Global stability of first-order methods for coercive tame functions

Cédric Josz[*] Lexiao Lai[†]

Abstract

We consider first-order methods with constant step size for minimizing locally Lipschitz coercive functions that are tame in an o-minimal structure on the real field. We prove that if the method is approximated by subgradient trajectories, then the iterates eventually remain in a neighborhood of a connected component of the set of critical points. Under suitable method-dependent regularity assumptions, this result applies to the subgradient method with momentum, the stochastic subgradient method with random reshuffling and momentum, and the random-permutations cyclic coordinate descent method.

Key words: differential inclusions, Kurdyka-Łojasiewicz inequality, semi-algebraic geometry

1 Introduction

Consider the unconstrained minimization problem

$$\inf_{x\in\mathbb{R}^n} f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x) \tag{1}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz for $i = 1, 2, \ldots, N$. Such unconstrained optimization problems are central in machine learning applications such as empirical risk minimization [19], low-rank matrix recovery [42, 46, 74], and the training of deep neural networks [40]. We study some widely used first-order methods, namely the subgradient method with momentum (Algorithm 1), the stochastic subgradient method with random reshuffling and momentum (Algorithm 2), and the random-permutations cyclic coordinate descent method (Algorithm 3). While they are

[*] `cj2638@columbia.edu`, IEOR, Columbia University, New York. Research supported by NSF EPCN grant 2023032 and ONR grant N00014-21-1-2282.

[†] `ll3352@columbia.edu`, IEOR, Columbia University, New York.

implemented by machine learning practitioners [64, 1, 56], the analysis of these methods with constant step sizes seems to be absent from the literature when the objective is neither convex nor differentiable with a locally Lipschitz gradient (see Section 2).

In this paper, we provide global stability guarantees for first-order methods with constant step size for objective functions that are locally Lipschitz, coercive, and tame in an o-minimal structure on the real field (Definition 4). In order to do so, we show that the function values and the iterates of an iterative method eventually stabilize around some critical value (Theorem 1) and the set of critical points (Corollary 1) respectively, given that the method is approximated by subgradient trajectories of the objective function (Definition 3). As it turns out, all of the aforementioned first-order methods are approximated by subgradient trajectories of locally Lipschitz functions under method-dependent regularity assumptions (Propositions 1 and 2) as summarized in Table 1 (random reshuffling with momentum is short for stochastic subgradient method with random reshuffling and momentum). Therefore, these methods fit into our framework and their stability is guaranteed by Theorem 1 and Corollary 1. To the best of our knowledge, these methods have not been studied before at such generality as in this paper. In particular, we do not require the objective function to be convex and we do not require it to be differentiable with a locally Lipschitz gradient.

The function class studied in this paper is well-suited for applications. Indeed, seemingly all continuous objective functions of interest nowadays are locally Lipshitz and tame in an o-minimal structure on the real field. Many objective functions arising in data science are coercive due to the use of regularizers. Some objectives are naturally coercive, such as in symmetric low-rank matrix recovery problems [30, 42]. For functions that are not coercive, our results can still be applied if the iterates are uniformly bounded. We discuss this extension in Remark 2.

Our results rely on the connection between the iterates of first-order methods and the subgradient trajectories of the objective function. The subgradient trajectories of a locally Lipschitz function are solutions to a differential inclusion (i.e., equation (2) with $c = 1$). Previous works used the theory of differential inclusions [3] to study the stochastic subgradient method. Most of them are in the setting where a stochastic subgradient oracle is available (i.e., it generates a subgradient of the objective function in expectation) and the step sizes are diminishing. It was shown that the iterates of stochastic subgradient method converge almost surely to an internally chain transitive set of a differential inclusion [7, Theorem 3.6] [18, Corollary 4], with the proviso that the iterates are bounded almost surely and that the step sizes are not summable but square summable, among other assumptions. By additionally assuming that the objective function is Whitney stratifiable, the iterates subsequentially converge to critical points and the function values converge to a critical value almost surely [23, Corollary 5.9].

In contrast to the above works, we are interested in random reshuffling or

Table 1: Standing assumption: f is coercive and tame.

Algorithm	Our assumption	Literature	Conclusion
Subgradient method with momentum	f locally Lipschitz	f differentiable with locally Lipschitz gradient [73, 54]	$f(x_k)$ and x_k eventually stay arbitrarily close to a critical value and a connected component of the set of critical points respectively, for all initial points in a bounded set X_0 and sufficiently small step sizes α
Random reshuffling with momentum	f_i locally Lipschitz and subdifferentially regular	no results	
Random-permutations cyclic coordinate descent method	f continuously differentiable	f differentiable with locally Lipschitz gradient [5]	

permutation, which includes sampling without replacement, and constant step sizes. The recent work of Bianchi *et al.* [13] uses differential inclusions to analyze the stochastic subgradient method with constant step size. When a stochastic subgradient oracle is available along with other assumptions involving a Markov kernel, the iterates of the stochastic subgradient method eventually lie in the neighborhood of the critical points with high probability [13, Theorem 3] for sufficiently small step sizes. The stochastic subgradient method with random reshuffling and diminishing step sizes was analyzed via differential inclusions recently in the work of Pauwels [57]. For Lipschitz continuous objectives, bounded iterates converge subsequentially to the set of critical points with respect to a conservative field [57, Corollary 6].

We next give the update rules of the first-order methods considered in this paper. Algorithm 1 is a generalization the framework proposed in the work of Kovachki and Stuart [35, (7)] from differentiable functions to locally Lipschitz functions. We denote by ∂f the Clarke subdifferential [20] (see Definition 1) of a locally Lipschitz function f. Algorithm 1 reduces to the heavy ball method [59] when $\gamma = 0$ and to the Nesterov's accelerated subgradient method [50, equation (2.2.22)] when $\beta = \gamma$ respectively. It also includes the vanilla subgradient method as a special case when $\beta = \gamma = 0$. Algorithm 2 is an extension of Algorithm 1 which exploits the composite nature of the objective function (1). Its update is the same as Algorithm 1 except that each step concerns only one component f_i, which is chosen at a random order at every iteration (epoch). This is exactly how stochastic subgradient method with momentum is implemented in practice (see for e.g., documentations from TensorFlow[1], PyTorch[2]

[1] https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/SGD

[2] https://pytorch.org/docs/stable/generated/torch.optim.SGD.html

and scikit-learn[3]). Last, Algorithm 3 is the random-permutations cyclic coordinate descent method, where $\nabla_i f(x) := [\nabla f(x)]_i e_i$, $[\nabla f(x)]_i$ is the ith entry of $\nabla f(x)$, and e_i is the ith canonical base. Similar to Algorithm 2, Algorithm 3 chooses a permutation of all the coordinates at every iteration and cycles through them.

The paper is organized as follows. Section 2 contains a literature review on the first-order methods and stochastic approximations with constant step size. Section 3 contains the global stability results for iterative methods that are approximated by subgradient trajectories. Finally, Section 4 explains how the first-order methods fit into the abstract framework of Section 3.

Algorithm 1 Subgradient method with momentum

```
choose step size $\alpha > 0$, momentum parameters $\beta \in (-1,1)$, $\gamma \in \mathbb{R}$, constant $\delta > 0$, $x_{-1}, x_0 \in \mathbb{R}^n$ with $\|x_{-1} - x_0\| \leqslant \delta\alpha$
for $k = 0, 1, \ldots$ do
    $y_k = x_k + \gamma(x_k - x_{k-1})$
    $x_{k+1} \in x_k + \beta(x_k - x_{k-1}) - \alpha\partial f(y_k)$
end for
```

Algorithm 2 Random reshuffling with momentum

```
choose step size $\alpha > 0$, momentum parameters $\beta \in (-1,1)$, $\gamma \in \mathbb{R}$, constant $\delta > 0$, $x_{-1,N-1}, x_0 \in \mathbb{R}^n$ with $\|x_{-1,N-1} - x_0\| \leqslant \delta\alpha$
for $k = 0, 1, \ldots$ do
    $x_{k,0} = x_k$
    $x_{k,-1} = x_{k-1,N-1}$
    choose a permutation $\sigma^k$ of $\{1, 2, \ldots, N\}$
    for $i = 1, 2, \ldots, N$ do
        $y_{k,i} = x_{k,i-1} + \gamma(x_{k,i-1} - x_{k,i-2})$
        $x_{k,i} \in x_{k,i-1} + \beta(x_{k,i-1} - x_{k,i-2}) - \alpha\partial f_{\sigma_i^k}(y_{k,i})$
    end for
    $x_{k+1} = x_{k,N}$
end for
```

2 Literature review

The gradient method with momentum with $\gamma = 0$ was introduced by Polyak [59]. It admits a nearly optimal local convergence rate for twice continuously differentiable strongly convex functions [59, Theorem 9]. Nesterov showed that it admits a globally

[3] `https://scikit-learn.org/stable/modules/sgd.html`

Algorithm 3 Random-permutations cyclic coordinate descent method

```
**choose** $x_0 \in \mathbb{R}^n$, step size $\alpha > 0$
**for** $k = 0, 1, \ldots$ **do**
    choose a permutation $\sigma^k$ of $\{1, 2, \ldots, n\}$
    $x_{k,0} = x_k$
    **for** $i = 1, 2, \ldots, n$ **do**
        $x_{k,i} = x_{k,i-1} - \alpha \nabla_{\sigma^k_i} f(x_{k,i-1})$
    **end for**
    $x_{k+1} = x_{k,n}$
**end for**
```

optimal convergence rate [50, Theorem 2.1.13] if one chooses $\beta = \gamma$ in an appropriate manner. With variable momentum parameters, it also has an optimal rate for convex functions with Lipschitz gradients whose infimum is attained [51]. If one relaxes the convexity assumption, then with a suitable choice of parameters α, β, and γ, the gradients $\nabla f(x_k)$ converge to zero [73, Lemmas 1,2,3] for any initial points $x_{-1}, x_0 \in \mathbb{R}^n$. If in addition f is coercive and satisfies the Kurdyka-Łojasiewicz inequality [36] at every point and $x_{-1} = x_0$, then the iterates have finite length [54, Theorem 4.9]. In the nonsmooth setting that we consider in this paper, there seems to be no results to the best of our knowledge.

The incremental subgradient method is a special of the stochastic subgradient method with random reshuffling where the components are visited in a fixed order. It can be traced back to the Widrow-Hoff least mean squares method [71] for minimizing a finite sum of convex quadratics in 1960. It was pointed out later by Kohonen that with sufficiently small constant step sizes, the limit points of the iterates of the least mean squares method are close to a minimum of the objective function [34]. With diminishing step sizes that are not summable but square summable, the least mean squares method converges to a minimum of the problem [44]. For convex objectives, the incremental subgradient method with constant step size $\alpha > 0$ satisfies $\liminf_{k\to\infty} f(x_k) \leqslant \inf_{\mathbb{R}^n} f + C\alpha$ for some $C > 0$ [48, Proposition 2.1], provided that $\inf_{\mathbb{R}^n} f > -\infty$ and the subgradients of the components f_i are uniformly bounded. We refer the readers to the survey paper [11], the textbook [10], and references therein for a more detailed discussion on the subject.

The stochastic subgradient method with random reshuffling is a stochastic version of the incremental subgradient method. It was shown recently that the stochastic gradient method with random reshuffling outperforms the incremental gradient method in expectation on strongly convex functions with quadratic components [31, Theorem 2], under certain choices of diminishing step sizes. If the objective function is strongly convex and differentiable with Lipschitz gradients among other assumptions, then the iterates and the corresponding function values of the stochastic gradient method with random reshuffling and constant step size eventually lie in a neighborhood of the minimizer [47, Theorem 1] and a neighborhood of the minimum [52, Theorem 1] respectively, both in expectation. By relaxing the strong convexity assumption to mere convexity, the function values evaluated at the average iterates $\hat{x}_k := (\sum_{l=0}^k x_l)/k$ eventually lie in a neighborhood of the minimum in expectation [47, Theorem 3] [52, Remark 1]. By further removing the convexity assumption, the minimum norm of the gradients eventually lies in a neighborhood of zero in expectation [47, Theorem 4] [57, Corollary 1, Corollary 3] (see also [52, Theorem 4] for a similar result). The long-term behavior of the iterates for nonconvex and nonsmooth objective functions has so far remained elusive.

Despite the empirical success of incorporating momentum into the incremental gradient method/stochastic gradient method with random reshuffling [64], the theoretical understanding of such methods is limited. So far, the only guarantees available are for modified versions [66, 67]. The work of Tran *et al.* [66] in 2021 studied a modified version of stochastic gradient method with random reshuffling and heavy ball. The momentum is constant within every iteration (epoch) and is equal to the average of the gradients evaluated in the previous epoch. With the modification, the norm of gradients of the average iterates $\hat{x}_k$ eventually lie in a neighborhood of zero in expectation [66, Corollary 1], under various assumptions [66, Assumption 1]. A modified stochastic gradient method with random reshuffling and Nesterov's momentum was studied recently [67]. The momentum is only applied at the level of the outer loop, at the end of each iteration (epoch). In this setting, the function values eventually lie a neighborhood of the minimum when the component functions are convex [67, Theorem 1], among other assumptions.

Coordinate descent methods are the object of the survey paper [72] by Wright in 2015. The idea of coordinate descent methods is to optimize with respect to one variable at a time. It was first studied under the framework of univariate relaxation [55, Section 14.6]. With exact line search and almost cyclic rule or Gauss-Southwell rule for cycling over the coordinates, the coordinate descent method converges linearly to a minimizer of a strongly convex objective that is twice differentiable [45, Theorem 2.1]. More recently, global convergence of random coordinate descent method was established for convex objectives with Lipschitz continuous partial derivatives [49]. In contrast to cyclic coordinate descent methods, random coordinate descent methods choose a coordinate randomly at each iteration instead of following a cycling rule. Similar to the stochastic subgradient method with random reshuffling, the random-permutations cyclic coordinate descent method considered in this work is easier to implement than the random coordinate descent method as it requires only sequential access of the data [32]. Using [5, Lemma 3.3, remark 3.2], the convergence of the random-permutations cyclic coordinate descent method can be deduced for coercive functions with locally Lipschitz gradients. The superior performance of the random-permutations cyclic coordinate descent method was observed in numerical experiments, and was supported by analysis for convex quadratic objectives [41, 32]. For objective functions without a locally Lipschitz gradient, the study of the method appears to be absent from the literature.

Stochastic approximations of differential inclusions with constant step size have led to recent advances on an oracle-based stochastic subgradient method [13]. Given differential equations with Lipschitz right-hand sides over a finite time horizon, Kurtz proposed a sequence of discrete time stochastic processes that approaches their solutions with a probability that goes to one [37, Theorem (4.7)] (see also [6, Proposition 3.1]). Over an infinite time horizon, the sequence of corresponding invariant measures concentrates around the Birkhoff center of the differential equations [6, Corollary 3.2]. Later, Roth and Sandholm [62] extended these results to differential inclusions with upper semicontinuous right-hand sides and compact supports, along with other assumptions. More recently, the work of Bianchi *et al.* [12] studied stochastic approximation with constant step size under a different set of assumptions, relaxing the compact support assumption from the previous literature. We refer the readers to the textbooks on Markov processes and stochastic approximations [27, 9] for more references on the subject. Although the above results cannot be directly applied to the settings of this work, readers will see that we adopt similar proof strategies when studying the relationship between the discrete and continuous dynamics. For example, we also study the subsequential convergence of the linear interpolation of the iterates to the solutions of a continuous-time system. More discussion on this matter is deferred to the following section in Remark 1. In addition, our analysis relies on the theory of set-valued analysis [20] and differential inclusion [3], which was also used in the aforementioned literature.

3 Global stability of first-order methods

We refer to an iterative method with constant step size as a set-valued mapping $\mathcal{M} : \mathbb{R}^{(\mathbb{R}^n)} \times (0,\infty) \times 2^{(\mathbb{R}^n)} \times \mathbb{N} \rightrightarrows (\mathbb{R}^n)^{\mathbb{N}}$ which, to an objective function $f : \mathbb{R}^n \to \mathbb{R}$, a constant step size $\alpha \in (0,\infty)$, a set $X_0 \subset \mathbb{R}^n$, and a natural number $\bar{k}$ associates a set of sequences in $\mathbb{R}^n$ whose $\bar{k}$th term is contained in X_0.

We next introduce several definitions. Let $\|\cdot\|$ be the induced norm of an inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^n$. Given a subset S of $\mathbb{R}^n$ and $x \in \mathbb{R}^n$, consider the distance of x to S defined by $d(x,S) := \inf\{\|x-y\| : y \in S\}$. Let $B(a,r)$ denote the closed ball of center $a \in \mathbb{R}^n$ and radius $r > 0$, and let $B(S,r) := \cup_{a\in S} B(a,r)$ where $S \subset \mathbb{R}^n$. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz if for all $a \in \mathbb{R}^n$, there exist $r > 0$ and $L > 0$ such that $\|f(x) - f(y)\| \leqslant L\|x-y\|$ for all $x, y \in B(a,r)$. We use $[f \leqslant \Delta] := \{x \in \mathbb{R}^n : f(x) \leqslant \Delta\}$ to denote a sublevel set of a function $f : \mathbb{R}^n \to \mathbb{R}$ where $\Delta \in \mathbb{R}$. A function $f : \mathbb{R}^n \to \mathbb{R}$ is coercive if $\lim_{\|x\|\to\infty} f(x) = \infty$.

Definition 1. *[20, Chapter 2] Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function. The Clarke subdifferential is the set-valued mapping $\partial f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined for all $x \in \mathbb{R}^n$*

by $\partial f(x) := \{s \in \mathbb{R}^n : f^\circ(x,d) \geqslant \langle s, d\rangle, \forall d \in \mathbb{R}^n\}$ *where*

$$f^\circ(x,d) := \limsup_{\substack{y \to x \\ t \searrow 0}} \frac{f(y+td) - f(y)}{t}.$$

We say that $x \in \mathbb{R}^n$ is critical if $0 \in \partial f(x)$, and that $v \in \mathbb{R}$ is a critical value if there exists $x \in \mathbb{R}^n$ such that $0 \in \partial f(x)$ and $v = f(x)$. If f is continuously differentiable, then $\partial f(x) = \{\nabla f(x)\}$ [20, 2.2.4 Proposition].

Definition 2. *[3, Definition 1 p. 12] Given some real numbers a and b such that $a < b$, a function $x(\cdot)$ defined from $[a,b]$ to $\mathbb{R}^n$ is absolutely continuous if for all $\epsilon > 0$, there exists $\delta > 0$ such that, for any finite collection of disjoint subintervals $[a_1,b_1],\ldots,[a_m,b_m]$ of $[a,b]$ such that $\sum_{i=1}^m b_i - a_i \leqslant \delta$, we have $\sum_{i=1}^m \|x(b_i) - x(a_i)\| \leqslant \epsilon$.*

By virtue of [53, Theorem 20.8], a function $x : [a,b] \to \mathbb{R}^n$ is absolutely continuous if and only if it is differentiable almost everywhere on (a,b), its derivative $x'(\cdot)$ is Lebesgue integrable, and $\forall t \in [a,b], x(t) - x(a) = \int_a^t x'(t)dt$.

The next definition is inspired by a series of works that resort to continuous-time dynamics to analyze discrete-time dynamics. The idea that discrete dynamics resemble their continuous counterpart dates back to Euler [28, 14]. He proposed discretizing ordinary differential equations to find approximate solutions. This technique is also used to prove the existence of solutions via the Cauchy Peano theorem [21, Theorem 1.2]. Ljung [43] and Kushner [39, 38] established a connection between the asymptotic behavior of discrete and continuous dynamics with noise, which is particularly useful when they are governed by conservative fields. Benaïm *et al.* [7, 8] strengthened this connection by relaxing some assumptions and incorporating set-valued dynamics. Due to its importance in analyzing optimization algorithms in recent years [18, 26, 23, 17, 63, 57], we elaborate on their contribution.

Benaïm, Hofbauer, and Sorin consider a closed set-valued mapping $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ with nonempty convex compact values for which there exists $C > 0$ such that $\sup\{\|s\| : s \in F(x)\} \leqslant C(1 + \|x\|)$ for all $x \in \mathbb{R}^n$. They show that discrete trajectories of F can be approximated by its continuous trajectories in the following sense. Let $(x_k)_{k\in\mathbb{N}}$ be a bounded sequence such that $x_{k+1} \in x_k + \alpha_k F(x_k)$ for all $k \in \mathbb{N}$ where $\alpha_k > 0$, $\sum_{k=0}^\infty \alpha_k = \infty$, and $\sum_{k=0}^\infty \alpha_k^2 < \infty$ (Ljung and Kushner also assume this, following Robbins and Monro [61]). Let $t_0 := 0$ and $t_k := \alpha_0 + \ldots + \alpha_{k-1}$ for $k \geqslant 1$. Consider the linear interpolation defined by

$$x(t) := x_k + \frac{t - t_k}{t_{k+1} - t_k}(x_{k+1} - x_k), \quad \forall t \in [t_k, t_{k+1})$$

as well as the time shifted interpolations $x^\tau(\cdot) := x(\tau + \cdot)$ where $\tau \geqslant 0$. The key insight is that for any sequence $\tau_k \to \infty$, the shifted interpolations x^{τ_k} subsequentially converge to a solution to the differential inclusion $x'(t) \in F(x(t))$ for almost every $t > 0$ in the topology of uniform convergence on compact intervals [7, Theorem 4.2]. When one specifies that $F := -\partial f$ where $f : \mathbb{R}^n \to \mathbb{R}$ is a locally Lipschitz function, this can used to derive asymptotic properties of some important algorithms in optimization.

In this work, we are interested in the constant step size regime for which it is hopeless to try to establish uniform convergence over compact intervals (for a fixed discrete trajectory, as above). We thus ask for something weaker from the algorithms we analyze, namely, that the continuous and discrete time dynamics are close in uniform norm up to a certain time. We ask that this holds for a set of time shifted trajectories to account for multistep methods. We next give a precise meaning to this notion. We will use $\lfloor t \rfloor$ to denote the floor of a real number t which is the unique integer such that $\lfloor t \rfloor \leqslant t < \lfloor t \rfloor + 1$.

Definition 3. *An iterative method $\mathcal{M}$ is approximated by subgradient trajectories of a locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ (up to a positive multiplicative constant) if there exists $c > 0$ such that for any compact sets $X_0, X_1 \subset \mathbb{R}^n$, there exists $T > 0$ such that for all $\epsilon > 0$, there exists $\bar{\alpha} > 0$ such that for all $\alpha \in (0, \bar{\alpha}]$, $\bar{k} \in \mathbb{N}$, and $(x_k)_{k \in \mathbb{N}} \in \mathcal{M}(f, \alpha, X_0, \bar{k})$ for which $x_0, \ldots, x_{\bar{k}} \in X_1$, there exists an absolutely continuous function $x : [0, T] \to \mathbb{R}^n$ such that*

$$x'(t) \in -c\partial f(x(t)), \quad \text{for almost every } t \in [0, T], \quad x(0) \in X_0, \tag{2}$$

and $\|x_k - x((k - \bar{k})\alpha)\| \leqslant \epsilon$ for $k = \bar{k}, \ldots, \bar{k} + \lfloor T/\alpha \rfloor$.

Remark 1. *In the next section, we show that Algorithms 1, 2, and 3 satisfy Definition 3 (see Propositions 1 and 2). In order to do so, we always use the same strategy which consists in taking sequences generated by a given method with smaller and smaller constant step size, and show that a subsequence of their linear interpolations converges uniformly to a subgradient trajectory up to a finite time.*

Several discretization methods of initial value problems with differential inclusions were studied in [65, 3, 20, 29] (see also a survey on the subject by Dontchev and Lempio [24]). Assume that the set-valued mapping underlying the differential inclusion is upper semicontinuous with nonempty compact convex values, such that the norm of their elements are upper bounded by a linear function of the norm of the argument. Then over any finite time horizon, a subsequence of linear interpolations of the Euler method with smaller and smaller step sizes converges uniformly to a solution to the initial value problem [24, Theorem 2.2]. If in addition the set-valued mapping is bounded, then a class of linear multistep methods has the same convergence property as above [65, p. 127, Theorem] (see also [24, Convergence Theorem 3.2]). We build on the techniques developed in the above works when checking Definition 3. We adapt them so that they can handle the case where ∂f is not accessible (as in Algorithms 2 and 3) and the set of initial points is a compact set.

The class of locally Lipschitz functions is too broad to obtain any meaningful results on the first-order methods [60, 22]. We thus consider functions that are tame in o-minimal structures. O-minimal structures (short for order-minimal) were originally considered by van den Dries, Pillay and Steinhorn [68, 58]. They are founded on the observation that many properties of semi-algebraic sets can be deduced from a few simple axioms [69]. Recall that a subset A of $\mathbb{R}^n$ is semi-algebraic [15] if it is a finite union of basic semi-algebraic sets, which are of the form $\{x \in \mathbb{R}^n : p_i(x) = 0,\ i = 1, \ldots, k;\ p_i(x) > 0,\ i = k+1, \ldots, m\}$ where $p_1, \ldots, p_m \in \mathbb{R}[X_1, \ldots, X_n]$ (i.e., polynomials with real coefficients).

Definition 4. *[70, Definition p. 503-506] 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 $\{(x_1, \ldots, x_k) \in \mathbb{R}^k : x_i = x_j\}$ for $1 \leqslant i < j \leqslant k$;*
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 $\pi : \mathbb{R}^{k+1} \to \mathbb{R}^k$ 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.*

Note that S_1 are the semi-algebraic subsets of $\mathbb{R}$ and by [70, 2.5 Examples (3)], S_k contains the semi-algebraic subsets of $\mathbb{R}^k$. 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 $f : \mathbb{R}^n \to \mathbb{R}$ is definable in an o-minimal structure if its graph, that is to say $\{(x,t) \in \mathbb{R}^{n+1} : f(x) = t\}$, is definable in that structure. A set $C \subset \mathbb{R}^n$ is tame [33] in an o-minimal structure $(S_k)_{k \in \mathbb{N}}$ if

$$\forall x \in \mathbb{R}^n,\ \forall r > 0, \quad C \cap B(x,r) \in S_n.$$

and a function $f : \mathbb{R}^n \to \mathbb{R}$ is tame if its graph is tame. With the above definitions, we are now ready to state two technical lemmas. The first relates a uniform neighborhood of a sublevel set with another sublevel set. The second is analogous to the descent lemma for smooth functions [50, Lemma 1.2.3] [4, Lemma 5.7].

Lemma 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function. Let $\Delta \in \mathbb{R}$ and let $L > 0$ be a Lipschitz constant of f in $[f \leqslant \Delta]$. For any $\epsilon' > 0$, $B([f \leqslant \Delta - \epsilon' L], \epsilon') \subset [f \leqslant \Delta]$.*

Proof. We show that $B(a, \epsilon') \subset [f \leqslant \Delta]$ for all $a \in [f \leqslant \Delta - \epsilon' L]$. Indeed, if $b \in B(a, \epsilon') \setminus [f \leqslant \Delta]$, then there exists c in the segment $[a, b)$ such that $f(c) = \Delta$ and $\epsilon' L = \Delta - (\Delta - \epsilon' L) \leqslant f(c) - f(a) \leqslant L\|c - a\| < \epsilon' L$. □

Lemma 2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz tame function and let $\mathcal{M}$ be an iterative method with constant step size. Let $X \subset \mathbb{R}^n$ and L be a Lipschitz constant of f on X. For all $T, \epsilon', \alpha, c > 0$, $\bar{k} \in \mathbb{N}$, $(x_k)_{k\in\mathbb{N}} \in (\mathbb{R}^n)^{\mathbb{N}}$, and for any subgradient trajectory $x : [0,T] \to \mathbb{R}^n$ of cf such that $x([0,T]) \subset X$, $x_k \in X$, and $\|x_k - x(\alpha(k-\bar{k}))\| \leqslant \epsilon'$ for $k = \bar{k}, \ldots, \bar{k} + \lfloor T/\alpha \rfloor$, we have*

$$f(x_k) \leqslant f(x((k-\bar{k})\alpha)) + \epsilon' L \leqslant f(x_{\bar{k}}) - c\int_0^{(k-\bar{k})\alpha} d(0, \partial f(x(s)))^2 \, ds + 2\epsilon' L$$

for $k = \bar{k}, \ldots, \bar{k} + \lfloor T/\alpha \rfloor$.

Proof. For $k = \bar{k}, \ldots, \bar{k} + \lfloor T/\alpha \rfloor$, we have

$$f(x_k) \leqslant f(x((k-\bar{k})\alpha)) + \epsilon' L \tag{3a}$$

$$= f(x(0)) - (f(x(0)) - f(x((k-\bar{k})\alpha))) + \epsilon' L \tag{3b}$$

$$\leqslant f(x_0) - (f(x(0)) - f(x((k-\bar{k})\alpha))) + 2\epsilon' L \tag{3c}$$

$$= f(x_0) - c\int_0^{(k-\bar{k})\alpha} d(0, \partial f(x(s)))^2 \, ds + 2\epsilon' L. \tag{3d}$$

In (3a) and (3c), we invoke the Lipschitz constant L of f on $X \ni x((k-\bar{k})\alpha), x_k$. (3d) is a consequence of [23, Lemma 5.2, Theorem 5.8] (see also [25]). □

We now turn to our main results, namely Theorem 1 and Corollary 1, which we prove using Lemmas 1 and 2.

Theorem 1 (Stability of function values)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz coercive tame function and let $\mathcal{M}$ be an iterative method with constant step size that is approximated by subgradient trajectories of f. For any bounded set $X_0 \subset \mathbb{R}^n$ and $\epsilon > 0$, there exist $\bar{\alpha}, \Delta > 0$ such that for all $(x_k)_{k\in\mathbb{N}} \in \mathcal{M}(f, (0, \bar{\alpha}], X_0, 0)$, we have $f(x_k) \leqslant \Delta$ for all $k \in \mathbb{N}$ and there exist a critical value f^* of f and $k_0 \in \mathbb{N}$ such that $|f(x_k) - f^*| \leqslant \epsilon$ for all $k \geqslant k_0$.*

Proof. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz coercive tame function. Since f is tame and coercive, there exists $\Delta > 0$ such that $X_0 \subset [f \leqslant \Delta/2]$ and Δ is not a critical value of f. By the definable Morse-Sard theorem [16, Corollary 9], f has finitely many critical values $f_1 > \cdots > f_p$ in $[f \leqslant \Delta]$ (and it has at least one since f is coercive and continuous). Since f is coercive and continuous, the compact sublevel sets $[|f - f_i| \leqslant \epsilon]$, $i = 1, \ldots, p$, are pairwise disjoint after possible reducing ϵ, which we may do without loss of generality. We may also assume that $f_1 + 2\epsilon \leqslant \Delta$. According to the Kurdyka-Łojasiewicz inequality [16, Theorem 14] (see also [2, Theorem 4.1]) and the monotonicity theorem [69, (1.2) p. 43] [36, Lemma 2], there exist $\rho > 0$ and a strictly increasing concave continuous definable function $\psi : [0, \rho) \to [0, \infty)$ that is continuously differentiable on $(0, \rho)$ with $\psi(0) = 0$ such that $d(0, \partial f(x)) \geqslant 1/\psi'(|f(x) - f_i|)$ for all $x \in [|f - f_i| \leqslant \epsilon]$ whenever $0 < |f(x) - f_i| < \rho$ for $i = 1, \ldots, p$. Without loss of generality, assume $\epsilon < \rho$ so that $d(0, \partial f(x)) \geqslant 1/\psi'(|f(x) - f_i|)$ for all $x \in [|f - f_i| \leqslant \epsilon]$ such that $f(x) \neq f_i$.

Consider a Lipschitz constant $L \geqslant 1$ of f in $[f \leqslant \Delta]$ and the quantity

$$M := \inf\{d(0, \partial f(x)) : |f(x) - f_i| \geqslant \epsilon/2, \ i = 1, \ldots, p, \ f(x) \leqslant \Delta\} > 0. \tag{4}$$

Since $\mathcal{M}$ is approximated by subgradient trajectories of f, by Definition 3 there exist $c, T > 0$, and $\bar{\alpha} \in (0, T/2)$ such that such that for all $\alpha \in (0, \bar{\alpha}]$, $\bar{k} \in \mathbb{N}$, and $(x_k)_{k \in \mathbb{N}} \in \mathcal{M}(f, \alpha, [f \leqslant \Delta/2], \bar{k})$ for which $x_0, \ldots, x_{\bar{k}} \in [f \leqslant \Delta]$, there exists a subgradient trajectory $x : [0, T] \to \mathbb{R}^n$ of f up to the multiplicative constant c for which $x(0) \in [f \leqslant \Delta/2]$ and $\|x_k - x(\alpha(k - \bar{k}))\| \leqslant \epsilon'$ for $k = \bar{k}, \ldots, \bar{k} + \lfloor T/\alpha \rfloor$ where

$$\epsilon' := \min\left\{\frac{\Delta}{4L}, \frac{cM^2T}{24L}, \frac{\epsilon}{8L}, \frac{cT}{2L\psi'(\epsilon/2)^2}\right\} > 0.$$

Since $[|f - f_1| \leqslant \epsilon], \ldots, [|f - f_p| \leqslant \epsilon]$ are compact, after possibly reducing T and $\bar{\alpha}$ the statement still holds if one replaces the initial set $[f \leqslant \Delta/2]$ by $[|f - f_1| \leqslant \epsilon], [|f - f_2| \leqslant \epsilon], \ldots,$ or $[|f - f_p| \leqslant \epsilon]$.

From now on, we fix a constant step size $\alpha \in (0, \bar{\alpha}]$. Consider a sequence $(x_k)_{k \in \mathbb{N}} \in \mathcal{M}(f, \alpha, X_0, 0) \subset \mathcal{M}(f, \alpha, [f \leqslant \Delta/2], 0)$ along with an associated subgradient trajectory $x : [0, T] \to \mathbb{R}^n$ of f up to the multiplicative constant c for which $x(0) \in [f \leqslant \Delta/2] \subset [f \leqslant \Delta - \epsilon' L]$ and $\|x_k - x(\alpha(k - \bar{k}))\| \leqslant \epsilon'$ for $k = \bar{k}, \ldots, \bar{k} + K$ where $\bar{k} = 0$ and $K := \lfloor T/\alpha \rfloor$. By Lemmas 1 and 2, for $k = 0, \ldots, K$, we have $f(x_k) \leqslant f(x(k\alpha)) + \epsilon' L \leqslant f(x(0)) + \epsilon' L \leqslant \Delta/2 + \epsilon' L \leqslant \Delta$ and

$$f(x_k) \leqslant f(x_0) - c \int_0^{k\alpha} d(0, \partial f(x(s)))^2 \, ds + 2\epsilon' L. \tag{5}$$

If $c \int_0^{K\alpha} d(0, \partial f(x(s)))^2 \, ds \geqslant 3\epsilon' L$, then we have $f(x_K) \leqslant f(x_0) - 3\epsilon' L + 2\epsilon' L \leqslant \Delta/2$ so that we may apply Lemmas 1 and 2 again with $\bar{k} = K$. Since the continuous function f is bounded below on the compact set $[f \leqslant \Delta/2]$, this process with constant decrease can only be repeated finitely many times. Thus there exist $v \in \mathbb{N}$ and an absolutely continuous function (again denoted $x(\cdot)$) such that $f(x_k) \leqslant f(x_{vK}) - c \int_0^{(k - vK)\alpha} d(0, \partial f(x(s)))^2 \, ds + 2\epsilon' L$ and $\|x_k - x(\alpha(k - vK))\| \leqslant \epsilon'$ for $k = vK, \ldots, (v+1)K$ where $c \int_0^{K\alpha} d(0, \partial f(x(s)))^2 \, ds < 3\epsilon' L$. Hence there exists $t' \in [0, K\alpha]$ such that $d(0, \partial f(x(t')))^2 \leqslant 3\epsilon' L/(cK\alpha) \leqslant 3\epsilon' L/(cT/2) \leqslant M^2/4$, where we use the fact that $\epsilon' \leqslant cM^2T/(24L)$. Since $d(0, \partial f(x(t'))) \leqslant M/2$ and $f(x(t')) \leqslant \Delta$, by definition of M in (4) there exists $i \in \{1, \ldots, p\}$ such that $|f(x(t')) - f_i| < \epsilon/2$. We also have that $f(x(t')) \leqslant f(x(0)) \leqslant f(x_{vK}) + \epsilon' L \leqslant \Delta/2 + \epsilon' L$. Thus $f_i < f(x(t')) + \epsilon/2 \leqslant \Delta/2 + \epsilon' L + \epsilon/2 \leqslant \Delta/2 + 3\epsilon/8$. For $k' = vK, \ldots, (v+1)K$, we have

$$|f(x_{k'}) - f_i| \leqslant |f(x_{k'}) - f(x(\alpha(k' - vK)))| + |f(x(\alpha(k' - vK))) - f(x(t'))| + \tag{6a}$$

$$\begin{aligned}
&|f(x(t')) - f_i| \tag{6b}\\
\leqslant & L\|x_{k'} - x(\alpha(k'-vK))\| + |f(x(0)) - f(x(K\alpha))| + \epsilon/4 \tag{6c}\\
\leqslant & L\epsilon' + 3\epsilon' L + \epsilon/2 \tag{6d}\\
\leqslant & \epsilon/8 + 3\epsilon/8 + \epsilon/2 \tag{6e}\\
= & \epsilon. \tag{6f}
\end{aligned}$$

Indeed, (6a) is due to the triangular inequality. We invoke the Lipschitz constant L of f on $[f \leqslant \Delta]$ in order to bound the first term in (6a). In order to bound the second term in (6a), we use the fact that the composition $f \circ x$ is decreasing and $0 \leqslant \alpha(k'-vK) \leqslant t' \leqslant K\alpha$. (6d) holds because $\|x_{k'} - x(\alpha(k'-vK))\| \leqslant \epsilon'$ and $|f(x(0)) - f(x(K\alpha))| = c\int_0^{K\alpha} d(0,\partial f(x(s)))^2 \; ds < 3\epsilon' L$. (6e) is due to $\epsilon' \leqslant \epsilon/(8L)$.

We next show that $f(x_k) \leqslant f_i + \epsilon$ for all $k \geqslant k' := vK$. Without loss of generality, we assume that $k' = 0$ so that by (6) we have $f(x_k) \leqslant f_i + \epsilon$ for $k = 0, \ldots, K$. We prove that $f(x_{K+1}) \leqslant f_i + \epsilon$, hence $f(x_k) \leqslant f_i + \epsilon$ for all $k \geqslant k'$ by induction. We distinguish two cases. If $f(x_1) < f_i - \epsilon$, then $f(x_{K+1}) \leqslant f(x_1) + 2\epsilon' L < f_i - \epsilon + \epsilon/4 \leqslant f_i + \epsilon$, where the first inequality follows from $x_1 \in [f \leqslant f_i - \epsilon] \subset [f \leqslant \Delta/2 + 3\epsilon/8 - \epsilon] \subset [f \leqslant \Delta/2]$ and Lemmas 1 and 2. If $x_1 \in [|f - f_i| \leqslant \epsilon]$, then let $x : [0,T] \to \mathbb{R}^n$ be an associated subgradient trajectory of f up to the multiplicative constant c such that $\|x_k - x(\alpha(k-1))\| \leqslant \epsilon'$ for $k = 1, \ldots, K+1$ and $x(0) \in [|f - f_i| \leqslant \epsilon]$. Note that for any $t \in [0, K\alpha]$, $f(x(K\alpha)) \leqslant f(x(t)) \leqslant f(x(0)) \leqslant f_i + \epsilon \leqslant \Delta - \epsilon \leqslant \Delta - \epsilon' L$. By Lemma 2, we have that $f(x_{K+1}) \leqslant f(x(K\alpha)) + \epsilon' L < f_i + \epsilon/2 + \epsilon/8 \leqslant f_i + \epsilon$, as desired. Otherwise, we have $f(x(t)) \in [f_i + \epsilon/2, f_i + \epsilon]$ for all $t \in [0, K\alpha]$. By the Kurdyka-Łojasiewicz inequality, we have $d(0, \partial f(x(t))) \geqslant 1/\psi'(f(x(t)) - f_i) \geqslant 1/\psi'(\epsilon/2) > 0$. According to [23, Lemma 5.2, Theorem 5.8] (see also [25]), it holds that

$$\begin{aligned}
f(x(K\alpha)) - f_i &\leqslant f(x(0)) - f_i - c\int_0^{K\alpha} d(0, \partial f(x(s)))^2 \; ds \tag{7a}\\
&\leqslant f(x(0)) - f_i - cK\alpha/\psi'(\epsilon/2)^2 \tag{7b}\\
&\leqslant f(x(0)) - f_i - cT/(2\psi'(\epsilon/2)^2) \tag{7c}\\
&\leqslant \epsilon - cT/(2\psi'(\epsilon/2)^2). \tag{7d}
\end{aligned}$$

Thus $f(x_{K+1}) - f_i \leqslant f(x(K\alpha)) - f_i + f(x_{K+1}) - f(x(K\alpha)) \leqslant \epsilon - cT/(2\psi'(\epsilon/2)^2) + \epsilon' L \leqslant \epsilon$, where we used the fact that $\epsilon' \leqslant (cT)/(2L\psi'(\epsilon/2)^2)$.

If $|f(x_k) - f_i| \leqslant \epsilon$ for all $k \geqslant k'$, then the conclusion of the theorem follows. Otherwise, there exists $\hat{k} \geqslant k'$ such that $f(x_{\hat{k}}) < f_i - \epsilon \leqslant \Delta/2 + 3\epsilon/8 - \epsilon \leqslant \Delta/2$. Following the same argument as in the paragraph below (5), there exists $v' \in \mathbb{N}$ and an absolutely continuous function (again denoted $x(\cdot)$) such that $f(x_k) \leqslant f(x_{v'K}) - c\int_0^{(k-\hat{k})\alpha} d(0, \partial f(x(s)))^2 \; ds + 2\epsilon' L$ and $\|x_k - x(\alpha(k - v'K))\| \leqslant \epsilon'$ for $k = \hat{k} + v'K, \ldots, \hat{k} + (v'+1)K$ where $c\int_0^{K\alpha} d(0, \partial f(x(s)))^2 \; ds < 3\epsilon' L$. As before, it follows that there exist $t'' \in [0,T]$ and $j \in \{1,2,\ldots,p\}$ such that $|f(x(t''))-f_j| \leqslant \epsilon/2$. Since $f(x(t'')) \leqslant f(x(0)) \leqslant f(x_{\hat{k}+v'K}) + \epsilon' L \leqslant f(x_{\hat{k}}) + 3\epsilon' L < f_i - \epsilon + 3\epsilon/8 = f_i - 5\epsilon/8$, it holds that $f_j < f_i$. Replicating (6a)-(6e), we get $|f(x_{k''}) - f_j| \leqslant \epsilon$ for $k'' = \hat{k}+v'K,\ldots,\hat{k}+(v'+1)K$. By the same argument as in the previous paragraph, we have $f(x_k) \leqslant f_j + \epsilon$ for all $k \geqslant k'' := \hat{k} + (v'+1)K$. Since f only has finitely many critical values, the conclusion of the theorem follows. □

Theorem 1 gives a "weak convergence" result, in the sense that the function values evaluated at the iterates eventually stabilize around some critical value. In fact, a "strong convergence" result regarding the distance between the iterates and the set of critical points can be obtained without any additional assumptions. This is the subject of the following corollary. Note that while Corollary 1 implies Theorem 1, it is not clear how to prove Corollary 1 without Theorem 1.

Corollary 1 (Stability of iterates)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz coercive tame function and let $\mathcal{M}$ be an iterative method with constant step size that is approximated by subgradient trajectories of f. For any bounded set $X_0 \subset \mathbb{R}^n$ and $\epsilon > 0$, there exists $\bar{\alpha} > 0$ such that for all $(x_k)_{k\in\mathbb{N}} \in \mathcal{M}(f,(0,\bar{\alpha}],X_0,0)$, there exist a connected component C of the set of critical points of f and $k_0 \in \mathbb{N}$ such that $d(x_k, C) \leqslant \epsilon$ for all $k \geqslant k_0$.*

Proof. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz coercive tame function and let $\mathcal{M}$ be an iterative method with constant step size that is approximated by subgradient trajectories of f. Let X_0 be a bounded subset of $\mathbb{R}^n$ and let $\epsilon > 0$. By Theorem 1, there exists $\alpha_1, \Delta > 0$ such that for all $(x_k)_{k\in\mathbb{N}} \in \mathcal{M}(f,(0,\alpha_1],X_0,0)$, $f(x_k) \leqslant \Delta$ for all $k \in \mathbb{N}$.

Let L denote a Lipschitz constant of f on the compact set $[f \leqslant \Delta]$ and consider the quantity

$$M := \inf\{d(0,\partial f(x)) : d(x,S) \geqslant \epsilon/2, f(x) \leqslant \Delta\} > 0, \tag{8}$$

where S is the set of critical points of f. Since $\mathcal{M}$ is approximated by subgradient trajectories of f, by Definition 3 there exist $c,T > 0$, and $\alpha_2 \in (0,\alpha_1]$ such that for all $\alpha \in (0,\alpha_2], \bar{k} \in \mathbb{N}$ and $(x_k)_{k\in\mathbb{N}} \in \mathcal{M}(f,\alpha,[f \leqslant \Delta],\bar{k})$ for which $x_0,\ldots,x_{\bar{k}} \in [f \leqslant \Delta]$, there exists a subgradient trajectory $x : [0,T] \to \mathbb{R}^n$ of f up to the multiplicative constant c for which $x(0) \in [f \leqslant \Delta]$ and $\|x_k - x((k-\bar{k})\alpha)\| \leqslant \epsilon'$ for $k = \bar{k},\ldots,\bar{k} + \lfloor T/\alpha \rfloor$ where $\epsilon' := \min\{\epsilon/4, cM^2T/(16(1+L)), \epsilon^2/(32(1+L)cT)\}$. Again by Theorem 1 there exists $\alpha_3 \in (0,\alpha_2]$ such that for all $(x_k)_{k\in\mathbb{N}} \in \mathcal{M}(f,(0,\alpha_3],X_0,0)$, there exist a critical value f^* of f and $k_0 \in \mathbb{N}$ such that $|f(x_k) - f^*| \leqslant \epsilon'$ for all $k \geqslant k_0$. Let $\bar{\alpha} := \min\{\alpha_3, \epsilon'/(2c(1+L)), T/2\}$.

Let $\alpha \in (0,\bar{\alpha}]$, $(x_k)_{k\in\mathbb{N}} \in \mathcal{M}(f,\alpha,X_0,0)$, and fix a corresponding f^* and k_0. We fix some $k \geqslant k_0$ from now on and show that $d(x_k,S) \leqslant \epsilon$. Since $(x_{k'})_{k'\in\mathbb{N}} \in \mathcal{M}(f,\alpha,[f \leqslant \Delta],k)$ and $(x_{k'})_{k'\in\mathbb{N}} \subset [f \leqslant \Delta]$, there exists a subgradient trajectory

$x : [0,T] \to \mathbb{R}^n$ of f up to the multiplicative constant c for which $x(0) \in [f \leqslant \Delta]$ and $\|x_{k'} - x(\alpha(k'-k))\| \leqslant \epsilon'$ for $k' = k, \ldots, k+K$ where $K := \lfloor T/\alpha \rfloor$. By Lemma 2, we have

$$c \int_0^{K\alpha} d(0, \partial f(x(s)))^2 \, ds \leqslant f(x_k) - f(x_{k+K}) + 2\epsilon' L \leqslant 2\epsilon'(1+L). \tag{9}$$

Thus there exists $t' \in [0, K\alpha]$ such that $d(0, \partial f(x(t')))^2 \leqslant 2\epsilon'(1+L)/(cK\alpha) \leqslant 2\epsilon'(1+L)/(cT/2) \leqslant M^2/4$, where we use the fact that $\epsilon' \leqslant cM^2T/(16(1+L))$. As $f(x(t')) \leqslant f(x(0)) \leqslant \Delta$, we have $d(x(t'), S) \leqslant \epsilon/2$. It now suffices to show that $\|x_k - x(t')\| \leqslant \epsilon/2$. Notice that $\|x_k - x(0)\| \leqslant \epsilon' \leqslant \epsilon/4$ and

$$\begin{align}
\|x(0) - x(t')\| &\leqslant \int_0^{t'} \|x'(s)\| \, ds \tag{10a}\\
&= \int_0^{t'} c \, d(0, \partial f(x(s))) \, ds \tag{10b}\\
&\leqslant \sqrt{\int_0^{t'} c \, ds} \sqrt{\int_0^{t'} c \, d(0, \partial f(x(s)))^2 \, ds} \tag{10c}\\
&\leqslant \sqrt{\int_0^{T} c \, ds} \sqrt{\int_0^{K\alpha} c \, d(0, \partial f(x(s)))^2 \, ds} \tag{10d}\\
&\leqslant \sqrt{cT} \sqrt{2\epsilon'(1+L)} \tag{10e}\\
&\leqslant \epsilon/4. \tag{10f}
\end{align}$$

Indeed, (10a) is due to triangular inequality. (10b) is a consequence of (2) and [23, Lemma 5.2, Theorem 5.8] (see also [25]). (10c) is due to the Cauchy-Schwarz inequality. (10f) is due $\epsilon' \leqslant \epsilon^2/(32(1+L)cT)$. Summing up, we have $|d(x_k, S) - d(x(t'), S)| \leqslant \|x_k - x(t')\| \leqslant \|x_k - x(0)\| + \|x(0) - x(t')\| \leqslant \epsilon/2$ and thus $d(x_k, S) \leqslant \epsilon$.

We have just shown that for all $d(x_k, S) \leqslant \epsilon$ for all $k \geqslant k_0$. Since $x_k \in [f \leqslant \Delta]$, we have that $d(x_k, S') = d(x_k, S) \leqslant \epsilon$, where $S' := S \cap B([f \leqslant \Delta], 2\epsilon)$. By the cell decomposition theorem [69, (2.11) p. 52], the definable compact set S' has finitely many compact connected components $C_1, \ldots, C_q$. Thus for each $k \geqslant k_0$, there exists $i_k \in \{1, \ldots, q\}$ such that $d(x_k, C_{i_k}) \leqslant \epsilon$. We next show that $d(x_{k+1}, C_{i_k}) \leqslant \epsilon$, so that i_k can actually be chosen independently of k. Naturally, we have $d(C_i, C_j) := \inf\{\|x - y\| : (x,y) \in C_i \times C_j\} > 0$ for all $i \neq j$, otherwise $C_i \cap C_j \neq \emptyset$. Without loss of generality, we may assume that $\epsilon \leqslant \min\{d(C_i, C_j) : i \neq j\}/4$. It follows that, for all $j \neq i_k$, we have $d(x_k, C_j) \geqslant d(C_{i_k}, C_j) - d(x_k, C_{i_k}) \geqslant 4\epsilon - \epsilon = 3\epsilon$. Similar to (10a)-(10f), we have $\|x(0) - x(\alpha)\| \leqslant \sqrt{c\alpha}\sqrt{2\epsilon'(1+L)} \leqslant \epsilon'$ since $\alpha \leqslant \bar{\alpha} \leqslant \epsilon'/(2c(1+L))$. Thus $\|x_{k+1} - x_k\| \leqslant \|x_{k+1} - x(\alpha)\| + \|x(\alpha) - x(0)\| + \|x(0) - x_k\| \leqslant 3\epsilon' \leqslant \epsilon$. Hence $d(x_{k+1}, C_j) \geqslant d(x_k, C_j) - \|x_k - x_{k+1}\| \geqslant 3\epsilon - \epsilon = 2\epsilon$ for all $j \neq i_k$. Since $d(x_{k+1}, S) = \min\{d(x_{k+1}, C_j) : j = 1, \ldots, q\} \leqslant \epsilon$, we conclude that $d(x_{k+1}, C_{i_k}) \leqslant \epsilon$.

□

Remark 2. *The assumption that f is coercive in Theorem 1 and Corollary 1 can be replaced by requiring the iterates to be uniformly bounded for all sufficiently small step sizes when initialized in X_0. In other words, we can ask for there to exist $\bar{\alpha}, r > 0$ such that $\mathcal{M}(f,(0,\bar{\alpha}],X_0,0) \subset B(0,r)^{\mathbb{N}}$. Indeed, one can then apply our results to a coercive function f_r which coincides with f in $B(0,2r)$, namely $f_r(x) := f(P_{B(0,2r)}(x)) + d(x,B(0,2r))$ for all $x \in \mathbb{R}^n$ where $P_{B(0,2r)}$ is the projection on $B(0,2r)$. It is clear that f_r is definable and coercive. In order to show that f_r is Lipschitz continuous, it suffices to prove $g_r(x) := f(P_{B(0,2r)}(x))$ is locally Lipschitz. Let $L > 0$ denote a Lipschitz constant of f in $B(0,2r)$. For all $x, y \in \mathbb{R}^n$, we have $\|g_r(x) - g_r(y)\| = \|f(P_{B(0,2r)}(x)) - f(P_{B(0,2r)}(y))\| \leqslant L\|P_{B(0,2r)}(x) - P_{B(0,2r)}(y)\| \leqslant L\|x-y\|$.*

4 Approximation of first-order methods by subgradient trajectories

The theory we developed in the previous section provides a unified framework under which global stability of iterative methods with constant step sizes can be established. In this section, we show that all of the first-order methods that we mentioned in Section 1 are approximated by subgradient trajectories under appropriate assumptions on the objective functions. As a result, Theorem 1 and Corollary 1 can be applied to conclude global stability of those methods. We need the following lemma in order to prove the approximation of random reshuffling with momentum.

Lemma 3. *Let $f_1,\ldots,f_N$ be locally Lipschitz, $X \subset \mathbb{R}^n$ be bounded, $\delta \geqslant 0$, $\beta \in (-1,1)$, and $\gamma \in \mathbb{R}$. There exist $\delta', \bar{\alpha} > 0$ such that for all $\alpha \in (0,\bar{\alpha}]$, $\bar{k} \in \mathbb{N}$, and sequence $(x_{k,i})_{(k,i)\in\mathbb{N}\times\{0,\ldots,N\}}$ generated by random reshuffling with momentum (Algorithm 2) for which $x_0,\ldots,x_{\bar{k}} \in X$, we have*

$$\|x_{k,i} - x_{k,i-1}\| \leqslant \delta'\alpha$$

for $k = 0,\ldots,\bar{k}$ and $i = 0,\ldots,N$.

Proof. Let $r > 0$ such that $x_{-1,0} \in B(0,r/2)$ and $X \subset B(0,r/2)$. Since $f_1, f_2, \ldots, f_N$ are locally Lipschitz, their corresponding Clarke subdifferentials $\partial f_1, \partial f_2, \ldots, \partial f_N$ are upper semicontinuous [20, 2.1.5 Proposition (d)] with compact values [20, 2.1.2 Proposition (a)]. Thus, by [3, Proposition 3 p. 42] there exists $r' > \delta$ such that $\cup_{i=1}^N \partial f_i(B(0,r)) \subset B(0,r')$. Let

$$\delta' := \frac{r'}{1-|\beta|} \quad \text{and} \quad \bar{\alpha} := \frac{r}{2\delta'(N+|\gamma|)}.$$

Fix any $\alpha \in (0, \bar{\alpha}]$, $\bar{k} \in \mathbb{N}$, and sequence $(x_{k,i})_{(k,i) \in \mathbb{N} \times \{0,\ldots,N\}}$ generated by random reshuffling with momentum (Algorithm 2) for which $x_0, \ldots, x_{\bar{k}} \in X$. We will prove the lemma using induction on (k,i) with the total order $\preccurlyeq$ defined by $(k_1, i_1) \preccurlyeq (k_2, i_2)$ if $k_1 < k_2$ or $k_1 = k_2$ and $i_1 \leqslant i_2$. For the base case, note that $\|x_{0,0} - x_{0,-1}\| \leqslant \delta\alpha < r'\alpha \leqslant \delta'\alpha$. Now fix any $(k,i) \in \{0, \ldots \bar{k}\} \times \{0, \ldots, N\}$ and assume that $\|x_{k',i'} - x_{k',i'-1}\| \leqslant \delta'\alpha$ for all $(k', i') \preccurlyeq (k, i-1)$ (we identify $(k'-1, N-1)$ with $(k', -1)$ for notational simplicity; possibly negative indices i are treated similarly throughout the paper). Then

$$\begin{aligned}
\|y_{k,i}\| &\leqslant \|y_{k,i} - x_{k,i-1}\| + \|x_{k,i-1} - x_{k,0}\| + \|x_{k,0}\| && \text{(12a)}\\
&\leqslant |\gamma| \|x_{k,i-1} - x_{k,i-2}\| + \sum_{j=1}^{i-1} \|x_{k,j} - x_{k,j-1}\| + \|x_k\| && \text{(12b)}\\
&\leqslant |\gamma|\delta'\alpha + (i-1)\delta'\alpha + \frac{r}{2} && \text{(12c)}\\
&\leqslant (|\gamma| + N - 1)\delta'\alpha + \frac{r}{2} && \text{(12d)}\\
&\leqslant r. && \text{(12e)}
\end{aligned}$$

Above, we use the triangular inequality in (12a). We apply the update rule and again the triangular inequality to obtain (12b). (12c) is a result of the inductive hypothesis. (12d) and (12e) follow from $i \leqslant N$ and $\alpha \leqslant \bar{\alpha} := r/(2\delta'(|\gamma| + N - 1))$ respectively.

Thus $x_{k,i} - x_{k,i-1} - \beta(x_{k,i-1} - x_{k,i-2}) \in -\alpha \partial f_{\sigma_i^k}(y_{k,i}) \subset \alpha B(0, r')$. Therefore,

$$\begin{aligned}
\|x_{k,i} - x_{k,i-1}\| &\leqslant |\beta| \|x_{k,i-1} - x_{k,i-2}\| + r'\alpha \\
&\leqslant |\beta|\delta'\alpha + r'\alpha \\
&= \delta'\alpha.
\end{aligned}$$

□

Recall that a locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ is subdifferentially regular [20, 2.3.4 Definition] if its generalized directional derivative agrees with the classical directional derivative, that is to say, we have

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

for all $x \in \mathbb{R}^n$ and $h \in \mathbb{R}^n$, and the limit on the right hand side exists. We assume subdifferential regularity in Propositon 1 in order to guarantee that $\partial(f_1 + \cdots + f_N) = \partial f_1 + \cdots + \partial f_N$, while in general we only know that $\partial(f_1 + \cdots + f_N) \subset \partial f_1 + \cdots + \partial f_N$ holds [20, 2.3.3 Proposition] (see Remark 3). If we do not assume subdifferential regularity, Proposition 1 still holds with the same proof if in the conclusion we replace "approximated by subgradient trajectories of f" by "approximated by trajectories of the conservative field $(\partial f_1 + \cdots + \partial f_N)/N$ [17, Definition 1, Corollary 4] of f". Theorem 1 and Corollary 1 then hold with critical values and points associated with the conservative field under the additional assumption that $f_1, \ldots, f_N$ are definable [17, Theorems 5 and 6]. On the other hand, since Algorithm 1 does not consider the composite structure of the objective function, we do not require subdifferential regularity in order to obtain approximation and stability guarantees, as can be seen in Table 1.

Proposition 1. *Random reshuffling with momentum (Algorithm 2) is approximated by subgradient trajectories of composite functions $f = (f_1 + \cdots + f_N)/N$ up to the multiplicative constant $N/(1-\beta)$ where $f_1, \ldots, f_N$ are locally Lipschitz and subdifferentially regular.*

Proof. Let $\mathcal{M}$ denote the random reshuffling with fixed momentum parameters $\beta \in (-1,1)$, $\gamma \in \mathbb{R}$, and $\delta > 0$. Let $X_0, X_1 \subset \mathbb{R}^n$ be compact sets and consider $r > 0$ such that $X_0, X_1 \subset B(0, r/2) \subset \mathbb{R}^n$. By Lemma 3, there exist $\delta', \bar{\alpha} > 0$ such that for all $\alpha \in (0, \bar{\alpha}]$, $\tilde{k} \in \mathbb{N}$, and sequence $(x_{k,i})_{(k,i)\in\mathbb{N}\times\{0,\ldots,N\}}$ generated by random reshuffling with momentum (Algorithm 2) for which $x_0, \ldots, x_{\tilde{k}} \in B(0,r)$, we have that $\|x_{k,i} - x_{k,i-1}\| \leqslant \delta'\alpha$ for $k = 0, \ldots, \tilde{k}$ and $i = 0, \ldots, N$. Let $T := r/(4\delta' N \max\{1, |\gamma|\}) > 0$ and $\bar{k} \in \mathbb{N}$. We next show that any sequence $(x_{k,i})_{(k,i)\in\mathbb{N}\times\{0,\ldots,N\}}$ and $(y_{k,i})_{(k,i)\in\mathbb{N}\times\{1,\ldots,N\}}$ generated by Algorithm 2 with step size $\alpha \in (0, \min\{T/2, \bar{\alpha}\}]$ such that $x_0, \ldots, x_{\bar{k}} \in X_1$ and $x_{\bar{k}} \in X_0$ satisfy $x_{k,0}, \ldots, x_{k,N}, y_{k,1}, \ldots, y_{k,N} \in B(0,r)$ for $k = \bar{k}, \ldots, \bar{k}+K-1$ where $K := \lfloor T/\alpha \rfloor + 1$.

Fix any such α and sequence generated by Algorithm 2. Note that $\alpha K = \alpha(\lfloor T/\alpha \rfloor + 1) \leqslant 2T$ and thus $\alpha \leqslant (2T)/K = r/(2K\delta' N \max\{1, |\gamma|\})$. As $x_0, \ldots, x_{\bar{k}_m} \in B(0,r)$, we have that $\|x_{\bar{k},i}\| \leqslant \|x_{\bar{k}}\| + \sum_{j=1}^{i} \|x_{k,j} - x_{k,j-1}\| \leqslant r/2 + i\delta'\alpha \leqslant r/2 + ir/(2NK) \leqslant r/2 + r/(2K)$ for $i = 1, \ldots, N$, where we apply Lemma 3 with $\tilde{k} = \bar{k}$ in the second last inequality. In particular, $x_{\bar{k}+1} = x_{\bar{k},N} \in B(0, r/2 + r/(2K))$. Apply the previous argument recursively, we have that $x_{k,i} \in B(0, r/2 + (k-\bar{k})r/(2K) + ir/(2NK)) \subset B(0,r)$ for $k = \bar{k}, \ldots, \bar{k}+K-1$. By the update rule of Algorithm 2, $\|y_{k,i}\| \leqslant \|y_{k,i} - x_{k,i-1}\| + \|x_{k,i-1}\| = |\gamma|\|x_{k,i-1} - x_{k,i-2}\| + \|x_{k,i-1}\| \leqslant |\gamma|\delta'\alpha + r/2 + (k-\bar{k})r/(2K) + ir/(2NK) \leqslant r/(2NK) + r/2 + (k-\bar{k})r/(2K) + ir/(2NK) \leqslant r$ for $k = \bar{k}, \ldots, \bar{k}+K-1$ and $i = 1, \ldots, N$.

Let $(\alpha_m)_{m\in\mathbb{N}}$ be a positive sequence that converges to zero and let $(\bar{k}_m)_{m\in\mathbb{N}}$ be a sequence of natural numbers. For each $m \in \mathbb{N}$, we attribute a sequence of iterates $(x_k^m)_{k\in\mathbb{N}} \in \mathcal{M}(f, \alpha_m, X_0, \bar{k}_m)$ such that $x_0, \ldots, x_{\bar{k}} \in X_1$. We may assume that $\alpha_m \in (0, \min\{T/2, \bar{\alpha}\}]$ for any m, then $x_{k,0}^m, \ldots, x_{k,N}^m, y_{k,1}^m, \ldots, y_{k,N}^m \in B(0,r)$ for $k = \bar{k}_m, \ldots, \bar{k}_m + \lfloor T/\alpha_m \rfloor$. Consider the linear interpolation of the iterates $x_{\bar{k}_m}^m, x_{\bar{k}_m+1}^m, \ldots, x_{\bar{k}_m+\lfloor T/\alpha_m\rfloor+1}^m$, that is to say, the function $\bar{x}^m(\cdot)$ defined from $[0,T]$ to $\mathbb{R}^n$ by

$$\bar{x}^m(t) := x_k^m + (t - \alpha_m(k - \bar{k}_m))\frac{x_{k+1}^m - x_k^m}{\alpha_m}$$

for all $t \in [\alpha_m(k-\bar{k}_m), \min\{\alpha_m(k-\bar{k}_m+1), T\}]$ and $k \in \{\bar{k}_m, \ldots, \bar{k}_m + \lfloor T/\alpha_m \rfloor\}$. Since $B(0,r)$ is convex, it holds that $\|\bar{x}^m(t)\| \leqslant r$ for all $t \in [0,T]$. We also have $\|(\bar{x}^m)'(t)\| = \|(x^m_{k+1} - x^m_k)/\alpha_m\| \leqslant \sum_{i=1}^N \|x^m_{k,i} - x^m_{k,i-1}\|/\alpha_m \leqslant N\delta'$ for all $t \in [\alpha_m(k-\bar{k}_m), \min\{\alpha_m(k-\bar{k}_m+1), T\}]$ and $k \in \{\bar{k}_m, \ldots, \bar{k}_m + \lfloor T/\alpha_m \rfloor\}$. By successively applying the Arzelà-Ascoli and the Banach-Alaoglu theorems (see [3, Theorem 4 p. 13]), there exist a subsequence (again denoted by $(\alpha_m)_{m\in\mathbb{N}}$) and an absolutely continuous function $x : [0,T] \to \mathbb{R}^n$ such that $\bar{x}^m(\cdot)$ converges uniformly to $x(\cdot)$ and $(\bar{x}^m)'(\cdot)$ converges weakly to $x'(\cdot)$ in $L^1([0,T],\mathbb{R}^n)$. We next verify that the limit $x(\cdot)$ is a solution to the differential inclusion with initial condition

$$x'(t) \in -\frac{1}{1-\beta}\sum_{i=1}^N \partial f_i(x(t)), \quad \text{for almost every } t \in [0,T], \quad x(0) \in X_0. \tag{14}$$

By subdifferential regularity of $f_1, \ldots, f_N$, we have $\partial(\sum_{i=1}^N f_i) = \sum_{i=1}^N \partial f_i$ [20, p. 40, Corollary 3]. It is thus easy to see that such $x(\cdot)$ is a subgradient trajectory of f up to the multiplicative constant $c := N/(1-\beta) > 0$.

For any fixed $m \in \mathbb{N}$, we have that

$$x^m_{k,i} - x^m_{k,i-1} - \beta(x^m_{k,i-1} - x^m_{k,i-2}) \in -\alpha_m \partial f_{\sigma^k_i}(y^m_{k,i}) \tag{15}$$

for all $k \in \{\bar{k}_m, \ldots, \bar{k}_m + \lfloor T/\alpha_m \rfloor \rfloor\}$ and $i \in \{0, \ldots, N\}$. For any fixed k, summing (15) up for $i = 1, \ldots, N$ yields

$$x^m_{k+1,0} - x^m_{k,0} - \beta(x^m_{k+1,-1} - x^m_{k,-1}) \in -\alpha_m \sum_{i=1}^N \partial f_{\sigma^k_i}(y^m_{k,i}).$$

Consider the linear interpolation of the iterates $x^m_{\bar{k}_m,-1}, x^m_{\bar{k}_m+1,-1}, \ldots,$ $x^m_{\bar{k}_m+\lfloor T/\alpha_m \rfloor+1,-1}$, that is to say, the function $\bar{x}^m_{-1}(\cdot)$ defined from $[0,T]$ to $\mathbb{R}^n$ by

$$\bar{x}^m_{-1}(t) := x^m_{k,-1} + (t - \alpha_m(k-\bar{k}_m))\frac{x^m_{k+1,-1} - x^m_{k,-1}}{\alpha_m}$$

for all $t \in [\alpha_m(k-\bar{k}_m), \min\{\alpha_m(k-\bar{k}_m+1), T\}]$ and $k \in \{\bar{k}_m, \ldots, \bar{k}_m + \lfloor T/\alpha_m \rfloor\}$.

For almost every $t \in (0,T)$ and any neighborhood $\mathcal{N}$ of 0, there exists $m_0 \in \mathbb{N}$ such that for any $m \geqslant m_0$, there exists $k \in \{\bar{k}_m, \ldots, \bar{k}_m + \lfloor T/\alpha_m \rfloor\}$ such that

$$\begin{aligned}
(\bar{x}^m)'(t) - \beta(\bar{x}^m_{-1})'(t) &= \frac{x^m_{k+1,0} - x^m_{k,0}}{\alpha_m} - \beta\frac{x^m_{k+1,-1} - x^m_{k,-1}}{\alpha_m} && \text{(16a)} \\
&\in -\sum_{i=1}^N \partial f_{\sigma^k_i}(y^m_{k,i}) && \text{(16b)} \\
&\subset -\sum_{i=1}^N \Big(\partial f_{\sigma^k_i}(x(t)) + \mathcal{N}/N\Big) && \text{(16c)}
\end{aligned}$$

$$\subset -\sum_{i=1}^{N} \partial f_i(x(t)) + \mathcal{N}, \tag{16d}$$

where (16c) follows from upper semi-continuity of ∂f_i [20, 2.1.5 Proposition (d)] and

$$\begin{aligned}\|y_{k,i}^m - x(t)\| &\leqslant \left\|y_{k,i}^m - \bar{x}^m(t)\right\| + \|\bar{x}^m(t) - x(t)\| \\ &= \left\|y_{k,i}^m - x_k^m - (t - \alpha_m(k - \bar{k}_m))\frac{x_{k+1}^m - x_k^m}{\alpha_m}\right\| + \|\bar{x}^m(t) - x(t)\| \\ &\leqslant \left\|y_{k,i}^m - x_k^m\right\| + \left\|x_{k+1}^m - x_k^m\right\| + \|\bar{x}^m(t) - x(t)\| \\ &\leqslant (|\gamma| + i)\delta'\alpha_m + N\delta'\alpha_m + \|\bar{x}^m(t) - x(t)\| \\ &\to 0\end{aligned}$$

as $m \to \infty$. It remains to show that $(\bar{x}_{-1}^m)'(\cdot)$ converges weakly to $x'(\cdot)$ in $L^1([0,T],\mathbb{R}^n)$. Indeed, by [3, Convergence Theorem p. 60], it then holds that $(x, (1-\beta)x') \in \mathrm{graph}(\sum_{i=1}^N \partial f_i)$ and thus (14) follows.

For any $m \in \mathbb{N}$ and almost every $s \in [0,T]$, $\|(\bar{x}_{-1}^m)'(s)\| = \|(x_{k+1,-1}^m - x_{k,-1}^m)/\alpha_m\| \leqslant \sum_{i=0}^{N-1} \|x_{k,i}^m - x_{k,i-1}^m\|/\alpha_m \leqslant N\delta'$ for some $k \in \mathbb{N}$. Thus it suffices to show that for all $t \in [0,T]$,

$$\int_0^t (\bar{x}_{-1}^m)'(s)\ ds \to \int_0^t x'(s)\ ds.$$

Indeed, $\left\|\int_0^t (\bar{x}_{-1}^m)'(s)\ ds - \int_0^t (\bar{x}^m)'(s)\ ds\right\| = \cdots$

$$\begin{aligned}&= \left\|\bar{x}_{-1}^m(t) - \bar{x}_{-1}^m(0) - (\bar{x}^m(t) - \bar{x}^m(0))\right\| \\ &= \Bigg\|x_{k,-1}^m + (t - \alpha_m(k - \bar{k}_m))\frac{x_{k+1,-1}^m - x_{k,-1}^m}{\alpha_m} - x_{\bar{k}_m,-1}^m \\ &\quad - \left(x_{k,0}^m + (t - \alpha_m(k - \bar{k}_m))\frac{x_{k+1,0}^m - x_{k,0}^m}{\alpha_m} - x_{\bar{k}_m,0}^m\right)\Bigg\| \\ &\leqslant \left\|x_{k+1,-1}^m - x_{k+1,0}^m\right\| + \left\|x_{k,-1}^m - x_{k,0}^m\right\| + \left\|x_{\bar{k}_m,-1}^m - x_{\bar{k}_m,0}^m\right\| \\ &\leqslant \delta'\alpha_m + \delta'\alpha_m + \delta'\alpha_m \to 0\end{aligned}$$

where $k = \bar{k}_m + \lfloor t/\alpha_m \rfloor$. As $x'(\cdot)$ is a weak limit of $(\bar{x}^m)'(\cdot)$, $(\bar{x}_{-1}^m)'(\cdot)$ converges weakly to $x'(\cdot)$.

To sum up, we have shown that for every sequence $(\alpha_m)_{m\in\mathbb{N}}$ of positive numbers converging to zero and every sequence $(\bar{k}_m)_{m\in\mathbb{N}}$ of natural numbers, there exists a subsequence of natural numbers for which the corresponding linear interpolations uniformly converge towards a solution of the differential inclusion (14). The conclusion of the proposition now easily follows. To see why, one can reason by contradiction and assume that there exists $\epsilon > 0$ such that for all $\bar{\alpha} > 0$, there exist $\hat{\alpha} \in (0, \bar{\alpha}]$, $\hat{k} \in \mathbb{N}$, and a sequence $(x_k^m)_{k\in\mathbb{N}} \in \mathcal{M}(f, \hat{\alpha}, X_0, \hat{k})$ such that $x_0^m, \ldots, x_{\hat{k}}^m \in X_1$, and for any solution $x(\cdot)$ to the differential inclusion (14), it holds that $\|x_k^m - x(\hat{\alpha}(k-\hat{k}))\| > \epsilon$ for some $k \in \{\hat{k}, \hat{k}+1, \ldots, \hat{k}+\lfloor T/\hat{\alpha} \rfloor\}$. We can then generate a sequence $(\alpha_m)_{m\in\mathbb{N}}$ of positive numbers converging to zero and a sequence $(\bar{k}_m)_{m\in\mathbb{N}}$ of natural numbers such that, for any solution $x(\cdot)$ to the differential inclusion (14), it holds that $\|x_k^m - x(\alpha_m(k-\bar{k}_m))\| > \epsilon$ for some $k \in \{\bar{k}_m, \bar{k}_m+1, \ldots, \bar{k}_m+\lfloor T/\alpha_m \rfloor\}$. Since there exists a subsequence $(\alpha_{\varphi(m)})_{m\in\mathbb{N}}$ such that $(\bar{x}^{\varphi(m)}(\cdot))_{m\in\mathbb{N}}$ uniformly converges to a solution to the differential inclusion (14), we obtain a contradiction. □

Remark 3. *To further see why subdifferential regularity is required for applying Definition 3 to Algorithm 2, consider a locally Lipschitz coercive semi-algebraic function $f := (f_1+f_2+f_3)/3$ where $f_1, f_2, f_3 : \mathbb{R} \to \mathbb{R}$ are defined by*

$$f_1(x) := \max\{x, 0\} \quad , \quad f_2(x) := \min\{x, 0\} \quad , \quad f_3(x) := x^2 \quad , \quad \forall x \in \mathbb{R}.$$

Notice that f_1, f_2, f_3 are all locally Lipschitz but f_1 is not subdifferentially regular. We have $0 \in \partial f_1(0) = [0,1], 0 \in \partial f_2(0) = [1,0]$, and $0 \in \partial f_3(0) = \{0\}$. Meanwhile $0 \notin \partial f(0) = \{1/3\}$. Thus random reshuffling with momentum can get stuck at 0. Therefore, the conclusion of Corollary 1 does not apply to this example.

Proposition 2. *Random-permutations cyclic coordinate descent method (Algorithm 3) is approximated by subgradient trajectories of continuously differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$.*

Proof. Similar to the proof of Proposition 1, let $X_0 \subset \mathbb{R}^n$ be a compact subset. Consider $r > 0$ such that $X_0 \subset B(0, r/2) \subset \mathbb{R}^n$. We would like to find $T > 0$ such that for any $\alpha \in (0, T/2], \bar{k} \in \mathbb{N}$, any sequence $(x_{k,i})_{(k,i)\in\mathbb{N}\times\{0,\ldots,n\}}$ generated by Algorithm 3 with step size α for which $x_{\bar{k},0} \in X_0$, we have that $x_{k,i} \in B(0,r)$ for all $k = \bar{k}, \ldots, \bar{k}+K-1$ and $i = 0, 1, \ldots, n$ where $K := \lfloor T/\alpha \rfloor + 1$. As ∇f is continuous, we have $M := \sup\{\|\nabla f(x)\| : x \in B(0,r)\} < \infty$. Let $T := r/(4Mn) > 0$. Then $\alpha K = \alpha(\lfloor T/\alpha \rfloor + 1) \leqslant 2T$ and $\alpha \leqslant 2T/K = r/(2KMn)$. For any $k = \bar{k}, \ldots, \bar{k}+K-1$ and $i = 0, \ldots, n-1$, if one assumes that $x_{k,i} \in B(0,r)$, then $\|x_{k,i+1} - x_{k,i}\| = \|\alpha \nabla_{\sigma_{i+1}^k} f(x_{k,i})\| \leqslant \alpha M \leqslant r/(2Kn)$. As $x_{\bar{k},0} \in B(0, r/2)$, by induction, we have $x_{k,i} \in B(0, r/2 + (kn+i)r/(2Kn)) \subset B(0,r)$ for any $k = \bar{k}, \ldots, \bar{k}+K-1$ and $i = 0, \ldots, n$.

We next show that Algorithm 3 is approximated by subgradient trajectories, which are solutions to the following differential equation with initial condition

$$x'(t) = -\nabla f(x(t)), \quad \text{for almost every } t \in [0,T], \quad x(0) \in X_0. \tag{19}$$

Denote by $\mathcal{M}$ the random-permutations cyclic coordinate descent method defined by Algorithm 3. Let $(\alpha_m)_{m\in\mathbb{N}}$ denote a sequence of positive numbers that converges to zero and $(\bar{k}_m)_{m\in\mathbb{N}}$ be a sequence of natural numbers. Without loss of generality, we may assume that $(\alpha_m)_{m\in\mathbb{N}} \subset (0, T/2]$. For each $m \in \mathbb{N}$, we attribute a sequence of iterates $(x_k^m)_{k\in\mathbb{N}} \in \mathcal{M}(f, \alpha_m, X_0, \bar{k}_m)$. Consider the linear interpolation of the iterates $x^m_{\bar{k}_m}, x^m_{\bar{k}_m+1}, \ldots, x^m_{\bar{k}_m+\lfloor T/\alpha_m \rfloor+1}$, that is to say, the function $\bar{x}^m(\cdot)$ defined from $[0,T]$ to $\mathbb{R}^n$ by

$$\bar{x}^m(t) := x^m_k + (t - \alpha_m(k - \bar{k}_m))\frac{x^m_{k+1} - x^m_k}{\alpha_m}$$

for all $t \in [\alpha_m k, \min\{\alpha_m(k+1), T\}]$ and $k \in \{\bar{k}_m, \ldots, \bar{k}_m + \lfloor T/\alpha_m \rfloor\}$. Recall that $x^m_k = x^m_{k,0} = x^m_{k-1,n}$. As we have shown in the first paragraph of the proof, $x_{k,i} \in B(0,r)$ for all $k = \bar{k}_m, \ldots, \bar{k}_m + \lfloor T/\alpha_m \rfloor$ and $i = 0, 1, \ldots, n$, thus $x^m_{\bar{k}_m}, x^m_{\bar{k}_m+1}, \ldots, x^m_{\bar{k}_m+\lfloor T/\alpha_m \rfloor+1} \in B(0,r)$. Since $B(0,r)$ is convex, it holds that $\|\bar{x}^m(t)\| \leqslant r$ for all $t \in [0,T]$. Observe that $(\bar{x}^m)'(t) = (x^m_{k+1} - x^m_k)/\alpha_m = -\sum_{i=1}^n \nabla_{\sigma^k_i} f(x_{k,i-1})$ for all $t \in (\alpha_m k, \min\{\alpha_m(k+1), T\})$ and $k \in \{\bar{k}_m, \ldots, \bar{k}_m + \lfloor T/\alpha_m \rfloor\}$. Hence, we have that

$$\|(\bar{x}^m)'(t)\| = \|\sum_{i=1}^n \nabla_{\sigma^k_i} f(x_{k,i-1})\| \leqslant \sum_{i=1}^n \|\nabla_{\sigma^k_i} f(x_{k,i-1})\| \leqslant nM$$

for almost every $t \in [0,T]$. By successively applying the Arzelà-Ascoli and the Banach-Alaoglu theorems (see [3, Theorem 4 p. 13]), there exist a subsequence (again denoted $(\alpha_m)_{m\in\mathbb{N}}$) and an absolutely continuous function $x : [0,T] \to \mathbb{R}^n$ such that $\bar{x}^m(\cdot)$ converges uniformly to $x(\cdot)$ and $(\bar{x}^m)'(\cdot)$ converges weakly to $x'(\cdot)$ in $L^1([0,T], \mathbb{R}^n)$.

For almost every $t \in [0,T]$ and any α_m in the sequence, $t \in (\alpha_m k, \min\{\alpha_m(k+1), T\})$ for some $k \in \{\bar{k}_m, \ldots, \bar{k}_m + \lfloor T/\alpha_m \rfloor\}$. We fix any such t and any $\xi > 0$ from now on. As ∇f is continuous, there exists $\delta > 0$ such that $\|\nabla_i f(y) - \nabla_i f(x(t))\| \leqslant \|\nabla f(y) - \nabla f(x(t))\| \leqslant \xi/(2n)$, for all $y \in B(x(t), \delta)$ and $i = 1, \ldots, n$. Since $(\bar{x}^m(\cdot))_{m\in\mathbb{N}}$ converges uniformly to $x(\cdot)$, there exists $m_0 \in \mathbb{N}$ such that $\|\bar{x}^m(t) - x(t)\| \leqslant \min\{\epsilon/2, \delta/2\}$ for any $m \geqslant m_0$. As $\lim_{m\to\infty} \alpha_m = 0$, there exists $m_1 \geqslant m_0$ such that $\alpha_m \leqslant \delta/(4nM)$ for all $m \geqslant m_1$. We next show that $\|(\bar{x}^m)'(t) - \nabla f(x(t))\| \leqslant \xi/2$ for all $m \geqslant m_1$. Indeed, if it is the case, then

$$\begin{aligned}(\bar{x}_m(t), \bar{x}'_m(t)) \in B\left(x(t), \min\left\{\frac{\xi}{2}, \frac{\delta}{2}\right\}\right) \times B\left(-\nabla f(x(t)), \frac{\xi}{2}\right) \\ \subset \text{graph}\,(-\nabla f) + B(0, \xi)\end{aligned}$$

and by [3, Convergence Theorem p. 60], it holds that $x'(t) = -\nabla f(x(t))$ for almost every $t \in [0,T]$. The sequence of initial iterates $(x^m(0))_{m\in\mathbb{N}}$ lies in the compact set X_0, hence its limit $x(0)$ lies in X_0 as well. As a result, $x(\cdot)$ is a solution to the differential inclusion (19).

Note that for any $i = 1, \ldots, n$,

$$\begin{aligned}&\|x^m_{k,i-1} - \bar{x}^m(t)\| \\ &= \left\| x^m_k - \alpha_m \sum_{j=1}^{i-1} \nabla_{\sigma^k_j} f(x^m_{k,j-1}) - x^m_k - (t - \alpha_m(k - \bar{k}_m))\frac{x^m_{k+1} - x^m_k}{\alpha_m} \right\|\end{aligned}$$

$$
\begin{aligned}
&= \left\| \alpha_m \sum_{j=1}^{i-1} \nabla_{\sigma_j^k} f(x_{k,j-1}^m) + (t - \alpha_m(k - \bar{k}_m)) \frac{x_{k+1}^m - x_k^m}{\alpha_m} \right\| \\
&\leqslant \alpha_m \left\| \sum_{j=1}^{i-1} \nabla_{\sigma_j^k} f(x_{k,j-1}^m) \right\| + \|x_{k+1}^m - x_k^m\| \\
&\leqslant \alpha_m \sum_{j=1}^{i-1} \left\| \nabla_{\sigma_j^k} f(x_{k,j-1}^m) \right\| + \alpha_m \|(\bar{x}^m)'(t)\| \\
&\leqslant \alpha_m (i-1) M + \alpha_m n M \\
&\leqslant 2\alpha_m n M \leqslant \delta/2.
\end{aligned}
$$

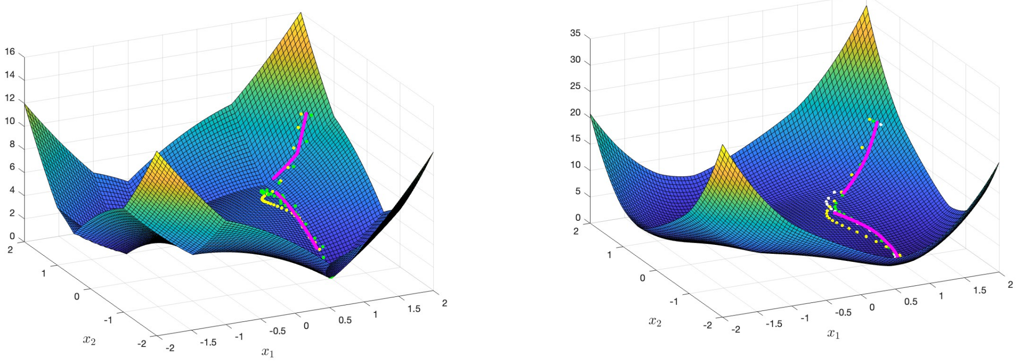
Thus, $\|x_{k,i-1}^m - x(t)\| \leqslant \|x_{k,i-1}^m - \bar{x}^m(t)\| + \|\bar{x}^m(t) - x(t)\| \leqslant \delta/2 + \min\{\xi/2, \delta/2\} \leqslant \delta$ for all $i = 1, \ldots, n$ and $m \geqslant m_1$. It follows that

$$
\begin{aligned}
\|(\bar{x}^m)'(t) - \nabla f(x(t))\| &= \left\| \sum_{i=1}^{n} \nabla_{\sigma_i^k} f(x_{k,i-1}^m) - \nabla f(x(t)) \right\| \\
&= \left\| \sum_{i=1}^{n} \Big(\nabla_{\sigma_i^k} f(x_{k,i-1}^m) - \nabla_{\sigma_i^k} f(x(t)) \Big) \right\| \\
&\leqslant \sum_{i=1}^{n} \left\| \nabla_{\sigma_i^k} f(x_{k,i-1}^m) - \nabla_{\sigma_i^k} f(x(t)) \right\| \\
&\leqslant \sum_{i=1}^{n} \frac{\xi}{2n} = \frac{\xi}{2}.
\end{aligned}
$$

To sum up, we have shown that for every sequence $(\alpha_m)_{m\in\mathbb{N}}$ of positive numbers converging to zero and every sequence $(\bar{k}_m)_{m\in\mathbb{N}}$ of natural numbers, there exists a subsequence of natural numbers for which the corresponding linear interpolations uniformly converge towards a solution of the differential equation (19). The conclusion of the proposition now easily follows using the same argument as the last paragraph in the proof of Proposition 1. □

Note that in the proof above, we do not make use of the set X_1 that appears in Definition 3. This is because that for any iterates $(x_k)_{k\in\mathbb{N}}$ generated by the random-permutations cyclic coordinate descent method, $(x_k)_{k\geqslant\bar{k}}$ is unrelated to $x_0, \ldots, x_{\bar{k}-1}$ if given $x_{\bar{k}}$. Contrary to Proposition 1, Proposition 2 cannot be relaxed to locally Lipschitz functions. For example, the coordinate descent method can get stuck at $(1,1)$, which is not a critical point of $f(x_1, x_2) = \max\{|x_1|, |x_2|\}$.

We conclude this paper by illustrating Theorem 1, Corollary 1, Proposition 1, and Proposition 2 on two examples. The first (Figure 1a) is nonsmooth and the second (Figure 1b) is continuously differentiable. One can see that the iterates indeed track a subgradient trajectory up to a certain time, then go on to track another subgradient trajectory, after which they stabilize around a critical point.



(a) $f(x_1, x_2) = |x_1^2 - 1| + 2|x_1 x_2 + 1| + |x_2^2 - 1|$.

(b) $f(x_1, x_2) = |x_1^2 - 1|^{3/2} + 2|x_1 x_2 + 1|^{3/2} + |x_2^2 - 1|^{3/2}$.

Figure 1: The subgradient method with momentum, random reshuffling with momentum, and random-permutations cyclic coordinate descent method are in yellow, green, and white respectively. Subgradient trajectories are in magenta.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] 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*, 35:438–457, 2010.

[3] J.-P. Aubin and A. Cellina. *Differential inclusions: set-valued maps and viability theory*, volume 264. Springer-Verlag, 1984.

[4] A. Beck. *First-order methods in optimization.* SIAM, 2017.

[5] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM journal on Optimization*, 23(4):2037–2060, 2013.

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

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

[8] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions, part ii: Applications. *Mathematics of Operations Research*, 31(4):673–695, 2006.

[9] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

[10] D. Bertsekas. *Convex optimization algorithms*. Athena Scientific, 2015.

[11] D. P. Bertsekas et al. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010(1-38):3, 2011.

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

[13] 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*, pages 1–31, 2022.

[14] J. Blanton. *Foundations of Differential Calculus*. Springer Science & Business Media, 2006.

[15] J. Bochnak, M. Coste, and M.-F. Roy. *Real algebraic geometry*, volume 36. Springer Science & Business Media, 2013.

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

[17] J. Bolte and E. Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, pages 1–33, 2020.

[18] V. S. Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

[19] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[20] F. H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM Classics in Applied Mathematics, 1990.

[21] E. A. Coddington and N. Levinson. *Theory of ordinary differential equations*. Tata McGraw-Hill Education, 1955.

[22] A. Daniilidis and D. Drusvyatskiy. Pathological subgradient dynamics. *SIAM Journal on Optimization*, 30(2):1327–1338, 2020.

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

[24] A. Dontchev and F. Lempio. Difference methods for differential inclusions: A survey. *SIAM review*, 34(2):263–294, 1992.

[25] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. Curves of descent. *SIAM Journal on Control and Optimization*, 53(1):114–138, 2015.

[26] J. C. Duchi and F. Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

[27] S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*. John Wiley & Sons, 2009.

[28] L. Euler. *Institutiones calculi integralis*, volume 1. impensis Academiae imperialis scientiarum, 1792.

[29] A. F. Filippov. *Differential equations with discontinuous righthand sides: control systems*, volume 18. Springer Science & Business Media, 2013.

[30] R. Ge, J. D. Lee, and T. Ma. Matrix Completion has No Spurious Local Minimum. *NIPS*, 2016.

[31] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1):49–84, 2021.

[32] M. Gürbüzbalaban, A. Ozdaglar, N. D. Vanli, and S. J. Wright. Randomness and permutations in coordinate descent methods. *Mathematical Programming*, 181(2):349–376, 2020.

[33] A. D. Ioffe. An invitation to tame optimization. *SIAM Journal on Optimization*, 19(4):1894–1917, 2009.

[34] T. Kohonen. An adaptive associative memory principle. *IEEE Transactions on Computers*, 100(4):444–445, 1974.

[35] N. B. Kovachki and A. M. Stuart. Continuous time analysis of momentum methods. *Journal of Machine Learning Research*, 22(17):1–40, 2021.

[36] K. Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pages 769–783, 1998.

[37] T. G. Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of applied Probability*, 7(1):49–58, 1970.

[38] H. Kushner. Convergence of recursive adaptive and identification procedures via weak convergence theory. *IEEE Transactions on Automatic Control*, 22(6):921–930, 1977.

[39] H. J. Kushner. General convergence results for stochastic approximations via weak convergence theory. *Journal of mathematical analysis and applications*, 61(2):490–503, 1977.

[40] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[41] C.-P. Lee and S. J. Wright. Random permutations fix a worst case for cyclic coordinate descent. *IMA Journal of Numerical Analysis*, 39(3):1246–1275, 2019.

[42] X. Li, Z. Zhu, A. M.-C. So, and R. Vidal. Nonconvex Robust Low-Rank Matrix Recovery. *SIAM Journal on Optimization*, 2019.

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

[44] Z.-Q. Luo. On the convergence of the lms algorithm with adaptive learning rate for linear feedforward networks. *Neural Computation*, 3(2):226–245, 1991.

[45] Z.-Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992.

[46] 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*, 24:1–84, 2023.

[47] K. Mishchenko, A. Khaled, and P. Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33:17309–17320, 2020.

[48] A. Nedic and D. P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.

[49] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[50] Y. Nesterov. *Lectures on Convex Optimization*, volume 137. Springer Science & Business Media, 2018.

[51] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

[52] 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*, 22(207):1–44, 2021.

[53] O. A. Nielsen. *An introduction to integration and measure theory*, volume 17. Wiley-Interscience, 1997.

[54] P. Ochs, Y. Chen, T. Brox, and T. Pock. iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

[55] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables.* Academic Press, New York, 1970.

[56] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[57] E. Pauwels. Incremental without replacement sampling in nonconvex optimization. *Journal of Optimization Theory and Applications*, pages 1–26, 2021.

[58] A. Pillay and C. Steinhorn. Definable sets in ordered structures. i. *Transactions of the American Mathematical Society*, 295(2):565–592, 1986.

[59] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[60] R. Ríos-Zertuche. Examples of pathological dynamics of the subgradient method for lipschitz path-differentiable functions. *Mathematics of Operations Research*, 47(4):3184–3206, 2022.

[61] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

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

[63] A. Salim. *Random monotone operators and application to stochastic optimization*. PhD thesis, Université Paris-Saclay (ComUE), 2018.

[64] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

[65] K. Taubert. Converging multistep methods for initial value problems involving multivalued maps. *Computing*, 27(2):123–136, 1981.

[66] T. H. Tran, L. M. Nguyen, and Q. Tran-Dinh. Smg: A shuffling gradient-based method with momentum. In *International Conference on Machine Learning*, pages 10379–10389. PMLR, 2021.

[67] T. H. Tran, K. Scheinberg, and L. M. Nguyen. Nesterov accelerated shuffling gradient method for convex optimization. In *International Conference on Machine Learning*, pages 21703–21732. PMLR, 2022.

[68] L. van den Dries. Remarks on Tarski's problem concerning ($\mathbb{R}$,+,$*$, exp). In *Studies in Logic and the Foundations of Mathematics*, volume 112, pages 97–121. Elsevier, 1984.

[69] L. van den Dries. *Tame topology and o-minimal structures*, volume 248. Cambridge university press, 1998.

[70] L. van den Dries and C. Miller. Geometric categories and o-minimal structures. *Duke Mathematical Journal*, 84(2):497–540, 1996.

[71] B. Widrow and M. E. Hoff. Adaptive switching circuits. Technical report, Stanford Univ Ca Stanford Electronics Labs, 1960.

[72] S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

[73] S. Zavriev and F. Kostyuk. Heavy-ball method in nonconvex optimization problems. *Computational Mathematics and Modeling*, 4(4):336–341, 1993.

[74] G. Zhang, H.-M. Chiu, and R. Y. Zhang. Accelerating SGD for Highly Ill-Conditioned Huge-Scale Online Matrix Completion. *Advances in Neural Information Processing Systems*, 35, 2022.