SIAM J. OPTIM.
Vol. 28, No. 4, pp. 3229–3259

STOCHASTIC METHODS FOR COMPOSITE AND WEAKLY CONVEX OPTIMIZATION PROBLEMS*

JOHN C. DUCHI† AND FENG RUAN†

Abstract. We consider minimization of stochastic functionals that are compositions of a (potentially) nonsmooth convex function h and smooth function c and, more generally, stochastic weakly convex functionals. We develop a family of stochastic methods—including a stochastic prox-linear algorithm and a stochastic (generalized) subgradient procedure—and prove that, under mild technical conditions, each converges to first order stationary points of the stochastic objective. We provide experiments further investigating our methods on nonsmooth phase retrieval problems; the experiments indicate the practical effectiveness of the procedures.

Key words. stochastic optimization, composite optimization, differential inclusion

AMS subject classification. 65K10

DOI. 10.1137/17M1135086

1. Introduction.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be the stochastic composite function

$$f(x) := \mathbb{E}_P[h(c(x; S); S)] = \int_{\mathcal{S}} h(c(x; s); s) dP(s), \tag{1}$$

where P is a probability distribution on a sample space $\mathcal{S}$ and for each $s \in \mathcal{S}$, the function $z \mapsto h(z; s)$ is closed convex and $x \mapsto c(x; s)$ is smooth. In this paper, we consider stochastic methods for minimization—or at least finding stationary points—of such composite functionals. The objective (1) is an instance of the more general problem of stochastic weakly convex optimization, where $f(x) := \mathbb{E}_P[f(x; S)]$ and for each x_0 and $s \in \mathcal{S}$, there is $\lambda(s, x_0)$ such that $x \mapsto f(x; s) + \frac{\lambda(s,x_0)}{2} \|x - x_0\|^2$ is convex in a neighborhood of x_0. (We show later how problem (1) falls in this framework.) Such functions have classical and modern applications in optimization [25, 18, 43, 47], for example, in phase retrieval [23] problems or training deep linear neural networks (e.g., [31]). We thus study the problem

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) + \varphi(x) = \mathbb{E}_P[f(x; S)] + \varphi(x) \\ \text{subject to} \quad & x \in X, \end{aligned} \tag{2}$$

where $X \subset \mathbb{R}^d$ is a closed convex set and $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a closed convex function.

Many problems are representable in the form (1). Taking the function c as the identity mapping, classical regularized stochastic convex optimization problems fall into this framework [40], including regularized least-squares and the Lasso [32, 51], with $s = (a, b) \in \mathbb{R}^d \times \mathbb{R}$ and $h(x; s) = \frac{1}{2}(a^T x - b)^2$ and φ typically some norm on x, or supervised learning objectives such as logistic regression or support vector

*Received by the editors June 19, 2017; accepted for publication (in revised form) September 17, 2018; published electronically November 27, 2018.

http://www.siam.org/journals/siopt/28-4/M113508.html

Funding: The author's work was partially supported by NSF award CCF-1553086 and an Alfred P. Sloan fellowship. The work of the second author was supported by an E.K. Potter Stanford Graduate Fellowship.

†Department of Statistics and Electrical Engineering, Stanford University, Stanford, CA 94305 (jduchi@stanford.edu, fengruan@stanford.edu).

machines [32]. The more general settings (1)–(2) include a number of important nonconvex problems. Examples include nonlinear least squares (cf. [42]), with $s = (a, b)$ and $b \in \mathbb{R}$, the convex term $h(t; s) \equiv h(t) = \frac{1}{2}t^2$ independent of the sampled s, and $c(x; s) = c_0(x; a) - b$, where c_0 is some smooth function a modeler believes predicts b well given x and data a. Another compelling example is the (robust) phase retrieval problem [8, 49]—which we explore in more depth in our numerical experiments—where the data $s = (a, b) \in \mathbb{R}^d \times \mathbb{R}_+$, $h(t; s) \equiv h(t) = |t|$ or $h(t; s) \equiv h(t) = \frac{1}{2}t^2$, and $c(x; s) = (a^T x)^2 - b$. In the case that $h(t) = |t|$, the form (1) is an exact penalty for the solution of a collection of quadratic equalities $(a_i^T x)^2 = b_i$, $i = 1, \ldots, N$, where we take P to be point masses on pairs (a_i, b_i).

Fletcher and Watson [29, 28] initiated work on composite problems to

$$\underset{x}{\text{minimize}} \quad h(c(x)) + \varphi(x) \text{ subject to } x \in X \tag{3}$$

for fixed convex h, smooth c, convex φ, and convex X. A motivation of this earlier work is nonlinear programming problems with the constraint that $x \in \{x : c(x) = 0\}$, in which case taking $h(z) = \|z\|$ functions as an exact penalty [33] for the constraint $c(x) = 0$. A more recent line of work, beginning with Burke [7] and continued by (among others) Druvyatskiy, Ioffe, Lewis, Pacquette, and Wright [38, 21, 19, 20], establishes convergence rate guarantees for methods that sequentially minimize convex surrogates for problem (3).

Roughly, these papers construct a model of the composite function $f(x) = h(c(x))$ as follows. Letting $\nabla c(x)$ be the transpose of the Jacobian of c at x, so $c(y) = c(x) + \nabla c(x)^T(y - x) + o(\|y - x\|)$, one defines the "linearized" model of f at x by

$$f_x(y) := h(c(x) + \nabla c(x)^T(y - x)), \tag{4}$$

which is convex in y. When h and ∇c are Lipschitzian, then $|f_x(y) - f(x)| = O(\|x - y\|^2)$, so that the model (4) is second-order accurate, which motivates the following *prox-linear* method. Beginning from some $x_0 \in X$, iteratively construct

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x) + \varphi(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}, \tag{5}$$

where $\alpha_k > 0$ is a stepsize that may be chosen by a line search. For small α_k, the iterates (5) guarantee decreasing $h(c(x_k)) + \varphi(x_k)$, the sequence of problems (5) are convex, and moreover, the iterates x_k converge to stationary points of problem (3) [19, sect. 5]. The prox-linear method is effective so long as minimizing the models $f_{x_k}(x)$ is reasonably computationally easy. More generally, minimizing a sequence of models f_{x_k} of f centered around the iterate x_k is natural, with examples including Rockafellar's proximal point algorithm [46] and general sequential convex programming approaches, such as trust region and other Taylor-like methods [42, 11, 19, 18].

In our problem (2), where $f(x) = \mathbb{E}[f(x; S)]$ for $f(\cdot, s)$ weakly convex or composite, the iterates (5) may be computationally challenging. Even in the case in which P is discrete so that problem (1) has the form $f(x) = \frac{1}{n}\sum_{i=1}^n h_i(c_i(x))$, which is evidently of the form (3), the iterations generating x_k may be prohibitively expensive for large n. When P is continuous or is unknown, because we can only simulate draws $S \sim P$ or in statistical settings where the only access to P is via observations $S_i \sim P$, then the iteration (5) is essentially infeasible. Given the wide applicability of the stochastic composite problem (1), it is of substantial interest to develop efficient online and stochastic methods to (approximately) solve it, or at least to find local optima.

In this work, we develop and study stochastic model-based algorithms, examples of which include a stochastic linear proximal algorithm, which is a stochastic analogue of problem (5), and a stochastic subgradient algorithm, both of whose definitions we give in section 2. The iterations of such methods are often computationally simple, and they require only individual samples $S \sim P$ at each iteration. Consider for concreteness the case when P is discrete and supported on $i = 1, \ldots, n$ (i.e., $f(x) = \frac{1}{n} \sum_{i=1}^{n} h_i(c_i(x))$). Then instead of solving the nontrivial subproblem (5), the stochastic prox-linear algorithm samples $i_0 \in [n]$ uniformly, then substitutes h_{i_0} and c_{i_0} for h and c in the iteration. Thus, as long as there is a prox-linear step for the individual compositions $h_i \circ c_i$, the algorithm is easy to implement and execute.

The main result of this paper is that the stochastic model-based methods we develop for the stochastic composite and weakly-convex optimization problems are convergent. More precisely, assuming that (i) with probability one, the iterates of the procedures are bounded, (ii) the objective function $F + \mathbb{I}_X$ is coercive, and (iii) second moment conditions on local-Lipschitzian and local-convexity parameters of the random functions $f(\cdot, s)$, any appropriate model-based stochastic minimization strategy has limit points taking values $f(x)$ in the set of stationary values of the function. If the image of noncritical points of the objective function is dense in $\mathbb{R}$, the methods converge to stationary points of the (potentially) nonsmooth, nonconvex objective (2) (Theorem 1 in section 2 and Theorem 5 in section 3.4). As gradients $\nabla f(x)$ may not exist (and may not even be zero at stationary points because of the nonsmoothness of the objective), demonstrating this convergence provides some challenge. To circumvent these difficulties, we show that the iterates are asymptotically equivalent to the trajectories of a particular ordinary differential inclusion [1] (a nonsmooth generalization of ordinary differential equations (ODEs)) related to problem (1), building off of the classical ODE method [39, 37, 6] (see section 3.2). By developing a number of analytic properties of the limiting differential inclusion using the weak convexity of f, we show that trajectories of the ODE must converge (section 3.3). A careful stability analysis then shows that limit properties of trajectories of the ODE are preserved under small perturbations, and viewing our algorithms as noisy discrete approximations to a solution of the ordinary differential inclusion gives our desired convergence (section 3.4).

Our results do not provide rates of convergence for the stochastic procedures, so to investigate the properties of the methods we propose, we perform a number of numerical simulations in section 4. We focus on a discrete version of problem (1) with the robust phase retrieval objective $f(x; a, b) = |(a^T x)^2 - b|$, which facilitates comparison with deterministic methods (5). Our experiments extend our theoretical predictions, showing the advantages of stochastic over deterministic procedures for some problems, and they also show that the stochastic prox-linear method may be preferable to stochastic subgradient methods because of robustness properties it enjoys (which our simulations verify, though our theory does not yet explain).

Related and subsequent work. The stochastic subgradient method has a substantial history. Early work due to Ermoliev and Norkin [25, 26, 27], Gupal [30], and Dorofeyev [17] identifies the basic assumptions sufficient for stochastic gradient methods to be convergent. Ruszczyński [48] provides a convergent gradient averaging-based optimization scheme for stochastic weakly convex problems. Our analytical approach is based on differential equations and inclusions, which have a long history in the study of stochastic optimization methods, where researchers have used a limiting differential equation or inclusion to exhibit convergence of stochastic approximation schemes [39, 1, 36, 6]; more recent work uses differential equations to model accelerated

gradient methods [50, 56]. Our approach gives similar convergence results to those for stochastic subgradient methods but allows us to study and prove convergence for a more general collection of model-based minimization strategies. Our results do not provide convergence rates, which is possible when the compositional structure (1) leaves the problem *convex* [53, 54]; the problems we consider are typically nonsmooth and nonconvex, so that these approaches do not apply.

Subsequent to the initial appearance of the current paper on the `arXiv` and inspired by our work,[1] Davis, Drusvyatskiy, and Grimmer have provided convergence rates for variants of stochastic subgradient, prox-linear, and related methods [14, 12, 13]. Here they show that the methods we develop in this paper satisfy nonasymptotic convergence guarantees. To make this precise, let $F_\lambda(x) = \inf_{y \in X}\{f(y) + \varphi(y) + \frac{\lambda}{2}\left\|y - x\right\|^2\}$ be the Moreau envelope of the objective (2), which is continuously differentiable and for which ∇F_λ being small is a proxy for near-stationarity of x (see [18, 12, 13]). Then they show that, with appropriate stepsizes, they can construct a (random) iterate $\widehat{x}_k$ such that $\mathbb{E}[\left\|\nabla F_\lambda(\widehat{x}_k)\right\|^2] = O(1/\sqrt{k})$. These convergence guarantees extend the with probability 1 convergence results we provide.

Notation and basic definitions. We collect here our (mostly standard) notation and basic definitions that we require. For $x, y \in \mathbb{R}$, we let $x \wedge y = \min\{x, y\}$. We let $\mathbb{B}$ denote the unit ℓ_2-ball in $\mathbb{R}^d$, where d is apparent from context and $\left\|\cdot\right\|$ denotes the operator ℓ_2-norm (the standard Euclidean norm on vectors). For a set $A \subset \mathbb{R}^d$ we let $\left\|A\right\| = \sup_{a \in A}\left\|a\right\|$. We say $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is λ-*weakly convex* (also known as lower-$\mathcal{C}^2$ or semiconvex [47, 5]) near x if there exists $\epsilon > 0$ such that for all $x_0 \in \mathbb{R}^d$,

$$y \mapsto f(y) + \frac{\lambda}{2}\left\|y - x_0\right\|_2^2, \quad y \in x + \epsilon\mathbb{B} \tag{6}$$

is convex (the vector x_0 is immaterial in (6), as holding at one x_0 is equivalent) [47, Chap. 10.G]. For a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, we let $\partial f(x)$ denote the Fréchet (or regular [47, Chap. 8.B]) subdifferential of f at the point x,

$$\partial f(x) := \left\{g \in \mathbb{R}^d : f(y) \geq f(x) + \langle g, y - x\rangle + o(\left\|y - x\right\|) \text{ as } y \to x\right\}.$$

The Fréchet subdifferential and standard (convex) subdifferential coincide for convex f [47, Ch. 8], and for weakly convex f, $\partial f(x)$ is nonempty for x in the relative interior of $\operatorname{dom} f$. The Clarke directional derivative of a function f at the point x in direction v is

$$f'(x; v) := \liminf_{t \downarrow 0, v' \to v} \frac{f(x + tv) - f(x)}{t},$$

and recall [47, Ex. 8.4] that $\partial f(x) = \{w \in \mathbb{R}^d : \langle v, w\rangle \leq f'(x; v) \text{ for all } v\}$.

We let $\mathcal{C}(A, B)$ denote the continuous functions from A to B. Given a sequence of functions $f_n : \mathbb{R}_+ \to \mathbb{R}^d$, we say that $f_n \to f$ in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ if $f_n \to f$ uniformly on all compact sets, that is, for all $T < \infty$ we have

$$\lim_{n \to \infty} \sup_{t \in [0,T]} \left\|f_n(t) - f(t)\right\| = 0.$$

This is equivalent to convergence in $\mathsf{d}(f, g) := \sum_{t=1}^\infty 2^{-t} \sup_{\tau \in [0,t]} \left\|f(\tau) - g(\tau)\right\| \wedge 1$, which shows the standard result that $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ is a Fréchet space. For a closed

[1] Based on personal communication with Damek Davis and Dmitriy Drusvyatskiy.

convex set X, we let $\mathbb{I}_X$ denote the $+\infty$-valued indicator for X, that is, $\mathbb{I}_X(x) = 0$ if $x \in X$ and $+\infty$ otherwise. The normal cone to X at x is

$$\mathcal{N}_X(x) := \{v \in \mathbb{R}^d : \langle v, y - x\rangle \le 0 \text{ for all } y \in X\}.$$

For closed convex C, $\pi_C(x) := \operatorname{argmin}_{y \in C} \|y - x\|$ denotes projection of x onto C.

2. Algorithms and main convergence result. In this section, we introduce the family of algorithms we study for problem (2). In analogy with the update (5), we first give a general form of our model-based approach, then exhibit three examples that fall into the broad scheme. We iterate

$$\begin{aligned} &\text{Draw } \; S_k \overset{\text{iid}}{\sim} P \\ &x_{k+1} := \operatorname*{argmin}_{y \in X} \left\{ f_{x_k}(y; S_k) + \varphi(y) + \frac{1}{2\alpha_k} \|y - x_k\|^2 \right\}. \end{aligned} \tag{7}$$

In the iteration (7), the function $f_{x_k}(\cdot; s)$ is an approximation, or model, of $f(\cdot; s)$ at the point x_k, and $\alpha_k > 0$ is a stepsize sequence.

For the model-based strategy (7) to be effective, we require that $f_x(\cdot; s)$ satisfy a few essential properties on its approximation quality.

C.(i) The function $y \mapsto f_x(y; s)$ is convex and subdifferentiable on its domain.

C.(ii) We have $f_x(x; s) = f(x; s)$.

C.(iii) At $y = x$ we have the containment

$$\partial_y f_x(y; s)|_{y=x} \subset \partial_x f(x; s).$$

In addition to conditions C.(i)–C.(iii), we require one additional technical condition on the models, which quantitatively guarantees they locally almost underestimate f.

C.(iv) There exists $\epsilon_0 > 0$ such that $0 < \epsilon \le \epsilon_0$ implies that for all $x_0 \in X$ there exists $\delta_\epsilon(x_0; s) \ge 0$ with

$$f(y; s) \ge f_x(y; s) - \frac{1}{2}\delta_\epsilon(x_0; s) \|y - x\|^2$$

for $x, y \in x_0 + \epsilon\mathbb{B}$, where $\mathbb{E}[\delta_\epsilon(x_0; S)] < \infty$.

2.1. Examples. We give four example algorithms for problems (1) and (2), each of which consists of a local model f_x satisfying conditions C.(i)–C.(iv). The conditions C.(i)–C.(iii) are immediate, while we defer verification of condition C.(iv) to after the statement of Theorem 1. The first example is the natural generalization of the classical subgradient method [25].

Example 1 (stochastic subgradient method). For this method, we let $\mathsf{g}(x; s) \in \partial f(x; s)$ be a (fixed) element of the Fréchet subdifferential of $f(x; s)$; in the case of the composite objective (1) this is $\mathsf{g}(x; s) \in \nabla c(x; s)\partial h(c(x; s); s)$. Then the model (7) for the stochastic (regularized and projected) subgradient method is

$$f_x(y; s) := f(x; s) + \langle \mathsf{g}(x; s), y - x\rangle .$$

The properties C.(i)–C.(iii) are immediate.

The stochastic prox-linear method applies to the structured family of convex composite problems (1), generalizing the deterministic prox-linear method [7, 19, 20].

Example 2 (stochastic prox-linear method). Here, we have $f(x;s) = h(c(x;s);s)$, and in analogy to the update (4) we linearize c without modifying h, defining

$$f_x(y;s) := h(c(x;s) + \nabla c(x;s)^T(y-x);s).$$

Again, conditions C.(i)–C.(iii) are immediate.

Lastly, we have stochastic proximal point methods for weakly convex functions.

Example 3 (stochastic proximal-point method). We assume that the instantaneous function $f(\cdot;s)$ is $\lambda(s)$-weakly convex over X. In this case, for the model in the update (7), we set $f_x(y;s) = f(y;s) + \frac{\lambda(s)}{2}\|y-x\|^2$.

Example 4 (guarded stochastic proximal-point method). We assume that for some $\epsilon > 0$ and all $x \in X$, the instantaneous function $f(\cdot;s)$ is $\lambda(s,x)$-weakly convex over $X \cap \{x + \epsilon\mathbb{B}\}$. In this case, for the model in the update (7), we set

$$f_x(y;s) = f(y;s) + \frac{\lambda(s,x)}{2}\|y-x\|^2 + \mathbb{I}_{x+\epsilon\mathbb{B}}(y), \tag{8}$$

which restricts the domain of the model function $f_x(\cdot;s)$ to a neighborhood of x so that the update (7) does not escape the region of convexity. Again, by inspection, this satisfies conditions C.(i)–C.(iii).

2.2. The main convergence result. The main theoretical result of this paper is to show that stochastic algorithms based on the update (7) converge almost surely to the stationary points of the objective function $F(x) = f(x) + \varphi(x)$ over X. To state our results formally, for $\epsilon > 0$ we we define the function $M_\epsilon : X \times \mathcal{S} \to \mathbb{R}_+$ by

$$M_\epsilon(x;s) := \sup_{y \in X, \|y-x\| \le \epsilon} \sup_{g \in \partial f(y;s)} \|g\|.$$

We then make the following local Lipschitzian and convexity assumptions on $f(\cdot;s)$.

Assumption A. There exists $\epsilon_0 > 0$ such that $0 < \epsilon \le \epsilon_0$ implies that

$$\mathbb{E}[M_\epsilon(x;S)^2] < \infty \quad \text{for all } x \in X.$$

Assumption B. There exists $\epsilon_0 > 0$ such that $0 < \epsilon \le \epsilon_0$ implies that for all $x \in X$, there exists $\lambda(s,x) \ge 0$ such that

$$y \mapsto f(y;s) + \frac{\lambda(s,x)}{2}\|y-x_0\|^2$$

is convex on the set $x + \epsilon\mathbb{B}$ for any x_0, and $\mathbb{E}[\lambda(S,x)] < \infty$.

As we shall see in Lemma 6 later, Assumptions A and B are sufficient to guarantee that $\partial f(x)$ exists, is nonempty for all $x \in X$, and is outer semicontinuous. In addition, it is immediate that for any $\lambda \ge \mathbb{E}[\lambda(S,x)]$, the function f is λ-weakly convex (6) on the ϵ-ball around x.

With the assumptions in place, we can now proceed to a (mildly) simplified version of our main result in this paper. Let $X^\star$ denote the set of stationary points for the objective function $F(x) = f(x) + \varphi(x)$ over X. Lemma 6 to come implies that $\partial F(x) = \partial f(x) + \partial\varphi(x)$ for all $x \in X$, so we can represent $X^\star$ as

$$X^\star := \{x \in X \mid \exists g \in \partial f(x) + \partial\varphi(x) \text{ with } \langle g, y-x\rangle \ge 0 \text{ for all } y \in X\}. \tag{9}$$

Equivalently, $\partial f(x)+\partial\varphi(x)\cap -\mathcal{N}_X(x) \neq \emptyset$, or $0 \in \partial f(x)+\partial\varphi(x)+\mathcal{N}_X(x)$. Important for us is the *image* of the set of stationary points, that is,

$$F(X^\star) := \{f(x) + \varphi(x) \mid x \in X^\star\}.$$

With these definitions, we have the following convergence result, which is a simplification of our main convergence result, Theorem 5, which we present in section 3.4.

THEOREM 1. *Let Assumptions* A *and* B *hold and assume X is compact. Let x_k be generated by any model-based update satisfying conditions* C.(i)–C.(iv) *with stepsizes $\alpha_k > 0$ satisfying $\sum_k \alpha_k = \infty$ and $\sum_k \alpha_k^2 < \infty$. Then with probability* 1*,*

$$\left[\liminf_k F(x_k), \limsup_k F(x_k)\right] \subset F(X^\star). \tag{10}$$

We provide a few remarks on the theorem, as well as elucidating Examples 1–4 in this context. The limiting inclusion (10) is familiar from the classical literature on stochastic subgradient methods [17, 26], though in our case, it applies to the broader family of model-based updates (7), including Examples 1–4.

To see that the theorem indeed applies to each of these examples, we must verify Condition C.(iv). For Examples 1, 3, and 4, this is immediate by taking the lower approximation function $\delta_\epsilon(x; s) = \lambda(s, x)$ from Assumption B, yielding the following.

Observation 1. Let Assumption B hold. Then Condition C.(iv) holds for each of Examples 1, 3, and 4.

We also provide conditions on the composite optimization problem (1), that is, when $f(x; s) = h(c(x; s); s)$, sufficient for Assumptions A–B and Condition C.(iv) to hold. Standard results [21] show that $\partial f(x; s) = \nabla c(x; s)\partial h(c(x; s); s)$, so Assumption A holds if $\sup_{\|y-x\|\le\epsilon} \|\nabla c(x; s)\partial h(c(x; s); s)\|$ is integrable (with respect to s). For Assumption B, we assume that there exists $\epsilon_0 > 0$ such that if $0 < \epsilon \le \epsilon_0$, there exist functions $\gamma_\epsilon : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}_+ \cup \{+\infty\}$ and $\beta_\epsilon : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}_+ \cup \{+\infty\}$ such that $c(\cdot; s)$ has $\beta_\epsilon(x; s)$-Lipschitz gradients in an ϵ neighborhood of x; that is,

$$\|\nabla c(y; s) - \nabla c(y'; s)\| \le \beta_\epsilon(x, s) \|y - y'\| \quad \text{for } \|y - x\|, \|y' - x\| \le \epsilon,$$

and that $h(\cdot; s)$ is $\gamma_\epsilon(x; s)$-Lipschitz continuous on the compact convex neighborhood

$$\operatorname{Conv}\left\{c(y; s) + \nabla c(y; s)^T(z - y) + v \mid y, z \in x + \epsilon\mathbb{B}, \|v\| \le \frac{\beta_\epsilon(x, s)}{2}\|y - z\|^2\right\}.$$

We then have the following claim; see Appendix A.1 for a proof.

CLAIM 1. *If $\mathbb{E}[\gamma_\epsilon(x; S)\beta_\epsilon(x; S)] < \infty$ for all $x \in X$, then Assumption* B *holds with $\lambda(s, x) = \gamma_\epsilon(x; s)\beta_\epsilon(s)$, and Condition* C.(iv) *holds with $\delta_\epsilon(x; s) = \gamma_\epsilon(x; s)\beta_\epsilon(x; s)$.*

Theorem 1 does not guarantee convergence of the iterates, though it does guarantee cluster points of $\{x_k\}$ have limiting *values* in the image of the stationary set. A slightly stronger technical assumption, which rules out pathological functions such as Whitney's construction [55], is the following assumption, which is related to Sard's results that the measure of critical values of $\mathcal{C}^d$-smooth $f : \mathbb{R}^d \to \mathbb{R}$ is zero.

Assumption C. The set $(F(X^\star))^c$ is dense in $\mathbb{R}$.

If f is convex then $(f + \varphi)(X^\star)$ is a singleton. Moreover, if the set of stationary points $X^\star$ consists of a (finite or countable) collection of sets $X_1^\star, X_2^\star, \ldots$ such that $f + \varphi$ is constant on each $X_i^\star$, then $F(X^\star)$ is at most countable and Assumption C

holds. In subsequent work to the first version of this paper, Davis et al. [15] give sufficient conditions for Assumption C to hold (see also [34, 35, 4]). We have the following.

COROLLARY 1. *In addition to the conditions of Theorem* 1*, let Assumption* C *hold. Then* $f(x_k) + \varphi(x_k)$ *converges, and all cluster points of the sequence* $\{x_k\}$ *belong to* $X^\star$.

3. Convergence analysis of the algorithm. In this section, we present the arguments necessary to prove Theorem 1 and its extensions, beginning with a heuristic explanation. By inspection and a strong faith in the limiting behavior of random iterations, we expect that the update scheme (7), as the stepsizes $\alpha_k \to 0$, are asymptotically approximately equivalent to iterations of the form

$$\frac{x_{k+1} - x_k}{\alpha_k} \approx - \left[\mathsf{g}(x_k) + v_k + w_k\right], \; \mathsf{g}(x_k) \in \partial f(x_k), \; v_k \in \partial \varphi(x_{k+1}), \; w_k \in \mathcal{N}_X(x_{k+1}),$$

and the correction w_k enforces $x_{k+1} \in X$. As $k \to \infty$ and $\alpha_k \to 0$, we may (again, deferring rigor) treat $\lim_k \frac{1}{\alpha_k}(x_{k+1} - x_k)$ as a continuous time process, suggesting that update schemes of the form (7) are asymptotically equivalent to a continuous time process $t \mapsto x(t) \in \mathbb{R}^d$ that satisfies the differential inclusion (a set-valued generalization of an ODE)

$$\dot{x} \in -\partial f(x) - \partial \varphi(x) - \mathcal{N}_X(x) = -\int \partial f(x; s) dP(s) - \partial \varphi(x) - \mathcal{N}_X(x). \tag{11}$$

We develop a general convergence result showing that this limiting equivalence is indeed the case and that the second equality of expression (11) holds. As part of this, we explore in the coming sections how the weak convexity structure of $f(\cdot; s)$ guarantees that the differential inclusion (11) is well behaved. We begin in section 3.1 with preliminaries on set-valued analysis and differential inclusions we require, which build on standard convergence results [1, 36]. Once we have presented these preliminary results, we show how the stochastic iterations (7) eventually approximate solution paths to differential inclusions (section 3.2), which builds off of a number of stochastic approximation results and the so-called "ODE method" Ljung develops [39], (see also [37, 2, 6]). We develop the analytic properties of the composite objective, which yields the uniqueness of trajectories solving (11) as well as a particular Lyapunov convergence inequality (section 3.3). Finally, we develop stability results on the differential inclusion (11), which allows us to prove convergence as in Theorem 1 (section 3.4).

3.1. Preliminaries: Differential inclusions and set-valued analysis. We now review a few results in set-valued analysis and differential inclusions [1, 36]. Our notation and definitions follow closely the references of Rockafellar and Wets [47] and Aubin and Cellina [1], and we cite a few results from the book of Kunze [36].

Given a sequence of sets $A_n \subset \mathbb{R}^d$, the limit supremum of the sets consists of limit points of subsequences $y_{n_k} \in A_{n_k}$, that is,

$$\limsup_n A_n := \left\{y : \exists \; y_{n_k} \in A_{n_k} \text{ s.t. } y_{n_k} \to y \text{ as } k \to \infty\right\}.$$

We let $G : X \rightrightarrows \mathbb{R}^d$ denote a set-valued mapping G from X to $\mathbb{R}^d$, and we define $\operatorname{dom} G := \{x : G(x) \neq \emptyset\}$. Then G is *outer semicontinuous (o.s.c.)* if for any sequence $x_n \to x \in \operatorname{dom} G$, we have $\limsup_n G(x_n) \subset G(x)$. One says that G is ϵ-δ

o.s.c. [1, Def. 1.1.5] if for all x and $\epsilon > 0$, there exists $\delta > 0$ such that $G(x + \delta\mathbb{B}) \subset G(x) + \epsilon\mathbb{B}$. These notions coincide when $G(x)$ is bounded. Two standard examples of o.s.c. mappings follow.

LEMMA 1 (see Hiriart-Urruty and Lemaréchal [33, Thm. VI.6.2.4]). *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be convex. Then the subgradient mapping $\partial f : \operatorname{int} \operatorname{dom} f \rightrightarrows \mathbb{R}^d$ is o.s.c.*

LEMMA 2 (see Rockafellar and Wets [47, Prop. 6.6]). *Let X be a closed convex set. Then the normal cone mapping $\mathcal{N}_X : X \rightrightarrows \mathbb{R}^d$ is o.s.c. on X.*

The *differential inclusion* associated with G beginning from the point x_0, denoted

$$\dot{x} \in G(x), \quad x(0) = x_0 \tag{12}$$

has a solution if there exists an absolutely continuous function $x : \mathbb{R}_+ \to \mathbb{R}^d$ satisfying $\frac{d}{dt}x(t) = \dot{x}(t) \in G(x(t))$ for all $t \geq 0$. For $G : \mathcal{T} \rightrightarrows \mathbb{R}^d$ and a measure μ on $\mathcal{T}$,

$$\int_{\mathcal{T}} G d\mu = \int_{\mathcal{T}} G(t) d\mu(t) := \left\{ \int g(t) d\mu(t) \mid g(t) \in G(t) \text{ for } t \in \mathcal{T}, \ g \text{ measurable} \right\}.$$

With these definitions, the following results (with minor extension) on the existence and uniqueness of solutions to differential inclusions are standard.

LEMMA 3 (see Aubin and Cellina [1, Thm. 2.1.4]). *Let $G : X \rightrightarrows \mathbb{R}^d$ be o.s.c. and compact-valued, and $x_0 \in X$. Assume there is $K < \infty$ such that $\operatorname{dist}(0, G(x)) \leq K$ for all x. Then there exists an absolutely continuous function $x : \mathbb{R}_+ \to \mathbb{R}^d$ such that $\dot{x}(t) \in G(x(t))$ and $x(t) \in x_0 + \int_0^t G(x(\tau))d\tau$ for all $t \in \mathbb{R}_+$.*

LEMMA 4 (see Kunze [36, Thm. 2.2.2]). *Let the conditions of Lemma 3 hold and assume there exists $c < \infty$ such that*

$$\langle x_1 - x_2, g_1 - g_2 \rangle \leq c \left\| x_1 - x_2 \right\|^2 \quad \text{for } g_i \in G(x_i) \text{ and all } x_i \in \operatorname{dom} G.$$

Then the solution to the differential inclusion (12) is unique.

We recall basic Lyapunov theory for differential inclusions. Let $V : X \to \mathbb{R}_+$ be a nonnegative function and $W : X \times \mathbb{R}^d \to \mathbb{R}_+$ be continuous with $v \mapsto W(x, v)$ convex in v for all x. A trajectory $\dot{x} \in G(x)$ is *monotone* for the pair V, W if

$$V(x(T)) - V(x(0)) + \int_0^T W(x(t), \dot{x}(t))dt \leq 0 \ \text{ for } T \geq 0.$$

The next lemma gives sufficient conditions for the existence of monotone trajectories.

LEMMA 5 (see Aubin and Cellina [1, Thm. 6.3.1]). *Let $G : X \rightrightarrows \mathbb{R}^d$ be o.s.c. and compact-convex valued. Assume that for each x there exists $v \in G(x)$ such that $V'(x; v) + W(x; v) \leq 0$. Then there exists a trajectory of the differential inclusion $\dot{x} \in G(x)$ such that*

$$V(x(T)) - V(x(0)) + \int_0^T W(x(t), \dot{x}(t))dt \leq 0.$$

Finally, we present a lemma on the subgradients of f using our set-valued integral definitions. The proof is somewhat technical and not the main focus of this paper, so we defer it to Appendix A.2.

LEMMA 6. *Let $f(\cdot; s)$ satisfy Assumptions* A *and* B. *Then*

$$\partial f(x) = \mathbb{E}_P[\partial f(x; S)],$$

and $\partial f(\cdot; s) : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ and $\partial f(\cdot) : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ are closed compact convex-valued and o.s.c.

Lemma 6 shows that $\partial f(x; s)$ is compact-valued and o.s.c., and we thus define the shorthand notation for the subgradients of $f + \varphi$ as

$$G(x; s) := \partial f(x; s) + \partial\varphi(x) \text{ and } G(x) := \int_{\mathcal{S}} \partial f(x; s) dP(s) + \partial\varphi(x), \tag{13}$$

both of which are o.s.c. in x and compact-convex valued because φ is convex.

3.2. Functional convergence of the iteration path. With our preliminaries in place, we now establish a general functional convergence theorem (Theorem 2) that applies to stochastic approximation-like algorithms that asymptotically approximate differential inclusions. By showing the generic algorithm (7) has the form our theorem requires, we conclude that each of Examples 1–4 converge to the appropriate differential inclusion (section 3.2.2).

3.2.1. A general functional convergence theorem. Let $\{g_k\}_{k\in\mathbb{N}}$ be a collection of set-valued mappings $g_k : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$, $\{\alpha_k\}_{k\in\mathbb{N}}$ be a sequence of positive stepsizes, $\{\xi_k\}_{k=1}^\infty$ be an arbitrary $\mathbb{R}^d$-valued sequence (the noise sequence). Consider the following iteration, which begins from the initial value $x_0 \in \mathbb{R}^d$:

$$x_{k+1} = x_k + \alpha_k[y_k + \xi_{k+1}], \quad \text{where } y_k \in g_k(x_k) \text{ for } k \geq 0. \tag{14}$$

For notational convenience, define the "times" $t_m = \sum_{k=1}^m \alpha_k$ as the partial stepsize sums, and let $x(\cdot)$ be the linear interpolation of the iterates x_k:

$$x(t) := x_k + \frac{t - t_k}{t_{k+1} - t_k}(x_{k+1} - x_k) \text{ and } y(t) = y_k \text{ for } t \in [t_k, t_{k+1}). \tag{15}$$

This path satisfies $\dot{x}(t) = y(t)$ for almost all t and is absolutely continuous on compact. For $t \in \mathbb{R}_+$, define the time-shifted process $x^t(\cdot) = x(t + \cdot)$. We have the following convergence theorem for the interpolation (15) of the iterative process (14), where we recall that we metrize $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ with $\mathsf{d}(f, g) = \sum_{t=1}^\infty 2^{-t} \sup_{\tau\in[0,t]} \|f(\tau) - g(\tau)\| \wedge 1$.

THEOREM 2. *Let the following conditions hold.*

(i) *The iterates are bounded, i.e.,* $\sup_k \|x_k\| < \infty$ *and* $\sup_k \|y_k\| < \infty$.
(ii) *The stepsizes satisfy* $\sum_{k=1}^\infty \alpha_k = \infty$ *and* $\sum_{k=1}^\infty \alpha_k^2 < \infty$.
(iii) *The weighted noise sequence converges:* $\lim_n \sum_{k=1}^n \alpha_k \xi_k = v$ *for some* $v \in \mathbb{R}^d$.
(iv) *There exists a closed-valued* $H : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ *such that for all* $\{z_k\} \subset \mathbb{R}^d$ *satisfying* $\lim_k z_k = z$ *and all increasing subsequences* $\{n_k\}_{k\in\mathbb{N}} \subset \mathbb{N}$, *we have*

$$\lim_{n\to\infty} \operatorname{dist}\left(\frac{1}{n}\sum_{k=1}^n g_{n_k}(z_k), H(z)\right) = 0.$$

Then for any sequence $\{\tau_k\}_{k=1}^\infty \subset \mathbb{R}_+$, *the sequence of functions* $\{x^{\tau_k}(\cdot)\}$ *is relatively compact in* $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$. *If* $\tau_k \to \infty$, *all limit points of* $\{x^{\tau_k}(\cdot)\}$ *are in* $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ *and there exists* $y : \mathbb{R}_+ \to \mathbb{R}^d$ *satisfying* $y(t) \in H(x(t))$ *for all* $t \in \mathbb{R}_+$, *where*

$$\bar{x}(t) = \bar{x}(0) + \int_0^t y(\tau) d\tau \quad \textit{for all } t \in \mathbb{R}_+.$$

The theorem is a generalization of Theorem 5.2 of Borkar [6], where the set-valued mappings g_k are identical for all k; our proof techniques are similar. For completeness, we provide a proof on the `arXiv` [22].

3.2.2. Differential inclusion for stochastic model-based methods. With Theorem 2 in place, we can now show how the update scheme (7) is representable by the general stochastic approximation (14). To do so, we must verify that any method satisfying Conditions C.(i)–C.(iv) satisfies the four conditions of Theorem 2. With this in mind, we introduce a bit of new notation before proceeding. In analogy to the gradient mapping from convex [41] and composite optimization [21], we define a stochastic gradient mapping G and consider its limits. For fixed x we define

(16)
$$x_\alpha^+(s) := \operatorname*{argmin}_{y\in X}\left\{f_x(y;s) + \varphi(y) + \frac{1}{2\alpha}\|y-x\|^2\right\} \quad \text{and} \quad \mathsf{G}_\alpha(x;s) := \frac{1}{\alpha}(x - x_\alpha^+(s)).$$

For any model $f_x(\cdot;s)$ we consider, the update is well behaved: it is measurable in s [45, Lem. 1], and it is bounded, as the next lemma shows.

LEMMA 7. *The update* (16) *guarantees that* $\|\mathsf{G}_\alpha(x;s)\| \le \|G(x;s)\|$, *where* $G(x;s)$ *is the subgradient* (13).

Proof. For shorthand, write $x^+ = x_\alpha^+(s)$ and let $g \in \partial f_x(x;s) \subset \partial f(x;s)$. By the definition of the optimality conditions for x^+, there exists a vector g^+ that $g^+ \in \partial f_x(x^+;s)$ and another vector $v^+ \in \partial\varphi(x^+)$ such that

$$\left\langle g^+ + \frac{1}{\alpha}(x^+ - x) + v^+, y - x^+ \right\rangle \ge 0 \quad \text{for all } y \in X.$$

Rearranging, we substitute $y = x$ to obtain

$$\left\langle g^+, x^+ - x\right\rangle + \frac{1}{\alpha}\left\|x - x^+\right\|^2 + \left\langle v^+, x^+ - x\right\rangle \le 0.$$

The subgradient mapping is monotone for $f_x(\cdot;s)$ and φ, so $\langle g^+, x - x^+\rangle \ge \langle g, x - x^+\rangle$ and $\langle v^+, x - x^+\rangle \ge \langle \partial\varphi(x), x - x^+\rangle$. Thus

$$\left\langle g, x^+ - x\right\rangle + \frac{1}{\alpha}\left\|x - x^+\right\|^2 + \left\langle v, x^+ - x\right\rangle \le 0$$

for all $v \in \partial\varphi(x)$. Cauchy–Schwarz implies $\|g + v\|\,\|x^+ - x\| \ge \frac{1}{\alpha}\|x - x^+\|^2$, which implies our desired result. ∎

To define the population counterpart of the gradient mapping G_α, we require a result showing that the gradient mapping is locally bounded and integrable. To that end, for $x \in X$ and $\epsilon > 0$, define the Lipschitz constants

$$L_\epsilon(x;s) := \sup_{x'\in X, \|x'-x\|\le\epsilon} \|G(x';s)\| \quad \text{and} \quad L_\epsilon(x) := \mathbb{E}_P[L_\epsilon(x;S)^2]^{\frac{1}{2}}.$$

The following lemma shows these are not pathological (see Appendix A.3 for a proof).

LEMMA 8. *Let Assumptions* A *and* B *hold. Then* $x \mapsto L_\epsilon(x;s)$ *and* $x \mapsto L_\epsilon(x)$ *are upper semicontinuous on* X *and* $L_\epsilon(x) < \infty$ *for all* $x \in X$.

As a consequence of this lemma and Lemma 7, $\mathsf{G}_\alpha(x; S)$ is locally bounded by $L_\epsilon(x; s)$ and we may define the mean subgradient mapping

$$\overline{\mathsf{G}}_\alpha(x) := \mathbb{E}_P\left[\mathsf{G}_\alpha(x; S)\right] = \int_{\mathcal{S}} \mathsf{G}_\alpha(x; s) dP(s).$$

Moreover, any update of the form (7) (e.g., Examples 1–4) has representation

$$x_{k+1} = x_k - \alpha_k \mathsf{G}_{\alpha_k}(x_k; S_k) = x_k - \alpha_k \overline{\mathsf{G}}_{\alpha_k}(x_k) - \alpha_k \xi_{\alpha_k}(x_k; S_k), \tag{17}$$

where the noise vector $\xi_\alpha(x; s) := \mathsf{G}_\alpha(x; s) - \overline{\mathsf{G}}_\alpha(x)$. Defining the filtration of σ-fields $\mathcal{F}_k := \sigma(x_0, S_1, \ldots, S_{k-1})$, we have $x_k \in \mathcal{F}_k$ and that ξ is a square-integrable martingale difference sequence adapted to $\mathcal{F}_k$. Indeed, for α and $\epsilon > 0$ we have

$$\|\mathsf{G}_\alpha(x; s)\| \le L_\epsilon(x; s) \quad \text{and} \quad \left\|\overline{\mathsf{G}}_\alpha(x)\right\| \le L_\epsilon(x)$$

by Lemma 7 and the definition of the Lipschitz constant, and for any x and $\alpha > 0$,

$$\mathbb{E}_P\left[\left\|\xi_\alpha(x; S)\right\|^2\right] \le \mathbb{E}_P\left[\left\|\mathsf{G}_\alpha(x; S)\right\|^2\right] \le \mathbb{E}\left[L_\epsilon^2(x; S)\right] = L_\epsilon(x)^2, \tag{18}$$

because $\mathbb{E}[\mathsf{G}_\alpha] = \overline{\mathsf{G}}_\alpha$. In the context of our iterative procedures, for any $\alpha > 0$,

$$\mathbb{E}[\xi_\alpha(x_k; S_k) \mid \mathcal{F}_k] = 0 \quad \text{and} \quad \mathbb{E}[\left\|\xi_\alpha(x_k; S_k)\right\|^2 \mid \mathcal{F}_k] \le L_\epsilon(x_k)^2.$$

The (random) progress of each iterate of the algorithm G is now the sum of a mean progress $\overline{\mathsf{G}}$ and a random noise perturbation ξ with (conditional) mean 0 and bounded second moments. The update form (17) shows that all of our examples—stochastic proximal point, stochastic prox-linear, and the stochastic gradient method—have the form (14) necessary for application of Theorem 2.

Functional convergence for the stochastic updates. Now that we have the representation (17), it remains to verify that the mean gradient mapping $\overline{\mathsf{G}}$ and errors ξ satisfy the conditions necessary for application of Theorem 2. That is, we verify (i) bounded iterates, (ii) nonsummable but square-summable stepsizes, (iii) convergence of the weighted error sequence, and (iv) the distance condition in the theorem. Condition (ii) is trivial. To address condition (i), we temporarily make the following assumption, noting that the compactness of X is sufficient for it to hold (we give other sufficient conditions in section 3.4, showing that it is not too onerous).

Assumption D. With probability 1, the iterates (7) are bounded,

$$\sup_k \|x_k\| < \infty.$$

A number of conditions, such as almost supermartingale convergence theorems [44], are sufficient to guarantee Assumption D. Whenever Assumption D holds, we have

$$\sup_k \sup_{\alpha > 0} \left\|\overline{\mathsf{G}}_\alpha(x_k)\right\| \le \sup_k L_\epsilon(x_k) < \infty,$$

by Lemmas 7 and 8, because the supremum of an upper semicontinuous function on a compact set is finite. That is, condition (i) of Theorem 2 on the boundedness of x_k and y_k holds.

The error sequences ξ_{α_k} are also well behaved for the model-based updates (7). That is, condition (iii) of Theorem 2 is satisfied.

LEMMA 9. *Let Assumptions* A, B, *and* D *hold. Then with probability 1, the limit* $\lim_{n\to\infty} \sum_{k=1}^n \alpha_k \xi_{\alpha_k}(x_k; S_k)$ *exists and is finite.*

Proof. Ignoring probability zero events, by Assumption D there is a random variable B, which is finite with probability 1, such that $\|x_k\| \le B$ for all $k \in \mathbb{N}$. As $L_\epsilon(\cdot)$ is upper semicontinuous (Lemma 8), we know that $\sup\{L_\epsilon(x) \mid \|x\| \le B, x \in X\} < \infty$. Hence, using inequality (18), we have

$$\sum_{k=1}^\infty \mathbb{E}\left[\alpha_k^2 \left\|\xi_{\alpha_k}(x_k; S_k)\right\|^2 \mid \mathcal{F}_k\right] \le \sum_{k=1}^\infty \alpha_k^2 \sup_{\|x\|\le B, x\in X} L_\epsilon(x)^2 < \infty.$$

Standard convergence results for ℓ_2-summable martingale difference sequences [16, Thm. 5.3.33] immediately give the result. ∎

Finally, we verify the fourth technical condition Theorem 2 requires by constructing an appropriate closed-valued mapping $H : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ for any update scheme of the form (7). Recall the definition (13) of the o.s.c. mapping $G(x) = \mathbb{E}_P[\partial f(x; S)] + \partial\varphi(x)$. We then have the following limiting inclusion, which is the key result allowing our limit statements.

LEMMA 10. *Let the sequence* $x_k \in X$ *satisfy* $x_k \to x \in X$ *and Assumptions* A *and* B *hold. Let* $\{i_k\} \subset \mathbb{N}$ *be an increasing sequence. Then, for updates* (7) *satisfying Conditions* C.(i)–C.(iv)*,*

$$\lim_{n\to\infty} \operatorname{dist}\left(\frac{1}{n}\sum_{k=1}^n \overline{\mathsf{G}}_{\alpha_{i_k}}(x_k), G(x) + \mathcal{N}_X(x)\right) = 0.$$

Proof. We begin with two intermediate lemmas on the continuity properties of the models f_x. Both lemmas assume the conditions of Lemma 10.

LEMMA 11. *There exists* $M'_\epsilon(x; s)$ *such that* $y \mapsto f_x(y; s)$ *is* $M'_\epsilon(x; s)$*-Lipschitz for* $y \in x + (\epsilon/2)\mathbb{B}$*, and* $\mathbb{E}[M'_\epsilon(x; S)] < \infty$*.*

Proof. Let $\epsilon > 0$, and let $g = \mathsf{g}(x; s) \in \partial f_x(x; s) \subset \partial f(x; s)$. We have that

$$f_x(y; s) \ge f_x(x; s) + \langle g, y - x\rangle \ge f(x; s) - M_\epsilon(x; s)\|y - x\|$$

by the local Lipschitz condition A on f. Condition C.(iv) and the Lipschitzian assumptions on f also guarantee that for $y \in x + \epsilon\mathbb{B}$,

$$f_x(y; s) \le f(y; s) + \frac{1}{2}\delta_\epsilon(x; s)\|y - x\|^2 \le f(x; s) + [M_\epsilon(x; s) + \delta_\epsilon(x; s)\|x - y\|]\|x - y\|.$$

These two boundedness conditions and convexity of the model f_x imply [33, Lem. IV.3.1.1] that $y \mapsto f_x(y; s)$ is $2M_\epsilon(x; s) + \delta_\epsilon(x; s)\epsilon$-Lipschitz for $y \in x + (\epsilon/2)\mathbb{B}$. ∎

LEMMA 12. *Let* $x_k, y_k \in X$ *satisfy* $x_k \to x, y_k \to x$*, and let* $g_k \in \partial f_{x_k}(y_k; s)$*. Then there exists an integrable function* $M(\cdot)$ *such that for large* k*,* $\operatorname{dist}(g_k, \partial f(x; s)) \le M(s)$ *for all* s*, and* $\operatorname{dist}(g_k, \partial f(x; s)) \to 0$*.*

Proof. By Lemma 11, we know that there exists an integrable M such that $\|g_k\| \le M(s)$ for all large enough k. This gives the first claim of the lemma, as $f(\cdot; s)$ is locally Lipschitz (Assumption A). Let g_∞ be any limit point of the sequence g_k; by moving to a subsequence if necessary, we assume without loss of generality that $g_k \to g_\infty \in \mathbb{R}^d$. Now let $y \in x + \epsilon\mathbb{B}$. Then for large k we have

$$f(y;s) \overset{(i)}{\geq} f_{x_k}(y;s) - \frac{\delta_\epsilon(x;s)}{2}\left\|y - x_k\right\|^2 \geq f_{x_k}(x_k;s) + \langle g_k, y - x_k\rangle - \frac{\delta_\epsilon(x;s)}{2}\left\|y - x_k\right\|^2$$
$$\to f(x;s) + \langle g_\infty, y - x\rangle - \frac{\delta_\epsilon(x;s)}{2}\left\|y - x\right\|^2,$$

where inequality (i) is a consequence of Condition C.(iv). By definition of the Fréchet subdifferential, we have $g_\infty \in \partial f(x;s)$ as desired. ∎

Now we return to the proof of Lemma 10. Let $x_k^+(s)$ be shorthand for the result of the update (7) when applied with the stepsize $\alpha = \alpha_{i_k}$. For any $\epsilon > 0$, Lemma 7 shows that $\left\|x_k^+(s) - x_k\right\| \leq \alpha_{i_k} L_\epsilon(x;s)$. By the (convex) optimality conditions for $x_k^+(s)$, there exists a vector $\mathsf{g}^+(x_k;s)$ such that

$$\mathsf{g}^+(x_k;s) \in \partial f_{x_k}(x_k^+(s);s)$$

and

$$\mathsf{G}_{\alpha_{i_k}}(x_k;s) \in \mathsf{g}^+(x_k;s) + \partial\varphi(x_k^+(s)) + \mathcal{N}_X(x_k^+(s)).$$

Let $v_k^+(s) \in \partial\varphi(x_k^+(s))$ and $w_k^+(s) \in \mathcal{N}_X(x_k^+(s))$ be the vectors such that

$$\mathsf{G}_{\alpha_{i_k}}(x_k;s) = \mathsf{g}^+(x_k;s) + v_k^+(s) + w_k^+(s).$$

The three set-valued mappings $x \mapsto \partial f(x;s)$, $x \mapsto \partial\varphi(x)$, and $x \mapsto \mathcal{N}_X(x)$ are o.s.c. (see Lemmas 1, 2, and 6). Since $x_k^+(s) \to x$ tends to x as $k \to \infty$ (as $x_k \to x$), this outer semicontinuity and Lemma 12 thus imply

(19)
$$\operatorname{dist}\big(\mathsf{g}^+(x_k;s), \partial f(x;s)\big) \to 0, \;\; \operatorname{dist}\big(v_k^+(s), \partial\varphi(x)\big) \to 0, \;\; \operatorname{dist}\big(w_k^+(s), \mathcal{N}_X(x)\big) \to 0$$

as $k \to \infty$. Because $x_k \to x$ and the Lipschitz constants $L_\epsilon(\cdot;s)$ are upper semicontinuous, (19) and Lemma 7 also imply that

$$\limsup_k \left\|\mathsf{g}^+(x_k;s) + v_k^+(s)\right\| \leq L_\epsilon(x;s) \;\; \text{and} \;\; \limsup_k \left\|\mathsf{G}_{\alpha_{i_k}}(x_k;s)\right\| \leq L_\epsilon(x;s).$$

By the triangle inequality, we thus obtain $\limsup_k \left\|w_k^+(s)\right\| \leq 2L_\epsilon(x;s)$, and hence,

$$\operatorname{dist}\left(w_k^+(s), \mathcal{N}_X(x) \cap 2L_\epsilon(x;s)\cdot\mathbb{B}\right) \to 0.$$

That $L_\epsilon(x) = \mathbb{E}[L_\epsilon(x;S)^2]^{\frac{1}{2}}$ yields $\mathcal{N}_X(x) \cap 2L_\epsilon(x)\mathbb{B} \supset \int(\mathcal{N}_X(x) \cap 2L_\epsilon(x;s)\cdot\mathbb{B})dP(s)$, and the definition of the set-valued integral and convexity of $\operatorname{dist}(\cdot,\cdot)$ imply that

$$\operatorname{dist}\left(\frac{1}{n}\sum_{k=1}^n \overline{\mathsf{G}}_{\alpha_{i_k}}(x_k), G(x) + \mathcal{N}_X(x) \cap 2L_\epsilon(x)\cdot\mathbb{B}\right)$$
$$(20) \quad \leq \frac{1}{n}\sum_{k=1}^n \int \operatorname{dist}\Big(\mathsf{G}_{\alpha_{i_k}}(x_k;s), \partial f(x;s) + \partial\varphi(x) + \mathcal{N}_X(x) \cap 2L_\epsilon(x;s)\cdot\mathbb{B}\Big)\, dP(s).$$

We now bound the preceding integral. By the definition of Minkowski addition and the triangle inequality, we have the pointwise convergence

$$\operatorname{dist}\Big(\mathsf{G}_{\alpha_{i_k}}(x_k;s), \partial f(x;s) + \partial\varphi(x) + \mathcal{N}_X(x) \cap 2L_\epsilon(x;s)\mathbb{B}\Big)$$
$$\leq \operatorname{dist}(\mathsf{g}(x_k;s), \partial f(x;s)) + \operatorname{dist}\big(v_k^+(s), \partial\varphi(x)\big) + \operatorname{dist}\big(w_k^+(s), \mathcal{N}_X(x) \cap 2L_\epsilon(x;s)\mathbb{B}\big) \to 0$$

as $k \to \infty$ by the earlier o.s.c. convergence guarantee (19). For suitably large k, the first term in the preceding sum is bounded by an integrable function $M'_\epsilon(x;s)$ by Lemma 12 and the latter two are bounded by $2L_\epsilon(x;s)$, which is square integrable by Lemma 8. Lebesgue's dominated convergence theorem thus implies that the individual summands in expression (20) converge to zero, and the analytic fact that the Cesáro mean $\frac{1}{n}\sum_{k=1}^n a_k \to 0$ if $a_k \to 0$ gives the result. □

With this lemma, we may now show the functional convergence of our stochastic model-based update schemes (7). We have verified that each of the conditions (i)–(iv) of Theorem 2 hold with the mapping $H(x) = -\mathcal{N}_X(x) - G(x)$. Indeed, H is closed-valued and o.s.c. as $G(\cdot)$ is convex compact o.s.c. and $\mathcal{N}_X(\cdot)$ is closed and o.s.c. Thus, with slight abuse of notation, let $x(\cdot)$ be the linear interpolation (15) of the iterates x_k for either the stochastic prox-linear algorithm or the stochastic subgradient algorithm, where we recall that $x^t(\cdot) = x(t + \cdot)$. We have the following.

THEOREM 3. *Let Assumptions* A, B, *and* D *hold. With probability one over the random sequence* $S_i \overset{\text{iid}}{\sim} P$ *we have the following. For any sequence* $\{\tau_k\}_{k=1}^\infty$, *the function sequence* $\{x^{\tau_k}(\cdot)\}$ *is relatively compact in* $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$. *In addition, for any sequence* $\tau_k \to \infty$, *any limit point of* $\{x^{\tau_k}(\cdot)\}$ *in* $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ *satisfies*

$$\bar{x}(t) = \bar{x}(0) + \int_0^t y(\tau)d\tau \quad \text{for all } t \in \mathbb{R}_+, \quad \text{where} \quad y(\tau) \in -G(x(\tau)) - \mathcal{N}_X(x(\tau)).$$

3.3. Properties of the limiting differential inclusion. Theorem 3 establishes that the updates (7), which include stochastic subgradient methods (Example 1), stochastic prox-linear methods (Example 2), or stochastic proximal point methods (Examples 3–4) have sample paths asymptotically approximated by the differential inclusion

$$\dot{x} \in -G(x) - \mathcal{N}_X(x) \quad \text{where} \quad G(x) = \partial f(x) + \partial\varphi(x)$$

for the objective $f(x) = \mathbb{E}[f(x;S)]$. To establish convergence of the iterates x_k themselves, we must understand the limiting properties of trajectories of the preceding differential inclusion.

We define the minimal subgradient

$$\mathsf{g}^\star(x) := \operatorname*{argmin}_g \left\{ \|g\|^2 \mid g \in \partial f(x) + \partial\varphi(x) + \mathcal{N}_X(x) \right\} = \pi_{G(x)+\mathcal{N}_X(x)}(0).$$

Before presenting the theorem on the differential inclusion, we need one regularity assumption on the objective function $F(x)$ and the constraint set X. Recall that a function f is coercive if $f(x) \to \infty$ as $\|x\| \to \infty$.

Assumption E. The function $x \mapsto F(x) + \mathbb{I}_X(x)$ is coercive.

This assumption ensures that the sublevel sets of the objective function $F + \mathbb{I}_X$ are compact. Now we have the following convergence theorem.

THEOREM 4. *Let Assumptions* A, B, *and* E *hold. Let* $x(\cdot)$ *be a solution to the differential inclusion* $\dot{x} \in -\partial f(x) - \partial\varphi(x) - \mathcal{N}_X(x)$ *initialized at* $x(0) \in X$. *Then* $x(t)$ *exists and is in* X *for all* $t \in \mathbb{R}_+$, $\sup_t \|x(t)\| < \infty$, $x(t)$ *is Lipschitz in* t, *and*

$$f(x(t)) + \varphi(x(t)) + \int_0^t \|\mathsf{g}^\star(x(\tau))\|^2 d\tau \le f(x(0)) + \varphi(x(0)).$$

We prove the theorem in section 3.3.1, giving a few corollaries to show that solutions to the differential inclusion converge to stationary points of $f + \varphi$.

COROLLARY 2. *Let $x(\cdot)$ be a solution to $\dot{x} \in -G(x) - \mathcal{N}_X(x)$ and assume that for some $t > 0$ we have $f(x(t)) = f(x(0))$. Then $\mathsf{g}^\star(x(\tau)) = 0$ for all $\tau \in [0, t]$.*

Proof. By Theorem 4, we have that $\int_0^t \left\|\mathsf{g}^\star(x(\tau))\right\|^2 d\tau = 0$, so that $\mathsf{g}^\star(x(\tau)) = 0$ for almost every $\tau \in [0, t]$. The continuity of $x(\cdot)$ and outer semicontinuity of G extend this to all τ. ∎

In addition, we can show that all cluster points of any trajectory solving the differential inclusion (11) are stationary. First, we recall the following definition.

DEFINITION 1. *Let $\{x(t)\}_{t\geq 0}$ be a trajectory. A point x_∞ is a* cluster point *of $x(t)$ if there exists an increasing sequence $t_n \to \infty$ such that $x(t_n) \to x_\infty$.*

We have the following observation.

COROLLARY 3. *Let $x(\cdot)$ be the trajectory of $\dot{x} \in -G(x) - \mathcal{N}_X(x)$, and let x_∞ be a cluster point of $x(\cdot)$. Then x_∞ is stationary, meaning that $\mathsf{g}^\star(x_\infty) = 0$.*

Proof. For $\epsilon > 0$, define $\mathcal{T}_\epsilon(x_\infty) = \{t \in \mathbb{R}_+ \mid \|x(t) - x_\infty\| \leq \epsilon\}$, and let μ denote Lebesgue measure on $\mathbb{R}$. Because the trajectory $x(\cdot)$ is Lipschitz, we have that $\mu(\mathcal{T}_\epsilon(x_\infty) \cap [T, \infty)) = \infty$ for all $\epsilon > 0$ and $T < \infty$ (cf. [1, Prop. 6.5.1]). Let ϵ_n, δ_n be sequences of positive numbers converging to 0. Because $f(x(t)) + \varphi(x(t))$ converges to $f(x_\infty) + \varphi(x_\infty)$ (the sequence is decreasing and $f + \varphi$ is continuous), we have $\int \left\|\mathsf{g}^\star(x(t))\right\|^2 dt < \infty$. Moreover, there exist increasing T_n such that

$$\int_{\mathcal{T}_{\epsilon_n}(x_\infty) \cap [T_n, \infty)} \left\|\mathsf{g}^\star(x(t))\right\|^2 dt \leq \delta_n.$$

As $\mu(\mathcal{T}_{\epsilon_n}(x_\infty) \cap [T_n, \infty)) = \infty$, there must exist an increasing sequence $t_n \geq T_n$, $t_n \in \mathcal{T}_{\epsilon_n}(x_\infty)$, such that $\left\|\mathsf{g}^\star(x(t_n))\right\|^2 \leq \delta_n$. By construction $x(t_n) \to x_\infty$, we have a subsequence $\mathsf{g}^\star(x(t_n)) \to 0$. The outer semicontinuity of $x \mapsto G(x) + \mathcal{N}_X(x)$ implies that $0 \in G(x_\infty) + \mathcal{N}_X(x_\infty)$. ∎

3.3.1. Proof of Theorem 4. Our argument proceeds in three main steps. For shorthand, we define $F(x) = f(x) + \varphi(x)$. Our first step shows that the function $V(x) := F(x) + \mathbb{I}_X(x) - \inf_{y \in X} F(y)$ is a Lyapunov function for the differential inclusion (11), where we take the function W in Lemma 5 to be $W(x, v) = \|v\|^2$. Once we have this, then we can use the existence result of Lemma 3 to show that a solution $x(\cdot)$ exists in a neighborhood of 0. The uniqueness of trajectories (Lemma 4) then implies that the trajectory x is nonincreasing for V, which then—combined with the assumption of coercivity of $F + \mathbb{I}_X$—implies that the trajectory x is bounded and we can extend uniquely it to all of $\mathbb{R}_+$.

Part 1: A Lyapunov function. To develop a Lyapunov function, we compute directional derivatives of $f + \varphi$.

LEMMA 13 (see [33, Chap. VI.1]). *Let h be convex and $g^\star = \operatorname{argmin}_{g \in \partial h(x)}\{\|g\|\}$. Then the directional derivative satisfies $h'(x; -g^\star) = -\left\|g^\star\right\|^2$.*

Now, take $\mathsf{g}^\star(x)$ as in the statement of the theorem and define the Lyapunov-like function $V(x) = f(x) + \varphi(x) + \mathbb{I}_X(x) - \inf_{y \in X}\{f(y) + \varphi(y)\}$; we claim that

$$V'(x; -\mathsf{g}^\star(x)) \leq -\left\|\mathsf{g}^\star(x)\right\|^2. \tag{21}$$

Before proving (21), we note that it is identical to that in Lemma 5 on monotone trajectories of differential inclusions. Thus there exists a solution $x(\cdot)$ to the differential inclusion $\dot{x} \in -G(x) - \mathcal{N}_X(x)$ defined on $[0,T]$ for some $T > 0$, where $x(\cdot)$ satisfies

$$f(x(t)) + \varphi(x(t)) + \mathbb{I}_X(x(t)) \le f(x(0)) + \varphi(x(0)) - \int_0^t \left\|\mathsf{g}^\star(x(\tau))\right\|^2 d\tau \tag{22}$$

for all $t \in [0,T]$. We return now to prove the claim (21). Let $x \in X$ and recall by Assumption B that for all $\lambda \ge \mathbb{E}[\lambda(S,x)]$ that $f + \frac{\lambda}{2}\left\|\cdot - x_0\right\|^2$ is convex in an ϵ-neighborhood of x. Now, define $F_x(y) = f(y) + \varphi(y) + \frac{\lambda}{2}\left\|y - x\right\|^2$, so that for v with $\|v\| = 1$ and $t \le \epsilon$, we have

$$\left|F(x+tv) - F(x)\right| \le \left|F_x(x+tv) - F(x)\right| + \frac{t^2\lambda^2}{2}\left\|v\right\|^2.$$

Because φ is convex and the error in the approximation f_x of f is second order, taking limits as $u \to v, t \to 0$, we have for any fixed $x \in X$ that

$$\begin{aligned}
&\liminf_{t \downarrow 0, u \to v} \frac{F(x+tu) + \mathbb{I}_X(x+tu) - F(x)}{t} \\
&= \liminf_{t \downarrow 0} \frac{F_x(x+tv) + \mathbb{I}_X(x+tv) - F_x(x)}{t} = \sup_{g \in \partial f(x) + \partial\varphi(x) + \mathcal{N}_X(x)} \left\langle g, v \right\rangle,
\end{aligned}$$

where $F(x) = f(x) + \varphi(x)$, and we have used that the subgradient set of $y \mapsto F_x(y)$ at $y = x$ is $\partial f(x) + \partial\varphi(x)$. Applying Lemma 13 with $v = -\mathsf{g}^\star(x)$ gives claim (21).

Part 2: Uniqueness of trajectories. Lemma 4 shows that solutions to $\dot{x} \in -G(x) - \mathcal{N}_X(x)$ have unique trajectories almost immediately. By Assumption B, for any $x \in X$, $f + \varphi + \frac{\lambda}{2}\left\|\cdot\right\|^2$ is convex on the set $X \cap \{x + \epsilon\mathbb{B}\}$ for all $\lambda \ge \mathbb{E}[\lambda(S,x)]$. Thus for points x_1, x_2 satisfying $\|x_i - x\| \le \epsilon$ and $g_i \in \partial f(x_i) + \partial\varphi(x_i) + \mathcal{N}_X(x_i)$,

$$\left\langle g_1 + \lambda x_1 - g_2 - \lambda x_2, x_1 - x_2 \right\rangle \ge 0 \quad \text{or} \quad \left\langle -g_1 + g_2, x_1 - x_2 \right\rangle \le \lambda \left\|x_1 - x_2\right\|^2,$$

because subgradients of convex functions are increasing [33, Chap. VI]. Now, suppose that on an interval $[0,T]$ the trajectory $x(t)$ satisfies $\|x(t)\| \le B$; that is, it lies in a compact subset of X. Then by taking a finite subcovering of $B\mathbb{B} \cap X$ as necessary, we may assume $f + \varphi + \frac{\lambda}{2}\left\|\cdot\right\|^2$ is convex over $B\mathbb{B} \cap X$. This preceding display is equivalent to the condition of Lemma 4, so that for any B and any interval $[0,T]$ for which the trajectory $x(t)$ satisfies $\|x(t)\| \le B$ on $t \in [0,T]$, the trajectory is unique. In particular, the Lyapunov inequality (22) is satisfied on the interval over which the trajectory $\dot{x} \in -G(x) - \mathcal{N}_X(x)$ is defined.

Part 3: Extension to all times. We argue that we may take $T \to \infty$. For any fixed $T < \infty$, we know that $f(x(T)) + \varphi(x(T)) \le f(x(0)) + \varphi(x(0))$, and the coercivity of $f + \varphi$ over X implies that there exists $B < \infty$ such that $\|x(t)\| \le B$ on this trajectory (i.e., $t \in [0,T]$). The compactness of $\partial f(x) + \partial\varphi(x)$ for $x \in X \cap \{y : \|y\| \le B\}$ implies that $\inf_g\{\|g\| \mid g \in \partial f(x) + \partial\varphi(x) + \mathcal{N}_X(x)\}$ is bounded (because $0 \in \mathcal{N}_X(x)$). The condition on existence of paths for all times T in Lemma 3 applies.

The Lipschitz condition on $x(t)$ is an immediate consequence of the boundedness of the subgradient sets $\partial f(x) + \partial\varphi(x)$ for bounded x.

3.4. Almost sure convergence to stationary points. Thus far we have shown that the limit points of the stochastic model-based iterations (7) are asymptotically equivalent to the differential inclusion (11) (Theorem 3) and that solutions

to the differential inclusion have certain uniqueness and convergence properties (Theorem 4). Based on those asymptotic equivalence results and convergence properties, this section shows that cluster points of the iterates x_k are stationary. To provide a starting point, we state the main theorem of the section, which applies to any sequence x_k generated by a model update (7) satisfying Conditions C.(i)–C.(iv).

THEOREM 5. *Let Assumptions* A, B, D*, and* E *hold. Then with probability* 1*,*

$$\left[\liminf_k F(x_k), \limsup_k F(x_k)\right] \subseteq F(X^\star) = \{f(x) : x \in X^\star\}. \tag{23}$$

Let us discuss the theorem briefly. Theorem 1 is an immediate consequence of Theorem 5, as Assumptions D and E are trivial when X is compact. To illustrate Theorem 5, we also establish convergence of the iterates of x_k to the stationary set $X^\star$ under the weak Sard-type Assumption C, giving Corollary 1 as a consequence.

COROLLARY 4. *Let Assumptions* A–E *hold. With probability* 1*, all cluster points of the sequence* $\{x_k\}_{k=1}^\infty$ *belong to the stationary set* $X^\star$*, and* $F(x_k) = f(x_k) + \varphi(x_k)$ *converges.*

Proof. By Assumption C (that $(F(X^\star))^c$ is dense), Theorem 5 implies that $F(x_k)$ converges. That all cluster points of x_k belong to $X^\star$ follows from Lemma 14 to come. □

Conditions for boundedness of the iterates. Key to our theorems is the boundedness of the iterates x_k, so it is important to give sufficient conditions that Assumption D holds even when X is unbounded. We may develop examples by considering the joint properties of the regularizer φ and objectives $f(x; S)$ in the stochastic updates of our methods. We mention two such examples, focusing for simplicity on the stochastic subgradient method (Example 1, using subgradient $\mathsf{g}(x; s) \in \partial f(x; s)$) in the unconstrained case $X = \mathbb{R}^d$. We first assume that $\varphi(x) = \frac{\lambda}{2}\|x\|^2$, i.e., ℓ_2 or Tikhonov regularization, common in statistical learning and inverse problems. In addition, let us assume that $f(x; s)$ is $L(s)$-Lipschitz in x, where $L := \mathbb{E}[L(S)^2]^{\frac{1}{2}} < \infty$, so that $\|\mathsf{g}(x; s)\| \le L(s)$. This regularization is sufficient to guarantee boundedness.

Observation 2. Let the conditions of the preceding paragraph hold. Assume that $\mathbb{E}[L(S)^2] < \infty$. Then with probability 1, $\sup_k \|x_k\| < \infty$.

We provide the proof of Observation 2 in Appendix A.4. More quickly growing regularization functions φ also yield bounded iterates. We begin with a definition.

DEFINITION 2. *A function* φ *is* β-coercive *if* $\lim_{\|x\|\to\infty} \varphi(x)/\|x\|^\beta = \infty$*, and it is* (λ, β)-regularly coercive *if it is* β*-coercive and* $\varphi(x) \ge \varphi(\lambda x)$ *for* $\|x\|$ *large.*

Observation 3. Let φ be (β, λ)-coercive with $\lambda \in [0, 1)$. Assume that for all $s \in \mathcal{S}$, $x \mapsto f(x; s)$ is $L(1 + \|x\|^\nu)$-Lipschitz in a neighborhood of x, where $L < \infty$ is some constant, and $\nu < \beta - 1$. Then $\sup_k \|x_k\| < \infty$.

The proof of Observation 3 is tangential to our main thrust; we provide it in [22].

3.4.1. Proof of Theorem 5. We prove the theorem using two intermediate results: in the first part (Lemma 14), we show that if a cluster point x_∞ of the sequence x_k is nonstationary, then the iterates $F(x_k)$ must decrease through $F(x_\infty)$ infinitely often. A consequence we show is that $\limsup_k F(x_k)$ and $\liminf_k F(x_k)$ belong to $F(X^\star)$. We then show (Lemma 15) that the interpolated path $x(\cdot)$ of the iterates x_k (recall definition (15)) cannot move too quickly (Lemma 15). We finally use this to show that all limting values of $f(x_k) + \varphi(x_k)$ belong to $F(X^\star)$. In the statements of the lemmas, we implicitly assume all of the conditions of the theorem (i.e., Assumptions A, B, D, and E).

We start with a result on the boundaries of the sequences $F(x_k)$ and the growth of the path $x(t)$ interpolating the iterates x_k (recall the definition (15)).

LEMMA 14. *With probability one,* $\liminf_k F(x_k) \in F(X^\star)$ *and* $\limsup_k F(x_k) \in F(X^\star)$. *For any increasing sequence* $\{h_k\} \subset \mathbb{R}$ *satisfying* $h_k \to \infty$ *and* $\lim_k x(h_k) = x_\infty \notin X^\star$ *and any sequence* $\tau_k \to \tau > 0$,

$$\liminf_k F(x(h_k - \tau_k)) > F(x_\infty) > \limsup_k F(x(h_k + \tau_k)). \tag{24}$$

Proof. We begin with the second claim (24) of the lemma, as the first is a nearly immediate consequence of the second. Let the probability 1 events of Theorem 3 hold; that is, the limit points of the shifted sequences $\{x^{\tau_k}(\cdot)\}$ satisfy the differential inclusion (11). We introduce the left- and right-shifted times

$$h_k^- = h_k - \tau_k \quad \text{and} \quad h_k^+ = h_k + \tau_k \quad \text{for } k \in \mathbb{N}.$$

To show the lemma, it suffices to show that, for any subsequence $\{h_{k(m)}\}$ of the sequence $\{h_k\}$, there exists a further subsequence $\{h_{k(m(n))}\}_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} F\left(x_{h^-_{k(m(n))}}\right) > F(x_\infty) > \lim_{n\to\infty} F\left(x_{h^+_{k(m(n))}}\right). \tag{25}$$

Now, fix a subsequence $\{h_{k(m)}\}_{m\in\mathbb{N}}$. By Assumption D, both sequences $\{x(h^-_{k(m)})\}$ and $\{x(h^+_{k(m)})\}$ are relatively compact in $\mathbb{R}^d$, and Theorem 3 implies that the sequence of shifted functions $\{x^{h^-_{k(m)}}(\cdot)\}_{m\in\mathbb{N}}$ is relatively compact in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$. As a consequence, there exists a further subsequence $\{h_{k(m(n))}\}_n$ such that for $u_n = x(h^-_{k(m(n))})$ and $v_n = x(h^+_{k(m(n))})$, there are points u_∞ and v_∞ and a function $\overline{x} \in \mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ such that

$$\lim_n u_n = u_\infty, \quad \lim_n v_n = v_\infty, \quad \text{and} \quad \lim_{n\to\infty} x^{h^-_{k(m(n))}}(\cdot) = \overline{x}(\cdot).$$

By this equation, that $\tau_k \to \tau$ as $k \to \infty$, and the assumption in the lemma that $\lim_k x(h_k) = \lim_k x^{h_k^-}(\tau_k) = x_\infty$, we have $\overline{x}(0) = u_\infty$, $\overline{x}(\tau) = x_\infty$, and $\overline{x}(2\tau) = v_\infty$. Theorem 3 shows that $\overline{x}$ satisfies the differential inclusion (11), which has monotone trajectory by Theorem 4. As $\overline{x}(\tau) = x_\infty \notin X^\star$, Corollary 2 implies the strict decrease

$$F(u_\infty) = F(\overline{x}(0)) > F(\overline{x}(\tau)) > F(\overline{x}(2\tau)) = F(v_\infty),$$

yielding inequality (25) and thus inequality (24).

Now we show the first claim of the lemma. Let $y = \liminf_k F(x_k)$ (the proof for case $y = \limsup_k F(x_k)$ is, mutatis mutandis, identical). As the sequence $\{x_k\}_{k=1}^\infty$ is bounded and the function F is continuous on X, there is a subsequence $\{x_{k(m)}\}_{m\in\mathbb{N}}$ with $x_{k(m)} \to x_\infty$ and $\lim_m F(x_{k(m)}) = F(x_\infty) = y$. Recall that $x_k = x(t_k)$ for $t_k = \sum_{i=1}^k \alpha_i$. If $x_\infty \notin X^\star$, then for any $\tau > 0$ and for $h_k = t_k$, inequality (24) implies $F(x_\infty) > \limsup_m F(x(h_{k(m)} + \tau)) \geq \liminf_k F(x_k)$, an absurdity, so we must have $x_\infty \in X^\star$. ∎

Our second intermediate result shows that the interpolated paths $x(\cdot)$ cannot move too quickly.

LEMMA 15. *For any two sequences* $\{h_k\}_{k=1}^\infty$ *and* $\{h'_k\}_{k=1}^\infty$ *satisfying* $h'_k > h_k$, $\lim_k h'_k = \lim_k h_k = \infty$ *and* $\liminf_k \|x(h'_k) - x(h_k)\| > 0$, *we have with probability* 1 *that* $\liminf_k (h'_k - h_k) > 0$.

Proof. As in the proof of Lemma 14, fix the sample $S_1, S_2, \ldots$ so that the probability 1 conclusions of Theorem 3 hold. Now, for $h \in \mathbb{R}_+$ define

$$k_<(h) = \max\{k \in \mathbb{N} : t_k \le h\} \quad \text{and} \quad k_>(h) = \min\{k \in \mathbb{N} : t_k \ge h\},$$

where we recall the interpolation times $t_k = \sum_{i=1}^k \alpha_i$. As $\alpha_k \to 0$, the statement $\liminf_{k\to\infty}(h_k - h'_k) > 0$ is equivalent to the statement $\liminf_{k\to\infty}(t_{k_>(h_k)} - t_{k_<(h'_k)}) > 0$. For any $m \le n \in \mathbb{N}$, we have

$$\begin{aligned}\|x(t_n) - x(t_m)\| &= \left\| \sum_{i=m+1}^n \alpha_i \overline{\mathsf{G}}_{\alpha_i}(x_i) + \sum_{i=m+1}^n \alpha_i \xi_i \right\| \\ &\le (t_n - t_m) \sup_i \left\| \overline{\mathsf{G}}_{\alpha_i}(x_i) \right\| + \left\| \sum_{i=m+1}^n \alpha_i \xi_i \right\|.\end{aligned}$$

Let $M = \sup_i \left\| \overline{\mathsf{G}}_{\alpha_i}(x_i) \right\| < \infty$ (use Lemmas 7 and 8 to see that $M < \infty$). Lemma 9 implies that $\lim_{m\to\infty} \sup_{n \ge m} \left\| \sum_{i=m+1}^n \alpha_i \xi_i \right\| = 0$. Thus, we obtain that for any $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $m, n \ge N$,

$$(t_n - t_m)M \ge \|x(t_n) - x(t_m)\| - \left\| \sum_{i=m+1}^n \alpha_i \xi_i \right\| \ge \|x(t_n) - x(t_m)\| - \epsilon. \tag{26}$$

As $x(\cdot)$ are linear interpolations of $x_k = x(t_k)$ and $h_k, h'_k \to \infty$, for any $\epsilon > 0$ there exists exists $K \in \mathbb{N}$ such that $k \ge K$ implies

$$\begin{aligned}\|x(h'_k) - x(h_k)\| &\le \max\left\{\|x(t_n) - x(t_m)\| : n, m \in [k_<(h'_k), k_>(h_k)]\right\} \\ &\le \left(t_{k_>(h_k)} - t_{k_>(h'_k)}\right) M + \epsilon.\end{aligned}$$

Since $\liminf_k \|x(h'_k) - x(h_k)\| > 0$, inequality (26) gives the result. ∎

To prove the theorem, we assume $(\liminf_k F(x_k), \limsup_k F(x_k))$ is nonempty, as otherwise the result is trivial. As in the proof of Lemma 14, fix the sample $S_1, S_2, \ldots$ so that the probability 1 conclusions of Theorem 3 hold.

Suppose for the sake of contradiction that $y^{\mathsf{hi}} \in (\liminf_k F(x_k), \limsup_k F(x_k))$ satisfies $y^{\mathsf{hi}} \notin F(X^\star)$. Let $y^{\mathsf{lo}} < y^{\mathsf{hi}}$, $y^{\mathsf{lo}} \in (\liminf_k F(x_k), \limsup_k F(x_k))$. We claim we may choose sequences $\{h_k^{\mathsf{lo}}\}$ and $\{h_k^{\mathsf{hi}}\}$ with $h_k^{\mathsf{lo}} < h_k^{\mathsf{hi}}$, $\lim_k h_k^{\mathsf{lo}} = \lim_k h_k^{\mathsf{hi}} = \infty$, and

$$F(x(h_k^{\mathsf{lo}})) = y^{\mathsf{lo}}, \quad F(x(h_k^{\mathsf{hi}})) = y^{\mathsf{hi}}, \quad \text{and} \quad y^{\mathsf{lo}} < F(x(t)) < y^{\mathsf{hi}} \quad \text{for} \quad t \in (h_k^{\mathsf{lo}}, h_k^{\mathsf{hi}}). \tag{27}$$

To see that sequences satisfying condition (27) exist, we consider traversals of the interval $[y^{\mathsf{lo}}, y^{\mathsf{hi}}]$ (see Figure 1). As $\liminf_k F(x_k) < y^{\mathsf{lo}} < y^{\mathsf{hi}} < \limsup_k F(x_k)$, there exist increasing sequences $\widetilde{h}'_k$ and $\widetilde{h}_k$ with

$$F(x(\widetilde{h}'_k)) = y^{\mathsf{lo}}, \quad F(x(\widetilde{h}_k)) = y^{\mathsf{hi}} \quad \text{and} \quad \widetilde{h}'_k < \widetilde{h}_k.$$

Then we define the last entrance and first subsequent exit times

$$h_k^{\mathsf{lo}} := \sup\{h \in [\widetilde{h}'_k, \widetilde{h}_k] : f(x(h)) \le y^{\mathsf{lo}}\} \quad \text{and} \quad h_k^{\mathsf{hi}} := \inf\{h \in [h_k^{\mathsf{lo}}, \widetilde{h}_k] : f(x(h)) \ge y^{\mathsf{hi}}\}.$$

The continuity of F and $x(\cdot)$ show that the conclusion (27) holds.

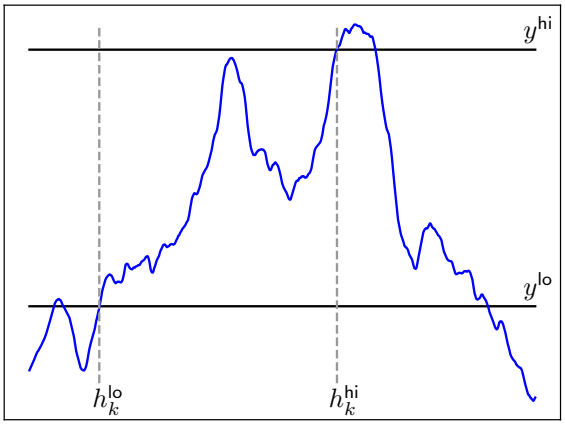


FIG. 1. *Illustration of proof of Theorem* 5. *The erratic line represents a trajectory* $F(x(t))$, *with last entrance time* h_k^{lo} *and first exit time* h_k^{hi}. *Such upcrossings must be separated in time by the strict decreases in Lemma* 14.

By taking a subsequence if necessary, we assume without loss of generality (w.l.o.g.) that $x(h_k^{\mathsf{hi}}) \to x_\infty$. By continuity, we have $y^{\mathsf{hi}} = F(x_\infty)$ and $x_\infty \notin X^\star$ as $y^{\mathsf{hi}} \notin F(X^\star)$. Now, fix some $\tau > 0$, and take $y_\tau = \liminf_k F(x(h_k^{\mathsf{hi}} - \tau))$, which satisfies $y_\tau > F(x_\infty) = y^{\mathsf{hi}}$ by Lemma 14, because $x_\infty \notin X^\star$. Consider the value gap $\Delta = \frac{1}{2}\min\{|y_\tau - y^{\mathsf{hi}}|, |y^{\mathsf{hi}} - y^{\mathsf{lo}}|\} > 0$. The continuity of F implies for some $\delta > 0$, we have $|F(x) - y^{\mathsf{hi}}| < \Delta$ for $x \in X \cap \{x_\infty + \delta\mathbb{B}\}$. As $\liminf_k |F(x(h_k^{\mathsf{lo}})) - F(x_\infty)| = |y^{\mathsf{lo}} - y^{\mathsf{hi}}| > \Delta$ and $\liminf_k |F(x(h_k^{\mathsf{hi}} - \tau)) - F(x_\infty)| = |y_\tau - y^{\mathsf{hi}}| > \Delta$, by continuity of F and our choice of δ, we must have the separation

$$\liminf_k \left\|x(h_k^{\mathsf{lo}}) - x_\infty\right\| > \delta, \text{ and } \liminf_k \left\|x(h_k^{\mathsf{hi}} - \tau) - x_\infty\right\| > \delta. \tag{28}$$

For this value $\delta > 0$, consider the sequence $\{h_k^\delta\}_{k=1}^\infty$ defined by

$$h_k^\delta = \max_t \left\{t \mid t < h_k^{\mathsf{hi}}, \left\|x(t) - x(h_k^{\mathsf{hi}})\right\| = \delta\right\}.$$

Then using $x(h_k^{\mathsf{hi}}) \to x_\infty$, we have $\liminf_k \left\|x(h_k^{\mathsf{lo}}) - x(h_k^{\mathsf{hi}})\right\| > \delta$ and so

$$h_k^\delta \in [h_k^{\mathsf{lo}}, h_k^{\mathsf{hi}}] \text{ eventually, and } F(x(h_k^\delta)) \in [y^{\mathsf{lo}}, y^{\mathsf{hi}}] \tag{29}$$

by definition (27) of the upcrossing times.

By (28) and that $x(h_k^{\mathsf{hi}}) \to x_\infty$, we have $h_k^\delta > \max\{h_k^{\mathsf{lo}}, h_k^{\mathsf{hi}} - \tau\}$ for large enough k. In particular, this implies that $\limsup_k (h_k^{\mathsf{hi}} - h_k^\delta) \le \tau$. Because the paths $x(\cdot)$ cannot move too quickly by Lemma 15, the quantity $\tau(\delta) := \liminf_k (h_k^{\mathsf{hi}} - h_k^\delta) \in (0, \tau]$. By taking subsequences if necessary, we may assume w.l.o.g. that the sequence $h_k^\delta - h_k^{\mathsf{hi}} \to \tau_\infty \in [\tau(\delta), \tau]$, so that $h_k^\delta = h_k^{\mathsf{hi}} - \tau_k$ for $\tau_k \to \tau_\infty > 0$. As $x(h_k^{\mathsf{hi}}) \to x_\infty \notin X^\star$, Lemma 14 implies that $\liminf_k F(x(h_k^\delta)) > y^{\mathsf{hi}}$, contradicting the containments (29). This is the desired contradiction, which gives the theorem.

4. Experiments. The asymptotic results in the previous sections provide somewhat limited guidance for application of the methods. To that end, in this section

we present experimental results explicating the performance of the methods as well as comparing their performance to the deterministic prox-linear method (5) (adapted from [19, sect. 5]). Drusvyatskiy and Lewis [19] provide a convergence guarantee for the deterministic method that after $O(1/\epsilon^2)$ iterations, the method can output an ϵ-approximate stationary point, that is, a point $\widehat{x}$ such that there exists x_0 with $\|\widehat{x} - x_0\| \le \epsilon$ and $\min\{\|g\| : g \in \partial f(x_0)\} \le \epsilon$. These comparisons provide us a somewhat better understanding of the practical advantages and disadvantages of the stochastic methods we analyze.

We consider the following problem. We have observations $b_i = \langle a_i, x^\star \rangle^2$, $i = 1, \ldots, n$, for an unknown vector $x^\star \in \mathbb{R}^d$, and we wish to find $x^\star$. This is a quadratic system of equations, which arises (for example) in phase retrieval problems in imaging science as well as in a number of combinatorial problems [10, 9]. The natural exact penalty form of this system of equations yields the minimization problem

$$\underset{x}{\text{minimize}} \;\; f(x) := \frac{1}{n} \sum_{i=1}^{n} |\langle a_i, x \rangle^2 - b_i|, \tag{30}$$

which is certainly of the form (1) with the function $h(t) = |t|$ and $c_i(x) = \langle a_i, x \rangle^2 - b_i$, so we may take the sample space $\mathcal{S} = \{1, \ldots, n\}$. In somewhat more general noise models, we may also assume we observe $b_i = \langle a_i, x^\star \rangle^2 + \xi_i$ for some noise sequence ξ_i; in this case the problem (30) is a natural robust analogue of the typical phase retrieval problem, which uses the smooth objective $(\langle a_i, x \rangle^2 - b_i)^2$. While there are a number of specialized procedures for solving such quadratic equations [10], we view problem (30) as a natural candidate for exploration of our algorithms' performance.

The stochastic prox-linear update of Example 2 is reasonably straightforward to compute for the problem (30). Indeed, as $\nabla_x(\langle a_i, x \rangle^2 - b_i) = 2 \langle a_i, x \rangle a_i$, by rescaling by the stepsize α_k we may simplify the problem to minimizing $|b + \langle a, x \rangle| + \frac{1}{2} \|x - x_0\|^2$ for some scalar b and vectors $a, x_0 \in \mathbb{R}^d$. A standard Lagrangian calculation shows that

$$\underset{x}{\text{argmin}} \left\{ |b + \langle a, x \rangle| + \frac{1}{2} \|x - x_0\|^2 \right\} = x_0 - \pi(\lambda) a, \quad \text{where } \lambda = \frac{\langle x_0, a \rangle + b}{\|a\|^2}$$

and $\pi(\cdot)$ is the projection of its argument into the interval $[-1, 1]$. The full proximal step (Example 3) is somewhat more expensive, and for general weakly convex functions, it may be difficult to estimate $\rho(s)$, the weak-convexity constant; nonetheless, in section 4.3 we use it to evaluate its merits relative to the prox-linear updates in terms of robustness to stepsize. Each iteration k of the deterministic prox-linear method [7, 19] requires solving the quadratic program

$$x_{k+1} = \underset{x}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} |\langle a_i, x_k \rangle^2 + 2 \langle a_i, x_k \rangle \langle a_i, x - x_k \rangle - b_i| + \frac{1}{2\alpha} \|x - x_k\|^2 \right\}, \tag{31}$$

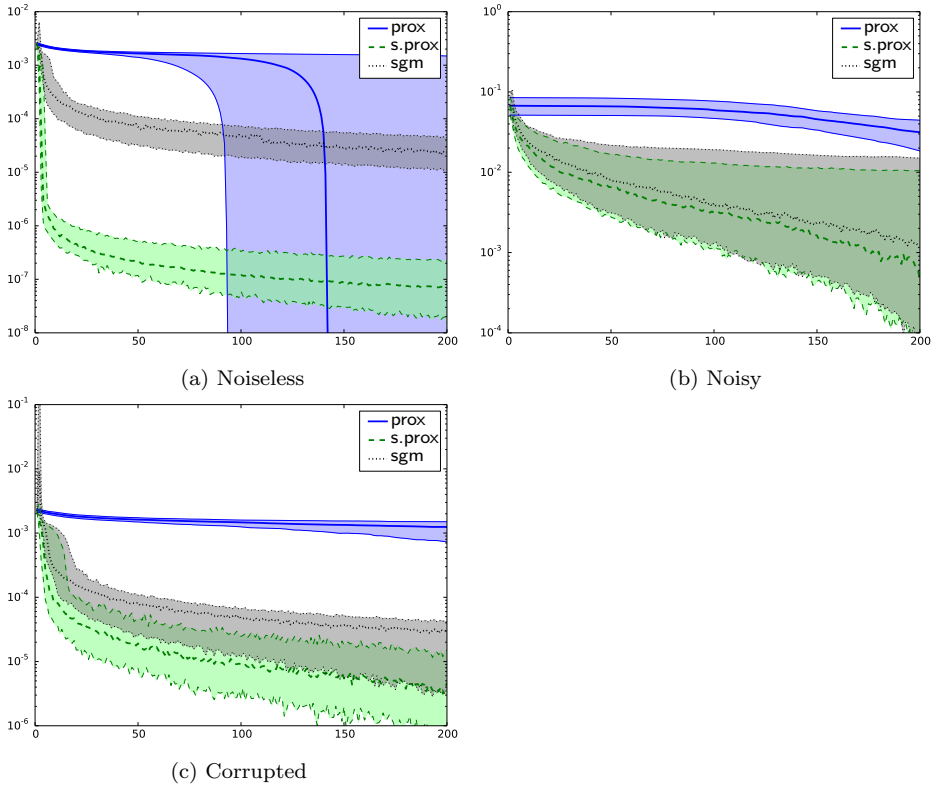
which we perform using Mosek via the `Convex.jl` package in `Julia` [52].

Before we present our results, we describe our choices for all parameters in our experiments. In each experiment, we let $n = 500$ and $d = 50$, and we choose $x^\star$ uniformly from the unit sphere $\mathbb{S}^{d-1}$. The noise variables ξ_i are independently and identically distributed Laplacian random variables with mean 0 and scale parameter σ, which we vary in our experiments. We construct the design matrix $A \in \mathbb{R}^{n \times d}$,

$A = [a_1 \ \cdots \ a_n]^T$, where each row is a measurement vector a_i, as follows: we choose $U \in \mathbb{R}^{n\times d}$ uniformly from the orthogonal matrices in $\mathbb{R}^{n\times d}$, i.e., $U^TU = I_{d\times d}$. We then make one of two choices for A, the first of which controls the condition number of A and the second the regularity of the norms of the rows a_i. In the former case, we set $A = UR$, where $R \in \mathrm{diag}(\mathbb{R}^d) \subset \mathbb{R}^{d\times d}$ is diagonal with linearly spaced diagonal elements in $[1, \kappa]$, so that $\kappa \geq 1$ gives the condition number of A. In the latter case, we set $A = RU$, where $R \in \mathrm{diag}(\mathbb{R}^n) \subset \mathbb{R}^{n\times n}$ is again diagonal with linearly spaced elements in $[1, \kappa]$. Finally, in each of our experiments, we set the stepsize for the stochastic methods as $\alpha_k = \alpha_0 k^{-\beta}$, where $\alpha_0 > 0$ is the initial stepsize and $\beta \in (\frac{1}{2}, 1)$ governs the rate of decrease in stepsize. We present three experiments in more detail in the coming subsections: (i) basic performance of the algorithms, (ii) the role of conditioning in the data matrix A, and (iii) an analysis of stepsize sensitivity for the different stochastic methods, that is, an exploration of the effects of the choices of α_0 and β in the stepsize choice.

4.1. Performance for well-conditioned problems. In our first group of experiments, we investigate the performance of the three algorithms under noiseless and noisy observational situations. In each of these experiments, we set the condition number $\kappa = \kappa(A) = 1$. We consider three experimental settings to compare the procedures: in the first, we have noiseless observations $b_i = \langle a_i, x^\star\rangle^2$; in the second, we set $b_i = \langle a_i, x^\star\rangle^2 + \xi_i$, where ξ_i are Laplacian with scale $\sigma = 1$; and in the third, we again have noiseless observations $b_i = \langle a_i, x^\star\rangle^2$, but for a fraction $p = 0.1$ of the observations, we replace b_i with an independent $\mathsf{N}(0, 25)$ random variable, so that $n/10$ of the observations provide no information. Within each experimental setting, we perform $N = 100$ independent tests, and in each individual test we allow the stochastic methods to perform $N = 200n$ iterations (so approximately 200 loops over the data). For the deterministic prox-linear method (31), we allow 200 iterations. Each deterministic iteration is certainly more expensive than n (sub)gradient steps or stochastic prox-linear steps, but it provides a useful benchmark for comparison. The stochastic methods additionally require specification of the initial stepsize α_0 and power β for $\alpha_k = \alpha_0 k^{-\beta}$, and to choose this, we let $\alpha_0 \in \{1, 10, 10^2, 10^3\}$ and $\beta \in \{0.6, 0.7, 0.8, 0.9\}$, perform $3n$ steps of the stochastic method with each potential pair (α_0, β), and then perform the full $N = 200n$ iterations with the best performing pair. We measure performance of the methods within each test by plotting the gap $f(x_k) - f(x^\star)$, where we approximate $x^\star$ by taking the best iterate x_k produced by any of those methods. While the problem is nonconvex and thus may have spurious local minima, these gaps provide a useful quantification of (relative) algorithm performance.

We summarize our experimental results in Figure 2. In each plot, we plot the median of the excess gap $f(x_k) - f(x^\star)$ as well as its 10% and 90% confidence intervals over our $N = 100$ tests. In order to compare the methods, the horizontal axis scales as iteration k divided by n for the stochastic methods and as iteration for the deterministic method (31). Each of the three methods is convergent in these experiments, and the stochastic methods exhibit fast convergence to reasonably accurate (say $\epsilon \approx 10^{-4}$) solutions after a few passes through the data. Eventually (though we do not always plot such results) the deterministic prox-linear algorithm achieves substantially better accuracy, though its progress is often slower. This corroborates substantial experience from the convex case with stochastic methods (c.f. [40, 24]).



(a) Noiseless

(b) Noisy

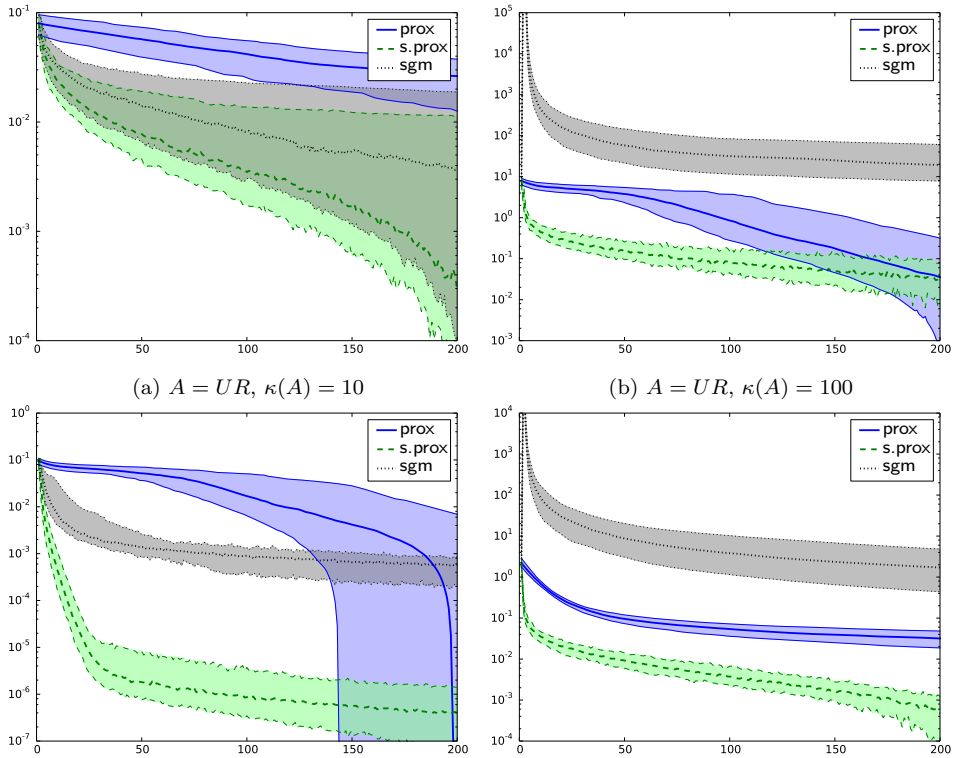
(c) Corrupted

FIG. 2. *Experiments with well-conditioned A matrices. The vertical axis shows* $f(x_k) - f(x^\star)$, *the horizontal axis iteration count for the prox-linear iteration* (31) *or the* k/n*th iteration for the stochastic methods. The key:* prox *is the prox-linear iteration* (5), s.prox *is the stochastic prox-linear method (Example* 2*),* sgm *is the stochastic subgradient method (Example* 1*).* (a) *Methods with no noise.* (b) $\xi_i \stackrel{\text{iid}}{\sim}$ *Laplacian with scale* $\sigma = 1$. (c) *Proportion* $p = 0.1$ *of observations* b_i *corrupted arbitrarily.*

There are differences in behavior for the different methods, which we can heuristically explain. In Figure 2(a), the stochastic prox-linear method (Example 2) converges substantially more quickly than the stochastic subgradient method. Intuitively, we expect this behavior because each data point (a_i, b_i) should have $\langle a_i, x\rangle^2 = b_i$ exactly, and the precise stepping of the prox-linear method achieves this more easily. In Figure 2(b), where $b_i = \langle a_i, x^\star\rangle^2 + \xi_i$ the two methods have similar behavior; in this case, the population expectation $f_{\text{pop}}(x) = \mathbb{E}[|b - \langle a, x^\star\rangle|^2]$ is smooth, because the noise ξ has a density, so gradient methods are likely to be reasonably effective. Moreover, with probability 1 we have $\langle a_i, x^\star\rangle^2 \neq b_i$, so that the precision of the prox-linear step is unnecessary. Finally, Figure 2(c) shows that the methods are robust to corruption, but because we have $\langle a_i, x^\star\rangle^2 = b_i$ for the majority of $i \in \{1, \ldots, n\}$, there is still benefit to using the more exact (stochastic) prox-linear iteration. We note in passing that the gap in function values $f(x_k)$ between the stochastic prox-linear method and stochastic subgradient method (SGM) is statistically significantly positive at the $p = 10^{-2}$ level for iterations $k = 1, \ldots, 20$, and that at each iteration k, the prox-linear method outperforms SGM for at least 77 of the $N = 100$ experiments (which is statistically significant for rejecting the hypothesis that each is equally likely to achieve lower objective value than the other at level $p = 10^{-6}$).

4.2. Problem conditioning and observation irregularity. In our second set of experiments, we briefly investigate conditioning of the problem (30) by modifying the condition number $\kappa = \kappa(A)$ of the measurement matrix $A \in \mathbb{R}^{n \times d}$ or by modifying the relative norms of the rows $\|a_i\|$ of A. In each of the experiments, we choose the initial stepsize α_0 and power β in $\alpha_k = \alpha_0 k^{-\beta}$ using the same heuristic as the previous experiment for the stochastic methods (by considering a grid of possible values and selecting the best after $3n$ iterations). We present four experiments, whose results we summarize in Figure 3. As in the previous experiments, we plot the gaps $f(x_k) - f(x^\star)$ versus iteration k (for the deterministic prox-linear method) and versus iteration k/n for the stochastic methods. In the first two, we use observations $b_i = \langle a_i, x^\star \rangle^2 + \xi_i$, where the noise variables are i.i.d. Laplacian with scale $\sigma = 1$, and we set $A = UR$ where R is diagonal, in the first (Figure 3(a)) scaling between 1 and $\kappa = 10$ and in the second (Figure 3(b)) scaling between 1 and $\kappa = 100$. Each method's performance degrades as the condition number $\kappa = \kappa(A)$ increases, as one would expect. The performance of SGM degrades substantially more quickly with the conditioning of the matrix A, in spite of the fact that noisy observations improve its performance relative to the other methods (in the case $\sigma = 0$, SGM's relative performance is worse).

In the second two experiments, we set $A = RU$, where R is diagonal with entries linearly spaced in $[1, \kappa]$ for $\kappa = 10$, so that the norms $\|a_i\|$ are irregular (varying by approximately a factor of $\kappa = 10$). In the first of the experiments (Figure 3(c)), we set



(a) $A = UR$, $\kappa(A) = 10$ (b) $A = UR$, $\kappa(A) = 100$

(c) $A = RU$, $b_i = \langle a_i, x^\star \rangle^2$ (d) $A = RU$, $b_i = \langle a_i, x^\star \rangle^2 + \xi_i$

FIG. 3. *Experiments with A matrices of varying condition number and irregularity in row norms.*

the observations $b_i = \langle a_i, x^\star \rangle^2$ with no noise, while in the second (Figure 3(d)) we set $b_i = \langle a_i, x^\star \rangle^2 + \xi_i$ for ξ_i i.i.d. Laplacian with scale $\sigma = 1$. In both cases, the stochastic prox-linear method has better performance—this is to be expected, because its more exact updates involving the linearization $h(c(x_k; s) + \nabla c(x_k; s)^T (x - x_k); s)$ are more robust to scaling of $\|a_i\|$. As we explore more carefully in the next set of experiments, one implication of these results is that the robustness and stability of the stochastic prox-linear algorithm with respect to problem conditioning is reasonably good, while the behavior of stochastic subgradient methods can be quite sensitive to conditioning behavior of the design matrix A.

4.3. Robustness of stochastic methods to stepsize. In our final experiment, we investigate the effects of stepsize parameters for the behavior of our stochastic methods. For stepsizes $\alpha_k = \alpha_0 k^{-\beta}$, the stochastic methods require specification of both the parameter α_0 and β, so it is interesting to investigate the robustness of the stochastic prox-linear method and SGM to various settings of α_0 and β. In each of these experiments, we set the condition number $\kappa(A) = 1$ and have no noise, i.e., $b_i = \langle a_i, x^\star \rangle^2$. We vary the initial stepsize $\alpha_0 \in \{2^{-1}, 2^1, 2^3, \ldots, 2^{11}\}$ and the power $\beta \in \{0.5, 0.55, 0.6, \ldots, 1\}$. In this experiment, we have $f(x^\star) = 0$, and we investigate the number of iterations

$$T(\epsilon) := \inf \{k \in \mathbb{N} \mid f(x_k) \le \epsilon\}$$

required to find and ϵ-optimal solution. (In our experiments, the stochastic methods always find such a solution eventually.) We perform $N = 250$ tests for each setting of the pairs α_0, β, and in each test, we implement all three of the stochastic gradient (Example 1), prox-linear (Example 2), and proximal-point (Example 3) methods, each for $k = 200n$ iterations, setting $T(\epsilon) = 200n$ if no iterate x_k satisfies $f(x_k) \le \epsilon$.

Figure 4 illustrates the results of these experiments, where the vertical axis gives the median time $T(\epsilon)$ to $\epsilon = 10^{-2}$-accuracy over all the $N = 250$ tests. The left plot demonstrates convergence time of the stochastic prox-linear and subgradient methods versus the initial stepsize α_0 and power β, indicated on the horizontal axes. The solid white-to-blue surface, with thin lines, corresponds to the iteration counts for the stochastic prox-linear method; the transparent surface with thicker lines corresponds

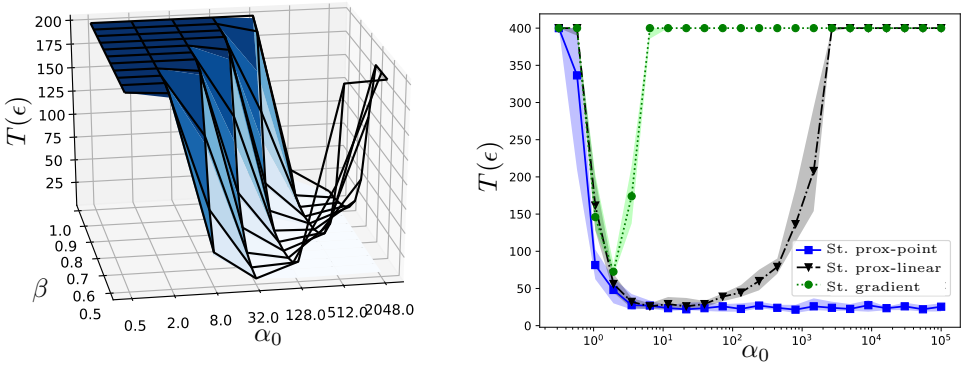


FIG. 4. *Iterations to achieve ϵ-accuracy. Left: time to convergence versus α_0 and β for stochastic subgradient (wireframe) and stochastic prox-linear (solid surface) methods. Right: time to convergence versus initial stepsize α_0 with $\beta = \frac{1}{2}$ for stochastic proximal, prox-linear, and gradient methods.*

to the iteration counts for the stochastic subgradient method. Figure 4 shows that the stochastic prox-linear algorithm consistently has comparable or better performance than SGM for the same choices of parameters α_0, β. The right plot shows convergence of the stochastic proximal-point method (see Example 3 in section 2.1), stochastic prox-linear method, and stochastic subgradient method versus stepsize on a log-plot of initial stepsizes, with $\beta = \frac{1}{2}$ fixed. The most salient aspect of the figures is that the stochastic prox-linear and proximal-point methods are more robust to stepsize (mis-)specification than is SGM. Indeed, Figure 4 makes apparent, the range of stepsizes yielding good performance for SGM is a relatively narrow valley, while the prox-linear and proximal-point methods enjoy reasonable performance for broad choices of (often large) stepsizes α_0, with less sensitivity to the rate of decrease β in the stepsize as well. This behavior is expected: the iterations of the stochastic prox-linear and proximal-point methods (Examples 2–3) guard more carefully against wild swings that result from aggressive stepsize choices, yielding more robust convergence and easier stepsize selection.

Appendix A. Technical proofs and results.

A.1. Proof of Claim 1. Fix $s \in \mathcal{S}$; we let $h = h(\cdot; s)$ and $c = c(\cdot; s)$ for notational simplicity. Then for any y, z with $\|y - x\| \le \epsilon$ and $\|z - x\| \le \epsilon$ and some vector v with $\|v\| \le \beta_\epsilon \|y - z\|^2 / 2$, we have

$$\begin{aligned} h(c(y)) &= h(c(z) + \nabla c(z)^T (y - z) + v) \\ &\overset{(i)}{\ge} h(c(z) + \nabla c(z)^T (y - z)) - \gamma_\epsilon(x) \|v\| \\ &\overset{(ii)}{\ge} h(c(z)) + \partial h(c(z))^T \nabla c(z)^T (y - z) - \frac{\gamma_\epsilon(x)\beta_\epsilon(x)}{2} \|z - y\|^2 , \end{aligned}$$

where inequality (i) follows from the local Lipschitz continuity of h and (ii) because h is subdifferentiable. Let $\lambda \ge \gamma_\epsilon(x)\beta_\epsilon(x)$. Then adding the quantity $\frac{\lambda}{2} \|y - x_0\|^2$ to both sides of the preceding inequalities, we obtain for any $g \in \partial h(c(x))$ that

$$\begin{aligned} h(c(y)) + \frac{\lambda}{2} \|y - x_0\|^2 &\ge h(c(z)) + (\nabla c(z) g)^T (y - z) - \frac{\lambda}{2} \|z - y\|^2 + \frac{\lambda}{2} \|y - x_0\|^2 \\ &= h(c(z)) + \frac{\lambda}{2} \|z - x_0\|^2 + \langle \nabla c(z) g, y - z \rangle + \lambda \langle z - x_0, y - z \rangle . \end{aligned}$$

That is, the function $y \mapsto h(c(y)) + \frac{\lambda}{2} \|y - x_0\|^2$ has subgradient $\nabla c(z) g + \lambda(z - x_0)$ at $y = z$ for all z with $\|z - x\| \le \epsilon$; any function with nonempty subdifferential everywhere on a compact convex set must be convex on that set [33]. In particular, we see that $y \mapsto f(y; s)$ is $\lambda(s, x) = \gamma_\epsilon(x, s)\beta_\epsilon(x, s)$-weakly convex in an ϵ-neighborhood of x, giving the result.

The final result on Condition C.(iv) is nearly immediate: we have

$$h(c(y; s); s) \ge h(c(x; s) + \nabla c(x; s)^T (y - x); s) - \frac{\gamma_\epsilon(x; s)\beta_\epsilon(x; s)}{2} \|y - x\|^2$$

for y in an ϵ-neighborhood of x by the Lipschitz continuity of h and ∇c.

A.2. Proof of Lemma 6. Recall Assumption B that for all $x \in X$ and some $\epsilon > 0$, there exists $\lambda(s, x)$ such that $y \mapsto f(y; s) + \frac{\lambda(s,x)}{2} \|y - x\|^2$ is convex for $\|y - x\| \le \epsilon$, and $\mathbb{E}[\lambda(S, x)] < \infty$ for all x. Then $f(\cdot; s)$ has a Fréchet subdifferential $\partial f(x; s)$ and the directional derivative of $f(\cdot; s)$ in the direction v is $f'(x; s; v) = \sup_{g \in \partial f(x;s)} \langle g, v \rangle$ (cf. [47, Chap. 8]). Let $\lambda(s) = \lambda(s, x)$ for shorthand, as x is fixed throughout our argument. Fix $v \in \mathbb{R}^d$, and let u be near v with $t < \epsilon / (\|u\| + \|v\|)$. Then

$$f(x+tu) = \int \left[f(x+tu;s) + \frac{\lambda(s)t^2}{2} \|u\|^2 \right] dP(s) - \frac{t^2}{2} \|u\|^2 \mathbb{E}[\lambda(S)].$$

Because $u \mapsto f(x+tu;s) + \frac{\lambda(s)}{2} \|tu\|^2$ is a normal convex integrand [47, Chap. 14], the dominated convergence theorem implies that

$$\begin{aligned} \frac{f(x+tu)-f(x)}{t} &= \int \left[\frac{f(x+tu;s)-f(x;s)}{t} + t\frac{\lambda(s)}{2} \|u\|^2 \right] dP(s) - \frac{t\|u\|^2}{2} \mathbb{E}[\lambda(S)] \\ &\to \int f'(x;s;v) dP(s) \quad \text{as } t \to 0, u \to v. \end{aligned}$$

That is, $f'(x;v) = \int f'(x,s;v) dP(s)$. An argument parallel to that of Bertsekas [3, Prop. 2.1–2.2] yields that $\partial f(x) = \int \partial f(x;s) dP(s)$ and that $\partial f(x)$ is compact.

Now we show that $\partial f(\cdot)$ is o.s.c. Because the support function of the subdifferential $\partial f(x)$ is the directional derivative of f, the outer semicontinuity of ∂f is equivalent to

$$\limsup_{k \to \infty} f'(x_k;v) \le f'(x;v) \quad \text{for all } \|v\| = 1 \text{ and } x_k \to x \in X \tag{32}$$

(cf. [33, Prop. V.3.3.9]). The sets $\partial_y (f(y;s) + (\lambda(s)/2)\|y-x\|^2)$ are bounded for y in a neighborhood of x because the function f is weakly convex, where $\lambda(\cdot)$ is P-integrable. Let $\lambda = \mathbb{E}_P[\lambda(S,x)] < \infty$, and define $g(y) = f(y) + \frac{\lambda}{2}\|y-x\|^2$. Then g is convex and continuous near x [3], and we have [33, Cor. VI.6.2.5] that

$$g'(x;v) = \limsup_{y \to x} g'(y;v)$$

for all $v \in \mathbb{R}^d$. But for convex g, we have $g'(x;v) = \lim_{t \downarrow 0} (g(x+tv) - g(x))/t$, and so the preceding display implies that as $y \to x$ we have

$$\begin{aligned} -o(1) &\le g'(x;v) - g'(y;v) \\ &= \lim_{t \downarrow 0} \left[\frac{f(x+tv)-f(x)}{t} + \frac{\lambda t \|v\|^2}{2} \right] - \lim_{t \downarrow 0} \left[\frac{f(y+tv)-f(y)}{t} + \lambda \langle v, y-x \rangle + \frac{\lambda t \|v\|^2}{2} \right] \\ &= \lim_{t \downarrow 0} \frac{f(x+tv)-f(x)}{t} - \lim_{t \downarrow 0} \frac{f(y+tv)-f(y)}{t} - \lambda \langle v, y-x \rangle . \end{aligned}$$

In particular, the preceding limits exist, we have $f'(x;v) = \lim_{t \downarrow 0} (f(x+tv) - f(x))/t$, and inequality (32) holds by taking $y \to x$. The preceding argument works, of course, for any weakly convex function, and so applies to $f(\cdot;s)$ as well.

The final claim of the lemma is a standard calculation [19, 21].

A.3. Proof of Lemma 8. That $L_\epsilon(x;s) < \infty$ for all x is immediate, as $G(x;s) = \partial\varphi(x) + \partial f(x;s)$, and subdifferentials of convex functions (φ) are compact convex sets.

We first show the upper semicontinuity of the function $L_\epsilon(\cdot;s)$. Suppose for the sake of contradiction that for some $x \in X$ and for some sequence $\{x_k\}_{k=1}^\infty \subset X$ converging to x, we have $\lim_{k\to\infty} L_\epsilon(x_k;s)$ exists and there is some $\delta > 0$ such that

$$\lim_{k \to \infty} L_\epsilon(x_k;s) \ge L_\epsilon(x;s) + \delta.$$

By definition of L_ϵ we may choose $x_k' \in X$ such that $\|x_k - x_k'\| \le \epsilon$ and subgradient vectors $p_k(s) \in \partial f(x_k';s)$ and $q_k(s) \in \partial\varphi(x_k')$ satisfying

$$L_\epsilon(x_k;s) \le \|p_k(s) + q_k(s)\| + \delta/2 \quad \text{for all } k.$$

Since the sequence $\{x_k'\} \subset X$ is bounded, it has accumulation points and we may assume w.l.o.g. that $x_k' \to x' \in X$, where x' satisfies $\|x' - x\| \le \epsilon$. The o.s.c. of the subdifferential for weakly convex functions (Lemma 1 or 6) shows that there must be a subsequence $\{n_k\}$ satisfying $p_{n_k}(s) \to p(s) \in \partial f(x'; s)$ and $q_{n_k}(s) \to q(s) \in \partial\varphi(x')$. In particular,

$$\begin{aligned}\lim_{k\to\infty} L_\epsilon(x_k; s) = \limsup_{k\to\infty} L_\epsilon(x_{n_k}; s) &\le \limsup_{k\to\infty} \|p_{n_k}(s) + q_{n_k}(s)\| + \frac{\delta}{2} \\ &= \|p(s) + q(s)\| + \frac{\delta}{2} \le L_\epsilon(x; s) + \frac{\delta}{2},\end{aligned}$$

which is a contradiction. Thus $L_\epsilon(\cdot; s)$ and $L_\epsilon^2(\cdot; s)$ are upper semicontinuous.

To see that $L_\epsilon(\cdot)$ is upper semicontinuous, we construct an integrable envelope for the function and then apply Fatou's lemma. Indeed, using the assumed $M_\epsilon(x; s)$-local Lipschitz continuity of $y \mapsto f(y; s)$ for y near x, we have

$$\|G(y; s)\| \le M_\epsilon(x; s) + \|\partial\varphi(y)\|$$

for y with $\|y - x\| \le \epsilon$. This quantity is integrable, and we may apply Fatou's lemma and Assumption A to obtain

$$\limsup_{y\to x} L_\epsilon(y) \le \mathbb{E}\left[\limsup_{y\to x} L_\epsilon(y)\right] \le \mathbb{E}[L_\epsilon(x; S)] \le \sqrt{\mathbb{E}[L_\epsilon(x; S)^2]}.$$

A.4. Proof of Observation 2. In the case that $\varphi(x) = \frac{\lambda}{2}\|x\|^2$, the stochastic update in Example 1 becomes $x_{k+1} = \frac{1}{1+\alpha_k\lambda} x_k - \frac{\alpha_k}{1+\alpha_k\lambda} g_k$, and we have the recursion

$$\|x_{k+1}\| \le \frac{\|x_k\|}{1+\alpha_k\lambda} + \frac{\alpha_k}{1+\alpha_k\lambda} L(S_k) \le \prod_{i=1}^k (1+\alpha_i\lambda)^{-1} \|x_1\| + \sum_{i=1}^k \alpha_i L(S_i) \prod_{j=i}^k (1+\alpha_j\lambda)^{-1}.$$

Let $L_i = L(S_i)$ for shorthand and define $\xi_i = L_i - \mathbb{E}[L(S_i)]$, noting ξ_i are i.i.d. and mean zero. Assume that $\lambda = 1$ and $\mathbb{E}[L(S)^2] = 1$ w.l.o.g. Defining $Z_k = \sum_{i=1}^k \alpha_i L_i \prod_{j=i}^k (1+\alpha_j\lambda)^{-1}$, so that $Z_k - \mathbb{E}[Z_k] = \sum_{i=1}^k \alpha_i \xi_i \prod_{j=i}^k (1+\alpha_j\lambda)^{-1}$ and noting that

$$\mathbb{E}[Z_{k+1}] = \frac{\mathbb{E}[Z_k] + \alpha_{k+1}}{1+\alpha_{k+1}} = \begin{cases} \le \mathbb{E}[Z_k] & \text{if } \mathbb{E}[Z_k] > 1, \\ \le 1 & \text{if } \mathbb{E}[Z_k] \le 1, \end{cases}$$

we have that $\sup_k \mathbb{E}[Z_k] < \infty$. Moreover, if we let $M_k = \sum_{i=1}^k \alpha_i \prod_{j=1}^{i-1}(1+\alpha_j)\xi_i$, then $Z_k - \mathbb{E}[Z_k] = \prod_{j=1}^k (1+\alpha_j)^{-1} M_k$, and M_k is a martingale adapted to the filtration $\mathcal{F}_k = \sigma(S_1, \ldots, S_k)$. Noting that $M_{k+1} - M_k = \alpha_{k+1} \prod_{j=1}^k (1+\alpha_j)\xi_{k+1}$, we have

$$\sum_{k=1}^\infty \frac{1}{\prod_{j=1}^k (1+\alpha_j)^2} \mathbb{E}[(M_{k+1} - M_k)^2 \mid \mathcal{F}_k] = \sum_{k=1}^\infty \alpha_{k+1}^2 \mathbb{E}[\xi_{k+1}^2] < \infty.$$

Applying standard L_2-martingale convergence results (e.g., [16, Exer. 5.3.35]) gives that $M_k / \prod_{j=1}^k (1+\alpha_j) \stackrel{a.s.}{\to} 0$, and thus $Z_k \stackrel{a.s.}{\to} \mathbb{E}[Z_k]$, while certainly $\limsup_k \|x_k\| \le \limsup_k Z_k$.

REFERENCES

[1] J.-P. Aubin and A. Cellina, *Differential Inclusions: Set-Valued Maps and Viability Theory*, Springer-Verlag, Berlin, 1984.

[2] M. Benaïm, J. Hofbauer, and S. Sorin, *Stochastic approximations and differential inclusions*, SIAM J. Control Optim., 44 (2005), pp. 328–348.

[3] D. P. Bertsekas, *Stochastic optimization problems with nondifferentiable cost functionals*, J. Optim. Theory Appl., 12 (1973), pp. 218–231.

[4] J. Bolte, A. Daniilidis, A. S. Lewis, and M. Shiota, *Clarke subgradients of stratifiable functions*, SIAM J. Optim., 18 (2007), pp. 556–572.

[5] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet, *Characterizations of Łojasiewicz inequalities: Subgradient flows, talweg, convexity*, Trans. Amer. Math. Soc., 362 (2010), pp. 3319–3363.

[6] V. Borkar, *Stochastic Approximation*, Cambridge University Press, Cambridge, UK, 2008.

[7] J. Burke, *Descent methods for composite nondifferentiable optimization problems*, Math. Program., 33 (1985), pp. 260–279.

[8] E. Candès, T. Strohmer, and V. Voroninski, *PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming*, Comm. Pure Appl. Math., 66 (8), 2013.

[9] E. J. Candès, X. Li, and M. Soltanolkotabi, *Phase retrieval via Wirtinger flow: Theory and algorithms*, IEEE Trans. Inform. Theory, 61 (2015), pp. 1985–2007.

[10] Y. Chen and E. Candès, *Solving random quadratic systems of equations is nearly as easy as solving linear systems*, Comm. Pure Appl. Math., 2015, to appear.

[11] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust Region Methods*, MOS-SIAM Ser. Optim., Philadelphia, 2000.

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

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

[14] D. Davis and B. Grimmer, *Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems*, preprint, arXiv:1707.03505 [cs.LG], 2017.

[15] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee, *Stochastic subgradient method converges on tame functions*, preprint, arXiv:1804.07795 [cs.LG], 2018.

[16] A. Dembo, Probability Theory: STAT310/MATH230, lecture notes, Stanford University, Stanford, CA, 2016, http://statweb.stanford.edu/~adembo/stat-310b/lnotes.pdf.

[17] P. A. Dorofeyev, *Some properties of the generalized gradient method*, Comput. Math. Math. Phys., 25 (1985), pp. 117–122.

[18] D. Drusvyatskiy, *The proximal point method revisited*, preprint, arXiv:1712.06038 [math.OC], 2018, http://www.arXiv.org/abs/1712.06038.

[19] D. Drusvyatskiy and A. Lewis, *Error bounds, quadratic growth, and linear convergence of proximal methods*, Math. Oper. Res., 43 (2018), pp. 919–948.

[20] D. Drusvyatskiy and C. Paquette, *Efficiency of minimizing compositions of convex functions and smooth maps*, preprint, arXiv:1605.00125 [math.OC], 2016.

[21] D. Drusvyatskiy, A. Ioffe, and A. Lewis, *Nonsmooth optimization using Taylor-like models: Error bounds, convergence, and termination criteria*, preprint, arXiv:1610.03446 [math.OS], 2016.

[22] J. C. Duchi and F. Ruan, *Stochastic methods for composite and weakly convex optimization problems*, preprint, arXiv:1703.08570 [math.OC], 2017.

[23] J. C. Duchi and F. Ruan. *Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval*, Inf. Inference, 2018, to appear.

[24] J. C. Duchi, E. Hazan, and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, J. Mach. Learning Res., 12 (2011), pp. 2121–2159.

[25] Y. M. Ermoliev, *On the stochastic quasi-gradient method and stochastic quasi-Feyer sequences*, Kibernetika, 2 (1969), pp. 72–83.

[26] Y. M. Ermoliev and V. Norkin, *Stochastic generalized gradient method for nonconvex nonsmooth stochastic optimization*, Cybernet. Systems Anal., 34 (1998), pp. 196–215.

[27] Y. M. Ermoliev and V. Norkin, *Solution of nonconvex nonsmooth stochastic optimization problems*, Cybernet. Systems Anal., 39 (2003), pp. 701–715.

[28] R. Fletcher, *A model algorithm for composite nondifferentiable optimization problems*, in Nondifferential and Variational Techniques in Optimization, D. C. Sorensen and R. J.-B. Wets, eds., Math. Program. Studies 17, Springer, Berlin, 1982, pp. 67–76.

[29] R. Fletcher and G. A. Watson, *First and second order conditions for a class of nondifferentiable optimization problems*, Math. Program., 18 (1980), pp. 291–307.

[30] A. M. Gupal, *Stochastic Methods for Solving Nonsmooth Extremal Problems*, Naukova Dumka, Kiev, 1979 (In Ukrainian).
[31] M. Hardt and T. Ma, *Identity matters in deep learning*, in Proceedings of the Fifth International Conference on Learning Representations, 2017.
[32] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, New York, 2009.
[33] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I & II, Springer, New York, 1993.
[34] A. D. Ioffe, *Critical values of set-valued maps with stratifiable graphs. extensions of Sard and Smale-Sard theorems*, Proc. Amer. Math. Soc., 136 (2008), pp. 3111–3119.
[35] A. D. Ioffe, *Variational Analysis of Regular Mappings*, Springer, New York, 2017.
[36] M. Kunze, *Non-Smooth Dynamical Systems*, Lect. Notes Math. 1744, Springer, New York, 2000.
[37] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed., Springer, New York, 2003.
[38] A. S. Lewis and S. J. Wright, *A proximal method for composite minimization*, preprint, arXiv:0812.0423 [math.OC], 2018.
[39] L. Ljung, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, 22 (1977), pp. 551–575.
[40] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.
[41] Y. Nesterov, *Introductory Lectures on Convex Optimization*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 2004.
[42] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 2006.
[43] R. Poliquin and R. T. Rockafellar, *Prox-regular functions in variational analysis*, Trans. AMS, 348 (1996), pp. 1805–1838.
[44] H. Robbins and D. Siegmund, *A convergence theorem for non-negative almost supermartingales and some applications*, in Optimizing Methods in Statistics, Academic Press, New York, 1971, pp. 233–257.
[45] R. T. Rockafellar, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.
[46] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[47] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Springer, New York, 1998.
[48] A. Ruszczyński, *A linearization method for nonsmooth stochastic programming problems*, Math. Oper. Res., 12 (1987), pp. 32–49.
[49] Y. Schechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, *Phase retrieval with application to optical imaging*, IEEE Signal Process. Magazine, 2015, pp. 87–109.
[50] W. Su, S. Boyd, and E. Candes, *A differential equation for modeling nesterovs accelerated gradient method: Theory and insights*, in Advances in Neural Information Processing Systems 27, 2014, pp. 2510–2518.
[51] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 58 (1996), pp. 267–288.
[52] M. Udell, K. Mohan, D. Zeng, J. Hong, S. Diamond, and S. Boyd, *Convex optimization in Julia*, in First Workshop on High Performance Technical Computing in Dynamic Languages, IEEE, Piscataway, NJ, 2014, pp. 18–28.
[53] M. Wang, J. Liu, and E. Fang, *Accelerating stochastic composition optimization*, in Advances in Neural Information Processing Systems 29, 2016, pp. 1714–1722.
[54] M. Wang, E. Fang, and H. Liu, *Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions*, Math. Program. Ser. A, 161 (2017), pp. 419–449.
[55] H. Whitney, *A function not constant on a connected set of critical points*, Duke Math. J., 1 (1935), pp. 514–517.
[56] A. Wibisono, A. Wilson, and M. I. Jordan, *A variational perspective on accelerated methods in optimization*, Proc. Natl. Acad. Sci. USA, 113 (2016), pp. 7351–7358.