Inexact subgradient methods for semialgebraic functions

Jérôme Bolte, Tam Le, Éric Moulines, Edouard Pauwels

Inexact subgradient methods for semialgebraic functions

Jérôme Bolte* Tam Le † Eric Moulines ‡ Edouard Pauwels§

May 12, 2025

Abstract

Motivated by the extensive application of approximate gradients in machine learning and optimization, we investigate inexact subgradient methods subject to persistent additive errors. Within a nonconvex semialgebraic framework, assuming boundedness or coercivity, we establish that the method yields iterates that eventually fluctuate near the critical set at a proximity characterized by an $O(\epsilon^\rho)$ distance, where ϵ denotes the magnitude of subgradient evaluation errors, and ρ encapsulates geometric characteristics of the underlying problem. Our analysis comprehensively addresses both vanishing and constant step-size regimes. Notably, the latter regime inherently enlarges the fluctuation region, yet this enlargement remains on the order of ϵ^ρ. In the convex scenario, employing a universal error bound applicable to coercive semialgebraic functions, we derive novel complexity results concerning averaged iterates. Additionally, our study produces auxiliary results of independent interest, including descent-type lemmas for nonsmooth nonconvex functions and an invariance principle governing the behavior of algorithmic sequences under small-step limits.

Keywords. Inexact subgradient, Clarke subdifferential, Nonsmooth nonconvex optimization, Path differentiable functions, First-order methods, Semialgebraic functions

AMS subject classifications: 68Q25, 49J52, 49J53, 90C70

1 Introduction

Let $f\colon \mathbb{R}^p \to \mathbb{R}$ be locally Lipschitz and consider the unconstrained global minimization problem

$$\min_{x\in\mathbb{R}^p} f(x).$$

*Toulouse School of Economics, Université Toulouse Capitole, Toulouse, France.
†LPSM, Université Paris Cité.
‡CMAP, Ecole polytechnique, Paris and Mohamed Bin Zayed University of AI
§Toulouse School of Economics, Université Toulouse Capitole et IUF, Toulouse, France.

An important tool for addressing such problems in high-dimensional nonsmooth settings, are subgradient algorithms, see e.g.[63, 64, 27, 57]. We focus here on the *inexact* or *biased subgradient method*, see [64]:

$$x_{k+1} \in x_k - \alpha_k \left[\partial^c f(x_k) + \bar{B}(0,\epsilon)\right],\ x_0 \in \mathbb{R}^n,$$

where for all $k \in \mathbb{N}$, $\alpha_k > 0$ is a sequence of positive step sizes, $\epsilon > 0$ is an error or a bias level and ∂^c denotes the Clarke subdifferential [23].

First, let us provide some motivations and insights into this method, particularly on its most distinctive features: inexact oracle evaluation and nonsmoothness.

There are many sources of error in the evaluation of subgradients. They stem from numerical errors or gradient approximation techniques, such as complex derivation oracles [39, 52, 14, 21] or low-accuracy calculations often used in neural network training [28]. The recent decade has seen a rapid proliferation of approximation methods, driven by the escalating dimensionality and structural complexity of optimization problems, such as large sums, composite functions, intrinsically complex neural network layers, differentiable programming, and federated learning frameworks. These developments have led to what can be broadly described as *sketchy calculations*. Our approach leverages sketch calculus methods to achieve computational efficiency, reduce memory requirements, and effectively manage heterogeneous data. Examples of sketching techniques include mini-batching [46], sparsification [68], sophisticated compression methods [53], and incremental computation strategies [64, 54]. Federated and distributed learning frameworks, discussed in [53], further exemplify scenarios involving inherently inexact subgradient computations. While stochastic optimization also extensively utilizes noisy gradient approximations due to random subsampling or Monte Carlo techniques [34, 55, 65, 37, 38, 45], we explicitly note that stochastic approximation techniques fall beyond the scope of this paper and warrant dedicated investigation.

Nonsmooth (and inexact) subgradient methods have deep roots in optimization with the many works of Shor [63], Ermoliev [34], Norkin [58], Nemirovskii [56] and also the pioneering work of Solodov-Zavriev [64]. Nonsmoothness is indeed prevalent for both convex and nonconvex problems. In the convex world, robustness questions naturally provide nonsmooth max problems [7], regularization techniques resort to nonsmooth regularizers [30, 67, 25, 6], while bundle methods rely on complex polyhedral oracles, see e.g. [52]. Modern machine learning provides a wealth of applications involving nonsmoothness. While this aspect has always been witnessed in deep learning through standard building blocks such as ReLU activations and MaxPooling, it appears in more recent contexts. For instance, sorting procedures may be used to promote sparsity [35], while optimization layers [2] may be used to refine learning abilities.

These modern situations make the analysis of gradient methods more difficult. In addition to approximation errors and nonsmoothness, let us mention the use non-vanishing stepsizes schedules as for instance in machine learning contexts [59, 49, 24]. Studying non-vanishing steps is more challenging since they no longer mitigate oscillations and errors induced by nonsmoothness.

The main contributions of this article, emphasizing the most novel aspects, can be summarized as follows:

— We develop an analytical framework to explicitly handle error terms, overcoming limitations of classical Lyapunov-based arguments typically used for subgradient analysis.

— We rigorously demonstrate that *all iterates eventually fluctuate around the critical set*, with fluctuations confined within an $O(\epsilon^\rho)$ bound, directly related to the local geometric structure of the optimization landscape.

— We introduce several intermediate results that provide valuable insights for vanishing and constant step-size regimes. In particular, Lemma 4.5 constitutes a general, independently significant result applicable to the ODE method, characterizing the limit points of sequences generated by small constant step-size recursions.

— Our analysis relies on weak and easily verifiable regularity assumptions on the cost function f. Specifically, our results hold broadly for local Lipschitz semialgebraic functions (or more generally, functions definable in an o-minimal structure), as detailed in Remark 2.2 and Proposition 3.1.

— In the convex setting, we avoid restrictive assumptions such as compactness or strong error bounds. Instead, we simply require convexity, coercivity, and semialgebraicity.

Our work follows a substantial body of research dedicated to nonsmooth and inexact (sub)gradient methods. We provide below a concise, though not exhaustive, overview of this literature.

In the convex setting, there has been extensive research dating back to early studies within the Russian-Soviet school, see e.g., [54, 41] and references therein. The inexact subgradient algorithm for Clarke regular nonconvex objective functions was first studied in the pioneering work of [64]. More recently, the analysis of stochastic gradient methods with biased errors has known several advances [66, 31, 60, 57]. These studies either focus on smooth objectives or rely on strong assumptions such as sharpness or metric regularity. [31] provides results qualitatively similar to ours, but exclusively for continuously differentiable targets; see also the recent review [29] for a broader perspective. For deterministic set-valued recursions, the seminal work [10] provides fundamental qualitative insights.

Our convergence analysis is based on the comparison of discrete-time sequences with the trajectories of continuous-time dynamical systems. This is a classical framework initiated by [48] and studied by [11, 44, 8], among others. This approach is particularly flexible, allowing to analyze stochastic approximation algorithms with constant or decreasing step sizes. Notable studies in the smooth case are [8, 36], [44, Chapter 8] and [22]. This framework was extended to situations where continuous-time dynamics are not determined by an ordinary differential equation but by upper semicontinuous differential inclusions, hence capturing nonsmooth algorithms. General results were first established in [9]; see also [10] for constant steps. The convergence of stochastic subgradient methods was then shown in [51] for Clarke regular functions, and in [27] for the broad class of definable functions.

Closely related to our work, [42] analyzes the global stability of subgradient methods with constant stepsizes applied to locally Lipschitz, coercive, and definable functions.

It is proved that iterates of methods approximated by subgradient trajectories stabilize around critical points. Results specifically cover the subgradient method with momentum, stochastic subgradient methods with random reshuffling and momentum, and cyclic coordinate descent methods with random permutations. The key difficulty in extending this latter result to include subgradient errors lies in characterizing iterates behavior near critical levels, where standard descent mechanisms fail. We explain this in more detail in the reading keys of Section 2.3.

Other studies worth mentioning, illustrating the scope of the dynamical system approach include [43] which establishes a discrete notion of Lyapunov stability for local minima of a locally Lipschitz and semialgebraic function, and [69] dealing with the boundedness of the iterates in the stochastic setting. A parallel line of works also consider the stochastic approximation sequence as a Markovian process in order to study the limit of invariant measures as the steps goes to zero [44, Section 8.4.3]. This strategy proves particularly useful when studying constant step stochastic and nonsmooth algorithms [61, 12, 13]. This literature is however quite distinct from our work and [42] in terms of convergence notions and proof techniques.

2 Preliminaries and statement of the main results

2.1 Notations.

For $f\colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz, we denote its Clarke subdifferential by $\partial^c f$, and, for $\epsilon > 0$, we define

$$\begin{aligned} \operatorname{crit}_\epsilon f &= \{x \ : \ \operatorname{dist}(0, \partial^c f(x)) \leq \epsilon\}, \\ \operatorname{vcrit}_\epsilon f &= f(\operatorname{crit}_\epsilon f), \end{aligned} \tag{1}$$

the ϵ-critical set, which is closed and the set of *ϵ-critical values*. When $l \notin \operatorname{vcrit}_\epsilon f$, it is called an *$\epsilon$-regular value*. We also denote by $\operatorname{crit} f$ the critical set and $\operatorname{vcrit} f$ the set of critical values. The set of minimizers is $\operatorname{argmin} f$.

For notational convenience, we write the algorithm as

$$x_{k+1} = x_k - \alpha_k v_\epsilon(x_k), \tag{2}$$

where $v_\epsilon : \mathbb{R}^p \to \mathbb{R}$ represent the biased oracle, i.e. satisfies $\operatorname{dist}(v_\epsilon(x), \partial^c f(x)) \leq \epsilon$ for all $x \in \mathbb{R}^p$.

Throughout the next sections, we use the shorthand $[a \leq g \leq b]$, for real numbers a, b and function $g\colon \mathbb{R}^p \to \mathbb{R}$, to denote $\{x \in \mathbb{R}^p \ : \ a \leq g(x) \leq b\}$. We also use the same notations for $=$ and $<$, $>$.

We denote the Euclidean norm by $\|\cdot\|$. For $A \subset \mathbb{R}^p$, we let $\|A\| = \sup_{a \in A} \|a\|$. For a subset, $A \subset \mathbb{R}^p$, $\overline{A}$ is its closure, and A^c its complement. For $\epsilon > 0$, $\bar{B}(0, \epsilon)$ is the closed ball of center 0 and radius ϵ. We denote the distance of a point x to a compact set $A \subset \mathbb{R}^p$

by $\operatorname{dist}(x, A) := \inf_{z \in A} \|x - z\|$. Given a set-valued map $Z : \mathbb{R}^p \rightrightarrows \mathbb{R}^q$, we denote the graph of Z by

$$\operatorname{graph}[Z] := \{(x, y) \in \mathbb{R}^p \times \mathbb{R}^q \ : \ y \in Z(x)\}.$$

We call a function f *coercive* when for all $a \in \mathbb{R}$, the sublevel sets $[f \leq a]$ are compact, this is equivalent to $f(x) \to \infty$ when $\|x\| \to \infty$.

2.2 Main results

The nonconvex setting We will work under the following assumption.

Assumption 1. *The function $f \colon \mathbb{R}^p \to \mathbb{R}$ is L-Lipschitz, lower-bounded, semialgebraic with $\operatorname{crit}_\epsilon f$ bounded for some $\epsilon > 0$.*

As stated in Lemma 3.4, Assumption 1 ensures that f is also coercive. Our first main result relates to vanishing step sizes.

Theorem 2.1 (Convergence for biased subgradient method with vanishing step size)**.** *Under Assumption 1, there is $\bar{\epsilon} > 0$, $C > 0$ $\rho > 0$ such that for any $\epsilon < \bar{\epsilon}$, $x_0 \in \mathbb{R}^p$, there is $\bar{\alpha} > 0$, such that for any $(x_k)_{k \in \mathbb{N}}$ given by (2) with $0 < \alpha_k \leq \bar{\alpha}$ for all $k \in \mathbb{N}$, $\alpha_k \to 0$ and $\sum_{k=0}^{\infty} \alpha_k = \infty$, we have*

$$\lim_{k \to \infty} \operatorname{dist}(f(x_k), \operatorname{vcrit}_\epsilon f) = 0,$$
$$\limsup_{k \to \infty} \operatorname{dist}(x_k, \operatorname{crit} f) \leq C\epsilon^\rho.$$

Our second main result relates to small constant step sizes.

Theorem 2.2 (Convergence for biased subgradient method with constant step size)**.** *Under Assumption 1, there is $\bar{\epsilon} > 0$, $C > 0$ $\rho > 0$ such that for any $\epsilon < \bar{\epsilon}$, $x_0 \in \mathbb{R}^p$, and $(x_k(\alpha))_{k \in \mathbb{N}}$ given by (2) with $\alpha_k = \alpha > 0$ for all $k \in \mathbb{N}$, we have*

$$\lim_{\alpha \to 0^+} \limsup_{k \to \infty} \operatorname{dist}(f(x_k(\alpha)), \operatorname{vcrit}_\epsilon f) = 0,$$
$$\limsup_{\alpha \to 0^+} \limsup_{k \to \infty} \operatorname{dist}(x_k(\alpha), \operatorname{crit} f) \leq C\epsilon^\rho.$$

Both these results are asymptotic regarding the iteration counter k. Obtaining non asymptotic convergence guarantees is open, even for the plain subgradient method ($\epsilon = 0$).

Remark 2.1 (Fluctuations with constant step sizes)**.** *In the previous theorem the limit with respect to α, can be interpreted as follows: for any $c > 0$, there exists $\bar{\alpha}$, such that for all $0 < \alpha \leq \bar{\alpha}$, the quantity $\operatorname{dist}(x_k(\alpha), \operatorname{crit} f)$ is of order $(C + c)\epsilon^\rho$ for sufficiently large $k \in \mathbb{N}$, where C and ρ are given in the theorems. In other words, the overall order of magnitude is ϵ^ρ with a constant depending on the step size, for small step sizes. Similar results were obtained by [42] when $\epsilon = 0$.*

Remark 2.2 (Local Lipschitz continuity)**.** *As it is classical in optimization, the global Lipschitz continuity of f in Assumption 1 can be relaxed to mere local Lipschitz continuity provided that there is a mechanism ensuring that the generated sequence remains in a compact set. This is the case for our biased subgradient method. Indeed, considering f as in Assumption 1, but only locally Lipschitz, the sequence remains bounded for small enough step sizes thanks to Lemma 4.3 since the objective is coercive as stated in Lemma 3.4 and* $\operatorname{vcrit}_\epsilon f$ *is upper bounded. Hence Theorem 2.1 and Theorem 2.2 actually holds if f in Assumption 1 is only locally Lipschitz rather than globally Lipschitz. We stick to global Lipschitz continuity of the objective f for simplicity of the presentation.*

The convex setting We complete these two results with an explicit estimate for the convex case. Our result is in the line of usual results, see, e.g., [54, 41], but does not use any compactness or strong error bounds.

Let us recall [18, Theorem 3]: for a coercive, convex, semialgebraic function $f\colon \mathbb{R}^p \to \mathbb{R}$, there exists $c > 0$ and $a \in (0, 1]$, such that

$$\frac{c}{2}\left((f(x) - \min f)^a + (f(x) - \min f)\right) \geq \operatorname{dist}(x, \operatorname{argmin} f) \quad \forall x \in \mathbb{R}^p. \tag{3}$$

Theorem 2.3 (Biased subgradient complexity: convex case)**.** *Let $f\colon \mathbb{R}^p \to \mathbb{R}$ be L-Lipschitz, semialgebraic, convex and coercive. Then with $a \in (0, 1]$ and $c > 0$, as in* (3)*, for any sequence generated by* (2)*, we have*

$$(2 - a - \epsilon c)\frac{\sum_{i=0}^k \alpha_i (f(x_i) - f^*)}{\sum_{i=0}^k \alpha_i} \leq (1-a)(\epsilon c)^{\frac{1}{1-a}} + \frac{\|x_0 - x_*\|^2 + (L+\epsilon)^2 \sum_{i=0}^k \alpha_i^2}{\sum_{i=0}^k \alpha_i}.$$

We understand for $a = 1$, the value $(1-a)(\epsilon c)^{\frac{1}{1-a}}$ as 0 if $\epsilon c < 1$ and undefined or infinite otherwise.

This allows to get convergence rates in value with various choices of step size. For example, if $\epsilon c < 1$, choosing $\alpha_i = \frac{1}{\sqrt{k+1}}$ for $i = 0, \ldots, k$, leads to

$$\min_{i=1,\ldots,k} f(x_k) - f^* \leq \frac{(1-a)(\epsilon c)^{\frac{1}{1-a}}}{2 - a - \epsilon c} + \frac{\|x_0 - x_*\|^2 + (L+\epsilon)^2}{(2 - a - \epsilon c)\sqrt{k+1}}.$$

Note that the bound is vacuous if for example $a = 1$ and $\epsilon \geq 1/c$ and provides effective guarantees only for small values of ϵ. This is somewhat unavoidable: if f is the absolute value and $\epsilon > 1$, one completely looses control of the resulting biased subgradient sequence. This is the so called "low error" setting.

2.3 Reading keys and natural extensions

For the two first theorems, we examine the continuous-time limit of the recursion (2), which leads to a differential inclusion given in (4). A special Lyapunov mechanism is

introduced in Section 3: Along the flow, the objective decreases only in certain regions of the space and may increase in other regions. This mechanism is used to describe the asymptotics of the system and provides explicit estimates in the Lipschitz coercive setting, under a Kurdyka-Lojasiewicz assumption for f and a metric regularity assumption for $\partial^c f$. This leads to an explicit estimate of the distance to crit f for any positively invariant set A such that $f(A) \subset \operatorname{vcrit}_\epsilon f$. Recall that A is positively invariant for the differential inclusion (4) if for any $x \in A$ there is a solution to (4) starting at x whose forward orbit is contained in A. The proof of Theorem 2.1 and Theorem 2.2 then consists of justifying the fact that relevant asymptotic sets fulfill this invariance property, which is done in Section 4. The main device is to use the small step limiting relation between the recursion (2) and its continuous-time counterpart for which the analysis was made in Section 3.

Our arguments rely in part on the repulsivity of regular values, the complement of $\operatorname{vcrit}_\epsilon f$, as described in Section 4.2, based on the continuous time analysis in Lemma 3.1. A repulsivity mechanism of the same nature was instrumental in the proof of [42, Theorem 1, Corollary 1], in a setting very close to ours, without errors. It turns out that, in the presence of errors, Lemma 3.1 does not provide a classical Lyapunov function, which should be decreasing, non-necessarily strictly, along all trajectories. Therefore, the presence of errors requires dedicated arguments and the repulsivity of regular values is not sufficient to conclude about Theorem 2.2 and Theorem 2.1.

Theorem 2.3 is obtained by using standard Lyapunov functions in convex optimization and by applying an error bound condition that is automatically satisfied in the semialgebraic case. These arguments are given in Section 5 and are completely independent of the arguments for Theorem 2.1 and Theorem 2.2, which make up the main technical part of this work.

These results could be readily extended in various ways.

— It can be seen directly that the same results apply beyond semialgebraicity to any polynomially bounded o-minimal structure, the prototypical example being the structure of globally subanalytic sets, see [32] for an overview.

— It is also directly possible to obtain qualitatively similar results, without the convergence rate estimates, in any o-minimal structure, replacing the exponent estimates by definable functions [32].

— While we consider the Clarke subdifferential, the main device used in the proof is the chain rule along Lipschitz curves, which is satisfied for the broader class of conservative gradient fields [19]. Examples are automatic differentiation oracles and generalized derivatives as in [33]. It is known that conservative gradients may induce spurious stationary points. The result in [17, Theorem 4.12] ensures that up to removing a zero measure meagre set of possible initialization $x_0 \in \mathbb{R}^p$ and a finite set of step sizes $\alpha \in \mathbb{R}$, one does actually manipulate the subdifferential oracle.

— We only consider the deterministic setting, it is possible to extend these results to the stochastic approximation setting, by adding zero mean stochastic perturbation terms which satisfy summability hypotheses for vanishing step sizes [9] or by resorting to

the study of convergence of stationary measures in the constant step size setting [61].

3 The continuous-time system and auxiliary results

Let $f\colon \mathbb{R}^p \to \mathbb{R}$ be locally Lipschitz. Consider for $\epsilon > 0$, the differential inclusion

$$\dot{x}(t) \in -\partial^c f(x(t)) + \bar{B}(0, \epsilon), \tag{4}$$

for almost every $t \in [0, +\infty)$, where the solution is to be found among locally Lipschitz curves. Existence of solutions on maximal intervals follows from classical results, see [5, 3]. In particular, if f is Lipschitz, we may consider solutions defined on $\mathbb{R}_+$. Equivalently, x is solution to (4) if for almost every t, $\operatorname{dist}(\dot{x}(t), -\partial^c f(x(t))) \leq \epsilon$.

3.1 Asymptotics of the biased dynamics

Following Valadier and [19], a locally Lipschitz function f is called path-differentiable if for any Lipschitz curve $\gamma\colon \mathbb{R} \to \mathbb{R}^p$, we have for almost every $t \in \mathbb{R}$,

$$\frac{d}{dt}(f \circ \gamma)(t) = \langle v, \dot{\gamma}(t) \rangle\,, \qquad \forall v \in \partial^c f(\gamma(t)).$$

Semialgebraic functions are path-differentiable [27, 19]. Under path-differentiability, we obtain the following results.

Lemma 3.1 (Descent properties)**.** *Let f be locally Lipschitz and path-differentiable. Let $\epsilon > 0$ and $x : \mathbb{R}_+ \to \mathbb{R}^p$ be a solution to the differential inclusion* (4) *assumed to be defined on $\mathbb{R}_+$ (see remark below). Then,*

1. (Weak Lyapunov) *There is a measurable selection v_x of $\partial^c f(x(\cdot))$ such that,*

$$f(x(t_2)) - f(x(t_1)) \leq -\int_{t_1}^{t_2} \|v_x(t)\|(\|v_x(t)\| - \epsilon)dt. \qquad \forall 0 \leq t_1 \leq t_2.$$

   In particular, if $f(x(0)) \notin \operatorname{vcrit}_\epsilon f$, there is $t > 0$ such that $f(x(s)) < f(x(0))$ for all $0 \leq s \leq t$.

2. (Decrease) *Let $l \notin \operatorname{cl} \operatorname{vcrit}_\epsilon f$ be such that $f(x(0)) \leq l$ then for all $t > 0$, $f(x(t)) < l$.*

3. (Asymptotics) *Either $\|x(t)\| \to +\infty$ as $t \to +\infty$, or we have both*

$$\liminf_{t\to\infty} \operatorname{dist}(0, \partial^c f(x(t))) \leq \epsilon$$
$$\lim_{t\to\infty} \operatorname{dist}(f(x(t)), \operatorname{vcrit}_\epsilon f) = 0.$$

4. (Quantitative estimates) *For any $0 \leq a < b$ and $\delta > \epsilon$, set $T = \frac{b-a}{\delta(\delta-\epsilon)}$ and assume that $f(x([0,T])) \subset [a,b]$, then there is $t \in [0,T]$ such that $\operatorname{dist}(0, \partial^c f(x(t))) \leq \delta$.*

5. (ϵ-stationarity) *For any* $a \in \mathbb{R}$ *and* $T > 0$, *if* $f(x([0,T])) = \{a\}$ *then* $x([0,T]) \subset \operatorname{crit}_\epsilon f$.

Remark 3.1. Actually Lemma 3.1 points 1, 2, 4, 5 hold on the interval of definition of the solution which does not need to be $\mathbb{R}_+$. In our proofs, we will apply Lemma 3.1 to solutions defined on an interval $[0,T] \subset \mathbb{R}_+$.

Proof :

1. Consider a measurable selection v_x such that for almost every $t \geq 0$,

$$v_x(t) \in \operatorname{argmin}_{v \in \partial^c f(x(t))} \|\dot{x}(t) + v\|.$$

Such a selection exists, see e.g. [1, Theorem 18.13]. Since $\dot{x}(t) \in -\partial^c f(x(t)) + \bar{B}(0,\epsilon)$, then $\|\dot{x}(t) + v_x(t)\| \leq \epsilon$ for almost every $t \geq 0$. Then, by path-differentiability of f, we have for almost every $t \geq 0$,

$$\begin{aligned}
\frac{d}{dt}(f \circ x)(t) &= \langle v_x(t), \dot{x}(t) \rangle \\
&= \langle v_x(t), -v_x(t) + v_x(t) + \dot{x}(t) \rangle \\
&= -\|v_x(t)\|^2 + \langle v_x(t), v_x(t) + \dot{x}(t) \rangle \\
&\leq -\|v_x(t)\|(\|v_x(t)\| - \|\dot{x}(t) + v_x(t)\|) \\
&\leq -\|v_x(t)\|(\|v_x(t)\| - \epsilon). \qquad (5)
\end{aligned}$$

Integrating from t_1 to t_2 gives the desired inequality. If $f(x(0)) \notin \operatorname{vcrit}_\epsilon f$, then $x(0) \notin \operatorname{crit}_\epsilon f$, and since $\operatorname{crit}_\epsilon f$ is closed, there is a compact set U with $x(0) \in \operatorname{int} U$ such that $U \cap \operatorname{crit}_\epsilon f = \emptyset$.We may therefore choose $t > 0$ small enough such that $x(s) \in U$ for all $s \in [0,t]$ and the result follows.

2. We distinguish two cases.

Case 1. Assume that $f(x(0)) \notin \operatorname{cl} \operatorname{vcrit}_\epsilon f$. Let us show that $f(x(t)) < f(x(0))$ for all $t > 0$.

Since $(\operatorname{cl} \operatorname{vcrit}_\epsilon f)^c$ is open and f is locally Lipschitz, there exists $t^* > 0$ small enough such that $f(x(s)) \notin \operatorname{cl} \operatorname{vcrit}_\epsilon f$ for all $s \in [0,t^*]$. In this case, 1 gives us $f(x(s)) \leq f(x(0)) - \int_0^{t^*} \|v_x(t)\|(\|v_x(t)\| - \epsilon\|)dt < f(x(0))$ for all $s \in (0,t^*]$.

Now we show that for all $t \geq t^*$, $f(x(t)) \leq f(x(t^*))$. Assume toward a contradiction that there exists $t \geq t^*$ such that $f(x(t^*)) < f(x(t))$. Since $f(x(t^*)) \notin \operatorname{cl} \operatorname{vcrit}_\epsilon f$, we may chose t so that $[f(x(t^*)), f(x(t))] \subset (\operatorname{cl} \operatorname{vcrit}_\epsilon f)^c$ by openness of the latter. Now, consider $t^- = \max\{s \ : \ s \in [t^*,t], f(x(s)) \leq f(x(t^*))\}$, and $t^+ = \min\{s \ : \ s \in [t^-,t], f(x(s)) \geq f(x(t))\}$. By continuity of $f \circ x$, t^- and t^+ are well defined with $t^+ \geq t^-$, and they satisfy for all $s \in [t^-,t^+]$, $f(x(s)) \in [f(x(t^-)), f(x(t^+))] \subset (\operatorname{cl} \operatorname{vcrit}_\epsilon f)^c \subset (\operatorname{vcrit}_\epsilon f)^c$, as well as $f(x(t^-)) = f(x(t^*))$ and $f(x(t^+)) = f(x(t))$. In particular, $f(x(t^+)) > f(x(t^-))$ hence $t^+ > t^-$. Let v_x be given by item 1. Then we have

$$f(x(t^+)) - f(x(t^-)) \leq -\int_{t^-}^{t^+} \|v_x(s)\|(\|v_x(s)\| - \epsilon)ds < 0,$$

where the last inequality comes from $t^+ > t^-$ and $\|v_x(s)\| > \epsilon$ for almost every $s \in [t^-, t^+]$ since $f(x(s)) \notin \operatorname{vcrit}_\epsilon f$. This yields a contradiction. We have shown that for all $t \geq t^*$, $f(x(t)) \leq f(x(t^*))$.

Finally, for $t \in (0, t^*]$, $f(x(t)) < f(x(0))$, and for $t > t^*$, $f(x(t)) \leq f(x(t^*)) < f(x(0))$ hence the desired result under the assumption that $f(x(0)) \notin \operatorname{cl} \operatorname{vcrit}_\epsilon f$.

Case 2. Now, assume that $f(x(0)) \in \operatorname{cl} \operatorname{vcrit}_\epsilon f$ and $f(x_0) \leq l$ for $l \notin \operatorname{cl} \operatorname{vcrit}_\epsilon f$. In this case we actually have $f(x(0)) < l$. Since $(\operatorname{cl} \operatorname{vcrit}_\epsilon f)^c$ is open, there is $l' \notin \operatorname{cl} \operatorname{vcrit}_\epsilon f$ such that $f(x(0)) < l' < l$. Then by continuity of $f \circ x$, either $f(x(t)) < l$ for all t or there exists $t > 0$ such that $f(x(t)) = l'$ and t can be chosen to be minimal since $[f \circ x = l']$ is closed and lower bounded. We have $f(x(s)) \leq l'$ for all $s \leq t$. Then by Case 1 shown previously, we have for all $s \geq t$, $f(x(s)) \leq l'$. Since $l' < l$, we have the desired result.

4. We now prove the fourth item, which does not depend on item 3 but is used to prove item 3. We have $a < b$. Assume toward a contradiction that $\operatorname{dist}(0, \partial^c f(x(t))) > \delta$ for all $t \in [0, T]$. Since $s \to \operatorname{dist}(0, \partial^c f(x(s)))$ is lower semi-continuous, there exists $\delta' > \delta$ such that $\operatorname{dist}(0, \partial^c f(x(t))) \geq \delta'$ for all $t \in [0, T]$. In this case, item 1 gives us $f(x(T)) \leq f(x(0)) - T\delta'(\delta' - \epsilon)$, hence

$$f(x(T)) \leq b - (b-a)\frac{\delta'(\delta' - \epsilon)}{\delta(\delta - \epsilon)} < b - (b-a) = a$$

This is a contradiction as we assumed that $f(x(T)) \geq a$.

3. Assume that $\|x(t)\|$ does not go to $+\infty$, this means that the trajectory has accumulation points. So $f(x(t))$ also has finite accumulation values and in particular $f(x(t))$ does not diverge to $-\infty$ or $+\infty$.

— Using item 1, we have that $\liminf_{t\to\infty} \operatorname{dist}(0, \partial^c f(x(t))) > \epsilon$ implies that $f(x(t)) \to -\infty$ as $t \to \infty$ and we obtain the first limit.

— Denote by I all the accumulation points of $f(x(t))$ as $t \to \infty$. I is a nonempty interval by continuity of $f \circ x$ and by the fact that $f \circ x$ does not go to $+\infty$ or to $-\infty$. We distinguish two cases.

First if I has empty interior, then $I = \{l\}$ and $f(x(t)) \to l$ for some $l \in \mathbb{R}$. Since the trajectory x has accumulation points, we may find a sequence $(t_k)_{k\in\mathbb{N}}$ in $\mathbb{R}_+$ such that $t_k \to \infty$ and $x(t_k) \to \bar{x} \in \mathbb{R}^p$ as $k \to \infty$, and we have $f(\bar{x}) = l$. By item 4, for any $\delta > \epsilon$, set $a = l - \delta(\delta - \epsilon)$, $b = l + \delta(\delta - \epsilon)$ we have $f(x(t_k + \mathbb{R}_+)) \in [a, b]$ for k sufficiently large and therefore, there exists $s_k \in [t_k, t_k + 2]$ such that $\operatorname{dist}(0, \partial^c f(x(s_k))) \leq \delta$. This can be repeated for smaller δ and we may find a sequence, $(s_k)_{k\in\mathbb{N}}$, such that for each k, $s_k \in [t_i, t_i + 2]$ for some $i \in \mathbb{N}$ and $\operatorname{dist}(0, \partial^c f(x(s_k)) \to \epsilon$. The restricted trajectory $\{x([t_i, t_i + 2])\}_{i\in\mathbb{N}}$ remains in a bounded set by local Lischicity of f, hence local boundedness of its subdifferential, so up to a subsequence, we may assume that $x(s_k)$ converges to a point $\tilde{x} \in \operatorname{crit}_\epsilon f$ with $f(\tilde{x}) = l$. This shows that $l \in \operatorname{vcrit}_\epsilon f$.

Secondly, let us assume that I has nonempty interior. Assume toward a contradiction that there exists $l \in \operatorname{int} I \cap (\operatorname{cl} \operatorname{vcrit}_\epsilon f)^c$. In this case, there is $u > 0$ such that $[l-u, l+u] \subset \operatorname{int} I \cap (\operatorname{cl} \operatorname{vcrit}_\epsilon f)^c$ and t such that $f(x(t)) \in [l-u, l]$. By item 2, we have $f(x(s)) < l$ for all $s > t$, but this is contradictory with the fact that $l+u \in \operatorname{int} I$, because this implies that $t \mapsto f(x(t))$ has accumulation values strictly greater than $l+u$. So I is an interval such that

$\emptyset \neq \operatorname{int} I \subset \operatorname{cl} \operatorname{vcrit}_\epsilon f$ hence $I \subset \operatorname{cl} \operatorname{vcrit}_\epsilon f$. This means that $\sup_{v \in I} \operatorname{dist}(v, \operatorname{vcrit}_\epsilon f) = 0$, which is the desired result.

5. If $T > 0$ and $f \circ x$ is constant on $[0, T]$, by item 1, we necessarily have $\operatorname{dist}(0, \partial^c f(x(t))) \leq \epsilon$ for almost every $t \in (0, T)$, and $x([0, T]) \in \operatorname{crit}_\epsilon f$ because $\operatorname{crit}_\epsilon f$ is closed and x is continuous.

□

3.2 Estimates under the nonsmooth KL inequality and a metric subregularity condition

We will obtain more precise estimates under the following assumption.

Assumption 2. f is L-Lipschitz, path-differentiable, the set $\operatorname{vcrit} f$ is nonempty finite, and there exists $\bar{\epsilon} \in (0, 1)$ such that for any $0 \leq \epsilon \leq \bar{\epsilon}$, $\operatorname{vcrit}_\epsilon f$ is a finite union of segments.

Furthermore, there exists $c > 0$ and $\theta \in [0, 1)$, and $\beta > 0$ such that for all $x \in f^{-1}(\operatorname{vcrit}_{\bar{\epsilon}} f)$

$$\operatorname{dist}(f(x), \operatorname{vcrit} f)^\theta \leq c \operatorname{dist}(0, \partial^c f(x)) \tag{KL}$$

$$\operatorname{dist}(x, \operatorname{crit} f) \leq c \operatorname{dist}(0, \partial^c f(x))^\beta. \tag{MR}$$

Property (KL) is some form of the nonsmooth Kurdyka-Lojasiewicz inequality [16], while (MR) is a form of metric sub-regularity of the subdifferential around the critical set, see [47, Proposition 3.1] for a general result, but also [40, 4, 62, 50] for concrete applications in optimization.

In our context, Assumption 2 is essential to control trajectories of (4). Some previous works on biased algorithms relied as well on similar conditions. For instance, KL inequality for real analytic functions was used in [31] and metric regularity ($\beta = 1$) appears in [57]. Assumption 2 is actually satisfied for all semialgebraic functions:

Lemma 3.2 (Semialgebraicity implies regularity)**.** *If Assumption 1 is satisfied, then Assumption 2 is also satisfied for some $\bar{\epsilon} > 0$.*

Proof : According to Sard's theorem for semialgebraic functions [16], $\operatorname{vcrit} f$ is finite. Assumption 1 ensures that there exists $\bar{\epsilon}$ such that $\operatorname{crit}_\epsilon f$ is compact for every $0 \leq \epsilon \leq \bar{\epsilon}$. For $0 \leq \epsilon \leq \bar{\epsilon}$ the sets $\operatorname{vcrit}_\epsilon f \subset \mathbb{R}$ are then compact and semialgebraic, i.e. they consist of a finite number of segments. Furthermore, by Lemma 3.4 (see next subsection), f is coercive so $f^{-1}(\operatorname{vcrit}_{\bar{\epsilon}} f)$ is compact. (KL) follows from Kurdyka-Lojasiewicz inequality for nonsmooth semialgebraic functions [15] and compactness. As for (MR), this is Hölder metric subregularity as given in [47, Proposition 3.1] on the compact set $f^{-1}(\operatorname{vcrit}_{\bar{\epsilon}} f)$.

□

We will therefore prove Theorem 2.1 and Theorem 2.2 under Assumption 2.

Remark 3.2 (Beyond semialgebraicity). Semialgebraicity in Assumption 1 can be replaced by global subanalyticity and all results would hold true in the exact same form. This allows to include the logarithm and exponential function restricted to compact segments for example. More generally, our results hold provided that f is definable in a polynomially bounded o-minimal structure [26], for which semialgebraic sets and globally subanalytic sets are the main examples. Our results could also be extended if in Assumption 1 we assume that f is definable in an o-minimal structure (not necessarily polynomially bounded) instead of being semialgebraic. Under this assumption, we would obtain the same results in Theorem 2.1 and Theorem 2.2, but the term of the form $C\epsilon^\rho$ would be replaced by a term of the form $e(\epsilon)$ for an abstract nonnegative increasing definable functions $e\colon \mathbb{R}^+ \to \mathbb{R}^+$ continuous at 0 with value 0.

The next lemma generalizes the following fact, "for a locally Lipschitz continuous semialgebraic f and a subgradient curve x ($\dot{x}(t) \in -\partial^c f(x(t))$ for all $t \in \mathbb{R}_+$), the inclusion $f(x(\mathbb{R}_+)) \subset \operatorname{vcrit} f$, implies $0 \in \partial^c f(x(t))$ for all $t \geq 0$". Indeed, $\operatorname{vcrit} f$ is made of a finite number of singletons, and if there is $t > 0$ such that $\operatorname{dist}(0, \partial^c f(x(t))) > 0$, this would hold locally and result in a strict decrease by path-differentiability. We would thus have a time $t' > 0$ such that $f(x(t')) \notin \operatorname{vcrit} f$. This fact, corresponding to the case when $\epsilon = 0$, can be extended to a general $\epsilon > 0$. We remind the reader that $f(x) \in \operatorname{vcrit}_\epsilon f$ does not imply that $x \in \operatorname{crit}_\epsilon f$ so that it is not sufficient to control the distance between $\operatorname{crit}_\epsilon f$ and $\operatorname{crit} f$.

Lemma 3.3 (Approximate stationarity of near-critical curves)**.** *Under Assumption 2, set* $\rho = \frac{\beta}{\max\{\theta(\beta+2),1\}}$. *There exists* $C > 0$, *such that for any* $\epsilon \in [0, \bar{\epsilon}]$ *and any solution curve* $x : \mathbb{R}_+ \to \mathbb{R}^p$ *such that* $f(x(\mathbb{R}_+)) \subset \operatorname{vcrit}_\epsilon f$, *we have for all* $t \geq 0$,

$$\operatorname{dist}(x(t), \operatorname{crit} f) \leq C\epsilon^\rho. \tag{6}$$

Proof : Fix an arbitrary $\epsilon \leq \bar{\epsilon}$, and $x\colon \mathbb{R}_+ \to \mathbb{R}^p$ an arbitrary solution. Since $f(x(\mathbb{R}_+))$ is connected, $f(x(\mathbb{R}_+))$ is contained in a single connected component of $\operatorname{vcrit}_\epsilon f$ of the form $[a, b] \subset \mathbb{R}$.

For all $z \in \operatorname{crit}_\epsilon f$, $\operatorname{dist}(0, \partial^c f(z)) \leq \epsilon$ implies $\operatorname{dist}(f(z), \operatorname{vcrit} f) \leq (c\epsilon)^{\frac{1}{\theta}}$ using (KL). Let $N \in \mathbb{N}$ such that $v_0 \leq v_1 \leq \cdots \leq v_{N+1}$ are the ordered critical values in $[a, b]$ to which we added $v_0 = a$ and $v_{N+1} = b$. This defines $N + 1$ segments which cover $[a, b]$, one of them has length at least $\frac{b-a}{N+1}$. Consequently, there is an open segment (u, v) such that $v - u = \frac{b-a}{N+1}$ which does not contain any critical value. Therefore, we may choose $z \in \operatorname{crit}_\epsilon f$ such that $f(z) = (u + v)/2$ and we have

$$\operatorname{dist}(f(z), \operatorname{vcrit} f) \geq \frac{v - u}{2} = \frac{b - a}{2(N + 1)}. \tag{7}$$

Combining (7) with (KL), we obtain

$$b - a \leq 2(N + 1)(c\epsilon)^{\frac{1}{\theta}} := K_1\epsilon^{\frac{1}{\theta}}. \tag{8}$$

We fix an arbitrary $0 < \alpha \leq 1$ and $t \geq 0$. We set $\delta = 2\epsilon^\alpha > \epsilon$, by point 4. of Lemma 3.1 and (8), there is $t' \geq t$, such that

$$\operatorname{dist}(0, \partial^c f(x(t'))) \leq 2\epsilon^\alpha \tag{9}$$

$$t' - t \leq \frac{b-a}{(2\epsilon^\alpha - \epsilon)2\epsilon^\alpha} \leq \frac{K_1 \epsilon^{\frac{1}{\theta}}}{(2\epsilon^\alpha - \epsilon)2\epsilon^\alpha} \leq \frac{K_1}{2}\epsilon^{\frac{1}{\theta}-2\alpha}. \tag{10}$$

It follows using the inclusion (4), the fact that f is L-Lipschitz so that its subgradient is bounded by L, and (10) that

$$\|x(t') - x(t)\| \leq \int_t^{t'} \|\dot{x}(s)\| ds \leq (t'-t)(L+\epsilon) = O\left(\epsilon^{\frac{1}{\theta}-2\alpha}\right) \tag{11}$$

Finally, using the estimates (11), (9) and (MR),

$$\operatorname{dist}(x(t), \operatorname{crit} f) \leq \|x(t) - x(t')\| + \operatorname{dist}(x(t'), \operatorname{crit} f) = O\left(\epsilon^{\frac{1}{\theta}-2\alpha}\right) + O\left(\epsilon^{\beta\alpha}\right) \tag{12}$$

Since $t \geq 0$ was arbitrary, the estimate (12) holds for all $t \geq 0$. Furthermore, since $0 < \alpha \leq 1$ was arbitrary, we may choose α freely. We distinguish two cases.

- If $\theta(\beta+2) > 1$, then we choose $\alpha = \frac{1}{\theta(2+\beta)} < 1$, and the right-hand side in (12) is of the form $O\big(\epsilon^{\frac{\beta}{\theta(2+\beta)}}\big)$.
- If $\theta(\beta+2) \leq 1$ then we may choose $\alpha = 1$. We have that $\epsilon^{\frac{1}{\theta}-2} \leq \epsilon^\beta$ hence the right-hand side in (12) is $O\left(\epsilon^\beta\right)$.

□

Remark 3.3 (On the initialization)**.** It may be puzzling not to see a condition on the initialization in Lemma 3.3. This is hidden in the condition $f(x(\mathbb{R}_+)) \subset \operatorname{vcrit}_\epsilon f$ which enforces x to start close enough to $\operatorname{crit} f$ so that $f(x(t))$ cannot leave $\operatorname{vcrit}_\epsilon f$ near $t = 0$.

Remark 3.4 (Power functions)**.** For a power function, $x \mapsto x^a$ on $\mathbb{R}_+$, for $a > 1$, we have $\theta = 1 - \frac{1}{a}$ and $\beta = \frac{1}{a-1}$. In this case, we have

$$\theta(\beta+2) = \frac{a-1}{a}\left(\frac{1}{a-1} + 2\right) = \frac{1}{a} + 2\frac{a-1}{a} = 2 - \frac{1}{a} \in (1, 2].$$

Hence we have $\rho > \frac{\beta}{2}$. This corresponds to the estimate obtained by [31] for analytic functions (if a is an integer). Indeed, in the notations of [31, Proposition 8.2] we have $\theta = 1$ and $\beta = r_Q$ and the resulting estimate of [31, Theorem 2.1 (iii)] corresponds to $\frac{\beta}{2}$ by combining estimates in [31, Proposition 8.2] and [31, Proposition 8.3]. In this univariate setting however, the correct estimate would be simply β and we leave the question of the optimality of our estimate in the nonsmooth multivariate case for future research.

Lemma 3.3 has the following direct consequence.

Corollary 3.1 (Invariant sets and biased dynamics)**.** *Under Assumption 2, let $\epsilon \leq \bar{\epsilon}$, $S \subset \mathbb{R}^p$ be a positively invariant set, that is for any $z \in S$, there is $x\colon \mathbb{R}_+ \to \mathbb{R}^p$, solution to (4) such that $x(\mathbb{R}_+) \subset S$ and $x(0) = z$. If $f(S) \subset \operatorname{vcrit}_\epsilon f$, then for any $z \in S$, $\operatorname{dist}(z, \operatorname{crit} f) \leq C\epsilon^\rho$, where C, ρ are given by Lemma 3.3.*

For the continuous time dynamics in (4), Lemma 3.1 point 3 ensures that accumulation points of bounded solutions to the differential inclusion (4) correspond to objective values in $\operatorname{vcrit}_\epsilon f$. Furthermore, it is known that the set of such accumulation points forms an invariant set (see for example [9, Theorem 3.6, Lemma 3.5]). We obtain the following.

Corollary 3.2 (Asymptotics of biased dynamics)**.** *Under Assumption 2, let $\epsilon \leq \bar{\epsilon}$, and $x\colon \mathbb{R}^+ \to \mathbb{R}$ be a bounded solution to the differential inclusion (4). For any $z \in \mathbb{R}^p$, accumulation point of the trajectory, we have $\operatorname{dist}(z, \operatorname{crit} f) \leq C\epsilon^\rho$, where C, ρ are given by Lemma 3.3.*

3.3 Coercivity and boundedness of ϵ critical points

Due to the importance of boundedness of curves in our results, we need to make a brief detour through coercivity properties and their link with the boundedness of $\operatorname{crit}_\epsilon f$. Let us recall first:

Theorem 3.1 (Ekeland's variational principle)**.** *Let X be a complete metric space with distance d. Let $f\colon X \mapsto \mathbb{R} \cup \{+\infty\}$ be lower semi-continuous, bounded below and finite at least at one point. Fix $\epsilon > 0$ and $x_0 \in X$ such that $f(x_0) \leq \epsilon + \inf_x f(x)$. Then for any $\lambda > 0$, there is $y_0 \in X$ such that*

$$f(y_0) \leq f(x_0) \qquad d(x_0, y_0) \leq \lambda \qquad f(x) + \frac{\epsilon}{\lambda} d(x, y_0) > f(y_0),\ \forall x \neq y_0.$$

The conclusion tells us that y_0 is a strict global minimizer of $g : x \to f(x) + \dfrac{\epsilon}{\lambda} d(x, y_0)$.

In our framework, $\mathbb{R}^p$ is endowed with the Euclidean distance and f is locally Lipschitz, so using the sum rule with the Clarke subdifferential yields the following fact:

$$0 \in \partial^c g(y_0) \subset \partial^c f(y_0) + \bar{B}\left(0, \frac{\epsilon}{\lambda}\right) \tag{13}$$

This has the following consequences:

Lemma 3.4 (Boundedness of the ϵ-critical set implies coercivity)**.** *Assume that $f\colon \mathbb{R}^p \to \mathbb{R}$ is locally Lipschitz, then for any $x \in \mathbb{R}^p$, $a \in (0, 1)$, there is $y \in \mathbb{R}^p$ such that*

$$\|y\| \geq \|x\| - \|x\|^a, \qquad \operatorname{dist}(0, \partial^c f(y))\|x\|^a \leq f(x) - \inf f.$$

In particular, if f is lower bounded and $\operatorname{crit}_{\bar{\epsilon}} f$ is bounded for some $\bar{\epsilon} > 0$, then f is coercive, that is $f(x) \to +\infty$ as $\|x\| \to +\infty$.

Proof : Fix $x \in \mathbb{R}^p$, if $f(x) = \inf f$ there is nothing to prove. Choose $\epsilon = f(x) - \inf f$ (assumed finite, otherwise there is nothing to prove) and $\lambda = \|x\|^a$. By Ekeland's variational principle in Theorem 3.1 there is $y \in \mathbb{R}^p$ such that

$$\|x - y\| \leq \|x\|^a,$$

hence $\|y\| \geq \|x\| - \|x\|^a$. Moreover using (13), i.e., $0 \in \partial^c f(y) + \bar{B}\left(0, \frac{\epsilon}{\|x\|^a}\right)$, we get $\operatorname{dist}(0, \partial^c f(y)) \leq \frac{\epsilon}{\|x\|^a}$.

If $\operatorname{crit}_{\bar{\epsilon}} f$ is bounded for some $\bar{\epsilon} > 0$, and f was not coercive, we could choose a sequence $\|x_k\| = k + 1$ so that $f(x_k) - \inf f$ is bounded by some M, and obtain an unbounded sequence y_k in $\operatorname{crit}_{\bar{\epsilon}} f$ for k large enough. □

Remark 3.5 (Coercivity and critical points)**.** (a) Of course crit f bounded and nonempty does not imply coercivity, e.g., take $s \mapsto (1 + s^2)^{-1}$.
(b) The fact that "$\operatorname{crit}_\epsilon f$ bounded (for some $\epsilon > 0$) implies f coercive" is a generalization of the classical result in convex analysis "argmin f bounded implies f coercive when f is convex".

In general, coercivity alone does not imply $\operatorname{crit}_{\bar{\epsilon}} f$ bounded, even for semialgebraic functions, take for instance $s \mapsto \sqrt{|s|}$ on $\mathbb{R} \setminus [-1, 1]$. The condition is however satisfied under sufficient growth, as showcased by the following result. Note the proposition below applies for example to objectives of the form of regularized risk $f(x) = \ell(x) + \frac{\lambda}{2}\|x\|^2$, where ℓ is bounded from below and semiagebraic, a widespread situation in machine learning.

Proposition 3.1. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a locally Lipschitz semialgebraic function. Assume there exists $\beta > 0$ such that $\liminf_{\|x\| \to \infty} \frac{f(x)}{\|x\|^\beta} > 0$. Then*

1. *If $\beta = 1$, there exists $\bar{\epsilon} > 0$ such that $\operatorname{crit}_{\bar{\epsilon}} f$ is bounded.*
2. *If $\beta > 1$, $\operatorname{crit}_\epsilon f$ is bounded for any $\epsilon > 0$.*

Proof : 1. We prove the case when $\beta = 1$. Assume toward a contradiction that for any $\epsilon > 0$, $\operatorname{crit}_\epsilon f$ is not bounded. Let $\tilde{\epsilon} : \mathbb{R}_+ \to \mathbb{R}$ be a semialgebraic function such that $\tilde{\epsilon}(t) \to 0$ as $t \to \infty$ and $\tilde{\epsilon}(t) > 0$ for all $t \geq 0$. The set $\{(t, \tilde{\epsilon}(t), x) \ : \ \operatorname{dist}(0, \partial^c f(x)) \leq \tilde{\epsilon}(t)\}$ is semialgebraic and unbounded. Hence, by the curve selection Lemma, there exists a C^1 semialgebraic path $\gamma : \mathbb{R}_+ \to \mathbb{R}^p$ such that $\operatorname{dist}(0, \partial^c f(\gamma(t)) \leq \tilde{\epsilon}(t)$ for any $t \geq 0$ and $\|\gamma(t)\| \to \infty$ as $t \to \infty$. By semialgebraicity, we may assume $\dot{\gamma}(t)$ does not vanish by considering only large t, and $\dot{\gamma}(t)/\|\dot{\gamma}(t)\| \to v$ for some unit vector $v \in \mathbb{R}^p$. Up to a strictly increasing time reparameterization, we may assume $\|\dot{\gamma}(t)\| = 1$ at any $t \geq 0$. Note that this time reparameterization does imply that $\dot{\gamma}(t) \to v$ and does not change the fact that $\tilde{\epsilon}(t) \to 0$ as $t \to \infty$.

Set $\lambda := \liminf_{\|x\| \to \infty} \frac{f(x)}{\|x\|} > 0$ and $f_1(x) := f(x) - \frac{\lambda}{2}\|x\|$ and $N := \partial^c \|\cdot\|$, the subdifferential of the norm. In particular, f_1 is coercive and semi-algebraic.

By applying path differentiability of f_1 along the C^1 curve γ, and using the sum rule for

path differentiable functions [19, Corollary 4], we may write

$$\begin{aligned} f_1(\gamma(T)) - f_1(\gamma(0)) &= \int_0^T \langle \dot{\gamma}(t), -\frac{\lambda}{2} N(\gamma(t)) \rangle dt \\ &\quad + \int_0^T \langle \dot{\gamma}(t), \partial^c f_1(\gamma(t)) + \frac{\lambda}{2} N(\gamma(t)) \rangle dt \qquad (14) \\ &= -\frac{\lambda}{2}(\|\gamma(T)\| - \|\gamma(0)\|) + \int_0^T \langle \dot{\gamma}(t), \operatorname{dist}(0, \partial^c f(\gamma(t))) \rangle dt \end{aligned}$$

Now, let us give some estimates. First, $\|\gamma(T)\| \geq \langle \gamma(T), v \rangle = \int_0^T \langle \dot{\gamma}(t), v \rangle dt$. And since $\dot{\gamma}(t) \to v$ we have $\|\gamma(T)\| \geq aT$ for some $a \in (0,1)$ for large values of T. Second, we have $\int_0^T \langle \dot{\gamma}(t), \operatorname{dist}(0, \partial^c f(\gamma(t))) \rangle dt = o(T)$, since its norm is bounded above by $\int_0^T \tilde{\epsilon}(t) dt = o(T)$ as $\tilde{\epsilon(t)} \to 0$. Thus we deduce from (14) that for T large enough, $f_1(\gamma(T)) = f_1(\gamma(0)) + \frac{\lambda}{2}\|\gamma(0)\| - (aT + b) + o(T)$. Therefore $f(\gamma(T)) \to -\infty$ which contradicts the coercivity of f_1.

2. Let us prove the case where $\beta > 1$ using similar ideas. Assume toward a contradiction that there exists $\bar{\epsilon} > 0$, such that $\operatorname{crit}_{\bar{\epsilon}} f$ is not bounded. We have a semialgebraic path γ such that $\|\gamma(t)\| \to \infty$ and $\operatorname{dist}(0, \partial^c f(\gamma(t))) \leq \bar{\epsilon}$ for all $t \geq 0$. We then rely on arguments that are analogous to the proof of 1: γ is chosen with unit speed, $\dot{\gamma}(t)/\|\dot{\gamma}(t)\|$ converges, and we set $\lambda := \liminf_{\|x\| \to \infty} \frac{f(x)}{\|x\|^\beta}$, $f_\beta(x) := f(x) - \frac{\lambda}{2}\|x\|^\beta$ and $N := \partial^c \|\cdot\|^\beta$. In particular, f_β is coercive. Applying similar arguments as in (14), there exists $c > 0$ such that for large values of $T > 0$

$$f_\beta(\gamma(T)) \leq f_\beta(\gamma(0)) - \frac{\lambda}{2}(\|\gamma(t)\|^\beta - \|\gamma(0)\|^\beta) + T\bar{\epsilon} = -cT^\beta + O(T)$$

where we used that $\frac{\lambda}{2}\|\gamma(T)\|^\beta \geq cT^\beta$ for some $c > 0$. Again we have $f_\beta(\gamma(T)) \to -\infty$ as $T \to \infty$ which contradicts its coercivity. □

4 Main consequences for the biased subgradient method

4.1 Link between discrete and continuous-time

In this section, we repeatedly use the connection between the discrete and continuous-time systems in the limit of small step sizes. This is a classical approach which we state for a general set-valued map, see [20, Lemma 21]. We use the following assumptions which guarantee the existence of solutions to the differential inclusion [5, Chapter 2, Section 1, Theorem 3].

Assumption 3. *Z is a set-valued map defined from $\mathbb{R}^p$ into the subsets of $\mathbb{R}^p$; it is nonempty convex valued and locally bounded with closed graph.*

Lemma 4.1 (Approximate solutions to differential inclusions)**.** *Let Z be as in Assumption 3. For each $i \in \mathbb{N}$, let $T_i > 0$ and assume that $T_i \to T$ as $i \to \infty$ for some $T > 0$. For each $i \in \mathbb{N}$, let $\gamma_i \colon [0, T_i] \to \mathbb{R}^p$ be a Lipschitz curve. Assume that the sequence*

$(\gamma_i)_{i\in\mathbb{N}}$ converges to some bounded and Lipschitz curve $\gamma\colon [0,T]\to\mathbb{R}$, in the sense that $\sup_{t\in[0,\min(T_i,T)]}\|\gamma(t)-\gamma_i(t)\|\to 0$, and

$$\lim_{i\to\infty}\int_0^{T_i}\operatorname{dist}((\gamma_i(t),\dot\gamma_i(t)),\operatorname{graph}[Z])\,dt=0. \tag{15}$$

Then $\dot\gamma(t)\in Z(\gamma(t))$ for almost every $t\in[0,T]$.

The following is classical and can be found for example in [20]. We will use it extensively to obtain discrete descent lemmas in the coming section.

Lemma 4.2 (Affine interpolants of discrete dynamics and their limit curves)**.** *Consider $T>0$ and sequences satisfying*

$$x_{k+1}-x_k\in\alpha_k Z(x_k),\qquad\qquad k=0,1,\dots,K,$$

where K is such that $\sum_{k=0}^{K-1}\alpha_k\geq T$ and $0<\alpha_k\leq\alpha$, $k=0,\dots,K-1$. Let $\gamma_\alpha\colon[0,T]\to\mathbb{R}^p$ be the interpolation of $(x_k)_{k=0,\dots,K}$ defined as follows: for all $k=0,1,\dots,K$, $\gamma_\alpha\left(\sum_{i=0}^{k-1}\alpha_k\right)=x_k$, and γ_α is affine between these points. The curve γ_α satisfies

$$\int_0^T\operatorname{dist}((\gamma_\alpha(t),\dot\gamma_\alpha(t)),\operatorname{graph}[Z])\,dt\leq\alpha T\sup_{0\leq k\leq K-1}\|Z(x_k)\|.$$

In particular, if Z is bounded and we have a uniformly bounded family of such curves γ_α, up to a subsequence, as $\alpha\to 0$, the curves converge uniformly to a solution of the underlying differential inclusion.

Proof : For all $\alpha>0$, the curve γ_α satisfies for any $t\in[0,T]$,

$$\operatorname{dist}((\gamma_\alpha(t),\dot\gamma_\alpha(t)),\operatorname{graph}[Z])\leq\alpha\sup_{0\leq k\leq K-1}\|Z(x_k)\|,$$

as $\dot\gamma_\alpha(t)=Z(x_k)$ where k is the closest point x_k on the curve corresponding to time smaller than t.

When Z is bounded by a constant L, one has a uniform bound since $\max_{0,\leq k\leq K}\|Z(x_k)\|\leq L$. In that case, the curves γ_α are L-Lipschitz continuous and thus equicontinuous. Whence, by letting $\alpha\to 0$, if the curves are bounded, Arzelà-Ascoli theorem applies and provides a subsequence of γ_α that uniformly converges to solutions of the continuous-time differential inclusion, thanks to Lemma 4.1. □

4.2 Descent lemmas

We now state some consequences of Lemma 4.1 which apply to our setting.

Lemma 4.3 (Quasi-descent Lemma)**.** *Let f be locally Lipschitz and path-differentiable, $\epsilon>0$ and $l\notin\operatorname{vcrit}_\epsilon f$. Then for any $\eta>0$ and $M>0$, there is $\bar\alpha>0$ such that for any $z\in\mathbb{R}^p$ satisfying $f(z)\leq l$, $\|z\|\leq M$, and for any sequence generated by (2), with $x_0=z$ and $\alpha_k\leq\bar\alpha$, we have $f(x_k)\leq l+\eta$ for all $k<\inf\{i\in\mathbb{N}:\|x_i\|>M\}$.*

In particular, if f is coercive, there is $M>0$, such that for α small enough, all sequences generated by (2) initialized with $f(x_0)\leq l$ are bounded by M.

Proof : Toward a contradiction, we assume that there are $\eta > 0$ and $M > 0$ such that for any $\alpha > 0$, there exists $z = x_0$ satisfying $f(z) \leq l$ and a sequence $(\alpha_k)_{k\in\mathbb{N}}$ smaller than α such that there exists $K > 0$ satisfying $f(x_K) > l + \eta$ and $\|x_k\| \leq M$ for all $k = 0, \dots, K$. Observe that $(\alpha_k)_{k\in\mathbb{N}}$ depends on α, and denote by L a Lipschitz constant of f on the ball of radius M.

We fix a time horizon $T < \eta/(L^2 + L\epsilon)$ and we may assume that $\alpha < T$. We may then also find $k \in \mathbb{N}$ such that $f(x_k) \leq l$ and $f(x_i) > l$ for all $i = k + 1, \dots, K$. We have

$$l + \eta < f(x_K) \leq f(x_k) + L \sum_{j=k}^{K-1} \|x_{j+1} - x_j\| \leq l + L \sum_{j=k}^{K-1} \alpha_j(L + \epsilon)$$
$$= l + (L^2 + L\epsilon) \sum_{j=k}^{K-1} \alpha_j.$$

We deduce that $\sum_{j=k}^{K-1} \alpha_j > T$. Let $K' \geq k$ be the smallest value such that $\sum_{j=k}^{K'} \alpha_j \geq T$, since $\alpha < T$ and by the argument above, we have $k < K' \leq K - 1$. We may then consider the truncated sequence $(x_i)_{i=k,\dots,K'+1}$, which is nonempty whenever $\alpha < T$. Consider the affine interpolation of this truncated sequence from $i = k$ to $i = K' + 1 \leq K$ and restricted to $[0, T]$ (which is possible because $\sum_{i=k}^{K'} \alpha_k \geq T$) as in Lemma 4.2, call it x_α, it is Lipschitz and satisfies:

- $\|x_\alpha(t)\| \leq M$ for all $t \in [0, T]$.
- $x_\alpha(0) = x_k$, $f(x_\alpha(0)) \leq l$ and $f(x_\alpha(t)) > l$ for some $t \in [0, \alpha]$.
- $l \leq f(x_\alpha(t)) \leq l + \eta + \alpha L(L + \epsilon)$ for all $t \in [\alpha, T]$.
- $\mathrm{dist}((x_\alpha(t), \dot{x}_\alpha(t)), \mathrm{graph}\left[-\partial^c f + \bar{B}(0, \epsilon)\right]) \leq \alpha(L + \epsilon)$ and $\|\dot{x}_\alpha(t)\| \leq L + \epsilon$ for almost every t.

Thus, these trajectories are uniformly bounded and equicontinuous. Applying Arzelà-Ascoli theorem allows to obtain a converging subsequence as $\alpha \to 0$. Lemma 4.1 ensures that the limit $x : [0, T] \to \mathbb{R}^p$ is a solution to (4) such that $f(x(0)) = l$ and $f(x([0, T])) \geq l$ which contradicts Lemma 3.1.1 (see Remark 3.1 for its validity for x defined on $[0, T]$).

The last remark follows because if f is coercive and locally Lipschitz, then it is Lipschitz on the (compact) sublevel set $l + \eta$, say with constant L. For a fixed step size threshold $\alpha_0 > 0$, choose $M_0 = \max\{\|x\| + \alpha_0(L+\epsilon)$, s.t. $f(x) \leq L+\eta\}$ and denote by $\tilde{L}$ a Lipschitz constant of f on the ball or radius M_0 centered at zero. By coercivity, we may choose $M > 0$ large enough such that $\inf_{\|x\|\geq M} f(x) > l + \eta + \alpha_0\tilde{L}(L + \epsilon)$. We apply the lemma for this choice of M, reducing the resulting $\bar{\alpha}$ if it is bigger than α_0. Fix any admissible sequence $(x_k)_{k\in\mathbb{N}}$ and k such that $f(x_k) \leq l + \eta$. It holds that $\|x_{k+1} - x_k\| \leq \alpha_0(L + \epsilon)$ and $\|x_k\| + \alpha_0(L + \epsilon) \leq M_0$ so that both x_k and x_{k+1} are contained in the ball of radius M_0. We deduce that $f(x_{k+1}) \leq l + \eta + \alpha_0\tilde{L}(L + \epsilon)$ so $\|x_{k+1}\| \leq M$. We deduce that $k + 1 < \inf\{i \in \mathbb{N} : \|x_i\| > M\}$ and the main statement of the lemma ensures that, $f(x_{k+1}) \leq l + \eta$. By induction this holds true for all $k \in \mathbb{N}$. □

Lemma 4.4 (ϵ-regular values are repulsive)**.** *Let $f\colon \mathbb{R}^p \to \mathbb{R}$ be a locally Lipschitz and path-differentiable function, $\epsilon > 0$, $l \notin \operatorname{vcrit}_\epsilon f$, and $\eta := \operatorname{dist}(l, \operatorname{vcrit}_\epsilon f)/16 > 0$. For any $M > 0$, there is $\bar{\alpha} > 0$ such that for any sequence generated by (2), with $\alpha_k \leq \bar{\alpha}$ for all $k \in \mathbb{N}$, and $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$, either $\{i \in \mathbb{N},\ \|x_i\| > M\}$ is nonempty or $\liminf_{k\to\infty} f(x_k) > l + 2\eta$ or $\limsup_{k\to\infty} f(x_k) < l - 2\eta$.*

In particular under Assumption 1, there is $\bar{\alpha} > 0$ such that regardless of the initial condition for any sequence generated by (2), with $\alpha_k \leq \bar{\alpha}$, and $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$ for all $k \in \mathbb{N}$, $\liminf_{k\to\infty} f(x_k) > l + 2\eta$ or $\limsup_{k\to\infty} f(x_k) < l - 2\eta$.

Proof : Fix $M > 0$ and denote by L a Lipschitz constant of f on the ball of radius M. By definition of η, we have that $[l - 8\eta, l + 8\eta] \subset (\operatorname{vcrit}_\epsilon f)^c$. Set $\delta := \min\{\|v\| : v \in \partial^c f(x),\ x \in \mathbb{R}^p,\ \|x\| \leq M,\ |f(x) - l| \leq 8\eta\} > \epsilon$ and $T \geq 12\frac{\eta}{\delta(\delta-\epsilon)}$. Any solution $x : [0, T] \to \mathbb{R}^p$ to (4) bounded by M such that $f(x(0)) \leq l + 4\eta$ satisfies $f(x(T)) \leq l - 8\eta$ by Lemma 3.1.1. We first construct a thresholds step size $\bar{\alpha} > 0$ which satisfies three desirable properties.

1. There is $\bar{\alpha} > 0$, such that any sequence generated by (2), bounded by M, with $f(x_0) \leq l + 4\eta$ satisfies $f(x_k) \leq l - 4\eta$ for some $k \in \mathbb{N}$. Indeed, if this was not the case, using Lemma 4.1 and the fact that $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$, we could construct a solution curve to the differential inclusion $x\colon \mathbb{R}^+ \to \mathbb{R}$ satisfying $f(x(0)) \leq l + 4\eta$ and $f(x(t)) \geq l - 4\eta$ for all $t \geq 0$. In particular $f(x(T)) \geq l - 4\eta$, which contradicts the previous claim stating that $f(x(T)) \leq l - 8\eta$ for any such curve.

2. We may assume that for any sequence bounded by M such that $\limsup_{k\to\infty} f(x_k) \geq l-2\eta$ and $\liminf_{k\to\infty} f(x_k) \leq l+2\eta$, $(f(x_k))_{k\in\mathbb{N}}$ takes infinitely many value in $[l-4\eta, l+4\eta]$. Indeed, for all k, we have $\|x_k\| \leq M$ and $\|x_{k+1}\| \leq M$, so that

$$|f(x_{k+1}) - f(x_k)| \leq L\|x_{k+1} - x_k\| \leq \bar{\alpha}L(L + \epsilon),$$

shrinking $\bar{\alpha}$ if necessary that this gap is less than η.

3. We may reduce $\bar{\alpha}$ further, as given by Lemma 4.3, so that any sequence initialized with $f(x_0) \leq l - 4\eta$ and bounded by M satisfies $f(x_k) \leq l - 3\eta$ for all $k \in \mathbb{N}$. The chosen $\bar{\alpha}$ satisfies the required properties.

Let us summarize the properties of the resulting $\bar{\alpha}$. Note that the properties claimed above can be shifted by initializing a sequence at an arbitrary $K \in \mathbb{N}$ and by considering $k \geq K$. Below, $(x_k)_{k\in\mathbb{N}}$ is any sequence generated by (2) with $\alpha_k \leq \bar{\alpha}$ for all $k \in \mathbb{N}$, bounded by M, and $K \in \mathbb{N}$ is arbitrary.

1. If $f(x_K) \leq l + 4\eta$, there is $k \geq K$ such that $f(x_k) \leq l - 4\eta$.
2. If $\limsup_{k\to\infty} f(x_k) \geq l - 2\eta$ and $\liminf_{k\to\infty} f(x_k) \leq l + 2\eta$, then there is $k \in \mathbb{N}$ such that $f(x_k) \in [l - 4\eta, l + 4\eta]$.
3. If $f(x_K) \leq l - 4\eta$ then $f(x_k) \leq l - 3\eta$ for all $k \geq K$.

For the choice of $\bar{\alpha}$ as above, assume toward a contradiction that there exists a sequence generated by (2), bounded by M, with $\alpha_k \leq \bar{\alpha}$ for all $k \in \mathbb{N}$, satisfying $\limsup_{k\to\infty} f(x_k) \geq$

$l - 2\eta$ and $\liminf_{k\to\infty} f(x_k) \leq l + 2\eta$. There would be $K \in \mathbb{N}$ such that $f(x_K) \in [l - 4\eta, l + 4\eta]$ by 2 and we deduce by 1 that there is $k \geq K$ such that $f(x_k) \leq l - 4\eta$. This implies by 3 that $\limsup_{k\to\infty} f(x_k) \leq l - 3\eta$ which is contradictory.

The last comment follows because by Lemma 4.3 reducing $\bar{\alpha}$ if necessary, there is $M > 0$ such that all sequences generated by (2) with $f(x_0) \leq l + 3\eta$ and $\alpha_k \leq \bar{\alpha}$ for all $k \in \mathbb{N}$ are bounded by M. If $\liminf_{k\to\infty} f(x_k) \leq l + 2\eta$, then there is $K > 0$ such that $f(x_K) \leq l + 3\eta$ and by considering $\tilde{x}_k = x_{k+K}$, bounded by M, the result follows. □

Under Assumption 1, f is coercive and $\operatorname{vcrit}_\epsilon f$ is closed. Combining the two previous lemma we obtain

Corollary 4.1 (Large-horizon descent Lemma)**.** *Under Assumption 1, fix x_0 such that $f(x_0) \notin \operatorname{vcrit}_\epsilon f$ and $\eta > 0$. There is $\bar{\alpha} > 0$ such that any sequence generated by (2) with $\alpha_k \leq \bar{\alpha}$ and $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$ we have*

(i) $f(x_k) \leq f(x_0) + \eta$ for all $k \in \mathbb{N}$,

(ii) there is κ such that

$$f(x_k) \leq c + \eta, \ \forall k \geq \kappa$$

where c is the first ϵ-critical value below $f(x_0)$.

Note that if $(1 + \rho)\eta < f(x_0) - c$ for some $\rho > 0$, the above implies

$$f(x_k) \leq f(x_0) - \rho\eta \text{ for } k \geq \kappa,$$

where κ is unknown.

4.3 Vanishing step sizes

For vanishing step sizes, the work of Benaim-Hofbauer-Sorin [9] ensures that the set of accumulation points is invariant. We additionally show that it has accumulation values in $\operatorname{vcrit}_\epsilon f$. The main result stated in Theorem 2.1 is a consequence of the following result and Lemma 3.2.

Theorem 4.1 (Convergence for biased subgradient with vanishing step)**.** *Under Assumption 2, let C, ρ be given by Lemma 3.3 and $\epsilon < \bar{\epsilon}$. Assume that $(x_k)_{k\in\mathbb{N}}$, is given by (2) with $\alpha_k \to 0$ and $\sum_{k=0}^{\infty} \alpha_k = \infty$. If $\sup_{k\in\mathbb{N}} \|x_k\| < \infty$,*

$$\lim_{k\to\infty} \operatorname{dist}(f(x_k), \operatorname{vcrit}_\epsilon f) = 0,$$
$$\limsup_{k\to\infty} \operatorname{dist}(x_k, \operatorname{crit} f) \leq C\epsilon^\rho.$$

The boundedness condition always holds for small enough step sizes if f is coercive.

Proof : By assumption, there is $M > 0$ such that for all $k \in \mathbb{N}$, $\|x_k\| \leq M$. Denote by L_f the set of limit points of $f(x_k)$, it is an interval since $\alpha_k \to 0$. We first prove the first statement which is equivalent to the fact that L_f does not contain any value

$l \notin \operatorname{vcrit}_\epsilon f$. Suppose toward a contradiction that this is the case, then there exists a value $l \in (\operatorname{vcrit}_\epsilon f)^c \cap L_f$. For such a value l, let $\eta > 0$ be given by Lemma 4.4 ($\operatorname{vcrit}_\epsilon f$ is closed by Assumption 2). Recall that $(\alpha_k)_{k\in\mathbb{N}}$ is vanishing so that either $\liminf_{k\to\infty} f(x_k) > l+2\eta$ or $\limsup_{k\to\infty} f(x_k) < l - 2\eta$, which is in contradiction with $l \in L_f$.

Denote by L the set of accumulation points of the sequence $(x_k)_{k\in\mathbb{N}}$. By the previous statement we have $f(L) \subset \operatorname{vcrit}_\epsilon f$. By [9, Theorem 4.2 and 4.3], L is an internally chain transitive set. In particular, L is invariant for the inclusion (4) by [9, Lemma 3.5]. This means that for any $\bar{x} \in L$, there is a solution x to (4), such that $x(0) = \bar{x}$ and $x(\mathbb{R}) \in L$. The second statement follows by Corollary 3.1.

The last comment on boundedness follows from Lemma 4.3. □

4.4 Constant step sizes

We start this section with a general result on constant step size discretization which is of independent interest.

Lemma 4.5 (Limits of accumulation points are invariant)**.** *Let $M > 0$ be a bound and $x_0 \in \mathbb{R}^p$ be fixed. Under Assumption 3, for any $s > 0$, denote by $L(s)$ the set of accumulation points of sequences satisfying $x_{k+1}(s) = x_k(s) + sh(x_k(s))$ for each $k \in \mathbb{N}$, where h is a selection in Z. Assume that $L(s)$ is nonempty and bounded by M for all s small enough. Set*

$$L = \bigcap_{\alpha>0} \operatorname{cl}\left\{ \bigcup_{0<s\leq\alpha} L(s) \right\}.$$

Then the set L is nonempty and invariant with respect to the flow induced by Z in the sense that for any $\bar{x} \in L$ there exists a solution to the differential inclusion $\dot{x}(t) \in Z(x(t))$ for almost every $t \in \mathbb{R}$ such that $x(\mathbb{R}) \subset L$ and $x(0) = \bar{x}$.

Proof : The assumption that $L(s)$ is bounded by a fixed M for small enough s ensures that all considered sequences are bounded and we may restrict all the asymptotics to happen in a fixed bounded set (for example of size $2M$), in particular, Z will be bounded by ζ on this set. This ensures that L is nonempty and this that all considered objects remain in a bounded set.

Let $\bar{x} \in L$. For a fixed $\alpha > 0$, $\bar{x} \in \operatorname{cl} \cup_{s\leq\alpha} L(s)$ means that for any $e > 0$, there is $s \leq \alpha$ such that $\operatorname{dist}(\bar{x}, L(s)) \leq e$. Therefore if $\bar{x} \in L$, we deduce that there exists a sequence $(s_j)_{j\in\mathbb{N}}$ which tends to 0 and a sequence $\bar{x}_j \in L(s_j)$ which tends to $\bar{x}$ as $j \to \infty$.

Fix an arbitrary $T > 0$ and $j \in \mathbb{N}$. We set $K_j = \lceil T/s_j \rceil$ and consider for $k \geq K_j$, $X_{j,k} = (x_i(s_j))_{i=k-K_j}^{i=k+K_j}$. Up to a subsequence, as $k \to \infty$, $X_{j,k}$ converges to $\bar{X}_j$ which contains $2K_j+1$ accumulation points in $L(s_j)$, the central one chosen to be $\bar{x}_j$. The affine interpolation of the (ordered) collection of points in $\bar{X}_j$ is denoted $\bar{\gamma}_j \colon [-T, T] \to \mathbb{R}^p$ with $\bar{\gamma}_j(0) = \bar{x}_j$.

This construction can be repeated for all $j \in \mathbb{N}$ and using Arzelà-Ascoli, there is a subsequence of $(\bar{\gamma}_j)_{j\in\mathbb{N}}$ which converges uniformly to $\bar{\gamma} \colon [-T, T] \to \mathbb{R}^p$. We have $\bar{\gamma}([-T, T]) \subset L$.

Indeed let $t \in [-T, T]$. For any $e > 0$, we have $\|\bar{\gamma}_j(t) - \bar{\gamma}(t)\| \leq e$ for j high enough. By construction, $\bar{\gamma}_j(s_j\lceil t/s_j \rceil) \in L(s_j)$. Since $s_j\lceil t/s_j \rceil$ converges to t and $\bar{\gamma}_j$ is Lipschitz with same constant for all j, then $\|\bar{\gamma}(t) - \bar{\gamma}_j(s_j\lceil t/s_j \rceil)\| \leq 2e$ for all j high enough. Since e was arbitrary, this means $\bar{\gamma}(t) \in L$.

We claim that $\bar{\gamma}$ is a solution to the differential inclusion $\frac{d}{dt}\bar{\gamma}(t) \in Z(\bar{\gamma}(t))$ for almost all t.

Indeed, for any $j \in \mathbb{N}$, there is $k_j \in \mathbb{N}$ such that $\|X_{j,k_j} - \bar{X}_j\| \leq 1/j$. Let $\gamma_j \colon [-T, T] \to \mathbb{R}^p$ be the affine interpolation of X_{j,k_j}. Up to the Arzelà-Ascoli subsequence, $\gamma_j \to \bar{\gamma}$ uniformly on $[-T, T]$. Furthermore, for each j

$$\int_{-T}^{T} \operatorname{dist}((\gamma_j(t), \dot{\gamma}_j(t)), \operatorname{graph}[Z])\, dt \leq 2Ts_j\zeta$$

By Lemma 4.1, this shows that $\bar{\gamma}$ is a solution as required.

The function $\bar{\gamma}$ may be extended to $[-2T, 2T]$, $[-3T, 3T]$, ... by taking further subsequences for each j and further Arzelà-Ascoli subsequences. The resulting function is defined on $\mathbb{R}$ and has to be a solution to the differential inclusion. This concludes the proof. □

The main result stated in Theorem 2.2 is a consequence of the following result and Lemma 3.2.

Theorem 4.2 (Convergence for biased subgradient with constant steps)**.** *Under Assumptions 2, let C, ρ be given by Lemma 3.3 and $\epsilon < \bar{\epsilon}$. Set for all $\alpha > 0$, $(x_k(\alpha))_{k\in\mathbb{N}}$, as given by* (2) *with $\alpha_k = \alpha$, for all $k \in \mathbb{N}$. Then if $\limsup_{\alpha\to 0} \sup_{k\to\infty} \|x_k(\alpha)\| < \infty$,*

$$\lim_{\alpha\to 0} \limsup_{k\to\infty} \operatorname{dist}(f(x_k(\alpha)), \operatorname{vcrit}_\epsilon f) = 0,$$
$$\limsup_{\alpha\to 0} \limsup_{k\to\infty} \operatorname{dist}(x_k(\alpha), \operatorname{crit} f) \leq C\epsilon^\rho.$$

The boundedness condition always holds if f is coercive.

Proof : By assumption, there is $M > 0$ such that for small enough α, $\|x_k(\alpha)\| \leq M$ for all $k \in \mathbb{N}$ large enough.

For each s denote by $L(s)$ the set of limit points of $x_k(s)$, which is closed. We set $L = \cap_{\alpha>0}\operatorname{cl} \cup_{s\leq\alpha} L(s)$ which is closed as an intersection of closed set, and bounded by M. Intuitively L is the set of accumulation points of accumulation points of $(x_k(\alpha))_{k\in\mathbb{N}}$ as $\alpha \to 0$. By Lemma 4.5, L is invariant. We will show that $f(L) \subset \operatorname{vcrit}_\epsilon f$ from which the first statement follows. The second is then a consequence of Corollary 3.1.

For any $l \notin \operatorname{vcrit}_\epsilon f$, since $\operatorname{vcrit}_\epsilon f$ is closed by Assumption 2, Lemma 4.4 ensures that there is $\bar{\alpha}$ and η, such that for any $s \leq \bar{\alpha}$, $[l - 2\eta, l + 2\eta] \cap f(L(s)) = \emptyset$. We deduce that

$$[l - 2\eta, l + 2\eta] \cap f\left(\cup_{s\leq\bar{\alpha}} L(s)\right) = \emptyset$$
$$[l - \eta, l + \eta] \cap f\left(\operatorname{cl} \cup_{s\leq\bar{\alpha}} L(s)\right) = \emptyset \quad \text{(use continuity)}.$$

Whence we see that for any $l \notin \operatorname{vcrit}_\epsilon f$, $l \notin f\left(\cap_{\alpha>0}\operatorname{cl}\cup_{s\leq\alpha} L(s)\right)$, and therefore $f(L) \subset \operatorname{vcrit}_\epsilon f$ through complementation. This proves the first claim.

The last comment on boundedness follows from Lemma 4.3. □

5 The convex case

Proof of Theorem 2.3: Consider a sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$ and a iterates $(x_k)_{k\in\mathbb{N}}$ such that for all $k \in \mathbb{N}$,

$$x_{k+1} = x_k - \alpha_k(v_k + b_k),$$

for some $v_k \in \partial^c f(x_k)$ and $\|b_k\| \leq \epsilon$. We have for any $k \in \mathbb{N}$ and any $x^* \in X^*$,

$$\begin{aligned}\frac{1}{2}\|x_{k+1} - x^*\|_2^2 &= \frac{1}{2}\|x_k - \alpha_k(v_k + b_k) - x^*\|_2^2 \\ &= \frac{1}{2}\|x_k - x^*\|_2^2 + \alpha_k(v_k + b_k)^T(x^* - x_k) + \frac{\alpha_k^2}{2}\|v_k + b_k\|_2^2 \\ &\leq \frac{1}{2}\|x_k - x^*\|_2^2 - \alpha_k(f(x_k) - f^*) + \alpha_k\epsilon\|x_k - x^*\| \\ &\quad + \frac{\alpha_k^2}{2}(L + \epsilon)^2.\end{aligned}$$

We deduce, using the error bound (3), that

$$\begin{aligned}&\frac{1}{2}\operatorname{dist}(x_{k+1}, X^*)^2 \qquad (16)\\ &\leq \frac{1}{2}\operatorname{dist}(x_k, X^*)^2 - \alpha_k(f(x_k) - f^*) + \alpha_k\epsilon\operatorname{dist}(x_k, X^*) + \frac{\alpha_k^2}{2}(L + \epsilon)^2 \\ &\leq \frac{1}{2}\operatorname{dist}(x_k, X^*)^2 + \alpha_k\left(\epsilon\frac{c}{2}\left((f(x_k) - f^*)^a + (f(x_k) - f^*)\right) - (f(x_k) - f^*)\right) \\ &\quad + \frac{\alpha_k^2}{2}(L + \epsilon)^2.\end{aligned}$$

If $a = 1$: This means that we have a sharp function, then, as long as $\epsilon c < 1$, we obtain the same global rate as the classical subgradient algorithm, modulo a constant $(1 - \epsilon c)$, and all accumulation points in the argmin. Indeed, summing the previous inequality from $i = 0$ to k leads to

$$(1 - \epsilon c)\frac{\sum_{i=0}^k \alpha_i(f(x_i) - f^*)}{\sum_{i=0}^k \alpha_i} \leq \frac{\|x_0 - x^*\|_2^2 + (L + \epsilon)^2\sum_{i=0}^k \alpha_i^2}{\sum_{i=0}^k \alpha_i}.$$

We remark that this is exactly the claimed formula for $a = 1$.

If $a < 1$: Then we can use the following lemma

Lemma 5.1. *Set $g(\delta) = s\delta^t - \delta$ for some $t \in (0,1)$ and $s > 0$, then for all $\delta \in \mathbb{R}_+$*

$$g(\delta) \leq -(1-t)(\delta - s^{\frac{1}{1-t}}).$$

Proof : The function g is concave on $\mathbb{R}_+$, set $\delta_0 = s^{\frac{1}{1-t}}$, we have $g(\delta_0) = 0$ and $g'(\delta_0) = t - 1 < 0$ and the result follows by concavity. □

We obtain setting $\delta = f(x_k) - f^*$ applying Lemma 5.1 to (16)

$$\begin{aligned}
&\frac{1}{2}\mathrm{dist}(x_{k+1}, X^*)^2 - \frac{1}{2}\mathrm{dist}(x_k, X^*)^2 \\
&\leq \frac{\alpha_k}{2}\left(-(1-a)\left((f(x_k) - f^*) - (\epsilon c)^{\frac{1}{1-a}}\right) + (\epsilon c - 1)(f(x_k) - f^*)\right) \\
&\quad + \frac{\alpha_k^2}{2}(L+\epsilon)^2 \\
&= \frac{\alpha_k}{2}\left((\epsilon c + a - 2)(f(x_k) - f^*) + (1-a)(\epsilon c)^{\frac{1}{1-a}}\right) + \frac{\alpha_k^2}{2}(L+\epsilon)^2
\end{aligned}$$

We deduce by summation

$$\begin{aligned}
(2 - a - \epsilon c)\frac{\sum_{i=0}^k \alpha_i (f(x_i) - f^*)}{\sum_{i=0}^k \alpha_i} &\leq (1-a)(\epsilon c)^{\frac{1}{1-a}} \\
&\quad + \frac{\|x_0 - x_*\|^2 + (L+\epsilon)^2 \sum_{i=0}^k \alpha_i^2}{\sum_{i=0}^k \alpha_i}.
\end{aligned}$$

□

Acknowledgements

The authors acknowledge the support of Centre Lagrange. JB, TL, EP thank AI Interdisciplinary Institute ANITI funding, through the French "Investments for the Future – PIA3" program under the grant agreement ANR-19-PI3A0004, Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant numbers FA8655-22-1-7012, ANR Chess, grant ANR-17-EURE-0010, ANR Regulia.

References

[1] Aliprantis, C., Border, K.: Infinite Dimensional Analysis. Springer Berlin, Heidelberg (2006). DOI 10.1007/3-540-29587-9

[2] Amos, B., Kolter, J.Z.: Optnet: Differentiable optimization as a layer in neural networks. In: International Conference on Machine Learning, pp. 136–145. PMLR (2017)

[3] Arscott, F., Filippov, A.: Differential Equations with Discontinuous Righthand Sides: Control Systems. Mathematics and its Applications. Springer Netherlands (1988)

[4] Aspelmeier, T., Charitha, C., Luke, D.R.: Local linear convergence of the admm/douglas–rachford algorithms without strong convexity and application to statistical imaging. SIAM Journal on Imaging Sciences **9**(2), 842–868 (2016)

[5] Aubin, J.P., Cellina, A.: Differential inclusions: set-valued maps and viability theory, vol. 264. Springer Science & Business Media (1983)

[6] Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Convex optimization with sparsity-inducing norms. Optimization for Machine Learning pp. 19–49 (2011)

[7] Ben-Tal, A., El Ghaoui, L., Nemirovski, A.: Robust optimization, vol. 28. Princeton university press (2009)

[8] Benaïm, M.: Recursive algorithms, urn processes and chaining number of chain recurrent sets. Ergodic Theory and Dynamical Systems **18**(1), 53–87 (1998)

[9] Benaim, M., Hofbauer, J., Sorin, S.: Stochastic approximations and differential inclusions. SIAM Journal on Control and Optimization **44**(1), 328–348 (2005)

[10] Benaïm, M., Hofbauer, J., Sorin, S.: Perturbations of set-valued dynamical systems, with applications to game theory. Dynamic Games and Applications **2**(2), 195–205 (2012). DOI 10.1007/s13235-012-0040-0. URL `https://doi.org/10.1007/s13235-012-0040-0`

[11] Benveniste, A., Métivier, M., Priouret, P.: Adaptive algorithms and stochastic approximations, vol. 22. Springer Science & Business Media (2012)

[12] Bianchi, P., Hachem, W., Salim, A.: Constant step stochastic approximations involving differential inclusions: stability, long-run convergence and applications. Stochastics **91**(2), 288–320 (2019)

[13] Bianchi, P., Hachem, W., Schechtman, S.: Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. Set-Valued and Variational Analysis **30**(3), 1117–1147 (2022)

[14] Blondel, M., Berthet, Q., Cuturi, M., Frostig, R., Hoyer, S., Llinares-López, F., Pedregosa, F., Vert, J.P.: Efficient and modular implicit differentiation. Advances in neural information processing systems **35**, 5230–5242 (2022)

[15] Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM Journal on Optimization **17**(4), 1205–1223 (2007). DOI 10.1137/050644641. URL `https://doi.org/10.1137/050644641`

[16] Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM Journal on Optimization **18**(2), 556–572 (2007)

[17] Bolte, J., Le, T., Pauwels, E.: Subgradient sampling for nonsmooth nonconvex minimization. SIAM Journal on Optimization **33**(4), 2542–2569 (2023). DOI 10.1137/22M1479178

[18] Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.W.: From error bounds to the complexity of first-order descent methods for convex functions. Mathematical Programming **165**, 471–507 (2017)

[19] Bolte, J., Pauwels, E.: Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. Mathematical Programming **188**(1), 19–51 (2021)

[20] Bolte, J., Pauwels, E., Rios-Zertuche, R.: Long term dynamics of the subgradient method for lipschitz path differentiable functions. Journal of the European Mathematical Society (2022)

[21] Bolte, J., Pauwels, E., Vaiter, S.: One-step differentiation of iterative algorithms. Advances in Neural Information Processing Systems **36** (2023)

[22] Borkar, V.S.: Stochastic approximation: a dynamical systems viewpoint, vol. 48. Springer (2009)

[23] Clarke, F.H.: Optimization and nonsmooth analysis. SIAM (1983)

[24] Cohen, J., Kaur, S., Li, Y., Kolter, J.Z., Talwalkar, A.: Gradient descent on neural networks typically occurs at the edge of stability. In: International Conference on Learning Representations (2020)

[25] Combettes, P.L., Pesquet, J.C.: Proximal splitting methods in signal processing. In: Fixed-point algorithms for inverse problems in science and engineering, pp. 185–212. Springer (2011)

[26] Coste, M.: An introduction to o-minimal geometry. Istituti editoriali e poligrafici internazionali Pisa (2000)

[27] Davis, D., Drusvyatskiy, D., Kakade, S., Lee, J.D.: Stochastic subgradient method converges on tame functions. Foundations of computational mathematics **20**(1), 119–154 (2020)

[28] De Sa, C., Feldman, M., Ré, C., Olukotun, K.: Understanding and optimizing asynchronous low-precision stochastic gradient descent. In: Proceedings of the 44th annual international symposium on computer architecture, pp. 561–574 (2017)

[29] Dieuleveut, A., Fort, G., Moulines, E., Wai, H.T.: Stochastic approximation beyond gradient for signal processing and machine learning. IEEE Transactions on Signal Processing (2023)

[30] Donoho, D.L.: Compressed sensing. IEEE Transactions on information theory **52**(4), 1289–1306 (2006)

[31] Doucet, A., Tadic, V.: Asymptotic bias of stochastic gradient search. Annals of Applied Probability **27**(6) (2017)

[32] van den Dries, L.: Tame topology and o-minimal structures, vol. 248. Cambridge university press (1998)

[33] Ermol'ev, Y.M., Norkin, V.: Stochastic generalized gradient method for nonconvex nonsmooth stochastic optimization. Cybernetics and Systems Analysis **34**(2), 196–215 (1998)

[34] Ermoliev, Y.M.: Stochastic quasigradient methods and their application to system optimization. Stochastics **9**, 1–36 (1983)

[35] Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)

[36] Fort, J.C., Pages, G.: Asymptotic behavior of a markovian stochastic algorithm with constant step. SIAM journal on control and optimization **37**(5), 1456–1482 (1999)

[37] Ghadimi, S., Lan, G.: Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization **23**(4), 2341–2368 (2013)

[38] Ghadimi, S., Lan, G., Zhang, H.: Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Mathematical Programming **155**(1), 267–305 (2016)

[39] Griewank, A., Faure, C.: Piggyback differentiation and optimization. In: Large-scale PDE-constrained optimization, pp. 148–164. Springer (2003)

[40] Hesse, R., Luke, D.R.: Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. SIAM Journal on Optimization **23**(4), 2397–2419 (2013)

[41] Huh, W.T., Rusmevichientong, P.: Online sequential optimization with biased gradients: theory and applications to censored demand. INFORMS Journal on Computing **26**(1), 150–159 (2014)

[42] Josz, C., Lai, L.: Global stability of first-order methods for coercive tame functions. Mathematical Programming pp. 1–26 (2023)

[43] Josz, C., Lai, L.: Lyapunov stability of the subgradient method with constant step size. Mathematical Programming **202**(1), 387–396 (2023). DOI 10.1007/s10107-023-01936-6. URL `https://doi.org/10.1007/s10107-023-01936-6`

[44] Kushner, H., Yin, G.: Stochastic Approximation and Recursive Algorithms and Applications. Stochastic Modelling and Applied Probability. Springer New York (2003)

[45] Lan, G., Zhou, Z.: Algorithms for stochastic optimization with function or expectation constraints. Computational Optimization and Applications **76**(2), 461–498 (2020)

[46] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)

[47] Lee, J.H., Pham, T.S.: Openness, hölder metric regularity, and hölder continuity properties of semialgebraic set-valued maps. SIAM Journal on Optimization **32**(1), 56–74 (2022). DOI 10.1137/20M1331901. URL `https://doi.org/10.1137/20M1331901`

[48] Ljung, L.: Analysis of recursive stochastic algorithms. IEEE transactions on automatic control **22**(4), 551–575 (1977)

[49] Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (2017)

[50] Luke, D.R., Teboulle, M., Thao, N.H.: Necessary conditions for linear convergence of iterated expansive, set-valued mappings. Mathematical Programming **180**, 1–31 (2020)

[51] Majewski, S., Miasojedow, B., Moulines, E.: Analysis of nonsmooth stochastic approximation: the differential inclusion approach. arXiv preprint arXiv:1805.01916 (2018)

[52] Mäkelä, M.: Survey of bundle methods for nonsmooth optimization. Optimization methods and software **17**(1), 1–29 (2002)

[53] Mishchenko, K., Gorbunov, E., Takáč, M., Richtárik, P.: Distributed learning with compressed gradient differences. arXiv preprint arXiv:1901.09269 (2019)

[54] Nedić, A., Bertsekas, D.P.: The effect of deterministic noise in subgradient methods. Mathematical programming **125**(1), 75–99 (2010)

[55] Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM Journal on optimization **19**(4), 1574–1609 (2009)

[56] Nemirovskij, A.S., Yudin, D.B.: Problem complexity and method efficiency in optimization. Wiley-Interscience (1983)

[57] Nguyen, N., Yin, G.: Stochastic approximation with discontinuous dynamics, differential inclusions, and applications. The Annals of Applied Probability **33**(1), 780–823 (2023)

[58] Norkin, V.I.: Nonlocal minimization algorithms of nondifferentiable functions. Cybernetics **14**(5), 704–707 (1978). DOI 10.1007/BF01069307

[59] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)

[60] Ramaswamy, A., Bhatnagar, S.: Analysis of gradient descent methods with nondiminishing bounded errors. IEEE Transactions on Automatic Control **63**(5), 1465–1471 (2017)

[61] Roth, G., Sandholm, W.H.: Stochastic approximations with constant step size and differential inclusions. SIAM Journal on Control and Optimization **51**(1), 525–555 (2013)

[62] Russell Luke, D., Thao, N.H., Tam, M.K.: Quantitative convergence analysis of iterated expansive, set-valued mappings. Mathematics of Operations Research **43**(4), 1143–1176 (2018)

[63] Shor, N.Z.: Minimization methods for non-differentiable functions, vol. 3. Springer Science & Business Media (2012)

[64] Solodov, M.V., Zavriev, S.: Error stability properties of generalized gradient-type algorithms. Journal of Optimization Theory and Applications **98**, 663–680 (1998)

[65] Spall, J.C.: Stochastic optimization. Handbook of computational statistics: Concepts and methods pp. 173–201 (2012)

[66] Tadić, V.B., Doucet, A.: Asymptotic bias of stochastic gradient search. In: 2011 50th IEEE Conference on Decision and Control and European Control Conference, pp. 722–727. IEEE (2011)

[67] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology **58**(1), 267–288 (1996)

[68] Wangni, J., Wang, J., Liu, J., Zhang, T.: Gradient sparsification for communication-efficient distributed optimization. Advances in Neural Information Processing Systems **31** (2018)

[69] Xiao, N., Hu, X., Toh, K.C.: Sgd-type methods with guaranteed global stability in nonsmooth nonconvex optimization. arXiv preprint arXiv:2307.10053 (2023)