ORIGINAL PAPER

Convergence of a stochastic subgradient method with averaging for nonsmooth nonconvex constrained optimization

Andrzej Ruszczyński[1] ^{ORCID}

Abstract

We prove convergence of a single time-scale stochastic subgradient method with subgradient averaging for constrained problems with a nonsmooth and nonconvex objective function having the property of generalized differentiability. As a tool of our analysis, we also prove a chain rule on a path for such functions.

Keywords Stochastic subgradient method · Nonsmooth optimization · Generalized differentiable functions · Chain rule

1 Introduction

We consider the problem

$$\min_{x \in X} \ f(x) \tag{1}$$

where $X \subset \mathbb{R}^n$ is convex and closed, and $f : \mathbb{R}^n \to \mathbb{R}$ is a Lipschitz continuous function, which may be neither convex nor smooth. The subgradients of $f(\cdot)$ are not available; instead, we postulate access to their random estimates.

Research on stochastic subgradient methods for nonsmooth and nonconvex functions started in late 1970's. Early contributions are due to Nurminski, who considered weakly convex functions and established a general methodology for studying convergence of non-monotonic methods [20], Gupal and his co-authors, who considered convolution smoothing (mollification) of Lipschitz functions and resulting finite-difference methods [11], and Norkin, who considered unconstrained problems with "generalized differentiable" functions [17, Ch. 3 and 7]. Recently, by an approach

✉ Andrzej Ruszczyński
  rusz@rutgers.edu

[1] Department of Management Science and Information Systems, Rutgers University, Piscataway, NJ 08854, USA

via differential inclusions, Duchi and Ruan [10] studied proximal methods for sum-composite problems with weakly convex functions, Davis *et al.* [8] proved convergence of the subgradient method for locally Lipschitz Whitney stratifiable functions with constraints, and Majewski et al. [15] studied several methods for subdifferentially regular Lipschitz functions.

Our objective is to show that a single time-scale stochastic subgradient method with direction averaging [21,22], is convergent for a broad class of functions enjoying the property of "generalized differentiability," which contains all classes of functions mentioned above, as well as their compositions.

Our analysis follows the approach of relating a stochastic approximation algorithm to a continuous-time dynamical system, pioneered in [13,14] and developed in many works (see, e.g., [12] and the references therein). Extension to multifunctions was proposed in [1] and further developed, among others, in [3,8,10,15].

For the purpose of our analysis, we also prove a chain rule on a path under generalized differentiability, which may be of independent interest.

Finally, we illustrate the use of the method for training a *ReLu* neural network.

2 The chain formula on a path

Norkin [19] introduced the following class of functions.

Definition 1 A function $f : \mathbb{R}^n \to \mathbb{R}$ is *differentiable in a generalized sense at a point* $x \in \mathbb{R}^n$, if an open set $U \subset \mathbb{R}^n$ containing x, and a nonempty, convex, compact valued, and upper semicontinuous multifunction $G_f : U \rightrightarrows \mathbb{R}^n$ exist, such that for all $y \in U$ and all $g \in G_f(y)$ the following equation is true:

$$f(y) = f(x) + \langle g(y), y - x \rangle + o(x, y, g),$$

with

$$\lim_{y \to x} \sup_{g \in G(y)} \frac{o(x, y, g)}{\|y - x\|} = 0.$$

The set $G_f(y)$ is the *generalized subdifferential* of f at y. If a function is differentiable in a generalized sense at every $x \in \mathbb{R}^n$ with the same generalized subdifferential mapping $G_f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, we call it *differentiable in a generalized sense*.

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable in a generalized sense, if each of its component functions, $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, has this property.

The class of such functions is contained in the set of locally Lipschitz functions [17, Thm. 1.1], and contains all subdifferentially regular functions [5], Whitney stratifiable Lipschitz functions [9], semismooth functions [16], and their compositions. In fact, if a function is differentiable in generalized sense and has directional derivatives at x in every direction, then it is semismooth at x. The Clarke subdifferential $\partial f(x)$ is an inclusion-minimal generalized subdifferential, but the generalized subdifferential mapping $G_f(\cdot)$ is not uniquely defined in Definition 1, which plays a role in our

considerations. For stochastic optimization, essential is the closure of the class of such functions with respect to expectation, which allows for easy generation of stochastic subgradients. In the Appendix we recall basic properties of functions differentiable in a generalized sense. For thorough exposition, see [17, Ch. 1 and 6].

Our interest is in a formula for calculating the increment of a function $f : \mathbb{R}^n \to \mathbb{R}$ along a path $p : [0, \infty) \to \mathbb{R}^n$, which is at the core of the analysis of nonsmooth and stochastic optimization algorithms (see [7,9] and the references therein). For an absolutely continuous function $p : [0, \infty) \to \mathbb{R}^n$ we denote by $\dot{p}(\cdot)$ its weak derivative, that is, a measurable function such that

$$p(t) = p(0) + \int_0^t \dot{p}(s) \, ds, \quad \forall \, t \geq 0.$$

Theorem 1 *If $f : \mathbb{R}^n \to \mathbb{R}$ and $p : [0, \infty) \to \mathbb{R}^n$ are differentiable in a generalized sense, then for every $T > 0$, any generalized subdifferential $G_f(\cdot)$, and every selection $g(p(t)) \in G_f(p(t))$, we have*

$$f(p(T)) - f(p(0)) = \int_0^T \left\langle g(p(t)), \dot{p}(t) \right\rangle dt. \tag{2}$$

Proof Consider the function

$$\varphi(\varepsilon) = \int_0^T f(p(t + \varepsilon)) \, dt, \quad \varepsilon \geq 0.$$

Its right derivative at 0 can be calculated in two ways:

$$\begin{aligned}
\varphi'_+(0) &= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left[\varphi(\varepsilon) - \varphi(0) \right] = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left[\int_0^T f(p(t + \varepsilon)) \, dt - \int_0^T f(p(t)) \, dt \right] \\
&= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left[\int_\varepsilon^{T+\varepsilon} f(p(\tau)) \, d\tau - \int_0^T f(p(t)) \, dt \right] \\
&= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left[\int_T^{T+\varepsilon} f(p(t)) \, dt - \int_0^\varepsilon f(p(t)) \, dt \right] = f(p(T)) - f(p(0)).
\end{aligned} \tag{3}$$

On the other hand,

$$\varphi'_+(0) = \lim_{\varepsilon \downarrow 0} \int_0^T \frac{1}{\varepsilon} \left[f(p(t + \varepsilon)) - f(p(t)) \right] dt. \tag{4}$$

By the generalized differentiability of $f(\cdot)$, the differential quotient under the integral can be expanded as follows:

$$\frac{1}{\varepsilon}\Big[f(p(t+\varepsilon))\,dt - f(p(t))\Big]$$

$$= \frac{1}{\varepsilon}\Big\langle g(p(t+\varepsilon)), p(t+\varepsilon) - p(t)\Big\rangle + \frac{1}{\varepsilon}o\Big(p(t), p(t+\varepsilon), g(p(t+\varepsilon))\Big),$$

$$\text{with} \quad \lim_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}o\Big(p(t), p(t+\varepsilon), g(p(t+\varepsilon))\Big) = 0. \tag{5}$$

Since $p(\cdot)$ is differentiable in a generalized sense, it is locally Lipschitz continuous [17, Thm. 1.1], hence absolutely continuous. Therefore, for almost all t, we have $\frac{1}{\varepsilon}\big[p(t+\varepsilon) - p(t)\big] = \dot{p}(t) + r(t, \varepsilon)$, with $\lim_{\varepsilon\downarrow 0} r(t, \varepsilon) = 0$. Combining this with (5), and using the local boundedness of generalized gradients, we obtain

$$\frac{1}{\varepsilon}\Big[f(p(t+\varepsilon)) - f(p(t))\Big] = \Big\langle g(p(t+\varepsilon)), \dot{p}(t)\Big\rangle + O(t, \varepsilon), \tag{6}$$

with $\lim_{\varepsilon\downarrow 0} O(t, \varepsilon) = 0$. By [17, Thm. 1.6] (Theorem 3), the function $\psi(t) = f(p(t))$ is differentiable in a generalized sense as well, and

$$G_\psi(t) = \Big\{\langle g, h\rangle : g \in G_f(p(t)),\ h \in G_p(t)\Big\}$$

is its generalized subdifferential. By virtue of [17, Cor. 1.5] (Theorem 5), any generalized subdifferential mapping $G_\psi(\cdot)$ is single-valued except for a countable number of points in $[0, 1]$. Since it is upper semicontinuous, it is continuous almost everywhere. By [17, Thm. 1.12] (Theorem 4), almost everywhere $G_p(t) = \{\dot{p}(t)\}$. Then for any $h(t+\varepsilon) \in G_p(t+\varepsilon)$ and for almost all t,

$$\lim_{\varepsilon\downarrow 0}\Big\langle g(p(t+\varepsilon)), h(t+\varepsilon)\Big\rangle = \Big\langle g(p(t)), \dot{p}(t)\Big\rangle.$$

Therefore, for almost all t,

$$\lim_{\varepsilon\downarrow 0}\Big\langle g(p(t+\varepsilon)), \dot{p}(t)\Big\rangle$$

$$= \Big\langle g(p(t)), \dot{p}(t)\Big\rangle + \lim_{\varepsilon\downarrow 0}\Big\langle g(p(t+\varepsilon)), \dot{p}(t) - h(t+\varepsilon)\Big\rangle = \Big\langle g(p(t)), \dot{p}(t)\Big\rangle,$$

where the last equation follows from the local boundedness of $G_f(\cdot)$ and the continuity of $G_p(\cdot)$ at the points of differentiability. Thus, for almost all t, we can pass to the limit in (6):

$$\lim_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}\Big[f(p(t+\varepsilon))\,dt - f(p(t))\Big] = \Big\langle g(p(t)), \dot{p}(t)\Big\rangle.$$

We can now use the Lebesgue theorem and pass to the limit under the integral in (4):

$$\varphi'_+(0) = \int_0^T \lim_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}\Big[f(p(t+\varepsilon))\,dt - f(p(t))\Big]\,dt = \int_0^T \Big\langle g(p(t)), \dot{p}(t)\Big\rangle\,dt.$$

Comparison with (3) yields (2). $\qquad\square$

3 The single time-scale method with subgradient averaging

We briefly recall from [21,22] a stochastic approximation algorithm for solving problem (1) where only random estimates of subgradients of f are available.

The method generates two random sequences: approximate solutions $\{x^k\}$ and path-averaged stochastic subgradients $\{z^k\}$, defined on a certain probability space $(\Omega, \mathscr{F}, P)$. We let $\mathscr{F}_k$ to be the σ-algebra generated by $\{x^0, \ldots, x^k, z^0, \ldots, z^k\}$. We assume that for each k, we can observe an $\mathscr{F}_k$-measurable random vector $g^k \in \mathbb{R}^n$, such that, for some $\mathscr{F}_k$-measurable vector r^k, we have $g^k - r^k \in G_f(x^k)$. Further assumptions on the errors r^k will be specified in Sect. 4.

The method proceeds for $k = 0, 1, 2 \ldots$ as follows ($a > 0$ and $\beta > 0$ are fixed parameters). We compute

$$y^k = \underset{y \in X}{\operatorname{argmin}} \left\{ \langle z^k, y - x^k \rangle + \frac{\beta}{2} \|y - x^k\|^2 \right\}, \tag{7}$$

and, with an $\mathscr{F}_k$-measurable stepsize $\tau_k \in \left(0, \min(1, 1/a) \right]$, we set

$$x^{k+1} = x^k + \tau_k (y^k - x^k). \tag{8}$$

Then we observe g^{k+1} at x^{k+1}, and update the averaged stochastic subgradient as

$$z^{k+1} = (1 - a\tau_k)z^k + a\tau_k g^{k+1}. \tag{9}$$

Convergence of the method was proved in [22] for weakly convex functions $f(\cdot)$. Unfortunately, this class does not contain functions with downward cusps, which are common in modern machine learning models (see Sect. 5).

4 Convergence analysis

We call a point $x^* \in \mathbb{R}^n$ *Clarke stationary* of problem (1), if

$$0 \in \partial f(x^*) + N_X(x^*), \tag{10}$$

where $N_X(x^*)$ denotes the normal cone to X at x^*. The set of Clarke stationary points of problem (1) is denoted by X^*.

We start from a useful property of the gap function $\eta : X \times \mathbb{R}^n \to (-\infty, 0]$,

$$\eta(x, z) = \min_{y \in X} \left\{ \langle z, y - x \rangle + \frac{\beta}{2} \|y - x\|^2 \right\}. \tag{11}$$

We denote the minimizer in (11) by $\bar{y}(x, z)$. Since it is a projection of $x - z/\beta$ on X, we observe that

$$\langle z, \bar{y}(x, z) - x \rangle + \beta \|\bar{y}(x, z) - x\|^2 \le 0. \tag{12}$$

Moreover, a point $x^* \in X^*$ if and only if $g^* \in \partial f(x^*)$ exists such that $\eta(x^*, g^*) = 0$.

We analyze convergence of the algorithm (7)–(9) under the following conditions, the first three of which are assumed to hold with probability 1:

(A1) All iterates x^k belong to a compact set;

(A2) $\tau_k \in \left(0, \min(1, 1/a)\right]$ for all k, $\lim_{k \to \infty} \tau_k = 0$, $\sum_{k=0}^{\infty} \tau_k = \infty$;

(A3) For all k, $r^k = e^k + \delta^k$, with $\sum_{k=0}^{\infty} \tau_k e^k$ convergent, and $\lim_{k \to \infty} \delta_k = 0$;

(A4) The set $\{f(x) : x \in X^*\}$ does not contain an interval of nonzero length.

Condition (A3) can be satisfied for a martingale $\sum_{k=0}^{\infty} \tau_k e^k$, but can also hold for broad classes of dependent "noise" sequences $\{e^k\}$ [12]. Condition (A4) is true for Whitney stratifiable functions [2, Cor. 5], but we need to assume it here.

We have the following elementary property of the sequence $\{z^k\}$.

Lemma 1 *Suppose the sequence $\{x^k\}$ is included in a set $A \subset \mathbb{R}^n$ and conditions* (A2) *and* (A3) *are satisfied. Then*

$$\lim_{k \to \infty} dist(z^k, B) = 0, \quad where \quad B = conv\left(\bigcup_{x \in A} \partial f(x) \right).$$

Proof Using (A2), we define the quantities $\tilde{z}^k = z^k + a \sum_{j=k}^{\infty} \tau_j e^j$ and establish the recursive relation

$$\tilde{z}^{k+1} = (1 - a\tau_k)\tilde{z}^k + a\tau_k g^k + \tau_k \Delta_k, \quad k = 0, 1, 2, \dots,$$

where $g^k \in B$ and $\Delta_k = a\delta^k + a\sum_{j=k}^{\infty} \tau_j e^j \to 0$ a.s.. The convexity of the distance function and (A2) yield the result. □

Theorem 2 *If assumptions* (A1)–(A4) *are satisfied, then, with probability 1, every accumulation point $\hat{x}$ of the sequence $\{x^k\}$ is Clarke stationary, and the sequence $\{f(x^k)\}$ is convergent.*

Proof Due to (A1), by virtue of Lemma 1, the sequence $\{z^k\}$ is bounded. We divide the proof into three standard steps.

Step 1: The Limiting Dynamical System. We define $p^k = (x^k, z^k)$, accumulated stepsizes $t_k = \sum_{j=0}^{k-1} \tau_j$, $k = 0, 1, 2 \dots$, and we construct the interpolated trajectory

$$P_0(t) = p^k + \frac{t - t_k}{\tau_k}(p^{k+1} - p^k), \quad t_k \le t \le t_{k+1}, \quad k = 0, 1, 2, \dots.$$

For an increasing sequence of positive numbers $\{s_k\}$ diverging to ∞, we define shifted trajectories $P_k(t) = P_0(t + s_k)$. Recall that $P_k(t) = \left(X_k(t), Z_k(t)\right)$.

By [15, Thm. 3.2], for any increasing sequence $\{n_k\}$ of positive integers, there exist a subsequence $\{\tilde{n}_k\}$ and absolutely continuous functions $X_\infty : [0, +\infty) \to X$ and $Z_\infty : [0, +\infty) \to \mathbb{R}^n$ such that for any $T > 0$

$$\lim_{k \to \infty} \sup_{t \in [0,T]} \left(\left\| X_{\tilde{n}_k}(t) - X_\infty(t) \right\| + \left\| Z_{\tilde{n}_k}(t) - Z_\infty(t) \right\| \right) = 0,$$

and $(X_\infty(\cdot), Z_\infty(\cdot))$ is a solution of the system of differential equations and inclusions:

$$\dot{x}(t) = \bar{y}\Big(x(t), z(t)\Big) - x(t), \tag{13}$$

$$\dot{z}(t) \in a\Big(\partial f(x(t)) - z(t)\Big). \tag{14}$$

Moreover, for any $t \geq 0$, the pair $(X_\infty(t), Z_\infty(t))$ is an accumulation point of the sequence $\{(x^k, z^k)\}$.

Step 2: Descent Along a Path. We use the Lyapunov function

$$W(x, z) = af(x) - \eta(x, z).$$

For any solution $(X(t), Z(t))$ of the system (13)–(14), and for any $T > 0$, we estimate the difference $W(X(T), Z(T)) - W(X(0), Z(0))$. We split $W(X(\cdot), Z(\cdot))$ into a generalized differentiable composition $f(X(\cdot))$ and the "classical" part $\eta(X(\cdot), Z(\cdot))$.

Since the path $X(\cdot)$ satisfies (13) and $\bar{y}(\cdot, \cdot)$ is continuous, $X(\cdot)$ is continuously differentiable. Thus, we can use Theorem 1 to conclude that for any $g(X(\cdot)) \in \partial f(X(\cdot))$

$$f(X(T)) - f(X(0)) = \int_0^T \Big\langle g(X(t)), \dot{X}(t)\Big\rangle \, dt$$

$$= \int_0^T \Big\langle g(X(t)), \bar{y}(X(t), Z(t)) - X(t)\Big\rangle \, dt. \tag{15}$$

On the other hand, since $\bar{y}(x, z)$ is unique, the function $\eta(\cdot, \cdot)$ is continuously differentiable. Therefore, the chain formula holds for it as well:

$$\eta(X(T), Z(T)) - \eta(X(0), Z(0))$$
$$= \int_0^T \Big\langle \nabla_x \eta(X(t), Z(t)), \dot{X}(t)\Big\rangle \, dt + \int_0^T \Big\langle \nabla_z \eta(X(t), Z(t)), \dot{Z}(t)\Big\rangle \, dt.$$

Substituting $\nabla_x \eta(x, z) = -z + \beta(x - \bar{y}(x, z))$, $\nabla_z \eta(x, z) = \bar{y}(x, z) - x$ and $\dot{Z}(t) = a\Big(\hat{g}(X(t)) - Z(t)\Big)$ with some $\hat{g}(X(\cdot)) \in \partial f(X(\cdot))$, and using (12) we obtain

$$\eta(X(T), Z(T)) - \eta(X(0), Z(0))$$
$$= \int_0^T \Big\langle -Z(t) + \beta(X(t) - \bar{y}(X(t), Z(t))), \, \bar{y}(X(t), Z(t)) - X(t)\Big\rangle \, dt$$
$$+ a \int_0^T \Big\langle \bar{y}(X(t), Z(t)) - X(t), \, \hat{g}(X(t)) - Z(t)\Big\rangle \, dt$$
$$\geq a \int_0^T \Big\langle \bar{y}(X(t), Z(t)) - X(t), \, \hat{g}(X(t)) - Z(t)\Big\rangle \, dt$$
$$\geq a \int_0^T \Big\langle \bar{y}(X(t), Z(t)) - X(t), \, \hat{g}(X(t))\Big\rangle \, dt + a\beta \int_0^T \Big\| \bar{y}(X(t), Z(t)) - X(t)\Big\|^2 \, dt.$$

We substitute the subgradient selector $g(X(t)) = \hat{g}(X(t))$ into (15) and combine it with the last inequality, concluding that

$$
\begin{aligned}
&W(X(T), Z(T)) - W(X(0), Z(0)) \\
&\quad \le -a\beta \int_0^T \left\| \bar{y}(X(t), Z(t)) - X(t) \right\|^2 dt = -a\beta \int_0^T \left\| \dot{X}(t) \right\|^2 dt. \quad (16)
\end{aligned}
$$

Step 3: Analysis of limit points Define the set $\mathscr{S} = \left\{ (x, z) \in X^* \times \mathbb{R}^n : \eta(x, z) = 0 \right\}$. Suppose $(\bar{x}, \bar{z})$ is an accumulation point of the sequence $\{(x^k, z^k)\}$. If $\eta(\bar{x}, \bar{z}) < 0$, then every solution $(X(t), Z(t))$ of the system (13)–(14), starting from $(X(0), Z(0)) = (\bar{x}, \bar{z})$ has $\dot{X}(0) \ne 0$. Using (16) and arguing as in [10, Thm. 3.20] or [15, Thm. 3.5], we obtain a contradiction. Therefore, we must have $\eta(\bar{x}, \bar{z}) = 0$. Suppose $\bar{x} \notin X^*$. Then

$$
\operatorname{dist}\left(0, \partial f(\bar{x}) + N_X(\bar{x})\right) > 0. \quad (17)
$$

Suppose $X(t) = \bar{x}$ for all $t \ge 0$. The inclusion (14) simplifies: $\dot{z}(t) \in a\left(\partial f(\bar{x}) - z(t)\right)$. By using the convex Lyapunov function $V(z) = \operatorname{dist}\left(z, \partial f(\bar{x})\right)$ and applying the classical chain formula on the path $Z(\cdot)$ [4], we deduce that

$$
\lim_{t \to \infty} \operatorname{dist}\left(Z(t), \partial f(\bar{x})\right) = 0. \quad (18)
$$

It follows from (17)–(18) that $T > 0$ exists, such that $-Z(T) \notin N_X(\bar{x})$, which yields $\dot{X}(T) \ne 0$. Consequently, the path $X(t)$ starting from $\bar{x}$ cannot be constant. But then again $T > 0$ exists, such that $\dot{X}(T) \ne 0$. By Step 1, the pair $(X(T), Z(T))$ would have to be an accumulation point of of the sequence $\{(x^k, z^k)\}$, a case already excluded. We conclude that every accumulation point $(\bar{x}, \bar{z})$ of the sequence $\{(x^k, z^k)\}$ is in $\mathscr{S}$. The convergence of the sequence $\left\{ W(x^k, z^k) \right\}$ then follows in the same way as [10, Thm. 3.20] or [15, Thm. 3.5]. As $\eta(x^k, z^k) \to 0$, the convergence of $\{f(x^k)\}$ follows as well. $\qquad \square$

Directly from Lemma 1 we obtain convergence of averaged stochastic subgradients.

Corollary 1 *If the sequence $\{x^k\}$ is convergent to a single point $\bar{x}$, then every accumulation point of $\{z^k\}$ is an element of $\partial f(\bar{x})$.*

5 Example

A *Rectified Linear Unit (ReLU)* neural network [18] predicts a random outcome $Y \in \mathbb{R}^m$ from random features $X \in \mathbb{R}^n$ by a nonconvex nonsmooth function $y(X, W)$, defined recursively as follows:

$$
s_1 = X, \quad s_{\ell+1} = (W_\ell s_\ell)_+, \ \ell = 1, 2, \ldots, L - 1, \quad y(X, W) = W_L s_L,
$$
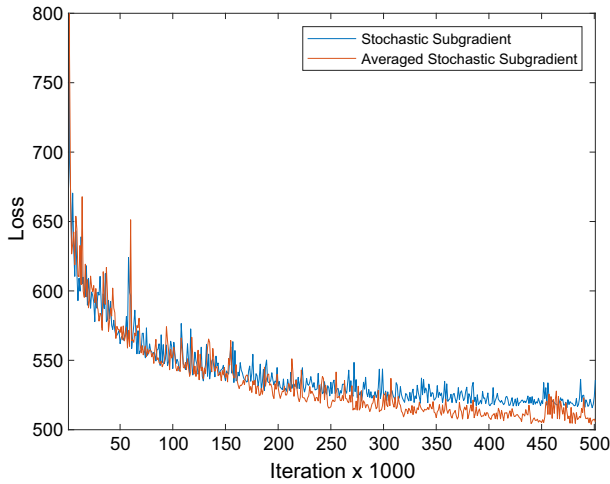
Fig. 1 Comparison of methods with (lower graph) and without averaging (upper graph)

where $(v)_+ = \max(0, v)$, componentwise. The decision variables are $W_1, \ldots, W_{L-1} \in \mathbb{R}^{n \times n}$ and $W_L \in \mathbb{R}^{m \times n}$. The simplest training problem is:

$$\min_{W \in \mathscr{W}} \ f(W) \triangleq \frac{1}{2} \mathbb{E}\Big[\|y(X, W) - Y\|^2\Big], \tag{19}$$

where $\mathscr{W}$ is a box about 0. The function $f(W)$ is not subdifferentially regular. It is not Whitney stratifiable, in general, because this property is not preserved under the expected value operator. However, we can use Theorems 3 and 6 to verify that it is differentiable in a generalized sense, and to calculate its stochastic subgradients. For a random data point (X^k, Y^k) we subdifferentiate the function under the expected value in (19) by recursive application of Theorem 3. In particular, for $L = 2$ and $m = 1$ we have $y(X, W) = W_2(W_1 X)_+$, and

$$g^k = \Big(y(X^k, W^k) - Y^k\Big)\big[\, D^k (W_2^k)^T (X^k)^T \ \ (W_1^k X^k)_+^T \,\big].$$

Here, D^k is a diagonal matrix with 1 on position (i, i), if $(W_1^k X^k)_i > 0$, and 0 otherwise. A typical run of the stochastic subgradient method and the method with direction averaging is shown in Fig. 1, on an example of predicting wine quality [6], with identical random starting points, sequences of observations, and schedules of stepsizes: $\tau_k = 0.03/(1 + 5k/N)$, where $N = 500,000$. The coefficient $a = 0.1$. For comparison, the loss of a linear regression model is 666.

Appendix A Generalized differentiability of functions

Compositions of generalized differentiable functions are crucial in our analysis.

Theorem 3 [17, Thm. 1.6] *If* $h : \mathbb{R}^m \to \mathbb{R}$ *and* $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, *are differentiable in a generalized sense, then the composition* $\psi(x) = h\Big(f_1(x), \ldots, f_m(x)\Big)$ *is differentiable in a generalized sense, and at any point* $x \in \mathbb{R}^n$ *we can define the generalized subdifferential of* ψ *as follows:*

$$
\begin{aligned}
G_\psi(x) = \ conv\Big\{ g \in \mathbb{R}^n : g = \big[g_1 \cdots g_m\big] g_0, \\
\text{with } g_0 \in G_h\Big(f_1(x), \ldots, f_m(x)\Big) \text{ and } g_j \in G_{f_j}(x), \ j = 1, \ldots, m\Big\}.
\end{aligned}
\tag{20}
$$

Even if we take $G_h(\cdot) = \partial h(\cdot)$ and $G_{f_j}(\cdot) = \partial f_j(\cdot)$, $j = 1, \ldots, m$, we may obtain $G_\psi(\cdot) \neq \partial \psi(\cdot)$, but G_ψ defined above satisfies Definition 1.

Theorem 4 [17, Thm. 1.12] *If* $f : \mathbb{R}^n \to \mathbb{R}$ *is differentiable in a generalized sense, then for almost all* $x \in \mathbb{R}^n$ *we have* $G_f(x) = \{\nabla f(x)\}$.

Functions of one variable have the following remarkable property.

Theorem 5 [17, Cor. 1.5] *If* $f : \mathbb{R} \to \mathbb{R}$ *is differentiable in a generalized sense, then the set of points* x *at which a generalized subdifferential* $G_f(x)$ *is not a singleton is at most countable.*

For stochastic optimization, essential is the closure of the class functions differentiable in a generalized sense with respect to expectation.

Theorem 6 [17, Thm. 23.1] *Suppose* $(\Omega, \mathscr{F}, P)$ *is a probability space and a function* $f : \mathbb{R}^n \times \Omega \to \mathbb{R}$ *is differentiable in a generalized sense with respect to* x *for all* $\omega \in \Omega$, *and integrable with respect to* ω *for all* $x \in \mathbb{R}^n$. *Let* $G_f : \mathbb{R}^n \times \Omega \rightrightarrows \mathbb{R}^n$ *be a multifunction, which is measurable with respect to* ω *for all* $x \in \mathbb{R}^n$, *and which is a generalized subdifferential mapping of* $f(\cdot, \omega)$ *for all* $\omega \in \Omega$. *If for every compact set* $K \subset \mathbb{R}^n$ *an integrable function* $L_K : \Omega \to \mathbb{R}$ *exists, such that* $\sup_{x \in K} \sup_{g \in G_f(x, \omega)} \|g\| \leq L_K(\omega)$, $\omega \in \Omega$, *then the function*

$$
F(x) = \int_\Omega f(x, \omega) \ P(d\omega), \quad x \in \mathbb{R}^n,
$$

is differentiable in a generalized sense, and the multifunction

$$
G_F(x) = \int_\Omega G_f(x, \omega) \ P(d\omega), \quad x \in \mathbb{R}^n,
$$

is its generalized subdifferential mapping.

References

1. Benaïm, M., Hofbauer, J., Sorin, S.: Stochastic approximations and differential inclusions. SIAM J. Control Optim. **44**(1), 328–348 (2005)
2. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM J. Optim. **18**(2), 556–572 (2007)
3. Borkar, V.S.: Stochastic Approximation: A Dynamical Systems Viewpoint. Springer, New York (2009)
4. Brézis, H.: Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations. In: Zarantonello, E.H. (ed.) Contributions to Nonlinear Functional Analysis, pp. 101–156. Elsevier, Amsterdam (1971)
5. Clarke, F.H.: Generalized gradients and applications. Trans. Am. Math. Soc. **205**, 247–262 (1975)
6. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J.: Modeling wine preferences by data mining from physicochemical properties. Decis. Support Syst. **47**(4), 547–553 (2009)
7. Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. SIAM J. Optim. **29**(1), 207–239 (2019)
8. Davis, D., Drusvyatskiy, D., Kakade, S., Lee, J.D.: Stochastic subgradient method converges on tame functions. Found. Comput. Math. (2019). https://doi.org/10.1007/s10208-018-09409-5
9. Drusvyatskiy, D., Ioffe, A.D., Lewis, A.S.: Curves of descent. SIAM J. Control Optim. **53**(1), 114–138 (2015)
10. Duchi, J.C., Ruan, F.: Stochastic methods for composite and weakly convex optimization problems. SIAM J. Optim. **28**(4), 3229–3259 (2018)
11. Gupal, A.M.: Stochastic Methods for Solving Nonsmooth Extremal Problems. Naukova Dumka, Kiev (1979)
12. Kushner, H., Yin, G.G.: Stochastic Approximation Algorithms and Applications. Springer, New York (2003)
13. Kushner, H.J., Clark, D.S.: Stochastic Approximation Methods for Constrained and Unconstrained Systems. Springer, New York (1978)
14. Ljung, L.: Analysis of recursive stochastic algorithms. IEEE Trans. Autom. Control **22**(4), 551–575 (1977)
15. Majewski, S., Miasojedow, B., Moulines, E.: Analysis of nonsmooth stochastic approximation: the differential inclusion approach. *arXiv preprint* arXiv:1805.01916 (2018)
16. Mifflin, R.: Semismooth and semiconvex functions in constrained optimization. SIAM J. Control Optim. **15**(6), 959–972 (1977)
17. Mikhalevich, V.S., Gupal, A.M., Norkin, V.I.: Nonconvex Optimization Methods. Nauka, Moscow (1987)
18. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814 (2010)
19. Norkin, V.I.: Generalized-differentiable functions. Cybern. Syst. Anal. **16**(1), 10–12 (1980)
20. Nurminski, E.A.: Numerical Methods for Solving Deterministic and Stochastic Minimax Problems. Naukova Dumka, Kiev (1979)
21. Ruszczyński, A.: A method of feasible directions for solving nonsmooth stochastic programming problems. In: Archetti, F., Di Pillo, G., Lucertini, M. (eds.) Stochastic Programming, pp. 258–271. Springer, Berlin (1986)
22. Ruszczyński, A.: A linearization method for nonsmooth stochastic programming problems. Math. Oper. Res. **12**(1), 32–49 (1987)