Stochastic subgradient method converges on tame functions

4 authors, including:


Damek Davis
Cornell University


Dmitriy Drusvyatskiy
University of Washington Seattle


Jason Lee
University of Southern California

Stochastic subgradient method converges on tame functions

Damek Davis* Dmitriy Drusvyatskiy† Sham Kakade‡
Jason D. Lee§

Abstract

This work considers the question: what convergence guarantees does the stochastic subgradient method have in the absence of smoothness and convexity? We prove that the stochastic subgradient method, on any semialgebraic locally Lipschitz function, produces limit points that are all first-order stationary. More generally, our result applies to any function with a Whitney stratifiable graph. In particular, this work endows the stochastic subgradient method, and its proximal extension, with rigorous convergence guarantees for a wide class of problems arising in data science—including all popular deep learning architectures.

1 Introduction

In this work, we study the long term behavior of the stochastic subgradient method on nonsmooth and nonconvex functions. Setting the stage, consider the optimization problem

$$\min_{x\in\mathbb{R}^d} \; f(x),$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is a locally Lipschitz continuous function. The stochastic subgradient method simply iterates the steps

$$x_{k+1} = x_k - \alpha_k\Big(y_k + M_{k+1}\Big) \qquad \text{with} \qquad y_k \in \partial f(x_k). \tag{1.1}$$

*School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850, USA; `people.orie.cornell.edu/dsd95/`.

†Department of Mathematics, University of Washington, Seattle, WA 98195; `www.math.washington.edu/~ddrusv`. Research of Drusvyatskiy was supported by the AFOSR YIP award FA9550-15-1-0237 and by the NSF DMS 1651851 and CCF 1740551 awards.

‡Departments of Statistics and Computer Science, University of Washington, Seattle, WA 98195; `homes.cs.washington.edu/~sham/`. Sham Kakade acknowledges funding from the Washington Research Foundation Fund for Innovation in Data-Intensive Discovery and the NSF CCF 1740551 award.

§Data Science and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089; `www-bcf.usc.edu/~lee715`. JDL acknowledges funding from the ARO MURI Award W911NF-11-1-0303.

Here $\partial f(x)$ denotes the Clarke subdifferential [7]. Informally, the set $\partial f(x)$ is the convex hull of limits of gradients at nearby differentiable points. In classical circumstances, the subdifferential reduces to more familiar objects. Namely, when f is C^1-smooth at x, the subdifferential $\partial f(x)$ consists only of the gradient $\nabla f(x)$, while for convex functions, it reduces to the subdifferential in the sense of convex analysis. The positive sequence $\{\alpha_k\}_{k\geq 0}$ is user specified, and it controls the step-sizes of the algorithm. As is typical for stochastic subgradient methods, we will assume that this sequence is square summable but not summable, meaning $\sum_k \alpha_k = \infty$ and $\sum_k \alpha_k^2 < \infty$. Finally, the stochasticity is modeled by the random (noise) sequence $\{M_k\}_{k\geq 1}$. We make the standard assumption that conditioned on the past, each random variable M_k has mean zero and its second moment is uniformly bounded.

Though variants of the stochastic subgradient method (1.1) date back to Robbins-Monro's pioneering 1951 work [26], their convergence behavior is still largely not understood in nonsmooth and nonconvex settings. For example, the following question remains open.

Does the (stochastic) subgradient method have any convergence guarantees on locally Lipschitz functions, which may be neither smooth nor convex?

That this question remains unanswered is somewhat concerning as the stochastic subgradient method forms a core numerical subroutine for several widely used solvers, including Google's TensorFlow [1] and the opensource PyTorch [25] library.

Convergence behavior of (1.1) is well understood when applied to convex, smooth, and more generally, weakly convex problems. In these three cases, almost surely, every limit point x^* of the iterate sequence is first-order critical [24], meaning $0 \in \partial f(x^*)$. Moreover, rates of convergence in terms of natural optimality/stationarity measures are available. In summary, the rates are $\mathbb{E}\left[f(x_k) - \inf f\right] = O(k^{-1/2})$, $\mathbb{E}\left[\|\nabla f(x_k)\|\right] = O(k^{-1/4})$, and $\mathbb{E}\left[\|\nabla f_{1/(2\rho)}(x_k)\|\right] = O(k^{-1/4})$, for functions that are convex [23], smooth [16], and ρ-weakly convex [11, 12], respectively. In particular, the convergence guarantee above for ρ-weakly convex functions appeared only recently in [11, 12], with the Moreau envelope $f_{1/(2\rho)}$ playing a central role.

Though widely applicable, these previous results on the convergence of the stochastic subgradient method do not apply to even relatively simple non-pathological functions, such as $f(x, y) = (|x| - |y|)^2$ and $f(x) = (1 - \max\{x, 0\})^2$. It is not only toy examples, however, that lack convergence guarantees, but the entire class of deep neural networks with nonsmooth activation functions (e.g., ReLU). Since such networks are routinely trained in practice, it is worthwhile to understand if indeed the iterates x_k tend to a meaningful limit.

In this paper, we provide a positive answer to this question for a wide class of locally Lipschitz functions; indeed, the function class we consider is virtually exhaustive in data scientific contexts (see Corollary 4.10 for consequences in deep learning). Aside from mild technical conditions, the only meaningful assumption we make is that f strictly decreases along any trajectory $x(\cdot)$ of the differential inclusion $\dot{x}(t) \in -\partial f(x(t))$ emanating from a noncritical point. Under this assumption, we will prove that every limit point of the stochastic subgradient method is critical for f, almost surely. Similar results also hold for the proximal stochastic subgradient method; see Section 5. Our proof is rooted in the "nonescape argument" for ODEs, using f as a Lyapunov function for the continuous dynamics. In particular, the proof we present is in the same spirit as that in [20, Theorem 5.2.1] and [15, Section 3.4.1].

The main question that remains therefore is which functions decrease along the continuous subgradient curves. Let us look for inspiration at convex functions, which are well-known to satisfy this property [5, 6]. Indeed, if f is convex and $x\colon [0,\infty) \to \mathbb{R}$ is any absolutely continuous curve, then the "chain rule" holds:

$$\tfrac{d}{dt}(f \circ x) = \langle \partial f(x), \dot{x} \rangle \qquad \text{for a.e. } t \geq 0. \tag{1.2}$$

An elementary linear algebraic argument then shows that if x satisfies $\dot{x}(t) \in -\partial f(x(t))$ a.e., then automatically $-\dot{x}(t)$ is the minimal norm element of $\partial f(x(t))$. Therefore, integrating (1.2) yields the desired descent guarantee

$$f(x(0)) - f(x(t)) = \int_{\tau=0}^{t} \mathrm{dist}^2(0; \partial f(x(\tau))) \qquad \text{for all } t \geq 0. \tag{1.3}$$

Evidently, exactly the same argument yields the chain rule (1.2) for *subdifferentially regular* functions. These are the functions f such that each subgradient $v \in \partial f(x)$ defines a linear lower-estimator of f up to first-order; see for example [8, Section 2.4] or [29, Definition 7.25]. Nonetheless, subdifferentially regular functions preclude "downwards cusps", and therefore still do not capture such simple examples as $f(x) = (1 - \max\{x, 0\})^2$. It is worthwhile to mention that one can not expect (1.3) to always hold. Indeed, there are pathological locally Lipschitz functions f that do not satisfy (1.3); one example is the nonconstant univariate 1-Lipschitz function whose Clarke subdifferential is identically the unit interval [4, 28].

In this work, we isolate a different structural property on the function f, which guarantees the validity of (1.2) and therefore of the descent condition (1.3). We will assume that the graph of the function f admits a partition into finitely many smooth manifolds, which fit together in a regular pattern. Formally, we require the graph of f to admit a so-called Whitney stratification, and we will call such functions Whitney stratifiable. Whitney stratifications have already figured prominently in optimization, beginning with the seminal work [2]. The main subclass of Whitney stratifiable functions consists of those functions that are semi-algebraic [21] – meaning those whose graphs can be written as a finite union of sets each defined by finitely many polynomial inequalities. Semialgebraicity is preserved under all the typical functional operations in optimization (e.g. sums, compositions, inf-projections) and therefore semi-algebraic functions are usually easy to recognize. More generally still, "semianalytic" functions [21] and those that are "definable in an o-minimal structure" are Whitney stratifiable [33]. The latter class of functions in particular shares all the robustness and analytic properties of semi-algebraic functions, while encompassing many more examples. Case in point, Wilkie [35] famously showed that there is an o-minimal structure that contains both the exponential $x \mapsto e^x$ and all semi-algebraic functions.[1]

The key observation for us, which originates in [14, Section 5.1], is that any locally Lipschitz Whitney stratifiable function necessarily satisfies the chain rule (1.2) along any absolutely continuous curve. Consequently, the descent guarantee (1.3) holds along any subgradient trajectory, and our convergence guarantees for the stochastic subgradient method become applicable. Since the composition of two definable functions is definable, it follows

[1]The term "tame" used in the title has a technical meaning. Van den Dries in [32] defines tame sets as those whose intersection with any ball is definable in some o-minimal structure. The manuscript [18] provides a nice exposition on the role of tame sets and functions in optimization.

immediately from Wilkie's o-minimal structure that nonsmooth deep neural networks built from definable pieces—such as quadratics t^2, hinge losses $\max\{0, t\}$, and log-exp $\log(1 + e^t)$ functions—are themselves definable. Hence, the results of this paper endow stochastic subgradient methods, applied to definable deep networks, with rigorous convergence guarantees.

Validity of the chain rule (1.2) for Whitney stratifiable functions is not new. It was already proved in [14, Section 5.1] for semi-algebraic functions, though identical arguments hold more broadly for Whitney stratifiable functions. These results, however, are somewhat hidden in the paper [14], which is possibly why they have thus far been underutilized. In this manuscript, we provide a self-contained review of the material from [14, Section 5.1], highlighting only the most essential ingredients and streamlining some of the arguments.

The outline of this paper is as follows. In Section 2, we fix our notation, describe basic results on differential inclusions, and specialize these results to subdifferential operators. In Section 3, we leverage results from the previous section to show that any limit point of the subgradient method is critical for f, provided that f decreases along subgradient curves. Finally, in Section 4, we verify the sufficient conditions for convergence (detailed in Section 3) for a broad class of locally Lipschitz functions, including those that are Clarke regular and Whitney stratifiable. In particular, we specialize our results to deep learning settings in Corollary 4.10. In the final Section 5, we extend the results of the previous sections to the proximal setting, by adapting the techniques of [15].

2 Preliminaries

Throughout, we will mostly use standard notation on differential inclusions, as set out for example in the monographs of Borkar [3], Clarke-Ledyaev-Stern-Wolenski [8], and Smirnov [30]. We will always equip the Euclidean space $\mathbb{R}^d$ with an inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\| := \sqrt{\langle x, x \rangle}$. The distance of a point x to a set $Q \subset \mathbb{R}^d$ will be written as $\operatorname{dist}(x; Q) := \min_{y \in Q} \|y - x\|$. The symbol $\mathcal{B}$ will denote the closed unit ball in $\mathbb{R}^d$, while $\mathcal{B}_\varepsilon(x)$ will stand for the closed ball of radius of $\varepsilon > 0$ around x. The symbol $\mathbb{R}_+$ will denote the set of nonnegative real numbers.

2.1 Absolutely continuous curves

Any continuous function $x \colon \mathbb{R}_+ \to \mathbb{R}^d$ is called a curve in $\mathbb{R}^d$. All curves in $\mathbb{R}^d$ comprise the set $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$. We will say that a sequence of function f_k converges to f in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ if f_k converges to f uniformly on compact intervals, that is, for all $T > 0$, we have

$$\lim_{k \to \infty} \sup_{t \in [0,T]} \|f_k(t) - f(t)\| = 0.$$

Recall that a curve $x \colon \mathbb{R}_+ \to \mathbb{R}^d$ is absolutely continuous if there exists a map $y \colon \mathbb{R}_+ \to \mathbb{R}^d$ that is integrable on any compact interval and satisfies

$$x(t) = x(0) + \int_0^t y(\tau)\, d\tau \qquad \text{for all } t \geq 0.$$

Moreover, then equality $y(t) = \dot{x}(t)$ holds for a.e. $t \geq 0$. Henceforth, for brevity, we will call absolutely continuous curves *arcs*. We will often use the observation that if $f\colon \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz and x is an arc, then the composition $f \circ x$ is absolutely continuous.

2.2 Differential inclusions and stochastic Euler trajectories

A set-valued map G from $\mathbb{R}^d$ to $\mathbb{R}^m$, denoted $G\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^m$, is a mapping from $\mathbb{R}^d$ to the powerset of $\mathbb{R}^m$. Thus $G(x)$ is a subset of $\mathbb{R}^m$, for each $x \in \mathbb{R}^d$. One standard extension of continuity to set-valued maps is as follows. The map G is called *upper semicontinuous* at a point x if for any $\varepsilon > 0$, there exists $\delta > 0$ satisfying

$$G(x + \delta\mathcal{B}) \subseteq G(x) + \varepsilon\mathcal{B}.$$

Notice that if $G(x)$ is a compact set for each $x \in \mathbb{R}^d$ and G is upper-semicontinuous on a compact set $C \subset \mathbb{R}^d$, then the image $G(C) := \{G(x) : x \in C\}$ is bounded.

Given a set-valued map $G\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$, an arc $x\colon \mathbb{R}_+ \to \mathbb{R}^d$ is called a *trajectory* of G if it satisfies the differential inclusion

$$\dot{x}(t) \in G(x(t)) \qquad \text{for a.e. } t \geq 0. \tag{2.1}$$

In this work, we will primarily focus on iterative algorithms that aim to asymptotically approximate the differential inclusion (2.1) using a noisy Forward Euler discretization and vanishing step-size. To this end, given a sequence of strictly positive step-sizes $\{\alpha_k\}_{k\geq 1}$, consider the iteration sequence:

$$x_{k+1} = x_k + \alpha_k\Big(y_k + M_{k+1}\Big) \qquad \text{with} \qquad y_k \in G(x_k), \tag{2.2}$$

where $\{M_k\}_{k\geq 1}$ is a sequence of random variables (the "noise") on some probability space. We aim to quantify the extent to which the iterates $\{x_k\}$ asymptotically follow a true trajectory of G. To this end, define $t_0 = 0$ and the partial sums $t_m = \sum_{k=1}^{m-1} \alpha_k$ for $m \geq 1$ and let $x(\cdot)$ be the linear interpolation of the discrete path:

$$x(t) := x_k + \frac{t - t_k}{t_{k+1} - t_k}(x_{k+1} - x_k) \qquad \text{for } t \in [t_k, t_{k+1}). \tag{2.3}$$

For each $\tau \geq 0$, define the the time-shifted curve $x^\tau(\cdot) = x(\tau + \cdot)$.

The following key theorem shows that under reasonable conditions, as $\tau_k \to \infty$, the shifted curves $\{x^{\tau_k}\}$ subsequentially converge in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ to a trajectory of (2.1). We first state the theorem and then comment on its justification.

Theorem 2.1. *Suppose that the following conditions hold.*

1. *For every $x \in \mathbb{R}^d$, the set $G(x)$ is nonempty, compact, and convex.*
2. *G is upper-semicontinous.*
3. *The sequence $\{\alpha_k\}$ is nonnegative, square summable, but not summable:*

$$\alpha_k \geq 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty, \qquad \textit{and} \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

4. Almost surely, we have $\sup_{k\geq 1}\|x_k\| < \infty$.

5. $\{M_k\}$ *is a martingale difference sequence w.r.t the increasing* σ*-fields*

$$\mathcal{F}_k = \sigma(x_j, y_j, M_j : j \leq k).$$

That is, almost surely there exists a real D *such that for all* $k \in \mathbb{N}$*, we have*

$$\mathbb{E}[M_{k+1}|\mathcal{F}_k] = 0 \qquad \text{and} \qquad \mathbb{E}[\|M_{k+1}\|^2|\mathcal{F}_k] \leq D(1 + \|x_k\|^2).$$

Then almost surely, the set of arcs $\{x^\tau(\cdot)\}_{\tau\geq 0}$ *is relatively compact in* $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$*. Moreover, almost surely, for any sequence* $\tau_n \to \infty$ *as* $n \to \infty$*, each limit point of the arcs* $\{x^{\tau_n}(\cdot)\}$ *in* $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ *is a trajectory of* G*.*

Theorem 2.1 is a small modification of [3, Theorem 5.2]. In particular, the conclusions of both theorems are the same, while the assumptions are slightly different. In [3, Theorem 5.2], condition 2 of Theorem 2.1 (upper-semicontinuity) is replaced by two alternative assumptions: (a) the graph of G is closed, and (b) there exists a constant $K > 0$ such that the inequality

$$\|y\| \leq K(1 + \|x\|) \qquad \text{holds for all } x \in \mathbb{R}^d, y \in G(x).$$

In the presence of (b), one may show that condition (a) is equivalent to the upper-semicontinuity of G. In the proof of [3, Theorem 5.2], the only purpose of assumption (b) is to ensure

$$\sup\left\{\|y\| : y \in \bigcup_{k\geq 1} G(x_k)\right\} < \infty,$$

which follows from the boundedness of $\{x_k\}$; see [3, Equation (5.2.3)]. Because we assume G is upper-semicontinuous, it must map compact sets to compact sets. Thus, the boundedness of $\{x_k\}$ automatically implies the above supremum is finite without any reference to (b). Taking this into consideration, the proof of Theorem 2.1 is therefore identical to the proof of [3, Theorem 5.2], save for the justification of the finiteness of the supremum.

It is worthwhile to mention that in the noiseless setting, $M_k = 0$ for all $k \in \mathbb{N}$, the conclusion of Theorem 2.1 holds under a weaker assumption. Namely, condition 3 can be weakened to $\alpha_k \searrow 0$ and $\sum_{k=1}^\infty \alpha_k = \infty$.

2.3 Subgradient dynamical system

In the current work, we focus on the optimization problem

$$\min_{x\in\mathbb{R}^d} f(x),$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is a locally Lipschitz function. The algorithms we consider are based on generalized derivatives, which we now review. Recall that by Rademacher's theorem, f is differentiable almost everywhere. The *Clarke subdifferential* of f at any point x is the set [8, Theorem 8.1]

$$\partial f(x) := \operatorname{conv}\left\{\lim_{i\to\infty} \nabla f(x_i) : x_i \xrightarrow{\Omega} x\right\},$$

where Ω is any full-measure subset of $\mathbb{R}^d$ such that f is differentiable at each of its points. A point $x \in \mathbb{R}^d$ is called *(Clarke) critical* for f if the inclusion $0 \in \partial f(x)$ holds. Equivalently, these are the points at which the Clarke directional derivative is nonnegative in every direction [8, Section 2.1]. A real number $r \in \mathbb{R}$ is called a *critical value* of f if there exists a critical point x satisfying $r = f(x)$.

It is standard that whenever f is locally Lipschitz continuous, the set-valued map $x \mapsto \partial f(x)$ is upper semicontinuous and its images $\partial f(x)$ are nonempty, compact, convex sets for each $x \in \mathbb{R}^d$; see for example [8, Proposition 1.5 (a,e)]. The stochastic subgradient method—the main topic of our work—is then simply the Euler recurrence (2.2) with $G = -\partial f$, that is

$$x_{k+1} = x_k - \alpha_k \Big(y_k + M_{k+1} \Big) \qquad \text{with} \qquad y_k \in \partial f(x_k), \tag{2.4}$$

where $\{\alpha_k\}_{k\geq 1}$ is a step-size sequence and $\{M_k\}_{k\geq 0}$ is a sequence of random variables (the "noise") on some probability space. As before, let us define $t_0 = 0$ and $t_m = \sum_{k=1}^{m-1} \alpha_k$ for $m \geq 1$, and let $x(\cdot)$ be the linear interpolation of the discrete path as in (2.3). For each $\tau \geq 0$, let $x^\tau(\cdot) = x(\tau + \cdot)$ be the time-shifted curves. For ease of reference, looking back at Theorem 2.1, let us isolate the following assumptions.

Assumption A (Standing Assumptions)**.**

1. *f is locally Lipschitz continuous.*
2. *The sequence $\{\alpha_k\}$ is nonnegative, square summable, but not summable:*

$$\alpha_k \geq 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty, \qquad \textit{and} \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

3. *Almost surely, the stochastic subgradient iterates are bounded:* $\sup_{k\geq 1} \|x_k\| < \infty$.
4. *$\{M_k\}$ is a martingale difference sequence w.r.t the increasing σ-fields*

$$\mathcal{F}_k = \sigma(x_j, y_j, M_j : j \leq k).$$

That is, almost surely there exists a real D such that for all $k \in \mathbb{N}$, we have

$$\mathbb{E}[M_{k+1}|\mathcal{F}_k] = 0 \qquad \textit{and} \qquad \mathbb{E}[\|M_{k+1}\|^2|\mathcal{F}_k] \leq D(1 + \|x_k\|^2).$$

Thus Theorem 2.1 directly implies the following subsequential convergence guarantee.

Corollary 2.2. *Suppose Assumption A holds. Then almost surely, the set of arcs $\{x^\tau(\cdot)\}_{\tau\geq 0}$ is relatively compact in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$. Moreover, almost surely, for any sequence $\tau_n \to \infty$ as $n \to \infty$, each limit point $z(\cdot)$ of the arcs $\{x^{\tau_n}(\cdot)\}$ in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ is a trajectory of the subgradient differential inclusion*

$$\dot{z}(t) \in -\partial f(z(t)) \qquad \textit{for a.e. } t \geq 0.$$

It is worthwhile to remark that when there is no noise in the subgradients, Assumption A can be weakened in Corollary 2.2 and in all results that follow.

Remark 1 (Weakened assumptions)**.** In Section 5, we employ a more refined variant of Theorem 2.1, found in Duchi-Ruan [15], to relax the conditions 3 and 4 in Assumption A. Indeed, boundedness may be induced by introducing proper regularization or constraints. On the other hand, the noise need only satisfy $\mathbb{E}[\|M_{k+1}\|^2|\mathcal{F}_k] \leq p(x_k)$ for some function $p : \mathbb{R}^d \to \mathbb{R}$, which is bounded on bounded sets.

In the noiseless setting, $M_k = 0$ for all $k \in \mathbb{N}$, the conclusion of Corollary 2.2 and all results that follow hold under a weaker assumption. Namely, condition 3 in Assumption A can be weakened to $\alpha_k \searrow 0$ and $\sum_{k=1}^{\infty} \alpha_k = \infty$.

3 Convergence of the stochastic subgradient method

In light of Corollary 2.2, we know that the stochastic subgradient trajectory $x(t)$ asymptotically approximates trajectories of the differential inclusion $\dot{z} \in -\partial f(z)$. Our goal is to now show that under mild regularity conditions on the continuous dynamics, every limit point of the stochastic subgradient method is critical for the problem, almost surely. To this end, we introduce the following assumption.

Assumption B.

1. **(Weak Sard)** *The set of noncritical values of f is dense in $\mathbb{R}$.*

2. **(Descent)** *Whenever $z\colon \mathbb{R}_+ \to \mathbb{R}^d$ is trajectory of the differential inclusion $\dot{z} \in -\partial f(z)$ and $z(0)$ is not a critical point of f, there exists a real $T > 0$ satisfying*

$$f(z(T)) < \sup_{t\in[0,T]} f(z(t)) \leq f(z(0)).$$

Some comments are in order. Recall that Sard's theorem guarantees that the set of critical values of any C^d-smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$ has measure zero. Thus property 1 in Assumption B asserts a very weak version of a nonsmooth Sard theorem. This is a very mild property, there mostly for technical reasons. It can fail, however, even for a C^1 smooth function on $\mathbb{R}^2$; see the famous example of [34]. Property 2 of Assumption B is more meaningful. It essentially asserts that f must locally strictly decrease along any subgradient trajectory emanating from a noncritical point.

We aim to prove the following theorem—the main result of the section.

Theorem 3.1. *Suppose that Assumptions A and B hold. Then almost surely, every limit point of $\{x_k\}_{k\geq 1}$ is critical for f and the function values $\{f(x_k)\}_{k\geq 1}$ converge.*

We prove Theorem 3.1 in Section 3.1. Upon first reading, the reader can safely skip to Section 4, which describes large classes of functions that automatically satisfy Assumption B.

3.1 Proof of Theorem 3.1

In this section, we will prove Theorem 3.1. The argument we present is rooted in the "non-escape argument" for ODEs, using f as a Lyapunov function for the continuous dynamics.

In particular, the proof we present is in the same spirit as that in [20, Theorem 5.2.1] and [15, Section 3.4.1].

Henceforth, we will suppose that Assumptions A and B hold. We first collect a few elementary lemmas.

Lemma 3.2. *Almost surely, we have* $\alpha_k \|M_{k+1}\| \to 0$ *as* $k \to \infty$.

Proof. Notice from Assumption A (items 2, 3, 4), we have

$$\mathbb{E}\left[M_{k+1} \mid \mathcal{F}_k\right] = 0 \quad \forall k \qquad\qquad \text{and} \qquad\qquad \sum_{i=0}^{\infty} \alpha_i^2 \mathbb{E}\left[\|M_{i+1}\|^2 \mid \mathcal{F}_i\right] < \infty.$$

Define the L^2 martingale $X_k = \sum_{i=1}^{k} \alpha_i M_{i+1}$. Thus the limit $\langle X \rangle_\infty$ of the predictable compensator

$$\langle X \rangle_k := \sum_{i=1}^{k} \alpha_i^2 \mathbb{E}\left[\|M_{i+1}\|^2 \mid \mathcal{F}_i\right],$$

exists. Applying [13, Theorem 5.3.33(a)], we deduce that almost surely X_k converges to a finite limit, which directly implies $\alpha_k \|M_{k+1}\| \to 0$ as $k \to \infty$, as claimed. □

Lemma 3.3. *Almost surely, we have* $\lim_{k\to\infty} \|x_{k+1} - x_k\| = 0$.

Proof. From the definition of the subgradient iteration, we have

$$\|x_{k+1} - x_k\| \leq \alpha_k \|y_k\| + \alpha_k \|M_{k+1}\|.$$

Assumption A (items 1,2,3) directly imply $\alpha_k \|y_k\| \to 0$ almost surely, while Lemma 3.2 guarantees $\alpha_k \|M_{k+1}\| \to 0$ almost surely. The result follows. □

Lemma 3.4. *Almost surely, it holds:*

$$\liminf_{t\to\infty} f(x(t)) = \liminf_{k\to\infty} f(x_k) \qquad \textit{and} \qquad \limsup_{t\to\infty} f(x(t)) = \limsup_{k\to\infty} f(x_k). \tag{3.1}$$

Proof. Clearly, the inequalities $\leq$ and $\geq$ hold in the two equalities above, respectively. We will argue that the reverse inequalities are valid. To this end, let $\tau_i \to \infty$ be an arbitrary sequence with $x(\tau_i)$ converging to some point x^* as $i \to \infty$.

For each index i, define the breakpoint $k_i = \max\{k \in \mathbb{N} : t_k \leq \tau_i\}$. Then by the triangle inequality, we have

$$\|x_{k_i} - x^*\| \leq \|x_{k_i} - x(\tau_i)\| + \|x(\tau_i) - x^*\| \leq \|x_{k_i} - x_{k_i+1}\| + \|x(\tau_i) - x^*\|$$

Lemma 3.3 implies that the right-hand-side tends to zero, and hence $x_{k_i} \to x^*$. Continuity of f then directly yields the guarantee $f(x_{k_i}) \to f(x^*)$.

In particular, we may take $\tau_i \to \infty$ to be a sequence realizing $\liminf_{t\to\infty} f(x(t))$. Since the curve $x(\cdot)$ is bounded, we may suppose that up to taking a subsequence, $x(\tau_i)$ converges to some point x^*. We therefore deduce

$$\liminf_{k\to\infty} f(x_k) \leq \lim_{i\to\infty} f(x_{k_i}) = f(x^*) = \liminf_{t\to\infty} f(x(t)),$$

thereby establishing the first equality in (A.1). The second equality follows analogously. □

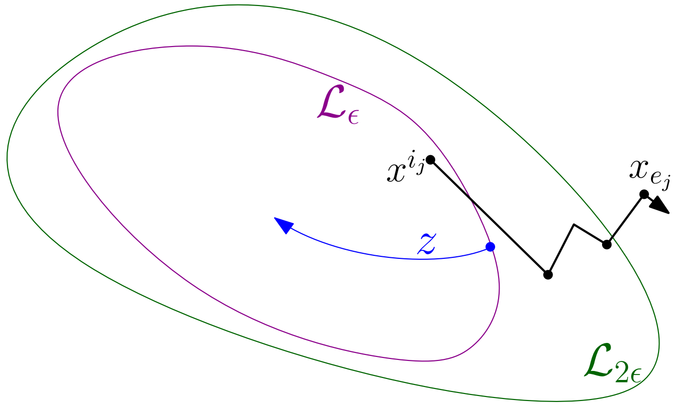

Figure 1: Illustration of the attraction argument

Before proving the main theorem, we prove the following preparatory lemma.

Lemma 3.5. *Almost surely, the values $f(x(t))$ have a limit as $t \to \infty$.*

Proof. Assume we are in the probability 1 event in which Lemmas 3.2, 3.3, and 3.4 hold. Without loss of generality, suppose $0 = \liminf_{t\to\infty} f(x(t))$. For each $r \in \mathbb{R}$, define the sublevel set

$$\mathcal{L}_r := \{x \in \mathbb{R}^d : f(x) \le r\}.$$

Let $\epsilon > 0$ be any real number such that ϵ is a non-critical value of f. Note that by Assumption B, we can let $\epsilon > 0$ be as small as we wish. By the first equality in (A.1), there are infinitely many indices k such that $f(x_k) < \epsilon$. The following elementary observation shows that for all large k, if x_k lies in $\mathcal{L}_\epsilon$ then the next iterate x_{k+1} lies in $\mathcal{L}_{2\epsilon}$.

Claim 1. For all sufficiently large indices $k \in \mathbb{N}$, the implication holds:

$$x_k \in \mathcal{L}_\epsilon \quad \Longrightarrow \quad x_{k+1} \in \mathcal{L}_{2\epsilon}.$$

Proof. Since the sequence $\{x_k\}_{k\ge0}$ is bounded, it is contained in some compact set $C \subset \mathbb{R}^n$. From continuity, we have

$$\mathrm{cl}\,(\mathbb{R}^d \setminus \mathcal{L}_{2\epsilon}) = \mathrm{cl}\,(f^{-1}(2\epsilon, \infty)) \subseteq f^{-1}[2\epsilon, \infty).$$

It follows that the two sets $C \cap \mathcal{L}_\epsilon$ and $\mathrm{cl}\,(\mathbb{R}^d \setminus \mathcal{L}_{2\epsilon})$ do not intersect. Since $C \cap \mathcal{L}_\epsilon$ is compact, we deduce that it is well separated from $\mathbb{R}^d \setminus \mathcal{L}_{2\epsilon}$; that is, there exists $\alpha > 0$ satisfying:

$$\min\{\|w - v\| : w \in C \cap \mathcal{L}_\epsilon,\ v \notin \mathcal{L}_{2\epsilon}\} \ge \alpha > 0.$$

In particular $\mathrm{dist}(x_k; \mathbb{R}^d \setminus \mathcal{L}_{2\epsilon}) \ge \alpha > 0$, whenever x_k lies in $\mathcal{L}_\epsilon$. Taking into account Lemma 3.3, we deduce $\|x_{k+1} - x_k\| < \alpha$ for all large k, and therefore $x_k \in \mathcal{L}_\epsilon$ implies $x_{k+1} \in \mathcal{L}_{2\epsilon}$, as claimed. □

Let us define now the following sequence of iterates. Let $i_1 \in \mathbb{N}$ be the first index satisfying

1. $x_{i_1} \in \mathcal{L}_\epsilon$,

2. $x_{i_1+1} \in \mathcal{L}_{2\epsilon} \setminus \mathcal{L}_\epsilon$, and

3. defining the exit time $e_1 := \min\{e \geq i_1 : x_e \notin \mathcal{L}_{2\epsilon} \setminus \mathcal{L}_\epsilon\}$, the iterate x_{e_1} lies in $\mathbb{R}^d \setminus \mathcal{L}_{2\epsilon}$.

Then let $i_2 > i_1$ be the next smallest index satisfying the same property, and so on. See Figure 1 for an illustration. The following claim will be key.

Claim 2. This process must terminate, that is $\{x_k\}$ exits $\mathcal{L}_{2\epsilon}$ only finitely many times.

Before proving the claim, let us see how it immediately yields the validity of the theorem. To this end, observe that Claims 1 and 2 immediately imply $x_k \in \mathcal{L}_{2\epsilon}$ for all large k. Since $\epsilon > 0$ can be made arbitrarily small, we deduce $\lim_{k\to\infty} f(x_k) = 0$. Equation (A.1) then directly implies $\lim_{t\to\infty} f(x(t)) = 0$, as claimed.

Proof of Claim 2. To verify the claim, suppose that the process does not terminate. Thus we obtain an increasing sequence of indices $i_j \in \mathbb{N}$ with $i_j \to \infty$ as $j \to \infty$. Set $\tau_j = t_{i_j}$ and consider the curves $x^{\tau_j}(\cdot)$ in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$. Then up to a subsequence, Corollary 2.2 shows that the curves $x^{\tau_j}(\cdot)$ converge to some arc $z(\cdot)$ satisfying

$$\dot{z}(t) \in -\partial f(z(t)) \qquad \text{for a.e. } t \geq 0.$$

By construction, we have $f(x_{i_j}) \leq \epsilon$ and $f(x_{i_j+1}) > \epsilon$. Letting L denote the Lipschitz constant of f on $\{x_k\}_{k\geq 0}$, we therefore deduce

$$\epsilon \geq f(x_{i_j}) \geq f(x_{i_j+1}) - L\|x_{i_j} - x_{i_j+1}\| \geq \epsilon - L\|x_{i_j} - x_{i_j+1}\|.$$

Lemma 3.3 directly implies $\|x_{i_j} - x_{i_j+1}\| \to 0$ as $j \to \infty$, and therefore by continuity $f(z(0)) = \lim_{j\to\infty} f(x_{i_j}) = \epsilon$. In particular, $z(0)$ is not a critical point of f. Hence, Assumption B yields real $T > 0$ such that

$$f(z(T)) < \sup_{t\in[0,T]} f(z(t)) \leq f(z(0)) = \epsilon.$$

In particular, there exists a real $\delta > 0$ satisfying $f(z(T)) \leq \epsilon - 2\delta$.

Appealing to uniform convergence on $[0, T]$, we conclude

$$\sup_{t\in[0,T]} |f(z(t)) - f(x^{\tau_j}(t))| < \epsilon,$$

for all large $j \in \mathbb{N}$, and therefore

$$\sup_{t\in[0,T]} f(x^{\tau_j}(t)) \leq \sup_{t\in[0,T]} f(z(t)) + \sup_{t\in[0,T]} |f(z(t)) - f(x^{\tau_j}(t))| \leq 2\epsilon.$$

Hence, for all large j, all the curves x^{τ_j} map $[0, T]$ into $\mathcal{L}_{2\epsilon}$. We conclude that the exit time satisfies

$$t_{e_j} > \tau_j + T \qquad \text{for all large } j.$$

We will show that the bound $f(z(T)) \leq \epsilon - 2\delta$ yields the opposite inequality $t_{e_j} \leq \tau_j + T$, which will lead to a contradiction.

To that end, let

$$\ell_j = \max\{\ell \in \mathbb{N} \mid \tau_j \leq t_\ell \leq \tau_j + T\},$$

be the last discrete index before T. Because $\alpha_k \to 0$ as $k \to \infty$, we have that $\ell_j \geq i_j + 1$ for all large j. We will now show that for all large j, we have

$$f(x_{t_{\ell_j}}) < \epsilon - \delta,$$

which implies $t_{e_j} < t_{\ell_j} \leq \tau_j + T$. Indeed, observe

$$f(x_{\ell_j}) = f(x^{\tau_j}(t_{\ell_j} - \tau_j)) = f(x^{\tau_j}(T)) + f(x^{\tau_j}(t_{\ell_j} - \tau_j)) - f(x^{\tau_j}(T)).$$

Letting L be the Lipschitz constant of f on a compact set containing $\{x(t)\}_{t\geq 0} \cup \{z(T)\}$, we deduce

$$\begin{aligned}
|f(x_{\ell_j}) - f(z(T))| &\leq |f(x^{\tau_j}(T)) - f(z(T))| + |f(x^{\tau_j}(t_{\ell_j} - \tau_j)) - f(x^{\tau_j}(T))| \\
&\leq L\|x^{\tau_j}(T) - z(T)\| + L\|x^{\tau_j}(t_{\ell_j} - \tau_j) - x^{\tau_j}(T)\| \\
&\leq L\|x^{\tau_j}(T) - z(T)\| + L\|x_{\ell_j} - x_{\ell_j+1}\|.
\end{aligned}$$

Uniform convergence of x^{τ_j} to z implies that the first term on the right-hand-side tends to zero, while the second term tends to zero by Lemma 3.3. Thus for all large j, the inequality $f(x_{\ell_j}) < \epsilon - \delta$ holds, which is the desired contradiction. □

The proof of the lemma is now complete. □

We can now prove the main convergence theorem.

Proof of Theorem 3.1. Let x^* be a limit point of $\{x_k\}$ and suppose for the sake of contradiction that x^* is not critical. Let i_j be the indices satisfying $x_{i_j} \to x^*$ as $j \to \infty$. Let $z(\cdot)$ be the subsequential limit of the curves $x^{t_{i_j}}(\cdot)$ in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ guaranteed to exist by Corollary 2.2. Since $z(0) = x^*$ is noncritical, Assumption B guarantees that there exists a real $T > 0$ satisfying

$$f(z(T)) < \sup_{t\in[0,T]} f(z(t)) \leq f(x^*).$$

On the other hand, we successively deduce

$$f(z(T)) = \lim_{j\to\infty} f(x^{t_{i_j}}(T)) = \lim_{t\to\infty} f(x(t)) = f(x^*),$$

where the last equality follows from Lemma 3.5 and continuity of f. We have thus arrived at a contradiction, and the theorem is proved. □

4 Verifying the assumptions

In light of Theorem 3.1, it is essential to isolate function classes that automatically satisfy Property 2 in Assumption B. In this section, we do exactly that, focusing on two problem classes: (1) subdifferentially regular functions and (2) those functions having Whitney stratifiable graphs. We will see that the latter problem class also automatically satisfies Property 1 in Assumption B.

The material in this section is not new. In particular, the results of this section have appeared in [14, Section 5.1]. These results, however, are somewhat hidden in the paper [14] and are difficult to parse. Moreover, at the time of writing [14, Section 5.1], there was no clear application of the techniques, in contrast to our current paper. Since we do not expect the readers to be experts in variational analysis and semialgebraic geometry, we provide here a self-contained treatment, highlighting only the most essential ingredients and streamlining some of the arguments.

Let us begin with the following definition, whose importance for verifying Property 2 in Assumption B will become clear shortly.

Definition 4.1 (Chain rule)**.** Consider a locally Lipschitz function f on $\mathbb{R}^d$. We will say that the subdifferential ∂f *admits a chain rule* if for any arc $z\colon \mathbb{R}_+ \to \mathbb{R}^d$, equality

$$(f \circ z)'(t) = \langle \partial f(z(t)), \dot{z}(t)\rangle \qquad \text{holds for a.e. } t \geq 0.$$

The importance of the chain rule becomes immediately clear with the following lemma.

Lemma 4.2. *Consider a locally Lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}$ and suppose that ∂f admits a chain rule. Let $z\colon \mathbb{R}_+ \to \mathbb{R}^d$ be any arc satisfying the differential inclusion*

$$\dot{z}(t) \in -\partial f(z(t)) \qquad \textit{for a.e. } t \geq 0.$$

Then equality $\|\dot{z}(t)\| = \operatorname{dist}(0, \partial f(z(t)))$ holds for a.e. $t \geq 0$, and therefore we have the estimate

$$f(z(0)) - f(z(t)) = \int_0^t \operatorname{dist}^2\left(0; \partial f(z(\tau))\right)\, d\tau, \qquad \forall t \geq 0. \tag{4.1}$$

In particular, property 2 of Assumption B holds.

Proof. Fix a real $t \geq 0$ satisfying $(f \circ z)'(t) = \langle \partial f(z(t)), \dot{z}(t)\rangle$. Observe then the equality

$$0 = \langle \partial f(z(t)) - \partial f(z(t)), \dot{z}(t)\rangle. \tag{4.2}$$

To simplify the notation, set $S := \partial f(z(t))$, $W := \operatorname{span}(S - S)$, and $y := -\dot{z}(t)$. Appealing to (4.2), we conclude $y \in W^\perp$, and therefore trivially we have

$$y \in (y + W) \cap W^\perp.$$

Basic linear algebra implies $\|y\| = \operatorname{dist}(0; y + W)$. Noting $\partial f(z(t)) \subset y + W$, we deduce $\|\dot{z}(t)\| \leq \operatorname{dist}(0; \partial f(z(t)))$ as claimed. Since the reverse inequality trivially holds, we obtain the claimed equality, $\|\dot{z}(t)\| = \operatorname{dist}(0; \partial f(z(t)))$.

Since ∂f admits a chain rule, we conclude for a.e. $\tau \geq 0$ the estimate

$$(f \circ z)'(\tau) = \langle \partial f(z(\tau)), \dot{z}(\tau)\rangle = -\|\dot{z}(\tau)\|^2 = -\operatorname{dist}^2\left(0; \partial f(z(\tau))\right).$$

Since f is locally Lipschitz, the composition $f \circ z$ is absolutely continuous. Hence integrating over the interval $[0, t]$ yields (5.4).

Suppose now that the point $z(0)$ is noncritical. Then by upper semi-continuity of ∂f, the exists $T > 0$ such that $z(\tau)$ is noncritical for any $\tau \in [0, T]$. It follows immediately that the value $\int_0^t \operatorname{dist}^2\left(0; \partial f(z(\tau))\right)\, d\tau$ is strictly increasing in $t \in [0, T]$, and therefore by (5.4) that $f \circ z$ is strictly decreasing. Hence item 2 of Assumption B holds, as claimed. □

Thus property 2 of Assumption B is sure to hold as long as ∂f admits a chain rule. In the following two sections, we identify two different function classes that indeed admit the chain rule.

4.1 Subdifferentially regular functions

The first function class we consider consists of subdifferentially regular functions. Such functions play a prominent role in variational analysis due to their similarity with convex functions; we refer the reader to the monograph [29] for details. In essence, subdifferential regularity forbids downward facing cusps in the graph of the function; e.g. $f(x) = -|x|$ is not subdifferentially regular. Here is the formal definition.

Definition 4.3 (Subdifferential regularity)**.** A locally Lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}$ is *subdifferentially regular* at a point $x \in \mathbb{R}^d$ if every subgradient $v \in \partial f(x)$ yields an affine minorant of f up to first-order:

$$f(y) \geq f(x) + \langle v, y - x\rangle + o(\|y - x\|) \qquad \text{as } y \to x.$$

The following lemma shows that if a locally Lipschitz function f is subdifferentially regular, then the subdifferential ∂f indeed admits a chain rule.

Lemma 4.4 (Chain rule under subdifferential regularity)**.** *If a locally Lipschitz function f is subdifferentially regular, then ∂f admits a chain rule and therefore item 2 of Assumption B holds.*

Proof. Consider an arc $x\colon \mathbb{R}_+ \to \mathbb{R}^d$. Since, x and $f \circ x$ are absolutely continuous, both are differentiable almost everywhere. Then for any such $t \geq 0$ and any subgradient $v \in \partial f(x(t))$, we conclude

$$\begin{aligned}(f\circ x)'(t) = \lim_{r \searrow 0} \frac{f(x(t+r)) - f(x(t))}{r} &\geq \lim_{r\searrow 0} \frac{\langle v, x(t+r) - x(t)\rangle + o(\|x(t+r) - x(t)\|)}{r} \\ &= \langle v, \dot{x}(t)\rangle.\end{aligned}$$

Instead, equating $(f \circ x)'(t)$ with the left limit of the difference quotient yields the reverse inequality $(f \circ x)'(t) \leq \langle v, \dot{x}(t)\rangle$. Thus ∂f admits a chain rule and item 2 of Assumption B holds by Lemma 4.2. □

Though subdifferentially regular functions are widespread in applications, they preclude "downwards cusps", and therefore do not capture such simple examples as $f(x, y) = (|x| - |y|)^2$ and $f(x) = (1 - \max\{x, 0\})^2$. The following section concerns a different function class that does capture these two nonpathological examples.

4.2 Stratifiable functions

As we saw in the previous section, subdifferential regularity is a local property that implies the desired item 2 of Assumption B. In this section, we instead focus on a broad class of functions satisfying a global geometric property, which eliminates pathological examples from consideration.

Before giving a formal definition, let us fix some notation. A set $M \subset \mathbb{R}^d$ is a C^p *smooth manifold* if there is an integer $r \in \mathbb{N}$ such that around any point $x \in M$, there is a neighborhood U and a C^p-smooth map $F\colon U \to \mathbb{R}^{d-r}$ with $\nabla F(x)$ of full rank and satisfying $M \cap U = \{y \in U : F(y) = 0\}$. If this is the case, the *tangent* and *normal spaces* to M at x are defined to be $T_M(x) := \text{Null}(\nabla F(x))$ and $N_M(x) := (T_M(x))^\perp$, respectively.

Definition 4.5 (Whitney stratification)**.** A *Whitney C^p-stratification* $\mathcal{A}$ of a set $Q \subset \mathbb{R}^d$ is a partition of Q into finitely many nonempty C^p manifolds, called *strata*, satisfying the following compatibility conditions.

1. **Frontier condition:** For any two strata L and M, the implication
$$L \cap \operatorname{cl} M \neq \emptyset \quad \Longrightarrow \quad L \subset \operatorname{cl} M \quad \text{holds.}$$

2. **Whitney condition (a):** For any sequence of points z_k in a stratum M converging to a point $\bar{z}$ in a stratum L, if the corresponding normal vectors $v_k \in N_M(z_k)$ converge to a vector v, then the inclusion $v \in N_L(\bar{z})$ holds.

A function $f \colon \mathbb{R}^d \to \mathbb{R}$ is *Whitney C^p-stratifiable* if its graph admits a Whitney C^p-stratification.

The definition of the Whitney stratification invokes two conditions, one topological and the other geometric. The frontier condition simply says that if one stratum L intersects the closure of another M, then L must be fully contained in the closure $\operatorname{cl} M$. In particular, the frontier condition endows the strata with a partial order $L \preceq M \Leftrightarrow L \subset \operatorname{cl} M$. The Whitney condition (a) is geometric. In short, it asserts that limits of normals along a sequence x_i in a stratum M are themselves normal to the stratum containing the limit of x_i.

The following discussion of Whitney stratifications follows that in [2]. Consider a Whitney C^p-stratification $\{M_i\}$ of the graph of a locally Lipschitz function $f \colon \mathbb{R}^d \to \mathbb{R}$. Let $\{\mathcal{M}_i\}$ be the manifolds obtained by projecting $\{M_i\}$ on $\mathbb{R}^d$. An easy argument using the constant rank theorem shows that the partition $\{\mathcal{M}_i\}$ of $\mathbb{R}^d$ is itself a Whitney C^p-stratification and the restriction of f to each stratum $\{\mathcal{M}_i\}$ is C^p-smooth. Whitney condition (a) directly yields the following consequence [2, Proposition 4]. For any stratum $\mathcal{M}$ and any point $x \in \mathcal{M}$, we have
$$(v, -1) \in N_M(x, f(x)) \qquad \text{for all} \quad v \in \partial f(x), \tag{4.3}$$
and
$$\partial f(x) \subset \nabla g(x) + N_{\mathcal{M}}(x), \tag{4.4}$$
where $g \colon \mathbb{R}^d \to \mathbb{R}$ is any C^1-smooth function agreeing with f on a neighborhood of x in $\mathcal{M}$.

The following theorem, which first appeared in [2, Corollary 5], shows that Whitney stratifiable functions automatically satisfy the weak Sard property of Assumption B. We present a quick argument here for completeness. It is worthwhile to mention that such a Sard type result holds more generally for any stratifiable set-valued map; see the original work [17] or the monograph [19, Section 8.4].

Lemma 4.6 (Stratified Sard)**.** *The set of critical values of any Whitney C^d-stratifiable locally Lipschitz function $f \colon \mathbb{R}^d \to \mathbb{R}$ has zero measure. In particular, item 1 of Assumption B holds.*

Proof. Let $\{M_i\}$ be the strata of a Whitney C^d stratification of the graph of f. Let $\pi_i \colon M_i \to \mathbb{R}$ be the restriction of the orthogonal projection $(x, r) \mapsto r$ to the manifold M_i. We claim that each critical value of f is a critical value (in the classical analytic sense) of π_i, for some index i. To see this, consider a critical point x of f and let M_i be the stratum of $\operatorname{gph} f$ containing $(x, f(x))$. Since x is critical for f, appealing to (4.3) yields the inclusion $(0, -1) \in N_{M_i}(x, f(x))$ and therefore the equality $\pi_i(T_{M_i}(x, f(x))) = \{0\} \subsetneq \mathbb{R}$. Hence

$(x, f(x))$ is a critical point of π_i and $f(x)$ its critical value, thereby establishing the claim. Since the set of critical values of each map π_i has zero measure by the standard Sard's theorem, and there are finitely many strata, it follows that the set of critical values of f also has zero measure. □

Next, we prove the chain rule for the subdifferential of any Whitney stratifiable function.

Theorem 4.7. *If a locally Lipschitz function f is Whitney C^1-stratifiable, then ∂f admits a chain rule and therefore item 2 of Assumption B holds.*

Proof. Let $\{M_i\}$ be the Whitney C^1-stratification of gph f and let $\{\mathcal{M}_i\}$ be its coordinate projection onto $\mathbb{R}^d$. Fix an arc $x\colon \mathbb{R}^d \to \mathbb{R}$. It is easy to see that for a.e. $t \geq 0$, both x and $f \circ x$ are differentiable at t and the implication holds:[2]

$$x(t) \in \mathcal{M}_i \quad \Longrightarrow \quad \dot{x}(t) \in T_{\mathcal{M}_i}(x(t)).$$

Fix such a real $t > 0$ and let $\mathcal{M}$ be a stratum containing $x(t)$. Since $\dot{x}(t)$ is tangent to $\mathcal{M}$ at $x(t)$, there exists a C^1-smooth curve $\gamma\colon (-1, 1) \to \mathcal{M}$ satisfying $\gamma(0) = x(t)$ and $\dot{\gamma}(0) = \dot{x}(t)$. Let $g\colon \mathbb{R}^d \to \mathbb{R}$ be any C^1 function agreeing with f on a neighborhood of $x(t)$ in $\mathcal{M}$. We claim that $(f \circ x)'(t) = (f \circ \gamma)'(0)$. Indeed, letting L be a Lipschitz constant of f around $x(t)$, we deduce

$$\begin{aligned}(f \circ x)'(t) = \lim_{r \to 0} \frac{f(x(t + r)) - f(x(t))}{r} &= \lim_{r \to 0} \frac{f(x(t + r)) - f(\gamma(r)) + f(\gamma(r)) - f(\gamma(0))}{r} \\ &= \lim_{r \to 0} \frac{f(x(t + r)) - f(\gamma(r))}{r} + (f \circ \gamma)'(0).\end{aligned}$$

Notice $\frac{|f(x(t+r)) - f(\gamma(r))|}{r} \leq L \left\| \frac{x(t+r) - x(t) + \gamma(0) - \gamma(r)}{r} \right\| \to L\|\dot{x}(t) - \dot{\gamma}(0)\| = 0$ as $r \to 0$. Thus

$$(f \circ x)'(t) = (f \circ \gamma)'(0) = (g \circ \gamma)'(0) = \langle \nabla g(x), \dot{\gamma}(0) \rangle = \langle \partial f(x(t)), \dot{x}(t) \rangle,$$

where the last equality follows from (4.4). □

Putting together Theorem 3.1 and 4.7 and Lemma 4.6, we arrive at the main result of our paper.

Corollary 4.8. *Suppose that a locally Lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfies Assumption A and is C^d-stratifiable, and let $\{x_k\}_{k \geq 0}$ be the iterates produced by the stochastic subgradient method. Then almost surely, every limit point of the iterates $\{x_k\}_{k \geq 0}$ is critical for f and the function values $\{f(x_k)\}_{k \geq 0}$ converge.*

Verifying Whitney stratifiability is often an easy task. Indeed, there are a number of well-known and easy to recognize function classes, whose members are automatically Whitney stratifiable. We now briefly review such classes, beginning with the semianalytic setting.

[2]Indeed, for any manifold $\mathcal{M}$ and any fixed t such that $x(t) \in \mathcal{M}$ but $\dot{x}(t) \notin T_{x(t)}\mathcal{M}$, there exists a finite open interval I_t with rational endpoints such that $x(I_t \setminus \{t\}) \cap \mathcal{M} = \emptyset$. Therefore, the set of such parameters t is covered by a countable collection of disjoint open intervals and is therefore itself countable, hence, zero measure.

A closed set Q is called *semianalytic* if it can be written as a finite union of sets, each having the form

$$\{x \in \mathbb{R}^d : p_i(x) \leq 0 \quad \text{for } i = 1, \ldots, \ell\}$$

for some real-analytic functions $p_1, p_2, \ldots, p_\ell$ on $\mathbb{R}^d$. If the functions $p_1, p_2, \ldots, p_\ell$ in the description above are polynomials, then Q is said to be a *semialgebraic* set. A well-known result of Łojasiewicz [21] shows that any semianalytic set admits a Whitney C^∞ stratification. Thus, the results of this paper apply to functions with semianalytic graphs. While the class of such functions is broad, it is sometimes difficult to recognize its members as semianalyticity is not preserved under some basic operations, such as projection onto a linear subspace. On the other hand, there are large subclasses of semianalytic sets that are easy to recognize.

For example, every semialgebraic set is semianalytic, but in contrast to the semianalytic case, semi-algebraic sets are stable with respect to all boolean operations and projections onto subspaces. The latter property is a direct consequence of the celebrated Tarski-Seidenberg Theorem. Moreover, semialgebraic sets are typically easy to recognize using quantifier elimination; see [10, Chapter 2] for a detailed discussion. Importantly, compositions of semialgebraic functions are semialgebraic.

A far reaching axiomatic extension of semialgebraic sets, whose members are also Whitney stratifiable, is built from "o-minimal structures". Loosely speaking, sets that are definable in an o-minimal structure share the same robustness properties and attractive analytic features as semialgebraic sets. For the sake of completeness, let us give a formal definition, following Coste [9] and van den Dries-Miller [33].

Definition 4.9 (o-minimal structure)**.** An *o-minimal structure* is a sequence of Boolean algebras $\mathcal{O}_d$ of subsets of $\mathbb{R}^d$ such that for each $d \in \mathbb{N}$:

(i) if A belongs to $\mathcal{O}_d$, then $A \times \mathbb{R}$ and $\mathbb{R} \times A$ belong to $\mathcal{O}_{d+1}$;

(ii) if $\pi \colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ denotes the coordinate projection onto $\mathbb{R}^d$, then for any A in $\mathcal{O}_{d+1}$ the set $\pi(A)$ belongs to $\mathcal{O}_d$;

(iii) $\mathcal{O}_d$ contains all sets of the form $\{x \in \mathbb{R}^d : p(x) = 0\}$, where p is a polynomial on $\mathbb{R}^d$;

(iv) the elements of $\mathcal{O}_1$ are exactly the finite unions of intervals (possibly infinite) and points.

The sets A belonging to $\mathcal{O}_d$, for some $d \in \mathbb{N}$, are called *definable in the o-minimal structure.*

As in the semialgebraic setting, any function definable in an o-minimal structure admits a Whitney C^p stratification, for any $p \geq 1$ (see e.g. [33]). Beyond semialgebraicity, Wilkie showed that that there is an o-minimal structure that simultaneously contains both the graph of the exponential function $x \mapsto e^x$ and all semi-algebraic sets [35].

A corollary for deep learning

Since the composition of two definable functions is definable, we conclude that nonsmooth deep neural networks built from definable pieces—such as ReLU, quadratics t^2, hinge losses $\max\{0, t\}$, and SoftPlus $\log(1 + e^t)$ functions—are themselves definable. Hence, the results of this paper endow stochastic subgradient methods, applied to definable deep networks, with rigorous convergence guarantees. Due to the importance of subgradient methods in deep learning, we make this observation precise in the following corollary which provides a rigorous convergence guarantee for a wide class of deep learning loss functions that are recursively defined, including convolutional neural networks, recurrent neural networks, and feed-forward networks.

Corollary 4.10 (Deep networks)**.** *For each given data pair (x_j, y_j) with $j = 1, \ldots, n$, recursively define:*

$$a_0 = x_j, \qquad a_i = \rho_i(V_i(w)a_{i-1}) \;\; \forall i = 1, \ldots, L, \qquad f(w; x_j, y_j) = \ell(y_j, a_L),$$

where

1. *$V_i(\cdot)$ are linear maps into the space of matrices.*
2. *$\ell(\cdot;\cdot)$ is any definable loss function, such as the logistic loss $\ell(y; z) = \log(1 + e^{-yz})$, the hinge loss $\ell(y; z) = \max\{0, 1 - yz\}$, absolute deviation loss $\ell(y; z) = |y - z|$, or the square loss $\ell(y; z) = \frac{1}{2}(y - z)^2$.*
3. *ρ_i are definable activation functions applied coordinate wise, such as those whose domain can be decomposed into finitely many intervals on which it coincides with $\log t$, $\exp(t)$, $\max(0, t)$, or $\log(1 + e^t)$.*

Let $\{w_k\}_{k\geq 1}$ be the iterates produced by the stochastic subgradient method on the deep neural network loss $f(w) := \sum_{j=1}^{n} f(w; x_j, y_j)$, and suppose that the standing assumption A holds.[3] *Then almost surely, every limit point w^* of the iterates $\{w_k\}_{k\geq 1}$ is critical for f, meaning $0 \in \partial f(w^*)$, and the function values $\{f(w_k)\}_{k\geq 1}$ converge.*

5 Proximal extentions

In this section, we explain how most of the results in the previous sections can be extended to a "proximal" setting and comment on sufficient conditions to ensure boundedness of subgradient iterates. Indeed, the arguments follow quickly by adapting the techniques developed by Duchi-Ruan [15]. For the sake of completeness, we have placed all the proofs in Appendix A.

Setting the stage, we consider the composite optimization problem

$$\min_{x \in \mathcal{X}} \; \varphi(x) := f(x) + g(x), \tag{5.1}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is a locally Lipschitz function, $g\colon \mathbb{R}^d \to \mathbb{R}$ is a finite convex function, and $\mathcal{X}$ is a closed convex set. As is standard in the literature on proximal methods, we will say that $x \in \mathcal{X}$ is a *critical point of* the composite problem (5.1) if the inclusion holds:

$$0 \in \partial f(x) + \partial g(x) + N_{\mathcal{X}}(x). \tag{5.2}$$

[3] In the assumption, replace x_k with w_k, since we now use w_k to denote the stochastic subgradient iterates.

Here, $\partial f(x)$ and $\partial g(x)$ denote the Clarke subdifferentials of f and g at x, respectively. Note that since g is convex, $\partial g(x)$ reduces to the subdifferential in the standard convex analytic sense [27, Section 23]. The symbol $N_{\mathcal{X}}(x)$ denotes the normal cone to the convex set $\mathcal{X}$ at x. It follows from [29, Corollary 10.9] that local minimizers of (5.1) are necessarily critical in the sense of (5.2). A real $r \in \mathbb{R}$ is called a *critical value of* (5.1) if equality, $r = f(x) + g(x)$, holds for some critical point x of (5.1).

In this section, we will take a slightly different perspective on the way that the noise impacts the subgradient computation, following the influential work [22]. Namely, fix a probability space $(\Omega, \mathcal{F}, P)$ and equip $\mathbb{R}^d$ with the Borel σ-algebra. We suppose that there exists a measurable mapping $\zeta\colon \mathbb{R}^d \times \Omega \to \mathbb{R}^d$ satisfying:

$$\mathbb{E}_\omega\left[\zeta(x,\omega)\right] \in \partial f(x) \qquad \text{for all } x \in \mathbb{R}^d.$$

In this section, we investigate the *proximal stochastic subgradient method*; given an iterate $x_k \in \mathcal{X}$, the scheme performs the update

$$\left\{\begin{array}{l} \text{Sample } \omega_k \sim P \\ x_{k+1} = \underset{x\in\mathcal{X}}{\operatorname{argmin}} \ \left\{g(x) + \langle \zeta(x_k,\omega_k), x - x_k\rangle + \frac{1}{2\alpha_k}\|x - x_k\|^2\right\} \end{array}\right\}. \tag{5.3}$$

Here $\{\alpha_k\}_{k\geq 1}$ is a positive control sequence. We will analyze the algorithm under the following assumptions, akin to Assumption A used in the previous sections.

Assumption C (Standing assumptions)**.**

1. *f is locally Lipschitz continuous.*

2. *g is a finite convex function, bounded from below, and $\mathcal{X}$ is a closed convex set.*

3. *The sequence $\{\alpha_k\}_{k\geq 1}$ is nonnegative, square summable, but not summable:*

$$\alpha_k \geq 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty, \qquad \textit{and} \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

4. *Almost surely, the iterates are bounded:* $\sup_{k\geq 1}\|x_k\| < \infty$.

5. *There exists a function $p\colon \mathbb{R}^d \to [0,\infty)$, that is bounded on bounded sets, such that*

$$\mathbb{E}_\omega\left[\zeta(x,\omega)\right] \in \partial f(x) \qquad \textit{and} \qquad \mathbb{E}_\omega\left[\|\zeta(x,\omega)\|^2\right] \leq p(x) \qquad \textit{for all } x \in \mathcal{X}.$$

6. *For every point $z \in \mathbb{R}^d$, there exists an $\epsilon > 0$ such that*

$$\mathbb{E}\left[\sup_{x\in\mathcal{B}_\epsilon(z)} \|\zeta(x,\omega)\|\right] < \infty.$$

As before, let us define $t_0 = 0$ and $t_m = \sum_{k=1}^{m-1} \alpha_k$ for $m \geq 1$, and let $x(\cdot)$ be the linear interpolation of the discrete path as in (2.3). For each $\tau \geq 0$, let $x^\tau(\cdot) = x(\tau + \cdot)$ be the time-shifted curves. Henceforth, define the set-valued map

$$H(x) = -\partial f(x) - \partial g(x) - N_{\mathcal{X}}(x).$$

The following is a direct extension of Corollary 2.2.

Corollary 5.1. *Suppose Assumption C holds. Then almost surely, the set of arcs $\{x^\tau(\cdot)\}_{\tau \geq 0}$ is relatively compact in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$. Moreover, almost surely, for any sequence $\tau_n \to \infty$ as $n \to \infty$, each limit point $z(\cdot)$ of the arcs $\{x^{\tau_n}(\cdot)\}$ in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ is absolutely continuous and satisfies the differential inclusion*

$$\dot{z}(t) \in H(z(t)) \qquad \text{for a.e. } t \geq 0.$$

Convergence guarantees of the proximal subgradient method will rely on the following assumption, akin to Assumption B.

Assumption D.

1. **(Weak Sard)** *The set of noncritical values of* (5.1) *is dense in* $\mathbb{R}$.
2. **(Descent)** *Whenever $z \colon \mathbb{R}_+ \to \mathbb{R}^d$ is an arc satisfying the differential inclusion*

$$\dot{z}(t) \in H(z(t)) \qquad \text{for a.e. } t \geq 0,$$

and $z(0)$ is not a critical point of (5.1)*, there exists a real $T > 0$ satisfying*

$$\varphi(z(T)) < \sup_{t \in [0,T]} \varphi(z(t)) \leq \varphi(z(0)).$$

Under the two assumptions, C and D, we obtain the following subsequential convergence guaranty.

Theorem 5.2. *Suppose that Assumptions C and D hold. Then almost surely, every limit point of $\{x_k\}_{k \geq 1}$ is critical for* (5.1) *and the function values $\{\varphi(x_k)\}_{k \geq 1}$ converge.*

Finally, let us observe that as in the unconstrained case, Assumption D is true as long as ∂f admits a chain rule.

Lemma 5.3. *Consider the optimization problem* (5.1) *and suppose that ∂f admits a chain rule. Let $z \colon \mathbb{R}_+ \to \mathbb{R}^d$ be any arc satisfying the differential inclusion*

$$\dot{z}(t) \in H(z(t)) \qquad \text{for a.e. } t \geq 0.$$

Then equality $\|\dot{z}(t)\| = \operatorname{dist}(0, H(z(t)))$ holds for a.e. $t \geq 0$, and therefore we have the estimate

$$\varphi(z(0)) - \varphi(z(t)) = \int_0^t \operatorname{dist}^2(0; H(z(\tau)))\, d\tau, \qquad \forall t \geq 0. \tag{5.4}$$

In particular, property 2 of Assumption D holds.

We now arrive at the main result of the section.

Corollary 5.4. *Suppose that Assumption C holds and that f, g, and $\mathcal{X}$ are definable in an o-minimal structure. Let $\{x_k\}_{k \geq 1}$ be the iterates produced by the proximal stochastic subgradient method* (5.3)*. Then almost surely, every limit point of the iterates $\{x_k\}_{k \geq 1}$ is critical for composite problem* (5.1) *and the function values $\{\varphi(x_k)\}_{k \geq 1}$ converge.*

5.1 Comments on boundedness

Thus far, all of our results have assumed that the subgradient iterates $\{x_k\}$ satisfy $\sup_{k\geq 1}\|x_k\| < \infty$ almost surely. One may enforce this assumption in several ways, most easily by assuming the constraint set $\mathcal{X}$ is bounded. Beyond boundedness of $\mathcal{X}$, proper choice of regularizer g may also ensure boundedness of $\{x_k\}$. Indeed, this observation was already made by Duchi-Ruan [15, Lemma 3.15]. Following their work, let us isolate the following assumption.

Assumption E (Regularizers that induce boundedness)**.**

1. *g is β-coercive, meaning* $\lim_{k\to\infty} g(x)/\|x\|^{\beta} = \infty$.
2. *There exists $\lambda \in (0,1]$ such that $g(x) \geq g(\lambda x)$ for x with sufficiently large norm.*

A natural regularizer satisfying this assumption is $\|x\|^{\beta+\epsilon}$ for any $\epsilon > 0$. The following theorem, whose proof is identical to that of [15, Lemma 3.15], shows that with Assumption E in place, the stochastic proximal subgradient methods produces bounded iterates.

Theorem 5.5 (Boundedness of iterates under coercivity)**.** *Suppose that Assumption E holds and that $\mathcal{X} = \mathbb{R}^d$. In addition, suppose there exists $L > 0$ and $\nu < \beta - 1$ such that $\|\zeta(x,\omega)\| \leq L(1 + \|x\|^{\nu})$ for all $x \in \mathbb{R}^d$ and $\omega \in \Omega$. Then* $\sup_{k\geq 1}\|x_k\| < \infty$.

We note that in the special (deterministic) case that $\zeta(x,\omega) \in \partial f(x)$ for all ω, the assumption on $\zeta(x,\omega)$ reduces to $\sup_{\zeta\in\partial f(x)}\|\zeta\| \leq L(1+\|x\|^{\nu})$, which stipulates that g grows more quickly than f.

A Proofs

Throughout the section, we suppose that Assumption C holds. To simplify notation, define the operator

$$T_\alpha(w) := \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ g(x) + \tfrac{1}{2\alpha}\|x - w\|^2 \right\},$$

and observe that each step of the stochastic subgradient methods reads:

$$x_{k+1} = T_{\alpha_k}(x_k - \alpha_k \zeta(x_k, \omega_k)).$$

Let $\mathcal{F}_k := \sigma(x_j, \omega_{j-1} : j \leq k)$ be the sigma algebra generated by the history of the algorithm.

A.1 Auxiliary lemma

In this subsection, we record auxiliary lemmas to be used in the sequel.

Lemma A.1. *For any $x, v \in \mathbb{R}^d$, and $\alpha > 0$, we have*

$$\alpha^{-1}\|x - T_\alpha(x - \alpha v)\| \leq 2\cdot \operatorname{dist}(0, \partial g(x)) + 2\cdot\|v\|.$$

Proof. Define $x_+ = T_\alpha(x - \alpha v)$. Then by definition, we

$$\tfrac{1}{2\alpha}\|x_+ - x\|^2 \leq g(x) - g(x_+) - \langle v, x_+ - x\rangle \leq \mathrm{dist}(0, \partial g(x))\|x_+ - x\| + \|v\| \cdot \|x_+ - x\|,$$

where the last inequality uses convexity of g and the Cauchy-Schwartz inequality. Dividing both sides by $\|x_+ - x\|$ yields the result. $\square$

Lemma A.2. *Suppose that a sequence $\{z_k\}_{k\geq 1}$ is bounded and $\{\beta_k\}$ is a nonnegative sequence satisfying $\sum_{k=1}^\infty \beta_k^2 < \infty$. Then $\beta_k \zeta(z_k, \cdot) \to 0$ almost surely.*

Proof. Notice that because $\{z_k\}_{k\geq 1}$ is bounded, it follows that $\{p(z_k)\}$ is bounded. Now consider the random variable $X_k = \beta_k^2\|\zeta(z_k, \cdot)\|^2$. Due to the estimate

$$\sum_{k=1}^{\infty} \mathbb{E}\left[X_k\right] \leq \sum_{k=1}^{\infty} \beta_k^2 p(x_k) < \infty,$$

standard results in measure theory (e.g., [31, Exercise 1.5.5]) imply that $X_k \to 0$ almost surely. $\square$

Lemma A.3. *Almost surely, we have $\alpha_k\|\zeta(x_k, \omega_k)\| \to 0$ as $k \to \infty$.*

Proof. From the variance bound ($\mathbb{E}\left[\|X - \mathbb{E}\left[X\right]\|^2\right] \leq \mathbb{E}\left[\|X\|^2\right]$) and Assumption C, we have

$$\mathbb{E}\left[\|\zeta(x_k, \omega_k) - \mathbb{E}\left[\zeta(x_k, \omega_k) \mid \mathcal{F}_k\right]\|^2 \mid \mathcal{F}_k\right] \leq \mathbb{E}\left[\|\zeta(x_k, \omega_k)\|^2 \mid \mathcal{F}_k\right] \leq p(x_k)$$

Therefore, the following infinite sum is finite:

$$\sum_{i=1}^{\infty} \alpha_i^2 \mathbb{E}\left[\|\zeta(x_i, \omega_i) - \mathbb{E}\left[\zeta(x_i, \omega_i) \mid \mathcal{F}_i\right]\|^2 \mid \mathcal{F}_i\right] \leq \sum_{i=1}^{\infty} \alpha_i^2 p(x_i) < \infty.$$

Define the L^2 martingale $X_k = \sum_{i=1}^k \alpha_i(\zeta(x_i, \omega_i) - \mathbb{E}\left[\zeta(x_i, \omega_i) \mid \mathcal{F}_i\right])$. Thus the limit $\langle X\rangle_\infty$ of the predictable compensator

$$\langle X\rangle_k := \sum_{i=1}^{k} \alpha_i^2 \mathbb{E}\left[\|\zeta(x_i, \omega_i) - \mathbb{E}\left[\zeta(x_i, \omega_i) \mid \mathcal{F}_i\right]\|^2 \mid \mathcal{F}_i\right],$$

exists. Applying [13, Theorem 5.3.33(a)], we deduce that almost surely X_k converges to a finite limit, which directly implies $\alpha_k\|\zeta(x_k, \omega_k) - \mathbb{E}\left[\zeta(x_k, \omega_k) \mid \mathcal{F}_k\right]\| \to 0$ almost surely as $k \to \infty$. Therefore, since $\alpha_k \mathbb{E}\left[\|\zeta(x_k, \omega_k)\| \mid \mathcal{F}_k\right] \leq \alpha_k\sqrt{p(x_k)} \to 0$ almost surely as $k \to 0$, it follows that $\alpha_k\|\zeta(x_k, \omega_k)\| \to 0$ almost surely as $k \to 0$. $\square$

Lemma A.4. *Almost surely, we have $\lim_{k\to\infty} \|x_{k+1} - x_k\| = 0$.*

Proof. Lemma A.1 implies that

$$\|x_{k+1} - x_k\| \leq 2\alpha_k \mathrm{dist}(0, \partial g(x_k)) + 2\alpha_k\|\zeta(x_k, \omega_k)\|.$$

Since g is a finite convex function, its subdifferential is bounded on bounded sets. Therefore, $\alpha_k \mathrm{dist}(0, \partial g(x_k)) \to 0$ almost surely. On the other hand, Lemma A.3 guarantees $\alpha_k\|\zeta(x_k, \omega_k)\| \to 0$ almost surely. The result follows. $\square$

The proof of the following Lemma is identical to the proof of Lemma 3.4.

Lemma A.5. *Almost surely, it holds:*

$$\liminf_{t\to\infty} \varphi(x(t)) = \liminf_{k\to\infty} \varphi(x_k) \qquad \text{and} \qquad \limsup_{t\to\infty} \varphi(x(t)) = \limsup_{k\to\infty} \varphi(x_k). \tag{A.1}$$

A.2 Proofs of Corollary 5.1 and Theorem 5.2

The main tool we use is the extension of Theorem 2.1 obtained by Duchi-Ruan in [15, Theorem 2].

Theorem A.6. *Consider a sequence of set-valued maps $G_k\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ and consider the following algorithm:*

$$x_{k+1} = x_k + \alpha_k \left[g_k + \xi_k\right], \qquad \text{where } g_k \in G_k(x_k) \qquad k \geq 0.$$

Suppose the following.

1. *The iterates are bounded, i.e.,* $\sup_{k\geq 1} \|x_k\| < \infty$ *and* $\sup_{k\geq 1} \|g_k\| < \infty$.
2. *The sequence $\{\alpha_k\}$ is nonnegative, square summable, but not summable:*

$$\alpha_k \geq 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty, \qquad \text{and} \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

3. *The weighted noise sequence is convergent: $\sum_{k=1}^{n} \alpha_k \xi_k \to v$ for some v as $k \to \infty$.*
4. *There exists a closed-valued mapping $H : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ such that for all sequences $\{w_k\} \subseteq \mathbb{R}^d$ such that $w_k \to w$ as $k \to \infty$, for any increasing sequence $\{n_k\} \subseteq \mathbb{N}$, and for any sequence $\{v_k\}$ satisfying $v_k \in G_{n_k}(w_k)$, the following holds:*

$$\lim_{n\to 0} \operatorname{dist}\left(\frac{1}{n}\sum_{k=1}^{n} v_k, H(w)\right) = 0.$$

Then for any sequence $\{\tau_k\}_{k=1}^{\infty} \subseteq \mathbb{R}_+$, the sequence of functions $\{x^{\tau_k}(\cdot)\}$ is relatively compact in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$. If in addition $\tau_k \to \infty$ as $k \to \infty$, all limit points $z(\cdot)$ of $\{x^{\tau_k}(\cdot)\}$ in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ are absolutely continuous and satisfy

$$\dot{z}(t) \in H(z(t)) \qquad \text{for a.e. } t \geq 0.$$

We are now ready to prove Corollary 5.1.

Proof of Corollary 5.1. Define the sequence of set-valued mappings

$$G_k(x) := -\partial f(x) - \alpha_k^{-1} \cdot \mathbb{E}_\omega \left[x - \alpha_k \zeta(x,\omega) - T_{\alpha_k}(x - \alpha_k \zeta(x,\omega))\right].$$

In addition, set $H = -\partial f - \partial g - N_{\mathcal{X}}$. We aim to apply Theorem A.6 to these two mappings. To that end, we first show that the sequence $\{x_k\}$ satisfies the desired recurrence involving G_k.

Claim 3. For all $k \geq 0$, we have that

$$x_{k+1} = x_k + \alpha_k \left[g_k + \xi_k\right] \qquad \text{for some } g_k \in G_k(x_k),$$

for some sequence $\{\xi_k\}$ such that the limit $\lim_{n\to\infty} \sum_{i=1}^{n} \alpha_i \xi_i$ exists almost surely.

Proof of Claim 3. Notice that for every k, we have

$$\begin{aligned}\tfrac{1}{\alpha_k}(x_k - x_{k+1}) &= \tfrac{1}{\alpha_k}\left[x_k - T_{\alpha_k}(x_k - \alpha_k\zeta(x_k,\omega_k))\right] \\ &= \mathbb{E}_\omega\left[\zeta(x_k,\omega)\right] + \tfrac{1}{\alpha_k}\mathbb{E}_\omega\left[x_k - \alpha_k\zeta(x_k,\omega) - T_{\alpha_k}(x_k - \alpha_k\zeta(x_k,\omega))\right] \\ &\quad + \tfrac{1}{\alpha_k}\left[\mathbb{E}_\omega\left[T_{\alpha_k}(x_k - \alpha_k\zeta(x_k,\omega))\right] - T_{\alpha_k}(x_k - \alpha_k\zeta(x_k,\omega_k))\right] \\ &\in -G_k(x_k) - \xi_k,\end{aligned}$$

where we define the noise sequence to be

$$\begin{aligned}\xi_k &:= \tfrac{1}{\alpha_k}\left[T_{\alpha_k}(x_k - \alpha_k\zeta(x_k,\omega_k)) - \mathbb{E}_\omega\left[T_{\alpha_k}(x_k - \alpha_k\zeta(x_k,\omega))\right]\right] \\ &= \tfrac{1}{\alpha_k}\left[T_{\alpha_k}(x_k - \alpha_k\zeta(x_k,\omega_k)) - x_k\right] - \tfrac{1}{\alpha_k}\left[\mathbb{E}_\omega\left[T_{\alpha_k}(x_k - \alpha_k\zeta(x_k,\omega)) - x_k\right]\right].\end{aligned}$$

Now we show that the limit of the weighted noise sums exist almost surely.

To that end, we first prove that $\{\alpha_k\xi_k\}$ is an L_2 martingale difference sequence, meaning that for all k, we have

$$\mathbb{E}\left[\alpha_k\xi_k \mid \mathcal{F}_k\right] = 0 \qquad \text{and} \qquad \sum_{k=1}^{\infty}\alpha_k^2\mathbb{E}\left[\|\xi_k\|^2 \mid \mathcal{F}_k\right] < \infty.$$

Clearly, ξ_k is zero mean, so we need only focus on the second property. By the variance bound ($\mathbb{E}\left[\|X - \mathbb{E}\left[X\right]\|^2\right] \leq \mathbb{E}\left[\|X\|^2\right]$) and Claim A.1, we have

$$\begin{aligned}\mathbb{E}\left[\|\xi_k\|^2 \mid \mathcal{F}_k\right] &\leq \frac{1}{\alpha_k^2}\mathbb{E}\left[\|T_{\alpha_k}(x_k - \alpha_k\zeta(x_k,\omega_k)) - x_k\|^2 \mid \mathcal{F}_k\right] \\ &\leq 4\cdot \operatorname{dist}(0,\partial g(x_k))^2 + 4\cdot\mathbb{E}\left[\|\zeta(x_k,\omega_k)\|^2 \mid \mathcal{F}_k\right].\end{aligned}$$

Notice that because $\{x_k\}$ is bounded a.s., it follows that $\operatorname{dist}(0,\partial g(x_k))$ and $\{p(x_k)\}$ are bounded a.s. Therefore, because

$$\sum_{k=1}^{\infty}\alpha_k^2\mathbb{E}\left[\|\zeta(x_k,\omega_k)\|^2 \mid \mathcal{F}_k\right] \leq \sum_{k=1}^{\infty}\alpha_k^2 p(x_k) < \infty,$$

it follows that $\sum_{k=1}^{\infty}\alpha_k^2\mathbb{E}\left[\|\xi_k\|^2 \mid \mathcal{F}_k\right] < \infty$, almost surely, as desired.

Now, define the L^2 martingale $X_k = \sum_{i=1}^{k}\alpha_i\xi_i$. Thus, the limit $\langle X\rangle_\infty$ of the predictable compensator

$$\langle X\rangle_k := \sum_{i=1}^{k}\alpha_i^2\mathbb{E}\left[\|\xi_i\|^2 \mid \mathcal{F}_i\right],$$

exists. Applying [13, Theorem 5.3.33(a)], we deduce that almost surely X_k converges to a finite limit, which completes the proof of the claim. □

Now we turn our attention to Item 1 of Theorem A.6.

Claim 4. The sequence $\{g_k\}$ is almost surely bounded.

Proof. Because the sequence $\{x_k\}$ is almost surely bounded, it follows that the

$$\sup\left\{\|v\| : v \in \bigcup_{k\geq 1} \partial f(x_k)\right\} < \infty,$$

almost surely. Thus, we need only show that

$$\sup_{k\geq 1}\left\{\left\|\frac{1}{\alpha_k}\mathbb{E}_\omega\left[x_k - \alpha_k\zeta(x_k,\omega) - T_{\alpha_k}(x_k - \alpha_k\zeta(x_k,\omega))\right]\right\|\right\} < \infty,$$

almost surely. Indeed, by the triangle inequality and Claim A.1, we have for any fixed $\omega \in \Omega$ the bound

$$\left\|\tfrac{1}{\alpha_k}\left[x_k - T_{\alpha_k}(x_k - \alpha_k\zeta(x_k,\omega))\right]\right\| \leq 2\cdot \operatorname{dist}(0, \partial g(x_k)) + 2\cdot\|\zeta(x_k,\omega)\|$$

Therefore, by Jensen's inequality, we have that

$$\begin{aligned}
&\left\|\tfrac{1}{\alpha_k}\mathbb{E}_\omega\left[x_k - \alpha_k\zeta(x_k,\omega) - T_{\alpha_k}(x_k - \alpha_k\zeta(x_k,\omega))\right]\right\| \\
&\leq 2\cdot\operatorname{dist}(0,\partial g(x_k)) + 3\cdot\mathbb{E}_\omega\left[\|\zeta(x_k,\omega)\|\right] \\
&\leq 2\cdot\operatorname{dist}(0,\partial g(x_k)) + 3\cdot\sqrt{p(x_k)},
\end{aligned}$$

which is almost surely bounded for all k. Taking the supremum yields the result. □

The proof will be complete once we verify Item 4 from Theorem A.6.

Claim 5. Define

$$H(x) = -\partial f(x) - \partial g(x) - N_{\mathcal{X}}(x).$$

Then Item 4 is true for the maps G_k.

Proof. Consider a sequence $\{z_k\}$ and a sequence $w_k^f \in \partial f(z_k)$. Let $\{n_k\}$ be an unbounded increasing sequence of indices. Observe that by Jensen's inequality, we have

$$\begin{aligned}
&\operatorname{dist}\left(\frac{1}{n}\sum_{k=1}^{n}\left(-w_k^f - \frac{1}{\alpha_{n_k}}\mathbb{E}_\omega\left[z_k - \alpha_{n_k}\zeta(z_k,\omega) - T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega))\right]\right), H(z)\right) \\
&\leq \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_\omega\left[\operatorname{dist}\left(-w_k^f - \frac{1}{\alpha_{n_k}}\left[z_k - \alpha_{n_k}\zeta(z_k,\omega) - T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega))\right], H(z)\right)\right].
\end{aligned}$$

Now suppose that $z_k \to z$. Fix an $\omega \in \Omega$. Our goal is to apply the dominated convergence theorem to each term in the above finite sum to conclude that each term converges to zero. Thus, we may assume without loss of generality that $\{z_k\}$ is completely contained within the ϵ ball around z over which Assumption C Item 6 guarantees the bound

$$\mathbb{E}\left[\sup_{x\in\mathcal{B}_\epsilon(z)}\|\zeta(x,\omega)\|\right] < \infty.$$

To that end, we must show two properties: for every fixed ω, each term in the sum tends to zero, and each term is bounded by an integrable function. We now prove both properties.

Subclaim 1. *We have that*

$$\operatorname{dist}\left(-w_k^f - \frac{1}{\alpha_{n_k}}\left[z_k - \alpha_{n_k}\zeta(z_k,\omega) - T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega))\right], H(z)\right) \to 0 \quad \text{as } k \to \infty.$$

Proof of Subclaim 1. Optimality conditions of the proximal subproblem imply

$$\tfrac{1}{\alpha_{n_k}}\left[z_k - \alpha_{n_k}\zeta(z_k,\omega) - T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega))\right] = w_k^g(\omega) + w_k^{\mathcal{X}}(\omega),$$

where $w_k^g(\omega) \in \partial g(T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega)))$ and $w_k^{\mathcal{X}}(\omega) \in N_{\mathcal{X}}(T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega)))$. Observe that by continuity and the fact that $\sum_{k=1}^{\infty} \alpha_{n_k}^2 < \infty$ and $\alpha_{n_k}\zeta(z_k,\omega) \to 0$ as $k \to \infty$ a.e. (see Lemma A.2), it follows that

$$T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega)) \to z.$$

Therefore, by outer semicontinuity of $\partial f, \partial g$, and $N_{\mathcal{X}}$, it follows that

$$\operatorname{dist}(w_k^f, \partial f(z)) \to 0; \qquad \operatorname{dist}(w_k^g(\omega), \partial g(z)) \to 0; \qquad \operatorname{dist}(w_k^{\mathcal{X}}(\omega), N_{\mathcal{X}}(z)) \to 0,$$

as $k \to \infty$. Consequently, almost surely we have that

$$\begin{aligned}
&\operatorname{dist}\left(-w_k^f - \tfrac{1}{\alpha_{n_k}}\left[z_k - \alpha_{n_k}\zeta(z_k,\omega) - T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega))\right], H(z)\right) \\
&\leq \operatorname{dist}(w_k^f, \partial f(z)) + \operatorname{dist}(w_k^g(\omega), \partial g(z)) + \operatorname{dist}(w_k^{\mathcal{X}}(\omega), N_{\mathcal{X}}(z)) \to 0,
\end{aligned}$$

as desired. □

Subclaim 2. *Let* $L_f := \sup_{k\geq 1} \operatorname{dist}(0, \partial f(z_k))$ *and* $L_g := \sup_{k\geq 1} \operatorname{dist}(0, \partial g(z_k))$. *Then*

$$\operatorname{dist}\left(-w_k^f - \tfrac{1}{\alpha_{n_k}}\left[z_k - \alpha_{n_k}\zeta(z_k,\omega) - T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega))\right], H(z)\right)$$

is dominated by an integrable function.

Proof of Subclaim 2. For each k, Lemma A.1 implies that

$$\left\|\tfrac{1}{\alpha_{n_k}}\left[z_k - \alpha_{n_k}\zeta(z_k,\omega) - T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega))\right]\right\| \leq 2L_g + 3\cdot\|\zeta(z_k,\omega)\|.$$

Consequently, we have

$$\begin{aligned}
&\operatorname{dist}\left(-w_k^f - \tfrac{1}{\alpha_{n_k}}\left[z_k - \alpha_{n_k}\zeta(z_k,\omega) - T_{\alpha_{n_k}r}(z_k - \alpha_{n_k}\zeta(z_k;\omega))\right], H(z)\right) \\
&\leq L_f + 2L_g + 3\cdot\|\zeta(z_k,\omega)\| + \operatorname{dist}(0, H(z)) \\
&\leq L_f + 2L_g + 3\cdot \sup_{x\in\mathcal{B}_\epsilon(z)} \|\zeta(x,\omega)\| + \operatorname{dist}(0, \partial f(z) + \partial g(z)),
\end{aligned}$$

which is integrable by Item 6 of Assumption C. □

Therefore, by the dominated convergence theorem, it follows that

$$\mathbb{E}_\omega\left[\operatorname{dist}\left(-w_k^f - \tfrac{1}{\alpha_{n_k}}\left[z_k - \alpha_{n_k}\zeta(z_k,\omega) - T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega))\right], H(z)\right)\right] \to 0$$

as $k \to \infty$. Thus, by the simple fact that $b_k \to 0$ as $k \to \infty$ implies $\frac{1}{n}\sum_{k=1}^n b_k \to 0$ as $n \to \infty$, we see that

$$\begin{aligned}
&\operatorname{dist}\left(\frac{1}{n}\sum_{k=1}^n \left(-w_k^f - \tfrac{1}{\alpha_{n_k}}\mathbb{E}_\omega\left[z_k - \alpha_{n_k}\zeta(z_k,\omega) - T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega))\right]\right), H(z)\right)\\
&\leq \frac{1}{n}\sum_{k=1}^n \mathbb{E}_\omega\left[\operatorname{dist}\left(-w_k^f - \tfrac{1}{\alpha_{n_k}}\left[z_k - \alpha_{n_k}\zeta(z_k,\omega) - T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega))\right], H(z)\right)\right] \to 0
\end{aligned}$$

as $n \to \infty$. □

We have now verified all parts of Theorem A.6. Therefore, the proof is complete. □

Proof of Theorem 5.2. The proof is identical to that of Theorem 3.1, with Lemmas A.3, A.4, A.5 playing the role of Lemmas 3.2, 3.3, 3.4. □

A.3 Verifying Assumption D for composite problems

Proof of Lemma 5.3. The argument is nearly identical to that of Lemma 4.2. Let $z\colon \mathbb{R}^d \to \mathcal{X}$ be an arc. Since ∂f admits a chain rule and so does ∂g (see Lemma 4.4), we deduce

$$(f \circ z)'(t) = \langle \partial f(z(t)), \dot z(t)\rangle \quad \text{and} \quad (g \circ z)'(t) = \langle \partial g(z(t)), \dot z(t)\rangle \qquad \text{for a.e. } t \geq 0. \tag{A.2}$$

Since $\mathcal{X}$ is convex, the same argument as in Lemma 4.4 shows

$$0 = \langle N_{\mathcal{X}}(z(t)), \dot z(t)\rangle \qquad \text{for a.e. } t \geq 0. \tag{A.3}$$

Adding equations (A.2) and (A.3) guarantees

$$(\varphi \circ z)'(t) = -\langle H(z(t)), \dot z(t)\rangle \qquad \text{for a.e. } t \geq 0.$$

The rest of the argument proceeds as in Lemma 4.2, with $H(z(t))$ playing the role of $-\partial f(z(t))$. □

Proof of Corollary 5.4. The result follows immediately from Lemma 5.2, once we show that Assumption D holds. Since f is definable in an o-minimal structure, Theorem 4.7 implies that ∂f admits the chain rule, and therefore Lemma 5.3 guarantees that the descent property of Assumption D holds. Thus we must only argue the weak Sard property of Assumption D. To this end, since f, g, and $\mathcal{X}$ are definable in an o-minimial structure, there exist Whitney C^d-stratifications $\mathcal{A}_f$, $\mathcal{A}_g$, and $\mathcal{A}_{\mathcal{X}}$ of $\operatorname{gph} f$, $\operatorname{gph} g$, and $\mathcal{X}$, respectively. Let $\Pi\mathcal{A}_f$ and $\Pi\mathcal{A}_g$ be the Whitney stratifications of $\mathbb{R}^d$ obtained by applying the coordinate projection $(x,r) \mapsto x$ to each stratum in $\mathcal{A}_f$ and $\mathcal{A}_g$. Appealing to [33, Theorem 4.8], we obtain a Whitney C^d-stratification $\mathcal{A}$ of $\mathbb{R}^d$ that is compatible with $(\Pi\mathcal{A}_f, \Pi\mathcal{A}_g, \mathcal{A}_{\mathcal{X}})$. That is, for every strata $M \in \mathcal{A}$ and $L \in \Pi\mathcal{A}_f \cup \Pi\mathcal{A}_g \cup \mathcal{A}_{\mathcal{X}}$, either $M \cap L = \emptyset$ or $M \subseteq L$.

Consider an arbitrary stratum $M \in \mathcal{A}$ intersecting $\mathcal{X}$ (and therefore contained in $\mathcal{X}$) and a point $x \in M$. Consider now the (unique) strata $M_f \in \Pi\mathcal{A}_f$, $M_g \in \Pi\mathcal{A}_g$, and $M_{\mathcal{X}} \in \mathcal{A}_{\mathcal{X}}$ containing x. Let $\widehat{f}$ and $\widehat{g}$ be C^d-smooth functions agreeing with f and g on a neighborhood of x in M_f and M_g, respectively. Appealing to (4.4), we conclude

$$\partial f(x) \subset \nabla \widehat{f}(x) + N_{M_f}(x) \qquad \text{and} \qquad \partial g(x) \subset \nabla \widehat{g}(x) + N_{M_g}(x).$$

The Whitney condition in turn directly implies $N_{\mathcal{X}}(x) \subset N_{M_{\mathcal{X}}}(x)$. Hence summing yields

$$\begin{aligned}\partial f(x) + \partial g(x) + N_{\mathcal{X}}(x) &\subset \nabla(\widehat{f} + \widehat{g})(x) + N_{M_f}(x) + N_{M_g}(x) + N_{M_{\mathcal{X}}}(x)\\ &\subset \nabla(\widehat{f} + \widehat{g})(x) + N_M(x),\end{aligned}$$

where the last inclusion follows from the compatibility $M \subset M_f$ and $M \subset M_g$. Notice that $\widehat{f} + \widehat{g}$ agrees with $f + g$ on a neighborhood of x in M. Hence if the inclusion, $0 \in \partial f(x) + \partial g(x) + N_{\mathcal{X}}(x)$, it must be that x is a critical point of the C^d-smooth function $f + g$ restricted to M, in the classical sense. Applying the standard Sard's theorem to each manifold M, the result follows. □

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

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

[3] V.S. Borkar. *Stochastic approximation*. Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi, 2008. A dynamical systems viewpoint.

[4] J.M. Borwein and X. Wang. Lipschitz func tions with maximal Clarke subdifferentials are generic. *Proc. Amer. Math. Soc.*, 128(11):3221–3229, 2000.

[5] H. Brézis. *Opérateurs maximaux monotones et semi-groupes de contraction dans des espaces de Hilbert*. North-Holland Math. Stud. 5, North-Holland, Amsterdam, 1973.

[6] R.E. Bruck, Jr. Asymptotic convergence of nonlinear contraction semigroups in Hilbert space. *J. Funct. Anal.*, 18:15–26, 1975.

[7] F.H. Clarke. Generalized gradients and applications. *Trans. Amer. Math. Soc.*, 205:247–262, Apr. 1975.

[8] F.H. Clarke, Y.S. Ledyaev, R.J. Stern, and P.R. Wolenski. *Nonsmooth analysis and control theory*, volume 178. Springer Science & Business Media, 2008.

[9] M. Coste. *An introduction to o-minimal geometry*. RAAG Notes, 81 pages, Institut de Recherche Mathématiques de Rennes, November 1999.

[10] M. Coste. *An Introduction to Semialgebraic Geometry*. RAAG Notes, 78 pages, Institut de Recherche Mathématiques de Rennes, October 2002.

[11] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *arXiv:1803.06523*, 2018.

[12] D. Davis and D. Drusvyatskiy. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. *arXiv:1802.02988*, 2018.

[13] A. Dembo. Probability theory: Stat310/math230 september 3, 2016. 2016. Available at `http://statweb.stanford.edu/~adembo/stat-310b/lnotes.pdf`.

[14] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. Curves of descent. *SIAM J. Control Optim.*, 53(1):114–138, 2015.

[15] J.C. Duchi and F. Ruan. Stochastic methods for composite optimization problems. *Preprint arXiv:1703.08570*, 2017.

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

[17] A.D. Ioffe. Critical values of set-valued maps with stratifiable graphs. Extensions of Sard and Smale-Sard theorems. *Proc. Amer. Math. Soc.*, 136(9):3111–3119, 2008.

[18] A.D. Ioffe. An invitation to tame optimization. *SIAM J. Optim.*, 19(4):1894–1917, 2008.

[19] A.D. Ioffe. *Variational analysis of regular mappings*. Springer Monographs in Mathematics. Springer, Cham, 2017. Theory and applications.

[20] H.J. Kushner and G.G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.

[21] S. Łojasiewicz. Ensemble semi-analytiques. *IHES Lecture Notes*, 1965.

[22] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2008.

[23] A.S. Nemirovsky and D.B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983.

[24] E. A. Nurminskii. Minimization of nondifferentiable functions in the presence of noise. *Cybernetics*, 10(4):619–621, Jul 1974.

[25] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[26] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.

[27] R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[28] R.T. Rockafellar. *The theory of subgradients and its applications to problems of optimization*, volume 1 of *R & E*. Heldermann Verlag, Berlin, 1981.

[29] R.T. Rockafellar and R.J-B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften, Vol 317, Springer, Berlin, 1998.

[30] G.V. Smirnov. *Introduction to the theory of differential inclusions*, volume 41 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002.

[31] Terence Tao. *An introduction to measure theory*, volume 126. American Mathematical Soc., 2011.

[32] L. van den Dries. *Tame topology and o-minimal structures*, volume 248 of *LMS Lecture Note Series*. Cambridge University Press, Cambridge, 1998.

[33] L. van den Dries and C. Miller. Geometric categories and o-minimal structures. *Duke Math. J.*, 84:497–540, 1996.

[34] H. Whitney. A function not constant on a connected set of critical points. *Duke Math. J.*, 1(4):514–517, 12 1935.

[35] A.J. Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function. *J. Amer. Math. Soc.*, 9(4):1051–1094, 1996.