_1) + 1)^2]. \quad (3.2)$$

The above function represents a one-hidden layer sigmoid neural network with two data $(x_1, y_1) = (1, 1)$ and $(x_2, y_2) = (-1, -1)$. By directly computing the gradient, one can easily see that (3.2) has only one critical point $(0, 0)$ which is a strict saddle with $f(0, 0) = 1$. The global minimum is

asymptotically attained as $w_1 \rightarrow \pm\infty$ and $w_2 \rightarrow 1 - 2(1 + e^{2w_1})^{-1}$, and its corresponding objective value approaches to $1/2$. In this case, the gradient trajectory of (3.2) starting at any point x_0 such that $f(x_0) < 1$ must be unbounded.

3.3.3 Matrix sensing

The landscape of (2.4) has been studied widely, for example, in [94, 95, 96]. Most of these work are based on the restrictive isometry property (RIP) of sensing matrices. A set of sensing matrices A_i for $i = 1, \dots, m$ are said to have (r, δ_r) -RIP [66] if there exists $\delta_r \in (0, 1)$ such that

$$(1 - \delta_r)\|\tilde{M}\|_F^2 \leq \frac{1}{m} \sum_{i=1}^m \langle A_i, \tilde{M} \rangle_F^2 \leq (1 + \delta_r)\|\tilde{M}\|_F^2$$

holds for any matrix \tilde{M} with $\text{rank}(\tilde{M}) \leq r$. To the best of our knowledge, the minimal assumptions to guarantee no spurious local minima for (2.4) is for the sensing matrices to satisfy $(4r, \delta_{4r})$ -RIP with $\delta_{4r} \leq 1/5$, as proposed in [96, Theorem III.1].

However, Theorem 3.1 is applicable to matrix sensing under a weaker condition than RIP. The first four assumptions in Assumption 3.1 hold because of exactly the same reasons as in the linear neural network case. For bounded gradient trajectories, Proposition 2.3 shows that lower bounded sensing matrices suffices. Therefore, Theorem 3.1 says that matrix sensing has no spurious setwise local minima if and only if it has no spurious local minima, given that the sensing matrices are lower bounded. Equipped with $(4r, \delta_{4r})$ -RIP where $\delta_{4r} \leq 1/5$, we conclude that matrix sensing has no spurious local minima at infinity, as shown in the third statement in Corollary 3.1.

3.3.4 Nonsmooth matrix factorization

There are few landscape results of (2.5) in the general rank case. However, if $\text{rank}(M) = 1 = r$, (2.5) is shown to have no spurious local minima if every entry M_{ij} of M is nonzero [97].

It is hard to analyze (2.5) because it is nonsmooth, nonconvex, and noncoercive. Despite all those “non” properties, we show that Theorem 3.1 is still applicable to (2.5) without any rank

assumption on M . As a corollary, when $\text{rank}(M) = 1$ and every entry of M is nonzero, (2.5) has no spurious setwise local minimum, hence no spurious local minima at infinity. Again, the first four assumptions in Assumption 3.1 are easy to check: f is bounded below because it is nonnegative; f is locally Lipschitz because of [40, Theorem 2.3.10]; since f is semi-algebraic, by [55, Corollary 5.4] and [42, Corollary 1.1], it admits a chain rule and has finitely many critical values.

Combined with Proposition 2.4, Theorem 3.1 shows that (2.5) has no spurious setwise local minimum if and only if it has no spurious local minima. Under the condition in [97, Theorem 1], i.e., $\text{rank}(M) = 1 = r$ and all the entries of M are non-zero, (2.5) reduces to

$$f(x, y) := \sum_{i=1}^m \sum_{j=1}^n |x_i y_j - M_{ij}|, \quad (3.3)$$

where $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$. In this case, (3.3) has no spurious local minima, thus it has no spurious local minima at infinity, and we obtain the last result in Corollary 3.1.

Chapter 4: Convergence of momentum methods

The landscape result (Theorem 3.1) in Chapter 3 establishes that, under the assumptions of bounded subgradient trajectories and the absence of spurious local minima, there are no spurious local minima at infinity. While this provides valuable insight into the global structure of certain optimization problems, it offers only a partial explanation for the convergence behavior of first-order methods. Specifically, it does not preclude the possibility that an algorithm may converge to a global minimum at infinity, nor does it address settings where spurious local minima are present, a situation frequently encountered in data science applications [90, 98, 99].

In this chapter, we address these limitations by analyzing a representative first-order algorithm: the momentum method. Leveraging the framework of bounded subgradient trajectories, we develop convergence guarantees regardless of the presence of spurious local minima. The results presented in this chapter are based on the following article:

C. Josz, L. Lai, and X. Li, “Convergence of the momentum method for semialgebraic functions with locally Lipschitz gradients,” *SIAM Journal on Optimization*, vol. 33, no. 4, pp. 3012–3037, 2023

The gradient method with constant momentum and constant step size (or momentum method for short [101, Equation (10)]) for minimizing a differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ consists in choosing initial points $x_{-1}, x_0 \in \mathbb{R}^n$ and generating a sequence of iterates according to the update rule

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha \nabla f(x_k + \gamma(x_k - x_{k-1})), \quad \forall k \in \mathbb{N}, \quad (4.1)$$

where $\alpha > 0$ is the step size and $\beta \in (-1, 1)$ and $\gamma \in \mathbb{R}$ are constant momentum parameters, as implemented in PyTorch [102, 103] and TensorFlow [104]. When $\gamma = 0$, this reduces to Polyak’s heavy ball method [105, Equation (9)] and when $\beta = \gamma$, this reduces to Nesterov’s accelerated

gradient method [16, Equation (2.2.22)]. If the objective function is strongly convex and satisfies some regularity assumptions, the former has a nearly optimal local convergence rate [105, Theorem 9] [16, Theorem 2.1.13], while the latter has a globally optimal convergence rate [16, Equation (2.2.23)]. This holds with a suitable choice of parameters α , β , and γ .

Various objective functions of interest nowadays are however not convex, including matrix factorization, matrix sensing, and linear neural networks. These problems have in common that they are semialgebraic and have locally Lipschitz continuous gradients. However, they do not have globally Lipschitz continuous gradients, they are not coercive, and whether they satisfy a global growth condition is unknown and hard to check for. In other words, the commonly used assumptions H1 (sufficient decrease), H2 (relative error), H3 (continuity) due to Attouch et al. [21], adapted to the momentum method in [106, 107], are not true, and H4 (global growth) is unknown. As a result, it is not known whether the momentum method — in particular the heavy ball method and Nesterov’s accelerated gradient method — would converge if the initial points lie close to a local minimizer of f . A fortiori, nothing is known if they are chosen arbitrarily or at random in \mathbb{R}^n .

Even if one assumes that the iterates are bounded, a common assumption in the literature which implies H3, it is not known whether the iterates would converge. Indeed, choose a step size $\alpha > 0$ and suppose that the iterates are bounded. Let $L > 0$ denote a Lipschitz constant of the gradient on the convex hull of the iterates. If $\alpha \geq 2/L$, then the argument employed in [20, Theorem 3.2] and [21, Theorem 3.2], which consists of taking a subsequence and invoking the Kurdyka-Łojasiewicz inequality [108, Proposition 1 p. 67] [44, Theorem 1], fails to establish convergence. Unfortunately, there is no way to control the size of L before choosing the step size α .

We next review the literature on the momentum method in the nonconvex setting. All the results in the literature require that f has an L -Lipschitz continuous gradient with $L > 0$, along with other assumptions that we next describe. First, we discuss convergence when the initial points are near a local minimizer. If the objective function f satisfies the Kurdyka-Łojasiewicz inequality at a local minimizer $x^* \in \mathbb{R}^n$, $\alpha \in (0, 2(1 - \beta)/L)$, $\beta \in [0, 1)$, $\gamma = 0$, and a global growth condition [107, (H4)] is satisfied, then the momentum method converges to a local minimizer

when initialized sufficiently close to x^* [107, Theorem 3.2]. The growth condition implies the existence of constants $a, b > 0$ such that $f(x) + b\|x - y\|^2 \geq f(x^*) - a\|y - x^*\|^2$ for all $x, y \in \mathbb{R}^n$, and in particular $f(x) \geq f(x^*) - a\|x - x^*\|^2$ for all $x \in \mathbb{R}^n$.

Second, we discuss convergence when the initial points are arbitrary. Under the same parameter settings for α, β, γ , if f is lower bounded, then the gradients $\nabla f(x_k)$ converge to zero [109, Lemmas 1, 2, 3] for any initial points $x_{-1}, x_0 \in \mathbb{R}^n$. If in addition the function is coercive and satisfies the Kurdyka-Łojasiewicz inequality [44] at every point and $x_{-1} = x_0$, then the iterates have finite length [106, Theorem 4.9]. If the Łojasiewicz gradient inequality holds [108, Proposition 1 p. 67], then a local convergence rate can be deduced [107, Theorem 3.3]. If instead the function satisfies an error bound and its level sets are properly separated, then with $\alpha \in (0, 1/L)$, $\beta = \gamma \in [0, 1/\sqrt{1 + L\alpha})$, and $x_{-1} = x_0$, the iterates and the function values converge linearly to a critical point and a critical value respectively [110, Theorem 3.7]. Finally, if the function satisfies the Kurdyka-Łojasiewicz inequality and the iterates are bounded, then they have finite length [111, Theorem 3.5] under the same parameter settings.

Third, we discuss convergence when the initial points are chosen outside a zero measure set. The momentum method is known to converge to a local minimizer for almost every initial point under several conditions. First, f should be coercive, twice differentiable, and should satisfy the Kurdyka-Łojasiewicz inequality. Second, the Hessian of f should have a negative eigenvalue at all critical points of f that are not local minimizers. Third, the parameters of the momentum method (4.1) should either satisfy $\alpha \in (0, 2(1 - \beta)/L)$, $\beta \in (0, 1)$ and $\gamma = 0$ [112, Lemma 2], or $\alpha \in (0, 4/L)$, $\beta \in (\max\{0, -1 + \alpha L/2\}, 1)$ and $\gamma = 0$ [113, Theorem 3].

Our contributions are as follows. We consider objective functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ that are semi-algebraic and differentiable with locally Lipschitz gradients. The generalization to arbitrary o-minimal structures on the real field [36] is immediate and omitted for the sake of brevity. We show that the length of the iterates generated by the momentum method is upper bounded by an expression depending on the objective function variation, the step size, and a desingularizing function. This length formula enables us to show that global Lipschitz continuity of the gradient and the

global growth condition are superfluous when establishing local convergence. It also enables us to establish global convergence under the assumption that the continuous-time gradient trajectories of f are bounded, which is satisfied by matrix factorization, matrix sensing, and linear neural networks, as discussed in Chapter 2. As a result, we bypass the need for coercivity and globally Lipschitz gradients. Finally, the length formula enables us to guarantee convergence to local minimizers almost surely, under second-order differentiability and the strict saddle property [83, 114].

This chapter is organized as follows. Section 4.1 contains the statement of the length formula and the ensuing local and global convergence results (Theorem 4.1 and Theorem 4.2). Section 4.2 contains the proof of the length formula. Section 4.3 contains the proof of tracking lemma. Section 4.4 contains the proof of result on saddle point avoidance (Theorem 4.3).

4.1 Convergence results

Given $k \in \mathbb{N}$, $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$, let $C^k(A, B)$ be the set of continuous functions $f : A \rightarrow B$ such that, if $k \geq 1$ then f is k times continuously differentiable on the interior of A . Let $C_{\text{loc}}^{1,1}(A, B)$ denote the set of functions in $C^1(A, B)$ whose first-order derivative is locally Lipschitz continuous on the interior of A . When $B = \mathbb{R}$, $C^k(A, B)$ and $C_{\text{loc}}^{1,1}(A, B)$ are abbreviated as $C^k(A)$ and $C_{\text{loc}}^{1,1}(A)$ respectively. A function $\psi : S \rightarrow S$ is a homeomorphism if it is a continuous bijection and the inverse function ψ^{-1} is continuous. $\psi : S \rightarrow S$ is a diffeomorphism if $\overset{\circ}{S} \neq \emptyset$, ψ is a homeomorphism, and both ψ and ψ^{-1} are continuously differentiable on $\overset{\circ}{S}$.

We are now ready to state the key lemma upon which rest all the convergence results in this manuscript. It is entirely new to the best of our knowledge.

Lemma 4.1 (Length formula). *Let $f \in C_{\text{loc}}^{1,1}(\mathbb{R}^n)$ be semialgebraic, $X \subset \mathbb{R}^n$ be bounded, $\beta \in (-1, 1)$, $\gamma \in \mathbb{R}$, and $\delta \geq 0$. There exist $\bar{\alpha}, \eta, \kappa > 0$ and a diffeomorphism $\psi : [0, \infty) \rightarrow [0, \infty)$ such that, for all $K \in \mathbb{N}$, $\alpha \in (0, \bar{\alpha}]$, and sequences $(x_k)_{k \in \{-1\} \cup \mathbb{N}}$ generated by the momentum method*

(4.1) for which $x_{-1}, \dots, x_K \in X$ and $\|x_0 - x_{-1}\| \leq \delta\alpha$, we have

$$\sum_{k=0}^K \|x_{k+1} - x_k\| \leq \psi(f(x_0) - f(x_K) + \eta\alpha) + \kappa\alpha.$$

The significance of this formula is that it relates the length of the iterates with the objective function variation, in spite of the fact that the objective function values generated by the momentum method are notoriously nonmonotonic. The proof of Lemma 4.1 is quite involved so we defer it to Section 4.2. There, the reader will learn that one can actually take ψ to be a desingularizing function of f on X (see Proposition 1.1).

We next provide some intuition on the constants in the length formula. They are constructed explicitly using the regularity of the objective function and the momentum parameters. Both η and κ increase with the number of critical values of f in X and the initial velocity δ in the momentum method. The constant η increases with the minimal Lipschitz constant of ∇f over a certain bounded set and with the magnitude of the momentum $|\beta|$. The constant κ increases with the minimal Lipschitz constant of f over a certain bounded set.

Before we proceed, we state the following simple fact regarding the gradient of the objective function at iterates produced by the momentum method.

Fact 4.1. *Let $f \in C_{\text{loc}}^{1,1}(\mathbb{R}^n)$, $X \subset \mathbb{R}^n$ be bounded, and $\beta, \gamma \in \mathbb{R}$. For all $\alpha > 0$, there exists $b_\alpha > 0$ such that for all $K \in \mathbb{N}$, if $x_{-1}, \dots, x_{K+1} \in X$ are iterates of the momentum method (4.1), then $\|\nabla f(x_k)\| \leq b_\alpha \|z_{k+1} - z_k\|$ for $k = 0, \dots, K$ where $z_k := (x_k, x_{k-1}) \in \mathbb{R}^{2n}$. If $M > 0$ is a Lipschitz constant of ∇f on $S + \max\{|\beta|, |\gamma|\}(S - S)$ where S is the convex hull of X , then one may take $b_\alpha := \sqrt{2} \max\{1/\alpha, |\beta|/\alpha + M|\gamma|\}$.*

Proof. By definition of y_k^β and y_k^γ in (4.13), for $k = 0, \dots, K$, we have

$$\begin{aligned}
\|\nabla f(x_k)\| &\leq \|\nabla f(y_k^\gamma)\| + \|\nabla f(x_k) - \nabla f(y_k^\gamma)\| \\
&\leq \|x_{k+1} - y_k^\beta\|/\alpha + M|\gamma|\|x_k - x_{k-1}\| \\
&\leq \|x_{k+1} - x_k\|/\alpha + (|\beta|/\alpha + M|\gamma|)\|x_k - x_{k-1}\| \\
&\leq \sqrt{2} \max\{1/\alpha, |\beta|/\alpha + M|\gamma|\}\|z_{k+1} - z_k\|,
\end{aligned}$$

□

We are now ready to state our first convergence result.

Theorem 4.1 (Local convergence). *Let $f \in C_{\text{loc}}^{1,1}(\mathbb{R}^n)$ be semialgebraic, $\beta \in (-1, 1)$, $\gamma \in \mathbb{R}$, $\delta \geq 0$, and $x^* \in \mathbb{R}^n$ be a local minimizer of f . For all $\epsilon > 0$, there exist $\bar{\alpha}, \xi > 0$ such that for all $\alpha \in (0, \bar{\alpha}]$ and for all sequence $(x_k)_{k \in \{-1\} \cup \mathbb{N}}$ generated by the momentum method (4.1) for which $\|x_0 - x_{-1}\| \leq \delta\alpha$ and $x_0 \in B(x^*, \xi)$, $(x_k)_{k \in \{-1\} \cup \mathbb{N}}$ converges to a local minimizer of f in $B(x^*, \epsilon)$.*

Proof. Without loss of generality, we may assume that $f(x) \geq f(x^*)$ for all $x \in B(x^*, 2\epsilon)$. Since f is continuous and has finitely many critical values by the semialgebraic Morse-Sard theorem (Lemma 1.1), we may also assume that $f(x^*)$ is the unique critical value in $B(x^*, 2\epsilon)$. By Lemma 4.1, there exist $\bar{\alpha}, \eta > 0$, $\kappa > \delta$, and a diffeomorphism $\psi : [0, \infty) \rightarrow [0, \infty)$ such that, for all $K \in \mathbb{N}$, $\alpha \in (0, \bar{\alpha}]$, and sequences $(x_k)_{k \in \{-1\} \cup \mathbb{N}}$ generated by the momentum method (4.1) for which $x_{-1}, \dots, x_K \in B(x^*, \epsilon)$ and $\|x_0 - x_{-1}\| \leq \delta\alpha$, we have

$$\sum_{k=0}^K \|x_{k+1} - x_k\| \leq \psi(f(x_0) - f(x_K) + \eta\alpha) + \kappa\alpha. \quad (4.2)$$

By the continuity of f , there exists $\xi \in (0, \epsilon/2]$ such that

$$f(x) - f(x^*) \leq \frac{1}{2} \psi^{-1}\left(\frac{\epsilon}{6}\right), \quad \forall x \in B(x^*, \xi). \quad (4.3)$$

Let

$$\bar{\alpha} := \min \left\{ \tilde{\alpha}, \frac{\epsilon}{3\kappa}, \frac{1}{3\eta} \psi^{-1} \left(\frac{\epsilon}{6} \right) \right\}. \quad (4.4)$$

We fix any $\alpha \in (0, \bar{\alpha}]$ from now on. Let $(x_k)_{k \in \{-1\} \cup \mathbb{N}}$ be a sequence generated by the momentum method (4.1) for which $\|x_0 - x_{-1}\| \leq \delta\alpha$ and $x_0 \in B(x^*, \xi)$. If $K := \inf\{k \in \mathbb{N} : x_k \notin B(x^*, \epsilon)\} < \infty$, then

$$\psi^{-1} \left(\frac{\epsilon}{6} \right) = \psi^{-1} \left(\frac{1}{2} \epsilon - \kappa \frac{\epsilon}{3\kappa} \right) \quad (4.5a)$$

$$\leq \psi^{-1} ((\epsilon - \xi) - \kappa \bar{\alpha}) \quad (4.5b)$$

$$\leq \psi^{-1} ((\|x_K - x^*\| - \|x_0 - x^*\|) - \kappa \bar{\alpha}) \quad (4.5c)$$

$$\leq \psi^{-1} (\|x_K - x_0\| - \kappa \alpha) \quad (4.5d)$$

$$\leq \psi^{-1} \left(\sum_{k=0}^{K-1} \|x_{k+1} - x_k\| - \kappa \alpha \right) \quad (4.5e)$$

$$\leq f(x_0) - f(x_{K-1}) + \eta \alpha \quad (4.5f)$$

$$\leq f(x_0) - f(x^*) + \eta \bar{\alpha} \quad (4.5g)$$

$$\leq \frac{1}{2} \psi^{-1} \left(\frac{\epsilon}{6} \right) + \frac{1}{3} \psi^{-1} \left(\frac{\epsilon}{6} \right) \quad (4.5h)$$

$$< \psi^{-1} \left(\frac{\epsilon}{6} \right). \quad (4.5i)$$

As $\psi^{-1}(\epsilon/6) > 0$, a contradiction occurs and thus $K = \infty$. Above, the arguments of ψ^{-1} in (4.5a) are equal. (4.5b) through (4.5e) rely on the fact that ψ^{-1} is an increasing function. (4.5b) is due to $\xi \leq \epsilon/2$ and $\bar{\alpha} \leq \epsilon/(3\kappa)$ by the definition of $\bar{\alpha}$ in (4.4). (4.5c) holds because $x_K \notin B(x^*, \epsilon)$ and $x_0 \in B(x^*, \xi)$. (4.5d) and (4.5e) are consequences of the triangular inequality. (4.5f) is due to the length formula (4.2) and the fact that $x_0, \dots, x_{K-1} \in B(x^*, \epsilon)$ and $x_{-1} \in B(x_0, \delta\alpha) \subset B(x_0, \delta\bar{\alpha}) \subset B(x^*, \xi + \delta\epsilon/(3\kappa)) \subset B(x^*, \epsilon/2 + \delta\epsilon/(3\delta)) \subset B(x^*, \epsilon)$ by the definition of $\bar{\alpha}$ in (4.4). (4.5g) is due to $f(B(x^*, \epsilon)) \subset [f(x^*), \infty)$. Finally, (4.5h) is due to $x_0 \in B(x^*, \xi)$, the choice of ξ as in (4.3), and $\bar{\alpha} \leq \psi^{-1}(\epsilon/6)/(3\eta)$ by definition of $\bar{\alpha}$ in (4.4).

We have shown that $(x_k)_{k \in \{-1\} \cup \mathbb{N}} \subset B(x^*, \epsilon)$. By the length formula (4.2), we have

$$\sum_{k=0}^{\infty} \|x_{k+1} - x_k\| \leq \psi \left(\max_{B(x^*, \xi)} f - \min_{B(x^*, \epsilon)} f + \eta\alpha \right) + \kappa\alpha.$$

Thus the sequence admits a limit $x^\sharp \in B(x^*, \epsilon)$. Combining with Fact 4.1, x^\sharp must be a critical point of f . As $f(x^*)$ is the unique critical value in $B(x^*, 2\epsilon)$, $f(x^\sharp) = f(x^*) \leq f(x)$ for all $x \in B(x^\sharp, \epsilon) \subset B(x^*, 2\epsilon)$. \square

Note that once local convergence is established, [107, Theorem 3.3] can be applied in order to obtain local convergence rates of the iterates. Indeed, the reader will be able to check later that the assumptions [107, (H1), (H2), (H3)] then hold (using Lemma 4.4 and Lemma 4.5 below). The rates also rely on the fact that one can take the diffeomorphism ψ to be of the form $\psi(t) = ct^\theta$ where $c > 0$ and $\theta \in (0, 1]$ for semialgebraic functions (using Proposition 4.2 below).

In order to go from local convergence to global convergence, we make an assumption regarding the continuous-time gradient trajectories of the objective function. Given $f \in C^1(\mathbb{R}^n)$, we refer to maximal solutions to $x'(t) = -\nabla f(x(t))$ for all $t \in (0, T)$ where $T \in (0, \infty]$ as continuous gradient trajectories (see [56, Chapter 17], [115], and [45, Section 3] for background and properties). We say that a continuous gradient trajectory $x : [0, T) \rightarrow \mathbb{R}^n$ is bounded if there exists $c > 0$ such that $\|x(t)\| \leq c$ for all $t \in [0, T)$. This assumption enables us to use a generalized version of a tracking lemma recently proposed by Kovachki and Stuart [101, Theorem 2].

Their result states that the momentum method tracks continuous gradient trajectories up to any given time for all sufficient small constant sizes. In Lemma 4.2 below, we relax their strong regularity assumptions which require the objective function to be thrice differentiable with bounded derivatives. We instead only require it to be differentiable with a locally Lipschitz gradient. In order to do so, we redefine the key quantity M_k in the proof of [101, Theorem 2], which regulates the tracking error and depends on the Hessian of the objective function, so that it depends only on the gradient. In contrast to [101, Theorem 2], we also make the tracking uniform with respect to the initial point in a bounded set. Choosing an upper bound on the step size in order to achieve

this requires some care and cannot be deduced from the proof of [101, Theorem 2]. Below, we use the notation $\lfloor t \rfloor$ to denote the floor of a real number t which is the unique integer such that $\lfloor t \rfloor \leq t < \lfloor t \rfloor + 1$.

Lemma 4.2 (Tracking). *Let $f \in C_{\text{loc}}^{1,1}(\mathbb{R}^n)$ be a lower bounded function, $\beta \in (-1, 1)$, $\gamma \in \mathbb{R}$, $\delta \geq 0$. For any bounded set $X_0 \subset \mathbb{R}^n$ and $\epsilon, T > 0$, there exists $\bar{\alpha} > 0$ such that for all $\alpha \in (0, \bar{\alpha}]$ and for any sequence $x_{-1}, x_0, x_1, \dots \in \mathbb{R}^n$ generated by the momentum method (4.1) for which $x_0 \in X_0$ and $\|x_0 - x_{-1}\| \leq \delta\alpha$, there exists $x(\cdot) \in C^1([0, T], \mathbb{R}^n)$ such that*

$$x'(t) = -\frac{1}{1-\beta} \nabla f(x(t)), \quad \forall t \in (0, T), \quad x(0) \in X_0,$$

for which $\|x_k - x(k\alpha)\| \leq \epsilon$ for $k = 0, \dots, \lfloor T/\alpha \rfloor$.

The proof of Lemma 4.2 is deferred to Section 4.3. We will also use the following simple fact in order to control the length of a single step of the momentum method as a function of the step size.

Fact 4.2. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable and Lipschitz continuous on $X \subset \mathbb{R}^n$, $\beta \in (-1, 1)$, $\gamma \in \mathbb{R}$, and $\delta_0 \geq 0$. There exists $\delta_1 \geq 0$ such that for all $\alpha > 0$, $K \in \mathbb{N}$, and sequence $(x_k)_{k \in \{-1\} \cup \mathbb{N}}$ generated by the momentum method (4.1) for which $x_{-1}, \dots, x_{K-1} \in X$ and $\|x_0 - x_{-1}\| \leq \delta_0\alpha$, we have*

$$\begin{aligned} \|x_k - x_{k-1}\| &\leq \delta_1\alpha, \quad k = 0, \dots, K, \\ \|z_k - z_{k-1}\| &\leq \sqrt{2}\delta_1\alpha, \quad k = 1, \dots, K, \end{aligned}$$

where $z_k := (x_k, x_{k-1}) \in \mathbb{R}^{2n}$. If $L > 0$ is a Lipschitz constant of $\bar{\beta}f$ on $S + \gamma(S - S)$ where S is the convex hull of X and $\bar{\beta} := (1 - \beta)^{-1}$, then we may take $\delta_1 := \delta_0 + L$.

Proof. For $k = -1, \dots, K - 1$, we have

$$\begin{aligned}
x_{k+1} - x_k &= \beta(x_k - x_{k-1}) - \alpha \nabla f(y_k^\gamma) \\
&= \beta^2(x_{k-1} - x_{k-2}) - \alpha(\beta \nabla f(y_{k-1}^\gamma) + \nabla f(y_k^\gamma)) \\
&\vdots \\
&= \beta^{k+1}(x_0 - x_{-1}) - \alpha \sum_{i=0}^k \beta^i \nabla f(y_{k-i}^\gamma)
\end{aligned}$$

where $y_k^\gamma := x_k + \gamma(x_k - x_{k-1})$. Let L be a Lipschitz constant of $\bar{\beta}f$ on $S + \gamma(S - S)$ where S is the convex hull of X . Since $\|x_0 - x_{-1}\| \leq \delta_0 \alpha$, we have

$$\|x_{k+1} - x_k\| \leq |\beta|^{k+1} \|x_0 - x_{-1}\| + \alpha L \bar{\beta}^{-1} \sum_{i=0}^k |\beta|^i \leq (\delta_0 |\beta|^{k+1} + L) \alpha.$$

Given that $\beta \in (-1, 1)$, it suffices to take $\delta := \delta_0 + L$. In addition,

$$\|z_k - z_{k-1}\| = (\|x_k - x_{k-1}\|^2 + \|x_{k-1} - x_{k-2}\|^2)^{1/2} \leq \sqrt{2} \delta \alpha$$

for $k = 1, \dots, K$. □

Finally, we will use the following result.

Lemma 4.3 ([45, Lemma 1]). *Let $f \in C^1(\mathbb{R}^n)$ be a semialgebraic function with bounded continuous gradient trajectories. If $X_0 \subset \mathbb{R}^n$ is bounded, then $\sigma(X_0) < \infty$ where*

$$\begin{aligned}
\sigma(X_0) &:= \sup_{x \in C^1(\mathbb{R}_+, \mathbb{R}^n)} \int_0^\infty \|x'(t)\| dt \\
&\text{subject to } \begin{cases} x'(t) = -\nabla f(x(t)), \forall t > 0, \\ x(0) \in X_0. \end{cases}
\end{aligned}$$

We are now ready to state our second convergence result. It shows that, similar to the gradient method [45, Theorem 1], the momentum method is endowed with global convergence if continuous

gradient trajectories are bounded. In the gradient method, one considers the supremum of the lengths of all discrete gradient trajectories over all possible initial points in a bounded set and over all possible step sizes [45, Equation (28)]. This enables one to reason by induction on the initial set and the upper bound on the step sizes.

When dealing with momentum, one needs to additionally consider an upper bound on the initial velocity $\|x_0 - x_{-1}\|/\alpha$ between two initial points in the inductive reasoning. Fact 4.2 guarantees that the velocity $\|x_k - x_{k-1}\|/\alpha$ remains bounded within each induction step. This enables one to reinitialize the momentum method after an arbitrary large number of iterations. Note that the length formula in Lemma 4.1 admits an error term $\eta\alpha$ that is not present in the gradient method [45, Proposition 9]. This requires additional care.

Theorem 4.2 (Global convergence). *Let $f \in C_{\text{loc}}^{1,1}(\mathbb{R}^n)$ be semialgebraic with bounded continuous gradient trajectories. Let $\beta \in (-1, 1)$, $\gamma \in \mathbb{R}$, $\delta \geq 0$ and X_0 be a bounded subset of \mathbb{R}^n . There exist $\bar{\alpha}, c > 0$ such that for all $\alpha \in (0, \bar{\alpha}]$, there exists $c_\alpha > 0$ such that any sequence $x_{-1}, x_0, x_1, \dots \in \mathbb{R}^n$ generated by the momentum method (4.1) that satisfies $x_0 \in X_0$ and $\|x_0 - x_{-1}\| \leq \delta\alpha$ obeys*

$$\sum_{i=0}^{\infty} \|x_{i+1} - x_i\| \leq c \quad \text{and} \quad \min_{i=0, \dots, k} \|\nabla f(x_i)\| \leq \frac{c_\alpha}{k+1}, \quad \forall k \in \mathbb{N}. \quad (4.6)$$

Proof. Let $\beta \in (-1, 1)$, $\gamma \in \mathbb{R}$, $\delta_0 \geq 0$, and X_0 be a bounded subset of \mathbb{R}^n . Without loss of generality, we may assume that $X_0 \neq \emptyset$. We will show that there exists $\bar{\alpha} > 0$ such that $\sigma(X_0, \bar{\alpha}, \delta_0) < \infty$ where

$$\sigma(X_0, \bar{\alpha}, \delta_0) := \sup_{\substack{x \in (\mathbb{R}^n)^\mathbb{N} \\ \alpha \in (0, \bar{\alpha}]}} \sum_{k=0}^{\infty} \|x_{k+1} - x_k\| \quad (4.7a)$$

$$\text{s.t.} \quad \begin{cases} x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha \nabla f(x_k + \gamma(x_k - x_{k-1})), \quad \forall k \in \mathbb{N}, \\ x_0 \in X_0, \quad \|x_0 - x_{-1}\| \leq \delta_0 \alpha. \end{cases} \quad (4.7b)$$

Letting $c := \sigma(X_0, \bar{\alpha}, \delta_0)$, the convergence rate is easily deduced. Indeed, for any feasible point $((x_k)_{k \in \{-1\} \cup \mathbb{N}}, \alpha)$ of (4.7), we have $x_{-1}, x_0, x_1, \dots \in B(X_0, \max\{c, \delta_0\alpha\}) := X_0 + B(0, \max\{c, \delta_0\alpha\})$

and thus $\|\nabla f(x_k)\| \leq b_\alpha \|z_{k+1} - z_k\|$ for all $k \in \mathbb{N}$ for some constant $b_\alpha > 0$ by Lemma 4.5, where $z_k := (x_k, x_{k-1})$. Hence

$$\begin{aligned} \sum_{k=0}^{\infty} \|\nabla f(x_k)\| &\leq \sum_{k=0}^{\infty} b_\alpha \|z_{k+1} - z_k\| \\ &\leq b_\alpha \|x_0 - x_{-1}\| + 2b_\alpha \sum_{k=0}^{\infty} \|x_{k+1} - x_k\| \\ &\leq b_\alpha (\delta_0 \alpha + 2c) =: c_\alpha \end{aligned}$$

and

$$\min_{i=0, \dots, k} \|\nabla f(x_i)\| \leq \frac{1}{k+1} \sum_{i=0}^k \|\nabla f(x_i)\| \leq \frac{c_\alpha}{k+1}.$$

Let $\Phi : \mathbb{R}_+ \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ be the continuous gradient flow of f defined for all $(t, x_0) \in \mathbb{R}_+ \times \mathbb{R}^n$ by $\Phi(t, x_0) := x(t)$ where $x(\cdot)$ is the unique continuous gradient trajectory of f initialized at x_0 . Uniqueness follows from the Picard–Lindelöf theorem [116, Theorem 3.1 p. 12]. Let $\Phi_0 := \Phi(\mathbb{R}_+, X_0)$ and let C be the set of critical points of f in $\overline{\Phi_0}$. C is compact by Lemma 4.3 and [40, 2.1.5 Proposition p. 29]. Thus there exists $\epsilon > 0$ such that either $X_0 \subset C$ or $X_0 \setminus \mathring{B}(C, \epsilon/6) \neq \emptyset$ where $\mathring{B}(C, \epsilon/6) := C + \mathring{B}(0, \epsilon/6)$.

Indeed, either $X_0 \subset C$ or there exists $x \in X_0 \cap C^c$ where the complement C^c of C is open since C is closed. Thus there exists $\epsilon > 0$ such that $B(x, \epsilon/6) \subset C^c$. Thus $x \notin \mathring{B}(C, \epsilon/6)$ (otherwise there exists $x' \in C$ such that $\|x - x'\| < \epsilon/6$, i.e., $C \ni x' \in B(x, \epsilon/6) \subset C^c$) and $x \in X_0 \setminus \mathring{B}(C, \epsilon/6)$.

By Fact 4.2, there exists $\delta_1 > \delta_0$ such that for all $\alpha > 0$, $K \in \mathbb{N}$, and sequence $(x_k)_{k \in \{-1\} \cup \mathbb{N}}$ generated by the momentum method (4.1) for which $x_{-1}, \dots, x_{K-1} \in B(\overline{\Phi_0}, \epsilon) := \overline{\Phi_0} + B(0, \epsilon)$ and $\|x_0 - x_{-1}\| \leq \delta_0 \alpha$, we have $\|x_k - x_{k-1}\| \leq \delta_1 \alpha$ for $k = 0, \dots, K$. By Lemma 4.1, there exist $\tilde{\alpha}, \eta > 0$, $\kappa \geq \delta_1$, and a diffeomorphism $\psi : [0, \infty) \rightarrow [0, \infty)$ such that for all $K \in \mathbb{N} \setminus \{0\}$, $\alpha \in (0, \tilde{\alpha}]$, and sequence $(x_k)_{k \in \{-1\} \cup \mathbb{N}}$ generated by the momentum method (4.1) for which $x_{-1}, \dots, x_{K-1} \in B(\overline{\Phi_0}, \epsilon)$ and $\|x_0 - x_{-1}\| \leq \delta_1 \alpha$, we have

$$\sum_{k=0}^{K-1} \|x_{k+1} - x_k\| \leq \psi(f(x_0) - f(x_{K-1}) + \eta\alpha) + \kappa\alpha. \quad (4.8)$$

Since f is continuous, there exists $\xi \in (0, \epsilon/2)$ such that

$$f(x) - \max_C f \leq \frac{1}{4} \psi^{-1} \left(\frac{\epsilon}{3} \right), \quad \forall x \in B(C, \xi). \quad (4.9)$$

Let $L > 0$ be a Lipschitz constant of $\bar{\beta}f$ on the convex hull of $B(\bar{\Phi}_0, \epsilon)$ and let

$$\hat{\alpha} := \min \left\{ \bar{\alpha}, \frac{\xi}{3L}, \frac{\epsilon}{6\kappa}, \frac{\psi^{-1}(\epsilon/3)}{4\eta} \right\} > 0 \quad (4.10)$$

where $\bar{\beta} := 1/(1-\beta)$. If $X_0 \subset C$, then let $\bar{\alpha} := \hat{\alpha}$ and $k^* := 0$. Otherwise $X_0 \setminus \mathring{B}(C, \epsilon/6) \neq \emptyset$. Since $\bar{\beta}\nabla f$ is continuous, its norm attains its infimum ν on the non-empty compact set $\bar{\Phi}_0 \setminus \mathring{B}(C, \xi/3)$. It is non-empty because $\bar{\Phi}_0 \setminus \mathring{B}(C, \xi/3) \supset X_0 \setminus \mathring{B}(C, \epsilon/6) \neq \emptyset$ and $\xi < \epsilon/2$. If $\nu = 0$, then there exists $x^* \in \bar{\Phi}_0 \setminus \mathring{B}(C, \xi/3)$ such that $\|\nabla f(x^*)\| = 0$. Then $x^* \in C \setminus \mathring{B}(C, \xi/3)$, which is a contradiction.

We thus have $\nu > 0$. Hence we may define $T := 2\sigma(X_0)/\nu$ where

$$\sigma(X_0) = \sup_{x \in C^1(\mathbb{R}_+, \mathbb{R}^n)} \int_0^\infty \|x'(t)\| dt$$

subject to $\begin{cases} x'(t) = -\bar{\beta}\nabla f(x(t)), \quad \forall t > 0, \\ x(0) \in X_0, \end{cases}$

is finite by Lemma 4.3. The factor $\bar{\beta} > 0$ does not change the optimal value because $x(\cdot)$ is a feasible point of the above problem if and only if $x(\cdot/\bar{\beta})$ is a feasible point of the problem in Lemma 2.7 and $\int_0^\infty \|x'(t/\bar{\beta})/\bar{\beta}\| dt = \int_0^\infty \|x'(t)\| dt$. Note that $\sigma(X_0) > 0$ and thus $T > 0$ because $X_0 \not\subset C$. In addition, since f is semialgebraic and has bounded continuous gradient trajectories, it is lower bounded by its smallest critical value¹. By Lemma 4.2, there exists $\bar{\alpha} \in (0, \hat{\alpha}]$ such that for any feasible point $((x_k)_{k \in \{-1\} \cup \mathbb{N}}, \alpha)$ of (4.7), there exists an absolutely continuous function

¹Indeed, assume to the contrary that there exists $x_0 \in \mathbb{R}^n$ such that $f(x_0)$ is less than the smallest critical value of f . The continuous gradient trajectory initialized at x_0 converges to a critical point x^* since it is bounded. This limit satisfies $f(x_0) \geq f(x^*)$, yielding a contradiction.

$x : [0, T] \rightarrow \mathbb{R}^n$ such that

$$x'(t) = -\bar{\beta}\nabla f(x(t)), \quad \forall t \in (0, T), \quad x(0) \in X_0, \quad (4.11)$$

for which $\|x_k - x(k\alpha)\| \leq \xi/3$ for $k = 0, \dots, \lfloor T/\alpha \rfloor$. Now suppose that $\|x'(t)\| \geq 2\sigma(X_0)/T$ for all $t \in (0, T)$. Then we obtain the following contradiction

$$\sigma(X_0) < T \frac{2\sigma(X_0)}{T} \leq \int_0^T \|x'(t)\| dt \leq \int_0^\infty \|x'(t)\| dt \leq \sigma(X_0).$$

Hence, there exists $t^* \in (0, T)$ such that $\|x'(t^*)\| = \|\bar{\beta}\nabla f(x(t^*))\| < 2\sigma(X_0)/T = 2\sigma(X_0)/(2\sigma(X_0)/\nu) = \nu$. Since $x(t^*) \in \bar{\Phi}_0$ and the infimum of the norm of $\bar{\beta}\nabla f$ on $\bar{\Phi}_0 \setminus \mathring{B}(C, \xi/3)$ is equal to ν , it must be that $x(t^*) \in \mathring{B}(C, \xi/3)$. Hence there exists $x^* \in C$ such that $\|x(t^*) - x^*\| \leq \xi/3$.

Since $\alpha \leq \hat{\alpha} \leq \xi/(3L)$, there exists $k^* \in \mathbb{N}$ such that $t_{k^*} := k^*\alpha \in [t^* - \xi/(3L), t^*]$. Thus $\|x_{k^*} - x^*\| \leq \|x_{k^*} - x(t_{k^*})\| + \|x(t_{k^*}) - x(t^*)\| + \|x(t^*) - x^*\| \leq \xi/3 + L|t_{k^*} - t^*| + \xi/3 \leq \xi/3 + \xi/3 + \xi/3 = \xi$. To obtain the second inequality, we used the fact that for all $t \geq 0$, we have $\|x'(t)\| = \|\bar{\beta}\nabla f(x(t))\| \leq L$ since $x(t) \in B(\bar{\Phi}_0, \epsilon)$.

Above, we defined $\bar{\alpha} \in (0, \hat{\alpha}]$ if $X_0 \subset C$ or $X_0 \not\subset C$ (in which case $X_0 \setminus \mathring{B}(C, \epsilon/6) \neq \emptyset$). We now consider a feasible point $((x_k)_{k \in \{-1\} \cup \mathbb{N}}, \alpha)$ of (4.7) regardless of whether $X_0 \subset C$. Based on the above and the fact that $X_0 \neq \emptyset$, there exists $k^* \in \mathbb{N}$ such that $x_{k^*} \in B(C, \xi)$. If $K := \inf\{k \geq$

$k^* : x_k \notin B(C, \epsilon)\} < \infty$, then

$$\psi^{-1}\left(\frac{\epsilon}{3}\right) = \psi^{-1}\left(\frac{1}{2}\epsilon - \kappa\frac{\epsilon}{6\kappa}\right) \quad (4.12a)$$

$$\leq \psi^{-1}\left((\epsilon - \xi) - \kappa\hat{\alpha}\right) \quad (4.12b)$$

$$\leq \psi^{-1}\left((\|x_K - x^*\| - \|x_{k^*} - x^*\|) - \kappa\bar{\alpha}\right) \quad (4.12c)$$

$$\leq \psi^{-1}\left(\|x_K - x_{k^*}\| - \kappa\alpha\right) \quad (4.12d)$$

$$\leq \psi^{-1}\left(\sum_{k=k^*}^{K-1} \|x_{k+1} - x_k\| - \kappa\alpha\right) \quad (4.12e)$$

$$\leq f(x_{k^*}) - f(x_{K-1}) + \eta\alpha \quad (4.12f)$$

$$\leq \max_C f + \frac{1}{4}\psi^{-1}\left(\frac{\epsilon}{3}\right) - f(x_{K-1}) + \eta\hat{\alpha} \quad (4.12g)$$

$$\leq \max_C f + \frac{1}{2}\psi^{-1}\left(\frac{\epsilon}{3}\right) - f(x_{K-1}) \quad (4.12h)$$

Above, the arguments of ψ^{-1} in (4.12a) are equal. (4.12b) through (4.12e) rely on the fact that ψ^{-1} is an increasing function. (4.12b) is due to $\xi < \epsilon/2$. (4.12c) holds because $x_K \notin B(C, \epsilon)$, $x^* \in C$, and $x_{k^*} \in B(x^*, \xi)$. (4.12d) and (4.12e) are consequences of the triangular inequality. (4.12f) is due to the length formula (4.8) and the fact that $x_{k^*}, \dots, x_{K-1} \in B(C, \epsilon) \subset B(\bar{\Phi}_0, \epsilon)$. (4.12g) is due to $x_{k^*} \in B(x^*, \xi)$ and (4.9). Finally, (4.12h) is due to $\hat{\alpha} \leq \psi^{-1}(\epsilon/3)/(4\eta)$ by definition of $\hat{\alpha}$ in (4.10). We remark that $K \geq k^* + 2$ since $\|x_{k^*+1} - x^*\| \leq \|x_{k^*+1} - x_{k^*}\| + \|x_{k^*} - x^*\| \leq \delta_1\alpha + \xi \leq \delta_1\epsilon/(6\kappa) + \epsilon/2 \leq \epsilon/6 + \epsilon/2 < \epsilon$. It also holds that $x_{-1}, \dots, x_{K-2} \in B(\bar{\Phi}_0, \epsilon)$, x_{K-1} belongs to

$$X_1 := B(C, \epsilon) \cap \left\{x \in \mathbb{R}^n : f(x) \leq \max_C f - \frac{1}{2}\psi^{-1}\left(\frac{\epsilon}{3}\right)\right\},$$

and $\|x_{K-1} - x_{K-2}\| \leq \delta_1\alpha$. Thus, by the length formula (4.8) and the definition of $\sigma(\cdot, \cdot, \cdot)$ in (4.7)

we have

$$\begin{aligned} \sum_{k=0}^{\infty} \|x_{k+1} - x_k\| &= \sum_{k=0}^{K-2} \|x_{k+1} - x_k\| + \sum_{k=K-1}^{\infty} \|x_{k+1} - x_k\| \\ &\leq \psi \left(\sup_{X_0} f - \min_{B(\bar{\Phi}_0, \epsilon)} f + \eta \bar{\alpha} \right) + \kappa \bar{\alpha} + \max\{0, \sigma(X_1, \bar{\alpha}, \delta_1)\}. \end{aligned}$$

Note that the inequality still holds if $K = \infty$, since in that case (4.8) implies

$$\sum_{k=0}^{\infty} \|x_{k+1} - x_k\| \leq \psi \left(\sup_{X_0} f - \min_{B(\bar{\Phi}_0, \epsilon)} f + \eta \bar{\alpha} \right) + \kappa \bar{\alpha}.$$

Hence

$$\sigma(X_0, \bar{\alpha}, \delta_0) \leq \psi \left(\sup_{X_0} f - \min_{B(\bar{\Phi}_0, \epsilon)} f + \eta \bar{\alpha} \right) + \kappa \bar{\alpha} + \max\{0, \sigma(X_1, \bar{\alpha}, \delta_1)\}.$$

It now suffices to replace X_0 by X_1 , δ_0 by δ_1 , and repeat the entire proof. Since $f(\Phi(t, x_1)) \leq f(\Phi(0, x_1)) \leq \max_C f - \psi^{-1}(\epsilon/3)/2 < \max_C f$ for all $t \geq 0$ and $x_1 \in X_1$, the maximal critical value of f in $\overline{\Phi(\mathbb{R}_+, X_1)}$ is less than the maximal critical value of f in $\overline{\Phi(\mathbb{R}_+, X_0)}$. By the semialgebraic Morse-Sard theorem (Lemma 1.1), f has finitely many critical values. Thus, it is eventually the case that one of the sets X_0, X_1, \dots is empty. In order to conclude, one simply needs to choose an upper bound on the step sizes $\bar{\alpha}'$ corresponding to X_1 that is less than or equal to the upper bound $\bar{\alpha}$ used for X_0 . $\sigma(X_0, \cdot, \delta_0)$ is finite when evaluated at the last upper bound thus obtained. Indeed, the recursive formula above still holds if we replace $\bar{\alpha}$ by any $\alpha \in (0, \bar{\alpha}]$. In particular, we may take $\alpha := \bar{\alpha}'$. \square

The inequalities in (4.6) imply that the iterates converge to a critical point of f (i.e., a point $x^* \in \mathbb{R}^n$ such that $\nabla f(x^*) = 0$). The previously known global convergence rate of the momentum method is $O(1/\sqrt{k})$ for coercive differentiable functions with a Lipschitz continuous gradient [106, Theorem 4.14] if $x_{-1} = x_0$ and $\gamma = 0$. Without the coercivity assumption, $O(1/\sqrt{k})$ is also the rate of a modified version of the momentum method which does not capture the heavy ball method and Nesterov's accelerated gradient method as special cases [117, Corollary 1]. Our third and final

convergence result gives sufficient conditions for convergence to a local minimizer.

Theorem 4.3 (Convergence to local minimizers). *Let $f \in C^2(\mathbb{R}^n)$ be semialgebraic with bounded continuous gradient trajectories. Let $\beta \in (-1, 1) \setminus \{0\}$, $\gamma \in \mathbb{R}$, and $\delta \geq 0$. If the Hessian of f has a negative eigenvalue at all critical points of f that are not local minimizers, then for any bounded subset X_0 of \mathbb{R}^n , there exists $\bar{\alpha} > 0$ such that, for all $\alpha \in (0, \bar{\alpha}]$ and for almost every $(x_{-1}, x_0) \in \mathbb{R}^n \times X_0$, any sequence $x_{-1}, x_0, x_1, \dots \in \mathbb{R}^n$ generated by the momentum method (4.1) that satisfies $\|x_0 - x_{-1}\| \leq \delta\alpha$ converges to a local minimizer of f .*

In practice, if X_0 has positive measure and $\delta > 0$, Theorem 4.3 means that one can generate x_0 uniformly at random in X_0 and generate x_{-1} uniformly at random in the ball of radius $\delta\alpha$ centered at x_0 in order to guarantee convergence to a local minimizer almost surely. In contrast to the gradient method [45, Corollary 1], we need to assume that the Hessian of f has a negative eigenvalue at local maxima of f . Indeed, the function values are not necessarily decreasing along the iterates. While the proof of Theorem 4.3 crucially depends on the length bound in Theorem 4.2, it mostly requires extending well-known arguments regarding the center and stable manifolds theorem [118, Theorem III.7]. For this reason, we defer its proof to Section 4.4.

4.2 Proof of the length formula

Given $\lambda > 0$, consider the following Lyapunov function proposed by Zavriev and Kostyuk [109]:

$$\begin{aligned} H_\lambda : \mathbb{R}^n \times \mathbb{R}^n &\longrightarrow \mathbb{R} \\ (x, y) &\longmapsto f(x) + \lambda\|x - y\|^2. \end{aligned}$$

For certain values of λ , it is known to be monotonic along the iterates if $\gamma = 0$ [109, Lemma 1] [106, Proposition 4.7 (a)] or $\beta = \gamma$ [110, Lemma 3.1 (ii)] [111, Lemma 3.2], but not for general $\beta \in (-1, 1)$ and $\gamma \in \mathbb{R}$. This justifies the need for Lemma 4.4 in which it will be convenient to

rewrite the update rule of the momentum method (4.1) as

$$y_k^\beta = x_k + \beta(x_k - x_{k-1}), \quad (4.13a)$$

$$y_k^\gamma = x_k + \gamma(x_k - x_{k-1}), \quad (4.13b)$$

$$x_{k+1} = y_k^\beta - \alpha \nabla f(y_k^\gamma), \quad (4.13c)$$

for all $k \in \mathbb{N}$. Likewise, bounds on the norm of the gradient of the objective function and the Lyapunov function are only known if $\gamma = 0$ [106, Theorem 4.9] or $\beta = \gamma$ [110, Equation (3.22)] [111, Lemma 3.3], which calls for Lemma 4.5.

Lemma 4.4. *Let $f \in C_{\text{loc}}^{1,1}(\mathbb{R}^n)$, $X \subset \mathbb{R}^n$ be bounded, $\beta \in (-1, 1)$, and $\gamma \in \mathbb{R}$. There exists $\bar{\alpha} > 0$ such that for all $\alpha \in (0, \bar{\alpha}]$, there exist $\lambda^+ > \lambda^- > 0$ such that for all $\lambda \in (\lambda^-, \lambda^+)$, there exists $c_1 > 0$ such that for all $K \in \mathbb{N}$, if $x_{-1}, \dots, x_{K+1} \in X$ are iterates of the momentum method (4.1), then for $k = 0, \dots, K$ we have*

$$H_\lambda(x_{k+1}, x_k) \leq H_\lambda(x_k, x_{k-1}) - c_1(\|x_{k+1} - x_k\|^2 + \|x_k - x_{k-1}\|^2).$$

If $M > 0$ is a Lipschitz constant of ∇f on $S + \max\{|\beta|, |\gamma|\}(S - S)$ where S is the convex hull of X , then one may take

$$\bar{\alpha} := \min \left\{ \frac{1}{M}, \frac{1 - \beta^2}{2(\beta^2 + 2|\beta - \gamma|M)} \right\}, \quad (4.14a)$$

$$\lambda^- := \left(\frac{1}{2\alpha} + \frac{M}{2} \right) \beta^2 + \frac{|\beta - \gamma|M}{2}, \quad \lambda^+ := \frac{1}{2\alpha} - \frac{|\beta - \gamma|M}{2}, \quad (4.14b)$$

$$\text{and } c_1 := \min\{\lambda - \lambda^-, \lambda^+ - \lambda\}. \quad (4.14c)$$

Proof. Consider $\bar{\alpha}$ as defined in (4.14a) and let $\alpha \in (0, \bar{\alpha}]$. Given $K \in \mathbb{N}$, let $x_{-1}, \dots, x_{K+1} \in X$ be

iterates generated by the momentum method (4.1). A bound on the Taylor expansion of f yields

$$f(x_{k+1}) \leq f(y_k^\beta) + \langle \nabla f(y_k^\beta), x_{k+1} - y_k^\beta \rangle + \frac{M}{2} \|x_{k+1} - y_k^\beta\|^2, \quad (4.15a)$$

$$f(x_k) \geq f(y_k^\beta) + \langle \nabla f(y_k^\beta), x_k - y_k^\beta \rangle - \frac{M}{2} \|x_k - y_k^\beta\|^2, \quad (4.15b)$$

where $k \in \{0, \dots, K\}$. Subtracting (4.15b) from (4.15a) yields

$$f(x_{k+1}) - f(x_k) \leq \langle \nabla f(y_k^\beta), x_{k+1} - x_k \rangle + \frac{M}{2} (\|x_{k+1} - y_k^\beta\|^2 + \|x_k - y_k^\beta\|^2) \quad (4.16a)$$

$$= \langle \nabla f(y_k^\beta) - \nabla f(y_k^\gamma), x_{k+1} - x_k \rangle + \langle \nabla f(y_k^\gamma), x_{k+1} - x_k \rangle \quad (4.16b)$$

$$+ \frac{M}{2} (\|x_{k+1} - y_k^\beta\|^2 + \|x_k - y_k^\beta\|^2) \quad (4.16c)$$

$$= \langle \nabla f(y_k^\beta) - \nabla f(y_k^\gamma), x_{k+1} - x_k \rangle + \frac{1}{\alpha} \langle x_{k+1} - y_k^\beta, x_k - x_{k+1} \rangle \quad (4.16d)$$

$$+ \frac{M}{2} (\|x_{k+1} - y_k^\beta\|^2 + \|x_k - y_k^\beta\|^2). \quad (4.16e)$$

By the Cauchy-Schwarz and AM-GM inequalities, we have

$$\langle \nabla f(y_k^\beta) - \nabla f(y_k^\gamma), x_{k+1} - x_k \rangle \leq \|\nabla f(y_k^\beta) - \nabla f(y_k^\gamma)\| \|x_{k+1} - x_k\| \quad (4.17a)$$

$$\leq M \|y_k^\beta - y_k^\gamma\| \|x_{k+1} - x_k\| \quad (4.17b)$$

$$= M |\beta - \gamma| \|x_k - x_{k-1}\| \|x_{k+1} - x_k\| \quad (4.17c)$$

$$\leq \frac{|\beta - \gamma| M}{2} (\|x_k - x_{k-1}\|^2 + \|x_{k+1} - x_k\|^2). \quad (4.17d)$$

By the cosine rule, for any $a, b, c \in \mathbb{R}^n$, it holds that

$$\langle a - b, c - a \rangle = \frac{1}{2} (\|b - c\|^2 - \|a - b\|^2 - \|c - a\|^2).$$

By letting $a := x_{k+1}$, $b := y_k^\beta$ and $c := x_k$, we have

$$\langle x_{k+1} - y_k^\beta, x_k - x_{k+1} \rangle = \frac{1}{2} (\|y_k^\beta - x_k\|^2 - \|x_{k+1} - y_k^\beta\|^2 - \|x_k - x_{k+1}\|^2). \quad (4.18)$$

Combining (4.16), (4.17) and (4.18), we find that

$$f(x_{k+1}) - f(x_k) \leq -\left(\frac{1}{2\alpha} - \frac{M}{2}\right) \|y_k^\beta - x_{k+1}\|^2 \quad (4.19a)$$

$$-\left(\frac{1}{2\alpha} - \frac{|\beta - \gamma|M}{2}\right) \|x_{k+1} - x_k\|^2 \quad (4.19b)$$

$$+\left[\left(\frac{1}{2\alpha} + \frac{M}{2}\right)\beta^2 + \frac{|\beta - \gamma|M}{2}\right] \|x_k - x_{k-1}\|^2. \quad (4.19c)$$

Let $\lambda \in (\lambda^-, \lambda^+)$ where λ^- and λ^+ are defined in (4.14b). Note that $\lambda^- < \lambda^+$ due to the fact that $\alpha \in (0, \bar{\alpha}]$. By definition of H_λ , it readily follows that

$$\begin{aligned} H_\lambda(x_{k+1}, x_k) - H_\lambda(x_k, x_{k-1}) &\leq -\left(\frac{1}{2\alpha} - \frac{M}{2}\right) \|y_k^\beta - x_{k+1}\|^2 \\ &\quad -\left(\frac{1}{2\alpha} - \frac{|\beta - \gamma|M}{2} - \lambda\right) \|x_{k+1} - x_k\|^2 \\ &\quad -\left[\lambda - \left(\frac{1}{2\alpha} + \frac{M}{2}\right)\beta^2 - \frac{|\beta - \gamma|M}{2}\right] \|x_k - x_{k-1}\|^2. \end{aligned}$$

The desired inequality is guaranteed by taking c_1 as defined in (4.14c). \square

Lemma 4.5. *Let $f \in C_{\text{loc}}^{1,1}(\mathbb{R}^n)$, $X \subset \mathbb{R}^n$ be bounded, and $\beta, \gamma \in \mathbb{R}$. For all $\alpha, \lambda > 0$, there exist $c_2 > 0$ such that for all $K \in \mathbb{N}$, if $x_{-1}, \dots, x_{K+1} \in X$ are iterates of the momentum method (4.1), then*

$$\max\{\|\nabla H_\lambda(z_k)\|, \|\nabla H_\lambda(z_{k+1})\|\} \leq c_2 \|z_{k+1} - z_k\|,$$

for $k = 0, \dots, K$ where $z_k := (x_k, x_{k-1}) \in \mathbb{R}^{2n}$. If $M > 0$ is a Lipschitz constant of ∇f on $S + \max\{|\beta|, |\gamma|\}(S - S)$ where S is the convex hull of X , then one may take

$$c_2 := \sqrt{2} \max\left\{\frac{1}{\alpha}, \frac{|\beta|}{\alpha} + M(|\gamma| + 1) + 4\lambda\right\}. \quad (4.20)$$

Proof. Using Fact 4.1, for $k = 0, \dots, K$ we have

$$\begin{aligned}
\|\nabla H_\lambda(z_k)\| &\leq \|\nabla f(x_k) + 2\lambda(x_k - x_{k-1})\| + \|2\lambda(x_k - x_{k-1})\| \\
&\leq \|\nabla f(x_k)\| + 4\lambda\|x_k - x_{k-1}\| \\
&\leq \sqrt{2} \max\{1/\alpha, |\beta|/\alpha + M|\gamma| + 4\lambda\}\|z_{k+1} - z_k\|.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\|\nabla H_\lambda(z_{k+1})\| &\leq \|\nabla f(x_{k+1}) + 2\lambda(x_{k+1} - x_k)\| + \|2\lambda(x_{k+1} - x_k)\| \\
&\leq \|\nabla f(x_{k+1})\| + 4\lambda\|x_{k+1} - x_k\| \\
&\leq \|\nabla f(x_k)\| + \|\nabla f(x_{k+1}) - \nabla f(x_k)\| + 4\lambda\|x_{k+1} - x_k\| \\
&\leq \|\nabla f(x_k)\| + (4\lambda + M)\|x_{k+1} - x_k\| \\
&\leq \sqrt{2} \max\{1/\alpha, |\beta|/\alpha + M|\gamma| + 4\lambda + M\}\|z_{k+1} - z_k\|. \quad \square
\end{aligned}$$

We say that $\psi : [0, \infty) \rightarrow [0, \infty)$ in Proposition 1.1 is a desingularizing function of f on X . The uniform Kurdyka-Łojasiewicz inequality (1.2) enables one to relate the length of the iterates of the gradient method in any bounded region to the function variation [45, Proposition 8]. If one uses the Kurdyka-Łojasiewicz inequality (1.1) instead, then the function values evaluated at the iterates would be restricted to a potentially small range around a critical value. If one uses the uniformized KL property [119, Lemma 6], then the iterates would need to lie in a uniform neighborhood of a compact subset of the critical points of f where f is constant.

In order to prove [45, Proposition 8], one uses the fact that the objective function is a Lyapunov function for all sufficiently small step sizes in the gradient method. However, in the momentum method the Lyapunov function depends on the step size, as can be seen in Lemma 4.4. The main challenge that we thus face is to obtain an upper bound on the length of the iterates that is independent of the step size. Otherwise, it could blow up as the step size gets small. Such is the object of the following results.

Proposition 4.1 takes a first step by showing that the length is bounded by a constant times a desingularizing function evaluated at the Lyapunov function variation plus a constant multiple of the step size. Both constants are independent of the step size. Proposition 4.2 ensures that the desingularizing function no longer depends on the step size. Finally, Lemma 4.1 gets rid of the dependence on the step size in the argument of the desingularizing function.

Proposition 4.1 below generalizes [45, Proposition 8] from the gradient method to the momentum method. While in the gradient method we have $c_3 = 2$ in (4.21), in the momentum method obtaining an expression for c_3 that does not depend on the step size requires some care. We will use the following simple lemma.

Lemma 4.6. *If $h : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is concave, $h(0) = 0$, and $a, b \geq 0$, then $|h(b) - h(a)| \leq h(|b - a|)$.*

Proof. We first show that h is increasing. Since h is concave, for any $0 \leq x < y < z$, we have

$$h(z) \leq \frac{h(y) - h(x)}{y - x}(z - y) + h(y).$$

If $(h(y) - h(x))(y - x) < 0$, then we obtain the contradiction $0 \leq \limsup_{z \rightarrow \infty} h(z) = -\infty$, establishing that h is increasing. It thus suffices to show that $h(b) - h(a) \leq h(b - a)$ for any $0 \leq a \leq b$. Since h is concave, we have

$$\frac{a}{b}h(0) + \frac{b-a}{b}h(b) \leq h(b-a) \quad \text{and} \quad \frac{b-a}{b}h(0) + \frac{a}{b}h(b) \leq h(a).$$

Summing these two inequalities and using the fact that $h(0) = 0$ yields the desired result. \square

Proposition 4.1. *Let $f \in C_{\text{loc}}^{1,1}(\mathbb{R}^n)$ be semialgebraic, $X \subset \mathbb{R}^n$ be bounded, $\beta \in (-1, 1)$, $\gamma \in \mathbb{R}$, $\delta \geq 0$, and $m \in \mathbb{N}^*$ be an upper bound on the number of critical values of f in \bar{X} . There exist $\bar{\alpha}, c_3, \zeta > 0$ such that for all $\alpha \in (0, \bar{\alpha}]$, there exists $\lambda > 0$ such that, for any desingularizing function ψ_λ of H_λ on $X \times X$ and for all $K \in \mathbb{N}$, if $(x_k)_{k \in \{-1\} \cup \mathbb{N}}$ are iterates of the momentum*

method (4.1) for which $x_{-1}, \dots, x_K \in X$ and $\|x_0 - x_{-1}\| \leq \delta\alpha$, then

$$\frac{1}{2m} \sum_{k=0}^K \|z_{k+1} - z_k\| \leq c_3 \psi_\lambda \left(\frac{H_\lambda(z_0) - H_\lambda(z_K)}{2m} \right) + \zeta \alpha \quad (4.21)$$

and $H_\lambda(z_0) \geq \dots \geq H_\lambda(z_K)$ where $z_k := (x_k, x_{k-1}) \in \mathbb{R}^{2n}$. If $L > 0$ and $M > 0$ are Lipschitz constants of $\bar{\beta}f$ and ∇f respectively on $S + \max\{|\beta|, |\gamma|\}(S - S)$ where S is the convex hull of X and $\bar{\beta} := (1 - \beta)^{-1}$, then one may take the same $\bar{\alpha}$ as in (4.14a),

$$c_3 := \frac{8\sqrt{2}(2 + |\gamma| + 3|\beta|)}{1 - \beta^2}, \quad \zeta := 2\sqrt{2}(\delta + L), \quad \text{and } \lambda := \frac{\beta^2 + 1 + M\beta^2\alpha}{4\alpha}. \quad (4.22)$$

Proof. Consider $\bar{\alpha}$ and c_3 as defined in (4.14a) and (4.22) respectively. Given $\alpha \in (0, \bar{\alpha}]$, let $\lambda \in (\lambda^-, \lambda^+)$ where λ^- and λ^+ are defined in (4.14b). Let ψ_λ be a desingularizing function of H_λ on $X \times X$. By Lemma 4.4 and Lemma 4.5, for $k = 0, \dots, K - 1$ we have

$$H_\lambda(z_{k+1}) - H_\lambda(z_k) \leq -c_1 \|z_{k+1} - z_k\|^2 \leq -\frac{c_1}{c_2} \|\nabla H_\lambda(z_k)\| \|z_{k+1} - z_k\| \quad (4.23)$$

and

$$H_\lambda(z_{k+1}) - H_\lambda(z_k) \leq -c_1 \|z_{k+1} - z_k\|^2 \leq -\frac{c_1}{c_2} \|\nabla H_\lambda(z_{k+1})\| \|z_{k+1} - z_k\|. \quad (4.24)$$

Since $\nabla H_\lambda(x, y) = (\nabla f(x) + 2\lambda(x - y), 2\lambda(y - x))^\top$, the critical values of f in \bar{X} are the same as those of H_λ in $\bar{X} \times \bar{X}$. We let V denote this set of critical values if they exist, otherwise $V := \{0\}$.

Assume that $[H_\lambda(z_K), H_\lambda(z_0))$ excludes the elements of V and the averages of any two consecutive elements of V .² If $H_\lambda(z_1) = H_\lambda(z_0)$, then $z_1 = z_0$ by (4.23). Thus $\nabla f(x_0) = 0$ and $z_K = \dots = z_0$ by induction. Otherwise, we have that $H_\lambda(z_1) < H_\lambda(z_0)$. With $\tilde{H}_\lambda := d(H_\lambda, V)$, we thus have $0 \notin \partial \tilde{H}_\lambda(z_k)$ and $1 \leq \|\nabla(\psi_\lambda \circ \tilde{H}_\lambda)(z_k)\| = \psi'_\lambda(\tilde{H}_\lambda(z_k)) \|\nabla \tilde{H}_\lambda(z_k)\|$ for $k = 1, \dots, K$ by the uniform Kurdyka-Łojasiewicz inequality (1.2). Let $k \in \{1, \dots, K - 1\}$. If $\tilde{H}_\lambda(z_k) \geq \tilde{H}_\lambda(z_{k+1})$,

²The point of excluding elements in V and the averages of two consecutive elements in V is to guarantee that there is a unique closest element in V that works for all $H_\lambda(z_K), \dots, H_\lambda(z_0)$ and this element is either greater than or equal to all of them or less than all of them.

multiplying (4.23) by $\psi'_\lambda(\tilde{H}_\lambda(z_k))$ and using concavity of ψ_λ , we find that

$$\begin{aligned}\|z_{k+1} - z_k\| &\leq \frac{c_2}{c_1} \psi'_\lambda(\tilde{H}_\lambda(z_k))(H_\lambda(z_k) - H_\lambda(z_{k+1})) \\ &= \frac{c_2}{c_1} \psi'_\lambda(\tilde{H}_\lambda(z_k))(\tilde{H}_\lambda(z_k) - \tilde{H}_\lambda(z_{k+1})) \\ &\leq \frac{c_2}{c_1} (\psi_\lambda(\tilde{H}_\lambda(z_k)) - \psi_\lambda(\tilde{H}_\lambda(z_{k+1}))).\end{aligned}$$

If $\tilde{H}_\lambda(z_k) \leq \tilde{H}_\lambda(z_{k+1})$, multiplying (4.24) by $\psi'_\lambda(\tilde{H}_\lambda(z_{k+1}))$ and using concavity of ψ_λ , we find that

$$\begin{aligned}\|z_{k+1} - z_k\| &\leq \frac{c_2}{c_1} \psi'_\lambda(\tilde{H}_\lambda(z_{k+1}))(H_\lambda(z_k) - H_\lambda(z_{k+1})) \\ &= \frac{c_2}{c_1} \psi'_\lambda(\tilde{H}_\lambda(z_{k+1}))(\tilde{H}_\lambda(z_{k+1}) - \tilde{H}_\lambda(z_k)) \\ &\leq \frac{c_2}{c_1} (\psi_\lambda(\tilde{H}_\lambda(z_{k+1})) - \psi_\lambda(\tilde{H}_\lambda(z_k))).\end{aligned}$$

As a result,

$$\|z_{k+1} - z_k\| \leq \frac{c_2}{c_1} |\psi_\lambda(\tilde{H}_\lambda(z_k)) - \psi_\lambda(\tilde{H}_\lambda(z_{k+1}))|, \quad k = 1, \dots, K-1.$$

We obtain the telescoping sum

$$\sum_{k=0}^K \|z_{k+1} - z_k\| \leq \|z_1 - z_0\| + \sum_{k=1}^{K-1} \frac{c_2}{c_1} \left| \psi_\lambda(\tilde{H}_\lambda(z_k)) - \psi_\lambda(\tilde{H}_\lambda(z_{k+1})) \right| + \|z_{K+1} - z_K\| \quad (4.25a)$$

$$= \|z_1 - z_0\| + \frac{c_2}{c_1} \left| \psi_\lambda(\tilde{H}_\lambda(z_0)) - \psi_\lambda(\tilde{H}_\lambda(z_K)) \right| + \|z_{K+1} - z_K\| \quad (4.25b)$$

$$\leq \sqrt{2}(\delta + L)\alpha + \frac{c_2}{c_1} \left(\psi_\lambda \left(\left| \tilde{H}_\lambda(z_0) - \tilde{H}_\lambda(z_K) \right| \right) - \psi_\lambda(0) \right) + \sqrt{2}(\delta + L)\alpha \quad (4.25c)$$

$$= \frac{c_2}{c_1} \psi_\lambda(H_\lambda(z_0) - H_\lambda(z_K)) + \zeta \alpha \quad (4.25d)$$

where ζ is defined in (4.22). Above, (4.25b) and (4.25d) are due to the monotonicity of $\tilde{H}_\lambda(z_0), \dots, \tilde{H}_\lambda(z_K)$.

We use Lemma 4.6 and Fact 4.2 to obtain (4.25c).

We next consider the general case where

$$[H_\lambda(z_K), H_\lambda(z_{K_{p+1}})] \cup \dots \cup [H_\lambda(z_{K_2}), H_\lambda(z_{K_1+1})] \cup [H_\lambda(z_{K_1}), H_\lambda(z_0)]$$

excludes the elements of V and the averages of any two consecutive elements of V . For notational convenience, let $K_0 := -1$ and $K_{p+1} := K$. Since $p \leq 2m - 1$, we have

$$\sum_{k=0}^K \|z_{k+1} - z_k\| = \sum_{i=0}^p \sum_{k=K_i+1}^{K_{i+1}} \|z_{k+1} - z_k\| \quad (4.26a)$$

$$\leq \sum_{j=0}^p \left(\frac{c_2}{c_1} \psi_\lambda(H_\lambda(z_{K_{i+1}}) - H_\lambda(z_{K_i+1}) + \zeta\alpha) \right) \quad (4.26b)$$

$$\leq \frac{c_2}{c_1} \sum_{i=0}^p \psi_\lambda(H_\lambda(z_{K_{i+1}}) - H_\lambda(z_{K_i+1})) + (p+1)\zeta\alpha \quad (4.26c)$$

$$\leq \frac{c_2}{c_1} (p+1) \psi_\lambda \left(\frac{1}{p+1} \sum_{i=0}^p (H_\lambda(z_{K_{i+1}}) - H_\lambda(z_{K_i+1})) \right) \quad (4.26d)$$

$$+ (p+1)\zeta\alpha \quad (4.26e)$$

$$\leq \frac{c_2}{c_1} (p+1) \psi_\lambda \left(\frac{H_\lambda(z_0) - H_\lambda(z_K)}{p+1} \right) + (p+1)\zeta\alpha \quad (4.26f)$$

$$\leq \frac{c_2}{c_1} 2m \psi_\lambda \left(\frac{L(z_0) - L(z_K)}{2m} \right) + 2m\zeta\alpha. \quad (4.26g)$$

Indeed, (4.26e) follows from Jensen's inequality and (4.26g) follows from the fact that $s \mapsto s\psi_\lambda(a/s)$ is increasing over $(0, \infty)$ for any constant $a > 0$. Substituting c_1 and c_2 using (4.14c), (4.14b), and (4.20), we find that

$$\begin{aligned} \frac{c_2}{c_1} &= \frac{\sqrt{2} \max \left\{ \frac{1}{\alpha}, \frac{|\beta|}{\alpha} + M(|\gamma| + 1) + 4\lambda \right\}}{\min \left\{ \lambda - \left(\frac{1}{2\alpha} + \frac{M}{2} \right) \beta^2 - \frac{|\beta - \gamma|M}{2}, \frac{1}{2\alpha} - \frac{|\beta - \gamma|M}{2} - \lambda \right\}} \\ &= \frac{2\sqrt{2} \max \{1, |\beta| + M(|\gamma| + 1)\alpha + 4\lambda\alpha\}}{\min \{2\lambda\alpha - (1 + \alpha M) \beta^2 - |\beta - \gamma|M\alpha, 1 - |\beta - \gamma|M\alpha - 2\lambda\alpha\}}. \end{aligned}$$

If we take λ to be the midpoint of (λ^-, λ^+) , i.e., $\lambda = (\beta^2 + 1 + M\beta^2\alpha)/(4\alpha)$, then this simplifies to

$$\frac{c_2}{c_1} = \frac{4\sqrt{2}(|\beta| + M(|\gamma| + 1)\alpha + \beta^2 + 1 + M\beta^2\alpha)}{1 - \beta^2 - (\beta^2 + 2|\beta - \gamma|)M\alpha}.$$

Notice that c_2/c_1 is an increasing function of α over $(0, \bar{\alpha}]$, where we recall that $\bar{\alpha} = \min\{1/M, (1 - \beta^2)/(2(\beta^2 + 2|\beta - \gamma|)M)\}$. As a result,

$$\begin{aligned} \frac{c_2}{c_1} &\leq \frac{4\sqrt{2}(|\beta| + M(|\gamma| + 1)\bar{\alpha} + \beta^2 + 1 + M\beta^2\bar{\alpha})}{1 - \beta^2 - (\beta^2 + 2|\beta - \gamma|)M\bar{\alpha}} \\ &\leq \frac{4\sqrt{2}(|\beta| + M(|\gamma| + 1)\frac{1}{M} + \beta^2 + 1 + M\beta^2\frac{1}{M})}{1 - \beta^2 - (\beta^2 + 2|\beta - \gamma|)M\frac{1-\beta^2}{2(\beta^2+2|\beta-\gamma|)M}} \\ &= \frac{4\sqrt{2}(|\beta| + |\gamma| + 1 + \beta^2 + 1 + \beta^2)}{(1 - \beta^2)/2} \\ &\leq \frac{8\sqrt{2}(2 + |\gamma| + 3|\beta|)}{1 - \beta^2} =: c_3 > 0. \quad \square \end{aligned}$$

If the objective function satisfies the Łojasiewicz gradient inequality, then the Lyapunov function also satisfies it according to [120, Theorem 3.6]. Proposition 4.2 below generalizes [120, Theorem 3.6] from functions satisfying the Łojasiewicz gradient inequality to functions satisfying the uniform Kurdyka-Łojasiewicz inequality. We show that a suitable choice of desingularizing function for the objective is a common desingularizing function for the Lyapunov functions for all sufficiently large parameters.

Proposition 4.2. *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a locally Lipschitz semialgebraic function and $X \subset \mathbb{R}^n$ be bounded. The family of functions $(H_\lambda)_{\lambda \geq 1/4}$ admits a common desingularizing function on $X \times X$.*

Proof. By Proposition 1.1, there exists a desingularizing function ψ of f on X . Without loss of generality, we may assume that $\psi'(t) \geq 1/\sqrt{t}$ for all $t > 0$, after possibly replacing ψ by $t \mapsto \int_0^t \max\{\psi'(s), 1/\sqrt{s}\} ds$, which is semialgebraic³ and concave since the integrand is decreasing.

³To see why, note that $\{s > 0 : \psi'(s) \geq 1/\sqrt{s}\}$ is semialgebraic and hence a finite union of open intervals and points. Thus the integral is equal to ψ up to a constant on finitely many intervals of \mathbb{R}_+ , and $t \mapsto 2\sqrt{t}$ up to a constant otherwise. The graph of such a function is hence semialgebraic.

We may also multiply ψ by $1/\min\{1, c\} \geq 1$ where

$$c := \inf \{ \psi'((v_2 - v_1)/2) \theta((v_1 + v_2)/2) : v_1, v_2 \in V, v_1 < v_2, (v_1, v_2) \cap V = \emptyset \} > 0, \quad (4.27)$$

$\theta(v) := \inf\{d(0, \partial f(x)) : x \in X, f(x) = v\}$ for all $v \in \mathbb{R}$, and V is the set of critical values of f in \bar{X} if it is non-empty, otherwise $V := \{0\}$. Note that $c > 0$ as it is the infimum of finitely many positive real numbers. Indeed, $\psi'(t) > 0$ for all $t > 0$ and $\theta(v) > 0$ for all $v \notin V$. To see why the latter statement holds, assume the contrary that $\theta(v) = 0$ for some $v \notin V$. Then there exists $(x_k, s_k)_{k \in \mathbb{N}} \subset X \times \mathbb{R}^n$ such that $f(x_k) = v$, $s_k \in \partial f(x_k)$, and $s_k \rightarrow 0$. As X is bounded, $(x_k)_{k \in \mathbb{N}}$ admits a limit point \bar{x} . We have that $f(\bar{x}) = v$ by continuity of f and $0 \in \partial f(\bar{x})$ by [40, 2.1.5 Proposition (b)]. Thus $v \in V$ and a contradiction occurs.

By [40, Corollary 1, p. 39] we have

$$\partial H_\lambda(x, y) = (\partial f(x) + \{2\lambda(x - y)\}) \times \{2\lambda(y - x)\}, \quad (4.28)$$

so that $0 \in \partial H_\lambda(x, y)$ if and only if $0 \in \partial f(x)$ and $x = y$. Therefore, the set of critical values of f in \bar{X} and the set of critical values of H_λ in $\overline{X \times X}$ coincide. Accordingly, let $\tilde{f} := d(f, V)$ and

$\tilde{H}_\lambda := d(H_\lambda, V)$. Now fix $\lambda \geq 1/4$. For all $x, y \in X$ such that $0 \notin \partial \tilde{H}_\lambda(x, y)$, we have

$$d(0, \partial \tilde{H}_\lambda(x, y)) = d(0, \partial H_\lambda(x, y)) \quad (4.29a)$$

$$= d(0, (\partial f(x) + \{2\lambda(x - y)\}) \times \{2\lambda(y - x)\}) \quad (4.29b)$$

$$= \sqrt{d(0, \partial f(x) + 2\lambda(x - y))^2 + \|2\lambda(x - y)\|^2} \quad (4.29c)$$

$$\geq \sqrt{\eta_1 d(0, \partial f(x))^2 - \eta_2 \|2\lambda(x - y)\|^2 + \|2\lambda(x - y)\|^2} \quad (4.29d)$$

$$= \sqrt{\eta_1 d(0, \partial f(x))^2 + (1 - \eta_2) 4\lambda^2 \|x - y\|^2} \quad (4.29e)$$

$$\geq \sqrt{\eta_1 d(0, \partial f(x))^2 + (1 - \eta_2) \lambda \|x - y\|^2} \quad (4.29f)$$

$$\geq \sqrt{\frac{\eta_1}{\psi'(\tilde{f}(x))^2} + \frac{1 - \eta_2}{\psi'(\lambda \|x - y\|^2)^2}} \quad (4.29g)$$

$$\geq \sqrt{\frac{\min\{\eta_1, 1 - \eta_2\}}{\psi'(\max\{\tilde{f}(x), \lambda \|x - y\|^2\})^2}} \quad (4.29h)$$

$$\geq \sqrt{\frac{\min\{\eta_1, 1 - \eta_2\}}{\psi' \left(\frac{\tilde{f}(x) + \lambda \|x - y\|^2}{2} \right)^2}} \quad (4.29i)$$

$$\geq \frac{\sqrt{\min\{\eta_1, 1 - \eta_2\}}}{\psi' \left(\frac{\tilde{H}_\lambda(x, y)}{2} \right)}. \quad (4.29j)$$

Above, (4.29a) holds because $0 \notin \partial \tilde{H}_\lambda(x, y)$ and thus $\tilde{H}_\lambda(x', y') - \tilde{H}_\lambda(x, y) = \pm(H_\lambda(x', y') - H_\lambda(x, y))$ for all (x', y') in neighborhood of (x, y) where the sign is constant. (4.29b) is due to (4.28). (4.29c) holds because the distance function is defined using the Euclidean norm. The existence of the constants $\eta_1 > 0$, $\eta_2 \in (0, 1)$ in (4.29d) are guaranteed by [120, Lemma 3.1]. (4.29e) comes from a factorization. (4.29f) is due to the fact that $\lambda \geq 1/4$. (4.29g) is due to the uniform Kurdyka-Łojasiewicz inequality (1.2) and the fact that $\psi'(t) \geq 1/\sqrt{t}$ for all $t > 0$. Indeed, if $0 \notin \partial \tilde{f}(x)$, then $d(0, \partial f(x)) = d(0, \partial \tilde{f}(x)) \geq 1/\psi'(\tilde{f}(x))$ by [40, 2.3.9 Theorem (Chain Rule D) (ii) p. 42]. If $0 \in \partial \tilde{f}(x)$, then $\tilde{f}(x) = 0$ or $f(x) = (v_1 + v_2)/2$ for some $v_1, v_2 \in V$ such that $v_1 < v_2$ and $(v_1, v_2) \cap V = \emptyset$. In the former case, $d(0, \partial f(x)) \geq 1/\psi'(\tilde{f}(x)) = 1/\psi'(0) = 1/\infty = 0$ where $\psi'(0) := \lim_{a \searrow 0} \psi'(a)$. In the latter case, we have $\tilde{f}(x) = (v_2 - v_1)/2$ and thus $d(0, \partial f(x)) \geq \theta((v_1 + v_2)/2) \geq 1/\psi'((v_2 - v_1)/2) = 1/\psi'(\tilde{f}(x))$ by (4.27). (4.29h) and

(4.29i) hold because ψ is concave and thus ψ' is decreasing. (4.29j) is due to the fact that $0 < \tilde{H}_\lambda(z) = d(f(x) + \lambda\|x - y\|^2, V) \leq d(f(x), V) + \lambda\|x - y\|^2 = \tilde{f}(x) + \lambda\|x - y\|^2$. We conclude that $t \in [0, \infty) \rightarrow 2\psi(t/2)/\sqrt{\min\{\eta_1, 1 - \eta_2\}}$ is a desingularizing function of H_λ on $X \times X$ for all $\lambda \geq 1/4$, which is actually also a desingularizing function of f on X . \square

Thanks to Proposition 4.1 and Proposition 4.2, we are now ready to prove the length formula.

Proof of Lemma 4.1. Let $m \in \mathbb{N}^*$ be an upper bound of the number of critical values of f in \bar{X} . We apply Proposition 4.1 to the set X and let $\bar{\alpha} \in (0, 1]$, $c_3 > 1$, and $\zeta > 0$ be given by the proposition. Let $\bar{\psi}$ be a common desingularizing function of $(H_\lambda)_{\lambda \geq 1/4}$ on $X \times X$ given by Proposition 4.2. Let $\alpha \in (0, \bar{\alpha}]$ and let $\lambda := (\beta^2 + 1 + M\beta^2\alpha)/(4\alpha) \geq 1/4$ as defined in (4.22). Since $c_3 > 1$, $\psi(t) := 2c_3m\bar{\psi}(t/(2m))$ is also a desingularizing function of H_λ on $X \times X$. Let $\kappa := 2m\zeta$ and $\eta := 2m\delta^2(\beta^2 + 1 + M\beta^2)/4$ where $M > 0$ is a Lipschitz constant of ∇f on $S + \max\{|\beta|, |\gamma|\}(S - S)$ and S is the convex hull of X . It follows from (4.21) that

$$\begin{aligned} \sum_{k=0}^K \|x_{k+1} - x_k\| &\leq \psi(H_\lambda(x_0, x_{-1}) - H_\lambda(x_K, x_{K-1})) + \kappa\alpha \\ &\leq \psi\left(f(x_0) - f(x_K) + \lambda\|x_0 - x_{-1}\|^2\right) + \kappa\alpha \\ &\leq \psi\left(f(x_0) - f(x_K) + \lambda\delta^2\alpha^2\right) + \kappa\alpha \\ &\leq \psi\left(f(x_0) - f(x_K) + \eta\alpha\right) + \kappa\alpha. \end{aligned} \quad \square$$

4.3 Proof of Lemma 4.2

The matrix

$$A := \begin{pmatrix} 1 + \beta & -\beta \\ 1 & 0 \end{pmatrix} \otimes I_n$$

is diagonalizable as the Kronecker product of two such matrices [121, Exercise 15 p. 265]. Thus there exist an invertible matrix $P \in \mathbb{R}^{2n \times 2n}$ and a diagonal matrix $D \in \mathbb{R}^{2n \times 2n}$ such that $A = PDP^{-1}$. Let $\|x\|_P := \|P^{-1}x\|$. Then for any $X \in \mathbb{R}^{2n \times 2n}$, we have $\|X\|_P := \sup\{\|Xv\|_P : \|v\|_P \leq 1\}$.

$1\} = \sup\{\|P^{-1}XPv\| : \|v\| \leq 1\} = \|P^{-1}XP\|$. By equivalence of norms, there exist $c_1, c_2, c_3 > 0$ such that $c_1\|x\| \leq \|x\|_P \leq c_2\|x\|$ and $\|X\|_P \leq c_3\|X\|$.

Similar to [101, Theorem 2], let $\bar{\beta} := 1/(1-\beta)$. Also, let $L > 0$ and $M > 0$ respectively denote Lipschitz constants of $\bar{\beta}f$ and $\bar{\beta}\nabla f$ with respect to the Euclidean norm $\|\cdot\|$ on $S + \gamma(S - S)$ where S denotes the convex hull of $B(X_0, \sigma_T(X_0) + \delta + 1) := X_0 + B(0, \sigma_T(X_0) + \delta + 1)$ and

$$\sigma_T(X_0) := \sup_{x \in C^1(\mathbb{R}_+, \mathbb{R}^n)} \int_0^T \|x'(t)\| dt \quad (4.30a)$$

$$\text{subject to } \begin{cases} x'(t) = -\bar{\beta}\nabla f(x(t)), \forall t > 0, \\ x(0) \in X_0. \end{cases} \quad (4.30b)$$

Let $c_4 := ML(1/2 + |\beta|/2 + |\gamma| - \beta|\gamma|)$ and $c_5 := c_2M\sqrt{1 + 2\gamma + 2\gamma^2}$. Without loss of generality, we may assume that $X_0 \neq \emptyset$. The feasible set of (4.30) is thus non-empty (i.e., $\sigma_T(X_0) > -\infty$) because f is lower bounded and belongs to $C_{\text{loc}}^{1,1}(\mathbb{R}^n)$ [56, Theorem 17.1.1]. Notice that we also have $\sigma_T(X_0) < \infty$. Indeed, by the Cauchy-Schwarz inequality any feasible point $x(\cdot)$ of (4.30) satisfies

$$\begin{aligned} \int_0^T \|x'(t)\| dt &\leq \sqrt{T} \sqrt{\int_0^T \|x'(t)\|^2 dt} \\ &= \sqrt{T} \sqrt{\int_0^T \langle -\bar{\beta}\nabla f(x(t)), x'(t) \rangle dt} \\ &= \sqrt{T} \sqrt{\bar{\beta}f(x(0)) - \bar{\beta}f(x(T))} \\ &\leq \sqrt{T\bar{\beta} \left(\sup_{X_0} f - \inf_{\mathbb{R}^n} f \right)} < \infty. \end{aligned}$$

It is easy to check that L and ML are respectively Lipschitz and gradient Lipschitz constants on $[0, T]$ of any feasible point $x(\cdot)$ of (4.30). Indeed, let $x(\cdot)$ be a feasible point of (4.30). Since $x(t) \in B(X_0, \sigma_T(X_0))$ for all $t \in [0, T]$, we have $\|x'(t)\| = \|\bar{\beta}\nabla f(x(t))\| \leq L$. By the mean value theorem, for all $s, t \in [0, T]$ we have $\|x'(t) - x'(s)\| = \|\bar{\beta}\nabla f(x(t)) - \bar{\beta}\nabla f(x(s))\| \leq$

$M\|x(t) - x(s)\| \leq ML|t - s|$. As a byproduct, we get the Taylor bound

$$\|x(t) - x(s) - x'(s)(t - s)\| \leq \frac{ML}{2}(t - s)^2. \quad (4.31)$$

Let $\epsilon \in (0, \delta + 1]$ and $\bar{\alpha} := \min\{1, \epsilon c_1 c_2^{-1} [e^{c_5 T} (|\beta| \delta + 2L - L\beta + c_4 c_5^{-1}) - c_4 c_5^{-1}]\}^{-1} > 0$. Let $x_{-1}, x_0, x_1, \dots \in \mathbb{R}^n$ be a sequence generated by the gradient method with momentum and step size $\alpha \in (0, \bar{\alpha}]$ for which $x_0 \in X_0$ and $\|x_0 - x_{-1}\| \leq \delta\alpha$. Let $x(\cdot)$ be a feasible point of (4.30) such that $x(0) = x_0$. Similar to the proof of [101, Theorem 2], let $\bar{x}_k := x(k\alpha)$ for all $k \in \mathbb{N}$. We next reason by induction. We have $\|x_0 - \bar{x}_0\| = 0 \leq \epsilon$. Assume that $\|x_k - \bar{x}_k\| \leq \epsilon$ for $k = 0, \dots, K$ for some index $K \in \mathbb{N}$. For $k = 1, \dots, K$, we have

$$\|\bar{x}_{k+1} - \bar{x}_k + \alpha \bar{\beta} \nabla f(\bar{x}_k)\| \leq ML\alpha^2/2, \quad (4.32a)$$

$$\|\bar{x}_{k-1} - \bar{x}_k - \alpha \bar{\beta} \nabla f(\bar{x}_k)\| \leq ML\alpha^2/2. \quad (4.32b)$$

Multiplying (4.32b) by $|\beta|$ and adding it to (4.32a) yields

$$\|\bar{x}_{k+1} - \bar{x}_k - \beta(\bar{x}_k - \bar{x}_{k-1}) + \alpha \nabla f(\bar{x}_k)\| \leq ML(1 + |\beta|)\alpha^2/2,$$

where we use the fact that $\bar{\beta} - \beta\bar{\beta} = 1$. We also have

$$\begin{aligned} \|\nabla f(\bar{x}_k + \gamma(\bar{x}_k - \bar{x}_{k-1})) - \nabla f(\bar{x}_k)\| &\leq M(1 - \beta)|\gamma|\|\bar{x}_k - \bar{x}_{k-1}\| \\ &\leq ML(1 - \beta)|\gamma|\alpha. \end{aligned}$$

Hence by combining the above two inequalities, we have

$$\|\bar{x}_{k+1} - \bar{x}_k - \beta(\bar{x}_k - \bar{x}_{k-1}) + \alpha \nabla f(\bar{x}_k + \gamma(\bar{x}_k - \bar{x}_{k-1}))\| \leq c_4 \alpha^2$$

where $c_4 = ML(1/2 + |\beta|/2 + |\gamma| - \beta|\gamma|)$. Let $e_k = x_k - \bar{x}_k$. We have

$$\|e_{k+1} - e_k - \beta(e_k - e_{k-1}) + \alpha[\nabla f(x_k + \gamma(x_k - x_{k-1})) - \nabla f(\bar{x}_k + \gamma(\bar{x}_k - \bar{x}_{k-1}))]\| \leq c_4\alpha^2$$

by using the update rule of momentum method (4.1). Thus

$$\|e_{k+1} - e_k - \beta(e_k - e_{k-1}) + \alpha M_k(e_k + \gamma(e_k - e_{k-1}))\| \leq c_4\alpha^2$$

where M_k is the linear application such that $M_k(a_k - b_k) := \nabla f(a_k) - \nabla f(b_k)$, $M_k x := 0$ for all $x \in \text{span}(a_k - b_k)^\perp$, $a_k := x_k + \gamma(x_k - x_{k-1})$, $b_k := \bar{x}_k + \gamma(\bar{x}_k - \bar{x}_{k-1})$ if $a_k \neq b_k$, otherwise $M_k := 0$.

Let $v_k = (e_k, e_{k-1}) \in \mathbb{R}^{2n}$. We have $\|v_{k+1} - Av_k + \alpha B_k v_k\| \leq c_4\alpha^2$ where

$$B_k := \begin{pmatrix} 1 + \gamma & -\gamma \\ 0 & 0 \end{pmatrix} \otimes M_k.$$

We also have

$$\|B_k\| = \left\| \begin{pmatrix} 1 + \gamma & -\gamma \\ 0 & 0 \end{pmatrix} \right\| \|M_k\| \leq M\sqrt{1 + 2\gamma + 2\gamma^2}$$

since $\bar{x}_k + \gamma(\bar{x}_k - \bar{x}_{k-1})$ and $x_k + \gamma(x_k - x_{k-1})$ belong to $S + \gamma(S - S)$. The latter inclusion follows from the induction hypothesis and the fact that $\epsilon \leq \delta + 1$. Hence $\|v_{k+1} - Av_k + \alpha B_k v_k\|_P \leq c_2 c_4 \alpha^2$ and thus

$$\begin{aligned} \|v_{k+1}\|_P &\leq (\|A\|_P + \alpha\|B_k\|_P)\|v_k\|_P + c_2 c_4 \alpha^2 \\ &\leq (\|A\|_P + \alpha c_3\|B_k\|)\|v_k\|_P + c_2 c_4 \alpha^2 \\ &\leq (1 + c_5 \alpha)\|v_k\|_P + c_2 c_4 \alpha^2 \end{aligned}$$

where $c_5 = c_3 M \sqrt{1 + 2\gamma + 2\gamma^2}$. By induction, we find that

$$\begin{aligned}
\|x_{k+1} - \bar{x}_{k+1}\| &= \|e_{k+1}\| \leq \|v_{k+1}\| \leq c_1^{-1} \|v_{k+1}\|_P \\
&\leq c_1^{-1} (1 + c_5 \alpha)^k \|v_1\|_P + c_1^{-1} c_2 c_4 \alpha^2 \sum_{i=0}^{k-1} (1 + c_5 \alpha)^i \\
&= c_1^{-1} c_2 (1 + c_5 \alpha)^k \|v_1\| + c_1^{-1} c_2 c_4 \alpha^2 \frac{(1 + c_5 \alpha)^k - 1}{c_5 \alpha} \\
&\leq c_1^{-1} c_2 e^{c_5 k \alpha} \|e_1\| + c_1^{-1} c_2 c_4 c_5^{-1} (e^{c_5 k \alpha} - 1) \alpha \\
&\leq c_1^{-1} c_2 e^{c_5 T} (\|x_1 - x_0\| + \|x_0 - \bar{x}_0\| + \|\bar{x}_0 - \bar{x}_1\|) \\
&\quad + c_1^{-1} c_2 c_4 c_5^{-1} (e^{c_5 T} - 1) \alpha \\
&\leq c_1^{-1} c_2 e^{c_5 T} (\|\beta(x_0 - x_{-1}) - \alpha \nabla f(x_0 + \gamma(x_0 - x_{-1}))\| \\
&\quad + L\alpha) + c_1^{-1} c_2 c_4 c_5^{-1} (e^{c_5 T} - 1) \alpha \\
&\leq c_1^{-1} c_2 e^{c_5 T} (|\beta| \delta \alpha + L(1 - \beta) \alpha + L\alpha) \\
&\quad + c_1^{-1} c_2 c_4 c_5^{-1} (e^{c_5 T} - 1) \alpha \\
&\leq c_1^{-1} c_2 [e^{c_5 T} (|\beta| \delta + 2L - L\beta + c_4 c_5^{-1}) - c_4 c_5^{-1}] \bar{\alpha} \\
&= \epsilon
\end{aligned}$$

since $\bar{\alpha} = \epsilon c_1 c_2^{-1} [e^{c_5 T} (|\beta| \delta + 2L - L\beta + c_4 c_5^{-1}) - c_4 c_5^{-1}]^{-1}$.

4.4 Proof of Theorem 4.3

It is known that if $F \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, X is an open subset of \mathbb{R}^n such that $\text{rank}(F'(x)) = n$ for all $x \in X$, and $F(X) \subset X$, then for almost every $x \in X$, $F^k(x)$ does not converge as $k \rightarrow \infty$ to any fixed point of F in X whose spectral radius is greater than one [122, Theorem 2]. The sequence $(F^k)_{k \in \mathbb{N}}$ is defined by $F^{k+1} := F \circ F^k$ for all $k \in \mathbb{N}$ where $F^0 : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the identity. In order to prove Proposition 4.3 below, and ultimately Theorem 4.3, we relax the assumption that $F(X) \subset X$ and instead only require that $x_k \in X$ for all $k \in \mathbb{N}$. Below, we let $\mu(\cdot)$ and $\rho(\cdot)$ denote the Lebesgue measure and the spectral radius respectively.

Lemma 4.7. *If $F \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ and X is an open subset of \mathbb{R}^n such that $\text{rank}(F'(x)) = n$ for all $x \in X$, then*

$$\mu \left(\left\{ x \in \mathbb{R}^n : \forall k \in \mathbb{N}, F^k(x) \in X, \lim_{k \rightarrow \infty} F^k(x) \in Y \right\} \right) = 0,$$

where $Y := \{x \in X : F(x) = x, \rho(F'(x)) > 1\}$.

Proof. Since $F \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ and $\text{rank}(F'(x)) = n$ for all $x \in X$, by the inverse function theorem [123, Theorem 1 p. 498] F is a local diffeomorphism over X . By the center-stable manifold theorem [118, Theorem III.7(2) p. 65], for all $y \in Y$, there exists an open neighborhood B_y of y such that its associated local center stable manifold $W_{\text{loc}}^{\text{sc}}(y) := \{x \in \mathbb{R}^n : \forall k \in \mathbb{N}, F^k(x) \in B_y\}$ has Lebesgue measure zero. Since $\{B_y : y \in Y\}$ is an open cover of Y , by Lindelöf's lemma [89, Theorem 30.3(a)] there exists $\{y_i\}_{i \in \mathbb{N}} \subset Y$ such that $Y \subset \cup_{i=0}^{\infty} B_{y_i}$.

We seek to show that the set

$$W := \left\{ x \in \mathbb{R}^n : \forall k \in \mathbb{N}, F^k(x) \in X, \lim_{k \rightarrow \infty} F^k(x) \in Y \right\}$$

has Lebesgue measure zero. In order to do so, we consider the sequence $V_0, V_1, V_2, \dots : Y \rightrightarrows X$ defined by $V_0(\cdot) := W_{\text{loc}}^{\text{sc}}(\cdot) \cap X$ and $V_{k+1} := (F|_X)^{-1} \circ V_k$ for all $k \in \mathbb{N}$ where $F|_X$ denotes the restriction of F to X . We will show that

$$W \subset \bigcup_{i=0}^{\infty} \bigcup_{k=0}^{\infty} V_k(y_i).$$

It is then easy to show by induction that $\mu(V_k(y_i)) = 0$ for all $k, i \in \mathbb{N}$. Indeed, on the one hand $\mu(V_0(y_i)) \leq \mu(W_{\text{loc}}^{\text{sc}}(y_i)) = 0$. On the other hand, if $\mu(V_k(y_i)) = 0$, then by [124, Theorem 1] $\mu(V_{k+1}(y_i)) = \mu((F|_X)^{-1}(V_k(y_i))) = 0$ since $\text{rank}(F'(x)) = n$ for all $x \in X$. We conclude that

$$\mu(W) \leq \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} \mu(V_k(y_i)) = 0.$$

Let $x \in W$ and $y := \lim_{k \rightarrow \infty} F^k(x)$. Since $y \in Y \subset \cup_{i=0}^{\infty} B_{y_i}$, there exists $j \in \mathbb{N}$ such that

$y \in B_{y_j}$. Since B_{y_j} is open, there exists $K \in \mathbb{N}$ such that $F^k(x) \in B_{y_j}$ for all $k \geq K$, or equivalently, $F^k(F^K(x)) \in B_{y_j}$ for all $k \in \mathbb{N}$. Thus $F^K(x) \in W_{\text{loc}}^{\text{sc}}(y_j)$ and in fact $F^K(x) \in W_{\text{loc}}^{\text{sc}}(y_j) \cap X = V_0(y_j)$. Since $x \in W$ and $F^K(x) \in V_0(y_j)$, we have $F^{K-1}(x) \in F^{-1}(F^K(x)) \cap X = (F|_X)^{-1}(F^K(x)) \subset (F|_X)^{-1}(V_0(y_j)) = V_1(y_j)$. By induction, it follows that $x \in V_K(y_j) \subset \bigcup_{i=0}^{\infty} \bigcup_{k=0}^{\infty} V_k(y_i)$. \square

Given an objective function $f \in C^2(\mathbb{R}^n)$ with an L -Lipschitz continuous gradient, the momentum method (4.1) does not converge to any critical point whose Hessian has a negative eigenvalue for almost every initial point if $\alpha \in (0, 2(1 - \beta)/L)$, $\beta \in (0, 1)$ and $\gamma = 0$ [112, Lemma 2], or $\alpha \in (0, 4/L)$, $\beta \in (\max\{0, -1 + \alpha L/2\}, 1)$ and $\gamma = 0$ [113, Theorem 3]. In order to prove Theorem 4.3, we enlarge the set of allowable momentum parameters. Below, we let $\lambda_{\min}(\cdot)$ denote the minimal real eigenvalue of a matrix.

Proposition 4.3. *Let $f \in C^2(\mathbb{R}^n)$, $X \subset \mathbb{R}^n$ be bounded, $\beta \in (-1, 1) \setminus \{0\}$, and $\gamma \in \mathbb{R}$. There exists $\bar{\alpha} > 0$ such that for all $\alpha \in (0, \bar{\alpha}]$ and for almost every $(x_{-1}, x_0) \in \mathbb{R}^{2n}$, the limit of any convergent sequence $x_{-1}, x_0, x_1, \dots \in X$ generated by the momentum method (4.1) does not belong to*

$$C^- := \{x \in \mathbb{R}^n : \nabla f(x) = 0, \lambda_{\min}(\nabla^2 f(x)) < 0\}.$$

Proof. Since X is bounded, there exists an open bounded set \tilde{X} such that $\bar{X} \subset \tilde{X}$. Let $M := \sup\{\rho(\nabla^2 f(x)) : x \in \tilde{X} + \gamma(\tilde{X} - \tilde{X})\} < \infty$, $\bar{\alpha} := |\beta|/(1 + |\gamma|M)$, and $\alpha \in (0, \bar{\alpha}]$. Any sequence $x_{-1}, x_0, x_1, \dots \in \mathbb{R}^n$ generated by the momentum method (4.1) follows the update rule $z_{k+1} = F(z_k)$ for all $k \in \mathbb{N}$ where $z_k := (x_k, x_{k-1})$ and $F : \mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$ is defined by

$$F(x, y) := \begin{pmatrix} x + \beta(x - y) - \alpha \nabla f(x + \gamma(x - y)) \\ x \end{pmatrix}$$

for all $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$. In order to prove the desired result, we claim that it suffices to check the two following facts:

1. $\text{rank}(F'(z)) = 2n$ for all $z \in \tilde{X} \times \tilde{X}$,

2. $\{(x, x) \in \mathbb{R}^{2n} : x \in C^-\} \subset Z$,

where

$$Z := \{z \in \mathbb{R}^{2n} : F(z) = z, \rho(F'(z)) > 1\}.$$

Indeed, by applying Lemma 4.7 to $F \in C^1(\mathbb{R}^{2n}, \mathbb{R}^{2n})$ and the open subset $\tilde{X} \times \tilde{X}$ of \mathbb{R}^{2n} , it follows that for almost every $(x_0, x_{-1}) \in \mathbb{R}^{2n}$, the limit of any convergent sequence $(x_0, x_{-1}), (x_1, x_0), \dots \in X \times X$ such that $F(x_k, x_{k-1}) = F(x_{k+1}, x_k)$ for all $k \in \mathbb{N}$ does not belong to Y where

$$\begin{aligned} Y &:= Z \cap (\tilde{X} \times \tilde{X}) \\ &\supset \{(x, x) \in \mathbb{R}^{2n} : x \in C^-\} \cap (\tilde{X} \times \tilde{X}) \\ &= \{(x, x) \in \mathbb{R}^{2n} : x \in C^- \cap \tilde{X}\}. \end{aligned}$$

In particular, for almost every $(x_0, x_{-1}) \in \mathbb{R}^{2n}$, the limit of any convergent sequence $x_{-1}, x_0, x_1, \dots \in X$ generated by the momentum method (4.1) does not belong to $C^- \cap \tilde{X}$. Since $\bar{X} \subset \tilde{X}$, such a limit must belong to \tilde{X} , and thus does not belong to C^- .

We next prove the two facts above. First, for any $(x, y) \in \tilde{X} \times \tilde{X}$, we have

$$F'(x, y) = \begin{pmatrix} (1 + \beta)I_n - \alpha(1 + \gamma)\nabla^2 f(x + \gamma(x - y)) & -\beta I_n + \alpha\gamma\nabla^2 f(x + \gamma(x - y)) \\ I_n & 0 \end{pmatrix}$$

where $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix. Since $|\beta| \geq \alpha(1 + |\gamma|M) > \alpha|\gamma|M$, by [125, Theorem 3] for all $(x, y) \in \tilde{X} \times \tilde{X}$ we have

$$\det(F'(x, y)) = \det(\beta I_n - \alpha\gamma\nabla^2 f(x + \gamma(x - y))) \neq 0$$

and thus $\text{rank}(F'(x, y)) = 2n$. Second, let $x \in C^-$ and $z := (x, x)$. We seek to show that $z \in Z$. Since $\nabla f(x) = 0$, we have $F(z) = F(x, x) = (x, x) = z$. Since $f \in C^2(\mathbb{R}^n)$, $\nabla^2 f(x)$ is symmetric and therefore admits an eigendecomposition $\nabla^2 f(x) = PDP^\top$ where $D = \text{diag}(d_1, \dots, d_n)$ and P

is an orthogonal matrix. Again by [125, Theorem 3], we have

$$\begin{aligned}
\det(\lambda I_{2n} - F'(x, x)) &= \det([\lambda^2 - (1 + \beta)\lambda + \beta]I_n - [\alpha\gamma - \lambda\alpha(1 + \gamma)]\nabla^2 f(x)) \\
&= \det([\lambda^2 - (1 + \beta)\lambda + \beta]I_n - [\alpha\gamma - \lambda\alpha(1 + \gamma)]PDP^\top) \\
&= \det([\lambda^2 - (1 + \beta)\lambda + \beta]I_n - [\alpha\gamma - \lambda\alpha(1 + \gamma)]D) \\
&= \prod_{i=1}^n ([\lambda^2 - (1 + \beta)\lambda + \beta] - [\alpha\gamma - \lambda\alpha(1 + \gamma)]d_i) \\
&= \prod_{i=1}^n (\lambda^2 + \underbrace{[\alpha(1 + \gamma)d_i - (1 + \beta)]\lambda + \beta - \alpha\gamma d_i}_{\varphi_i(\lambda)}).
\end{aligned}$$

Since $x \in C^-$, there exists $j \in \{1, \dots, n\}$ such that $d_j < 0$. Since φ_j is a quadratic function whose leading coefficient is positive and $\varphi_j(1) = \alpha d_j < 0$, φ_j has a root that is greater than 1. Thus $\rho(F'(z)) > 1$. \square

We are now ready to prove Theorem 4.3.

Proof of Theorem 4.3. Let $f \in C^2(\mathbb{R}^n)$ be a semialgebraic function with bounded continuous gradient trajectories. Let $\beta \in (-1, 1) \setminus \{0\}$, $\gamma \in \mathbb{R}$, and $\delta \geq 0$. Assume that the Hessian of f has a negative eigenvalue at all critical points of f that are not local minimizers. Let X_0 be a bounded subset of \mathbb{R}^n . By Theorem 4.2, there exist $\bar{\alpha}, c > 0$ such that for all $\alpha \in (0, \bar{\alpha}]$, there exists $c_\alpha > 0$ such that any sequence $x_{-1}, x_0, x_1, \dots \in \mathbb{R}^n$ generated by the momentum method (4.1) that satisfies $x_0 \in X_0$ and $\|x_0 - x_{-1}\| \leq \delta\alpha$ obeys

$$\sum_{i=0}^{\infty} \|x_{i+1} - x_i\| \leq c \quad \text{and} \quad \min_{i=0, \dots, k} \|\nabla f(x_i)\| \leq \frac{c_\alpha}{k+1}, \quad \forall k \in \mathbb{N}.$$

It hence converges to a critical point of f and belongs to the bounded set $B(X_0, c)$. By Proposition 4.3, after possibly reducing $\bar{\alpha} > 0$, for all $\alpha \in (0, \bar{\alpha}]$ and for almost every $(x_{-1}, x_0) \in \mathbb{R}^{2n}$, the limit of any convergent sequence $x_{-1}, x_0, x_1, \dots \in B(X_0, c)$ generated by the momentum method (4.1) is not a critical point of f where the Hessian admits a negative eigenvalue. We conclude for all $\alpha \in (0, \bar{\alpha}]$ and for almost every $(x_{-1}, x_0) \in \mathbb{R}^n \times X_0$, any sequence $x_{-1}, x_0, x_1, \dots \in \mathbb{R}^n$ gener-

ated by the momentum method (4.1) that satisfies $\|x_0 - x_{-1}\| \leq \delta\alpha$ converges to a local minimizer of f . □

Chapter 5: Proximal random reshuffling algorithm

In Chapter 4, we demonstrated how the notion of bounded subgradient trajectories can be used to analyze the convergence properties of momentum methods applied to locally smooth semialgebraic functions (Theorems 4.1 to 4.3). In this chapter, we extend these ideas to a broader setting by considering stochastic and constrained optimization problems. The results presented in this chapter are based on the following article:

C. Jozs, L. Lai, and X. Li, “Proximal random reshuffling under local lipschitz continuity,” *arXiv preprint arXiv:2408.07182*, 2024

Specifically, we consider proximal random reshuffling (PRR, i.e., Algorithm 1) for solving the composite model

$$\inf_{x \in \mathbb{R}^n} \Phi(x) := \sum_{i=1}^N f_i(x) + g(x) \quad (5.1)$$

where $f_1, \dots, f_N : \mathbb{R}^n \rightarrow \mathbb{R}$ are either locally Lipschitz or differentiable with locally Lipschitz gradients, and $g : \mathbb{R}^n \rightarrow \overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ is proper, convex, and locally Lipschitz on its closed domain. We allow the user to choose any conservative field $D_i : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ for each f_i , for example the gradient ∇f_i if f_i is continuously differentiable, or the Clarke subdifferential ∂f_i if f_i is semialgebraic, but it could also be the output of an automatic differentiation procedure (see, e.g., [28, Theorem 8]).

Inspired by the pioneering works of [24, 25, 127, 128, 129, 26], we consider the differential inclusion

$$x'(t) \in -(D_1 + \dots + D_N + \partial g)(x(t)), \quad \text{for a.e. } t > 0, \quad (5.2)$$

where $x : [0, \infty) \rightarrow \mathbb{R}^n$ is an absolutely continuous function. Their insight, extended to the proximal setting by [22] (i.e., $N = 1$, $D_1 = \partial f_1$), is that trajectories of (5.2) track the iterates

Algorithm 1 Proximal random reshuffling (PRR)

choose $\alpha_0, \alpha_1, \alpha_2, \dots > 0, x_0 \in \text{dom } \Phi$.

for $k = 0, 1, 2, \dots$ **do**

$x_{k,0} = x_k$

 choose a permutation σ^k of $\llbracket 1, N \rrbracket$

for $i = 1, \dots, N$ **do**

$x_{k,i} \in x_{k,i-1} - \alpha_k D_{\sigma_i^k}(x_{k,i-1})$

end for

$x_{k+1} = \text{prox}_{\alpha_k g}(x_{k,N})$

end for

$(x_k)_{k \in \mathbb{N}}$ generated by Algorithm 1 in the following sense. For any accuracy $\epsilon > 0$ and any time $T > 0$, after sufficiently many iterations, say l , we have $\|x_k - x(t_k - t_l)\| \leq \epsilon$ for some solution $x(\cdot)$ to (5.2), where the times $t_k := \alpha_0 + \dots + \alpha_{k-1}$ lie in $[t_l, t_l + T]$. This requires bounded iterates, a regularity condition that is slightly stronger than local Lipschitz continuity, and a specific choice of step sizes.

However, such bounded iterates assumption can be artificial and unnecessary for understanding the fundamental behavior of the algorithm. In the absence of a global gradient Lipschitz constant, even if iterates remain bounded in practice, gradient descent (a special case of Algorithm 1) may fail to converge (see Figure 5.1 below). This occurs because the iterates are not uniformly bounded

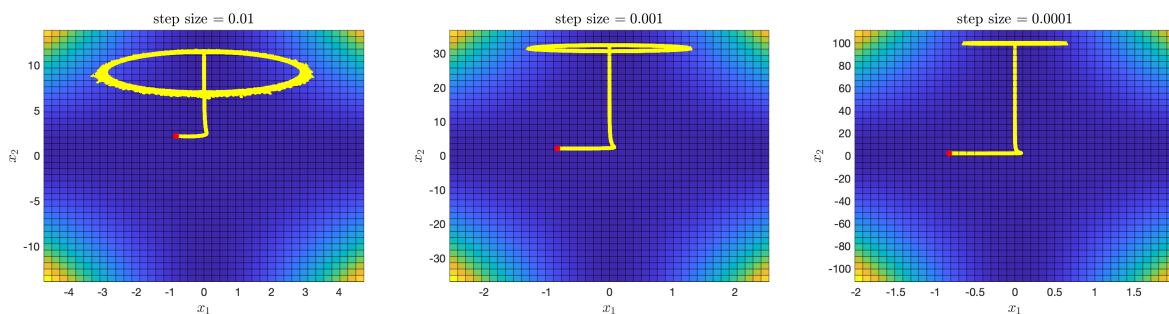


Figure 5.1: Iterates of the gradient descent with different constant step sizes applied to $f(x, y) = x^2y^2 - x - y$.

for all sufficiently small step sizes, and the sufficient decrease condition [21, H1] does not hold. A similar issue arises in machine learning problems such as ℓ_1 matrix completion (Example 5.2 without constraints), where constant step sizes lead to non-uniformly bounded iterates (see Figure 5.2). Moreover, while many well-tuned algorithms exhibit bounded iterates in practice, there

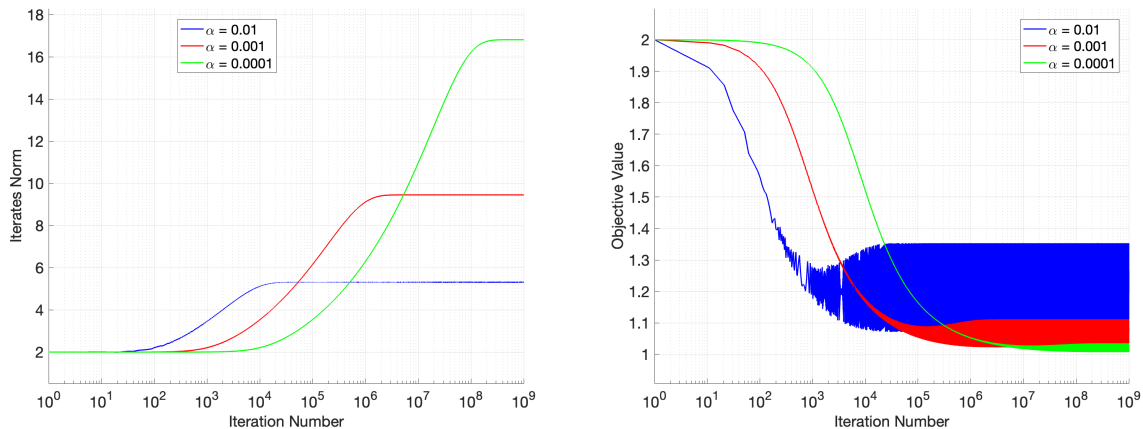


Figure 5.2: Iterate norms and function values of the subgradient method with constant step sizes $\alpha_k = \alpha$ applied to ℓ_1 matrix completion instance $f(x_1, x_2, y_1, y_2) := |1 - x_1 y_1| + |1 - x_2 y_1| + |1 - x_2 y_2|$.

exist natural scenarios where unbounded iterates arise. For instance, in the same ℓ_1 matrix completion problem, using diminishing step sizes can lead to unbounded iterates (see Figure 5.3). Finally, verifying boundedness empirically is inherently difficult. As shown in Figure 5.4, even

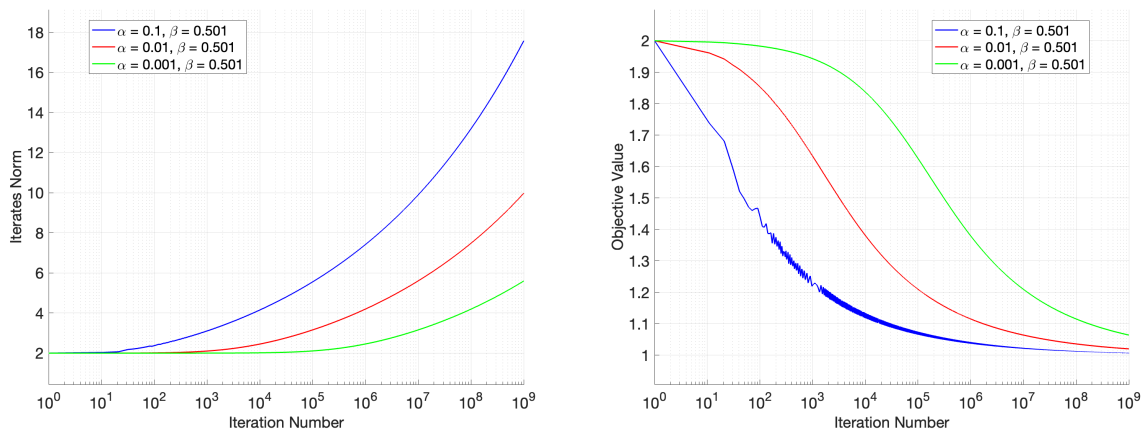


Figure 5.3: Iterate norms and function values of the subgradient method with non-summable diminishing step sizes $\alpha_k = \alpha / (k + 1)^\beta$ applied to robust matrix completion instance.

with summable step sizes, after running 10^9 iterations, the iterate norm appears to increase, but it does so at such a slow rate that it is unclear whether it is truly unbounded or simply growing at a diminishing rate and will eventually stabilize. This highlights the need for theoretical results that do not rely on bounded iterate assumptions.

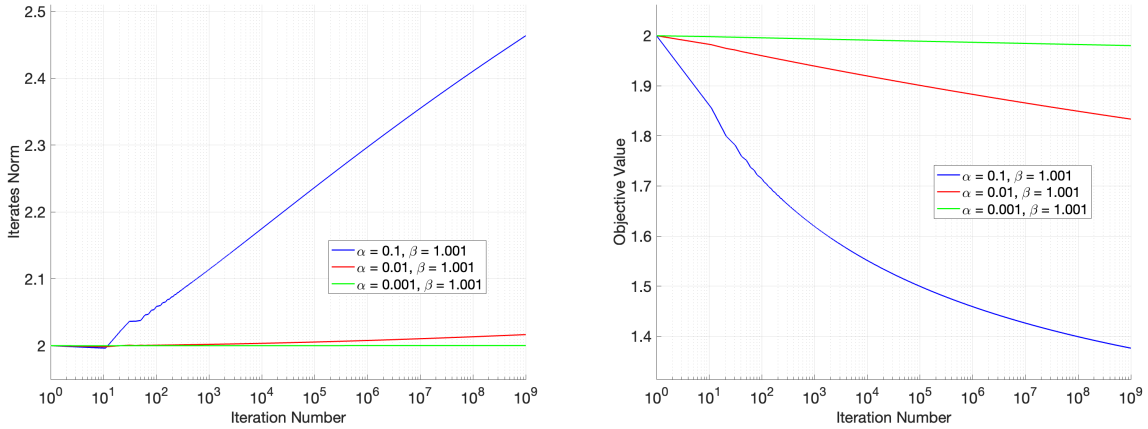


Figure 5.4: Iterate norms and function values of the subgradient method with summable diminishing step sizes $\alpha_k = \alpha/(k + 1)^\beta$ applied to robust matrix completion instance.

In this chapter, we propose a tracking lemma devoid of such assumptions, which is at the same time more intuitive. For any initial iterate in some bounded set, we show that $\|x_k - x(t_k)\| \leq \epsilon$ for all times t_k in $[0, T]$, provided that the step sizes are sufficiently small. Notably, this result extends [23] in a nontrivial way, even if we ignore the constraints, as it applies over any finite time horizon without requiring boundedness of iterates. In the general case, the dynamics is governed by the set-valued mapping $D_1 + \dots + D_N + \partial g$ with possibly unbounded values. This leads us to extend the notion due to [28] initially conceived with compact values. Some set-valued analysis reveals that the local boundedness of $x \in \mathbb{R}^n \mapsto \inf\{\|y\| : y \in (D_1 + \dots + D_N + \partial g)(x)\}$ plays a role instead. The idea of selecting the minimal norm element from unbounded set-valued mappings is well-established in convex subgradient dynamics [130]. In nonconvex settings, some authors [131, 74] have instead truncated unbounded set-valued mappings, a concept closely linked to the local boundedness of their distance to the origin.

The tracking lemma enables us to ensure that Algorithm 1 can recover stationary points of the objective function Φ , with varying degrees of precision depending on the regularity of f_1, \dots, f_N and g . If f_1, \dots, f_N are locally Lipschitz, by a direct application of the tracking lemma, we establish that Algorithm 1 reaches (ϵ, δ) -near approximate stationary points (Definition 5.2) under

mild conditions (Theorem 5.1). If the objective is weakly convex, then the algorithm converges to (ϵ, δ) -near approximate stationary points when certain summable step sizes are used (Theorem 5.2). If additionally requiring the uniqueness and convergence of the solution to (5.2), then the algorithm converges to $(\epsilon, 0)$ -near approximate stationary points (Theorem 5.3). The power of the tracking lemma becomes especially apparent when f_1, \dots, f_N are locally smooth. Inspired by [45] and [132], we establish that if Φ is definable and has bounded subgradient trajectories, then Algorithm 1 converges to stationary points at a rate approaching $o(1/\sqrt{k})$ (Theorem 5.4). This rate improves to $o(1/k)$ in the deterministic setting ($N = 1$), where Algorithm 1 reduces to the proximal gradient method. Applied to nonnegative matrix factorization (NMF) [72, 18, 133], it provides the first convergence guarantee to stationary points for solving NMF via the projected gradient method under the Euclidean metric with constant step sizes (see Example 5.4 and its subsequent discussion for more details).

This chapter is organized as follows. Section 5.1 provides a detailed review of related work, comparing our results and techniques with existing approaches to highlight our contributions. Section 5.2 contains the main results and is divided into several subsections. They respectively contain the definitions, assumptions, theorems, and examples. Section 5.3 contains the proofs. It begins with main technical lemma regarding tracking by trajectories of conservative fields. The four main results are then proved successively using this lemma. Finally, we prove all intermediate results in Section 5.4 with a summary of our findings, a discussion of limitations, and potential directions for future research.

5.1 Literature review

This section reviews the literature closely related to Algorithm 1, comparing their results and assumptions to ours. We divide the discussion into two subsections based on the regularity of the component functions f_i 's: nonsmooth and locally smooth. At the end of each subsection, we also compare the proof techniques used in our analysis with those in the related works.

5.1.1 Nonsmooth component functions

We begin by reviewing related work on algorithms for solving problem (5.1), where the component functions f_i are assumed to be locally Lipschitz. To the best of our knowledge, the PRR algorithm has not yet been analyzed under such generality. However, several related algorithms have been studied under varying assumptions, and many of the techniques developed in those works have inspired our own approach. A recent study by [134] investigates a class of proximal algorithms that includes a special case of PRR without the reshuffling step. Under the assumptions that the component functions are Clarke regular [40] and that the iterates are bounded, they prove subsequential convergence to critical points. The proximal stochastic subgradient method with replacement [135, 136, 22] is closely related to PRR, consisting of a stochastic subgradient step on $f := f_1 + \dots + f_N$, followed by a proximal step with respect to g . In the convex setting, convergence was established in [135] under certain Lipschitz conditions [135, equation (6)]. If one relaxes the convexity and assume that f_1, \dots, f_N are Lipschitz continuous and Φ is weakly convex instead, then the method generates a point with small gradient norm of a Moreau envelope of Φ in expectation [136, Theorem 3.4]. If one further relaxes the regularity assumptions and only assume that Φ is Whitney stratifiable, it has been shown that the iterates converge subsequentially to composite critical points, provided that they remain bounded almost surely and satisfy additional technical assumptions (see, e.g., [22, Theorem 6.2] and [131]).

We next turn to the other subgradient-based algorithms that apply to general nonsmooth and nonconvex optimization problems, which might not admit the same composite structure as in (5.1). This includes variants of the subgradient method that allow stochasticity [22, 23], inexact subgradient oracles [137], and/or momentum terms in the updates [23, 138, 139]. These algorithms are analyzed in settings where Assumption 5.1 always hold. Moreover, it is worth noticing that all of these works assume either the iterates are bounded or the objective function is coercive. Under such assumptions, the algorithms are shown to converge subsequentially to critical points with diminishing step sizes [22, 138, 137], or eventually stay close to the set of critical points with non-diminishing step sizes [23, 139, 137]. In terms of the stationary measure in this chapter (see

Definition 5.2), for any $\epsilon > 0$, these algorithms return $(\epsilon, 0)$ -near approximate stationary (NAS) points after sufficiently large number of iterates, given that the step sizes are sufficiently small. The analyses in these works rely on the continuous-time limit of the iterates as we have discussed in the introduction.

We discuss the relevance of Theorems 5.1 to 5.3 in light of the previous results. First, we avoid assuming coercivity of the objective function or boundedness of the iterates, which allows us to apply these theorems to applications in data science (see Section 5.2.4). Specifically, Theorem 5.1 establishes that an approximate stationary point can be reached under mild assumptions (Assumption 5.1). While this result may seem limited, it cannot be strengthened without additional hypotheses. To analyze the asymptotic behavior of Algorithm 1, we rely on summable step sizes in Theorems 5.2 and 5.3 under stricter assumptions. These theorems follow directly from the tracking lemma (Lemma 5.1), proven in Section 5.3.1. Notably, Theorem 5.2 provides new asymptotic guarantees for locally Lipschitz weakly convex functions, advancing beyond prior works [136, 134]. Meanwhile, Theorem 5.3 appears to be the first result leveraging bounded trajectory assumptions for subgradient-based methods, whereas prior studies applied this condition only to gradient and momentum methods [45, 100] for smooth functions.

5.1.2 Locally smooth component functions

We now discuss the contributions of our results when the component functions f_i are locally smooth, and relate them to existing work. We divide our discussion into the deterministic case (Algorithm 1 with $N = 1$) and the stochastic case (Algorithm 1 with $N > 1$), with the latter one further divided based on the presence or absence of the nonsmooth term g . We conclude by highlighting the technical contributions of Theorem 5.4 in comparison to the literature.

In the deterministic case (Algorithm 1 with $N = 1$), the algorithm reduces to the proximal gradient method. Theorem 5.4 establishes that, if Φ is locally smooth definable and has bounded subgradient trajectories, the iterates converge to stationary points at a rate of $o(1/k)$ under sufficiently small constant step sizes. This rate is faster than the general $O(1/\sqrt{k})$ convergence for

nonconvex globally smooth functions when Φ attains its infimum ([17, Theorem 10.15], [140, Theorem 3]), and is consistent with the known $O(1/k)$ rate in the convex case ([17, Theorem 10.26], [140, Theorem 4]). The global smoothness assumption can be relaxed to local smoothness if a sufficient decrease condition [21, H1] is satisfied. Under this condition, convergence to critical points holds [141, Theorem 3.1], and the iterates have finite length when a limit point exists and Φ satisfies the Kurdyka-Łojasiewicz inequality ([21, Theorem 5.1], [142, Theorem 3.1], [143, Theorem 4.5]). Furthermore, explicit convergence rates can be obtained if the desingularizing function in the KŁ inequality is a power function ([142, Theorem 3.4], [143, Theorem 4.6]).

If $N > 1$, Theorem 5.4 shows that if Φ is locally smooth definable and has bounded subgradient trajectories, then the PRR algorithm (Algorithm 1) with diminishing step sizes $\alpha_k = \alpha/(k+1)^\beta$ converges to stationary points at a rate $o(1/k^{1-\beta})$ for any $\beta \in (1/2, 1)$ and sufficiently small α . When $g = 0$, Algorithm 1 corresponds to the random reshuffling (RR) algorithm, also known as stochastic gradient descent (SGD) without replacement [144, 145, 146]. If a fixed permutation σ^k is used for all $k \in \mathbb{N}$, the method reduces to the classical incremental gradient (IG) algorithm, whose asymptotic behavior has been extensively studied [60, 147, 148, 149]. RR has been shown to converge faster than IG or SGD with replacement when the f_i are convex with globally Lipschitz gradients and Hessians [146, 150]. In the nonconvex setting with globally smooth f_i , RR achieves $O(1/\epsilon^3)$ complexity using $O(\epsilon)$ constant step sizes under certain expected smoothness conditions [151], as shown in [145, Corollary 3] and [152, Corollary 1]. The same complexity is obtained using optimized diminishing step sizes [152, Theorem 6]. When f_i are semialgebraic and iterates are bounded, RR with the same step sizes as in Theorem 5.4 converges to stationary points [132, Corollary 3.8] at the rate $O(k^{-(3\beta-1)/2})$ for β close to $1/2$ [153, Theorem 5.3], approaching $O(k^{-1/4})$, which can be improved with knowledge of a Łojasiewicz exponent.

When $g \neq 0$, related results for PRR and proximal SGD are more limited. For convex, globally smooth f_i , [154] shows proximal SGD converges at nearly $O(1/\sqrt{k})$, matching our rate for nonconvex f_i . [155] studies projected proximal SGD with $C^1 f$, and g locally Lipschitz, regular, and lower bounded, assuming a compact constraint set. Their framework handles broader function classes

without requiring definability or convexity of g , but critically relies on boundedness of the domain, making it inapplicable to problems like NMF. In contrast, PRR achieves better iteration complexity, $O(1/\epsilon^3)$ versus $O(1/\epsilon^4)$ for proximal SGD, when f_i are globally smooth and additional regularity conditions hold ([156, Section 4], [136]). A normal-map-based PRR algorithm was proposed in [153], showing similar improvements over the proximal stochastic subgradient method under analogous assumptions. By further assuming bounded iterates and a Kurdyka-Łojasiewicz property, the asymptotic behavior of the iterates is also established [153, Sections 4 and 5].

We conclude by highlighting the contributions of Theorem 5.4 in the context of related work. Global convergence of deterministic first-order methods for smooth definable functions is well established [20, 21, 106, 45, 100], while convergence of SGD with replacement has been studied in [157] and [158]. The analysis of RR is more recent in [132]. Although [45] and [100] also leverage bounded subgradient trajectories, their results are restricted to deterministic methods with a local descent property. In contrast, the PRR algorithm analyzed in Theorem 5.4 does not satisfy such a property and instead relies on an approximate descent condition (Lemma 5.7), which complicates the proof of a length formula (Proposition 5.4 and corollary 5.2). The inclusion of a convex extended-value term g further introduces difficulties due to the potential unboundedness of ∂g . We also demonstrate that Theorem 5.4 applies to NMF, which requires a separate proof of bounded subgradient trajectories due to a weak balanceness condition (Lemma 2.1) in contrast to the strong balanceness condition [69, Theorem 2.2] used for matrix factorization (MF) in [45, 100]. Finally, unlike [132], we remove the assumptions of global smoothness and bounded iterates, and extend the analysis from semialgebraic to general definable functions. This generalization is nontrivial in the RR setting, where previous work relies on a quasi-additivity condition for the desingularizing function, an assumption that fails for general definable functions. Our approach circumvents this via a new uniform Kurdyka-Łojasiewicz inequality (Lemma 5.8).

5.2 Main results

Given two integers $a \leq b$, we use the notation $\llbracket a, b \rrbracket := \{a, a + 1, \dots, b\}$. A function $\Phi : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ is weakly convex [159, 160] if there exists $\rho \geq 0$ such that $\Phi + \rho \|\cdot\|^2$ is convex. Given $g : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$, the proximal mapping $\text{prox}_g : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is defined by $\text{prox}_g(x) := \arg \min\{y \in \mathbb{R}^n : g(y) + \|y - x\|^2/2\}$ for all $x \in \mathbb{R}^n$ [161, 3.b]. If g is proper, lower semicontinuous and convex, then $\text{prox}_g(x)$ is a singleton for any $x \in \mathbb{R}^n$ [17, Theorem 6.3] and it lies in the domain of the convex subdifferential of g [17, Theorem 6.39].

5.2.1 Definitions

We begin by stating the definition of a solution to a differential inclusion.

Definition 5.1. Let I be an interval of \mathbb{R}_+ and $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a set-valued mapping. We say that $x : I \rightarrow \mathbb{R}^n$ is a solution to $x' \in -D(x)$ if $x(\cdot)$ is absolutely continuous and $x'(t) \in -D(x(t))$ for almost every $t \in I$.

We say that $x(\cdot)$ is a D -trajectory if it is maximal (see, e.g., [45, Definition 5]). Also, we say that $x(\cdot)$ is a subgradient trajectory of $\Phi : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ if it is a $\partial\Phi$ -trajectory, where $\partial\Phi : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ denotes the Clarke subdifferential (see, e.g., [38, p. 336], [40, Chapter 2]).

In order to measure the optimality of the iterates generated by Algorithm 1, we will borrow the notion of (ϵ, δ) -near approximate stationarity (see, e.g., [162, Definition 4], [163, Definition 2.7], [164, 136]). It was introduced due to the intractability of finding near stationary points for Lipschitz functions, in the sense that the Clarke subdifferential admits small elements [165, Theorem 5]. It is thus well suited for this chapter. We remark that this is a stronger stationary notion compared to the one based on the Goldstein subdifferential ([165, Definition 4], [166]). Given $S \subset \mathbb{R}^n$, $x \in \mathbb{R}^n$, and $r \geq 0$, let $B(x, r) := \{y \in \mathbb{R}^n : \|x - y\| \leq r\}$, $d(x, S) := \inf\{\|x - y\| : y \in S\}$, and $P_S(x) := \arg \min\{\|x - y\| : y \in S\}$.

Definition 5.2. Given $\epsilon, \delta \geq 0$ and a set-valued mapping $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, a point $x \in \mathbb{R}^n$ is (ϵ, δ) -near approximate D -stationary if $d(0, D(B(x, \epsilon))) \leq \delta$.

Remark 5.1. In particular, a $(0, 0)$ -near approximate D -stationary point is also called a D -stationary point or a D -critical point.

We next recall a generalization of semialgebraic sets ([36], [167, Definition p. 503-506]) which is quite relevant for practical applications in optimization. It will be useful for establishing convergence of Algorithm 1 to stationary points.

Definition 5.3. An o-minimal structure on the real field is a sequence $S = (S_k)_{k \in \mathbb{N}}$ such that for all $k \in \mathbb{N}$:

1. S_k is a Boolean algebra of subsets of \mathbb{R}^k , with $\mathbb{R}^k \in S_k$;
2. S_k contains the diagonal $\{(x_1, \dots, x_k) \in \mathbb{R}^k : x_i = x_j\}$ for $1 \leq i < j \leq k$;
3. If $A \in S_k$, then $A \times \mathbb{R}$ and $\mathbb{R} \times A$ belong to S_{k+1} ;
4. If $A \in S_{k+1}$ and $\pi : \mathbb{R}^{k+1} \rightarrow \mathbb{R}^k$ is the projection onto the first k coordinates, then $\pi(A) \in S_k$;
5. S_3 contains the graphs of addition and multiplication;
6. S_1 consists exactly of the finite unions of open intervals and singletons.

A subset A of \mathbb{R}^n is definable in an o-minimal structure $(S_k)_{k \in \mathbb{N}}$ if $A \in S_n$. A function $\Phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is definable in an o-minimal structure if its epigraph (or equivalently its graph) is definable in that structure. Similarly, a set-valued mapping $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is definable in that structure if its graph is definable in that structure. Throughout this chapter, we fix an arbitrary o-minimal structure on the real field, and say that the sets or functions are definable if they are definable in this structure.

The last definition extends the notion of conservative field (see [168, Definition 3.7], [28]) to allow for unbounded and empty values. The domain and graph of a set-valued mapping $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ are respectively defined by $\text{dom } D := \{x \in \mathbb{R}^n : D(x) \neq \emptyset\}$ and $\text{graph } D := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : D(x) \ni y\}$. A set-valued mapping is proper if $\text{dom } D \neq \emptyset$. It is locally bounded if for all $x \in \mathbb{R}^n$, there exists a neighborhood U of x such that $D(U)$ is bounded.

Definition 5.4. Given $\Phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ that is locally Lipschitz continuous on its domain, a set-valued mapping $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a conservative field for Φ if it has closed graph, $\text{dom } D \subset \text{dom } \Phi$, and for any absolutely continuous function $x : [0, 1] \rightarrow \text{dom } D$, we have $(\Phi \circ x)'(t) = \langle v, x'(t) \rangle$ for all $v \in D(x(t))$ and almost every $t \in (0, 1)$.

If $\Phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is proper, lower semicontinuous, definable, locally Lipschitz continuous on its domain, and the Clarke subdifferential has closed graph, then the Clarke subdifferential is a conservative field for Φ [55, Corollary 5.4]. This is also true if $\Phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is subdifferentially regular and locally Lipschitz continuous on its domain, and the Clarke subdifferential has closed graph [55, Lemma 4.11]. In particular, if $\Phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is convex and locally Lipschitz continuous on its domain, then the convex subdifferential (equal to the Clarke subdifferential $\partial\Phi$ due to [169, (3.19) Proposition]) is a conservative field for Φ . The following properties, valid for conservative fields of real-valued definable functions, continue to hold for those of extended real-valued definable functions:

1. projection formula [28, Theorem 4] (see Theorem 5.1),
2. Morse-Sard theorem [28, Theorem 5] (see Theorem 5.2),
3. Kurdyka-Łojasiewicz inequality ([28, Theorem 6], [43, Theorem 14]) (see Theorem 5.3),
4. uniform Kurdyka-Łojasiewicz inequality [45, Proposition 5] (see Lemma 5.8),
5. length formula ([45, Proposition 7], [44, Theorem 2]) (see Proposition 5.1),
6. uniform boundedness [45, Lemma 1] (assuming $d(0, D)$ is bounded over bounded sets, see Proposition 5.3).

The proofs are generally similar and omitted for brevity.

While our main results in Section 5.2.3 can be obtained by specializing the general unbounded conservative field D to the form $\sum_{i=1}^n D_i + \partial g$ in Assumption 5.1, it remains meaningful to retain the general definition in Definition 5.4. The intermediate results listed above, which underpin

our main theorems, hold under this broader framework, provided definability is assumed. These findings may be of independent interest and could be useful for future research.

We acknowledge that Definition 5.4 can lead to an uninformative conservative field in the most general case, particularly when Φ is not definable. For instance, if f is the indicator function of the Koch curve, there exists no nonconstant absolutely continuous (a.c.) curve entirely contained within $\text{dom } \Phi$. As a result, Definition 5.4 would classify the entire space \mathbb{R}^2 as a conservative field of Φ at every point in $\text{dom } \Phi$, rendering the conservative field uninformative.

However, within the class of definable functions, which is the primary focus of this chapter, Definition 5.4 remains meaningful. By [37, Theorem 3.9], it follows that if Φ is definable, then for all but finitely many points $\hat{x} \in \text{dom } \Phi$, there exists a nonconstant a.c. curve passing through \hat{x} . This ensures that the pathological scenario exemplified by the Koch curve does not occur for a broad class of functions of interest, further justifying the validity of Definition 5.4.

5.2.2 Assumptions

We first introduce the standing assumption for all the main results.

Assumption 5.1.

1. $f_1, \dots, f_N : \mathbb{R}^n \rightarrow \mathbb{R}$ are locally Lipschitz.
2. $D_i : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a locally bounded conservative field with nonempty convex values for f_i for all $i \in \llbracket 1, N \rrbracket$.
3. $g : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is proper, convex, and locally Lipschitz in its closed domain.
4. $\Phi := f + g$ is lower bounded, where $f := f_1 + \dots + f_N$, and $D_\Phi := \sum_{i=1}^N D_i + \partial g$.

Most of the conditions above are mild and standard in the literature, as discussed in the literature review. The convex value requirement for D_i can be relaxed, by substituting D_i with its convex hull $\text{co } D_i$. Since g is locally Lipschitz in its closed domain, it holds that $\text{dom } \Phi = \text{dom } g = \text{dom } \partial g$ and ∂g has closed graph [170, Theorem 24.4]. Also, thanks to the outer sum rule of conservative

fields [28, Corollary 4] and the fact that D_Φ has closed graph (Lemma 5.2), D_Φ is a conservative field for Φ .

Remark 5.2. A common strategy in the literature (see, e.g., [155], [22], [131]) for handling unbounded set-valued mappings is to explicitly decompose g as $g = \tilde{g} + \delta_C$, where \tilde{g} is a locally Lipschitz convex function and δ_C is the indicator function of a closed convex set. However, this approach implicitly assumes that g admits a globally defined convex extension, which is not always guaranteed. For instance, consider

$$g(x, y) := \begin{cases} -2\sqrt{xy} & \text{if } xy \geq 1 \text{ and } x > 0, \\ +\infty & \text{elsewhere.} \end{cases}$$

It is clear that $\text{dom}(g)$ is a closed convex set and that g is smooth on its domain. Nevertheless, as shown in [171], g does not possess a convex extension over \mathbb{R}^2 . This example illustrates the limitations of the decomposition approach and motivates our choice to adopt a more general framework that directly handles convex functions that are locally Lipschitz on their domains.

The following assumption is needed for Theorem 5.2, where we assume access to Clarke sub-differentials of f_i . Recall that a locally Lipschitz function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is subdifferentially regular [40, 2.3.4 Definition] if the classical directional derivative exists and agrees with the generalized directional derivative, that is to say, we have

$$\lim_{t \searrow 0} \frac{f(x+th) - f(x)}{t} = \limsup_{\substack{y \rightarrow x \\ t \searrow 0}} \frac{f(y+th) - f(y)}{t}$$

for all $x, h \in \mathbb{R}^n$.

Assumption 5.2.

1. f_1, \dots, f_N are subdifferentially regular.
2. $D_i := \partial f_i$ for all $i \in \llbracket 1, N \rrbracket$.

3. Φ is weakly convex.

When both Assumptions 5.1 and 5.2 hold, we have that $D_{\Phi} = \sum_{i=1}^N \partial f_i + \partial g = \partial \Phi$ from [38, 10.9 Corollary and p. 337]. We remark that many functions arising in practice are weakly convex [172, 173, 174]. The next assumption will be used to establish convergence to stationary points of Φ (Theorem 5.4).

Assumption 5.3.

1. f_1, \dots, f_N are differentiable with locally Lipschitz continuous gradients.
2. $D_i := \{\nabla f_i\}$ for all $i \in \llbracket 1, N \rrbracket$.

5.2.3 Theorems

We are now ready to state our main theorems concerning iterates generated by proximal random reshuffling (Algorithm 1). The four theorems in this subsection treat objective functions with varying degrees of regularity, where different step size strategies are needed. None of the theorems assume global Lipschitz continuity or require prior knowledge on the iterates. The results are new even when $g = 0$, in which case Algorithm 1 reduces to random reshuffling [144, 145, 146, 30]. We discuss some applications of these results in Section 5.2.4. The proofs of the theorems are deferred to Section 5.3.

We begin by a theorem that applies to the most general setting in this chapter, where Algorithm 1 is guaranteed to reach a near approximate stationary point.

Theorem 5.1. *Let Assumption 5.1 hold and $\delta, \epsilon > 0$. For any bounded set $X_0 \subset \text{dom } \Phi$, there exists $\bar{\alpha} > 0$ such that for any sequence generated by Algorithm 1 initialized in X_0 such that*

$$\alpha_k \in (0, \bar{\alpha}] \quad \text{and} \quad \sum_{k=0}^{\infty} \alpha_k = \infty,$$

at least one iterate is (ϵ, δ) -near approximate D_{Φ} -stationary.

In the above setting, we require the step sizes to be nonsummable, which is standard for first-order methods [175, 176]. We also need the step sizes to be small, otherwise the iterates may diverge and never reach a near approximate stationary point (consider $\Phi(x) := x^4$). Examples of such step sizes include those that are eventually constant or diminishing at certain rates (for e.g., $\alpha_k = \alpha/(k+1)^p$, $\alpha \in (0, \bar{\alpha}]$, $p \in (0, 1]$).

The conclusion of Theorem 5.1 may appear weak at first glance, but as discussed in our literature review, the PRR algorithm has not been studied under such general assumptions. The closest related result we are aware of is [22, Theorem 6.2], which applies to the stochastic proximal subgradient method with replacement. Unlike the unbiased subgradient estimates assumed in [22, p. 141], PRR relies on biased estimates. This key complication is resolved by Lemma 5.1. Moreover, Theorem 5.1 only requires nonsummable step sizes, thus accommodating constant step sizes. In this regime, the result offers new insights even when iterates are bounded. The closest related work we know of considers the random reshuffling algorithm without constraints and requires uniform boundedness of iterates for all sufficiently small step sizes [23, Corollary 1, Remark 2]. However, this condition is difficult to verify and is known to hold only in special cases such as coercive objectives. As illustrated by the example in Figure 5.2, for robust matrix completion, iterates remain bounded but over increasingly large regions as step sizes decrease. Nonetheless, Theorem 5.1 remains applicable in such cases.

In the remaining theorems of this subsection, we address the convergence of Algorithm 1. By additionally assuming that Φ is weakly convex, we establish convergence of the iterates to an (ϵ, δ) -near approximate stationary point in the following theorem.

Theorem 5.2. *Let Assumptions 5.1 and 5.2 hold and $\delta, \epsilon > 0$. For all $x_0 \in \text{dom } \Phi$, there exists $T_0 \geq 0$ such that for all $T \geq T_0$, there exists $\bar{\alpha} > 0$ such that any sequence generated by Algorithm 1 such that*

$$\alpha_k \in (0, \bar{\alpha}] \quad \text{and} \quad \sum_{k=0}^{\infty} \alpha_k = T$$

converges to an (ϵ, δ) -near approximate $\partial\Phi$ -stationary point.

In this theorem, weak convexity plays a crucial role by ensuring the uniqueness of the subgra-

dient trajectory and that its velocity eventually becomes arbitrarily small (see Lemma 5.6). These two properties, together with the tracking lemma and the proposed step sizes, guarantee that the iterates converge to a (ϵ, δ) -NAS.

We next elaborate on the choice of step sizes. Given $\bar{\alpha}, T > 0$, one option is to run the algorithm with constant step size and then stop after finitely many iterations. More precisely, if we let $K \in \mathbb{N}$ be such that $T/K \in (0, \bar{\alpha}]$, then we may implement Algorithm 1 with $\alpha_k = T/K$ for all $k \in \llbracket 0, K - 1 \rrbracket$ and then terminate. In this case, the K th iterate is guaranteed to be an (ϵ, δ) -near approximate $\partial\Phi$ -stationary point. Another option is to use a sequence of geometrically decaying step sizes, i.e., $\alpha_k := \alpha\rho^k$ for some $\alpha > 0$ and $\rho \in (0, 1)$, which are known to speed up the local convergence of the (projected) subgradient method [177, 178, 179, 180], if certain sharpness condition holds around the set of global minima. In order for such step sizes to satisfy the criterion in the above theorem, it suffices to take $\alpha \in (0, \bar{\alpha}]$ and $\rho = 1 - \alpha/T$. In order to manage the overall magnitude of the sum in applications such as the nonnegative ℓ_1 matrix completion problem in Example 5.2, our theorem suggests increasing T and potentially reducing $\bar{\alpha}$. The theorem ensures that by doing so one reaches regions of interest in the state space.

In Theorem 5.2, we show the convergence to an (ϵ, δ) -near approximate stationary point, which is not necessarily close to a stationary point of Φ (i.e., a $(0, 0)$ -near approximate $\partial\Phi$ -stationary point). In fact, simple examples (for e.g., $\Phi := \exp$) can satisfy all assumptions in Theorem 5.2 yet fail to admit any stationary points. In the following theorem, we prove convergence to somewhere near a stationary point, if an assumption on D_Φ -trajectories holds.

Theorem 5.3. *Let Assumption 5.1 hold, $\epsilon > 0$, and $x_0 \in \text{dom } \Phi$. Assume that there exists a unique D_Φ -trajectory initialized at x_0 and that it converges. There exists $T_0 \geq 0$ such that for all $T \geq T_0$, there exists $\bar{\alpha} > 0$ such that any sequence generated by Algorithm 1 initialized at x_0 with*

$$\alpha_k \in (0, \bar{\alpha}] \quad \text{and} \quad \sum_{k=0}^{\infty} \alpha_k = T$$

converges to an $(\epsilon, 0)$ -near approximate D_Φ -stationary point.

We remark that D_Φ -stationary points (i.e., $(0,0)$ -near approximate D_Φ -stationary points) exist in the setting of the above theorem, as bounded D_Φ -trajectories converge to D_Φ -stationary points. We use summable step sizes in both Theorems 5.2 and 5.3 because Algorithm 1 with non-summable step sizes might not converge in either setting. Nonconvergence can easily be verified for $\Phi(x, y) := |x|/y + \delta_C(x, y)$ where δ_C is the indicator function of the set $C := \{(x, y) \in \mathbb{R}^2 : y \geq 1\}$.

In our final convergence theorem, we provide sufficient conditions for Algorithm 1 to converge to stationary points of Φ with certain nonsummable step sizes. While the regularity assumption is stronger than in the previous theorems, this theorem is connected to them due to the use of the same tracking lemma.

Theorem 5.4. *Let Assumptions 5.1 and 5.3 hold and $\beta = 0$ if $N = 1$, else $\beta \in (1/2, 1)$. If Φ is definable and has bounded subgradient trajectories, then for any bounded set $X_0 \subset \text{dom } \Phi$, there exist $\bar{\alpha}, c > 0$ such that any sequence generated by Algorithm 1 initialized in X_0 with*

$$\alpha \in (0, \bar{\alpha}] \quad \text{and} \quad \alpha_k = \frac{\alpha}{(k+1)^\beta}, \quad \forall k \in \mathbb{N}, \quad (5.3)$$

converges to a $(0,0)$ -near approximate $\partial\Phi$ -stationary point,

$$\sum_{i=0}^{\infty} \|x_{i+1} - x_i\| \leq c, \quad \text{and} \quad \min_{i \in \llbracket 0, k \rrbracket} d(0, \partial\Phi(x_i)) = o(k^{\beta-1}), \quad \forall k \in \mathbb{N}^*.$$

While one might expect to first establish boundedness of the iterates by combining Lemma 5.1 with the assumption of bounded subgradient trajectories, and then apply classical arguments to derive convergence, this approach is generally invalid. Tracking bounded subgradient trajectories does not directly imply bounded iterates. In fact, in the case $N = 1$ and $\beta = 0$, bounded iterates alone are insufficient for convergence, as illustrated in Figure 5.2. As shown in our proof in Section 5.3.5, establishing bounded iterates in this setting is just as challenging as proving convergence directly. In contrast, for $N > 1$ and $\beta \in (1/2, 1)$, the boundedness of iterates is sufficient to establish convergence, with the same $O(1/k^{1-\beta})$ rate. However, merely assuming bounded iterates

offers no control over how the constants hidden in the big-O notation vary with the initialization. By instead assuming bounded subgradient trajectories, we obtain convergence constants that are uniform across all initial points in a fixed compact set. While one could alternatively assume uniform boundedness of the iterates to achieve similar uniformity, verifying such an assumption is highly nontrivial and computationally expensive, as illustrated in Figure 5.4. In contrast, bounded subgradient trajectories are often easier to verify in practice, as shown in Example 5.4.

Remark 5.3. In practice, for Theorems 5.1 and 5.4, the constant $\bar{\alpha}$ needs to be estimated by tuning step sizes, typically by decreasing them. For Theorems 5.2 and 5.3, both $\bar{\alpha}$ and T_0 require estimation, implying adjustments of step sizes along with an increase in the number of iterations. In general, these constants depend on how accurately the iterates approximate the subgradient flow described in Lemma 5.1, making their precise determination challenging. Nevertheless, the practical tuning of step sizes and iteration counts remains feasible, and the theorems ensure convergence or desired outcomes when appropriate parameters are chosen.

Remark 5.4. Inspired by the practical stopping criteria proposed in [181], we suggest using the following criterion:

$$s_k := \frac{\|x_{k+1} - x_k\|}{\alpha_k} \leq \eta\delta, \quad (5.4)$$

where $\eta > 0$ is a tolerance parameter chosen by the user. This criterion provides a practical measure of stationarity, ensuring algorithm termination when close to a stationary point under the assumptions of Theorem 5.4. Defining the error term

$$e_k := \nabla f(x_k) - \sum_{i=1}^N \nabla f_{\sigma_i^k}(x_{k,i-1}), \quad (5.5)$$

it follows from Corollary 5.1 that $\|s_k\| \leq 2d(0, \partial\Phi(x_k)) + 2\|e_k\|$. Given that $\|e_k\| = O(\alpha_k)$ by (5.15), the stopping criterion is guaranteed to be triggered once a δ -stationary point ($(0, \delta)$ -NAS) is reached, provided sufficiently small step sizes and sufficiently large η are used. Moreover, we also obtain $d(0, \Phi(x_{k+1})) \leq 2\|s_k\| + \|e_k\|$, ensuring that the triggered stopping criterion indeed identifies a δ -stationary point. Therefore, this criterion is both practical and theoretically justified.

5.2.4 Examples

The four theorems can respectively be applied to four examples with increasing degrees of regularity. In the first example, we show that Theorem 5.1 can be applied in deep learning [182].

Example 5.1 (training of neural networks). In a typical setting of supervised learning using neural networks, we are given a training set $\{(x_1, y_1), \dots, (x_N, y_N)\}$ where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}^q$ are the feature and the label for the i th sample respectively. The goal is to find a neural network $h(\cdot, \theta) : \mathbb{R}^p \rightarrow \mathbb{R}^q$ by solving

$$\inf_{\theta \in \mathbb{R}^n} \ell(h(x_1, \theta), y_1) + \dots + \ell(h(x_N, \theta), y_N).$$

In most circumstances, h is a composition of affine functions and nonlinear activation functions parametrized by the weights $\theta \in \mathbb{R}^n$. The loss function $\ell : \mathbb{R}^q \times \mathbb{R}^q \rightarrow \mathbb{R}$ (for e.g., mean squared loss or cross entropy loss) is always lower bounded. Let $f_i := \ell(h(x_i, \cdot), y_i)$, which might not be differentiable everywhere. In practice, “derivatives” of f_i are obtained via back propagation, whose outputs constitute a conservative field D_i for f_i [28, Theorem 8].

Theorem 5.1 can readily be applied to provide a guarantee for the implementation of the stochastic subgradient method in practice [103]. Indeed, Assumption 5.1 is verified by letting $g := 0$ and Algorithm 1 reduces to random reshuffling. We would like to highlight that we do not make any assumption on coercivity of the objective function or boundedness of the iterates, in contrast to [23]. The theorem can also be applied to handle nonsmooth regularizers such as the group sparse/ ℓ_1 regularizers [183, 184].

In the next example, we see that if the objective function is weakly convex, then Theorem 5.2 can be applied to guarantee the asymptotic behavior of Algorithm 1.

Example 5.2 (nonnegative ℓ_1 matrix completion). Let $M \in \mathbb{R}^{m \times n}$ and $\Omega \subset \llbracket 1, m \rrbracket \times \llbracket 1, n \rrbracket$ be a

collection of observed entries. We seek to solve

$$\inf_{X, Y \geq 0} \sum_{(i,j) \in \Omega} |(XY - M)_{ij}|$$

where $X \in \mathbb{R}^{m \times r}$ and $Y \in \mathbb{R}^{r \times n}$. This is a nonconvex formulation of nonnegative robust principal component analysis with partial observations [185], with the rank-one case studied in [186]. Let f_1, \dots, f_N denote the summands in the objective and g denote the indicator of $\mathbb{R}_+^{m \times r} \times \mathbb{R}_+^{r \times n}$. Each function f_i is weakly convex according to [187, Lemma 4.2], hence subdifferentially regular [40, 2.5.6 Proposition]. The sum $\Phi := f_1 + \dots + f_N + g$ is then weakly convex. We may thus apply Theorem 5.2 to guarantee the convergence of Algorithm 1 to an (ϵ, δ) -near approximate $\partial\Phi$ -stationary point.

Recall that the strengthened guarantees in Theorems 5.3 and 5.4 are due to the fact that Φ has bounded and convergent subgradient trajectories. We next show two examples where this assumption can be verified.

Example 5.3 (ℓ_1 matrix sensing). Given sensing matrices $A_1, \dots, A_N \in \mathbb{R}^{m \times n}$ and measurements $b_1, \dots, b_N \in \mathbb{R}$, we aim to recover a low-rank matrix [179, 188] by solving

$$\inf_{X, Y} \sum_{i=1}^N |\langle A_i, XY \rangle - b_i|$$

where $X \in \mathbb{R}^{m \times r}$ and $Y \in \mathbb{R}^{r \times n}$. Assuming the lower bound in the ℓ_1/ℓ_2 -restricted isometry property [189, 179], the objective function has convergent subgradient trajectories, following similar arguments as in the proof of Propositions 2.3 and 2.4. Uniqueness of the subgradient trajectories is due to weak convexity of the objective function [48]. This allows us to apply Theorem 5.3 with D_i 's being the Clarke subdifferentials of the summands and $g := 0$. Therefore, with the step sizes in Theorem 5.3, Algorithm 1 is guaranteed to converge near a critical point.

Example 5.4 (nonnegative ℓ_p matrix factorization). Given $M \in \mathbb{R}^{m \times n}$ and $p \geq 2$, we aim to solve

$$\inf_{X, Y \geq 0} \|XY - M\|_p^p$$

where $(X, Y) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{r \times n}$ and $\|\cdot\|_p$ is the entrywise ℓ_p -norm. One possible way is to set each f_i to be one of $|\sum_{k=1}^r X_{\ell k} Y_{kj} - M_{\ell j}|^p$ for every $\ell = 1, \dots, m, j = 1, \dots, n$ and g to be the indicator of $\mathbb{R}_+^{m \times r} \times \mathbb{R}_+^{r \times n}$. The summands f_i 's and the function g readily satisfy Assumptions 5.1 and 5.3. According to Proposition 2.5, the objective function has bounded subgradient trajectories. Therefore, Algorithm 1 is guaranteed to converge to a stationary point by Theorem 5.4. When $N = 1$, convergence happens at the rate $o(1/k)$. When $N > 1$, convergence happens at the rate $o(1/k^{1/2-\epsilon})$ if $\beta = 1/2 + \epsilon$ for any $\epsilon \in (0, 1/2)$. We actually proved boundedness of subgradient trajectories for any $p \geq 1$, so that Theorem 5.3 can be applied when $p \in [1, 2)$.

Various algorithms have proposed for solving nonnegative ℓ_2 matrix factorization over the past three decades [73]. We briefly summarize their convergence guarantees:

1. convergence in function value: multiplicative update [190], hierarchical alternating least square [191, 192] (if the columns of the factors remain nonzero);
2. stationarity of limit points: active-set method [193], proximal gradient method [194], alternating direction method of multipliers [195];
3. convergence of bounded iterates to stationary points: proximal alternating linearized minimization [119], Bregman proximal gradient method [196, 197].

By modifying the above algorithms, one can obtain convergence guarantees to modified versions of nonnegative matrix factorization. For example, $X, Y \geq 0$ can be replaced by $X, Y \geq \epsilon$ for some small $\epsilon > 0$, or by $u \geq X, Y \geq 0$ for some large $u > 0$. A regularizer can also be added to the objective. In this vein, modified multiplicative update [198] and modified hierarchical alternating least square [199] yield bounded iterates and subsequential convergence to modified problems. The projected gradient method [200, 194] with line search and box constraints produces bounded

iterates whose limit points are stationary points of the modified problem. A similar result holds with norm-based regularizers [201].

Despite the extensive convergence results available for various algorithms applied to NMF, it may be surprising that no existing work establishes local or global convergence for solving NMF using the projected gradient method in its standard form: without alternating updates, Bregman divergences, line search, or modifications to the objective or constraints, and using only constant step sizes. This is because convergence of the forward-backward algorithm typically requires both bounded iterates and the sufficient decrease property (H1 in [21]). However, under only local gradient Lipschitz continuity and a constant step size, H1 may fail even when iterates are bounded. To address this, some authors (e.g., [197]) introduce Bregman divergences that ensure global relative smoothness, allowing the use of the H1-H3 framework with boundedness to establish convergence. On the other hand, local convergence results without boundedness rely on a quadratic growth-type condition (H4 in [21]), which is difficult to verify and likely fails in the NMF setting due to its quartic structure.

Remark 5.5. For Example 5.4, we provide a more detailed heuristic for estimating the step size $\bar{\alpha}$ according to the general principles mentioned in Remark 5.3. Specifically, given an initial point $(X_0, Y_0) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{r \times n}$ with $X_0, Y_0 > 0$, an initial step size can be chosen as $\bar{\alpha} = 1/L$, where L is an upper bound on the Hessian of $\|XY - M\|_F^2$. If this initial step size proves too large and causes divergence within a few iterations, we recommend iteratively reducing $\bar{\alpha}$ by factors of 10 until stable convergence occurs. Such tuning is commonly necessary for problems lacking gradient Lipschitz continuity and when adaptive step sizes are unavailable.

5.3 Proofs of main results

In this section, we prove the results in Section 5.2.

5.3.1 Tracking lemma

The object of this subsection is to show that the iterates generated by Algorithm 1 can be tracked by D_Φ -trajectories. This result, stated below, is one of the key technical contributions of this chapter. All of the theorems pertaining to Algorithm 1 rely on it.

Lemma 5.1. *Let Assumption 5.1 hold. For any compact set $X_0 \subset \text{dom } \Phi$ and $\epsilon, T > 0$, there exists $\bar{\alpha} > 0$ such that for any sequence generated by Algorithm 1 initialized in X_0 with $\alpha_0, \alpha_1, \dots \in (0, \bar{\alpha}]$, there exists a solution $x : [0, T] \rightarrow \text{dom } \Phi$ to $x' \in -D_\Phi(x)$ such that $x(0) \in X_0$ and*

$$\forall k \in \mathbb{N}^*, \quad \alpha_0 + \dots + \alpha_{k-1} \leq T \quad \implies \quad \|x_k - x(\alpha_0 + \dots + \alpha_{k-1})\| \leq \epsilon.$$

We clarify the contribution of Lemma 5.1 in details. In recent years, there has been significant progress in analyzing optimization algorithms through their continuous counterparts [27, 22, 28, 29, 30, 31, 202, 45, 23], while such idea traces back to the stochastic approximation literature in the 1970s [24, 25, 127]. To the best of our knowledge, Lemma 5.1 is the first result to demonstrate that such approximations remain valid over any finite time period, even without the assumption of a smooth objective function [45, Proposition 4] or bounded iterates [202, Lemma 1]. This nontrivial extension is made possible thanks to Proposition 5.3, which asserts uniform boundedness of the iterates produced by Algorithm 1 when initialized in a bounded set. In addition, we have to overcome hurdles introduced by the variable step sizes, despite the use of similar techniques as in [202, Lemma 1] and [23, Proposition 1] at several places.

We would like to further comment that, while the proof of Proposition 5.3 shares some similarities with that of Lemma 5.1, extracting a common core to streamline both proofs is not straightforward. The key distinction lies in how truncation is applied: the former relies on linear interpolations truncated at the point where the iterates exit a large ball, whereas the latter truncates at the time specified in the statement. Developing a unified approach that integrates these proofs into a single argument remains an interesting direction for future investigation.

An immediate consequence of Lemma 5.1, regarding the existence and uniqueness of trajecto-

ries associated with D_Φ can be obtained as follows.

Proposition 5.1. *Let Assumption 5.1 hold. Then for every $x_0 \in \text{dom } \Phi$, there exists a solution $x : \mathbb{R}_+ \rightarrow \text{dom } \Phi$ to $x' \in -D_\Phi(x)$ such that $x(0) = x_0$. If Assumption 5.3 also holds, then there is only one such solution.*

Proof. From Lemma 5.1, given any initial point $x_0 \in \text{dom } \Phi$, there exists a solution on $[0, T]$ to $x' \in -D_\Phi(x)$ up to any time $T \geq 0$. Let $T = 1$, then we have that there exists a solution x defined on $[0, 1]$ that is initialized at x_0 . Now we treat $x(1)$ as the new initial point and apply Lemma 5.1 again, then we obtain a solution defined on $[1, 2]$. By repeating this process, a solution defined on \mathbb{R}_+ can be constructed. If Assumption 5.3 also holds, then Φ is primal lower nice [75] at every point in $\text{dom } \Phi$. Thus, the uniqueness follows from [48, Theorem 2.9]. \square

We next provide a roadmap for the remainder of this subsection. We first study convergence towards solutions to differential inclusions with unbounded right-hand side (Proposition 5.2). It requires closedness of the graph of the sum of two set-valued mappings (Lemma 5.2). Note that Proposition 5.2 generalizes [46, Theorem 1, p. 60] by relaxing the locally boundedness requirement, and is used later for the convergence to solutions of $x' \in -D_\Phi(x)$. After that, we turn our attention to the iterates generated by Algorithm 1. We show that the displacement generated by the proximal operator is upper bounded by the distance from 0 to the subdifferential of g (Lemma 5.3), which is locally bounded (Lemma 5.4). It then follows that the distances between consecutive iterates are well controlled by the step sizes (Corollary 5.1). Combining these results, we show that iterates are uniformly bounded, given that the sum of step sizes is bounded (Proposition 5.3). Finally, we prove the approximation of iterates by D_Φ -trajectories (Lemma 5.1).

We first state some results regarding of the sum of two set-valued mappings and the convergence to its solutions.

Lemma 5.2. *If $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is locally bounded with closed graph and $G : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ has closed graph, then $F + G$ has closed graph.*

Proof. Suppose $\text{graph}(F + G) \ni (x_k, y_k) \rightarrow (x, y)$, i.e., $y_k = u_k + v_k$, for some $u_k \in F(x_k)$, $v_k \in G(x_k)$. Since F is locally bounded, u_k admits a subsequence which converges to $u \in \mathbb{R}^n$. As F has closed graph, $u \in F(x)$. Then $G(x_k) \ni v_k = y_k - u_k \rightarrow y - u$. Since G has closed graph, we have $G(x) \ni y - u$, i.e., $y \in u + G(x) \subset F(x) + G(x)$. \square

Proposition 5.2. *Let $F, G : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ have closed graphs with convex values and F be locally bounded. Assume that $F + G$ is proper. Let $I \subset \mathbb{R}$ be a bounded interval and $x_k, y_k : I \rightarrow \mathbb{R}^n$ be measurable functions such that for almost every $t \in I$ and for any neighborhood U of 0 in \mathbb{R}^{2n} , there exists $k_0 \in \mathbb{N}$ such that*

$$(x_k(t), y_k(t)) \in \text{graph}(F + G) + U, \quad \forall k \geq k_0.$$

Assume that

1. $(x_k(\cdot))_{k \in \mathbb{N}}$ is uniformly bounded in $L^\infty(I, \mathbb{R}^n)$ and $x_k(\cdot) \rightarrow x(\cdot)$ a.e. on I ;
2. $(y_k(\cdot))_{k \in \mathbb{N}}$ is uniformly bounded in $L^\infty(I, \mathbb{R}^n)$ and $y_k(\cdot) \rightarrow y(\cdot)$ weakly in $L^1(I, \mathbb{R}^n)$.

Then $(x(t), y(t)) \in \text{graph}(F + G)$ for almost all $t \in I$.

For the proof of Proposition 5.2, refer to Section 5.4.1. The proof follows a similar approach to [131, Lemma 2], which truncates the unbounded-valued component of the system to a bounded set-valued mapping. However, our result generalizes this idea, extending beyond the case where F is the subdifferential of a locally Lipschitz function and G is a normal cone. To the best of our knowledge, Proposition 5.2 does not have an exact counterpart in the literature. Consequently, we retain the statement for convenience and ease of citation.

We now move on to the study of iterates generated by Algorithm 1. We begin with a simple upper bound on the length traveled after one iteration.

Lemma 5.3. *Let $g : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ be proper and convex. For any $x \in \text{dom } \partial g$, $\alpha > 0$, $d \in \mathbb{R}^n$, it holds that*

$$\frac{1}{\alpha} \|\text{prox}_{\alpha g}(x - \alpha d) - x\| \leq 2\|d\| + 2d(0, \partial g(x)).$$

Proof. Let $y := \text{prox}_{\alpha g}(x - \alpha d)$. Applying definition of the proximal operator, it holds that

$$g(y) + \frac{1}{2\alpha} \|y - (x - \alpha d)\|^2 \leq g(x) + \frac{1}{2\alpha} \|x - (x - \alpha d)\|^2.$$

Rearranging the above inequality yields that

$$\frac{1}{2\alpha} \|y - x\|^2 \leq g(x) - g(y) + \langle x - y, d \rangle. \quad (5.6)$$

Meanwhile, by convexity of g , for any $v \in \partial g(x)$, we have $g(y) \geq g(x) + \langle v, y - x \rangle$. Together with (5.6), this yields that $\|y - x\|^2 / (2\alpha) \leq \langle x - y, d + v \rangle$. By Cauchy-Schwarz and triangular inequalities, we have

$$\frac{1}{2\alpha} \|y - x\|^2 \leq \langle x - y, d + v \rangle \leq \|y - x\| \|d + v\| \leq \|y - x\| (\|d\| + \|v\|).$$

We conclude by taking the infimum with respect to $v \in \partial g(x)$ on the right hand side of the above inequality. \square

The bound on the length after one iteration previously obtained will prove useful once we can control the distance between the origin and ∂g . This is the object of the next result, which can be regarded as a simple corollary of the norm preserving extension result of convex Lipschitz functions in [203]. Let $B(S, r) := S + B(0, r)$.

Lemma 5.4. *If $g : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is proper, convex, locally Lipschitz on its domain, and has closed domain, then $d(0, \partial g)$ is locally bounded over $\text{dom } \partial g$ and $\text{dom } \partial g = \text{dom } g$.*

Proof. Let $X \subset \mathbb{R}^n$ be any compact set. Notice that $Y := \text{dom } g \cap B(\text{co } X, 1)$ is a convex compact set, so the restriction of g on Y , denoted as $g|_Y : Y \rightarrow \mathbb{R}$, is an L -Lipschitz convex function on Y . Applying [203, Theorem 1], we obtain an L -Lipschitz convex function $\hat{g} : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $\hat{g}(x) = g|_Y(x)$ for all $x \in Y$. Thus, $g(x) = \hat{g}(x) + \delta_Y(x)$ for all $x \in B(\text{co } X, 1)$ and $\partial g(x) = \partial \hat{g}(x) + N_Y(x)$ for all $x \in X$, where $N_Y(x)$ denotes the normal cone of Y at x . Since $0 \in N_Y(x)$ for

all $x \in X$, we have that $d(0, \partial g(x)) \leq d(0, \partial \hat{g}(x)) \leq L$ for all $x \in X$. Thus, $d(0, \partial g)$ is locally bounded over $\text{dom } g$. \square

Combining Lemmas 5.3 and 5.4, one can control the length traveled after one iteration directly via the step sizes. To see why, it will be convenient to use the notation $\|S\| := \sup\{\|s\| : s \in S\}$ for any $S \subset \mathbb{R}^n$ and $r \geq 0$.

Corollary 5.1. *Suppose Assumption 5.1 holds. Let $X \subset \mathbb{R}^n$ be bounded and x_0, \dots, x_{K+1} be generated by Algorithm 1. Consider the constants*

$$L_i := \sup_{B(X,1)} \|D_i\|, \quad L := \max\{L_1, \dots, L_N\}, \quad L_g := \sup_{B(X,1) \cap \text{dom } g} d(0, \partial g),$$

and assume that $x_0, \dots, x_K \in X$.

1. If $\alpha_0, \dots, \alpha_K \leq 1/(NL)$, then $\{x_{k,i}\}_{k \in \llbracket 0, K \rrbracket}^{i \in \llbracket 1, N \rrbracket} \subset B(X, 1)$.
2. If $\alpha_0, \dots, \alpha_K \leq 1/(2NL + 2L_g)$, then $\|x_{k+1} - x_k\| \leq 2(NL + L_g)\alpha_k$ for all $k \in \llbracket 0, K \rrbracket$ and $x_{K+1} \in B(X, 1)$.

Proof. We first note that $L_g < \infty$ by Lemma 5.4 and a standard compactness argument. The first part can be proved by induction, with hypothesis $H_i : “\forall k \in \llbracket 0, K \rrbracket, \|x_{k,i-1} - x_k\| \leq (i-1)/N”$. By Algorithm 1, $x_{k,0} = x_k \in X$ for all $k \in \llbracket 0, K \rrbracket$, hence H_1 . If H_i holds, then $x_{k,i-1} \in B(X, 1)$ and

$$\begin{aligned} \|x_{k,i} - x_k\| &\leq \|x_{k,i} - x_{k,i-1}\| + \|x_{k,i-1} - x_k\| \leq \alpha_k \|D_{\sigma_i^k}(x_{k,i-1})\| + \frac{i-1}{N} \\ &\leq \frac{1}{NL} \cdot L_i + \frac{i-1}{N} \leq \frac{i}{N}, \end{aligned}$$

hence H_{i+1} holds. As for the second part, by Lemma 5.3 and part one, we have

$$\|x_{k+1} - x_k\| \leq 2\alpha_k \left(\left\| \sum_{i=1}^N D_{\sigma_i^k}(x_{k,i-1}) \right\| + d(0, \partial g(x_k)) \right) \leq 2\alpha_k(NL + L_g) \leq 1.$$

Finally, since $x_K \in X$, it holds that $x_{K+1} \in B(X, 1)$. \square

We next record a property of D_Φ -trajectories.

Lemma 5.5. *Let Assumption 5.1 hold and let I be an interval of \mathbb{R}_+ . If $x : I \rightarrow \mathbb{R}^n$ is a solution to $x' \in -D_\Phi(x)$, then $\Phi \circ x$ is differentiable almost everywhere on I with*

$$(\Phi \circ x)'(t) = -\|x'(t)\|^2 \quad \text{and} \quad \|x'(t)\| = d(0, D_\Phi(x(t))), \quad \text{for almost every } t \in I.$$

Proof of the lemma is omitted as it follows immediately from the definition of conservative fields and [55, Proposition 4.10 & Lemma 4.11]. We next prove the final result needed for the proof of Lemma 5.1. We show that iterates generated by Algorithm 1 remain in a bounded set, whose radius only depends on the sum of step sizes.

Proposition 5.3. *Let Assumption 5.1 hold. Then for any bounded set $X_0 \subset \text{dom } \Phi$ and $T > 0$, there exist $\bar{\alpha}, r > 0$ such that for any sequence generated by Algorithm 1 initialized in X_0 with $\alpha_0, \alpha_1, \dots \leq \bar{\alpha}$, we have*

$$\forall k \in \mathbb{N}^*, \quad \alpha_0 + \dots + \alpha_{k-1} \leq T \implies x_0, \dots, x_k \in B(0, r).$$

Proof. We assume without loss of generality that $X_0 \subset \text{dom } \Phi$ is nonempty and compact. Fix $T > 0$. There exists $Q > 0$ such that for any absolutely continuous $x : [0, T] \rightarrow \mathbb{R}^n$ that satisfies $x(0) \in X_0$ and is a solution to

$$x' \in -D_\Phi(x) = -\sum_{i=1}^N D_i(x) - \partial g(x), \tag{5.7}$$

it holds that $d(x(\tilde{T}), X_0) \leq \|x(\tilde{T}) - x(0)\| \leq \int_0^{\tilde{T}} \|x'(t)\| dt \leq \int_0^T \|x'(t)\| dt \leq Q$ for all $\tilde{T} \in [0, T]$.

Indeed, let $x(\cdot)$ be any such function, we have

$$\int_0^T \|x'(t)\| dt \leq \sqrt{T} \sqrt{\int_0^T \|x'(t)\|^2 dt} \quad (5.8a)$$

$$= \sqrt{T} \sqrt{\int_0^T -(\Phi \circ x)'(t) dt} \quad (5.8b)$$

$$= \sqrt{T} \sqrt{\Phi(x(0)) - \Phi(x(T))} \quad (5.8c)$$

$$\leq \sqrt{T} \sqrt{\sup_{x_0 \in X_0} \Phi(x_0) - \inf_{y \in \mathbb{R}^n} \Phi(y)} =: Q. \quad (5.8d)$$

Above, (5.8a) follows from the Cauchy-Schwarz inequality and (5.8b) follows from Lemma 5.5. Let $R > 0$ such that $X_0 \subset B(0, R)$ and $L > 0$ such that $D_i(B(0, R + Q + 2)) \subset B(0, L)$ for all $i \in \llbracket 1, N \rrbracket$. By Lemma 5.4 and a standard compactness argument, there exists $L_g > 0$ such that $d(0, \partial g(x)) \leq L_g$ for all $x \in B(0, R + Q + 2) \cap \text{dom } \Phi$.

We next reason by contradiction and assume that for any $r > 0$, there exist a positive sequence $\bar{\alpha}_m \rightarrow 0$, a sequence $(K_m)_{m \in \mathbb{N}}$ of natural numbers, and sequences of iterates $(x_k^m)_{k \in \mathbb{N}}$ generated by Algorithm 1 with step sizes $\alpha_0^m, \alpha_1^m, \dots \leq \bar{\alpha}_m$, $\sum_{k=0}^{K_m-1} \alpha_k^m \leq T$, and $x_0^m \in X_0$ such that $\max\{\|x_k^m\| : k = 0, \dots, K_m\} > r$ for any $m \in \mathbb{N}$. Take $r = R + Q + 2$ and assume that $\bar{\alpha}_m \leq 1/2(NL + L_g)$ without loss of generality. For each $m \in \mathbb{N}$, let $k_m := \min\{k \in \mathbb{N} : x_k^m \in B(0, r), x_{k+1}^m \notin B(0, r)\}$. We have that $k_m \leq K_m - 1$ following our assumption. Thus $\sum_{k=0}^{k_m-1} \alpha_k^m \leq T$ for any $m \in \mathbb{N}$ and $\bar{T} := \liminf_{m \rightarrow \infty} \sum_{k=0}^{k_m-1} \alpha_k^m \in [0, T]$. By taking a subsequence if necessary, assume that $\lim_{m \rightarrow \infty} \sum_{k=0}^{k_m-1} \alpha_k^m = \bar{T}$.

Let $T_0^m := 0$ and $T_k^m := \sum_{i=0}^{k-1} \alpha_i^m$ for all $k \in \mathbb{N}^*$ and $m \in \mathbb{N}$. For each sequence $x_0^m, x_1^m, \dots, x_{k_m}^m$, consider the (extended) linear interpolation $\bar{x}^m : [0, \max\{T_{k_m}^m, \bar{T}\}] \rightarrow \mathbb{R}^n$ defined by

$$\bar{x}^m(t) = x_k^m + (t - T_k^m) \frac{x_{k+1}^m - x_k^m}{\alpha_k^m}$$

for any $t \in [T_k^m, T_{k+1}^m]$ and $k \in \{0, 1, \dots, k_m - 1\}$. Also, $\bar{x}^m(t) = x_{k_m}^m$ for $t \in [T_{k_m}^m, \bar{T}]$ if $T_{k_m}^m < \bar{T}$. As $x_k^m \in B(0, r)$ for $k = 0, \dots, k_m$, we know that $\bar{x}^m(t) \in B(0, r)$ for all $t \in [0, \max\{T_{k_m}^m, \bar{T}\}]$

by the convexity of $B(0, r)$. For any $t \in (T_k^m, T_{k+1}^m)$ and $k \in \{0, 1, \dots, k_m - 1\}$, it holds that $(\bar{x}^m)'(t) = (x_{k+1}^m - x_k^m)/\alpha_k^m \in B(0, 2(NL + L_g))$ by Corollary 5.1. Also, $(\bar{x}^m)'(t) = 0$ for any $t \in (T_{k_m}^m, \bar{T})$ if $T_{k_m}^m < \bar{T}$. By successively applying the Arzelà-Ascoli and the Banach-Alaoglu theorems [46, Theorem 4 p. 13], there exist a subsequence (again denoted $(\bar{x}^m(\cdot))_{m \in \mathbb{N}}$) and an absolutely continuous function $x : [0, \bar{T}] \rightarrow \mathbb{R}^n$ such that $\bar{x}_{[[0, \bar{T}]]}^m(\cdot)$ converges uniformly to $x(\cdot)$ and $(\bar{x}_{[[0, \bar{T}]]}^m)'(\cdot)$ converges weakly to $x'(\cdot)$ in $L^1([0, \bar{T}], \mathbb{R}^n)$. In addition, for almost every $t \in (0, \bar{T})$, since $T_{k_m}^m \rightarrow \bar{T}$, for sufficiently large m , it holds that $t \in (T_k^m, T_{k+1}^m)$ for some $k \in \llbracket 0, k_m - 1 \rrbracket$. Therefore, for any neighborhood U of 0 and for all sufficiently large m , it holds that

$$(\bar{x}^m)'(t) = \frac{x_{k+1}^m - x_k^m}{\alpha_k^m} \quad (5.9a)$$

$$= \sum_{i=1}^N \frac{x_{k,i}^m - x_{k,i-1}^m}{\alpha_k^m} + \frac{x_{k+1}^m - x_{k,N}^m}{\alpha_k^m} \quad (5.9b)$$

$$\in - \sum_{i=1}^N D_i(x_{k,i-1}^m) - \partial g(x_{k+1}^m) \cap B(0, 2L_2) \quad (5.9c)$$

$$\subset - \sum_{i=1}^N D_i(x(t)) - \partial g(x(t)) + U. \quad (5.9d)$$

Above, (5.9c) follows from [17, Theorem 6.39] and Lemma 5.3. (5.9d) is due to the fact that $x_{k,i}^m \rightarrow x(t)$ for all $i \in \llbracket 0, N - 1 \rrbracket$, $x_{k+1}^m \rightarrow x(t)$ as $m \rightarrow \infty$, and that D_i and $\partial g \cap B(0, 2L_2)$ are locally bounded with closed graphs by Assumption 5.1.

By Lemma 5.2, $D_1 + \dots + D_N$ is locally bounded with closed graph. According to Proposition 5.2, $x(\cdot)$ satisfies (5.7) for almost every $t \in (0, \bar{T})$. Also, $x(0) = \lim_{m \rightarrow \infty} \bar{x}^m(0) \in X_0$. Recall that $x(\bar{T}) \in B(X_0, Q) \subset B(0, R + Q)$. Notice that $\lim_{m \rightarrow \infty} \bar{x}^m(\bar{T}) = x(\bar{T}) \in B(0, R + Q)$ and $\|\bar{x}^m(\bar{T}) - x_{k_m}^m\| = \|\bar{x}^m(\bar{T}) - \bar{x}^m(T_{k_m}^m)\| \leq (2 + 2L_2)|\bar{T} - T_{k_m}^m| \rightarrow 0$ as $m \rightarrow \infty$. Thus $x_{k_m}^m \in B(0, R + Q + 1)$ for all sufficiently large m . By Corollary 5.1, we have that $\|x_{k_m+1}^m - x_{k_m}^m\| \leq 2\alpha_k^m(NL + L_g) \leq 1$, contradicting the assumption that $x_{k_m+1}^m \notin B(0, R + Q + 2)$. \square

We are now ready to prove the desired result.

Proof of Lemma 5.1. Let $X_0 \subset \text{dom } \Phi$ be nonempty and compact. Let $T > 0$. By Proposition 5.3,

there exist $\hat{\alpha}, r > 0$ such that for any sequence generated by Algorithm 1 initialized in X_0 with $\alpha_0, \alpha_1, \dots \leq \bar{\alpha}$, we have

$$\forall k \in \mathbb{N}^*, \quad \alpha_0 + \dots + \alpha_{k-1} \leq T \implies x_0, \dots, x_k \in B(0, r).$$

Let $(\bar{\alpha}_m)_{m \in \mathbb{N}}$ be a positive sequence that converges to zero. We assume that $\bar{\alpha}_m \leq \min\{1, \hat{\alpha}\}$. To each $\bar{\alpha}_m$, we attribute a sequence of iterates $(x_k^m)_{k \in \mathbb{N}}$ generated by Algorithm 1 with step sizes $\alpha_0^m, \alpha_1^m, \dots \leq \bar{\alpha}_m$ and $x_0^m \in X_0$.

Let $T_0^m := 0$ and $T_k^m := \sum_{i=0}^{k-1} \alpha_i^m$ for all $k \in \mathbb{N}^*$ and $m \in \mathbb{N}$. Let $T^m := \min\{\sum_{k=0}^{\infty} \alpha_k^m, T\}$. By taking a subsequence if necessary, we have that $T^m \rightarrow \bar{T} \in [0, T]$ as $m \rightarrow \infty$. For each sequence $(x_k^m)_{k \in \mathbb{N}}$, consider the (extended) linear interpolation $\bar{x}^m : [0, \max\{T^m, \bar{T}\}] \rightarrow \mathbb{R}^n$ defined by

$$\bar{x}^m(t) = x_k^m + (t - T_k^m) \frac{x_{k+1}^m - x_k^m}{\alpha_k^m}$$

for any $t \in [T_k^m, \min\{T_{k+1}^m, T^m\}]$ and $k \in \mathbb{N}$ such that $T_k^m < T^m$. If $T^m < \bar{T}$, then $\bar{x}^m(t) := \lim_{s \nearrow T^m} \bar{x}^m(s)$ for all $t \in [T^m, \bar{T}]$. For any $m \in \mathbb{N}$, let $k_m := \sup\{k \in \mathbb{N} : T_k^m \leq T^m\}$. If $k_m = \infty$, then we have that $x_k^m \in B(0, r)$ for all $k \in \mathbb{N}$. Otherwise if $k_m < \infty$, then $T_{k_m+1}^m = T_{k_m}^m + \alpha_{k_m}^m \leq T^m + \bar{\alpha}_m \leq T + 1$, and thus $x_0^m, \dots, x_{k_m+1}^m \in B(0, r)$. In both cases, we know that $\bar{x}^m(t) \in B(0, r)$ for all $t \in [0, \max\{T^m, \bar{T}\}]$ by the convexity of $B(0, r)$. Using arguments similar as in the proof of Proposition 5.3, by passing to a subsequence if necessary, $(\bar{x}_{[0, \bar{T}]}^m(\cdot))_{m \in \mathbb{N}}$ converges uniformly to a solution $x : [0, \bar{T}] \rightarrow \mathbb{R}^n$ to $x' \in -D_\Phi(x)$ with $x(0) = \lim_{m \rightarrow \infty} \bar{x}^m(0) = \lim_{m \rightarrow \infty} x_0^m \in X_0$.

The conclusion of the lemma now follows. To see why, assume the contrary that there exists $\epsilon > 0$ such that for any $\bar{\alpha} > 0$, there exists $(x_k)_{k \in \mathbb{N}}$ generated by Algorithm 1 with step sizes $(\alpha_k)_{k \in \mathbb{N}} \subset (0, \bar{\alpha}]$ and $x_0 \in X_0$ such that for any absolutely continuous function $x : [0, T] \rightarrow \mathbb{R}^n$ is a solution to $x' \in -D_\Phi(x)$ with $x(0) \in X_0$, there exists $K \in \mathbb{N}^*$ such that

$$\alpha_0 + \dots + \alpha_{K-1} \leq T \quad \text{and} \quad \|x_K - x(\alpha_0 + \dots + \alpha_{K-1})\| > \epsilon.$$

Based on this, we may construct a sequence of iterates $(x_k^m)_{k \in \mathbb{N}}$ generated by Algorithm 1 with step sizes $(\alpha_k^m)_{k \in \mathbb{N}}$ and $x_0 \in X_0$, where $\sup_k \alpha_k^m \rightarrow 0$ as $m \rightarrow \infty$. Moreover, for each $(x_k^m)_{k \in \mathbb{N}}$ and any solution $x : [0, T] \rightarrow \mathbb{R}^n$ to $x' \in -D_\Phi(x)$ with $x(0) \in X_0$, there exists $K_m \in \mathbb{N}^*$ such that

$$\alpha_0^m + \cdots + \alpha_{K_m-1}^m \leq T \quad \text{and} \quad \|x_{K_m}^m - x(\alpha_0^m + \cdots + \alpha_{K_m-1}^m)\| > \epsilon.$$

This is in contradiction with what we have shown above, namely the uniform convergence of the linear interpolations to a solution to $x' \in -D_\Phi(x)$. \square

5.3.2 Reachability of (ϵ, δ) -near approximate stationarity

Proof of Theorem 5.1. We assume without loss of generality that $X_0 \subset \text{dom } \Phi$ is nonempty and compact. Fix any $\delta, \epsilon > 0$ and let $T := (\sup_{y \in X_0} \Phi(y) - \inf_{x \in \mathbb{R}^n} \Phi(x)) / \delta^2 \in \mathbb{R}_+$. Let $x : [0, T] \rightarrow \mathbb{R}^n$ be a solution to $x' \in -D_\Phi(x)$ with $x(0) \in X_0$. Following the same arguments as (5.8), there exists $r > 0$ such that $x([0, T]) \subset \text{dom } \Phi \cap B(X_0, r)$ for any such $x(\cdot)$. According to Lemma 5.4 and a standard compactness argument, $d(0, \partial g)$ is bounded over compact subsets of $\text{dom } \Phi$. Thus, there exists $L > 0$ such that $d(0, D_\Phi(y)) \leq L$ for all $y \in \text{dom } \Phi \cap B(X_0, r)$.

By Lemma 5.1, there exists $\bar{\alpha} \in (0, \epsilon/(2L)]$ such that for any sequence generated by Algorithm 1 with $\alpha_0, \alpha_1, \dots \leq \bar{\alpha}$, $\sum_{k=0}^\infty \alpha_k = \infty$, and $x_0 \in X_0$, there exists a solution $x : [0, T] \rightarrow \mathbb{R}^n$ to $x' \in -D_\Phi(x)$ such that $x(0) \in X_0$ and

$$\forall k \in \mathbb{N}^*, \quad \alpha_0 + \cdots + \alpha_{k-1} \leq T \quad \implies \quad \|x_k - x(\alpha_0 + \cdots + \alpha_{k-1})\| \leq \epsilon/2.$$

By Lemma 5.5 it holds that

$$\int_0^T d(0, D_\Phi(x(t)))^2 dt = \Phi(x(0)) - \Phi(x(T)) \leq \sup_{y \in X_0} \Phi(y) - \inf_{x \in \mathbb{R}^n} \Phi(x).$$

Thus, there exists $t \in [0, T]$ such that $d(0, D_\Phi(x(t))) \leq \delta$. As $\alpha_k \leq \bar{\alpha}$, there exists $k \in \mathbb{N}$ such that $t_k := \sum_{i=0}^{k-1} \alpha_i \in [\max\{0, t - \bar{\alpha}\}, t]$.

We next show that $\|x_k - x(t)\| \leq \epsilon$, then the conclusion of the theorem follows. Indeed,

$$\|x_k - x(t)\| \leq \|x_k - x(t_k)\| + \|x(t_k) - x(t)\| \quad (5.10a)$$

$$\leq \epsilon/2 + \int_{t_k}^t \|x'(s)\| ds \quad (5.10b)$$

$$= \epsilon/2 + \int_{t_k}^t d(0, D\Phi(x(s))) ds \quad (5.10c)$$

$$\leq \epsilon/2 + (t - t_k)L \quad (5.10d)$$

$$\leq \epsilon/2 + \bar{\alpha}L \leq \epsilon. \quad (5.10e)$$

Above, (5.10a) and (5.10b) follow from the triangular inequality. (5.10c) is a result of again by Lemma 5.5 and (5.10d) follows from the fact that $d(0, D\Phi)$ is locally bounded over $\text{dom } \Phi$. \square

5.3.3 Convergence to (ϵ, δ) -near approximate stationarity

The proof of Theorem 5.2 requires the following lemma, whose proof is omitted as it essentially follows from the proof of [204, Theorem 6.2]. We remark that weak convexity implies that the objective function is primal lower nice [75] everywhere in its domain, with the same constants.

Lemma 5.6. *Let Φ be proper, weakly convex, and lower bounded. For any subgradient trajectory $x : \mathbb{R}_+ \rightarrow \mathbb{R}^n$ of Φ and for any $\epsilon > 0$, there exists $T \geq 0$ such that $\|x'(t)\| \leq \epsilon$ for almost every $t \geq T$.*

We proceed to prove the theorem.

Proof of Theorem 5.2. Fix any $\delta, \epsilon > 0$. As Φ is weakly convex and lower bounded, for any $x_0 \in \text{dom } \Phi$, there exists a unique subgradient trajectory $x : \mathbb{R}_+ \rightarrow \mathbb{R}^n$ of Φ initialized at x_0 [48]. By Lemmas 5.5 and 5.6, there exists $T_0 > 0$ such that $d(0, \partial\Phi(x(t))) = \|x'(t)\| \leq \delta$ for almost every $t \geq T_0$. Since $\text{graph } \partial\Phi$ is closed, $\text{epi } d(0, \partial\Phi)$ is closed. Fix $t \geq T_0$ and let $t_k \rightarrow t$ be such that $(x(t_k), \delta) \in \text{epi } d(0, \partial\Phi)$. By continuity of $x(\cdot)$, $(x(t_k), \delta) \rightarrow (x(t), \delta) \in \text{epi } d(0, \partial\Phi)$. In other words, $d(0, \partial\Phi(x(t))) \leq \delta$ actually holds for all $t \geq T_0$.

By the subdifferential regularity of f_1, \dots, f_N and the convexity of g , it holds that $\partial\Phi = \partial f_1 + \dots + \partial f_N + \partial g$ [38, 10.9 Corollary, p. 430]. Thus by Lemma 5.1, for any $T \geq T_0$, there exists $\bar{\alpha} > 0$ such that for any sequence generated by Algorithm 1 with $\alpha_0, \alpha_1, \dots \leq \bar{\alpha}$, $\sum_{k=0}^{\infty} \alpha_k = T$, and $x_0 \in \text{dom } \Phi$, it holds that

$$\forall k \in \mathbb{N}^*, \quad \|x_k - x(\alpha_0 + \dots + \alpha_{k-1})\| \leq \epsilon.$$

As a result, it holds that $\{x_k\}_{k \in \mathbb{N}} \subset B(x([0, T]), \epsilon)$, and that any limit point x^* of the sequence must lie in $B(x(T), \epsilon)$. As $d(0, \partial\Phi(x(T))) \leq \delta$, it remains to show that the sequence $(x_k)_{k \in \mathbb{N}}$ is convergent. Indeed, by Corollary 5.1, there exists $C > 0$ such that $\|x_{k+1} - x_k\| \leq C\alpha_k$ for all $k \in \mathbb{N}$. Therefore, $\sum_{k=0}^{\infty} \|x_{k+1} - x_k\| \leq C \sum_{k=0}^{\infty} \alpha_k = CT$, thus the sequence $(x_k)_{k \in \mathbb{N}}$ is convergent. \square

5.3.4 Convergence to $(\epsilon, 0)$ -near approximate stationarity

The proof of Theorem 5.3 is a relatively direct consequence of the tracking lemma (Lemma 5.1).

Proof. Proof of Theorem 5.3.

Fix any $\epsilon > 0$ and $x_0 \in \text{dom } \Phi$. According to the assumption, there exists a unique solution $x : \mathbb{R}_+ \rightarrow \mathbb{R}^n$ to $x'(t) \in -D\Phi(x(t))$ such that $x(0) = x_0$. By Lemma 5.5 and the fact that $x(\cdot)$ is convergent, we know that $x^\sharp := \lim_{t \rightarrow \infty} x(t)$ must be a $D\Phi$ -critical point. Also, there exists a $T_0 > 0$ such that $x([T_0, \infty)) \subset B(x^\sharp, \epsilon/2)$.

For any $T \geq T_0$, by Lemma 5.1, there exists $\bar{\alpha} > 0$ such that for any sequence generated by Algorithm 1 with $\alpha_0, \alpha_1, \dots \leq \bar{\alpha}$ and $\sum_{k=0}^{\infty} \alpha_k = T$, we have

$$\|x_k - x(\alpha_0 + \dots + \alpha_{k-1})\| \leq \epsilon/2 \tag{5.11}$$

for any $k \in \mathbb{N}^*$. Using similar arguments as in the proof of Theorem 5.2, the sequence $(x_k)_{k \in \mathbb{N}}$ is convergent. Taking $k \rightarrow \infty$ in (5.11), we have that $\|x^* - x(T)\| \leq \epsilon/2$, where $x^* := \lim_{k \rightarrow \infty} x_k$. Therefore, $\|x^* - x^\sharp\| \leq \|x^* - x(T)\| + \|x(T) - x^\sharp\| \leq \epsilon/2 + \epsilon/2 = \epsilon$. \square

5.3.5 Convergence to $(0, 0)$ -near approximate stationarity

The proof of Theorem 5.4 is derived from the following steps. We first establish an approximate descent lemma for the iterates generated by Algorithm 1 in Lemma 5.7. A length formula is then proved in Proposition 5.4 by using the approximate descent lemma (Lemma 5.7), the uniform Kurdyka-Łojasiewicz inequality (Lemma 5.8) and the upper bound on the distance between two successive iterations of Algorithm 1 (Corollary 5.1). We then obtain the convergence of iterates in Theorem 5.4.

Different versions of approximate descent for Algorithm 1 have been established in the literature, even though it does not fit into the H1-H2-H3 framework in [21] or [20]. Indeed, the sufficient decrease and relative error condition cannot be guaranteed due to random reshuffling. An approximate descent property of $\mathbb{E}[\Phi(x_k)]$ for Algorithm 1 was proved in [156, E.2]. The approximate descent of $\Phi(x_k)$ for the special case of Algorithm 1 where $g = 0$ was proved in [132, Lemma 3.2]. In contrast to existing work, we do not require global Lipschitz continuity of ∇f_i 's nor $g = 0$. For our purposes, it is sufficient to restrict the iterates to a bounded subset, which incidentally makes the proof more direct.

Lemma 5.7. *Suppose Assumptions 5.1 and 5.3 hold. Let X be a bounded set. There exists $\bar{\alpha} > 0$ such that if $(x_0, \dots, x_{K+1}) \in X \times \dots \times X \times \mathbb{R}^n$ is generated by Algorithm 1 with $\alpha_0, \dots, \alpha_K \leq \bar{\alpha}$, then for all $k \in \llbracket 0, K \rrbracket$ we have*

$$\Phi(x_{k+1}) \leq \Phi(x_k) - \frac{\alpha_k}{4} d(0, \partial\Phi(x_{k+1}))^2 - \frac{1}{8\alpha_k} \|x_{k+1} - x_k\|^2 + \frac{1}{12} (N-1)^2 N (2N-1) L^2 M^2 \alpha_k^3$$

where L and M are respectively Lipschitz constants of f_1, \dots, f_N on $B(X, 2)$ and $\nabla f_1, \dots, \nabla f_N$ on $\text{co } B(X, 1)$.

Proof. Let $L_g := \sup_{x \in B(X, 1) \cap \text{dom } g} d(0, \partial g(x))$ and $\bar{\alpha} := \min\{1/(2NL + 2L_g), 1/(2NM)\}$. Suppose $(x_0, \dots, x_{K+1}) \in X \times \dots \times X \times \mathbb{R}^n$ is generated by Algorithm 1 with $\alpha_0, \dots, \alpha_K \leq \bar{\alpha}$. By Corollary 5.1, $x_{K+1} \in B(X, 1)$. Let $k \in \llbracket 0, K \rrbracket$. Since NM is a Lipschitz constant of ∇f over

co $B(X, 1)$, we have

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{NM}{2} \|x_{k+1} - x_k\|^2.$$

It holds also that $g(x_{k+1}) \leq g(x_k) + \langle \partial g(x_{k+1}), x_{k+1} - x_k \rangle$ by convexity of g . Added to the previous inequality, this yields

$$\Phi(x_{k+1}) \leq \Phi(x_k) + \langle \nabla f(x_k) + \partial g(x_{k+1}), x_{k+1} - x_k \rangle + \frac{NM}{2} \|x_{k+1} - x_k\|^2. \quad (5.12)$$

We proceed the second term on the right hand side. For all $g'(x_{k+1}) \in \partial g(x_{k+1})$, it holds that

$$\langle \nabla f(x_k) + g'(x_{k+1}), x_{k+1} - x_k \rangle = \alpha_k \langle \nabla f(x_k) + g'(x_{k+1}), (x_{k+1} - x_k)/\alpha_k \rangle \quad (5.13a)$$

$$= \alpha_k \|\nabla f(x_k) + g'(x_{k+1}) + (x_{k+1} - x_k)/\alpha_k\|^2/2 \quad (5.13b)$$

$$- \alpha_k \|\nabla f(x_k) + g'(x_{k+1})\|^2/2 - \|x_{k+1} - x_k\|^2/(2\alpha_k). \quad (5.13c)$$

On the one hand,

$$\begin{aligned} d(0, \partial \Phi(x_{k+1}))^2 &= d(0, \nabla f(x_{k+1}) + \partial g(x_{k+1}))^2 \\ &\leq 2\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 + 2d(0, \nabla f(x_k) + \partial g(x_{k+1}))^2 \\ &\leq 2 \left(\sum_{i=1}^N \|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\| \right)^2 + 2\|\nabla f(x_k) + g'(x_{k+1})\|^2 \\ &\leq 2N^2 M^2 \|x_{k+1} - x_k\|^2 + 2\|\nabla f(x_k) + g'(x_{k+1})\|^2. \end{aligned} \quad (5.14)$$

On the other hand, since $x_{k+1} = \text{prox}_{\alpha_k g}(x_{k,N})$, by Fermat's rule [38, Theorem 10.1] we have $0 \in \alpha_k \partial g(x_{k+1}) + x_{k+1} - x_{k,N}$. Recall that $x_{k,N} = x_k - \alpha_k \sum_{i=1}^N \nabla f_{\sigma_i^k}(x_{k,i-1})$. Thus there exists $g'(x_{k+1}) \in \partial g(x_{k+1})$ such that

$$0 = g'(x_{k+1}) + \frac{x_{k+1} - x_k}{\alpha_k} + \sum_{i=1}^N \nabla f_{\sigma_i^k}(x_{k,i-1}).$$

By the QM-AM inequality [205], we hence have

$$\begin{aligned}
\|\nabla f(x_k) + g'(x_{k+1}) + (x_{k+1} - x_k)/\alpha_k\|^2 &= \left\| \nabla f(x_k) - \sum_{i=1}^N \nabla f_{\sigma_i^k}(x_{k,i-1}) \right\|^2 \\
&= \left\| \sum_{i=1}^N \nabla f_{\sigma_i^k}(x_k) - \nabla f_{\sigma_i^k}(x_{k,i-1}) \right\|^2 \\
&\leq (N-1) \sum_{i=2}^N \|\nabla f_{\sigma_i^k}(x_k) - \nabla f_{\sigma_i^k}(x_{k,i-1})\|^2 \\
&= (N-1)M^2 \sum_{i=2}^N \|x_k - x_{k,i-1}\|^2 \\
&= (N-1)M^2\alpha_k^2 \sum_{i=2}^N \left\| \sum_{j=1}^{i-1} \nabla_{\sigma_j^k} f(x_{k,j-1}) \right\|^2 \\
&\leq (N-1)M^2\alpha_k^2 \sum_{i=2}^N (i-1) \sum_{j=1}^{i-1} \|\nabla_{\sigma_j^k} f(x_{k,j-1})\|^2 \\
&\leq (N-1)L^2M^2\alpha_k^2 \sum_{i=1}^{N-1} i^2 \\
&\leq \frac{1}{6}(N-1)^2N(2N-1)L^2M^2\alpha_k^2. \tag{5.15}
\end{aligned}$$

Above, we use the fact that $x_{k,i-1} \in B(X, 1)$ for all $i \in \llbracket 1, N \rrbracket$ and $k \in \llbracket 0, K \rrbracket$ since $\alpha_k \leq 1/(NL)$ by Corollary 5.1.

Substituting (5.13) into (5.12), and applying (5.14) and (5.15) to eliminate terms involving $g'(x_{k+1})$, we obtain that

$$\begin{aligned}
\Phi(x_{k+1}) &\leq \Phi(x_k) - \frac{\alpha_k}{4}d(0, \partial\Phi(x_{k+1}))^2 + \left(\frac{NM}{2} + \frac{N^2M^2}{2}\alpha_k - \frac{1}{2\alpha_k} \right) \|x_{k+1} - x_k\|^2 \\
&\quad + \frac{1}{12}(N-1)^2N(2N-1)L^2M^2\alpha_k^3.
\end{aligned}$$

This yields the desired inequality since $\alpha_k \leq 1/(2NM)$. □

With the approximate descent property, [132] first establish the convergence of $f(x_k)$ and $\|\nabla f(x_k)\|$ to obtain the convergence of x_k provided that $g = 0$ and the f_i 's have globally Lips-

chitz gradients. However, without the global Lipschitz gradient continuity of f_i 's nor boundedness of x_k , we cannot expect to first obtain the convergence of function value and gradient norm. Thus, a different approach inspired by [45] is taken to directly show the convergence of x_k . In [45, Proposition 8], the author proved a length formula for the gradient method, which is a special case of Algorithm 1 where $g = 0$ and $N = 1$. It is inspired by Kurdyka's original length formula [44, Theorem 2].

In order to establish a length formula for Algorithm 1, we next state the uniform Kurdyka–Łojasiewicz inequality [45, Proposition 5] for D_Φ . The motivation behind the uniform Kurdyka–Łojasiewicz inequality as opposed to the Kurdyka–Łojasiewicz inequality is to extend the inequality to all points of a bounded set, without restricting the function value. We in fact propose a strengthened version of the uniform Kurdyka–Łojasiewicz inequality for reasons discussed below. Also, we allow the inequality to hold at D_Φ -critical points, which is new even for the Clarke subdifferential, and simplifies the subsequent analysis. Given $\Phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ and $X \subset \mathbb{R}^n$, a scalar v is called a D_Φ -critical value of Φ in X if there exists $x \in X$ such that $0 \in D_\Phi(x)$ and $v = \Phi(x)$. In particular, if $D_\Phi = \partial\Phi$, we call this v a critical value of Φ in X .

Lemma 5.8. *Let Assumption 5.1 hold. Assume that Φ and D_Φ are definable. Let X be a bounded subset of $\text{dom } \Phi$ with $\overline{X} \subset \text{dom } \Phi$. Define V to be the set of D_Φ -critical values of Φ in \overline{X} if it is nonempty; otherwise $V := \{0\}$. Let $\theta \in (0, 1)$. There exist a concave definable diffeomorphism $\psi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ and $\xi > 0$ such that*

$$\forall x \in X, \quad d(0, D_\Phi(x)) \geq \frac{1}{\psi'(d(\Phi(x), V))}, \quad (5.16a)$$

$$\forall s, t \geq 0, \quad \frac{1}{\psi'(s+t)} \leq \frac{1}{\psi'(s)} + \max \left\{ \frac{1}{\psi'(t)}, \xi t \right\}, \quad (5.16b)$$

$$\forall t \geq 0, \quad \psi'(t) \geq t^{-\theta}. \quad (5.16c)$$

Proof. In the prompt, $\psi'(0)$ is the right derivative of ψ at 0, and we use the convention $1/\infty = 0$. Following the arguments in [28, Theorem 6] (see also [43, Corollary 15]) and the linear extension

construction in [45, Proposition 5], there exists a concave definable diffeomorphism $\psi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ that satisfies (5.16a). We may assume that $\psi'(t) \geq t^{-\theta}$ for all $t \geq 0$, after possibly replacing ψ by $t \mapsto \int_0^t \max\{\psi'(s), t^{-\theta}\} ds$, which is concave definable. It is concave since the integrand is decreasing. To see why it is definable, note that $\{s > 0 : \psi'(s) \geq t^{-\theta}\}$ is definable and hence a finite union of open intervals and points. Thus the integral is equal to ψ up to a constant on finitely many intervals of \mathbb{R}_+ , and equal to $t \mapsto t^{1-\theta}/(1-\theta)$ up to a constant elsewhere. The graph of such a function is hence definable. By the monotonicity theorem [36, (1.2) p. 43] and adapting the linear extension in [45, Proposition 5], we may assume that $1/\psi'$ is continuously differentiable on $(0, T)$ and is constant on (T, ∞) for some $T \in (0, \infty)$, with $(1/\psi)'$ monotone on $(0, T)$ and $\lim_{t \nearrow T} (1/\psi)''(t) \in (0, \infty)$. If $(1/\psi)''(t) \rightarrow \infty$ as $t \rightarrow 0$, then $(1/\psi)'$ is decreasing on $(0, T)$. By concavity of ψ , $(1/\psi)''(t) = -\psi''(t)/(\psi'(t))^2 \geq 0$ for all $t \in (0, T)$. As $(1/\psi)''(t) = 0$ for all $t \in (T, \infty)$, $1/\psi'$ is concave on \mathbb{R}_+ . Therefore, (5.16b) holds by Lemma 4.6. Otherwise if $\lim_{t \searrow 0} (1/\psi)''(t) < \infty$, then there exists $\xi > 0$ such that $|(1/\psi)''(t)| \leq \xi$ for all $t \in (0, T) \cup (T, \infty)$. Therefore, $|1/\psi'(s+t) - 1/\psi'(s)| \leq \xi t$ for any $s, t \in \mathbb{R}_+$, and (5.16b) follows. \square

We say that ψ in Lemma 5.8 is a desingularizing function of Φ over X if it satisfies (5.16a) with $D := \partial\Phi$. The existing analysis of random reshuffling in [132, Theorem 3.6] requires $1/\psi'$ to satisfy a quasi-additivity-type property, namely

$$\forall s, t \in (0, \eta), \quad s + t < \eta \implies \frac{1}{\psi'(s+t)} \leq C_\psi \left(\frac{1}{\psi'(s)} + \frac{1}{\psi'(t)} \right)$$

for some constants $\eta, C_\psi > 0$. This is true for power functions with $C_\psi = 1$ and any $\eta > 0$. It is hence satisfied in polynomially bounded o-minimal structures. However, it is not clear why this should in general o-minimal structures. Thankfully, Lemma 5.8 shows that one can actually get a somewhat weaker property for free, namely (5.16b), which is sufficient for proving the length formula below. This is one of the key technical contributions of this chapter.

Proposition 5.4. *Suppose Assumptions 5.1 and 5.3 hold and Φ is definable. Let $X \subset \text{dom } \Phi$ be bounded, $r \geq 1$, $\theta \in (0, 1)$, and $m \in \mathbb{N}^*$ be an upper bound on the number of critical values of*

Φ in \bar{X} . Let ψ be a desingularizing function of Φ over X such that there exists $\xi \geq 2$ satisfying (5.16b) and $\psi'(t) \geq t^{-\theta}$ for all $t \geq 0$. There exist $c_1 \geq 0$ and $\bar{\alpha}, c_2 > 0$ such that for all $K \in \mathbb{N}$, if $(x_0, \dots, x_{K+1}) \in X \times \dots \times X \times \mathbb{R}^n$ is generated by Algorithm 1 with $\alpha_0, \dots, \alpha_K \leq \bar{\alpha}$ and $\alpha_1/\alpha_0, \dots, \alpha_{K+1}/\alpha_K \leq r$, then

$$\frac{1}{2m} \sum_{k=0}^K \|x_{k+1} - x_k\| \leq 2r\psi \left(\frac{1}{2m} \left[\Phi(x_0) - \Phi(x_K) + c_1 \sum_{k=0}^{K-1} \alpha_k^3 \right] \right) \quad (5.17)$$

$$+ \frac{c_1}{m} \sum_{k=0}^K \alpha_k \max \left\{ \left(\sum_{i=k}^K \alpha_i^3 \right)^\theta, \sum_{i=k}^K \alpha_i^3 \right\} + c_2 \max_{k \in \llbracket 0, K \rrbracket} \alpha_k. \quad (5.18)$$

If $L, M \geq 1$ are respectively Lipschitz constants of f_1, \dots, f_N on $B(X, 2)$ and $\nabla f_1, \dots, \nabla f_N$ on $\text{co } B(X, 1)$, and $L_g := \sup_{x \in B(X, 1) \cap \text{dom } g} d(0, \partial g(x))$, then one may choose $\bar{\alpha} := 1/(2 \max\{NL + L_g, NM\})$,

$$c_1 := \xi(N-1)^2 N(2N-1)L^2 M^2 / 12 \quad \text{and} \quad c_2 := 4(NL + L_g). \quad (5.19)$$

Proof. Let $K \in \mathbb{N}$ and $(x_0, \dots, x_{K+1}) \in X \times \dots \times X \times \mathbb{R}^n$ be generated by Algorithm 1 with $\alpha_0, \dots, \alpha_K \leq \bar{\alpha}$ and $\alpha_1/\alpha_0, \dots, \alpha_{K+1}/\alpha_K \leq r$. Define $y_k := c_1 \sum_{i=k}^K \alpha_i^3 / \xi$ and $z_k := \Phi(x_k) + y_k$ for all $k \in \llbracket 0, K \rrbracket$. By Lemma 5.7, for all $k \in \llbracket 0, K \rrbracket$ we have

$$z_{k+1} - z_k \leq -\frac{\alpha_k}{4} d(0, \partial \Phi(x_{k+1}))^2 - \frac{1}{8\alpha_k} \|x_{k+1} - x_k\|^2. \quad (5.20)$$

Let V be the set of critical values of Φ in \bar{X} if it is nonempty, and otherwise let $V := \{0\}$. Since Φ is definable, V has finitely many elements by the definable Morse-Sard theorem [43, Corollary 9]. Define $\tilde{z}_k := d(z_k, V)$ for all $k \in \llbracket 0, K \rrbracket$.

Assume that $[z_K, z_0)$ excludes the elements of V and the averages of any two consecutive elements of V . Since $z_0 \geq \dots \geq z_K$, either $\tilde{z}_0 \leq \dots \leq \tilde{z}_K$ or $\tilde{z}_0 \geq \dots \geq \tilde{z}_K$. In the first case, for all

$k \in \llbracket 0, K - 1 \rrbracket$, we have

$$\psi(\tilde{z}_{k+1}) - \psi(\tilde{z}_k) \geq \psi'(\tilde{z}_{k+1})(\tilde{z}_{k+1} - \tilde{z}_k) \quad (5.21a)$$

$$= \psi'(\tilde{z}_{k+1})(z_k - z_{k+1}) \quad (5.21b)$$

$$\geq \psi'(d(\Phi(x_{k+1}) + y_{k+1}, V)) \left(\frac{\alpha_k}{4} d(0, \partial\Phi(x_{k+1}))^2 + \frac{1}{8\alpha_k} \|x_{k+1} - x_k\|^2 \right) \quad (5.21c)$$

$$\geq \psi'(d(\Phi(x_{k+1}), V) + y_{k+1}) \left(\frac{\sqrt{\alpha_k}}{2\sqrt{2}} d(0, \partial\Phi(x_{k+1})) + \frac{1}{4\sqrt{\alpha_k}} \|x_{k+1} - x_k\| \right)^2 \quad (5.21d)$$

$$\geq \frac{\left(\frac{\sqrt{\alpha_k}}{2\sqrt{2}} d(0, \partial\Phi(x_{k+1})) + \frac{1}{4\sqrt{\alpha_k}} \|x_{k+1} - x_k\| \right)^2}{\frac{1}{\psi'(d(\Phi(x_{k+1}), V))} + \max \left\{ \frac{1}{\psi'(y_{k+1})}, \xi y_{k+1} \right\}} \quad (5.21e)$$

$$\geq \frac{\left(\frac{\sqrt{\alpha_k}}{2\sqrt{2}} d(0, \partial\Phi(x_{k+1})) + \frac{1}{4\sqrt{\alpha_k}} \|x_{k+1} - x_k\| \right)^2}{d(0, \partial\Phi(x_{k+1})) + \max \left\{ \frac{1}{\psi'(y_{k+1})}, \xi y_{k+1} \right\}}. \quad (5.21f)$$

Indeed, (5.21a) holds because ψ is concave. (5.21b) is due to the existence of $v \in V$ such that $\tilde{z}_k = v - z_k$ for all $k \in \llbracket 0, K \rrbracket$. (5.21c) follows by the descent property (5.20). (5.21d) uses the fact that ψ' is decreasing and

$$\begin{aligned} d(\Phi(x_{k+1}) + y_{k+1}, V) &= \min_{v \in V} |\Phi(x_{k+1}) + y_{k+1} - v| \\ &\leq \min_{v \in V} |\Phi(x_{k+1}) - v| + y_{k+1} = d(\Phi(x_{k+1}), V) + y_{k+1} \end{aligned}$$

for all $k \in \llbracket 0, K - 1 \rrbracket$. It also uses the QM-AM inequality [205]). (5.21e) is an application of (5.16b). Finally, (5.21f) is a consequence of the uniform Kurdyka-Łojasiewicz inequality (5.16a).

Using the AM-GM inequality, (5.21) yields

$$\frac{\sqrt{\alpha_k}}{2\sqrt{2}}d(0, \partial\Phi(x_{k+1})) + \frac{1}{4\sqrt{\alpha_k}}\|x_{k+1} - x_k\| \quad (5.22a)$$

$$\leq \sqrt{(\psi(\tilde{z}_{k+1}) - \psi(\tilde{z}_k)) \left(d(0, \partial\Phi(x_{k+1})) + \max \left\{ \frac{1}{\psi'(y_{k+1})}, \xi y_{k+1} \right\} \right)} \quad (5.22b)$$

$$\leq \frac{1}{2(1-\iota)\sqrt{\alpha_k}}(\psi(\tilde{z}_{k+1}) - \psi(\tilde{z}_k)) + \frac{(1-\iota)\sqrt{\alpha_k}}{2} \left(d(0, \partial\Phi(x_{k+1})) + \max \left\{ \frac{1}{\psi'(y_{k+1})}, \xi y_{k+1} \right\} \right) \quad (5.22c)$$

where $\iota \in [0, 1)$ can be arbitrary. By letting $\iota = 0$ and multiplying by $4\sqrt{\alpha_k}$ on both sides, we obtain

$$\|x_{k+1} - x_k\| \leq 2(\psi(\tilde{z}_{k+1}) - \psi(\tilde{z}_k)) + 2 \max \left\{ \frac{\alpha_k}{\psi'(y_{k+1})}, \xi \alpha_k y_{k+1} \right\}.$$

Telescoping yields

$$\sum_{k=0}^K \|x_{k+1} - x_k\| \leq 2(\psi(\tilde{z}_K) - \psi(\tilde{z}_0)) + 2 \sum_{k=0}^K \max \left\{ \frac{\alpha_k}{\psi'(y_{k+1})}, \xi \alpha_k y_{k+1} \right\} + \|x_{K+1} - x_K\| \quad (5.23a)$$

$$\leq 2\psi(\tilde{z}_K - \tilde{z}_0) + 2 \sum_{k=0}^K \max \left\{ \frac{\alpha_k}{\psi'(y_k)}, \xi \alpha_k y_k \right\} + 2(NL + L_g)\alpha_K \quad (5.23b)$$

$$\leq 2\psi(z_0 - z_K) + 2 \sum_{k=0}^K \max \left\{ \frac{\alpha_k}{\psi'(y_k)}, \xi \alpha_k y_k \right\} + 2(NL + L_g) \max_{k \in \llbracket 0, K \rrbracket} \alpha_k \quad (5.23c)$$

where (5.23b) uses Lemma 4.6, $y_k \geq y_{k+1}$, and Corollary 5.1.

In the second case, i.e., $\tilde{z}_0 > \dots > \tilde{z}_K$, (5.21) becomes

$$\psi(\tilde{z}_k) - \psi(\tilde{z}_{k+1}) \geq \frac{\left(\frac{\sqrt{\alpha_k}}{2\sqrt{2}}d(0, \partial\Phi(x_{k+1})) + \frac{1}{4\sqrt{\alpha_k}}\|x_{k+1} - x_k\| \right)^2}{d(0, \partial\Phi(x_k)) + \max \left\{ \frac{1}{\psi'(y_k)}, \xi y_k \right\}}$$

and (5.22) becomes

$$\begin{aligned} & \frac{\sqrt{\alpha_k}}{2\sqrt{2}}d(0, \partial\Phi(x_{k+1})) + \frac{1}{4\sqrt{\alpha_k}}\|x_{k+1} - x_k\| \\ & \leq \frac{1}{2(1-\iota)\sqrt{\alpha_k}}(\psi(\tilde{z}_k) - \psi(\tilde{z}_{k+1})) + \frac{(1-\iota)\sqrt{\alpha_k}}{2} \left(d(0, \partial\Phi(x_k)) + \max \left\{ \frac{1}{\psi'(y_k)}, \xi y_k \right\} \right). \end{aligned}$$

Setting $\iota = 1 - r^{-1}/\sqrt{2}$ and multiplying $4\sqrt{\alpha_k}$ on both sides yields

$$\begin{aligned} & \sqrt{2}r^{-1}\alpha_{k+1}d(0, \partial\Phi(x_{k+1})) + \|x_{k+1} - x_k\| \\ & \leq 2r(\psi(\tilde{z}_k) - \psi(\tilde{z}_{k+1})) + \sqrt{2}r^{-1}\alpha_k \left(d(0, \partial\Phi(x_k)) + \max \left\{ \frac{1}{\psi'(y_k)}, \xi y_k \right\} \right) \end{aligned}$$

where we use the fact that $\alpha_{k+1}/\alpha_k \leq r$. This can be simplified to

$$\begin{aligned} \|x_{k+1} - x_k\| & \leq 2r[\psi(\tilde{z}_k) - \psi(\tilde{z}_{k+1})] + \sqrt{2}r^{-1}[\alpha_k d(0, \partial\Phi(x_k)) - \alpha_{k+1}d(0, \partial\Phi(x_{k+1}))] \\ & \quad + \sqrt{2} \max \left\{ \frac{\alpha_k}{\psi'(y_k)}, \xi \alpha_k y_k \right\}. \end{aligned}$$

Telescoping and $r \geq 1$ yield

$$\begin{aligned}
\sum_{k=0}^K \|x_{k+1} - x_k\| &\leq 2r(\psi(\tilde{z}_0) - \psi(\tilde{z}_K)) + \sqrt{2}r^{-1}[\alpha_0 d(0, \partial\Phi(x_0)) - \alpha_K d(0, \partial\Phi(x_K))] \\
&\quad + \sqrt{2} \sum_{k=0}^K \max \left\{ \frac{\alpha_k}{\psi'(y_k)}, \xi \alpha_k y_k \right\} + \|x_{K+1} - x_K\| \\
&\leq 2r(\psi(\tilde{z}_0) - \psi(\tilde{z}_K)) + \sqrt{2}r^{-1} \alpha_0 d(0, \partial\Phi(x_0)) + \sqrt{2} \sum_{k=0}^K \max \left\{ \frac{\alpha_k}{\psi'(y_k)}, \xi \alpha_k y_k \right\} \\
&\quad + \|x_{K+1} - x_K\| \\
&\leq 2r\psi(\tilde{z}_0 - \tilde{z}_K) + \sqrt{2}\alpha_0(\|\nabla f(x_0)\| + d(0, \partial g(x_0))) + \sqrt{2} \sum_{k=0}^K \max \left\{ \frac{\alpha_k}{\psi'(y_k)}, \xi \alpha_k y_k \right\} \\
&\quad + 2(NL + L_g)\alpha_K \\
&\leq 2r\psi(z_0 - z_K) + \sqrt{2}(NL + L_g)\alpha_0 + \sqrt{2} \sum_{k=0}^K \max \left\{ \frac{\alpha_k}{\psi'(y_k)}, \xi \alpha_k y_k \right\} + 2(NL + L_g)\alpha_K \\
&\leq 2r\psi(z_0 - z_K) + 2 \sum_{k=0}^K \max \left\{ \frac{\alpha_k}{\psi'(y_k)}, \xi \alpha_k y_k \right\} + 4(NL + L_g) \max_{k \in \llbracket 0, K \rrbracket} \alpha_k.
\end{aligned}$$

Compared with the upper bound obtained in the first case (5.23c), the above bound is larger. We can hence use it as a common bound for both cases.

We next consider the case where $z_0 > \dots > z_K$ and there exist $0 \leq K_1 \leq \dots \leq K_p \leq K$ such that

$$[z_K, z_{K_p+1}) \cup \dots \cup [z_{K_2}, z_{K_1+1}) \cup [z_{K_1}, z_0)$$

excludes the elements of V and the averages of any two consecutive elements of V . For notational

convenience, let $K_0 := -1$ and $K_{p+1} := K$. We have

$$\begin{aligned}
\sum_{k=0}^K \|x_{k+1} - x_k\| &= \sum_{i=0}^p \sum_{k=K_{i+1}}^{K_{i+1}} \|x_{k+1} - x_k\| \\
&\leq \sum_{i=0}^p \left(2r\psi(z_{K_{i+1}} - z_{K_{i+1}}) + 2 \sum_{k=K_{i+1}}^{K_{i+1}} \max \left\{ \frac{\alpha_k}{\psi'(y_k)}, \xi \alpha_k y_k \right\} \right. \\
&\quad \left. + 4(NL + L_g) \max_{k \in \llbracket K_{i+1}, K_{i+1} \rrbracket} \alpha_k \right) \\
&\leq 2r(p+1)\psi \left(\frac{1}{p+1} \sum_{i=0}^p z_{K_{i+1}} - z_{K_{i+1}} \right) + 2 \sum_{k=0}^K \max \left\{ \frac{\alpha_k}{\psi'(y_k)}, \xi \alpha_k y_k \right\} \\
&\quad + 4(p+1)(NL + L_g) \max_{k \in \llbracket 0, K \rrbracket} \alpha_k \\
&\leq 2r(p+1)\psi \left(\frac{z_0 - z_K}{p+1} \right) + 2 \sum_{k=0}^K \max \left\{ \frac{\alpha_k}{\psi'(y_k)}, \xi \alpha_k y_k \right\} + 4(p+1)(NL + L_g) \max_{k \in \llbracket 0, K \rrbracket} \alpha_k \\
&\leq 4mr \psi \left(\frac{z_0 - z_K}{2m} \right) + 2 \sum_{k=0}^K \max \left\{ \frac{\alpha_k}{\psi'(y_k)}, \xi \alpha_k y_k \right\} + 8m(NL + L_g) \max_{k \in \llbracket 0, K \rrbracket} \alpha_k,
\end{aligned}$$

where the second inequality uses concavity of ψ and the last inequality uses $p \leq 2m - 1$ and the fact that $s \mapsto s\psi(a/s)$ is increasing over $(0, \infty)$ for any constant $a > 0$ [206, Proposition 2.1].

We now consider the general case where $z_0 \geq \dots \geq z_K$. Observe that if $z_{k+1} = z_k$ at some iteration $k \in \llbracket 0, K \rrbracket$, then $x_{k+1} = x_k$ and $0 \in \partial\Phi(x_{k+1})$ by (5.20). Hence such iterations do not contribute to the length, and happen only at critical points of Φ . We can thus remove them and obtain the above length formula with the remaining indices, to which we can add back the discarded indices. Indeed, the added terms on the left hand side would be equal to zero, and the added terms on the right hand size would only make the bound greater.

The desired formula now follows from simple observations. First,

$$z_0 - z_K = \Phi(x_0) + y_0 - \Phi(x_K) - y_K = \Phi(x_0) - \Phi(x_K) + \frac{c_1}{\xi} \sum_{k=0}^{K-1} \alpha_k^3.$$

Second, since $\psi'(t) \geq t^{-\theta}$ for all $t \geq 0$ and $y_k \leq \xi y_k = c_1 \sum_{i=k}^K \alpha_i^3$, we have

$$\begin{aligned} \max \left\{ \frac{\alpha_k}{\psi'(y_k)}, \xi \alpha_k y_k \right\} &= \alpha_k \max \left\{ \frac{1}{\psi'(y_k)}, \xi y_k \right\} \\ &\leq \alpha_k \max \{ y_k^\theta, \xi y_k \} \\ &\leq \alpha_k \max \left\{ c_1^\theta \left(\sum_{i=k}^K \alpha_i^3 \right)^\theta, c_1 \sum_{i=k}^K \alpha_i^3 \right\} \\ &\leq c_1 \alpha_k \max \left\{ \left(\sum_{i=k}^K \alpha_i^3 \right)^\theta, \sum_{i=k}^K \alpha_i^3 \right\}. \end{aligned}$$

The last inequality holds because either $c_1 = 0$ or $c_1 \geq 1$. Indeed, recall that $c_1 := \xi(N - 1)^2 N(2N - 1)L^2 M^2 / 12$. If $N = 1$, then $c_1 = 0$ and both sides of the last inequality are zero, so it holds trivially; otherwise $N \geq 2$ and $c_1 \geq \xi \cdot 1^2 \cdot 2 \cdot 3L^2 M^2 / 12 = \xi L^2 M^2 / 2$. Recall that we assume $\xi \geq 2$ and $M, L \geq 1$ in the statement of this Proposition, so $c_1 \geq 2 \cdot 1^2 \cdot 1^2 / 2 = 1$ and $c_1^\theta \leq c_1$ for all $\theta \in (0, 1)$ yields the last inequality. \square

The above length formula finds a particularly simple form if we borrow the step sizes used in [132, Lemma 3.7]. It then agrees with the one we derived for the momentum method Lemma 4.1. It is hence suitable for obtaining global convergence.

Corollary 5.2. *Suppose Assumptions 5.1 and 5.3 hold and that Φ is definable. Let $\beta = 0$ if $N = 1$, else $\beta \in (1/2, 1)$. Let $X \subset \text{dom } \Phi$ be a bounded set. There exist $\bar{\alpha} > 0$, $\eta, \kappa \geq 0$, and a desingularizing function ψ of Φ over X such that for all $K \in \mathbb{N}$, $\alpha \in (0, \bar{\alpha}]$, and $\gamma \in \mathbb{N}^*$, if $(x_0, \dots, x_{K+1}) \in X \times \dots \times X \times \mathbb{R}^n$ is generated by Algorithm 1 with $\alpha_k = \alpha / (k + \gamma)^\beta$, then*

$$\sum_{k=0}^K \|x_{k+1} - x_k\| \leq \psi(\Phi(x_0) - \Phi(x_K) + \eta\alpha) + \kappa\alpha.$$

Proof. Since $\alpha_{k+1}/\alpha_k \leq \max\{1, [(k + \gamma)/(k + \gamma + 1)]^\beta\} \leq 1$ for all $k \in \llbracket 0, K \rrbracket$, by Proposition 5.4 the formula (5.17) holds with $r := 1$ and whatever $\theta \in (0, 1)$ we desire, with some corresponding

$\xi \geq 2$ and $m \in \mathbb{N}^*$. Replacing ψ with $4m\psi(\cdot/(2m))$, it holds that

$$\begin{aligned} \sum_{k=0}^K \|x_{k+1} - x_k\| \leq & \psi \left(\Phi(x_0) - \Phi(x_K) + c_1 \sum_{k=0}^{K-1} \alpha_k^3 \right) + 2c_1 \sum_{k=0}^K \alpha_k \max \left\{ \left(\sum_{i=k}^K \alpha_i^3 \right)^\theta, \sum_{i=k}^K \alpha_i^3 \right\} \\ & + 2mc_2 \max_{k \in \llbracket 0, K \rrbracket} \alpha_k. \end{aligned}$$

If $N = 1$, then it suffices to take $\eta := c_1 = 0$ and $\kappa := 2mc_2$. If $N > 1$, then let $\theta \in (\max\{(1 - \beta)/(3\beta - 1), 1/2\}, 1)$. For any $i \in \mathbb{N}$, using classical series integral comparison arguments (see, e.g. [132, Appendix B]), we have

$$\sum_{k=i}^{\infty} \alpha_k^3 = \sum_{k=i}^{\infty} \frac{\alpha^3}{(k + \gamma)^{3\beta}} \leq \alpha_i^3 + \int_{i+\gamma}^{\infty} \frac{\alpha^3}{v^{3\beta}} dv \leq \alpha_i^3 + \alpha^3 \frac{(i + \gamma)^{1-3\beta}}{3\beta - 1}. \quad (5.24)$$

Without loss of generality, we may assume that $\bar{\alpha} \leq 1/2$. Since $\beta \in (1/2, 1)$, we obtain an upper bound on the argument of Φ in (5.17):

$$\sum_{k=0}^K \alpha_k^3 \leq \sum_{k=0}^{\infty} \alpha_k^3 \leq \frac{\alpha^3}{\gamma^{3\beta}} + \alpha^3 \frac{\gamma^{1-3\beta}}{3\beta - 1} \leq \alpha^3 + \alpha^3 \frac{1}{3\beta - 1} \leq 3\alpha^3 \leq \alpha,$$

where the second inequality is obtained by setting $i = 0$ in (5.24) and $k = 0$ in $\alpha_k = \alpha/(k + \gamma)^\beta$. We next upper bound the second term on the right hand side of (5.17). Note that $\sum_{i=k}^K \alpha_i^3 < (\sum_{i=k}^K \alpha_i^3)^\theta$ as $\sum_{i=k}^K \alpha_i^3, \theta \in (0, 1)$. Therefore,

$$\begin{aligned} \sum_{k=0}^K \alpha_k \max \left\{ \left(\sum_{i=k}^K \alpha_i^3 \right)^\theta, \sum_{i=k}^K \alpha_i^3 \right\} &= \sum_{k=0}^K \alpha_k \left(\sum_{i=k}^K \alpha_i^3 \right)^\theta \\ &\leq \sum_{k=0}^K \frac{\alpha}{(k + \gamma)^\beta} \left(\alpha_k^3 + \alpha^3 \frac{(k + \gamma)^{1-3\beta}}{3\beta - 1} \right)^\theta \\ &\leq \sum_{k=0}^K \frac{\alpha}{(k + \gamma)^\beta} \left(\alpha_k^{3\theta} + \alpha^{3\theta} \frac{(k + \gamma)^{(1-3\beta)\theta}}{(3\beta - 1)^\theta} \right) \\ &\leq \sum_{k=0}^K \frac{\alpha^{3\theta+1}}{(k + \gamma)^{\beta(1+3\theta)}} + \frac{\alpha^{3\theta+1}}{(3\beta - 1)^\theta} \sum_{k=0}^K (k + \gamma)^{(1-3\beta)\theta - \beta}. \end{aligned}$$

Similar to the upper bound of $\sum_{i=0}^K \alpha_i^3$ in (5.24), we have

$$\sum_{k=0}^K \frac{\alpha^{3\theta+1}}{(k+\gamma)^{\beta(1+3\theta)}} \leq \alpha^{3\theta+1} + \alpha^{3\theta+1} \frac{1}{(3\theta+1)\beta-1}$$

due to $\theta \in (1/2, 1)$, and

$$\begin{aligned} \frac{\alpha^{3\theta+1}}{(3\beta-1)^\theta} \sum_{k=0}^K (k+\gamma)^{(1-3\beta)\theta-\beta} &\leq 2^\theta \alpha^{3\theta+1} \left(\frac{1}{\gamma^{\beta+(3\beta-1)\theta}} + \frac{\gamma^{1-\beta-(3\beta-1)\theta}}{\beta+(3\beta-1)\theta-1} \right) \\ &\leq 2\alpha^{3\theta+1} \left(1 + \frac{1}{\beta+(3\beta-1)\theta-1} \right). \end{aligned}$$

Since $\alpha \in (0, 1/2]$, we have $\alpha^{3\theta+1} \leq \alpha$ and

$$\begin{aligned} \sum_{k=0}^K \alpha_k \max \left\{ \left(\sum_{i=k}^K \alpha_i^3 \right)^\theta, \sum_{i=k}^K \alpha_i^3 \right\} &\leq \alpha^{3\theta+1} \left(3 + \frac{1}{(3\theta+1)\beta-1} + \frac{2}{\beta+(3\beta-1)\theta-1} \right) \\ &\leq \alpha \left(3 + \frac{1}{(3\theta+1)\beta-1} + \frac{2}{\beta+(3\beta-1)\theta-1} \right) \\ &\leq 3\alpha \left(1 + \frac{1}{\beta+(3\beta-1)\theta-1} \right). \end{aligned}$$

The desired inequality then follows by posing

$$\eta := c_1 \quad \text{and} \quad \kappa := 6c_1 \left(1 + \frac{1}{\beta+(3\beta-1)\theta-1} \right) + 2mc_2.$$

□

We are now ready to prove the main result of this subsection. Leveraging the tracking lemma (Lemma 5.1) and the length formula (Corollary 5.2), we employ a framework similar to that used in Theorem 4.2 to establish the desired result, Theorem 5.4. To provide a clear overview, we outline the proof below, with the full details available in Section 5.4.2.

Sketch proof of Theorem 5.4. Let X_0 be a bounded subset of $\text{dom } \Phi$. We can first use the tracking lemma (Lemma 5.1) and the length formula (Corollary 5.2) to prove the existence of $\bar{\alpha} > 0$ such

that $\Gamma(X_0, \bar{\alpha}) < \infty$ where

$$\Gamma(X_0, \bar{\alpha}) := \sup_{\substack{x \in (\mathbb{R}^n)^{N \times \llbracket 0, N \rrbracket}, \sigma \in \mathfrak{S}_N \\ (\alpha, \gamma) \in (0, \bar{\alpha}] \times \mathbb{N}^*}} \sum_{k=0}^{\infty} \|x_{k+1,0} - x_{k,0}\| \quad (5.25a)$$

$$\text{s.t.} \quad \begin{cases} x_{k+1,0} = \text{prox}_{\alpha_k g}(x_{k,N}), \\ x_{k,i} = x_{k,i-1} - \alpha_k \nabla f_{\sigma_i^k}(x_{k,i-1}), \quad \forall i \in \llbracket 1, N \rrbracket, \\ \alpha_k = \alpha / (k + \gamma)^\beta, \quad \forall k \in \mathbb{N}, \quad x_{0,0} \in X_0. \end{cases} \quad (5.25b)$$

Above, \mathfrak{S}_N denotes the symmetric group of degree N .

The finiteness of $\Gamma(X_0, \bar{\alpha})$ naturally means that $\|x_k - x_0\| \leq \sum_{i=0}^k \|x_i - x_{i-1}\| \leq \Gamma(X_0, \bar{\alpha})$ for all $k \in \mathbb{N}$, that is, $x_k \in B(X_0, \Gamma(X_0, \bar{\alpha}))$ (where $x_k := x_{k,0}$). Let L_0, M respectively be Lipschitz constants of f_1, \dots, f_N and $\nabla f_1, \dots, \nabla f_N$ on $B(X_0, \Gamma(X_0, \bar{\alpha}) + 2)$. We then establish

$$d(0, \partial\Phi(x_{k+1})) \leq \left(\frac{1}{\alpha_k} + M \right) \|x_{k+1} - x_k\| + (N-1) \sqrt{\frac{N(2N-1)}{6}} ML_0 \alpha_k.$$

Note that for all nonnegative decreasing sequence u_0, u_1, u_2, \dots , we have $u_k \leq 2(\sum_{i=\lfloor k/2 \rfloor}^{\infty} u_i) / (k+2)$ as explained in [45, Footnote 1]. In particular, for $u_k := \min_{i \in \llbracket 0, k \rrbracket} \alpha_i d(0, \partial\Phi(x_{i+1}))$, we have

$$\min_{i \in \llbracket 0, k \rrbracket} \alpha_i d(0, \partial\Phi(x_{i+1})) \leq \frac{2}{k+2} \sum_{i=\lfloor k/2 \rfloor}^{\infty} (1 + \alpha M) \|x_{i+1} - x_i\| + (N-1) \sqrt{\frac{N(2N-1)}{6}} ML_0 \alpha_i^2.$$

Thus, using $\alpha_k = \alpha / (k+1)^\beta$ and series integral comparison arguments again, we can finally obtain

$$\min_{i \in \llbracket 0, k \rrbracket} d(0, \partial\Phi(x_{i+1})) \leq \frac{1}{(k+1)^{1-\beta}} \left(2(\alpha^{-1} + M) \sum_{i=\lfloor k/2 \rfloor}^{\infty} \|x_{i+1} - x_i\| + \frac{4(N-1)\sqrt{N(2N-1)}ML_0\alpha}{\sqrt{6}(2\beta-1)(\lfloor k/2 \rfloor + 1)^{2\beta-1}} \right).$$

□

It is worth noting that our proof of Theorem 5.4 is based on a worst-case analysis, which holds for any permutation of indices used in the stochastic gradient updates. As a consequence, the resulting complexity bound may not be tight with respect to the sample size N , and the step

size upper bound $\bar{\alpha}$ requires a conservative choice to ensure the iterates track the subgradient trajectories accurately. In particular, while Lemma 5.7 sets $\bar{\alpha} := \min\{1/(2NL + 2L_g), 1/(2NM)\}$, which yields the typical $\Theta(1/N)$ scaling seen in nonconvex settings (see, e.g., [181, Eq. (2.16)]), the final value of $\bar{\alpha}$ used in the proof of Theorem 5.4 may be significantly smaller. This tighter restriction is necessary to guarantee the recursive tracking of subgradient trajectories, particularly in the absence of additional assumptions.

One natural question is whether a high-probability analysis could improve the dependence on N or allow for a larger step size. Indeed, such improvements have been demonstrated in the literature. For example, applying the approach in [181] to PRR under their global smoothness assumptions, one can use a constant step size and improve the last-iterate complexity from $O(N^{3/2}\epsilon^{-3})$ (deterministic) to $O(\max\{N\epsilon^{-2}, N^{1/2}\epsilon^{-3}\})$ (high probability). Correspondingly, the step size improves from $O(1/N^{3/2})$ to $O(1/N)$.

However, in our setting, we do not assume a global Lipschitz constant for ∇f_i or a uniform bound on the iterates. To ensure boundedness and thereby control the local Lipschitz constants, we need to adopt step sizes that diminish faster than $1/\sqrt{k}$. Under such conditions, high-probability analysis does not lead to an explicit improvement in either the allowable step size or the last-iterate complexity. Indeed, we explored both worst-case and high-probability analyses and found that, when using step sizes close to $1/\sqrt{k}$, the last-iterate complexity remains approximately $O(\epsilon^{-6})$. This complexity is quite poor, reflecting significantly slower convergence of the last iterate compared to the convergence rate in terms of minimum gradient norm. Moreover, the dependence on N remains unclear. For these reasons, we chose not to include any details in this work.

5.4 Proof of intermediate results

5.4.1 Proof of Proposition 5.2

For simplicity, we can consider setting of $U \subset B(0, 1)$. By assumption, there exist $r_x, r_y > 0$ such that for every $k \in \mathbb{N}$, $x_k(t) \in B(0, r_x)$ and $y_k(t) \in B(0, r_y)$ for all $t \in I \setminus Z_k$, where Z_k has zero measure. Let $Z := \bigcup_{k \in \mathbb{N}} Z_k$, then Z also has zero measure. Thus, $x_k(t) \in B(0, r_x)$

and $y_k(t) \in B(0, r_y)$ for all $k \in \mathbb{N}$ and $t \in I \setminus Z$. In addition, there exists $r_F > 0$ such that $F(x_k(t)) \subset F(B(0, r_x)) \subset B(0, r_F)$ for all $k \in \mathbb{N}$ and $t \in I \setminus Z$ as F is locally bounded.

Furthermore, there is a set Z' with zero measure such that for all $t \in I \setminus Z'$, for any neighborhood U of 0 such that $U \subset B(0, 1)$, there exists $k_0 \in \mathbb{N}$ satisfying

$$(x_k(t), y_k(t)) \in \text{graph}(F + G) + U, \quad \forall k \geq k_0.$$

Since $y_k(t) \in F(x_k(t)) + G(x_k(t)) + B(0, 1)$, there exist $\alpha_k(t) \in F(x_k(t))$, $\beta_k(t) \in G(x_k(t))$ and $\gamma_k(t) \in B(0, 1)$ such that $y_k(t) = \alpha_k(t) + \beta_k(t) + \gamma_k(t)$ for all $k \geq k_0$ and $t \in I \setminus (Z \cup Z')$. Notice that $\|\beta_k(t)\| \leq r_y + r_F + 1$. Set $r := r_y + r_F + 1$, and define the cut set-valued mapping G_r as $G_r(\cdot) := G(\cdot) \cap B(0, r)$. Note that $F + G_r$ is proper, locally bounded, and has convex values. In addition, $F + G_r$ has closed graph by Lemma 5.2. Using similar arguments as in [40, 2.1.5(d) Proposition], it follows that $F + G_r$ is upper semicontinuous [46, p. 59], and thus upper hemicontinuous [46, p. 60]. Since $\beta_k(t) \in G_r(x_k(t))$, for all $t \in I \setminus (Z \cup Z')$, we have

$$(x_k(t), y_k(t)) \in \text{graph}(F + G_r) + U, \quad \forall k \geq k_0.$$

It is easy to see the above also holds for neighborhood not necessarily a subset of $B(0, 1)$ by taking the same k_0 as the one for neighborhood $U \cap \mathring{B}(0, 1)$. Therefore, applying [46, Theorem 1, p. 60] to $F + G_r$, we can conclude that for a.e. $t \in I$,

$$(x(t), y(t)) \in \text{graph}(F + G_r) \subset \text{graph}(F + G).$$

□

5.4.2 Proof of Theorem 5.4

Let X_0 be a bounded subset of $\text{dom } \Phi$. Our goal is to show that there exist $\bar{\alpha} > 0$ such that $\Gamma(X_0, \bar{\alpha}) < \infty$, where

$$\Gamma(X_0, \bar{\alpha}) := \sup_{\substack{x \in (\mathbb{R}^n)^{\mathbb{N}} \times \llbracket 0, N \rrbracket, \sigma \in \mathfrak{S}_N \\ (\alpha, \gamma) \in (0, \bar{\alpha}] \times \mathbb{N}^*}} \sum_{k=0}^{\infty} \|x_{k+1,0} - x_{k,0}\| \quad (5.26a)$$

$$\text{s.t. } \begin{cases} x_{k+1,0} = \text{prox}_{\alpha_k g}(x_{k,N}), \\ x_{k,i} = x_{k,i-1} - \alpha_k \nabla f_{\sigma_i^k}(x_{k,i-1}), \quad \forall i \in \llbracket 1, N \rrbracket, \\ \alpha_k = \alpha / (k + \gamma)^\beta, \quad \forall k \in \mathbb{N}, \quad x_{0,0} \in X_0, \end{cases} \quad (5.26b)$$

and \mathfrak{S}_N denotes the symmetric group of degree N .

Let $\Sigma : \mathbb{R}_+ \times \text{dom } g \rightarrow \text{dom } g$ defined by $\Sigma(t, x_0) := x(t)$ where $x(\cdot)$ is the unique solution of

$$x'(t) \in -\partial\Phi(x(t)) = -\nabla f(x(t)) - \partial g(x(t)) = -\sum_{i=1}^N \nabla f_i(x(t)) - \partial g(x(t)), \quad x(0) = x_0.$$

Notice that the uniqueness can be easily derived from [74, Theorem 4.3(b)]. Let C be the set of critical points of Φ in $\overline{\Sigma(\mathbb{R}_+, X_0)}$. Note that C is bounded due to Proposition 5.3 and closed due to [46, Proposition 2(a), p. 141]. Thus, C is compact, and either $X_0 \subset C$ (in this case just take any $\epsilon \in (0, 1]$) or there exists $\epsilon \in (0, 1]$ such that $X_0 \setminus \mathring{B}(C, \epsilon/6) \neq \emptyset$.

To prove $\Gamma(X_0, \bar{\alpha}) < \infty$ for some $\bar{\alpha} > 0$, we follow the two steps below. First, we will show that the iterates can go arbitrarily close to a critical point of Φ . Second, we will show that either the iterates always stay within a neighborhood of the critical point, or they admit a sufficient decrease in function value when step out of the neighborhood of the critical point.

For the first step, we show that for every X_0 and every $\delta \in (0, \epsilon/2)$, there exists $\hat{\alpha}$ such that for any $(x_k)_{k \in \mathbb{N}}$ generated by Algorithm 1 with $x_0 \in X_0$ and $(\alpha_k)_{k \in \mathbb{N}}$ satisfying $\alpha_k \in (0, \hat{\alpha}]$ for all $k \in \mathbb{N}$, there exist $k^* \in \mathbb{N}$ and $x^* \in C$ such that $\|x_{k^*} - x^*\| \leq \delta$. If $X_0 \subset C$, then the desired result trivially holds because we can simply take any $\hat{\alpha} > 0$ with $k^* = 0$ and $x^* = x_0$. If

$X_0 \setminus \mathring{B}(C, \epsilon/6) \neq \emptyset$, then the argument can be further divided into two parts. The first part is to show that the subgradient trajectory of Φ can go arbitrarily close to a critical point of Φ , i.e., we show that for fixed X_0 and δ , there exists $T > 0$ such that for any solution $x(\cdot)$ of $x'(t) \in -\partial\Phi(x(t))$ with $x_0 \in X_0$, we have that $\|x(t^*) - x^*\| \leq \delta/3$ for some $t^* \in [0, T]$ and $x^* \in C$. To prove it, let

$$\nu := \inf_{x \in \bar{\Sigma}_0 \setminus \mathring{B}(C, \delta/3)} \|\partial\Phi(x)\| = \inf \left\{ \|y\| : y \in \partial\Phi \left(\bar{\Sigma}_0 \setminus \mathring{B}(C, \delta/3) \right) \right\}.$$

Recall that $\partial\Phi := \nabla f(x) + \partial g(x)$ has closed graph by Lemma 5.2. Since $\bar{\Sigma}_0 \setminus \mathring{B}(C, \delta/3)$ is a nonempty compact subset of $\text{dom } g$, the set $\partial\Phi(\bar{\Sigma}_0 \setminus \mathring{B}(C, \delta/3))$ is nonempty and closed. Thus, ν is finite and attained. Combined with the definition of C , it holds that $\nu > 0$. Let $T := 2\Gamma(X_0)/\nu$ where $\Gamma(X_0)$ is defined by

$$\Gamma(X_0) := \sup_{x \in \mathcal{A}(\mathbb{R}_+, \mathbb{R}^n)} \int_0^\infty \|x'(t)\| dt$$

subject to $\begin{cases} x'(t) \in -\partial\Phi(x(t)), \text{ for a.e. } t > 0, \\ x(0) \in X_0. \end{cases}$

and $\Gamma(X_0) < \infty$ by Proposition 5.3. Note that $T > 0$ because $X_0 \not\subset C$ and $\Gamma(X_0) > 0$. Furthermore, it follows that there exists $t^* \in [0, T]$ such that $d(0, \partial\Phi(x(t^*))) = \|x'(t^*)\| < 2\Gamma(X_0)/T$. We know from Lemma 5.5 that $d(0, \partial\Phi(x(t))) = \|x'(t)\|$ for almost every $t \in [0, T]$. Assume the contrary that $\|x'(t)\| \geq 2\Gamma(X_0)/T$ for all such t , then

$$\int_0^\infty \|x'(t)\| dt \geq \int_0^T \|x'(t)\| dt \geq 2\Gamma(X_0) > \Gamma(X_0)$$

contradicts the definition of $\Gamma(X_0)$. According to the definition of ν , we have $x(t^*) \in \mathring{B}(C, \delta/3)$. Hence, $\|x(t^*) - x^*\| \leq \delta/3$ for this $x^* \in C$ and the claim follows.

We next show that there exists $\hat{\alpha} > 0$ such that for any $(x_k)_{k \in \mathbb{N}}$ generated by Algorithm 1 with $x_0 \in X_0$ and $(\alpha_k)_{k \in \mathbb{N}}$ satisfying $\alpha_k \in (0, \hat{\alpha}]$ for all $k \in \mathbb{N}$, there exists $k^* \in \mathbb{N}$ such that

$\|x_{k^*} - x(t^*)\| \leq 2\delta/3$. To see this, let

$$L_\Phi := \sup_{x \in B(\bar{\Sigma}_0, \epsilon)} \|\nabla f(x)\| + \sup_{x \in B(\bar{\Sigma}_0, \epsilon) \cap \text{dom } g} d(0, \partial g(x)),$$

where $L_\Phi < \infty$ by Lemma 5.4. By Lemma 5.1, there exists $\hat{\alpha} \in (0, \delta/(3L_\Phi)]$ such that for any $(x_k)_{k \in \mathbb{N}}$ generated by Algorithm 1 with $x_0 \in X_0$ and $(\alpha_k)_{k \in \mathbb{N}}$ satisfying $\alpha_k \in (0, \hat{\alpha}]$ for all $k \in \mathbb{N}$, we have that for all $k \in \mathbb{N}^*$,

$$\sum_{i=0}^{k-1} \alpha_i \leq T \implies \left\| x_k - x \left(\sum_{i=0}^{k-1} \alpha_i \right) \right\| \leq \frac{\delta}{3},$$

Since $\alpha_k \leq \hat{\alpha} \leq \delta/(3L_\Phi)$ and $\sum_{k=0}^{\infty} \alpha_k = \infty$, there exists $k^* \in \mathbb{N}$ such that $t_{k^*} := \sum_{k=0}^{k^*-1} \alpha_k \in [t^* - \delta/(3L_\Phi), t^*]$, where the convention $\sum_{k=0}^{-1} \alpha_k = 0$ is used. Thus,

$$\begin{aligned} \|x_{k^*} - x(t^*)\| &\leq \|x_{k^*} - x(t_{k^*})\| + \|x(t_{k^*}) - x(t^*)\| \leq \frac{\delta}{3} + \int_{t_{k^*}}^{t^*} \|x'(\tau)\| d\tau \\ &= \frac{\delta}{3} + \int_{t_{k^*}}^{t^*} d(0, \partial \Phi(x(\tau))) d\tau \leq \frac{\delta}{3} + L_\Phi(t^* - t_{k^*}) \leq \frac{2\delta}{3}. \end{aligned}$$

Now claim of the first step of the proof follows easily from combining the above two parts, i.e.,

$$\|x_{k^*} - x^*\| \leq \|x_{k^*} - x(t^*)\| + \|x(t^*) - x^*\| \leq \frac{2\delta}{3} + \frac{\delta}{3} = \delta.$$

For the second step, define $K^* := \inf\{k \geq k^* : x_k \notin \mathring{B}(C, \epsilon)\}$. Clearly, we have that $K^* \geq k^* + 1$. Furthermore, let $L_g := \sup_{x \in B(\bar{\Sigma}_0, \epsilon) \cap \text{dom } g} d(0, \partial g(x))$ and $L := \max\{L_1, \dots, L_N\}$, where $L_i := \sup_{x \in B(\bar{\Sigma}_0, \epsilon+1)} \|\nabla f_i(x)\|$ for $i = 1, \dots, N$. Since Φ is continuous over $\text{dom } g$ and C is compact, there exists $\delta \in (0, \epsilon/2)$ such that

$$\Phi(x) - \max_C \Phi \leq \frac{1}{4} \psi^{-1} \left(\frac{\epsilon}{3} \right), \quad \forall x \in B(C, \delta) \cap \text{dom } g. \quad (5.27)$$

Apply the result in the first step to this δ , we obtain the desired $\hat{\alpha} \in (0, \delta/(3L_\Phi)]$, k^* and x^* . Since

$B(\bar{\Sigma}_0, \epsilon)$ is bounded, by Corollary 5.2, there exist $\eta, \kappa > 0$ and $\tilde{\alpha} \in (0, \hat{\alpha}]$ such that for every $\alpha \in (0, \tilde{\alpha}]$,

$$\sum_{k=0}^K \|x_{k+1} - x_k\| \leq \psi(\Phi(x_0) - \Phi(x_K) + \eta\alpha) + \kappa\alpha. \quad (5.28)$$

for all $K = 0, \dots, K^* - 1$ holds for any $(x_0, \dots, x_{K^*}) \in (B(\bar{\Sigma}_0, \epsilon))^{K^*} \times \mathbb{R}^n$ generated by Algorithm 1 with $\alpha_k = \alpha/(k + \gamma)^\beta$ where $\alpha \in (0, \tilde{\alpha}]$.

If $K^* = \infty$, then $x_k \in \mathring{B}(C, \epsilon)$ for all $k \geq k^*$. Combined with the fact that $x_k \in B(\bar{\Sigma}_0, \delta/3)$ for all $k = 0, \dots, k^* - 1$, we have that $x_k \in B(\bar{\Sigma}_0, \epsilon)$ for $k \in \mathbb{N}$. In this case, a limiting argument shows

$$\sum_{k=0}^{\infty} \|x_{k+1} - x_k\| \leq \psi \left(\sup_{X_0} \Phi - \inf_{B(\bar{\Sigma}_0, \epsilon)} \Phi + \eta\alpha \right) + \kappa\alpha.$$

Thus, $\Gamma(X_0, \tilde{\alpha}) < \infty$ immediately follows in this case.

If $K^* < \infty$, then consider any $\alpha \in (0, \bar{\alpha}_0]$, where

$$\bar{\alpha}_0 := \min \left\{ \tilde{\alpha}, \frac{\epsilon}{4(NL + L_g)}, \frac{\epsilon}{6\kappa}, \frac{1}{2\eta} \psi^{-1} \left(\frac{\epsilon}{3} \right) \right\}.$$

Furthermore, $K^* \geq k^* + 2$ because Corollary 5.1 implies that

$$\begin{aligned} \|x_{k^*+1} - x^*\| &\leq \|x_{k^*+1} - x_{k^*}\| + \|x_{k^*} - x^*\| \leq 2\alpha_{k^*}(NL + L_g) + \delta \\ &\leq 2 \frac{\epsilon}{4(NL + L_g)}(NL + L_g) + \delta < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

We claim that at iteration $K^* - 1$, the objective value has admitted a sufficient decrease from the largest critical value in C , i.e.,

$$\Phi(x_{K^*-1}) \leq \max_C \Phi - \frac{1}{4} \psi^{-1} \left(\frac{\epsilon}{3} \right). \quad (5.29)$$

To verify the claim, first notice that

$$\sum_{k=k^*}^{K^*-1} \|x_{k+1} - x_k\| \geq \|x_{K^*} - x_{k^*}\| \geq \|x_{K^*} - x^*\| - \|x_{k^*} - x^*\| \geq \epsilon - \delta \geq \frac{\epsilon}{2}.$$

Furthermore, using the range of α and the fact that ψ^{-1} is increasing, we have

$$\psi^{-1} \left(\sum_{k=k^*}^{K^*-1} \|x_{k+1} - x_k\| - \kappa\alpha \right) \geq \psi^{-1} \left(\frac{\epsilon}{2} - \frac{\epsilon}{6} \right) = \psi^{-1} \left(\frac{\epsilon}{3} \right).$$

On the other hand, (5.28) imply that

$$\Phi(x_{k^*}) - \Phi(x_{K^*-1}) \geq \psi^{-1} \left(\sum_{k=k^*}^{K^*-1} \|x_{k+1} - x_k\| - \kappa\alpha \right) - \eta\alpha$$

Therefore, it yields that

$$\Phi(x_{k^*}) - \Phi(x_{K^*-1}) \geq \psi^{-1} \left(\frac{\epsilon}{3} \right) - \eta\alpha \geq \frac{1}{2} \psi^{-1} \left(\frac{\epsilon}{3} \right). \quad (5.30)$$

Finally, since $x_{k^*} \in B(C, \delta) \cap \text{dom } g$, combining (5.27) and (5.30), we obtain the desired claim (5.29).

Define the successive initial set of X_0 by

$$X_1 := B(C, \epsilon) \cap \left\{ x \in \text{dom } g : \Phi(x) \leq \max_C \Phi - \frac{1}{4} \psi^{-1} \left(\frac{\epsilon}{3} \right) \right\}. \quad (5.31)$$

From (5.29), we know that $x_{K^*-1} \in X_1$. By (5.28), for any feasible points $(x, \sigma, \alpha, \gamma)$ of (5.26), we

have

$$\begin{aligned}
\sum_{k=0}^{\infty} \|x_{k+1} - x_k\| &= \sum_{k=0}^{K^*-2} \|x_{k+1} - x_k\| + \sum_{k=K^*-1}^{\infty} \|x_{k+1} - x_k\| \\
&\leq \psi(\Phi(x_0) - \Phi(x_{K^*}) + \eta\alpha) + \kappa\alpha + \sum_{k=K^*-1}^{\infty} \|x_{k+1} - x_k\| \\
&\leq \psi\left(\sup_{X_0} \Phi - \inf_{B(\bar{\Sigma}_0, \epsilon)} \Phi + \eta\bar{\alpha}_0\right) + \kappa\bar{\alpha}_0 + \Gamma(X_1, \bar{\alpha}_0).
\end{aligned}$$

where in the last inequality we use the fact that if $(x, \sigma, \alpha, \gamma)$ is feasible for (5.26), then for any $\bar{k} \in \mathbb{N}$, $((x_k, \cdot)_{k \geq \bar{k}}, (\sigma_k)_{k \geq \bar{k}}, \alpha, \bar{k} + \gamma)$ is again feasible. Using the convention $\Gamma(\emptyset, \cdot) = -\infty$, by taking supremum over all feasible points $(x, \sigma, \alpha, \gamma)$ of (5.26), we can combine the case when $K^* = \infty$ and $K^* < \infty$ into

$$\Gamma(X_0, \bar{\alpha}_0) \leq \psi\left(\sup_{X_0} \Phi - \inf \Phi + \eta\bar{\alpha}_0\right) + \kappa\bar{\alpha}_0 + \max\{\Gamma(X_1, \bar{\alpha}_0), 0\}.$$

Note that the above inequality is also valid if $\bar{\alpha}_0$ is replaced by any $\alpha \in (0, \bar{\alpha}_0]$.

If $X_1 \neq \emptyset$, we may repeat the above arguments by replacing X_0 with X_1 . Recursively, we obtain sequences of $X_\ell, \bar{\alpha}_\ell, \psi_\ell, \eta_\ell, \kappa_\ell$ such that

$$\Gamma(X_\ell, \bar{\alpha}_\ell) \leq \psi_\ell\left(\sup_{X_\ell} \Phi - \inf \Phi + \eta_\ell\bar{\alpha}_\ell\right) + \kappa_\ell\bar{\alpha}_\ell + \max\{\Gamma(X_{\ell+1}, \bar{\alpha}_\ell), 0\}, \quad (5.32)$$

for any $\ell \in \mathbb{N}$ as long as $X_\ell \neq \emptyset$, where $\psi_0 := \psi, \eta_0 := \eta, \kappa_0 = \kappa$. Again, the above inequality remains valid when reducing $\bar{\alpha}_\ell$.

We finally show that $X_\ell = \emptyset$ for some $\ell \in \mathbb{N}$. Let v_ℓ be the maximum critical value of Φ in $(-\infty, \max_{X_\ell} \Phi]$. By the construction of the sequence of initial sets (5.31), it is evident that if neither X_ℓ nor $X_{\ell+1}$ is empty, then $v_\ell > \max_{X_{\ell+1}} \Phi \geq v_{\ell+1}$. As Φ has finitely many critical values by [43, Corollary 9], eventually $X_\ell = \emptyset$. Let $\bar{\ell} \geq 1$ be the smallest index such that $X_{\bar{\ell}} = \emptyset$. Telescoping

(5.32) with $\bar{\alpha}_\ell$ replaced by $\bar{\alpha} := \min\{\bar{\alpha}_\ell : \ell = 0, \dots, \bar{\ell} - 1\}$ yields

$$\Gamma(X_0, \bar{\alpha}) \leq \sum_{\ell=0}^{\bar{\ell}-1} \psi_\ell \left(\sup_{X_\ell} \Phi - \inf \Phi + \eta_\ell \bar{\alpha}_\ell \right) + \kappa_\ell \bar{\alpha}_\ell < \infty.$$

The finiteness of $\Gamma(X_0, \bar{\alpha})$ naturally means that $\|x_k - x_0\| \leq \sum_{i=0}^k \|x_i - x_{i-1}\| \leq \Gamma(X_0, \bar{\alpha})$ for all $k \in \mathbb{N}$, that is, $x_k \in B(X_0, \Gamma(X_0, \bar{\alpha}))$ (where $x_k := x_{k,0}$). Let L_0, M respectively be Lipschitz constants of f_1, \dots, f_N and $\nabla f_1, \dots, \nabla f_N$ on $B(X_0, \Gamma(X_0, \bar{\alpha}) + 2)$. After possibly reducing $\bar{\alpha}$, we have $\bar{\alpha} \leq 1/(NL_0)$. Since $x_{k+1} = \text{prox}_{\alpha_k g}(x_{k,N})$, Fermat's rule [38, Theorem 10.1] implies that $0 \in \alpha_k \partial g(x_{k+1}) + x_{k+1} - x_{k,N}$ and hence

$$\partial \Phi(x_{k+1}) = \nabla f(x_{k+1}) + \partial g(x_{k+1}) \ni \nabla f(x_{k+1}) - \frac{x_{k+1} - x_{k,N}}{\alpha_k}.$$

Since $x_{k,N} = x_k - \alpha_k \sum_{i=1}^N \nabla f_{\sigma_i^k}(x_{k,i-1})$, by Corollary 5.1 it follows that

$$\begin{aligned} d(0, \partial \Phi(x_{k+1})) &\leq \left\| \nabla f(x_{k+1}) - \frac{x_{k+1} - x_{k,N}}{\alpha_k} \right\| \\ &\leq \frac{\|x_{k+1} - x_k\|}{\alpha_k} + \left\| \nabla f(x_{k+1}) - \sum_{i=1}^N \nabla f_{\sigma_i^k}(x_{k,i-1}) \right\| \\ &\leq \frac{\|x_{k+1} - x_k\|}{\alpha_k} + \|\nabla f(x_{k+1}) - \nabla f(x_k)\| + \left\| \nabla f(x_k) - \sum_{i=1}^N \nabla f_{\sigma_i^k}(x_{k,i-1}) \right\| \\ &\leq \left(\frac{1}{\alpha_k} + M \right) \|x_{k+1} - x_k\| + (N-1) \sqrt{\frac{N(2N-1)}{6}} ML_0 \alpha_k. \end{aligned}$$

The last inequality admits a similar derivation as in the proof of Lemma 5.7. Note that for all nonnegative decreasing sequence u_0, u_1, u_2, \dots , we have $u_k \leq 2(\sum_{i=\lfloor k/2 \rfloor}^{\infty} u_i)/(k+2)$ as explained

in [45, Footnote 1]. In particular, for $u_k := \min_{i \in \llbracket 0, k \rrbracket} \alpha_i d(0, \partial\Phi(x_{i+1}))$, we have

$$\begin{aligned} \min_{i \in \llbracket 0, k \rrbracket} \alpha_i d(0, \partial\Phi(x_{i+1})) &\leq \frac{2}{k+2} \sum_{i=\lfloor k/2 \rfloor}^{\infty} \min_{j \in \llbracket 0, i \rrbracket} \alpha_j d(0, \partial\Phi(x_{j+1})) \\ &\leq \frac{2}{k+2} \sum_{i=\lfloor k/2 \rfloor}^{\infty} \alpha_i d(0, \partial\Phi(x_{i+1})) \\ &\leq \frac{2}{k+2} \sum_{i=\lfloor k/2 \rfloor}^{\infty} (1 + \alpha M) \|x_{i+1} - x_i\| + (N-1) \sqrt{\frac{N(2N-1)}{6}} ML_0 \alpha_i^2. \end{aligned}$$

Thus, using $\alpha_k = \alpha/(k+1)^\beta$, we have

$$\begin{aligned} \min_{i \in \llbracket 0, k \rrbracket} d(0, \partial\Phi(x_{i+1})) &\leq \frac{2}{\alpha_k(k+2)} \sum_{i=\lfloor k/2 \rfloor}^{\infty} (1 + \alpha M) \|x_{i+1} - x_i\| + (N-1) \sqrt{\frac{N(2N-1)}{6}} ML_0 \alpha_i^2 \\ &\leq \frac{2\alpha^{-1}}{(k+1)^{1-\beta}} \sum_{i=\lfloor k/2 \rfloor}^{\infty} (1 + \alpha M) \|x_{i+1} - x_i\| + (N-1) \sqrt{\frac{N(2N-1)}{6}} ML_0 \frac{\alpha^2}{(i+1)^{2\beta}} \\ &\leq \frac{2\alpha^{-1}}{(k+1)^{1-\beta}} \left((1 + \alpha M) \sum_{i=\lfloor k/2 \rfloor}^{\infty} \|x_{i+1} - x_i\| \right. \\ &\quad \left. + (N-1) \sqrt{\frac{N(2N-1)}{6}} ML_0 \left(\frac{\alpha^2}{(\lfloor k/2 \rfloor + 1)^{2\beta}} + \sum_{i=\lfloor k/2 \rfloor + 1}^{\infty} \int_i^{i+1} \frac{\alpha^2}{v^{2\beta}} dv \right) \right) \\ &\leq \frac{2\alpha^{-1}}{(k+1)^{1-\beta}} \left((1 + \alpha M) \sum_{i=\lfloor k/2 \rfloor}^{\infty} \|x_{i+1} - x_i\| \right. \\ &\quad \left. + (N-1) \sqrt{\frac{N(2N-1)}{6}} ML_0 \left(\frac{\alpha^2}{(\lfloor k/2 \rfloor + 1)^{2\beta}} + \int_{\lfloor k/2 \rfloor + 1}^{\infty} \frac{\alpha^2}{v^{2\beta}} dv \right) \right) \\ &\leq \frac{1}{(k+1)^{1-\beta}} \left(2(\alpha^{-1} + M) \sum_{i=\lfloor k/2 \rfloor}^{\infty} \|x_{i+1} - x_i\| \right. \\ &\quad \left. + (N-1) \sqrt{\frac{N(2N-1)}{6}} ML_0 \left(\frac{2\alpha}{(\lfloor k/2 \rfloor + 1)^{2\beta}} + \frac{2\alpha}{(2\beta-1)(\lfloor k/2 \rfloor + 1)^{2\beta-1}} \right) \right) \\ &\leq \frac{1}{(k+1)^{1-\beta}} \left(2(\alpha^{-1} + M) \sum_{i=\lfloor k/2 \rfloor}^{\infty} \|x_{i+1} - x_i\| + \frac{4(N-1)\sqrt{N(2N-1)}ML_0\alpha}{\sqrt{6}(2\beta-1)(\lfloor k/2 \rfloor + 1)^{2\beta-1}} \right). \end{aligned}$$

□

5.4.3 Projection formula

In this subsection, we present a variational stratification property for conservative fields and discuss its consequences. Specifically, we aim to recover a “projection formula” that relates the projection of generalized conservative fields to the Riemannian gradient of the potential function when restricted to certain smooth manifolds. Similar properties have been shown for definable conservative fields with locally Lipschitz definable potential functions [28, Theorem 4] and for Clarke subdifferentials of lower semicontinuous functions [43, Corollary 9].

Theorem 5.1. *Let $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a definable conservative field for a definable lower semicontinuous function $\Phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$. Then there is a Whitney stratification $\{M_1, \dots, M_m\}$ of $\text{dom } D$ such that for all $x \in \text{dom } D$, we have that $P_{T_{M_i}(x)}D(x) = \{\nabla_{M_i}\Phi(x)\}$ if $x \in M_i$ for some $i = 1, \dots, m$.*

Proof. We first claim that for any C^r definable manifold $M \subset \text{dom } \Phi$ satisfying $D(x) \subset T_x M$, there exists a Whitney stratification $\{M_i\}_{i=1}^m$ of M such that $\Phi|_{M_i}$ is C^r on M_i and R_i defined by

$$R_i := \{x \in M_i : \nabla_{M_i}\Phi(x) \neq P_{T_{M_i}(x)}D(x)\}, \quad \forall i = 1, \dots, m$$

satisfies $\dim R_i < \dim M_i$.

With the claim, the theorem can be proved by induction. Since $\text{dom } D$ is definable, there is a Whitney stratification $\{S_i^0\}_{i=1}^{m_0}$ of $\text{dom } D$ such that $\Phi|_{S_i^0}$ is C^p smooth on the C^p smooth manifold S_i^0 for all $i = 1, \dots, m_0$ [167]. Apply the claim to each manifold $M := S_i^0$ and conservative field $P_{T_{S_i^0}(x)}D(x)$ with potential $\Phi|_{S_i^0}$, we obtain a Whitney stratification $\{S_{ij}^0\}_{j=1}^{m_{0i}}$ of S_i^0 such that $\Phi|_{S_{ij}^0}$ is C^r on S_{ij}^0 and R_{ij}^0 defined by

$$R_{ij}^0 := \{x \in S_{ij}^0 : \nabla_{S_{ij}^0}\Phi(x) \neq P_{T_{S_{ij}^0}(x)}D(x)\}, \quad \forall j = 1, \dots, m_{0i}, \quad \forall i = 1, \dots, m_0$$

satisfies $\dim R_{ij}^0 < \dim S_{ij}^0$. It is clear that each R_{ij}^0 is a definable set, so again there is a Whitney stratification $\{R_{ijk}^0\}_{k=1}^{m_{0ij}}$ of R_{ij}^0 such that $\Phi|_{R_{ijk}^0}$ is C^p smooth on the C^p smooth manifold R_{ijk}^0 for

all $k = 1, \dots, m_{0ij}$. By renumbering the index, we can define a Whitney stratification

$$\{S_i^1\}_{i=1}^{m_1} := \{R_{ijk}^0 : k = 1, \dots, m_{0ij}, j = 1, \dots, m_{0i}, i = 1, \dots, m_0\}, \quad m_1 := \sum_{i=1}^{m_0} \sum_{j=1}^{m_{0i}} m_{0ij}.$$

Notice that $\{S_i^1\}_{i=1}^{m_1}$ is a refinement of $\{S_i^0\}_{i=1}^{m_0}$ with $\max_{i=1}^{m_1} \dim S_i^1 < \max_{i=1}^{m_0} \dim S_i^0$. One can continue this process by recursively until obtaining a Whitney stratification $\{S_i^L\}_{i=1}^{m_L}$ at the L -th iteration, where $\max_{i=1}^{m_L} \dim S_i^L = 0$. In this case, we can apply the claim once again to each manifold $M := S_i^L$ and obtain a Whitney stratification $\{S_{ij}^L\}_{j=1}^{m_{Lj}}$ such that $\Phi|_{S_{ij}^L}$ is C^r on S_{ij}^L and R_{ij}^L defined by

$$R_{ij}^L := \{x \in S_{ij}^L : \nabla \Phi_{S_{ij}^L}(x) \neq P_{T_{S_{ij}^L}(x)} D(x)\}, \quad \forall j = 1, \dots, m_{Li}, \quad \forall i = 1, \dots, m_L$$

satisfies $\dim R_{ij}^L < \dim S_{ij}^L$. Since $\dim S_{ij}^L = 0$, the only possibility is that $R_{ij}^L = \emptyset$. This terminates the inductive process. Now the desired Whitney stratification $\{M_i\}_{i \in I}$ can be defined by renumbering the index

$$\{M_i\}_{i \in I} := \{S_i^m : i = 1, \dots, m_m, m = 0, \dots, L\}.$$

With this Whitney stratification, we have $P_{T_x M_x} = \{\nabla_{M_x} f(x)\}$ for all $x \in \text{dom } D$, where M_x is the active strata of x .

Therefore, it remains to prove the claim. Since $\Phi|_{\text{dom } D}$ is definable, there is a Whitney stratification $\{M_i\}_{i=1}^m$ such that $\Phi|_{M_i}$ is C^p smooth on M_i and $\text{rank } \Phi|_{M_i}$ is constant over each M_i . Since $\Phi(x) \in \mathbb{R}$ for $x \in \text{dom } D$, we have $\text{rank } \Phi|_{M_i} \in \{0, 1\}$. It suffices to show that for any M_i , the corresponding R_i defined in the claim satisfies $\dim R_i < \dim M_i$.

We first deal with the case when $\text{rank } \Phi|_{M_i} = 0$. In this case Φ is a constant on M_i and it suffices to show for a.e. $x \in M_i$, we have $\sup\{\|v\| : v \in P_{T_{M_i}(x)} D(x)\} = 0$. Assume for the sake of contradiction, there exists $\delta, \rho > 0, \bar{x} \in M_i$ such that $\sup\{\|v\| : v \in P_{T_{M_i}(x)} D(x)\} > \delta$ for all $x \in B_{M_i}(\bar{x}, \rho)$. Consider a set-valued mapping G defined by $G(x) := \{v \in P_{T_{M_i}(x)} D(x) : \|v\| \geq \delta/2\}$.

Note that G is definable with nonempty values over $B_{M_i}(\bar{x}, \rho)$. By [37, Theorem 3.1, p. 25], there is a definable function $g : B_{M_i}(\bar{x}, \rho) \rightarrow \mathbb{R}^n$ such that $g(x) \in G(x)$ for all $x \in B_{M_i}(\bar{x}, \rho)$. Since g is a definable function, there is a Whitney stratification $\{M_{ij}\}_{j=1}^{m_i}$ of $B_{M_i}(\bar{x}, \rho)$ such that $g|_{M_{ij}}$ is C^p over M_{ij} . Since there is at least one $j \in \{1, \dots, m_i\}$ such that $\dim M_{ij} = \dim M_i$, we can find $\hat{x} \in B_{M_i}(\bar{x}, \rho)$ and $\epsilon \in (0, \rho)$ such that $B_{M_i}(\hat{x}, \epsilon) \subset B_{M_i}(\bar{x}, \rho)$ and $g|_{B_{M_i}(\hat{x}, \epsilon)}$ is C^p over $B_{M_i}(\hat{x}, \epsilon)$. Consider the absolutely continuous curve $\gamma : I \rightarrow B_{M_i}(\hat{x}, \epsilon)$ defined by

$$\gamma'(t) = g(\gamma(t)), \quad \gamma(0) = \hat{x}.$$

Since $\gamma'(t) \in T_{M_i}(x)$, for a.e. $t \in I$ we have

$$\langle \Phi \circ \gamma \rangle'(t) = \langle D(\gamma(t)), \gamma'(t) \rangle = \langle v^*(t), \gamma'(t) \rangle = \|\gamma'(t)\|^2 \geq \delta^2/4,$$

where $v^* \in D(\gamma(t))$ is chosen to satisfy $P_{T_{M_i}(x)}v^*(t) = \gamma'(t)$. This contradicts with the fact that Φ is a constant on M_i . Therefore, $P_{T_{M_i}(x)}D(x) = \{0\} = \{\nabla_{M_i}\Phi(x)\}$ except for x in a lower dimensional set.

For the case when $\text{rank}(\Phi|_{M_i}) = 1$, consider $\tilde{D} := D - \nabla_{M_i}\Phi$. Notice that for any absolutely continuous curve $\gamma : \mathbb{R}_+ \rightarrow M_i$,

$$\langle \tilde{D}(\gamma(t)), \gamma'(t) \rangle = \langle D(\gamma(t)), \gamma'(t) \rangle - \langle \nabla_{M_i}\Phi(\gamma(t)), \gamma'(t) \rangle = (\Phi \circ \gamma)'(t) - (\Phi \circ \gamma)'(t) = 0$$

by Definition 5.4. Thus, the constant function 0 is a potential function of the conservative field \tilde{D} . Obviously, the rank of constant function 0 is 0 on M_i , so the result in the previous paragraph can be applied to \tilde{D} , and we can obtain that for a.e. $x \in M_i$, we have $P_{T_{M_i}(x)}\tilde{D}(x) = \{0\}$. Note that $\nabla_{M_i}\Phi(x) \in T_{M_i}(x)$, so $P_{T_{M_i}(x)}D(x) = \{\nabla_{M_i}\Phi(x)\}$ except for x in a lower dimensional set. \square

We next state several consequences of the above projection formula. Thanks to the notion of generalized conservative fields, these results apply to a general setting, extending from those in [43] and [28]. The proofs are omitted as they follow essentially the same arguments in [43,

Corollary 5 and Theorem 14] or [28, Theorems 5 and 6]. The first consequence is the definable Morse-Sard theorem, which asserts that there are at most finitely many critical values associated with a definable conservative field.

Theorem 5.2. *Assume that $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a definable conservative field for $\Phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, where Φ is lower semicontinuous and definable. Then Φ has finitely many D -critical values.*

The projection formula also helps to generalize the nonsmooth Kurdyka–Łojasiewicz inequality ([28, Theorem 6], [44, Theorem 1], [43, Theorem 14]).

Theorem 5.3. *Assume that $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a definable conservative field for $\Phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, where Φ is lower semicontinuous and definable. Then there exists $\epsilon > 0$, $\phi : [0, \epsilon) \rightarrow \mathbb{R}_+$, definable strictly increasing, continuously differentiable on $(0, \epsilon)$ with $\phi(0) = 0$ and a continuous definable function $\chi : \mathbb{R}_+ \rightarrow (0, \infty)$, such that for all $x \in \text{dom } \Phi$ with $0 < |\Phi(x)| \leq \chi(\|x\|)$, it holds that*

$$d(0, D(x)) \geq \frac{1}{\phi'(|\Phi(x)|)}.$$

5.4.4 Continuous length formula

We now examine the lengths of subgradient trajectories under the assumption that both the conservative field and its potential function are definable in an o-minimal structure. In this setting, the lengths of bounded trajectories can be controlled by function variations through a desingularizing function. Similar results on the finiteness of trajectory lengths have been established for subgradient and conservative field dynamics when the objective function is locally Lipschitz ([44, Theorem 2], [28, Theorem 7], [207]). The following proposition extends [45, Proposition 6], which addresses the special case where f is locally Lipschitz and D is the Clarke subdifferential of f . The proof follows similar arguments as in [45, Proposition 6].

Proposition 5.1. *Let $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a definable conservative field for Φ , where $\Phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is a proper lower semicontinuous definable function that is continuous over its domain. Let $X \subset \text{dom } D$ be bounded. Assume that Φ has at most $m \in \mathbb{N}^*$ D -critical values in \overline{X} . Let ψ be a desingularizing*

function of (Φ, D) on X . If $x : [0, T] \rightarrow X$ with $T \geq 0$ is a solution to $x' \in -D(x)$, then

$$\frac{1}{2m} \int_0^T \|x'(t)\| dt \leq \psi \left(\frac{\Phi(x(0)) - \Phi(x(T))}{2m} \right).$$

Before we present the proof of Proposition 5.1, we need to first show a descent property of these trajectories. The proof of it is similar to [22, Lemmas 5.2 and 6.3] (see also [55]).

Proposition 5.2. *Let D be a conservative field for $\Phi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$. Let $T \in (0, \infty]$, and suppose $x : [0, T] \rightarrow \mathbb{R}^n$ is a solution to $x' \in -D(x)$. Then $\Phi \circ x$ is differentiable almost everywhere on $[0, T]$ with*

$$(\Phi \circ x)'(t) = -\|x'(t)\|^2 \quad \text{and} \quad \|x'(t)\| = d(0, D(x(t))).$$

Proof. By Definition 5.4, we have that $\Phi \circ x$ is differentiable a.e. on $[0, T]$ with

$$(\Phi \circ x)'(t) = \langle -x'(t), x'(t) \rangle = -\|x'(t)\|^2$$

by taking $v = -x'(t) \in D(x(t))$. We next show that $\|x'(t)\| = d(0, D(x(t)))$. Observe that we have $\langle D(t) - D(t), x'(t) \rangle = 0$. Let $W := \text{aff}(F(t) - F(t))$ and $y := -x'(t)$. Then it is easy to see that $y \in W^\perp$ and $y \in y + W$. By [55, Lemma 4.7], we have that $\|y\| = d(0, y + W)$. Notice that $D(t) \subset y + W$, and thus

$$d(0, D(t)) \geq d(0, y + W) = \|y\| = \|x'(t)\|$$

Since $-x'(t) \in D(t)$, we also have that $\|x'(t)\| \geq d(0, D(t))$, which proves the desired equality. □

Now we are ready to deliver the proof of Section 5.4.4. Note that it is similar to the one of [45, Proposition 6], despite the use of Theorem 5.2 and lemma 5.8.

Proof of Proposition 5.1. Let V be the set of D -critical values of Φ over \overline{X} if there exist, and otherwise let $V := \{0\}$. Since D and Φ are definable, by Theorem 5.2, we can write $V = \{v_1, \dots, v_\ell\}$,

where $\ell \leq m$ and $v_1 > \dots > v_\ell$. Furthermore, let $u_i = (v_i + v_{i+1})/2$ for $i = 1, \dots, \ell - 1$. By Proposition 5.2, Φ is decreasing along any solution to $x'(t) \in -D(x(t))$. Consider all v_i 's and u_i 's that are in $[\Phi(x(T)), \Phi(x(0))]$ and relabel them as $w_1 > \dots > w_p$ for some integer $0 \leq p \leq 2\ell - 1$. Note that $p = 0$ means $[\Phi(x(T)), \Phi(x(0))]$ does not contain any v_i or u_i . Now for $i = 1, \dots, p$, we define

$$t_i^- := \inf \{t \in [0, T] : \Phi(x(t)) = w_i\}, \quad t_i^+ := \sup \{t \in [0, T] : \Phi(x(t)) = w_i\}.$$

For convenience, we denote $t_0^+ := 0$ and $t_{p+1}^- := T$. Notice that according to the above definition, we have $0 \leq t_1^- \leq t_1^+ \leq \dots \leq t_p^- \leq t_p^+ \leq T$.

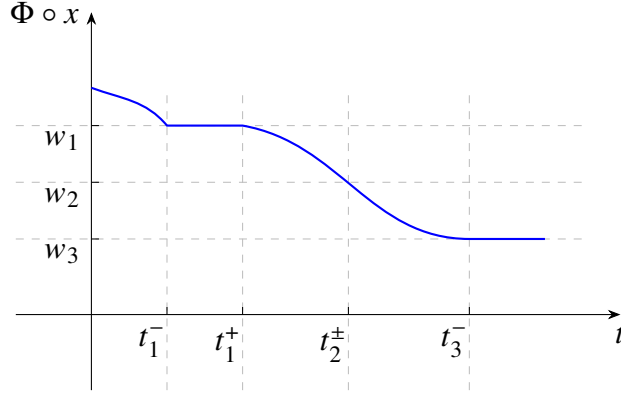


Figure 5.1: Visualization of function evolution in the proof of Proposition 5.1.

By Proposition 5.2, for any $0 \leq s \leq t \leq T$, we have

$$\Phi(x(s)) - \Phi(x(t)) = \int_s^t \|x'(\tau)\|^2 d\tau \quad \text{and} \quad \|x'(\tau)\| = d(0, D(x(\tau))) \quad \text{a.e. on } [s, t].$$

Thus, for every $i = 1, \dots, p$, we have

$$\int_{t_i^-}^{t_i^+} \|x'(\tau)\| d\tau = \Phi(x(t_i^-)) - \Phi(x(t_i^+)) = 0.$$

On the other hand, for each $i = 1, \dots, p + 1$, $\tilde{z}(\cdot) := d(\Phi(x(\cdot)), V)$ is either strictly increasing or strictly decreasing over $[t_{i-1}^+, t_i^-]$. If $\tilde{z}(\cdot)$ is increasing, then $\tilde{z}(\cdot) = w_{i-1} - \Phi(x(\cdot))$ over $[t_{i-1}^+, t_i^-]$

and hence $\tilde{z}'(\cdot) = -(\Phi \circ x)'(\cdot)$ a.e. on $[t_{i-1}^+, t_i^-]$; if $\tilde{z}(\cdot)$ is decreasing, then $\tilde{z}(\cdot) = \Phi(x(\cdot)) - w_i$ over $[t_{i-1}^+, t_i^-]$ and hence $\tilde{z}'(\cdot) = (\Phi \circ x)'(\cdot)$ a.e. on $[t_{i-1}^+, t_i^-]$. In both cases, for any $s, t \in [t_{i-1}^+, t_i^-]$, we have

$$\begin{aligned}
|\psi(\tilde{z}(t)) - \psi(\tilde{z}(s))| &= \left| \int_{\tilde{z}(s)}^{\tilde{z}(t)} \psi'(r) dr \right| = \left| \int_s^t \psi'(\tilde{z}(\tau)) d\tilde{z}(\tau) \right| && (r \leftarrow \tilde{z}(\tau)) \\
&= \left| \int_s^t \psi'(\tilde{z}(\tau)) d(\Phi \circ x)(\tau) \right| \\
&= \int_s^t \psi'(\tilde{z}(\tau)) \|x'(\tau)\| d(0, D(x(\tau))) d\tau && \text{(Proposition 5.2)} \\
&\geq \int_s^t \|x'(\tau)\| d\tau. && \text{(Lemma 5.8)}
\end{aligned}$$

In conclusion,

$$\int_0^T \|x'(\tau)\| d\tau = \sum_{i=1}^{p+1} \int_{t_{i-1}^+}^{t_i^-} \|x'(\tau)\| d\tau + \sum_{i=1}^p \int_{t_i^-}^{t_i^+} \|x'(\tau)\| d\tau \quad (5.33a)$$

$$= \sum_{i=1}^{p+1} |\psi(\tilde{z}(t_i^-)) - \psi(\tilde{z}(t_{i-1}^+))| + 0 \quad (5.33b)$$

$$\leq \sum_{i=1}^{p+1} \psi(|\tilde{z}(t_i^-) - \tilde{z}(t_{i-1}^+)|) \quad (5.33c)$$

$$= \sum_{i=1}^{p+1} \psi(\Phi(x(t_{i-1}^+)) - \Phi(x(t_i^-))) \quad (5.33d)$$

$$\leq (p+1)\psi\left(\frac{1}{p+1} \sum_{i=1}^{p+1} (\Phi(x(t_{i-1}^+)) - \Phi(x(t_i^-)))\right) \quad (5.33e)$$

$$= (p+1)\psi\left(\frac{\Phi(x(0)) - \Phi(x(T))}{p+1}\right) \quad (5.33f)$$

$$\leq 2m\psi\left(\frac{\Phi(x(0)) - \Phi(x(T))}{2m}\right), \quad (5.33g)$$

where (5.33c) is due to [100, Lemma 3.5] and (5.33g) is because of $p+1 \leq 2\ell \leq 2m$ and [206, Proposition 2.1]. \square

As a consequence of the above proposition, we show in the following corollary that any

bounded D -trajectory converges to a D -critical point.

Corollary 5.3. *Under the same assumptions as in Proposition 5.1, then for any bounded D -trajectory $x : \mathbb{R}_+ \rightarrow \mathbb{R}^n$ initialized at $x_0 \in \text{dom } D$, it holds that $x(t)$ converges to a D -critical point in $\text{dom } D$ as $t \rightarrow \infty$.*

Proof. Since $x(\cdot)$ is bounded, there exist $x^* \in \mathbb{R}^n$ and a positive sequence $(t_k)_{k \in \mathbb{N}}$ satisfying $t_k \rightarrow \infty$ and $x(t_k) \rightarrow x^*$. Fix any t , there exists a large enough k such that $t < t_k$ and

$$\|x(t) - x(t_k)\| \leq \int_t^{t_k} \|x'(\tau)\| d\tau.$$

Let $k \rightarrow \infty$ on both sides, we obtain that

$$\|x(t) - x^*\| \leq \int_t^\infty \|x'(\tau)\| d\tau.$$

By Proposition 5.1, we have that

$$\int_0^\infty \|x'(\tau)\| d\tau < \infty$$

and thus,

$$\int_t^\infty \|x'(\tau)\| d\tau \rightarrow 0, \quad \text{as } t \rightarrow \infty.$$

Therefore, $x(t) \rightarrow x^*$ as $t \rightarrow \infty$. Let $Z := \{t \in \mathbb{R}_+ : x(t) \notin -D(x(t))\}$, then Z is a set with zero measure. By the integrability of $\|x'\|$ on $(0, \infty)$, there is a positive sequence $(s_k)_{k \in \mathbb{N}}$ satisfying $s_k \notin Z$ for all $k \in \mathbb{N}$ and $s_k \rightarrow \infty$ as $k \rightarrow \infty$ with $x'(s_k) \rightarrow 0$. Notice that $x(s_k) \rightarrow x^*$ and $-x'(s_k) \in D(x(s_k))$, thus $0 \in D(x^*)$ due to the assumption that D has closed graph. It shows that $x^* \in \text{dom } D$ is a D -critical point of Φ and the desired result follows. \square

5.4.5 Uniform boundedness of subgradient trajectories

The following proposition establishes the uniform boundedness of trajectories for a definable conservative field, a key result in the proof of Theorem 5.4. It extends [45, Lemma 1] using similar proof techniques.

Proposition 5.3. *Under the same assumptions as in Proposition 5.1, if $d(0, D)$ is bounded over bounded subsets and Φ has bounded D -trajectories, then the D -trajectories initialized in a bounded subset of $\text{dom } D$ have uniformly bounded lengths.*

Proof. Let $X_0 \subset \text{dom } D$ be nonempty and bounded. It suffices to show that $\sigma(X_0) < \infty$ where

$$\sigma(X_0) := \sup_{x \in \mathcal{A}(\mathbb{R}_+, \mathbb{R}^n)} \int_0^\infty \|x'(t)\| dt \quad (5.34a)$$

$$\text{subject to } \begin{cases} x'(t) \in -D(x(t)), \text{ for a.e. } t > 0, \\ x(0) \in X_0. \end{cases} \quad (5.34b)$$

By Proposition 5.1, the feasible set of (5.34) is non-empty and $\sigma(X_0) > -\infty$.

We first consider the case with finite time horizon $T \geq 0$. By Theorem 5.2, Φ has finitely many D -critical values. Notice that $\sigma_T(X_0) \leq \sqrt{T(\sup_{X_0} \Phi - m(\Phi))}$ where $m(\Phi)$ is the smallest D -critical value of Φ and

$$\sigma_T(X_0) := \sup_{x \in \mathcal{A}(\mathbb{R}_+, \mathbb{R}^n)} \int_0^T \|x'(t)\| dt \quad (5.35a)$$

$$\text{subject to } \begin{cases} x'(t) \in -D(x(t)), \text{ for a.e. } t > 0, \\ x(0) \in X_0. \end{cases} \quad (5.35b)$$

Indeed, by the Cauchy-Schwarz inequality and Proposition 5.2, for any feasible point $x(\cdot)$ of (5.35),

$$\int_0^T \|x'(t)\| dt \leq \sqrt{T} \sqrt{\int_0^T \|x'(t)\|^2 dt} \quad (5.36a)$$

$$\leq \sqrt{T} \sqrt{\int_0^\infty \|x'(t)\|^2 dt} \quad (5.36b)$$

$$= \sqrt{T} \sqrt{\Phi(x(0)) - \Phi\left(\lim_{t \rightarrow \infty} x(t)\right)} \quad (5.36c)$$

$$\leq \sqrt{T \left(\sup_{X_0} \Phi - m(\Phi) \right)}. \quad (5.36d)$$

We next treat the case with infinite time horizon. Consider a sequence of feasible points $x_0(\cdot), x_1(\cdot), x_2(\cdot), \dots$ of (5.34) such that $\int_0^\infty \|x'_k(t)\| dt$ converges to $\sigma(X_0)$. We proceed to show that the sequence is equicontinuous. Let $\epsilon > 0$ and $t \geq 0$. Consider problem (5.35) with finite time horizon $T := t + \epsilon$. Since $x_0(\cdot), x_1(\cdot), x_2(\cdot), \dots$ are feasible points of (5.35), for all $s \in [0, T]$ and $k = 0, 1, 2, \dots$,

$$\|x_k(s)\| \leq \|x_k(s) - x_k(0)\| + \|x_k(0)\| \quad (5.37a)$$

$$\leq \int_0^s \|x'_k(\tau)\| d\tau + \|x_k(0)\| \quad (5.37b)$$

$$\leq \sigma_T(X_0) + \sup_{x_0 \in X_0} \|x_0\| =: r. \quad (5.37c)$$

All the trajectories $x_0(\cdot), x_1(\cdot), x_2(\cdot), \dots$ hence belong to $B(0, r)$ up to time T . We define $L := \sup_{x \in B(0, r) \cap \text{dom } D} d(0, D(x))$. Note that $L < \infty$ because $d(0, D)$ is bounded over bounded subsets of $\text{dom } D$. As a result, for all $s \in [t - \delta, t + \delta]$ with $\delta := \epsilon / (2L)$ and for all $k \in \mathbb{N}$, by Proposition 5.2,

$$\|x_k(t) - x_k(s)\| = \left\| \int_s^t x'_k(\tau) d\tau \right\| \leq \int_{t-\delta}^{t+\delta} \|x'_k(\tau)\| d\tau = \int_{t-\delta}^{t+\delta} d(0, D(x_k(\tau))) d\tau \leq 2\delta L = \epsilon.$$

It follows that $x_0(\cdot), x_1(\cdot), x_2(\cdot), \dots$ is equicontinuous on \mathbb{R}_+ . In addition, (5.37a)-(5.37c) imply that, for all $t \geq 0$, the sequence $x_0(t), x_1(t), x_2(t), \dots$ is bounded. The Arzelà-Ascoli theorem [46, Theorem 1, p. 13] implies that there exists a subsequence (again denoted $x_k(\cdot)$) converging

uniformly over compact intervals to a continuous function $x : \mathbb{R}_+ \rightarrow \mathbb{R}^n$.

We next show that $x(\cdot)$ is a D -trajectory of Φ . Let $T \geq 0$ be a real number. Similar to the above paragraph, one can obtain that $(x'_k(\tau))_{k \in \mathbb{N}}$ is bounded in $L^\infty([0, T], \mathbb{R}^n)$. By the Banach-Alaoglu theorem [46, Theorem 3 p. 13], there exists a subsequence (again denoted $x'_k(\cdot)$) that converges weakly* to a function $y(\cdot)$ in $L^\infty([0, T], \mathbb{R}^n)$. Together with $L^\infty([0, T], \mathbb{R}^n) \subset L^1([0, T], \mathbb{R}^n)$, we find that $x'_k(\cdot)$ converges weakly to $y(\cdot)$ in $L^1([0, T], \mathbb{R}^n)$. Since $x_k(t) - x_k(s) = \int_s^t x'_k(\tau) d\tau$ for all $0 \leq s \leq t \leq T$, we have $x(t) - x(s) = \int_s^t y(\tau) d\tau$. As a result, $x'(t) = y(t)$ for almost every $t \in (0, T)$. Since D has closed graph, it follows that $x'(t) \in -D(x(t))$ for almost every $t \in (0, T)$. As $T \geq 0$ was arbitrary, we have $x'(t) \in -D(x(t))$ for almost every $t > 0$.

Since Φ is definable and has bounded D -trajectories, by Corollary 5.3, $x(\cdot)$ converges to a D -critical point $x^* \in \text{dom } D$ of Φ . Let $\epsilon > 0$ and let $m \in \mathbb{N}^*$ be the number of D -critical values of Φ in $B(\overline{x(\mathbb{R}_+)}, \epsilon)$. Let ψ be a desingularizing function of Φ on $B(\overline{x(\mathbb{R}_+)}, \epsilon)$. Since Φ is continuous over its domain and $\text{dom } D \subset \text{dom } \Phi$, there exists $\delta \in (0, \epsilon/2)$ such that

$$\Phi(x) - \Phi(x^*) \leq m \psi^{-1}\left(\frac{\epsilon}{4m}\right), \quad \forall x \in B(x^*, \delta) \cap \text{dom } D. \quad (5.38)$$

Let $t^* \geq 0$ be such that $\|x(t) - x^*\| \leq \delta/2$ for all $t \geq t^*$. By the uniform convergence of $x_k(\cdot)$, we have that $\|x_k(t) - x(t)\| \leq \delta/2$ for all $t \in [0, t^*]$ for all k large enough. Hence

$$\|x_k(t^*) - x^*\| \leq \|x_k(t^*) - x(t^*)\| + \|x(t^*) - x^*\| \leq \delta/2 + \delta/2 = \delta.$$

If $T_k := \inf\{t \geq t^* : x_k(t) \notin \overset{\circ}{B}(x^*, \epsilon)\} < \infty$, then

$$\int_{t^*}^{T_k} \|x'_k(t)\| dt \geq \|x_k(T_k) - x_k(t^*)\| \geq \|x_k(T_k) - x^*\| - \|x_k(t^*) - x^*\| \geq \epsilon - \delta \geq \frac{\epsilon}{2}.$$

Combined with Proposition 5.1, it yields

$$\frac{\epsilon}{2} \leq \int_{t^*}^{T_k} \|x'_k(t)\| dt \leq 2m\psi\left(\frac{\Phi(x_k(t^*)) - \Phi(x_k(T_k))}{2m}\right).$$

Since ψ^{-1} is increasing, it is equivalent to

$$2m \psi^{-1} \left(\frac{\epsilon}{4m} \right) \leq \Phi(x_k(t^*)) - \Phi(x_k(T_k)).$$

Since $x_k(t^*) \in B(x^*, \delta) \cap \text{dom } D$, by (5.38), it follows that

$$\Phi(x_k(T_k)) \leq \Phi(x_k(t^*)) - 2m \psi^{-1} \left(\frac{\epsilon}{4m} \right) \leq \Phi(x^*) - m \psi^{-1} \left(\frac{\epsilon}{4m} \right).$$

Since $x_k(t) \in B(\overline{x(\mathbb{R}_+)}, \epsilon)$ for all $t \in [0, T_k]$ and $x_k(T_k)$ belongs to

$$X_1 := B(x^*, \epsilon) \cap \left\{ x \in \text{dom } D : \Phi(x) \leq \Phi(x^*) - m \psi^{-1} \left(\frac{\epsilon}{4m} \right) \right\},$$

by Proposition 5.1 and the definition of $\sigma(\cdot)$ in (5.34) we have

$$\begin{aligned} \int_0^\infty \|x'_k(t)\| dt &= \int_0^{T_k} \|x'_k(t)\| dt + \int_{T_k}^\infty \|x'_k(t)\| dt \\ &\leq 2m \psi \left(\frac{1}{2m} \left(\sup_{X_0} \Phi - \inf_{B(x^*, \epsilon)} \Phi \right) \right) + \max\{0, \sigma(X_1)\}. \end{aligned}$$

Note that the inequality still holds if $T_k = \infty$. By taking the limit, we get

$$\sigma(X_0) \leq 2m \psi \left(\frac{1}{2m} \left(\sup_{X_0} \Phi - \inf_{B(x^*, \epsilon)} \Phi \right) \right) + \max\{0, \sigma(X_1)\}.$$

It now suffices to replace X_0 by X_1 and repeat the proof starting below (5.36d). A maximizing sequence $\bar{x}_k(\cdot)$ corresponding to $\sigma(X_1)$ converges to a D -trajectory $\bar{x}(\cdot)$ whose initial point lies in the compact set X_1 . If $X_1 \neq \emptyset$, then the D -critical value $\Phi(\lim_{t \rightarrow \infty} \bar{x}(t))$ is less than or equal to $\Phi(x^*) - m\psi^{-1}(\epsilon/(4m)) < \Phi(x^*)$. By Theorem 5.2, Φ has finitely many D -critical values. Thus, it is eventually the case that one of the sets X_0, X_1, \dots is empty. We conclude that $\sigma(X_0) < \infty$ by the above recursive formula. \square

References

- [1] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” *Computer*, vol. 42, no. 8, pp. 30–37, 2009.
- [2] S. Zhang, L. Yao, A. Sun, and Y. Tay, “Deep learning based recommender system: A survey and new perspectives,” *ACM computing surveys (CSUR)*, vol. 52, no. 1, pp. 1–38, 2019.
- [3] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, “Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer,” in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1441–1450.
- [4] M Turk and A Pentland, “Eigenfaces for recognition,” *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
- [5] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.
- [6] S. Balaban, “Deep learning and face recognition: The state of the art,” *Biometric and surveillance technology for human and activity identification XII*, vol. 9457, pp. 68–75, 2015.
- [7] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, *et al.*, “End to end learning for self-driving cars,” *arXiv preprint arXiv:1604.07316*, 2016.
- [8] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, “A survey of deep learning techniques for autonomous driving,” *Journal of field robotics*, vol. 37, no. 3, pp. 362–386, 2020.
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

- [11] V. Cevher, S. Becker, and M. Schmidt, “Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics,” *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 32–43, 2014.
- [12] L. Bottou, F. E. Curtis, and J. Nocedal, “Optimization methods for large-scale machine learning,” *SIAM review*, vol. 60, no. 2, pp. 223–311, 2018.
- [13] E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for modern deep learning research,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 13 693–13 696.
- [14] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, and V. Gadepally, “From words to watts: Benchmarking the energy costs of large language model inference,” in *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, IEEE, 2023, pp. 1–9.
- [15] D. Bertsekas, *Nonlinear Programming* (Athena Scientific optimization and computation series). Athena Scientific, 1995, ISBN: 9781886529144.
- [16] Y. Nesterov, *Lectures on Convex Optimization*. Springer Science & Business Media, 2018, vol. 137.
- [17] A. Beck, *First-order methods in optimization*. SIAM, 2017.
- [18] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
- [19] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proceedings of the fourteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings*, 2011, pp. 315–323.
- [20] P.-A. Absil, R. Mahony, and B. Andrews, “Convergence of the iterates of descent methods for analytic cost functions,” *SIAM Journal on Optimization*, vol. 16, no. 2, pp. 531–547, 2005.
- [21] H. Attouch, J. Bolte, and B. F. Svaiter, “Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods,” *Mathematical Programming*, vol. 137, no. 1, pp. 91–129, 2013.
- [22] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee, “Stochastic subgradient method converges on tame functions,” *Foundations of computational mathematics*, vol. 20, no. 1, pp. 119–154, 2020.
- [23] C. Josz and L. Lai, “Global stability of first-order methods for coercive tame functions,” *Mathematical Programming*, pp. 1–26, 2023.

- [24] L. Ljung, “Analysis of recursive stochastic algorithms,” *IEEE transactions on automatic control*, vol. 22, no. 4, pp. 551–575, 1977.
- [25] H. J. Kushner, “General convergence results for stochastic approximations via weak convergence theory,” *Journal of mathematical analysis and applications*, vol. 61, no. 2, pp. 490–503, 1977.
- [26] M. Benaïm, “Dynamics of stochastic approximation algorithms,” in *Seminaire de probabilites XXXIII*, Springer, 2006, pp. 1–68.
- [27] J. C. Duchi and F. Ruan, “Stochastic methods for composite and weakly convex optimization problems,” *SIAM Journal on Optimization*, vol. 28, no. 4, pp. 3229–3259, 2018.
- [28] J. Bolte and E. Pauwels, “Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning,” *Mathematical Programming*, vol. 188, pp. 19–51, 2021.
- [29] A. Salim, “Random monotone operators and application to stochastic optimization,” Ph.D. dissertation, Université Paris-Saclay (ComUE), 2018.
- [30] E. Pauwels, “Incremental without replacement sampling in nonconvex optimization,” *Journal of Optimization Theory and Applications*, pp. 1–26, 2021.
- [31] P. Bianchi, W. Hachem, and S. Schechtman, “Convergence of constant step stochastic gradient descent for non-smooth non-convex functions,” *Set-Valued and Variational Analysis*, vol. 30, no. 3, pp. 1117–1147, 2022.
- [32] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, “Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality,” *Mathematics of operations research*, vol. 35, no. 2, pp. 438–457, 2010.
- [33] M. Korda, “Stability and performance verification of dynamical systems controlled by neural networks: Algorithms and complexity,” *IEEE Control Systems Letters*, vol. 6, pp. 3265–3270, 2022.
- [34] A. Daniilidis and D. Drusvyatskiy, “Pathological subgradient dynamics,” *SIAM Journal on Optimization*, vol. 30, no. 2, pp. 1327–1338, 2020.
- [35] J. Bolte and E. Pauwels, “Curiosities and counterexamples in smooth convex optimization,” *Mathematical Programming*, vol. 195, no. 1, pp. 553–603, 2022.
- [36] L. Van den Dries, *Tame topology and o-minimal structures*. Cambridge university press, 1998, vol. 248.

- [37] M. Coste, *An introduction to o-minimal geometry*. Istituti editoriali e poligrafici internazionali Pisa, 2000.
- [38] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009, vol. 317.
- [39] L. C. Evans and R. F. Garzepy, *Measure theory and fine properties of functions*. Oxfordshire: Routledge, 2018.
- [40] F. H. Clarke, *Optimization and Nonsmooth Analysis*. Philadelphia: SIAM Classics in Applied Mathematics, 1990.
- [41] J. Bochnak, M. Coste, and M.-F. Roy, *Real algebraic geometry*. Springer Science & Business Media, 2013, vol. 36.
- [42] T. S. Pham and H. H. Vui, *Genericity in polynomial optimization*. London: World Scientific, 2016, vol. 3.
- [43] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota, “Clarke subgradients of stratifiable functions,” *SIAM Journal on Optimization*, vol. 18, no. 2, pp. 556–572, 2007.
- [44] K. Kurdyka, “On gradients of functions definable in o-minimal structures,” in *Annales de l’institut Fourier*, vol. 48, 1998, pp. 769–783.
- [45] C. Josz, “Global convergence of the gradient method for functions definable in o-minimal structures,” *Mathematical Programming*, pp. 1–29, 2023.
- [46] J.-P. Aubin and A. Cellina, *Differential inclusions: set-valued maps and viability theory*. Berlin: Springer-Verlag, 1984, vol. 264.
- [47] O. A. Nielsen, *An introduction to integration and measure theory*. New York: Wiley-Interscience, 1997, vol. 17.
- [48] S. Marcellin and L. Thibault, “Evolution problems associated with primal lower nice functions,” *Journal of convex Analysis*, vol. 13, no. 2, p. 385, 2006.
- [49] F. Santambrogio, “{Euclidean, metric, and wasserstein} gradient flows: An overview,” *Bulletin of Mathematical Sciences*, vol. 7, no. 1, pp. 87–154, 2017.
- [50] W. Rudin *et al.*, *Principles of mathematical analysis*. McGraw-hill New York, 1964, vol. 3.
- [51] P. M. Fitzpatrick and H. L. Royden, *Real Analysis*, 4th ed. Upper Saddle River, NJ: Pearson, Jan. 2010.

- [52] I. Ekeland, “On the variational principle,” *Journal of Mathematical Analysis and Applications*, vol. 47, no. 2, pp. 324–353, 1974.
- [53] J.-B. Hiriart-Urruty, “A short proof of the variational principle for approximate solutions of a minimization problem,” *The American Mathematical Monthly*, vol. 90, no. 3, pp. 206–207, 1983.
- [54] T. X. D. Ha, “The ekeland variational principle for set-valued maps involving coderivatives,” *Journal of mathematical analysis and applications*, vol. 286, no. 2, pp. 509–523, 2003.
- [55] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis, “Curves of descent,” *SIAM Journal on Control and Optimization*, vol. 53, no. 1, pp. 114–138, 2015.
- [56] H. Attouch, G. Buttazzo, and G. Michaille, *Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization*. Philadelphia: SIAM, 2014.
- [57] J. Bolte and E. Pauwels, “Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning,” *Mathematical Programming*, pp. 1–33, 2020.
- [58] J. Borwein and X. Wang, “Lipschitz functions with maximal clarke subdifferentials are generic,” *Proceedings of the American Mathematical Society*, vol. 128, no. 11, pp. 3221–3229, 2000.
- [59] A. Daniilidis and G. Flores, “Linear structure of functions with maximal clarke subdifferential,” *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 511–521, 2019.
- [60] Z. Luo, “On the convergence of the lms algorithm with adaptive learning rate for linear feedforward networks,” *Neural Computation*, vol. 3, no. 2, pp. 226–245, 1991.
- [61] V. Elser, T.-Y. Lan, and T. Bendory, “Benchmark problems for phase retrieval,” *SIAM Journal on Imaging Sciences*, vol. 11, no. 4, pp. 2429–2455, 2018.
- [62] J. Miao, T. Ishikawa, Q. Shen, and T. Earnest, “Extending x-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes,” *Annu. Rev. Phys. Chem.*, vol. 59, no. 1, pp. 387–410, 2008.
- [63] C Fienup and J Dainty, “Phase retrieval and image reconstruction for astronomy,” *Image recovery: theory and application*, vol. 231, p. 275, 1987.
- [64] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, “Phase retrieval with application to optical imaging: A contemporary overview,” *IEEE signal processing magazine*, vol. 32, no. 3, pp. 87–109, 2015.

- [65] Y. Chi, Y. M. Lu, and Y. Chen, “Nonconvex optimization meets low-rank matrix factorization: An overview,” *IEEE Transactions on Signal Processing*, vol. 67, no. 20, pp. 5239–5269, 2019.
- [66] B. Recht, M. Fazel, and P. A. Parrilo, “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.
- [67] N. Gillis and S. A. Vavasis, “On the complexity of robust pca and ℓ_1 -norm low-rank matrix approximation,” *Mathematics of Operations Research*, vol. 43, no. 4, pp. 1072–1084, 2018.
- [68] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy, “Low-rank matrix recovery with composite optimization: Good conditioning and rapid convergence,” *Foundations of Computational Mathematics*, pp. 1–89, 2021.
- [69] S. S. Du, W. Hu, and J. D. Lee, “Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [70] W.-J. Zeng and H. C. So, “Outlier-robust matrix completion via ℓ_p -minimization,” *IEEE Transactions on Signal Processing*, vol. 66, no. 5, pp. 1125–1140, 2017.
- [71] R. M. Wallace, “Analysis of absorption spectra of multicomponent systems,” *The Journal of Physical Chemistry*, vol. 64, no. 7, pp. 899–901, 1960.
- [72] P. Paatero and U. Tapper, “Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values,” *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.
- [73] N. Gillis, *Nonnegative matrix factorization*. SIAM, 2020.
- [74] B. Cornet, “Existence of slow solutions for a class of differential inclusions,” *Journal of mathematical analysis and applications*, vol. 96, no. 1, pp. 130–147, 1983.
- [75] R. Poliquin, “Integration of subdifferentials of nonconvex functions,” *Nonlinear Analysis: Theory, Methods & Applications*, vol. 17, no. 4, pp. 385–398, 1991.
- [76] B. Bah, H. Rauhut, U. Terstiege, and M. Westdickenberg, “Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers,” *Information and Inference: A Journal of the IMA*, vol. 11, no. 1, pp. 307–353, 2022.
- [77] A. Eftekhari, “Training linear neural networks: Non-local convergence and complexity results,” in *International Conference on Machine Learning*, PMLR, 2020, pp. 2836–2847.
- [78] K. Chen, D. Lin, and Z. Zhang, “On non-local convergence analysis of deep linear networks,” in *International Conference on Machine Learning*, PMLR, 2022, pp. 3417–3443.

- [79] C. Jozs and X. Li, “Certifying the absence of spurious local minima at infinity,” *SIAM Journal on Optimization*, vol. 33, pp. 1416–1439, 3 2023.
- [80] K. L. Blackmore, R. C. Williamson, and I. M. Mareels, “Local minima and attractors at infinity for gradient descent learning algorithms,” *Journal of Mathematical Systems Estimation and Control*, vol. 6, pp. 231–234, 1996.
- [81] S. Liang, R. Sun, J. D. Lee, and R. Srikant, “Adding one neuron can eliminate all bad local minima,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [82] J. Sohl-Dickstein and K. Kawaguchi, “Eliminating all bad local minima from loss landscapes without even adding an extra unit,” *arXiv preprint arXiv:1901.03909*, 2019.
- [83] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, “Gradient Descent Only Converges to Minimizers,” *COLT*, 2016.
- [84] C. Jozs, Y. Ouyang, R. Y. Zhang, J. Lavaei, and S. Sojoudi, “A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization,” *NeurIPS*, Dec. 2018.
- [85] L. Venturi, A. S. Bandeira, and J. Bruna, “Spurious valleys in one-hidden-layer neural network optimization landscapes,” *Journal of Machine Learning Research*, vol. 20, p. 133, 2019.
- [86] C. D. Freeman and J. Bruna, “Topology and geometry of half-rectified network optimization,” in *International Conference on Learning Representations*, 2017.
- [87] T. Ding, D. Li, and R. Sun, “Suboptimal local minima exist for wide neural networks with smooth activations,” *Mathematics of Operations Research*, 2022.
- [88] D. Li, T. Ding, and R. Sun, “On the benefit of width for neural networks: Disappearance of basins,” *SIAM Journal on Optimization*, vol. 32, no. 3, pp. 1728–1758, 2022.
- [89] J. R. Munkres, “Topology,” *Prentice Hall, US*, 2000.
- [90] K. Kawaguchi, “Deep learning without poor local minima,” in *Advances in Neural Information Processing Systems*, PMLR, vol. 29, 2016.
- [91] T. Laurent and J. Brecht, “Deep linear networks with arbitrary loss: All local minima are global,” in *International conference on machine learning*, PMLR, 2018, pp. 2902–2907.
- [92] L. Zhang, “Depth creates no more spurious local minima,” *arXiv preprint arXiv:1901.09827*, 2019.

- [93] A. J. Wilkie, “Model completeness results for expansions of the ordered field of real numbers by restricted pfaffian functions and the exponential function,” *Journal of the American Mathematical Society*, vol. 9, no. 4, pp. 1051–1094, 1996.
- [94] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, “Global optimality in low-rank matrix optimization,” *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3614–3628, 2018.
- [95] D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi, “Non-square matrix sensing without spurious local minima via the burer-monteiro approach,” in *Artificial Intelligence and Statistics*, PMLR, 2017, pp. 65–74.
- [96] S. Li, Q. Li, Z. Zhu, G. Tang, and M. B. Wakin, “The global geometry of centralized and distributed low-rank matrix recovery without regularization,” *IEEE Signal Processing Letters*, vol. 27, pp. 1400–1404, 2020.
- [97] C. Jozs and L. Lai, “Nonsmooth rank-one matrix factorization landscape,” *Optimization Letters*, pp. 1–21, 2021.
- [98] I. Safran and O. Shamir, “Spurious local minima are common in two-layer relu neural networks,” in *International conference on machine learning*, PMLR, 2018, pp. 4433–4441.
- [99] R. Y. Zhang, S. Sojoudi, and J. Lavaei, “Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery,” *Journal of Machine Learning Research*, vol. 20, no. 114, pp. 1–34, 2019.
- [100] C. Jozs, L. Lai, and X. Li, “Convergence of the momentum method for semialgebraic functions with locally lipschitz gradients,” *SIAM Journal on Optimization*, vol. 33, no. 4, pp. 3012–3037, 2023.
- [101] N. B. Kovachki and A. M. Stuart, “Continuous time analysis of momentum methods,” *Journal of Machine Learning Research*, vol. 22, no. 17, pp. 1–40, 2021.
- [102] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” in *NIPS 2017 Workshop on Autodiff*, Long Beach, California, USA, 2017.
- [103] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimeshein, L. Antiga, *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.
- [104] M. Abadi, “Tensorflow: Learning functions at scale,” in *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming*, 2016, pp. 1–1.
- [105] B. T. Polyak, “Some methods of speeding up the convergence of iteration methods,” *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

- [106] P. Ochs, Y. Chen, T. Brox, and T. Pock, “Ipiano: Inertial proximal algorithm for nonconvex optimization,” *SIAM Journal on Imaging Sciences*, vol. 7, no. 2, pp. 1388–1419, 2014.
- [107] P. Ochs, “Local convergence of the heavy-ball method and ipiano for non-convex optimization,” *Journal of Optimization Theory and Applications*, vol. 177, no. 1, pp. 153–180, 2018.
- [108] S. Łojasiewicz, “Ensembles semi-analytiques,” *IHES notes*, 1965.
- [109] S. Zavriev and F. Kostyuk, “Heavy-ball method in nonconvex optimization problems,” *Computational Mathematics and Modeling*, vol. 4, no. 4, pp. 336–341, 1993.
- [110] B. Wen, X. Chen, and T. K. Pong, “Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems,” *SIAM Journal on Optimization*, vol. 27, no. 1, pp. 124–145, 2017.
- [111] Z. Jia, Z. Wu, and X. Dong, “An inexact proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth optimization problems,” *Journal of Inequalities and Applications*, vol. 2019, no. 1, pp. 1–16, 2019.
- [112] T. Sun, D. Li, Z. Quan, H. Jiang, S. Li, and Y. Dou, “Heavy-ball algorithms always escape saddle points,” *IJCAI*, 2019.
- [113] M. O’Neill and S. J. Wright, “Behavior of accelerated gradient methods near critical points of nonconvex functions,” *Mathematical Programming*, vol. 176, no. 1, pp. 403–427, 2019.
- [114] R. Pemantle, “Nonconvergence to unstable points in urn models and stochastic approximations,” *The Annals of Probability*, vol. 18, no. 2, pp. 698–712, 1990.
- [115] G. Garrigos, “Descent dynamical systems and algorithms for tame optimization, and multi-objective problems,” Ph.D. dissertation, Université Montpellier; Universidad técnica Federico Santa María (Valparaiso), 2015.
- [116] E. A. Coddington and N. Levinson, *Theory of ordinary differential equations*. Tata McGraw-Hill Education, 1955.
- [117] S. Ghadimi and G. Lan, “Accelerated gradient methods for nonconvex nonlinear and stochastic programming,” *Mathematical Programming*, vol. 156, no. 1-2, pp. 59–99, 2016.
- [118] M. Shub, *Global stability of dynamical systems*. Springer Science & Business Media, 2013.
- [119] J. Bolte, S. Sabach, and M. Teboulle, “Proximal alternating linearized minimization for nonconvex and nonsmooth problems,” *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.

- [120] G. Li and T. K. Pong, “Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods,” *Foundations of computational mathematics*, vol. 18, no. 5, pp. 1199–1232, 2018.
- [121] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.
- [122] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, “First-order methods almost always avoid strict saddle points,” *Mathematical programming*, vol. 176, no. 1, pp. 311–337, 2019.
- [123] V. Zorich and R. Cooke, *Mathematical Analysis I* (Mathematical Analysis). Springer, 2004, ISBN: 9783540403869.
- [124] S. P. Ponomarev, “Submersions and preimages of sets of measure zero,” *Siberian Mathematical Journal*, vol. 28, no. 1, pp. 153–163, 1987.
- [125] J. R. Sylvester, “Determinants of block matrices,” *The Mathematical Gazette*, vol. 84, no. 501, pp. 460–467, 2000.
- [126] C. Josz, L. Lai, and X. Li, “Proximal random reshuffling under local lipschitz continuity,” *arXiv preprint arXiv:2408.07182*, 2024.
- [127] H Kushner, “Convergence of recursive adaptive and identification procedures via weak convergence theory,” *IEEE Transactions on Automatic Control*, vol. 22, no. 6, pp. 921–930, 1977.
- [128] M. Benaïm, J. Hofbauer, and S. Sorin, “Stochastic approximations and differential inclusions,” *SIAM Journal on Control and Optimization*, vol. 44, no. 1, pp. 328–348, 2005.
- [129] M. Benaïm, J. Hofbauer, and S. Sorin, “Stochastic approximations and differential inclusions, part ii: Applications,” *Mathematics of Operations Research*, vol. 31, no. 4, pp. 673–695, 2006.
- [130] H Brézis, “Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de hilbert. number 5 in north holland math,” *Studies. North-Holland, Amsterdam*, 1973.
- [131] S. Schechtman, “Stochastic proximal subgradient descent oscillates in the vicinity of its accumulation set,” *Optimization Letters*, vol. 17, no. 1, pp. 177–190, 2023.
- [132] X. Li, A. Milzarek, and J. Qiu, “Convergence of Random Reshuffling under the Kurdyka–Łojasiewicz Inequality,” *SIAM Journal on Optimization*, vol. 33, no. 2, pp. 1092–1120, 2023.

- [133] Y.-X. Wang and Y.-J. Zhang, “Nonnegative matrix factorization: A comprehensive review,” *IEEE Transactions on knowledge and data engineering*, vol. 25, no. 6, pp. 1336–1353, 2012.
- [134] M. V. Solodov, “Convergence properties of proximal (sub) gradient methods without convexity or smoothness of any of the functions,” *SIAM Journal on Optimization*, vol. 35, no. 1, pp. 28–41, 2025.
- [135] J. Duchi and Y. Singer, “Efficient online and batch learning using forward backward splitting,” *The Journal of Machine Learning Research*, vol. 10, pp. 2899–2934, 2009.
- [136] D. Davis and D. Drusvyatskiy, “Stochastic model-based minimization of weakly convex functions,” *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 207–239, 2019.
- [137] J. Bolte, T. Le, É. Moulines, and E. Pauwels, “Inexact subgradient methods for semialgebraic functions,” *arXiv preprint arXiv:2404.19517*, 2024.
- [138] K. Ding, N. Xiao, and K.-C. Toh, “Adam-family methods with decoupled weight decay in deep learning,” *arXiv preprint arXiv:2310.08858*, 2023.
- [139] N. Xiao, X. Hu, and K.-C. Toh, “Stochastic subgradient methods with guaranteed global stability in nonsmooth nonconvex optimization,” *arXiv preprint arXiv:2307.10053*, 2023.
- [140] Y. Nesterov, “Gradient methods for minimizing composite functions,” *Mathematical programming*, vol. 140, no. 1, pp. 125–161, 2013.
- [141] C. Kanzow and P. Mehlitz, “Convergence properties of monotone and nonmonotone proximal gradient methods revisited,” *Journal of Optimization Theory and Applications*, vol. 195, no. 2, pp. 624–646, 2022.
- [142] P. Frankel, G. Garrigos, and J. Peypouquet, “Splitting methods with variable metric for kurdyka–łojasiewicz functions and general convergence rates,” *Journal of Optimization Theory and Applications*, vol. 165, pp. 874–900, 2015.
- [143] X. Jia, C. Kanzow, and P. Mehlitz, “Convergence Analysis of the Proximal Gradient Method in the Presence of the Kurdyka–Łojasiewicz Property without Global Lipschitz Assumptions,” *SIOPT*, vol. 33, no. 4, pp. 3038–3056, 2023.
- [144] D. P. Bertsekas *et al.*, “Incremental gradient, subgradient, and proximal methods for convex optimization: A survey,” *Optimization for Machine Learning*, vol. 2010, no. 1-38, p. 3, 2011.
- [145] K. Mishchenko, A. Khaled, and P. Richtárik, “Random reshuffling: Simple analysis with vast improvements,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 309–17 320, 2020.

- [146] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo, “Why random reshuffling beats stochastic gradient descent,” *Mathematical Programming*, vol. 186, no. 1, pp. 49–84, 2021.
- [147] O. Mangasariany and M. Solodovy, “Serial and parallel backpropagation convergence via nonmonotone perturbed minimization,” *Optimization Methods and Software*, 1994.
- [148] D. P. Bertsekas and J. N. Tsitsiklis, “Gradient convergence in gradient methods with errors,” *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 627–642, 2000.
- [149] M. Gurbuzbalaban, A. Ozdaglar, and P. A. Parrilo, “Convergence rate of incremental gradient and incremental newton methods,” *SIAM Journal on Optimization*, vol. 29, no. 4, pp. 2542–2565, 2019.
- [150] J. Haochen and S. Sra, “Random shuffling beats sgd after finite epochs,” in *International Conference on Machine Learning*, PMLR, 2019, pp. 2624–2633.
- [151] A. Khaled and P. Richtárik, “Better theory for sgd in the nonconvex world,” *arXiv preprint arXiv:2002.03329*, 2020.
- [152] L. M. Nguyen, Q. Tran-Dinh, D. T. Phan, P. H. Nguyen, and M. van Dijk, “A unified convergence analysis for shuffling-type gradient methods,” *Journal of Machine Learning Research*, vol. 22, no. 207, pp. 1–44, 2021.
- [153] X. Li, A. Milzarek, and J. Qiu, “A new random reshuffling method for nonsmooth nonconvex finite-sum optimization,” *arXiv preprint arXiv:2312.01047*, 2023.
- [154] L. Rosasco, S. Villa, and B. C. Vũ, “Convergence of stochastic proximal gradient algorithm,” *Applied Mathematics & Optimization*, vol. 82, pp. 891–917, 2020.
- [155] S. Majewski, B. Miasojedow, and E. Moulines, “Analysis of nonsmooth stochastic approximation: The differential inclusion approach,” *arXiv preprint arXiv:1805.01916*, 2018.
- [156] K. Mishchenko, A. Khaled, and P. Richtárik, “Proximal and federated random reshuffling,” in *International Conference on Machine Learning*, PMLR, 2022, pp. 15 718–15 749.
- [157] V. B. Tadić, “Convergence and convergence rate of stochastic gradient search in the case of multiple and non-isolated extrema,” *Stochastic Processes and their Applications*, vol. 125, no. 5, pp. 1715–1755, 2015.
- [158] S. Dereich and S. Kassing, “Convergence of stochastic gradient descent schemes for lojasiewicz-landscapes,” *arXiv preprint arXiv:2102.09385*, 2021.
- [159] E. A. Nurminkii, “The quasigradient method for the solving of the nonlinear programming problems,” *Cybernetics*, vol. 9, no. 1, pp. 145–150, 1973.

- [160] J.-P. Vial, “Strong and weak convexity of sets and functions,” *Mathematics of Operations Research*, vol. 8, no. 2, pp. 231–259, 1983.
- [161] J.-J. Moreau, “Proximité et dualité dans un espace hilbertien,” *Bulletin de la Société mathématique de France*, vol. 93, pp. 273–299, 1965.
- [162] L. Tian and A. M.-C. So, “No dimension-free deterministic algorithm computes approximate stationarities of Lipschitzians,” *Mathematical Programming*, pp. 1–24, 2024.
- [163] L. Tian and A. M.-C. So, “On the hardness of computing near-approximate stationary points of clarke regular nonsmooth nonconvex problems and certain dc programs,” in *ICML Workshop on Beyond First-Order Methods in ML Systems*, 2021.
- [164] D. Davis and B. Grimmer, “Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems,” *SIAM Journal on Optimization*, vol. 29, no. 3, pp. 1908–1930, 2019.
- [165] J. Zhang, H. Lin, S. Jegelka, S. Sra, and A. Jadbabaie, “Complexity of finding stationary points of nonconvex nonsmooth functions,” in *International Conference on Machine Learning*, PMLR, 2020, pp. 11 173–11 182.
- [166] A. Goldstein, “Optimization of Lipschitz continuous functions,” *Mathematical Programming*, vol. 13, pp. 14–22, 1977.
- [167] L. Van den Dries and C. Miller, “Geometric categories and o-minimal structures,” *Duke Mathematical Journal*, vol. 84, no. 2, pp. 497–540, 1996.
- [168] J. Bolte, T. Le, and E. Pauwels, “Subgradient sampling for nonsmooth nonconvex minimization,” *SIAM Journal on Optimization*, vol. 33, no. 4, pp. 2542–2569, 2023.
- [169] F. H. Clarke, “Generalized gradients and applications,” *Transactions of the American Mathematical Society*, vol. 205, pp. 247–262, 1975.
- [170] R. Rockafellar, *Convex Analysis* (Princeton mathematical series). Princeton University Press, 1970, ISBN: 9780691080697.
- [171] B. Schwartz, “Finite extensions of convex functions,” *Mathematische Operationsforschung und Statistik. Series Optimization*, vol. 10, no. 4, pp. 501–509, 1979.
- [172] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, “Rank-sparsity incoherence for matrix decomposition,” *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

- [173] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer, “Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery,” *IEEE Transactions on Network Science and Engineering*, vol. 1, no. 1, pp. 10–22, 2014.
- [174] D. Davis, D. Drusvyatskiy, and C. Paquette, “The nonsmooth landscape of phase retrieval,” *IMA Journal of Numerical Analysis*, vol. 40, no. 4, pp. 2652–2695, 2020.
- [175] H. Robbins and S. Monro, “A stochastic approximation method,” *The annals of mathematical statistics*, pp. 400–407, 1951.
- [176] B. T. Polyak, “Minimization of unsmooth functionals,” *USSR Computational Mathematics and Mathematical Physics*, vol. 9, no. 3, pp. 14–29, 1969.
- [177] J.-L. Goffin, “On convergence rates of subgradient optimization methods,” *Mathematical programming*, vol. 13, pp. 329–347, 1977.
- [178] D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette, “Subgradient methods for sharp weakly convex functions,” *Journal of Optimization Theory and Applications*, vol. 179, pp. 962–982, 2018.
- [179] X. Li, Z. Zhu, A. Man-Cho So, and R. Vidal, “Nonconvex robust low-rank matrix recovery,” *SIAM Journal on Optimization*, vol. 30, no. 1, pp. 660–686, 2020.
- [180] D. Davis, D. Drusvyatskiy, and V. Charisopoulos, “Stochastic algorithms with geometric step decay converge linearly on sharp functions,” *Mathematical Programming*, vol. 207, no. 1, pp. 145–190, 2024.
- [181] H. Yu and X. Li, “High probability guarantees for random reshuffling,” *arXiv preprint arXiv:2311.11841*, 2023.
- [182] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [183] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini, “Group sparse regularization for deep neural networks,” *Neurocomputing*, vol. 241, pp. 81–89, 2017.
- [184] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 68, no. 1, pp. 49–67, 2006.
- [185] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011.

- [186] S. Fattahi and S. Sojoudi, “Exact guarantees on the absence of spurious local minima for non-negative rank-1 robust principal component analysis,” *Journal of machine learning research*, 2020.
- [187] D. Drusvyatskiy and C. Paquette, “Efficiency of minimizing compositions of convex functions and smooth maps,” *Mathematical Programming*, vol. 178, pp. 503–558, 2019.
- [188] J. Ma and S. Fattahi, “Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization,” *Journal of Machine Learning Research*, vol. 24, no. 96, pp. 1–84, 2023.
- [189] Y. Chen, Y. Chi, and A. J. Goldsmith, “Exact and stable covariance estimation from quadratic sampling via convex programming,” *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 4034–4059, 2015.
- [190] D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” *Advances in neural information processing systems*, vol. 13, 2000.
- [191] A. Cichocki, R. Zdunek, and S.-i. Amari, “Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization,” in *International conference on independent component analysis and signal separation*, Springer, 2007, pp. 169–176.
- [192] J. Kim, Y. He, and H. Park, “Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework,” *Journal of Global Optimization*, vol. 58, pp. 285–319, 2014.
- [193] H. Kim and H. Park, “Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method,” *SIAM journal on matrix analysis and applications*, vol. 30, no. 2, pp. 713–730, 2008.
- [194] C.-J. Lin, “Projected gradient methods for nonnegative matrix factorization,” *Neural computation*, vol. 19, no. 10, pp. 2756–2779, 2007.
- [195] D. Hajinezhad, T.-H. Chang, X. Wang, Q. Shi, and M. Hong, “Nonnegative matrix factorization using admm: Algorithm and convergence analysis,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 4742–4746.
- [196] M. C. Muckamala and P. Ochs, “Beyond alternating updates for matrix factorization with inertial bregman proximal gradient algorithms,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [197] M. Teboulle and Y. Vaisbourd, “Novel proximal gradient methods for nonnegative matrix factorization with sparsity constraints,” *SIAM Journal on Imaging Sciences*, vol. 13, no. 1, pp. 381–421, 2020.

- [198] N. Takahashi and R. Hibi, “Global convergence of modified multiplicative updates for nonnegative matrix factorization,” *Computational Optimization and Applications*, vol. 57, pp. 417–440, 2014.
- [199] T. Kimura and N. Takahashi, “Global convergence of a modified HALS algorithm for non-negative matrix factorization,” in *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, IEEE, 2015, pp. 21–24.
- [200] P. H. Calamai and J. J. Moré, “Projected gradient methods for linearly constrained problems,” *Mathematical programming*, vol. 39, no. 1, pp. 93–116, 1987.
- [201] A. Rakotomamonjy, “Direct optimization of the dictionary learning problem,” *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5495–5506, 2013.
- [202] C. Jozs and L. Lai, “Lyapunov stability of the subgradient method with constant step size,” *Mathematical Programming*, pp. 1–10, 2023.
- [203] S. Cobzas and C. Mustata, “Norm preserving extension of convex lipschitz functions,” *J. Approx. theory*, vol. 24, no. 3, pp. 236–244, 1978.
- [204] A. Lewis and T. Tian, “Identifiability, the KL property in metric spaces, and subgradient curves,” *arXiv preprint arXiv:2205.02868*, 2022.
- [205] P. W. Gwanyama, “The HM-GM-AM-QM inequalities,” *College Mathematics Journal*, pp. 47–50, 2004.
- [206] P. Maréchal, “On a functional operation generating convex functions, part 1: Duality,” *Journal of Optimization Theory and Applications*, vol. 126, no. 1, pp. 175–189, 2005.
- [207] S. Lojasiewicz, “Sur les trajectoires du gradient d’une fonction analytique,” *Seminari di geometria*, vol. 1983, pp. 115–117, 1982.