Foundations of Computational Mathematics (2020) 20:119–154
https://doi.org/10.1007/s10208-018-09409-5

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Stochastic Subgradient Method Converges on Tame Functions

Damek Davis[1] · Dmitriy Drusvyatskiy[2] · Sham Kakade[3] · Jason D. Lee[4]

Received: 26 May 2018 / Revised: 15 October 2018 / Accepted: 8 November 2018 /
Published online: 7 January 2019

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.

Keywords Subgradient · Proximal · Stochastic subgradient method · Differential inclusion · Lyapunov function · Semialgebraic · Tame

Mathematics Subject Classification 65K05 · 65K10 · 34A60 · 90C15

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

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

Communicated by Michael Overton.

Research of Dmitriy Drusvyatskiy was supported by the AFOSR YIP Award FA9550-15-1-0237 and by the NSF DMS 1651851 and CCF 1740551 Awards. Sham Kakade acknowledges funding from the Washington Research Foundation Fund for Innovation in Data-Intensive Discovery and the NSF CCF 1740551 Award. Jason D. Lee acknowledges funding from the ARO MURI Award W911NF-11-1-0303.

Extended author information available on the last page of the article

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

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

Here $\partial f(x)$ denotes the Clarke subdifferential [10]. 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 $\{\xi_k\}_{k\geq 1}$. We make the standard assumption that conditioned on the past, each random variable ξ_k has mean zero and its second moment grows at a controlled rate.[1]

Though variants of the stochastic subgradient method (1.1) date back to Robbins–Monro's pioneering 1951 work [34], their convergence behavior is still largely not understood in nonsmooth and nonconvex settings. In particular, 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 open-source PyTorch [33] 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 [32], meaning $0 \in \partial f(x^*)$. Moreover, rates of convergence in terms of natural optimality/stationarity measures are available. In summary, the rates are $\mathbb{E}\,[f(x_k) - \inf f] = O(k^{-1/2})$, $\mathbb{E}\,[\|\nabla f(x_k)\|] = O(k^{-1/4})$, and $\mathbb{E}\left[\|\nabla f_{1/2\rho}(x_k)\|\right] = O(k^{-1/4})$, for functions that are convex [31], smooth [20], and ρ-weakly convex [15,16], respectively. In particular, the convergence guarantee above for ρ-weakly convex functions appeared only recently in [15,16], 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 nonpathological 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

[1] The zero mean assumption on ξ_k is not for free when f is given in expectation form $f(x) = \mathbb{E}\,[f(x, \omega)]$ and we choose $y_k + \xi_k \in \partial f(x_k, \omega_k)$ with $\omega_k \sim P$. It is true under certain circumstances [11, Theorem 2.7.2], [9], but verifying its validity in general remains an open and difficult question. In deterministic settings, a principled automatic differentiation approach for computing Clarke subgradients has been proposed in [24–26].

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 5.11 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, a standard Lyapunov-type argument shows that every limit point of the stochastic subgradient method is critical for f, almost surely. Techniques of this type can be found for example in the monograph of Kushner–Yin [27, Theorem 5.2.1] and the landmark papers of Benaïm et al. [2,3]. Here, we provide a self-contained treatment, which facilitates direct extensions to "proximal" variants of the stochastic subgradient method.[2] In particular, our analysis follows closely the recent work of Duchi–Ruan [19, Section 3.4.1] on convex composite minimization.

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 [7,8]. Indeed, if f is convex and $x\colon [0,\infty) \to \mathbb{R}$ is any absolutely continuous curve, then the "chain rule" holds:

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

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

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

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 [12, Section 2.4] or [36, 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 univariate 1-Lipschitz function whose Clarke subdifferential is the unit interval at every point [6,35].

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,

[2] Concurrent to this work, the independent preprint [29] also provides convergence guarantees for the stochastic projected subgradient method, under the assumption that the objective function is "subdifferentially regular" and the constraint set is convex. Subdifferential regularity rules out functions with downward kinks and cusps, such as deep networks with the Relu$(\cdot)$ activation functions. Besides subsuming the subdifferentially regular case, the results of the current paper apply to the broad class of Whitney stratifiable functions, which includes all popular deep network architectures.

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 [4]. An important subclass of Whitney stratifiable functions consists of semialgebraic functions [28]—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 semialgebraic functions are usually easy to recognize. More generally still, "semianalytic" functions [28] and those that are "definable in an o-minimal structure" are Whitney stratifiable [39]. The latter function class, in particular, shares all the robustness and analytic properties of semialgebraic functions, while encompassing many more examples. Case in point, Wilkie [41] famously showed that there is an o-minimal structure that contains both the exponential $x \mapsto e^x$ and all semialgebraic functions.[3]

The key observation for us, which originates in [18, 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 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 [18, Section 5.1] for semialgebraic functions, though identical arguments hold more broadly for Whitney stratifiable functions. These results, however, are somewhat hidden in the paper [18], which is possibly why they have thus far been underutilized. In this manuscript, we provide a self-contained review of the material from [18, Section 5.1], highlighting only the most essential ingredients and streamlining some of the arguments.

Though the discussion above is for unconstrained problems, the techniques we develop apply much more broadly to constrained problems of the form

$$\min_{x \in \mathcal{X}} \; f(x) + g(x).$$

Here f and g are locally Lipschitz continuous functions and $\mathcal{X}$ is an arbitrary closed set. The popular *proximal stochastic subgradient method* simply iterates the steps

$$\left\{ \begin{array}{l} \text{Sample an estimator } \zeta_k \text{ of } \partial f(x_k) \\ \text{Select } x_{k+1} \in \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle \zeta_k, x \rangle + g(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\} \end{array} \right\}. \tag{1.4}$$

[3] The term "tame" used in the title has a technical meaning. Tame sets are those whose intersection with any ball is definable in some o-minimal structure. The manuscript [22] provides a nice exposition on the role of tame sets and functions in optimization.

Combining our techniques with those in [19] quickly yields subsequential convergence guarantees for this algorithm. Note that we impose no convexity assumptions on f, g, or $\mathcal{X}$.

The outline of this paper is as follows. In Sect. 2, we fix the notation for the rest of the manuscript. Section 3 provides a self-contained treatment of asymptotic consistency for discrete approximations of differential inclusions. In Sect. 4, we specialize the results of the previous section to the stochastic subgradient method. Finally, in Sect. 5, we verify the sufficient conditions for subsequential convergence for a broad class of locally Lipschitz functions, including those that are subdifferentially regular and Whitney stratifiable. In particular, we specialize our results to deep learning settings in Corollary 5.11. In the final Sect. 6, we extend the results of the previous sections to the proximal setting.

2 Preliminaries

Throughout, we will mostly use standard notation on differential inclusions, as set out for example in the monographs of Borkar [5], Clarke et al. [12], and Smirnov [37]. 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 indicator function of Q, denoted δ_Q, is defined to be zero on Q and $+\infty$ off it. 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. We will use $\mathbb{R}_+$ to denote the set of nonnegative real numbers. We denote the graph of a function f by $\operatorname{gph} f = \{(x, f(x)) \colon x \in \mathbb{R}, f(x) < \infty\}$.

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 converge 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

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

Moreover, if this is the case, 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 continuous and x is an arc, then the composition $f \circ x$ is absolutely continuous.

2.2 Set-Valued Maps and the Clarke Subdifferential

A set-valued map $G\colon \mathcal{X} \rightrightarrows \mathbb{R}^m$ is a mapping from a set $\mathcal{X} \subseteq \mathbb{R}^d$ to the powerset of $\mathbb{R}^m$. Thus $G(x)$ is a subset of $\mathbb{R}^m$, for each $x \in \mathcal{X}$. We will use the notation

$$G^{-1}(v) := \{x \in \mathcal{X} : v \in G(x)\}$$

for the preimage of a vector $v \in \mathbb{R}^m$. The map G is *outer-semicontinuous* at a point $x \in \mathcal{X}$ if for any sequences $x_i \xrightarrow{\mathcal{X}} x$ and $v_i \in G(x_i)$ converging to some vector $v \in \mathbb{R}^m$, the inclusion $v \in G(x)$ holds.

The most important set-valued map for our work will be the generalized derivative in the sense of Clarke [10]–a notion we now review. Consider a locally Lipschitz continuous function $f\colon \mathbb{R}^d \to \mathbb{R}$. The well-known Rademacher's theorem guarantees that f is differentiable almost everywhere. Taking this into account, the *Clarke subdifferential* of f at any point x is the set [12, 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. It is standard that the map $x \mapsto \partial f(x)$ is outer-semicontinuous and its images $\partial f(x)$ are nonempty, compact, convex sets for each $x \in \mathbb{R}^d$; see for example [12, Proposition 1.5 (a,e)].

Analogously to the smooth setting, a point $x \in \mathbb{R}^d$ is called *(Clarke) critical* for f whenever the inclusion $0 \in \partial f(x)$ holds. Equivalently, these are the points at which the Clarke directional derivative is nonnegative in every direction [12, 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 $r = f(x)$.

3 Differential Inclusions and Discrete Approximations

In this section, we discuss the asymptotic behavior of discrete approximations of differential inclusions. All the elements of the analysis we present, in varying generality, can be found in the works of Benaïm et al. [2,3], Borkar [5], and Duchi–Ruan [19]. Out of these, we most closely follow the work of Duchi–Ruan [19].

3.1 Functional Convergence of Discrete Approximations

Let $\mathcal{X}$ be a closed set and let $G\colon \mathcal{X} \rightrightarrows \mathbb{R}^d$ be a set-valued map. Then an arc $z\colon \mathbb{R}_+ \to \mathbb{R}^d$ is called a *trajectory* of G if it satisfies the differential inclusion

$$\dot{z}(t) \in G(z(t)) \quad \text{for a.e. } t \geq 0. \tag{3.1}$$

Notice that the image of any arc z is automatically contained in $\mathcal{X}$, since arcs are continuous and $\mathcal{X}$ is closed. In this work, we will primarily focus on iterative algorithms

that aim to asymptotically track a trajectory of the differential inclusion (3.1) using a noisy discretization with vanishing step-sizes. Though our discussion allows for an arbitrary set-valued map G, the reader should keep in mind that the most important example for us will be $G = -\partial f$, where f is a locally Lipschitz function.

Throughout, we will consider the following iteration sequence:

$$x_{k+1} = x_k + \alpha_k(y_k + \xi_k). \tag{3.2}$$

Here $\alpha_k > 0$ is a sequence of step-sizes, y_k should be thought of as an approximate evaluation of G at some point near x_k, and ξ_k is a sequence of "errors".

Our immediate goal is to isolate reasonable conditions, under which the sequence $\{x_k\}$ asymptotically tracks a trajectory of the differential inclusion (3.1). Following the work of Duchi–Ruan [19] on stochastic approximation, we stipulate the following assumptions.

Assumption A (*Standing assumptions*)

1. All limit points of $\{x_k\}$ lie in $\mathcal{X}$.
2. The iterates are bounded, i.e., $\sup_{k\geq 1} \|x_k\| < \infty$ and $\sup_{k\geq 1} \|y_k\| < \infty$.
3. The sequence $\{\alpha_k\}$ is nonnegative, square summable, but not summable:

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

4. The weighted noise sequence is convergent: $\sum_{k=1}^{n} \alpha_k \xi_k \to v$ for some v as $n \to \infty$.
5. For any unbounded increasing sequence $\{k_j\} \subset \mathbb{N}$ such that x_{k_j} converges to some point $\bar{x}$, it holds:

$$\lim_{n\to\infty} \operatorname{dist}\left(\frac{1}{n}\sum_{j=1}^{n} y_{k_j}, G(\bar{x})\right) = 0.$$

Some comments are in order. Conditions 1, 2, and 3 are in some sense minimal, though the boundedness condition must be checked for each particular algorithm. Condition 4 guarantees that the noise sequence ξ_k does not grow too quickly relative to the rate at which α_k decrease. The key Condition 5 summarizes the way in which the values y_k are approximate evaluations of G, up to convexification.

To formalize the idea of asymptotic approximation, let us define the time points $t_0 = 0$ and $t_m = \sum_{k=1}^{m-1} \alpha_k$, for $m \geq 1$. Let $x(\cdot)$ now be the linear interpolation of the discrete path:

$$x(t) := x_k + \frac{t - t_k}{t_{k+1} - t_k}(x_{k+1} - x_k) \quad \text{for } t \in [t_k, t_{k+1}). \tag{3.3}$$

For each $\tau \geq 0$, define the time-shifted curve $x^\tau(\cdot) = x(\tau + \cdot)$. As $x(\cdot)$ and $x^\tau(\cdot)$ linearly interpolate subsequences of $\{x_k\}$, they are bounded whenever $\{x_k\}$ is bounded.

The following result of Duchi–Ruan [19, Theorem 2] shows that under the above conditions, for any sequence $\tau_k \to \infty$, the shifted curves $\{x^{\tau_k}\}$ subsequentially converge in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ to a trajectory of (3.1). Results of this type under more stringent assumptions, and with similar arguments, have previously appeared for example in Benaïm et al. [2,3] and Borkar [5].

Theorem 3.1 (Functional approximation) *Suppose that Assumption A holds. Then for any sequence $\{\tau_k\}_{k=1}^{\infty} \subseteq \mathbb{R}_+$, the set 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 trajectories of the differential inclusion* (3.1).

3.2 Subsequential Convergence to Equilibrium Points

A primary application of the discrete process (3.2) is to solve the inclusion

$$0 \in G(z). \tag{3.4}$$

Indeed, one can consider the points satisfying (3.4) as equilibrium (constant) trajectories of the differential inclusion (3.1). Ideally, one would like to find conditions guaranteeing that every limit point $\bar{x}$ of the sequence $\{x_k\}$, produced by the recursion (3.2), satisfies the desired inclusion (3.4). Making such a leap rigorous typically relies on combining the asymptotic convergence guarantee of Theorem 3.1 with existence of a Lyapunov-like function $\varphi(\cdot)$ for the continuous dynamics; see, e.g., [2,3]. Let us therefore introduce the following assumption.

Assumption B *(Lyapunov condition)* There exists a continuous function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$, which is bounded from below, and such that the following two properties hold.

1. (Weak Sard) For a dense set of values $r \in \mathbb{R}$, the intersection $\varphi^{-1}(r) \cap G^{-1}(0)$ is empty.
2. (Descent) Whenever $z\colon \mathbb{R}_+ \to \mathbb{R}^d$ is a trajectory of the differential inclusion (3.1) and $0 \notin G(z(0))$, there exists a real $T > 0$ satisfying

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

The weak Sard property is reminiscent of the celebrated Sard's theorem in real analysis. Indeed, consider the classical setting $G = -\nabla f$ for a smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$. Then the weak Sard property stipulates that the set of noncritical values of f is dense in $\mathbb{R}$. By Sard's theorem, this is indeed the case, as long as f is C^d smooth. Indeed, Sard's theorem guarantees the much stronger property that the set of noncritical values has full measure. We will comment more on the weak Sard property in Sect. 4, once we shift focus to optimization problems. The descent property, says that φ eventually strictly decreases along the trajectories of the differential inclusion $\dot{z} \in G(z)$ emanating from any nonequilibrium point. This Lyapunov-type condition is standard in the literature and we will verify that it holds for a large class of optimization problems in Sect. 5.

As we have alluded to above, the following theorem shows that under Assumptions A and B, every limit point $\bar{x}$ of $\{x_k\}$ indeed satisfies the inclusion $0 \in G(\bar{x})$. We were unable to find this result stated and proved in this generality. Therefore, we record a complete proof in Sect. 3.3. The idea of the proof is of course not new, and can already be seen for example in [2,19,27]. Upon first reading, the reader can safely skip to Sect. 4.

Theorem 3.2 *Suppose that Assumptions A and B hold. Then every limit point of* $\{x_k\}_{k\geq 1}$ *lies in* $G^{-1}(0)$ *and the function values* $\{\varphi(x_k)\}_{k\geq 1}$ *converge.*

3.3 Proof of Theorem 3.2

In this section, we will prove Theorem 3.2. The argument we present is rooted in the "nonescape argument" for ODEs, using φ as a Lyapunov function for the continuous dynamics. In particular, the proof we present is in the same spirit as that in [27, Theorem 5.2.1] and [19, Section 3.4.1].

Henceforth, we will suppose that Assumptions A and B hold, and let $x(t)$ be defined as in (3.3). We first collect two elementary lemmas.

Lemma 3.3 *The equality* $\lim_{k\to\infty} \|x_{k+1} - x_k\| = 0$ *holds.*

Proof From the recurrence (3.2), we have $\|x_{k+1} - x_k\| \leq \alpha_k \|y_k\| + \alpha_k \|\xi_k\|$. Assumption A guarantees $\alpha_k \to 0$ and $\{y_k\}$ are bounded, and therefore $\alpha_k \|y_k\| \to 0$. Moreover, since the sequence $\sum_{k=1}^{n} \alpha_k \xi_k$ is convergent, we deduce $\alpha_k \|\xi_k\| \to 0$. The result follows. □

Lemma 3.4 *Equalities hold:*

$$\liminf_{t\to\infty} \varphi(x(t)) = \liminf_{k\to\infty} \varphi(x_k) \quad \textit{and} \quad \limsup_{t\to\infty} \varphi(x(t)) = \limsup_{k\to\infty} \varphi(x_k). \tag{3.5}$$

Proof Clearly, the inequalities $\leq$ and $\geq$ hold in (3.5), respectively, as the curve $x(t)$ interpolates the sequence $\{x_k\}$. 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

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

where the second inequality follows because $x(\tau_i)$ is on the line segment between x_{k_i} and x_{k_i+1}; see Eq. (3.3). Lemma 3.3 implies that the right-hand side tends to zero, and hence $x_{k_i} \to x^*$. Continuity of φ then directly yields the guarantee $\varphi(x_{k_i}) \to \varphi(x^*)$.

In particular, we may take $\tau_i \to \infty$ to be a sequence realizing $\liminf_{t\to\infty} \varphi(x(t))$. 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

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

thereby establishing the first equality in (3.5). The second equality follows analogously. □

The proof of Theorem 3.3 will follow quickly from the following proposition.

Proposition 3.5 *The values $\varphi(x(t))$ have a limit as $t \to \infty$.*

Proof Without loss of generality, suppose $0 = \liminf_{t\to\infty} \varphi(x(t))$. For each $r \in \mathbb{R}$, define the sublevel set

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

Choose any $\epsilon > 0$ satisfying $\epsilon \notin \varphi(G^{-1}(0))$. Note that by Assumption B, we can let $\epsilon > 0$ be as small as we wish. By the first equality in (3.5), there are infinitely many indices k such that $\varphi(x_k) < \epsilon$. 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:

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

Proof of Claim 1 Since the sequence $\{x_k\}_{k\geq 1}$ is bounded by Assumption A, it is contained in some compact set $C \subset \mathbb{R}^d$. From continuity, we have

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

It follows that the two closed 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:

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

In particular $\mathrm{dist}(x_k; \mathbb{R}^d \setminus \mathcal{L}_{2\epsilon}) \geq \alpha > 0$, 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. Define $i_1 \in \mathbb{N}$ to 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. the iterate x_{e_1} lies in $\mathbb{R}^d \setminus \mathcal{L}_{2\epsilon}$, where e_1 is the exit time $e_1 := \min\{e \geq i_1 : x_e \notin \mathcal{L}_{2\epsilon} \setminus \mathcal{L}_\epsilon\}$.

Then let $i_2 > i_1$ be the next smallest index satisfying the same property, and so on. See Fig. 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.

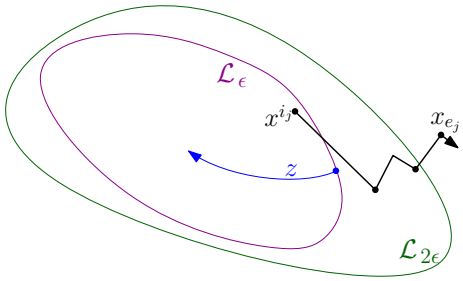

Fig. 1 Illustration of the nonescape argument

Before proving the claim, let us see how it immediately yields the validity of the proposition. 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 $\lim_{k\to\infty} \varphi(x_k) = 0$. Equation (3.5) then directly implies $\lim_{t\to\infty} \varphi(x(t)) = 0$, 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, Theorem 3.1 shows that the curves $x^{\tau_j}(\cdot)$ converge in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ to some arc $z(\cdot)$ satisfying

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

By construction, we have $\varphi(x_{i_j}) \leq \epsilon$ and $\varphi(x_{i_j+1}) > \epsilon$. We therefore deduce

$$\begin{aligned} \epsilon &\geq \varphi(x_{i_j}) \geq \varphi(x_{i_j+1}) + (\varphi(x_{i_j}) - \varphi(x_{i_j+1})) \\ &\geq \epsilon + [\varphi(x_{i_j}) - \varphi(z(0))] - [\varphi(x_{i_j+1}) - \varphi(z(0))]. \end{aligned} \tag{3.6}$$

Recall $x_{i_j} \to z(0)$ as $j \to \infty$. Lemma 3.3 in turn implies $\|x_{i_j} - x_{i_j+1}\| \to 0$ and therefore $x_{i_j+1} \to z(0)$ as well. Continuity of φ then guarantees that the right-hand side of (3.6) tends to ϵ, and hence $\varphi(z(0)) = \lim_{j\to\infty} \varphi(x_{i_j}) = \epsilon$. In particular, $z(0)$ is not an equilibrium point of G because we have assumed that $\epsilon \notin \varphi(G^{-1}(0))$. Hence, Assumption B yields a real $T > 0$ such that

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

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

Appealing to uniform convergence of x^{τ_j} to z on $[0, T]$, we conclude

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

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

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

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

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

We will show that the bound $\varphi(z(T)) \leq \epsilon - 2\delta$ 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

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

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

$$\|x_{\ell_j} - x^{\tau_j}(T)\| = \|x^{\tau_j}(t_{\ell_j} - \tau_j) - x^{\tau_j}(T)\| \leq \|x_{\ell_j} - x_{\ell_j+1}\| \to 0,$$

where the last inequality follows because $x^{\tau_j}(t_{\ell_j} - \tau_j)$ and $x^{\tau_j}(T)$ lie on the line segment between x_{ℓ_j} and x_{ℓ_j+1}; see Eq. (3.3). Hence $x_{\ell_j} \to z(T)$ as $j \to \infty$. Continuity of φ then guarantees $\lim_{j\to\infty} \varphi(x_{\ell_j}) = \varphi(z(T))$. Consequently, the inequality $\varphi(x_{\ell_j}) < \epsilon - \delta$ holds for all large j, which is the desired contradiction. □

The proof of the proposition is now complete. □

We can now prove the main convergence theorem.

Proof of Theorem 3.2 Let x^* be a limit point of $\{x_k\}$ and suppose for the sake of contradiction that $0 \notin G(x^*)$. 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 Theorem 3.1. Assumption B guarantees that there exists a real $T > 0$ satisfying

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

On the other hand, we successively deduce

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

where the last two equalities follow from Proposition 3.5 and continuity of φ. We have thus arrived at a contradiction, and the theorem is proved. □

4 Subgradient Dynamical System

Assumptions A and B, taken together, provide a powerful framework for proving subsequential convergence of algorithms to a zero of the set-valued map G. Note that the two assumptions are qualitatively different. Assumption A is a property of both the algorithm (3.2) and the map G, while Assumption B is a property of G alone.

For the rest of our discussion, we apply the differential inclusion approach outlined above to optimization problems. Setting the notation, consider the optimization task

$$\min_{x\in\mathbb{R}^d} \; f(x), \tag{4.1}$$

where $f:\mathbb{R}^d \to \mathbb{R}$ is a locally Lipschitz continuous function. Seeking to apply the techniques of Sect. 3, we simply set $G = -\partial f$ in the notation therein. Thus we will be interested in algorithms that, under reasonable conditions, track solutions of the differential inclusion

$$\dot{z}(t) \in -\partial f(z(t)) \quad \text{for a.e. } t \geq 0, \tag{4.2}$$

and subsequentially converge to critical points of f. Discrete processes of the type (3.2) for the optimization problem (4.1) are often called stochastic approximation algorithms. Here, we study two such prototypical methods: the stochastic subgradient method in this section and the stochastic proximal subgradient in Sect. 6. Each fits under the umbrella of Assumption A.

Setting the stage, the *stochastic subgradient method* simply iterates the steps:

$$x_{k+1} = x_k - \alpha_k(y_k + \xi_k) \quad \text{with} \quad y_k \in \partial f(x_k), \tag{4.3}$$

where $\{\alpha_k\}_{k\geq 1}$ is a step-size sequence and $\{\xi_k\}_{k\geq 1}$ is now a sequence of random variables (the "noise") on some probability space. Let us now isolate the following standard assumptions (e.g., [5,27]) for the method and see how they immediately imply Assumption A.

Assumption C *(Standing assumptions for the stochastic subgradient method)*

1. The sequence $\{\alpha_k\}$ is nonnegative, square summable, but not summable:

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

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

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

That is, there exists a function $p:\mathbb{R}^d \to [0,\infty)$, which is bounded on bounded sets, so that almost surely, for all $k \in \mathbb{N}$, we have

$$\mathbb{E}[\xi_k|\mathcal{F}_k] = 0 \quad \text{and} \quad \mathbb{E}[\|\xi_k\|^2|\mathcal{F}_k] \leq p(x_k).$$

The following is true.

Lemma 4.1 *Assumption C guarantees that almost surely Assumption A holds.*

Proof Suppose Assumption C holds. Clearly, A.1 and A.3 hold vacuously, while A.2 almost surely follows immediately from C.2 and local Lipschitz continuity of f. Assumption A.5 follows quickly from the fact the ∂f outer-semicontinuous and compact-convex valued; we leave the details to the reader. Thus we must only verify A.4, which follows quickly from standard martingale arguments. Indeed, notice from Assumption C, we have

$$\mathbb{E}\left[\xi_k \mid \mathcal{F}_k\right] = 0 \quad \forall k \qquad \text{and} \qquad \sum_{i=0}^{\infty} \alpha_i^2 \mathbb{E}\left[\|\xi_i\|^2 \mid \mathcal{F}_i\right] \leq \sum_{i=0}^{\infty} \alpha_i^2 p(x_i) < \infty.$$

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 [17, Theorem 5.3.33(a)], we deduce that almost surely X_k converges to a finite limit. □

Thus applying Theorem 3.1, we deduce that under Assumption C, almost surely, the stochastic subgradient path tracks a trajectory of the differential inclusion (4.2). As we saw in Sect. 3, proving subsequential convergence to critical points requires existence of a Lyapunov-type function φ for the continuous dynamics. Henceforth, let us assume that the Lyapunov function φ is f itself. Section 5 is devoted entirely to justifying this assumption for two broad classes of functions that are virtually exhaustive in data scientific contexts.

Assumption D *(Lyapunov condition in unconstrained minimization)*

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 the classical 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 D 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 Whitney

[40]. Property 2 of Assumption D is more meaningful. It essentially asserts that f must locally strictly decrease along any subgradient trajectory emanating from a noncritical point.

Thus applying Theorem 3.2, we have arrived at the following guarantee for the stochastic subgradient method.

Theorem 4.2 *Suppose that Assumptions C and D hold. Then almost surely, every limit point of stochastic subgradient iterates $\{x_k\}_{k\geq 1}$ is critical for f and the function values $\{f(x_k)\}_{k\geq 1}$ converge.*

5 Verifying the Descent Condition

In light of Theorems 3.2 and 4.2, it is important to isolate a class of functions that automatically satisfy Assumption D.2. In this section, we do exactly that, focusing on two problem classes: (1) subdifferentially regular functions and (2) those functions whose graphs are Whitney stratifiable. We will see that the latter problem class also satisfies D.1.

The material in this section is not new. In particular, the results of this section have appeared in [18, Section 5.1]. These results, however, are somewhat hidden in the paper [18] and are difficult to parse. Moreover, at the time of writing [18, 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 D will become clear shortly.

Definition 5.1 *(Chain rule)* Consider a locally Lipschitz function f on $\mathbb{R}^d$. We will say that f *admits a chain rule* if for any arc $z\colon \mathbb{R}_+ \to \mathbb{R}^d$, equality

$$(f\circ z)'(t) = \langle \partial f(z(t)), \dot{z}(t)\rangle \quad \text{holds for a.e. } t\geq 0, \tag{5.1}$$

In long form, Eq. (5.1) means that for almost every $t \geq 0$ and all subgradients $v \in \partial f(z(t))$, the identity $\langle v, \dot{z}(t)\rangle = (f\circ z)'(t)$ holds. The importance of the chain rule becomes immediately clear with the following lemma.

Lemma 5.2 *Consider a locally Lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}$ that admits a chain rule. Let $z\colon \mathbb{R}_+ \to \mathbb{R}^d$ be any arc satisfying the differential inclusion*

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

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

$$f(z(0)) - f(z(t)) = \int_0^t \operatorname{dist}^2(0; \partial f(z(\tau)))\, d\tau, \quad \forall t\geq 0. \tag{5.2}$$

In particular, property 2 of Assumption D 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

$$0 = \langle \partial f(z(t)) - \partial f(z(t)), \dot{z}(t) \rangle. \tag{5.3}$$

To simplify the notation, set $S := \partial f(z(t))$, $W := \operatorname{span}(S - S)$, and $y := -\dot{z}(t)$. Note that W is the parallel subspace to $\partial f(z(t))$, while $y + W$ is the affine hull of $\partial f(z(t))$, since we have $y \in \partial f(z(t))$. Appealing to (5.3), 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

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

Since f is locally Lipschitz, the composition $f \circ z$ is absolutely continuous. Hence integrating over the interval $[0, t]$ yields (5.2).

Suppose now that the point $z(0)$ is noncritical. Then by outer semicontinuity of ∂f, there exists $T > 0$ such that $z(\tau)$ is noncritical for any $\tau \in [0, T]$. It follows immediately that the value $\int_0^t \operatorname{dist}^2(0; \partial f(z(\tau)))\, d\tau$ is strictly increasing in $t \in [0, T]$, and therefore by (5.2) that $f \circ z$ is strictly decreasing. Hence item 2 of Assumption D holds, as claimed. □

Thus property 2 of Assumption D 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.

5.1 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 close connection with convex functions; we refer the reader to the monograph [36] 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. We now present the formal definition.

Definition 5.3 *(Subdifferential regularity)* A locally Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ 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:

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

The following lemma shows that any locally Lipschitz function that is subdifferentially regular indeed admits a chain rule.

Lemma 5.4 (Chain rule under subdifferential regularity) *Any locally Lipschitz function that is subdifferentially regular admits a chain rule and therefore item 2 of Assumption D holds.*

Proof Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz and subdifferentially regular function. Consider an arc $x : \mathbb{R}_+ \to \mathbb{R}^d$. 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 D holds by Lemma 5.2. □

Thus we have arrived at the following corollary. For ease of reference, we state subsequential convergence guarantees both for the general process (3.2) and for the specific stochastic subgradient method (4.3).

Corollary 5.5 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz function that is subdifferentially regular and such that its set of noncritical values is dense in $\mathbb{R}$.*

- *(Stochastic approximation) Consider the iterates $\{x_k\}_{k\geq 1}$ produced by (3.2) and suppose that Assumption A holds with $G = -\partial f$. Then every limit point of the iterates $\{x_k\}_{k\geq 1}$ is critical for f and the function values $\{f(x_k)\}_{k\geq 1}$ converge.*
- *(Stochastic subgradient method) Consider the iterates $\{x_k\}_{k\geq 1}$ produced by the stochastic subgradient method (4.3) and suppose that Assumption C holds. Then almost surely, every limit point of the iterates $\{x_k\}_{k\geq 1}$ is critical for f and the function values $\{f(x_k)\}_{k\geq 1}$ converge.*

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.

5.2 Stratifiable Functions

As we saw in the previous section, subdifferential regularity is a local property that implies the desired item 2 of Assumption D. 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 $F : U \to \mathbb{R}^{d-r}$ 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 5.6 *(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
$$L \cap \text{cl}\, M \neq \emptyset \implies L \subset \text{cl}\, M \quad \text{holds.}$$
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 $f : \mathbb{R}^d \to \mathbb{R}$ 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 $\text{cl}\, M$. In particular, the frontier condition endows the strata with a partial order $L \preceq M \Leftrightarrow L \subset \text{cl}\, M$. 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 [4]. Consider a Whitney C^p-stratification $\{M_i\}$ of the graph of a locally Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$. 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 [4, Proposition 4]. For any stratum $\mathcal{M}$ and any point $x \in \mathcal{M}$, we have

$$(v, -1) \in N_M(x, f(x)) \quad \text{for all} \quad v \in \partial f(x), \tag{5.4}$$

and

$$\partial f(x) \subset \nabla g(x) + N_{\mathcal{M}}(x), \tag{5.5}$$

where $g : \mathbb{R}^d \to \mathbb{R}$ is any C^1-smooth function that coincides with f on some neighborhood $U \subseteq \mathcal{M}$ containing x.

The following theorem, which first appeared in [4, Corollary 5], shows that Whitney stratifiable functions automatically satisfy the weak Sard property of Assumption D.

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 [21] or the monograph [23, Section 8.4].

Lemma 5.7 (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 D 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 gph f containing $(x, f(x))$. Since x is critical for f, appealing to (5.4) yields the inclusion $(0, -1) \in N_{M_i}(x, f(x))$ and therefore the equality $\pi_i\left(T_{M_i}(x, f(x))\right) = \{0\} \subsetneq \mathbb{R}$. 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 any Whitney stratifiable function.

Theorem 5.8 *Any locally Lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}$ that is Whitney C^1-stratifiable admits a chain rule, and therefore item 2 of Assumption D 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 $x\colon \mathbb{R}^d \to \mathbb{R}$. Clearly, both x and $f \circ x$ are differentiable at a.e. $t \geq 0$. Moreover, we claim that for a.e. $t \geq 0$, the implication holds:

$$x(t) \in \mathcal{M}_i \implies \dot{x}(t) \in T_{\mathcal{M}_i}(x(t)). \tag{5.6}$$

To see this, fix a manifold $\mathcal{M}_i$ and let Ω_i be the set of all $t \geq 0$ such that $x(t) \in \mathcal{M}_i$, the derivative $\dot{x}(t)$ exists, and we have $\dot{x}(t) \notin T_{\mathcal{M}_i}(x(t))$. If we argue that Ω_i has zero measure, then so does the union $\cup_i \Omega_i$ and the claim is proved. Fix an arbitrary $t \in \Omega_i$. There exists a closed interval I around t such that x restricted to I intersects $\mathcal{M}_i$ only at $x(t)$, since otherwise we would deduce that $\dot{x}(t)$ lies in $T_{\mathcal{M}_i}(x(t))$ by definition of the tangent space. We may further shrink I such that its endpoints are rational. It follows that Ω_i may be covered by disjoint closed intervals with rational endpoints. Hence Ω_i is countable and therefore zero measure, as claimed.

Fix now a real $t > 0$ such that x and $f \circ x$ are differentiable at t and the implication (5.6) holds. 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

$$(f \circ x)'(t) = \lim_{r \to 0} \frac{f(x(t + r)) - f(x(t))}{r}$$

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

Notice $\frac{|f(x(t+r))-f(\gamma(r))|}{r} \le L \left\| \frac{x(t+r)-x(t)+\gamma(0)-\gamma(r)}{r} \right\| \to L\|\dot{x}(t) - \dot{\gamma}(0)\| = 0$ 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 (5.5). □

Putting together Theorems 3.2, 4.2, 5.8 and Lemma 5.7, we arrive at the main result of our paper. Again for ease of reference, we state subsequential convergence guarantees both for the general process (3.2) and for the specific stochastic subgradient method (4.3).

Corollary 5.9 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz function that is C^d-stratifiable.*

- **(Stochastic approximation)** *Consider the iterates $\{x_k\}_{k\ge 1}$ produced by (3.2) and suppose that Assumption A holds with $G = -\partial f$. Then every limit point of the iterates $\{x_k\}_{k\ge 1}$ is critical for f and the function values $\{f(x_k)\}_{k\ge 1}$ converge.*
- **(Stochastic subgradient method)** *Consider the iterates $\{x_k\}_{k\ge 1}$ produced by the stochastic subgradient method (4.3) and suppose that Assumption C holds. Then almost surely, every limit point of the iterates $\{x_k\}_{k\ge 1}$ is critical for f and the function values $\{f(x_k)\}_{k\ge 1}$ 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.

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

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

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 [28] 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, semialgebraic 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 [14, 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 [13] and van den Dries–Miller [39].

Definition 5.10 *(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 $\pi : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ 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., [39]). Beyond semialgebraicity, Wilkie showed that there is an o-minimal structure that simultaneously contains both the graph of the exponential function $x \mapsto e^x$ and all semialgebraic sets [41].

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

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

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 C holds.[4] *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.*

6 Proximal Extensions

In this section, we extend most of the results in Sects. 4 and 5 on unconstrained problems to a "proximal" setting and comment on sufficient conditions to ensure boundedness of the iterates. The arguments follow quickly by combining the techniques developed by Duchi–Ruan [19] with those presented in Sect. 5. Consequently, all the proofs are in "Appendix A".

Setting the stage, consider the composite optimization problem

$$\min_{x\in\mathcal{X}} \varphi(x) := f(x) + g(x), \tag{6.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ are locally Lipschitz functions and $\mathcal{X}$ is a closed set. As is standard in the literature on proximal methods, we will say that $x \in \mathcal{X}$ is a *composite critical point* of the problem (6.1) if the inclusion holds:

$$0 \in \partial f(x) + \partial g(x) + N_{\mathcal{X}}(x). \tag{6.2}$$

Here, the symbol $N_{\mathcal{X}}(x)$ denotes the Clarke normal cone to the closed set $\mathcal{X} \subseteq \mathbb{R}^d$ at $x \in \mathcal{X}$. We refer the reader to Appendix A for a formal definition. We only note that when $\mathcal{X}$ is a closed convex set, $N_{\mathcal{X}}$ reduces to the normal cone in the sense of convex analysis, while for a C^1-smooth manifold $\mathcal{X}$, it reduces to the normal space in the sense of differential geometry. It follows from [36, Corollary 10.9] that local minimizers of (6.1) are necessarily composite critical. A real number $r \in \mathbb{R}$ is called a *composite critical value* if equality, $r = f(x) + g(x)$, holds for some composite critical point x.

Thus we will be interested in an extension of the stochastic subgradient method that tracks a trajectory of the differential inclusion

$$\dot{z}(t) \in -(\partial f + \partial g + N_{\mathcal{X}})(z(t)) \quad \text{for a.e. } t \geq 0, \tag{6.3}$$

and subsequentially converges to a composite critical point of (6.1). Seeking to apply the techniques of Sect. 3, we can simply set $G = -\partial f - \partial g - N_{\mathcal{X}}$ in the notation

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

therein. Note that G thus defined is not necessarily a subdifferential of a single function because equality in the subdifferential sum rule [36, Corollary 10.9] can fail when the summands are not subdifferentially regular.

We now aim to describe the proximal stochastic subgradient method for the problem (6.1). There are two ingredients we must introduce: a stochastic subgradient oracle for f and the proximity map of $g + \delta_{\mathcal{X}}$, where $\delta_{\mathcal{X}}$ is the indicator function of $\mathcal{X}$. We describe the two ingredients in turn.

Stochastic subgradient oracle Our model of the stochastic subgradient oracle follows that of the influential work [30]. 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 $\zeta : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$ satisfying:

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

Thus after sampling $\omega \sim P$, the vector $\zeta(x, \omega)$ can serve as a stochastic estimator for a true subgradient of f.

Proximal map Standard deterministic proximal splitting methods utilize the proximal map of $g + \delta_{\mathcal{X}}$, namely:

$$z \mapsto \operatorname*{argmin}_{x \in \mathcal{X}} \{g(x) + \tfrac{1}{2\alpha}\|x - z\|^2\}.$$

Since we do not impose convexity assumptions on g and $\mathcal{X}$, this map can be set-valued. Thus we must pass to a measurable selection. Indeed, supposing that g is bounded from below on $\mathcal{X}$, the result [36, Exercise 14.38] guarantees that there exists a measurable selection $T_{(\cdot)}(\cdot) : (0, \infty) \times \mathbb{R}^d \to \mathbb{R}^d$, such that

$$T_\alpha(z) \in \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ g(x) + \tfrac{1}{2\alpha}\|x - z\|^2 \right\} \quad \text{for all } \alpha > 0, z \in \mathbb{R}^d.$$

We can now formally state the algorithm. Given an iterate $x_k \in \mathcal{X}$, the *proximal stochastic subgradient method* performs the update

$$\left\{ \begin{array}{l} \text{Sample } \omega_k \sim P \\ x_{k+1} = T_{\alpha_k}(x_k - \alpha_k \zeta(x_k, \omega_k)). \end{array} \right\}. \tag{6.4}$$

Here $\{\alpha_k\}_{k \geq 1}$ is a positive control sequence. We will analyze the algorithm under the following two assumptions, akin to Assumptions C and D of Sect. 4. Henceforth, define the set-valued map $G : \mathcal{X} \rightrightarrows \mathbb{R}^d$ by

$$G(x) = -\partial f(x) - \partial g(x) - N_{\mathcal{X}}(x). \tag{6.5}$$

Assumption E *(Standing assumptions for the proximal stochastic subgradient method)*

1. $\mathcal{X}$ is closed, f and g are locally Lipschitz, and g is bounded from below on $\mathcal{X}$.
2. There exists a function $L\colon \mathbb{R}^d \to \mathbb{R}$, which is bounded on bounded sets, satisfying

$$L(x) \geq \sup_{z:g(z)\leq g(x)} \frac{g(x)-g(z)}{\|x-z\|}.$$

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

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

4. Almost surely, the iterates are bounded: $\sup_{k\geq 1} \|x_k\| < \infty$.
5. There exists a function $p\colon \mathbb{R}^d \to \mathbb{R}_+$, that is bounded on bounded sets, such that

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

6. For every convergent sequence $\{z_k\}_{k\geq 1}$, we have

$$\mathbb{E}_\omega\left[\sup_{k\geq 1} \|\zeta(z_k,\omega)\|\right] < \infty.$$

Assumption F *(Lyapunov condition in proximal minimization)*

1. **(Weak Sard)** The set of composite noncritical values of (6.1) is dense in $\mathbb{R}$.
2. **(Descent)** Whenever $z\colon \mathbb{R}_+ \to \mathcal{X}$ is an arc satisfying the differential inclusion

$$\dot{z}(t) \in -(\partial f + \partial g + N_{\mathcal{X}})(z(t)) \quad \text{for a.e. } t \geq 0,$$

and $z(0)$ is not a composite critical point of (6.1), there exists a real $T > 0$ satisfying

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

Let us make a few comments. Properties E.1, E.3, E.4, and E.5 are mild and completely expected in light of the results in the previous sections. Property E.6 is a technical condition ensuring that the expected maximal noise in the stochastic subgradient is bounded along any convergent sequence. Finally, Property E.2 is a mild technical condition on function g that we allow. In particular, it holds for any convex, globally Lipschitz, or coercive locally Lipschitz function. We record this observation in the following lemma.

Lemma 6.1 *Consider any function $g\colon \mathbb{R}^d \to \mathbb{R}_+$ that is either convex, globally Lipschitz, or locally Lipschitz and coercive. Then g satisfies property 2 in Assumption E.*

Proof Fix a point x and a point z satisfying $g(z) \leq g(x)$. Let us look at each case and upper bound the error $g(x) - g(z)$. Suppose first that g is convex. Then for the vector $v \in \partial g(x)$ of minimal norm, we have

$$g(x) - g(z) \leq \langle v, x - z \rangle \leq \|v\| \|x - z\| = L(x) \cdot \|x - z\|.$$

where $L(x) := \operatorname{dist}(0, \partial g(x))$. If f is globally Lipschitz, then clearly we have

$$g(x) - g(z) \leq L(x) \cdot \|x - z\|.$$

where $L(x)$ is identically equal to the global Lipschitz constant of g. Finally, in the third case, suppose that g is coercive and locally Lipschitz. We deduce

$$g(x) - g(z) \leq L(x) \cdot \|x - z\|,$$

where $L(x)$ is the Lipschitz modulus of g on the compact sublevel set $[g \leq g(x)]$. Since g is locally Lipschitz continuous, in all three cases, the function $L(\cdot)$ is bounded on bounded sets. □

Under the two Assumptions, E and F, we obtain the following subsequential convergence guarantee. The argument in appendix is an application of Theorem 3.2. To this end, we show that Assumption E implies Assumption A almost surely, while Assumption F is clearly equivalent to Assumption B.

Theorem 6.2 *Suppose that Assumptions E and F hold. Then almost surely, every limit point of the iterates $\{x_k\}_{k \geq 1}$ produced by the proximal stochastic subgradient method (6.4) is composite critical for (6.1) and the function values $\{\varphi(x_k)\}_{k \geq 1}$ converge.*

Finally, we must now understand problem classes that satisfy Assumption F. To this end, analogously to Definition 5.1, we say that $\mathcal{X}$ *admits a chain rule* if for any arc $z : \mathbb{R}_+ \to \mathcal{X}$, equality holds

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

Whenever $\mathcal{X}$ is Clarke regular or Whitney stratifiable, $\mathcal{X}$ automatically admits a chain rule. Indeed, the argument is identical to that of Lemma 5.4 and Theorem 5.8. As in the unconstrained case, Assumption F.2 is true as long as f, g, and $\mathcal{X}$ admit a chain rule.

Lemma 6.3 *Consider the optimization problem (6.1) and suppose that f, g, and $\mathcal{X}$ admit a chain rule. Let $z : \mathbb{R}_+ \to \mathcal{X}$ be any arc satisfying the differential inclusion*

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

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

$$\varphi(z(0)) - \varphi(z(t)) = \int_0^t \operatorname{dist}^2(0; G(z(\tau)))\, d\tau, \quad \forall t \geq 0. \tag{6.6}$$

In particular, property 2 of Assumption F holds.

We now arrive at the main result of the section.

Corollary 6.4 *Suppose that Assumption E 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 (6.4). Then almost surely, every limit point of the iterates $\{x_k\}_{k\geq 1}$ is composite critical for the problem (6.1) and the function values $\{\varphi(x_k)\}_{k\geq 1}$ converge.*

6.1 Comments on Boundedness

Thus far, all of our results have assumed that the subgradient iterates $\{x_k\}$ satisfy the inequality $\sup_{k\geq 1}\|x_k\| < \infty$ 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 [19, Lemma 3.15]. Following their work, let us isolate the following assumption.

Assumption G *(Regularizers that induce boundedness)*

1. g is convex and 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 [19, Lemma 3.15], shows that with Assumption G in place, the stochastic proximal subgradient methods produce bounded iterates.

Theorem 6.5 (Boundedness of iterates under coercivity) *Suppose that Assumption G 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 $\sup_{k\geq 1}\|x_k\| < \infty$ almost surely.*

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 for the Proximal Extension

In this section, we follow the notation of Sect. 6. Namely, we let $\zeta : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$ be the stochastic subgradient oracle and $T_{(\cdot)}(\cdot) : (0,\infty) \times \mathbb{R}^d \to \mathbb{R}^d$ the proximal selection. Throughout, we let x_k and ω_k be generated by the proximal stochastic subgradient method (6.4) and suppose that Assumption E holds. Let $\mathcal{F}_k := \sigma(x_j, \omega_{j-1} : j \leq k)$ be the sigma algebra generated by the history of the algorithm.

Let us now formally define the normal cone constructions of variational analysis. For any point $x \in \mathcal{X}$, the *proximal normal cone* to $\mathcal{X}$ at x is the set

$$N^P_{\mathcal{X}}(x) := \{\lambda v \in \mathbb{R}^d : x \in \mathrm{proj}_{\mathcal{X}}(x + v), \lambda \geq 0\},$$

where $\mathrm{proj}_{\mathcal{X}}(\cdot)$ denotes the nearest point map to $\mathcal{X}$. The *limiting normal cone* to $\mathcal{X}$ at x, denoted $N^L_{\mathcal{X}}(x)$, consists of all vector $v \in \mathbb{R}^d$ such that there exist sequences $x_i \in \mathcal{X}$ and $v_i \in N^P_{\mathcal{X}}(x_i)$ satisfying $(x_i, v_i) \to (x, v)$. The *Clarke normal cone* to $\mathcal{X}$ at x is then simply

$$N_{\mathcal{X}}(x) := \mathrm{cl}\,\mathrm{conv}\, N^L_{\mathcal{X}}(x).$$

A. 1 Auxiliary Lemmas

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

Lemma A.1 *There exists a function $L : \mathbb{R}^d \to \mathbb{R}_+$, which is bounded on bounded sets, such that for any $x, v \in \mathbb{R}^d$, and $\alpha > 0$, we have*

$$\alpha^{-1}\|x - x_+\| \le 2 \cdot L(x) + 2 \cdot \|v\|,$$

where we set $x_+ := T_\alpha(x - \alpha v)$.

Proof Let $L(\cdot)$ be the function from Property E.2. From the definition of the proximal map, we deduce

$$\tfrac{1}{2\alpha}\|x_+ - x\|^2 \le g(x) - g(x_+) - \langle v, x_+ - x\rangle \le L(x) \cdot \|x_+ - x\| + \|v\| \cdot \|x_+ - x\|,$$

where the second inequality follows by Property E.2 if $g(x_+) \le g(x)$; otherwise if $g(x_+) \ge g(x)$, then $g(x) - g(x_+) \le 0 \le L(x)\|x_+ - x\|$ trivially. Dividing both sides by $\frac{1}{2}\|x_+ - x\|$ yields the result. □

Lemma A.2 *Let $\{z_k\}_{k\ge1}$ be a bounded sequence in $\mathbb{R}^d$ and let $\{\beta_k\}_{k\ge1}$ be a nonnegative sequence satisfying $\sum_{k=1}^{\infty} \beta_k^2 < \infty$. Then almost surely over $\omega \sim P$, we have $\beta_k \zeta(z_k, \omega) \to 0$.*

Proof Notice that because $\{z_k\}_{k\ge1}$ 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

$$\sum_{k=1}^{\infty} \mathbb{E}[X_k] \le \sum_{k=1}^{\infty} \beta_k^2 p(z_k) < \infty,$$

standard results in measure theory (e.g., [38, Exercise 1.5.5]) imply that $X_k \to 0$ almost surely. □

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, $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \le \mathbb{E}\left[\|X\|^2\right]$, and Assumption E, we have

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

Therefore, the following infinite sum is a.s. 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 [17, 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 $\alpha_k \| \mathbb{E}\left[\zeta(x_k, \omega_k) \mid \mathcal{F}_k\right] \| \leq \alpha_k \mathbb{E}\left[\| \zeta(x_k, \omega_k) \| \mid \mathcal{F}_k \right] \leq \alpha_k \sqrt{p(x_k)} \to 0$ almost surely as $k \to \infty$, it follows that $\alpha_k \| \zeta(x_k, \omega_k) \| \to 0$ almost surely as $k \to \infty$. □

A.2 Proof Theorem 6.2

In addition to Assumption E, let us now suppose that Assumption F holds. Define the set-valued map $G \colon Q \rightrightarrows \mathbb{R}^d$ by $G = -\partial f - \partial g - N_{\mathcal{X}}$. We aim to apply Theorem 3.2, which would immediately imply the validity of Theorem 6.2. To this end, notice that Assumption F is exactly Assumption B for our map G. Thus we must only verify that Assumption A holds almost surely. Note that Properties A.1 and A.3 hold vacuously. Thus, we must only show that A.2, A.4, and A.5 hold. The argument we present is essentially the same as in [19, Section 3.2.2] .

For each index k, define the set-valued map

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

Note that G_k is a deterministic map, with k only signifying the dependence on the deterministic sequence α_k. Define now the noise sequence

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

Let us now write the proximal stochastic subgradient method in the form (3.2).

Lemma A.4 (Recursion relation) *For all $k \geq 0$, we have*

$$x_{k+1} = x_k + \alpha_k \left[y_k + \xi_k \right] \quad \textit{for some } y_k \in G_k(x_k).$$

Proof Notice that for every index $k \geq 0$, we have

$$\tfrac{1}{\alpha_k}(x_k - x_{k+1}) = \tfrac{1}{\alpha_k} \left[x_k - T_{\alpha_k}(x_k - \alpha_k \zeta(x_k, \omega_k)) \right]$$

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

as desired. □

The following lemma shows that A.4 holds almost surely.

Lemma A.5 (Weighted noise sequence) *The limit* $\lim_{n\to\infty}\sum_{i=1}^{n}\alpha_i\xi_i$ *exists almost surely.*

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

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

Clearly, ξ_k has zero mean conditioned on the past, and so we need only focus on the second property. By the variance bound, $\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \le \mathbb{E}\left[\|X\|^2\right]$, and Lemma A.1, we have

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

Notice that because $\{x_k\}$ is bounded a.s., it follows that $\{L(x_k)\}$ and $\{p(x_k)\}$ are bounded a.s. Therefore, because

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

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 [17, 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 A.2.

Lemma A.6 *Almost surely, the sequence* $\{y_k\}$ *is bounded.*

Proof Because the sequence $\{x_k\}$ is almost surely bounded and f is locally Lipschitz, clearly we have

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

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. To this end, by the triangle inequality and Lemma A.1, we have for any fixed $\omega \in \Omega$ the bound

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

Therefore, by Jensen's inequality, we have that

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

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

As the last step, we verify Item A.5.

Lemma A.7 *Item 5 of Assumption A is true.*

Proof Assumption A.5 requires us to bound all subsequential averages of the direction vectors $\{y_k\}$. Here it is more convenient to prove a more general statement, namely, that for any sequence of points z_k converging to z and $y_{n_k} \in G_{n_k}(z_k)$, we have $\operatorname{dist}\left(\frac{1}{n}\sum_{k=1}^n y_{n_k}, G(z)\right) \to 0$. Assumption A.5 is then an immediate consequence.

Thus, consider any sequence $\{z_k\} \subseteq \mathcal{X}$ converging to a point $z \in \mathcal{X}$ and an arbitrary sequence $w_k^f \in \partial f(z_k)$. Let $\{n_k\}$ be an unbounded increasing sequence of indices. Observe that since $G(z)$ is convex and using Jensen's inequality, we have

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

Our goal is to prove that the right-hand side tends to zero almost surely, which directly implies validity of A.5

Our immediate goal is to apply the dominated convergence theorem to each term in the above finite sum to conclude that each term converges to zero. To that end, we must show two properties: for every fixed ω, each term in the sum tends to zero, and that each term is bounded by an integrable function. We now prove both properties.

Claim 3 Almost surely in $\omega \sim P$, we have that

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

Proof of Subclaim 3 Optimality conditions [36, Exercise 10.10] of the proximal subproblem imply

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

for some $w_k^g(\omega) \in \partial g(T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega)))$ and $w_k^{\mathcal{X}}(\omega) \in N_{\mathcal{X}}^L(T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega)))$, and where $N_{\mathcal{X}}^L$ denotes the limiting normal cone. 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

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

Indeed, setting $z_k^+ = T_{\alpha_{n_k}}(z_k - \alpha_{n_k}\zeta(z_k,\omega))$, we have that by Lemma A.1,

$$\|z_k - z_k^+\| \le 2\alpha_{n_k}L(z_k) + 2\alpha_{n_k}\|\zeta(z_k,\omega)\| \to 0 \quad \text{as } k \to \infty,$$

which implies that $\lim_{k\to\infty} z_k^+ = \lim_{k\to\infty} z_k = z$.

We furthermore deduce that $w_k^{\mathcal{X}}(\omega)$ and $w_k^g(\omega)$ are bounded almost surely. Indeed, $w_k^g(\omega)$ is bounded since g is locally Lipschitz and z_k^+ are bounded. Moreover, Lemma A.1 implies

$$\|w_k^g(\omega) + w_k^{\mathcal{X}}(\omega)\| = \left\|\tfrac{1}{\alpha_{n_k}}\left[z_k - \alpha_{n_k}\zeta(z_k,\omega) - z_k^+\right]\right\| \le 2\cdot L(z_k) + 3\cdot\sup_{k\ge 1}\|\zeta(z_k,\omega)\|.$$

Observe that the right-hand side is a.s. bounded by item 6 of Assumption E. Thus, since $w_k^g(\omega) + w_k^{\mathcal{X}}(\omega)$ and $w_k^g(\omega)$ are a.s. bounded, it follows that $w_k^{\mathcal{X}}(\omega)$ must also be a.s. bounded, as desired.

Appealing to outer semicontinuity of ∂f, ∂g, and $N_{\mathcal{X}}^L$ (e.g., [36, Proposition 6.6]), the inclusion $N_{\mathcal{X}}^L \subset N_{\mathcal{X}}$, and the boundedness of $\{w_k^f\}$, $\{w_k^g(\omega)\}$, and $\{w_k^{\mathcal{X}}(\omega)\}$, it follows that

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

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

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

as desired. □

Claim 4 Let $L_f := \sup_{k\ge 1}\operatorname{dist}(0, \partial f(z_k))$ and $L_g := \sup_{k\ge 1} L(z_k)$. Then for all $k \ge 0$, the functions

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

are uniformly dominated by an integrable function in ω.

Proof of Subclaim 4 For each k, Lemma A.1 implies the bound

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

Consequently, we have

$$\begin{aligned}&\operatorname{dist}\left(-w_k^f - \tfrac{1}{\alpha_{n_k}}\left[z_k - \alpha_{n_k}\zeta(z_k,\omega) - T_{\alpha_{n_k}r}(z_k - \alpha_{n_k}\zeta(z_k,\omega))\right], G(z)\right)\\ &\le L_f + 2L_g + 3\cdot\|\zeta(z_k,\omega)\| + \operatorname{dist}(0, G(z))\\ &\le L_f + 2L_g + 3\cdot\sup_{k\ge 1}\|\zeta(z_k,\omega)\| + \operatorname{dist}(0, \partial f(z) + \partial g(z)),\end{aligned}$$

which is integrable by Item 6 of Assumption E. □

Applying the dominated convergence theorem, it follows that

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

as $k \to \infty$. Notice the simple fact that for any real sequence $b_k \to 0$, it must be that $\frac{1}{n}\sum_{k=1}^n b_k \to 0$ as $n \to \infty$. Consequently

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

as $n \to \infty$. This completes the proof. □

We have now verified all parts of Theorem 3.1. Therefore, the proof is complete.

A.3 Verifying Assumption F for Composite Problems

Proof of Lemma 6.3 The argument is nearly identical to that of Lemma 5.2, with one additional subtlety that G is not necessarily outer-semicontinuous. Let $z\colon \mathbb{R}^d \to \mathcal{X}$ be an arc. Since f, g, and $\mathcal{X}$ admit a chain rule, we deduce

$$(f\circ z)'(t) = \langle \partial f(z(t)), \dot{z}(t)\rangle \quad (g\circ z)'(t) = \langle \partial g(z(t)), \dot{z}(t)\rangle, \quad \text{and} \quad 0 = \langle N_{\mathcal{X}}(z(t)), \dot{z}(t)\rangle,$$

for a.e. $t \geq 0$. Adding the three equations yields

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

Suppose now that $z(\cdot)$ satisfies $\dot{z}(t) \in -G(z(t))$ for a.e. $t \geq 0$. Then the same linear algebraic argument as in Lemma 5.2 yields the equality $\|\dot{z}(t)\| = \operatorname{dist}(0; G(z(t)))$ for a.e. $t \geq 0$ and consequently the Eq. (6.6).

To complete the proof, we must only show that property 2 of Assumption F holds. To this end, suppose that $z(0)$ is not composite critical and let $T > 0$ be arbitrary. Appealing to (6.6), clearly $\sup_{t\in[0,T]} \varphi(z(t)) \leq \varphi(z(0))$. Thus we must only argue $\varphi(z(T)) < \varphi(z(0))$. According to (6.6), if this were not the case, then we would deduce $\operatorname{dist}(0; G(z(t))) = 0$ for a.e. $t \in [0, T]$. Appealing to the equality $\|\dot{z}\| = \operatorname{dist}(0; G(z(t)))$, we therefore conclude $\|\dot{z}\| = 0$ for a.e. $t \in [0, T]$. Since $z(\cdot)$ is absolutely continuous, it must therefore be constant $z(\cdot) \equiv z(0)$, but this is a contradiction since $0 \notin G(z(0))$. Thus property 2 of Assumption F holds, as claimed. □

Proof of Corollary 6.4 The result follows immediately from Lemma 6.2, once we show that Assumption F holds. Since f and g are definable in an o-minimal structure, Theorem 5.8 implies that f and g admit the chain rule. The same argument as in Theorem 5.8 moreover implies $\mathcal{X}$ admits the chain rule as well. Therefore, Lemma 6.3 guarantees that the descent property of Assumption F holds. Thus we must only argue the weak Sard property of Assumption F. To this end, since f, g, and $\mathcal{X}$ are definable in an o-minimal structure, there exist Whitney C^d-stratifications $\mathcal{A}_f$, $\mathcal{A}_g$, and $\mathcal{A}_{\mathcal{X}}$ of gph f, 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 [39, 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 (5.5), we conclude

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

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, $0 \in \partial f(x) + \partial g(x) + N_{\mathcal{X}}(x)$, holds 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. M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM J. Control Optim.*, 44(1):328–348, 2005.
3. M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. II. Applications. *Math. Oper. Res.*, 31(4):673–695, 2006.
4. J. Bolte, A. Daniilidis, A.S. Lewis, and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.
5. V.S. Borkar. *Stochastic approximation*. Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi, 2008. A dynamical systems viewpoint.
6. J.M. Borwein and X. Wang. Lipschitz func tions with maximal Clarke subdifferentials are generic. *Proc. Amer. Math. Soc.*, 128(11):3221–3229, 2000.
7. H. Brézis. *Opérateurs maximaux monotones et semi-groupes de contraction dans des espaces de Hilbert*. North-Holland Math. Stud. 5, North-Holland, Amsterdam, 1973.
8. R.E. Bruck, Jr. Asymptotic convergence of nonlinear contraction semigroups in Hilbert space. *J. Funct. Anal.*, 18:15–26, 1975.
9. J.V. Burke, X. Chen, and H. Sun. Subdifferentiation and smoothing of nonsmooth integral functionals. *Preprint, Optimization-Online*, May 2017.
10. F.H. Clarke. Generalized gradients and applications. *Trans. Amer. Math. Soc.*, 205:247–262, 1975.
11. F.H. Clarke. *Optimization and nonsmooth analysis*, volume 5 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 1990.
12. 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.
13. M. Coste. *An introduction to o-minimal geometry*. RAAG Notes, 81 pages, Institut de Recherche Mathématiques de Rennes, November 1999.
14. M. Coste. *An Introduction to Semialgebraic Geometry*. RAAG Notes, 78 pages, Institut de Recherche Mathématiques de Rennes, October 2002.
15. D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *To Appear in SIAM J. Optim.,* arXiv:1803.06523, 2018.
16. D. Davis and D. Drusvyatskiy. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. arXiv:1802.02988, 2018.
17. A. Dembo. Probability theory: Stat310/math230 september 3, 2016. 2016. Available at http://statweb.stanford.edu/~adembo/stat-310b/lnotes.pdf.
18. D. Drusvyatskiy, A.D. Ioffe, and A.S. Lewis. Curves of descent. *SIAM J. Control Optim.*, 53(1):114–138, 2015.
19. J.C. Duchi and F. Ruan. Stochastic methods for composite optimization problems. Preprint arXiv:1703.08570, 2017.
20. S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.*, 23(4):2341–2368, 2013.

21. 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.
22. A.D. Ioffe. An invitation to tame optimization. *SIAM J. Optim.*, 19(4):1894–1917, 2008.
23. A.D. Ioffe. *Variational analysis of regular mappings*. Springer Monographs in Mathematics. Springer, Cham, 2017. Theory and applications.
24. S. Kakade and J.D. Lee. Provably correct automatic subdifferentiation for qualified programs. arXiv preprint arXiv:1809.08530, 2018.
25. K.A. Khan and P.I. Barton. Evaluating an element of the Clarke generalized Jacobian of a composite piecewise differentiable function. *ACM Trans. Math. Software*, 39(4):Art. 23, 28, 2013.
26. K.A. Khan and P.I. Barton. A vector forward mode of automatic differentiation for generalized derivative evaluation. *Optimization Methods and Software*, 30(6):1185–1212, 2015.
27. 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.
28. S. Łojasiewicz. Ensemble semi-analytiques. *IHES Lecture Notes*, 1965.
29. S. Majewski, B. Miasojedow, and E. Moulines. Analysis of nonsmooth stochastic approximation: the differential inclusion approach. Preprint arXiv:1805.01916, 2018.
30. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2008.
31. 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.
32. E.A. Nurminskii. Minimization of nondifferentiable functions in the presence of noise. *Cybernetics*, 10(4):619–621, Jul 1974.
33. 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.
34. H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.
35. R.T. Rockafellar. *The theory of subgradients and its applications to problems of optimization*, volume 1 of *R & E*. Heldermann Verlag, Berlin, 1981.
36. R.T. Rockafellar and R.J-B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften, Vol 317, Springer, Berlin, 1998.
37. G.V. Smirnov. *Introduction to the theory of differential inclusions*, volume 41 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002.
38. T. Tao. *An introduction to measure theory*, volume 126. American Mathematical Soc., 2011.
39. L. van den Dries and C. Miller. Geometric categories and o-minimal structures. *Duke Math. J.*, 84:497–540, 1996.
40. H. Whitney. A function not constant on a connected set of critical points. *Duke Math. J.*, 1(4):514–517, 12 1935.
41. 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.

Affiliations

Damek Davis[1] · Dmitriy Drusvyatskiy[2] · Sham Kakade[3] · Jason D. Lee[4]

✉ Dmitriy Drusvyatskiy
ddrusv@uw.edu
https://www.math.washington.edu/~ddrusv

Damek Davis
https://people.orie.cornell.edu/dsd95/

Sham Kakade
https://homes.cs.washington.edu/∼sham/

Jason D. Lee
http://www-bcf.usc.edu/∼lee715

1 School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850, USA

2 Department of Mathematics, University of Washington, Seattle, WA 98195, USA

3 Departments of Statistics and Computer Science, University of Washington, Seattle, WA 98195, USA

4 Data Science and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA