SIAM J. OPTIM.
Vol. 29, No. 3, pp. 2257–2290

STOCHASTIC (APPROXIMATE) PROXIMAL POINT METHODS: CONVERGENCE, OPTIMALITY, AND ADAPTIVITY*

HILAL ASI† AND JOHN C. DUCHI‡

Abstract. We develop model-based methods for solving stochastic convex optimization problems, introducing the approximate-proximal point, or APROX, family, which includes the stochastic subgradient, proximal point, and bundle methods. When the modeling approaches we propose are appropriately accurate, the methods enjoy stronger convergence and robustness guarantees than classical approaches, even though the model-based methods typically add little to no computational overhead over stochastic subgradient methods. For example, we show that improved models converge with probability 1 and enjoy optimal asymptotic normality results under weak assumptions; these methods are also adaptive to a natural class of what we term easy optimization problems, achieving linear convergence under appropriate strong growth conditions on the objective. Our substantial experimental investigation shows the advantages of more accurate modeling over standard subgradient methods across many smooth and nonsmooth optimization problems.

Key words. stochastic optimization, proximal methods, robustness, adaptivity

AMS subject classifications. 65K10, 90C06

DOI. 10.1137/18M1230323

1. Introduction.

In this paper, we develop and analyze a family of model-based methods, moving beyond naive stochastic gradient methods, for solving the stochastic convex optimization problem

$$\begin{aligned} &\text{minimize} \;\; F(x) = \mathbb{E}_P[f(x;S)] = \int_{\mathcal{S}} f(x;s)dP(s) \\ &\text{subject to} \;\; x \in \mathcal{X}. \end{aligned} \tag{1}$$

In problem (1), the set $\mathcal{S}$ is a sample space, and for each $s \in \mathcal{S}$, the function $f(\cdot; s) : \mathbb{R}^n \to \mathbb{R}$ is a closed convex function, subdifferentiable on the closed convex set $\mathcal{X} \subset \mathbb{R}^n$.

Stochastic minimization problems, in which an optimizer has access to samples S_i drawn independently and identically distributed from P and uses these samples to minimize F, have applications in numerous fields, including machine learning, statistical estimation, and simulation-based optimization [65, 23, 55]. The current accepted methodology for such problems is the stochastic (sub)gradient method [66, 65, 43, 10, 54], which Robbins and Monro [49] originally developed for smooth stochastic approximation problems, and which iterates as follows: beginning at an initial point x_1, iteratively draw $S_k \overset{\text{iid}}{\sim} P$ and update

$$x_{k+1} := x_k - \alpha_k g_k \quad \text{for some } g_k \in \partial f(x_k; S_k). \tag{2}$$

*Received by the editors December 4, 2018; accepted for publication (in revised form) May 29, 2019; published electronically September 5, 2019.
https://doi.org/10.1137/18M1230323
Funding: Research partially supported by NSF-CAREER Award 1553086, by ONR-YIP N00014-19-1-2288, and by the Sloan Foundation.
†Department of Electrical Engineering, Stanford University, Stanford, CA 94305 (asi@stanford.edu).
‡Department of Statistics and Electrical Engineering, Stanford University, Stanford, CA 94305 (jduchi@stanford.edu).

The stochastic gradient method enjoys convergence guarantees [66, 43] and widespread empirical success in large-scale convex and nonconvex stochastic optimization [65, 10, 54, 17, 35]. In spite of this success, there are notable difficulties with the stochastic subgradient method (2): it is sensitive to step-size selection; it can diverge on objectives, such as $F(x) = x^4$, that do not obey its convergence criteria; and it is rarely adaptive to nuanced aspects of problem difficulty. Engineers thus waste time and computation dealing with these issues and finding appropriate step sizes, which cascades into additional practical challenges.

An alternative to treating the stochastic gradient method (2) (SGM) as a noisy approximation to gradient descent is to view it as minimizing a sequence of random *models* of the functions F and f, and we leverage this view here. In this context, SGM makes a linear approximation to the instantaneous function f around the point x_k, setting

$$f_{x_k}(x; S_k) := f(x_k; S_k) + \langle g_k, x - x_k \rangle$$

and choosing x_{k+1} to minimize the regularized model $f_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2$. More sophisticated models are plausible. Most familiar is the stochastic proximal point method [51, 31, 8, 29, 9] (the least-mean-squares algorithm [63] for quadratic f), which makes no approximation, using $f_{x_k}(x; s) = f(x; s)$ and iterating

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ f(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\}. \tag{3}$$

This modeling perspective is important in nonstochastic optimization, where (for example) Newton, Gauss–Newton, bundle, and trust-region methods (see, e.g., [25, 26, 11, 44, 45]) explicitly build sequences of easier-to-minimize models while minimizing the global function F. A substantial body of work investigates this modeling perspective [12, 18], and recent work by Duchi and Ruan [19] and Davis and Drusvyatskiy [15] demonstrates convergence for appropriate models in weakly convex stochastic optimization, motivating our approach.

We show how to extend this modeling perspective to stochastic convex optimization problems, leveraging it to build a new family of algorithms for solving problem (1), which, in homage to the stochastic proximal point iteration (3), we call the aProx (approximate proximal point) algorithms, with substantially better theoretical guarantees and empirical performance than naive stochastic subgradient methods. The aProx algorithms iterate as follows: for $k = 1, 2, \ldots$, we draw a random $S_k \overset{\text{iid}}{\sim} P$, then update the iterate x_k by minimizing a regularized approximation to $f(\cdot; S_k)$, setting

$$x_{k+1} := \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ f_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\}. \tag{4}$$

The function $f_x(\cdot; s)$ is a *model* of $f(\cdot; s)$ at the point x, meaning that f_x satisfies the following conditions on its structure and local approximation properties for f:

(C.i) The function $y \mapsto f_x(y; s)$ is convex and subdifferentiable on $\mathcal{X}$.

(C.ii) The model f_x satisfies the equality $f_x(x; s) = f(x; s)$ and

$$f_x(y; s) \le f(y; s) \quad \text{for all } y.$$

By the first-order conditions for convexity, for any $g \in \partial_y f_x(y; s)|_{y=x}$ conditions (C.i) and (C.ii) imply $f(y; s) \ge f_x(y; s) \ge f_x(x; s) + \langle g, y - x \rangle = f(x; s) + \langle g, y - x \rangle$,

yielding the containment

$$\partial_y f_x(y; s)\big|_{y=x} \subset \partial_x f(x; s). \tag{5}$$

Davis and Drusvyatskiy [15] and Duchi and Ruan [19] consider similar modeling conditions, and they inspire our treatment here. See section 2 and Figure 1 for examples.

The aProx methodology (4) is flexible in that it allows many possible modeling choices. As we shall see, though stochastic gradient and proximal point methods (see (2) and (3), respectively) are both special cases, they possess quite different behaviors. Thus, it is interesting to provide conditions on the accuracy of the models f_x, in addition to (C.i) and (C.ii), that imply stronger convergence guarantees than those available for stochastic gradient and other simple methods. To describe our contributions at a high level, we list two assumptions that we frequently make. As $f(\cdot; s)$ is subdifferentiable on $\mathcal{X}$, there exist measurable selections $f'(x; s) \in \partial f(x; s)$, and $\mathbb{E}[\partial f(x; S)] = \partial F(x)$ for $x \in \mathcal{X}$ (cf. [7, section 2]).

Assumption A1. The set $\mathcal{X}^\star := \operatorname{argmin}_{x\in\mathcal{X}}\{F(x)\}$ is nonempty, and there exists $\sigma^2 < \infty$ such that $\mathbb{E}[\|f'(x^\star; S)\|_2^2] \le \sigma^2$ for $x^\star \in \mathcal{X}^\star$ and all measurable selections $f'(x^\star; s) \in \partial f(x^\star; s)$.

Assumption A2. There exists a nondecreasing function $\mathsf{G}_{\text{big}} : \mathbb{R}_+ \to [0, \infty)$ such that for all $x \in \mathcal{X}$ and measurable selections $f'(x; s) \in \partial f(x; s)$, $\mathbb{E}[\|f'(x; S)\|^2] \le \mathsf{G}_{\text{big}}(\operatorname{dist}(x, \mathcal{X}^\star))$.

Assumption A2 makes no restrictions on the growth of the function G_{big}, so the second moment of the subgradient $f'(x; S)$ may grow arbitrarily. This contrasts with typical assumptions for stochastic subgradient methods (see, e.g., [66, 43]), which assume uniform boundedness or second-moment conditions on subgradients. Within this context, we take a three-pronged approach.

1. First, in section 3, we develop conditions for the *stability* of iterates (4) under Assumption A1, meaning that they remain in a bounded neighborhood of the optimal solution set $\mathcal{X}^\star$ of problem (1). We leverage this stability to prove convergence for the aProx iteration (4) even for functions with substantial variation in their gradient estimates (e.g., the gradient may grow superexponentially in $\|x\|$) to which standard results do not apply. As a consequence, we extend Polyak and Juditsky's analysis of averaged stochastic gradient methods to all aProx models (4)—showing asymptotic normality with optimal covariance under weaker conditions than those necessary for classical situations—so long as the iterates are bounded, highlighting the importance of this stability.

2. Second, in section 4, we study the performance of aProx methods (4) for what we term *easy* problems. In these problems there exists a shared minimizer $x^\star$ common to all the sampled functions. This assumption is strong, yet many problems are easy: in statistical learning, Belkin, Rakhlin, and Tsybakov [5] and Belkin, Hsu, and Mitra [4] show that functions that *perfectly interpolate* the observed data (suffering zero loss on the observations) can achieve optimal statistical convergence; Kaczmarz algorithms solve consistent over-parameterized linear systems [56, 42, 41]; the problem of finding a point in the intersection of convex sets assumes there exists a point in each of them [36, 3]. By incorporating a simple lower bound condition—basically, that if f is nonnegative, any model f_x should also be nonnegative—in addition to conditions (C.i) and (C.ii), we show how aProx adapts to these easy problems and

achieves (near) linear convergence, even in stochastic settings, using methods with no additional computational complexity beyond stochastic gradient methods.

3. Finally, in section 5, we present representative nonasymptotic convergence guarantees. Any aProx method (4) recovers standard convergence guarantees of stochastic gradient and proximal point methods [43, 8] (see [15] for the basic convergence guarantee). We show how stochastic proximal point methods enjoy fast convergence under (restricted) strong convexity with only weak moment conditions on $\partial f(x; s)$, further emphasizing the advantages of accurate modeling.

In addition to our theoretical results, we perform substantial simulations. Our experiments consider a wide range of smooth, nonsmooth, and superpolynomially growing convex problems: regression with squared and absolute losses, logistic and Poisson regression, and projection problems onto intersections of half-spaces (relating these to classification problems). The common refrain in each of these is that even slightly improved aProx models are much more robust to step-size choice than stochastic gradient methods, and more careful modeling allows fast convergence in a much broader range of problems, including those with moderately poor conditioning where stochastic gradient methods fail.

1.1. Related work. We situate our paper in relation to classical and modern work on stochastic optimization problems. Stochastic gradient methods are classical, beginning with the development by Robbins and Monro [49] in the 1950s [47, 48, 66, 43, 65, 32]. A number of authors recognize the challenges associated with step-size selection and instability of stochastic gradient methods: in the case of smooth strongly convex minimization, Nemirovski et al. [43] show how a slightly misspecified step size can cause arbitrarily slow convergence guarantees. More recent work (e.g., [2]) shows that even when assumptions sufficient for convergence hold, stochastic gradient methods can exhibit transient divergent behavior.

In effort to alleviate some of these issues, there is recent work on more careful approaches to stochastic optimization problems. Of most relevance to our work are *stochastic proximal point methods* (3), which use the true function $f_x(y; s) = f(y; s)$ in the iteration (4). Bertsekas [8] analyzes stochastic proximal point algorithms in an incremental framework (when $\mathcal{S} = \{1, \ldots, m\}$ is a finite set), showing convergence results similar to subgradient methods, while Kulis and Bartlett [31] and Karampatziakis and Langford [29] give theoretical and empirical results in online convex optimization settings, demonstrating regret bounds similar to classical results [66]. Toulis and Airoldi [58] and Toulis, Tran, and Airoldi [59] study stochastic proximal point algorithms and convergence guarantees for their final iterates, a different approach to the one that we take. Their results, however, assume that the functions under consideration are both globally Lipschitz and globally strongly convex, a contradiction severely limiting the applicability of their results. (Their analysis explicitly and frequently uses both assumptions; under weaker assumptions, their convergence guarantees exhibit the same potential for exponential divergence of stochastic gradient methods.) Patrascu and Necoara [46] also analyze stochastic proximal point algorithms, providing nonasymptotic convergence results under the assumption that each function $f(\cdot; s)$ is Lipschitz or strongly convex [46, Assumptions 1 and 9]; these assumptions fail for many problems (including linear regression with $f(x; (a, b)) = \frac{1}{2}(\langle a, x\rangle - b)^2$), though their results also apply to intersections of sets $\mathcal{X} = \mathcal{X}_1 \cap \cdots \cap \mathcal{X}_m$.

Ryu and Boyd [52] also investigate the stochastic proximal point method, making arguments on its stability stronger than classical results for stochastic gradient methods. Under Assumption A1, Ryu and Boyd show that the stochastic proximal point

method guarantees $\mathbb{E}[\|x_{k+1} - x^\star\|_2] \leq \mathbb{E}[\|x_1 - x^\star\|_2] + \sigma \sum_{i=1}^k \alpha_i$, so that the iterates do not diverge exponentially. Yet this result is not enough to explain or provide empirical or theoretical stability, boundedness, or convergence guarantees.

Notation. For a convex function f, $\partial f(x)$ denotes its subgradient set at x, and $f'(x) \in \partial f(x)$ denotes an arbitrary element of the subdifferential. Throughout, $x^\star$ denotes a minimizer of problem (1) and $\mathcal{X}^\star = \operatorname{argmin}_{x \in \mathcal{X}} F(x)$ its optimal set. We let $\mathcal{F}_k := \sigma(S_1, \ldots, S_k)$ be the σ-field generated by the first k random variables S_i, so $x_k \in \mathcal{F}_{k-1}$ for all k under iteration (4).

2. Methods. We begin our contributions by introducing different natural models for stochastic convex optimization problems, as well as a few conditions in addition to (C.i) and (C.ii) that we can use to demonstrate new aspects of convergence for the aProx family. While the stochastic proximal point method (3) satisfies all the conditions in the paper, in some situations it may be expensive or challenging to implement exactly. With that in mind, we provide a catalogue of a few models to serve as a reference for the remainder of the paper.

Stochastic subgradient methods. The starting point for any model-based methods are the simple first-order models. As we discuss in the introduction, the stochastic subgradient method uses the model

$$f_x(y; s) := f(x; s) + \langle f'(x; s), y - x \rangle, \tag{6}$$

where $f'(x; s) \in \partial f(x; s)$ is an arbitrary element of the subdifferential. The linear model (6) satisfies conditions (C.i) and (C.ii) by convexity.

Proximal point methods. The stochastic proximal point method uses the "model"

$$f_x(y; s) := f(y; s), \tag{7}$$

that is, the true function. The model (7) satisfies all the conditions we provide.

Truncated models. The first condition beyond (C.i) and (C.ii) builds out of the simple observation that if one is minimizing a nonnegative function (for example, in most machine learning and statistical applications with a loss function), then a priori a model of the function that takes negative values cannot be accurate. If $f(x; s) \geq 0$ for all x, a better approximation to f than the linear model (6) is to take $f'(x; s) \in \partial f(x; s)$ and define

$$f_x(y; s) := \left[f(x; s) + \langle f'(x; s), y - x \rangle \right]_+ ,$$

where $f'(x; s) \in \partial f(x; s)$. More generally, we may consider models that provide a lower guarantee:

(C.iii) For all $s \in \mathcal{S}$, the models $f_x(\cdot; s)$ satisfy

$$f_x(y; s) \geq \inf_{z \in \mathcal{X}} f(z; s).$$

Thus, if we are given an oracle that, for each fixed $s \in \mathcal{S}$, can compute the minimal value $\inf_{z \in \mathcal{X}} f(z; s)$, we may consider the truncated models

$$f_x(y; s) := \max \left\{ f(x; s) + \langle f'(x; s), y - x \rangle, \inf_{z \in \mathcal{X}} f(z; s) \right\}. \tag{8}$$

See Figure 1(a) for an illustration of this model.

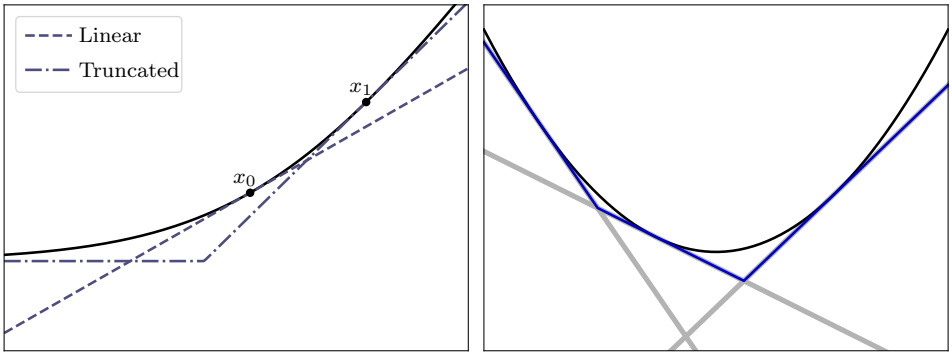

FIG. 1. (a) *Models of the function* $f(x) = \log(1+e^{-x})$*: a linear model* (6) *built around the point* x_0 *and truncated model* (8) *built around the point* x_1. (b) *The bundle model (maximum of linear functions), as in the iteration* (9). *The lighter lines represent individual linear approximations, the darker line their maximum.*

Many statistical, machine learning, and signal-processing examples support this model, because for any individual sample s we have $\inf_{z\in\mathcal{X}} f(z;s) = 0$. For example, in linear regression, $s = (a,b) \in \mathbb{R}^n \times \mathbb{R}$, and $\inf_z(\langle a,z\rangle - b)^2 = 0$ when $a \neq 0$. In logistic regression [23], we have $s = (a,b) \in \mathbb{R}^n \times \{-1,1\}$, and $f(x;(a,b)) = \log(1+\exp(-b\langle a,x\rangle))$ satisfies $\inf_z f(z;(a,b)) = 0$ unless $a = 0$. Support vector machines use $f(x;(a,b)) = [1-b\langle a,x\rangle]_+$, which again has infimal value 0. Even in more complex scenarios, these infimal values may be easy to compute; see, for example, our discussion of Poisson regression in section 3.3, Example 3.

Relatively accurate models. Now we consider an additional condition on accuracy, which allows less accurate models than the exact model (3); for example, this allows the bundle model, which approximates $f(\cdot;s)$ by the maximum of affine lower bounds, which we discuss later. As we will see in what follows (Theorem 3.2), this condition is sufficient for strong stability and convergence guarantees for any APROX method using the model-based updates (4). We require a bit more notation. Let $f_{x_0}(\cdot;s)$ be a model centered at x_0 satisfying conditions (C.i) and (C.ii). For $\alpha > 0$ define

$$x_\alpha := \operatorname*{argmin}_{x\in\mathcal{X}} \left\{ f_{x_0}(x;s) + \frac{1}{2\alpha}\|x-x_0\|_2^2 \right\},$$

which is the result of a single update (leaving dependence on s implicit). Then we consider the following:

(C.iv) For some $\epsilon > 0$, there exists a function $C : \mathcal{S} \to \mathbb{R}_+$ with $\mathbb{E}[C(S)] < \infty$ such that for all $x_0 \in \mathcal{X}$, the updated point x_α and model $f_{x_0}(\cdot;s)$ satisfy

$$f(x_\alpha;s) \le f_{x_0}(x_\alpha;s) + \frac{1-\epsilon}{2\alpha}\|x_\alpha - x_0\|_2^2 + C(s)\alpha.$$

The lower bound condition (C.ii) guarantees that $f(x;s) \ge f_{x_0}(x;s)$ for all x, x_0, so (C.iv) provides a complementary upper bound. It is clear that the full (proximal point) model (7) satisfies condition (C.iv) with $\epsilon = 1$ and $C(s) = 0$, as $f_x = f$. We term condition (C.iv) a "relative" accuracy because the necessary approximation scales with α so that higher accuracy is necessary as $\alpha \downarrow 0$, though $\|x_\alpha - x_0\| = O(\alpha)$ by standard results [18].

While aside from proximal-point models, it is not clear a priori how to guarantee that condition (C.iv) holds, one approach is to use bundle methods [26, 57], which are identical to Kelley's cutting plane method [30]. In this situation, we begin from the linear model $f_x^0(y;s) = f(x;s) + \langle f'(x;s), y - x\rangle$, and iteratively construct the lower piecewise-linear models

$$x_\alpha^i := \operatorname*{argmin}_{y\in\mathcal{X}} \left\{ f_x^{i-1}(y) + \frac{1}{2\alpha}\|y-x\|_2^2 \right\} \quad \text{and} \quad f_x^i(y) := \max\left\{ f_x^{i-1}(y;s), f(x_\alpha^i;s) + \langle f'(x_\alpha^i;s), y - x_\alpha^i\rangle \right\}. \tag{9}$$

Whenever the iterate x_α^i satisfies condition (C.iv), we may terminate the iteration, as $f_x^{i-1}(y)$ satisfies conditions (C.i) and (C.ii) by construction. While we do not address this in this paper, the number of iterations to achieve a solution satisfying condition (C.iv) is at most $O(\|f'(x_0;s)\|^2)$, as each step solves a strongly convex optimization problem (see [57, section 2.4]).

3. Stability and its consequences. The first of our main strands of approach, upon which we focus in this section, is the stability and boundedness of the iterates for stochastic proximal point methods and their relatives in the aProx family of methods. These stability guarantees are in strong contrast to standard stochastic subgradient methods, which may diverge for problems on which aProx methods converge.

We begin with our definition of stability. In this definition, we let $\mathcal{A}$ denote the set of positive step-size sequences $\{\alpha_k\}$ with $\sum_k \alpha_k^2 < \infty$. We call a pair $(\mathcal{F}, \mathcal{P})$ a *collection of problems* if $\mathcal{P}$ is a collection of probability measures on a sample space $\mathcal{S}$, and $\mathcal{F}$ is a collection of functions $f : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$, where $f(\cdot; s)$ is convex. We make the following definition.

Definition 3.1. *An algorithm generating iterates x_k according to the model-based update* (4) *is* stable in probability *for the collection of problems $(\mathcal{F}, \mathcal{P})$ if for all $f \in \mathcal{F}$ and $P \in \mathcal{P}$ defining $F(x) = \mathbb{E}_P[f(x;S)]$ and $\mathcal{X}^\star = \operatorname{argmin}_{x\in\mathcal{X}} F(x)$, and for all step-size sequences $\{\alpha_k\} \in \mathcal{A}$,*

$$\sup_k \operatorname{dist}(x_k, \mathcal{X}^\star) < \infty \quad \textit{with probability } 1. \tag{10}$$

Classical results coupled with the Robbins–Siegmund supermartingale convergence theorem [50] guarantee that the stochastic subgradient method has bounded iterates, $\sup_k \operatorname{dist}(x_k, \mathcal{X}^\star) < \infty$, whenever

$$\mathbb{E}\left[\|f'(x;S)\|_2^2\right] \le C_0 + C_1 \operatorname{dist}(x, \mathcal{X}^\star)^2 \quad \text{for all } x \in \mathcal{X},$$

which typical smoothness or boundedness conditions imply (cf. [50, 48, 6]). Stochastic (approximate) proximal point approaches allow us to move beyond these quadratic growth assumptions; in contrast, for objectives for which the gradients grow more than quadratically, condition (10) typically fails for gradient methods.

Example 1 (divergence for nonquadratics). Let $F(x) = \frac{1}{4}x^4$, consider any sequence of step sizes $\alpha_k > 0$ satisfying $\alpha_{k+1} \ge \frac{1}{4}\alpha_k$ for all k, and let $x_{k+1} = x_k - \alpha_k F'(x_k)$ be generated by the gradient method. Then whenever the initial iterate x_1 satisfies $|x_1| \ge \sqrt{3/\alpha_1}$, we have $|x_k| \ge 2|x_{k-1}|$ for all k, so $|x_k| \ge 2^k|x_1|$ for all $k \in \mathbb{N}$. ∎

When the objective grows faster than polynomially, even worse behavior is possible; for example, for the objective $F(x) = (e^x + e^{-x})$, for any polynomially decreasing

step-size sequence, if x_1 is large enough we have the superexponential divergence $|x_k| \geq 2^{2^k}|x_1|$. Standard stochastic gradient methods may also exhibit undesirable behavior for easier problems; even in situations for which the objective is smooth and in which there is no noise, (sub)gradient methods may suffer transient exponential growth, as the following example demonstrates.

Example 2 (instability for quadratics). Let $F(x) = \frac{1}{2}x^2$. Then the gradient method iterates as $x_{k+1} = (1 - \alpha_k)x_k$. Let us assume that $\alpha_k = \alpha_0 k^{-\beta}$, and let $\alpha_0 \geq 3K^\beta$ for some $K \in \mathbb{N}$. Then assuming $x_1 \neq 0$, for all $k \leq K$ we have $|x_{k+1}| = |1 - \alpha_k||x_k| \geq 2|x_k|$, so that $|x_{k+1}| \geq 2^k$ for $k \leq K$. Classical guarantees for the (stochastic) gradient method show that x_k will converge eventually, but even for smooth quadratics, gradient descent suffers from exponential growth behavior if the step size is misspecified. ∎

While stylized, these examples highlight the difficulty of naive application of gradient methods; in what follows, we show how accurate models alleviate the issues in the examples.

3.1. Stability of (approximate) proximal methods. The starting point of almost all of what follows are sufficient conditions on the models we use to guarantee stability as in Definition 3.1. For this first result, in addition to the two conditions (C.i) and (C.ii), we assume the models are accurate at their updated points, that is, that condition (C.iv) holds. The full model (stochastic proximal point method) satisfies these conditions, and so too do bundle models (9). In the theorem, recall the σ-field $\mathcal{F}_k := \sigma(S_1, \ldots, S_k)$.

THEOREM 3.2. *Let Assumption* A1 *hold and x_k be generated by the iteration* (4) *with any model satisfying conditions* (C.i), (C.ii), *and* (C.iv). *Then for all $x^\star \in \mathcal{X}^\star$,*

$$\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2 \mid \mathcal{F}_{k-1}\right] \leq \|x_k - x^\star\|_2^2 + \alpha_k^2 \left(\frac{\sigma^2}{\epsilon} + \mathbb{E}[C(S)]\right).$$

Before providing the proof of the theorem (see section 3.1.1), we present a few of its consequences for stability. By taking $x^\star$ to be the projection of x_k onto $\mathcal{X}^\star$, Theorem 3.2 implies

$$\mathbb{E}[\operatorname{dist}(x_{k+1}, \mathcal{X}^\star)^2 \mid \mathcal{F}_{k-1}] \leq \operatorname{dist}(x_k, \mathcal{X}^\star)^2 + \alpha_k^2 \left(\frac{\sigma^2}{\epsilon} + \mathbb{E}[C(S)]\right).$$

A few somewhat more consequential corollaries, at least from the perspective of our stability definition (Definition 3.1), follow. We first have that the expected distance is nondivergent.

COROLLARY 3.3. *Let the conditions of Theorem* 3.2 *hold. For each $k \in \mathbb{N}$,*

$$\mathbb{E}\left[\operatorname{dist}(x_{k+1}, \mathcal{X}^\star)^2\right] \leq \mathbb{E}\left[\operatorname{dist}(x_1, \mathcal{X}^\star)^2\right] + \left(\frac{\sigma^2}{\epsilon} + \mathbb{E}[C(S)]\right) \sum_{i=1}^k \alpha_i^2.$$

A second corollary establishes that the iterates of appropriately accurate APROX methods are stable. We require the Robbins–Siegmund almost supermartingale convergence lemma.

LEMMA 3.4 (see [50]). *Let $A_k, B_k, C_k, D_k \geq 0$ be nonnegative random variables adapted to the filtration $\mathcal{F}_k$ and satisfying $\mathbb{E}[A_{k+1} \mid \mathcal{F}_k] \leq (1 + B_k)A_k + C_k - D_k$.*

Then on the event $\{\sum_k B_k < \infty, \sum_k C_k < \infty\}$, *there is a random* $A_\infty < \infty$ *such that* $A_k \stackrel{a.s.}{\to} A_\infty$ *and* $\sum_k D_k < \infty$.

By applying Theorem 3.2 with $A_k = \operatorname{dist}(x_{k+1}, \mathcal{X}^\star)^2$, $C_k = \alpha_{k+1}^2(\sigma^2/\epsilon + \mathbb{E}[C(S)])$, and $B_k = D_k = 0$ in Lemma 3.4, we have the following corollary.

COROLLARY 3.5. *Let the conditions of Theorem* 3.2 *hold and assume* $\sum_k \alpha_k^2 < \infty$. *Then*

$$\sup_{k \in \mathbb{N}} \operatorname{dist}(x_k, \mathcal{X}^\star) < \infty$$

and $\operatorname{dist}(x_k, \mathcal{X}^\star)$ *converges to some finite value with probability* 1.

Combining Corollaries 3.3 and 3.5, we see that the stochastic proximal point method and its APROX variants (as long as they satisfy the accuracy condition (C.iv)) are stable according to Definition 3.1. This is in strong contrast to stochastic gradient methods and their relatives, which can be unstable even for relatively simple problems.

3.1.1. Proof of Theorem 3.2. We now return to the promised proof of Theorem 3.2. In giving the proof, we present lemmas on the progress of individual iterates for subsequent use. These results are typical of Lyapunov-type arguments for convergence of stochastic gradient methods [66, 43].

LEMMA 3.6. *Let* h *be convex and subdifferentiable on a closed convex set* $\mathcal{X}$ *and let* $\beta > 0$. *Then for all* $x_0, x_1, y \in \mathcal{X}$, *and* $h'(y) \in \partial h(y)$,

$$h(y) - h(x_1) \le \langle h'(y), y - x_0 \rangle + \frac{1}{2\beta} \|x_1 - x_0\|^2 + \frac{\beta}{2} \|h'(y)\|^2 .$$

Proof. By the first-order conditions for convexity, we have

$$\begin{aligned} h(y) - h(x_1) &\le \langle h'(y), y - x_1 \rangle = \langle h'(y), y - x_0 \rangle + \langle h'(y), x_0 - x_1 \rangle \\ &\le \langle h'(y), y - x_0 \rangle + \frac{1}{2\beta} \|x_1 - x_0\|^2 + \frac{\beta}{2} \|h'(y)\|^2 , \end{aligned}$$

where the second line uses Young's inequality. ∎

We also have the following lemma, which gives a one-step progress guarantee for any algorithm using models satisfying conditions (C.i) and (C.ii).

LEMMA 3.7. *Let condition* (C.i) *hold. In each step of the method* (4), *for any* $x \in \mathcal{X}$,

$$\frac{1}{2} \|x_{k+1} - x\|_2^2 \le \frac{1}{2} \|x_k - x\|_2^2 - \alpha_k \left[f_{x_k}(x_{k+1}; S_k) - f_{x_k}(x; S_k) \right] - \frac{1}{2} \|x_k - x_{k+1}\|_2^2 .$$

Proof. By the first-order conditions for convex optimization, we have that

$$\langle \alpha_k g_k + (x_{k+1} - x_k), y - x_{k+1} \rangle \ge 0$$

for all $y \in \mathcal{X}$ and some $g_k \in \partial f_{x_k}(x_{k+1}; S_k)$. Setting $y = x$, we obtain

$$\begin{aligned} \alpha_k \langle g_k, x_{k+1} - x \rangle &\le \langle x_{k+1} - x_k, x - x_{k+1} \rangle \\ &= \frac{1}{2} \left[\|x_k - x\|_2^2 - \|x_{k+1} - x\|_2^2 - \|x_{k+1} - x_k\|_2^2 \right] . \end{aligned}$$

As $f_{x_k}(x; S_k) \ge f_{x_k}(x_{k+1}; S_k) + \langle g_k, x - x_{k+1} \rangle$ by condition (C.i), this gives the result. ∎

With Lemmas 3.6 and 3.7 in place, we can prove the theorem. Let $x^\star \in \mathcal{X}^\star$ be an otherwise arbitrary optimal point. Applying Lemma 3.7 with $x = x^\star$, we have

$$\begin{aligned}
&\frac{1}{2}\left\|x_{k+1} - x^\star\right\|_2^2 \\
&\leq \frac{1}{2}\left\|x_k - x^\star\right\|_2^2 - \alpha_k\left[f_{x_k}(x_{k+1}; S_k) - f_{x_k}(x^\star; S_k)\right] - \frac{1}{2}\left\|x_k - x_{k+1}\right\|_2^2 \\
&\overset{\text{(i)}}{\leq} \frac{1}{2}\left\|x_k - x^\star\right\|_2^2 - \alpha_k\left[f(x_{k+1}; S_k) - f_{x_k}(x^\star; S_k)\right] - \frac{\epsilon}{2}\left\|x_k - x_{k+1}\right\|_2^2 + C(S_k)\alpha_k^2 \\
&\overset{\text{(ii)}}{\leq} \frac{1}{2}\left\|x_k - x^\star\right\|_2^2 - \alpha_k\left[f(x_{k+1}; S_k) - f(x^\star; S_k)\right] - \frac{\epsilon}{2}\left\|x_k - x_{k+1}\right\|_2^2 + C(S_k)\alpha_k^2,
\end{aligned}$$

where inequality (i) is a consequence of the accurate model condition (C.iv) and (ii) holds because $f_x(x^\star; s) \leq f(x^\star; s)$ by the lower model condition (C.ii). Now, we apply Lemma 3.6 with $x_1 = x_{k+1}$, $x_0 = x_k$, $y = x^\star$, and $\beta = \frac{\alpha_k}{\epsilon}$ to find

$$\begin{aligned}
&\frac{1}{2}\left\|x_{k+1} - x^\star\right\|_2^2 \\
&\leq \frac{1}{2}\left\|x_k - x^\star\right\|_2^2 + \alpha_k\langle f'(x^\star; S_k), x^\star - x_k\rangle + \frac{\alpha_k^2}{2\epsilon}\left\|f'(x^\star; S_k)\right\|_2^2 + C(S_k)\alpha_k^2
\end{aligned}$$

for all $f'(x^\star; S_k) \in \partial f(x^\star; S_k)$.

For some $F'(x^\star) \in \partial F(x^\star)$, we have $\langle F'(x^\star), y - x^\star\rangle \geq 0$ for all $y \in \mathcal{X}$. As our choice of $f'(x^\star; s) \in \partial f(x^\star; s)$ above was arbitrary, we may take $f'(x^\star; S_k)$ so that $\mathbb{E}[f'(x^\star; S_k)] = F'(x^\star)$ for any desired $F'(x^\star) \in \partial F(x^\star)$ (cf. [7]). Thus, taking expectations with respect to $\mathcal{F}_{k-1}$,

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}[\left\|x_{k+1} - x^\star\right\|_2^2 \mid \mathcal{F}_{k-1}] \\
&\leq \frac{1}{2}\left\|x_k - x^\star\right\|_2^2 + \frac{\alpha_k^2}{2\epsilon}\mathbb{E}\left[\left\|f'(x^\star; S)\right\|_2^2\right] + \mathbb{E}[C(S)]\alpha_k^2 + \alpha_k\langle F'(x^\star), x^\star - x_k\rangle.
\end{aligned}$$

As $\langle F'(x^\star), x^\star - x_k\rangle \leq 0$, we obtain the theorem.

3.2. Convergence of aProx methods. The key consequence of Theorem 3.2 is that the iterates x_k are stable in probability (see (10)), remaining bounded with probability 1. Iterate boundedness of algorithms does not guarantee convergence in general; however, in this section, we show that as a consequence of this boundedness, any algorithm satisfying conditions (C.i) and (C.ii) is convergent. Assumptions A1 and A2 are insufficient to guarantee convergence of subgradient methods, which—as our examples show—may diverge without uniform boundedness conditions on the subgradients. In contrast, any aProx method that is stable in probability (see (10)) guarantees convergence with probability 1. Throughout this section and section 3.3, we without comment make the assumption that the step sizes α_k satisfy the standard summability conditions

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

Proposition 3.8. *Let Assumptions* A1 *and* A2 *hold. Let the iterates x_k be generated by any method satisfying conditions* (C.i), (C.ii), *and $F^\star = \inf_{x\in\mathcal{X}} F(x)$. On the event that $\sup_k \operatorname{dist}(x_k, \mathcal{X}^\star) < \infty$, with probability* 1 *both $\sum_k \alpha_k(F(x_k) - F^\star) < \infty$*

and the iterates x_k are convergent: there exists $x^\star \in \mathcal{X}^\star$ such that $\|x_k - x^\star\| \stackrel{a.s.}{\to} 0$ and $F(x_k) \stackrel{a.s.}{\to} F(x^\star)$.

Proposition 3.8 implies an asymptotic convergence rate on (weighted averages of) the iterates x_k. Indeed, let $\{\gamma_k\}_{k=1}^\infty \subset \mathbb{R}_+$ be a nondecreasing sequence with $\gamma_k > 0$, and $\gamma_k \to \infty$. Defining the weighted averages $\overline{x}_k = \sum_{i=1}^k \gamma_i \alpha_i x_i / (\sum_{i=1}^k \gamma_i \alpha_i)$, we have the following corollary.

COROLLARY 3.9. *Let the conditions of Proposition* 3.8 *hold. Then with probability* 1*,*

$$\lim_{k\to\infty} \frac{1}{\gamma_k} \left(\sum_{i=1}^k \gamma_i \alpha_i \right) [F(\overline{x}_k) - F^\star] = 0.$$

Proof. We have $(\sum_{i=1}^k \gamma_i \alpha_i)(F(\overline{x}_k) - F^\star) \le \sum_{i=1}^k \gamma_i \alpha_i (F(x_i) - F^\star)$ by Jensen's inequality. Kronecker's lemma gives the result. ∎

For example, taking $\gamma_k = \alpha_k^{-1}$, we obtain that the average $\overline{x}_k = \frac{1}{k} \sum_{i=1}^k x_i$ satisfies

$$k\alpha_k \left(F(\overline{x}_k) - F^\star\right) \stackrel{\text{a.s.}}{\to} 0.$$

To prove Proposition 3.8, we present a lemma giving a one-step progress guarantee for any method satisfying conditions (C.i) and (C.ii).

LEMMA 3.10. *Let conditions* (C.i) *and* (C.ii) *hold and let x_k be generated by the updates* (4)*. Then for any $x \in \mathcal{X}$,*

$$\frac{1}{2} \|x_{k+1} - x\|_2^2 \le \frac{1}{2} \|x_k - x\|_2^2 - \alpha_k [f(x_k; S_k) - f(x; S_k)] + \frac{\alpha_k^2}{2} \|f'(x_k; S_k)\|_2^2 .$$

Proof. Using Lemma 3.7, it suffices to show that for any $\alpha > 0$ and $x_0, x_1, x \in \mathcal{X}$,

$$-\alpha [f_{x_0}(x_1; s) - f_{x_0}(x; s)] - \frac{1}{2} \|x_1 - x_0\|_2^2 \le -\alpha [f(x_0; s) - f(x; s)] + \frac{\alpha^2}{2} \|f'(x_0; s)\|_2^2 .$$

To see this, recall that conditions (C.i) and (C.ii) imply the containment (5), which in turn implies

$$\begin{aligned} -f_{x_0}(x_1; s) + f_{x_0}(x; s) &= -[f_{x_0}(x_0; s) - f_{x_0}(x; s)] + f_{x_0}(x_0; s) - f_{x_0}(x_1; s) \\ &\le -[f_{x_0}(x_0; s) - f_{x_0}(x; s)] + \langle f'(x_0; s), x_0 - x_1 \rangle \\ &\stackrel{\text{(C.ii)}}{\le} -[f(x_0; s) - f(x; s)] + \langle f'(x_0; s), x_0 - x_1 \rangle. \end{aligned}$$

Then we use that for any vector v, $\alpha \langle v, \Delta \rangle - \frac{1}{2} \|\Delta\|_2^2 \le \frac{\alpha^2}{2} \|v\|_2^2$, which gives the result. ∎

Proof of Proposition 3.8. By Assumption A2, $\mathbb{E}[\|f'(x; S)\|^2] \le \mathsf{G}_{\text{big}}(r)$ for all x such that $\operatorname{dist}(x, \mathcal{X}^\star) \le r$. Lemma 3.10 implies that for any $x^\star \in \mathcal{X}^\star$,

$$\mathbb{E}[\|x_{k+1} - x^\star\|^2 \mid \mathcal{F}_{k-1}] \le \|x_k - x^\star\|^2 - 2\alpha_k (F(x_k) - F^\star) + \alpha_k^2 \mathsf{G}_{\text{big}}(\operatorname{dist}(x_k, \mathcal{X}^\star)). \tag{11}$$

On the event that $\sup_k \operatorname{dist}(x_k, \mathcal{X}^\star) < \infty$, we have $\sum_k \alpha_k^2 \mathsf{G}_{\text{big}}(\operatorname{dist}(x_k, \mathcal{X}^\star)) < \infty$, and so the Robbins–Siegmund lemma (Lemma 3.4) implies that for any $x^\star \in \mathcal{X}^\star$ there is

some (random) $V(x^\star) < \infty$ such that $\|x_k - x^\star\| \stackrel{\text{a.s.}}{\to} V(x^\star)$, and

$$\sum_k \alpha_k (F(x_k) - F(x^\star)) < \infty.$$

We now show that $\operatorname{dist}(x_k, \mathcal{X}^\star) \stackrel{\text{a.s.}}{\to} 0$. Letting $x^\star$ be the projection of x_k onto $\mathcal{X}^\star$, inequality (11) gives $\mathbb{E}[\operatorname{dist}(x_{k+1}, \mathcal{X}^\star)^2 \mid \mathcal{F}_{k-1}] \le \operatorname{dist}(x_k, \mathcal{X}^\star)^2 - 2\alpha_k(F(x_k) - F^\star) + \alpha_k^2 \mathsf{G}_{\text{big}}(\operatorname{dist}(x_k, \mathcal{X}^\star))$. Thus using Lemma 3.4, there exists a random $D_\infty < \infty$ such that $\operatorname{dist}(x_k, \mathcal{X}^\star) \stackrel{\text{a.s.}}{\to} D_\infty$. For $\epsilon_i \in \mathbb{R}_+$, define the gap function

$$\Gamma(x^\star, \epsilon_1, \epsilon_2) := \inf_{x \in \mathcal{X}} \left\{F(x) - F^\star \mid \epsilon_1 \le \|x - x^\star\| \le 4\epsilon_1, \operatorname{dist}(x, \mathcal{X}^\star) \ge \epsilon_2\right\}.$$

The set $\{x \in \mathcal{X} \mid \|x - x^\star\| \in [\epsilon_1, 4\epsilon_1],\ \operatorname{dist}(x, \mathcal{X}^\star) \ge \epsilon_2\}$ is compact, so the infimum in the definition of the gap Γ is attained and $\Gamma(x^\star, \epsilon_1, \epsilon_2) > 0$ for $\epsilon_i > 0$. When the a.s. limits $V(x^\star)$ and D_∞ are positive, there exists a (random) $K < \infty$ such that $\|x_k - x^\star\| \in [V(x^\star)/2, 2V(x^\star)]$ and $\operatorname{dist}(x_k, \mathcal{X}^\star) \ge D_\infty/2$ for all $k \ge K$. Thus with probability 1,

$$\infty > \sum_k \alpha_k (F(x_k) - F^\star) \ge \sum_{k \ge K} \alpha_k \Gamma(x^\star, V(x^\star)/2, D_\infty/2).$$

As $\sum_k \alpha_k = \infty$ and $\Gamma(x^\star, \epsilon_1, \epsilon_2) > 0$ for $\epsilon_i > 0$, we have $D_\infty = 0$ with probability 1.

Finally, we show the sequence x_k converges. We begin by noting that $V(x^\star) = \lim_k \|x_k - x^\star\|$ is 1-Lipschitz. Indeed, let $x_1^\star, x_2^\star \in \mathcal{X}^\star$. By the a.s. limit definition of V,

$$|V(x_1^\star) - V(x_2^\star)| \le \limsup_k \left| \|x_k - x_1^\star\| - \|x_k - x_2^\star\| \right| \le \|x_1^\star - x_2^\star\|.$$

Let us now show that for some $x^\star \in \mathcal{X}^\star$, we have $V(x^\star) = 0$. Let

$$\mathbb{B} = \{x \in \mathbb{R}^n \mid \|x\| \le 1\}$$

be the ball. As $V(x^\star) < \infty$, there exists a (random) $r < \infty$ such that $x_k \in r\mathbb{B}$ for all k. As $\operatorname{dist}(x_k, \mathcal{X}^\star) \stackrel{\text{a.s.}}{\to} 0$, the compactness of $\mathbb{B}$ implies that $r\mathbb{B} \cap \mathcal{X}^\star \ne \emptyset$. Let $x^\star \in r\mathbb{B} \cap \mathcal{X}^\star$. For any $x \in r\mathbb{B}$, the projection $\pi(x)$ of x onto $\mathcal{X}^\star$ satisfies $\|\pi(x) - x\| \le \|x^\star - x\| \le 2r$, and so $\|\pi(x)\| \le 3r$, and defining the set $C = 3r\mathbb{B}$, we have $\operatorname{dist}(x_k, \mathcal{X}^\star) = \operatorname{dist}(x_k, \mathcal{X}^\star \cap C)$ for all k. Now, fix $\epsilon > 0$, and let $\{x_i^\star\}_{i=1}^N$ be an ϵ-net of $C \cap \mathcal{X}^\star$, where $N < \infty$. As $x_k \in C$ for all k,

$$\min_{i \in [N]} \|x_k - x_i^\star\| - \epsilon \le \operatorname{dist}(x_k, \mathcal{X}^\star \cap C) = \operatorname{dist}(x_k, \mathcal{X}^\star) \stackrel{\text{a.s.}}{\to} 0,$$

and $\min_{i \in [N]} \|x_k - x_i^\star\| \stackrel{\text{a.s.}}{\to} \min_{i \in [N]} V(x_i^\star)$. Thus for any $\epsilon > 0$, there exists $x_\epsilon \in \mathcal{X}^\star \cap C$ such that $V(x_\epsilon) \le \epsilon$, so that $\inf_{x \in C \cap \mathcal{X}^\star} V(x) = 0$. By the continuity of V, the infimum is attained and there is $x^\star$ such that $V(x^\star) = 0$. The sequence x_k thus converges to this $x^\star \in \mathcal{X}^\star$, and $F(x_k) \stackrel{\text{a.s.}}{\to} F(x^\star) = F^\star$. ∎

3.3. Asymptotic normality. Without additional conditions, it is challenging to provide more precise convergence rate guarantees than those of Corollary 3.9, which (at best) provides an asymptotic rate scaling as $1/(\alpha_k k)$ for $\alpha_k \asymp k^{-\beta}$. With this in mind, we introduce an additional assumption of smoothness and strong convexity in a

neighborhood of the optimal point, and we consider distributional convergence to establish asymptotic normality of the averaged iterates, extending results of Polyak and Juditsky [48]. In most analyses of stochastic convex optimization problems yielding asymptotic normality, typical assumptions are that the random functions $f(\cdot; s)$ have globally Lipschitz gradients (e.g., [48, section 5] and [22, 33, 52]), which is reasonable when $\mathcal{X}$ is compact, but may fail for noncompact $\mathcal{X}$. In contrast, we consider the following assumption.

Assumption A3. The functions F and f satisfy the following.

(i) The function F is $\mathcal{C}^2$ in a neighborhood of

$$x^\star = \operatorname*{argmin}_{x \in \mathcal{X}} F(x)$$

and $\nabla^2 F(x^\star) \succ 0$.

(ii) There exists $\epsilon > 0$ such that $f(\cdot; s)$ is $L(s)$-smooth on the set

$$\mathcal{X}_\epsilon^\star := \{x \mid \|x - x^\star\| \le \epsilon\},$$

meaning that $x \mapsto \nabla f(x; s)$ is $L(s)$-Lipschitz on $\mathcal{X}_\epsilon^\star$, and $\mathbb{E}[L(S)^2] = L^2 < \infty$.

Assumption A3 says that in a neighborhood of $x^\star$, the random functions f have Lipschitz gradients with probability 1. We will apply Assumption A3 in conjunction with Assumption A2, which enforces a type of local Lipschitz continuity of f and F. Typical Lipschitz conditions on ∇f imply Assumption A2: if $\nabla f(\cdot; s)$ is $L_r(s)$-Lipschitz on $\mathcal{X}_r^\star = \{x \in \mathcal{X} \mid \|x - x^\star\| \le r\}$, where the smoothness constant L_r is square integrable for finite r, Assumption A2 holds whenever $\mathbb{E}[\|\nabla f(x^\star; S)\|^2] < \infty$. Indeed,

$$\|\nabla f(x; s)\| \le \|\nabla f(x^\star; s)\| + \|\nabla f(x^\star; s) - \nabla f(x; s)\| \le \|\nabla f(x^\star; s)\| + L_r(s) \|x - x^\star\|,$$

so we have the moment bound $\mathsf{G}_{\text{big}}(r) \le 2\mathbb{E}[\|\nabla f(x^\star; S)\|^2] + 2\mathbb{E}[L_r(S)^2] r^2$. To further motivate Assumption A3, we provide a brief example.

Example 3 (Poisson regression). In problems with count data $b_1, b_2, \ldots, b_m \in \mathbb{N}$, we may wish to predict counts based on a covariate vector $a_i \in \mathbb{R}^n$. A standard model is Poisson regression, a generalized linear model [39, 23], where we model $b \in \mathbb{N}$ conditional on $a \in \mathbb{R}^n$ as coming from a Poisson distribution with parameter $\lambda = e^{\langle a, x \rangle}$, so $p(b \mid a, x) = e^{-\lambda} \lambda^b / b!$. The negative log likelihood is $f(x; (a, b)) = -\log p(b \mid a, x) = \log(b!) + \exp(\langle a, x \rangle) - b\langle a, x \rangle$, and it is easy to compute the truncated model (8), as f satisfies

$$\inf_z f(x; (a, b)) = \log(b!) + \inf_t \{e^t - bt\} = \log(b!) + b - b \log b.$$

Because $f'(x; (a, b)) = ae^{\langle a, x \rangle} - ba$, using $|e^t - e^s| \le e^{\max\{s,t\}} |s - t|$ shows that f satisfies Assumption A3 as soon as we have the covariance condition $\operatorname{Cov}(a) \succ 0$ and $\mathbb{E}[e^{r\|a\|_2}] < \infty$ for $r < \infty$. Such moment conditions are not sufficient for SGM to converge. ∎

We have the following asymptotic normality theorem; we present the proof in Appendix A.

THEOREM 3.11. *Let Assumptions* A1–A3 *hold. Let the iterates* x_k *be generated by any method satisfying conditions* (C.i) *and* (C.ii) *with step sizes* $\alpha_k = \alpha_0 k^{-\beta}$ *for some* $\beta \in (\frac{1}{2}, 1)$ *and* $\alpha_0 > 0$. *Assume additionally that the iterates are bounded: with probability* 1, $\sup_k \|x_k\| < \infty$. *Then*

$$\frac{1}{\sqrt{k}} \sum_{i=1}^{k} (x_i - x^\star) \xrightarrow{d} \mathsf{N}\big(0, \nabla^2 F(x^\star)^{-1} \operatorname{Cov}(\nabla f(x^\star; S)) \nabla^2 F(x^\star)^{-1}\big).$$

To prove Theorem 3.11, we use two main insights. The first is that if the iterates remain bounded, then Proposition 3.8 guarantees convergence. The second is a gradient approximation result that shows that even if the models in the iterations (4) are nonsmooth, they locally behave as first-order Taylor approximations to the functions f, and thus eventually the iterates approximate the stochastic gradient method on quadratics. From this, we can apply the techniques of Polyak and Juditsky [48] to guarantee asymptotic normality.

The asymptotic convergence guarantee in Theorem 3.11 cannot be improved. It achieves the local asymptotic minimax bound for stochastic optimization [20] (the analogue of the standard Fisher information in classical statistical problems; see [34] and [61, section 8.7]). In contrast to stochastic gradient schemes, however, Theorem 3.11 requires essentially *only* that the iterates remain bounded. In the end, then, the important consequence of the APROX family is that, with appropriately accurate models, we can guarantee stability (Definition 3.1). By leveraging these stability guarantees, we can then show that model-based iteration schemes (4) are (asymptotically) optimal.

4. Fast convergence for easy problems. The stability and asymptotic results in section 3 provide some evidence for the benefits of using better models within stochastic optimization problems: if the models are accurate enough that the iterates remain bounded, then we obtain asymptotic optimality results even when standard gradient methods diverge. In this section, we study a different collection of problems, which we term *easy* optimization problems. More precisely, we say that a stochastic minimization problem is easy if there are *shared* global minimizers.

DEFINITION 4.1. *Let* $F(x) := \mathbb{E}_P[f(x; S)]$. *Then* F *is* easy to optimize *if for each* $x^\star \in \mathcal{X}^\star := \operatorname{argmin}_{x \in \mathcal{X}} F(x)$ *and* P*-almost all* $s \in \mathcal{S}$ *we have*

$$\inf_{x \in \mathcal{X}} f(x; s) = f(x^\star; s).$$

Definition 4.1 places strong restrictions on the class of functions we consider, but a number of examples satisfy its conditions. Other researchers have considered similar conditions to Definition 4.1; for example, Schmidt and Le Roux [53] study stochastic optimization problems of the form $F(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)$, where $\nabla f_i(x^\star) = 0$ for all i. We briefly enumerate a few examples to motivate what follows, returning in section 4.3 to flesh them out fully.

Example 4 (overdetermined linear systems; see section 4.3.1). In an overdetermined linear system, we have a matrix $A \in \mathbb{R}^{m \times n}$, with $m \geq n$, and we wish to solve $Ax = b$, where we assume the system of equalities is feasible. Letting $a_i \in \mathbb{R}^n$ denote the rows of A, both objectives $F(x) = \frac{1}{2m} \|Ax - b\|_2^2$ and $F(x) = \frac{1}{m} \|Ax - b\|_1$ satisfy Definition 4.1, where we take samples $s = (a_i, b_i) \in \mathbb{R}^n \times \mathbb{R}$, and any solution $x^\star$ to $Ax = b$ satisfies $f(x^\star; (a_i, b_i)) = 0$ for all i. ∎

Example 5 (finding a point in the intersection of convex sets; see section 4.3.2). Let $C_1, C_2, \ldots, C_m$ be closed convex sets with nonempty intersection $\mathcal{X}^\star := \bigcap_{i=1}^m C_i$. Then the objective $F(x) = \frac{1}{m}\sum_{i=1}^m \operatorname{dist}(x, C_i)$ is convex, and treating the sample space $\mathcal{S} = \{1, \ldots, m\}$, we have $F(x) = 0$ and $f(x; i) := \operatorname{dist}(x, C_i) = 0$ for all i if and only if $x \in \mathcal{X}^\star$. ∎

Example 6 (data interpolation; see section 4.3.3). A more involved example arises out of recent results in statistical machine learning. In this area, substantial recent success in deep learning [35, 64] arises out of models that fit a training sample of data *perfectly*. In settings more amenable to analysis, Belkin and coauthors [4, 5] study statistical algorithms that minimize convex losses and perfectly interpolate the data, that is, given a sample $\{S_1, \ldots, S_m\}$ drawn i.i.d. from an underlying population, they find points $x^\star$ satisfying $f(x^\star; S_i) = \inf_x f(x; S_i)$ for each $i = 1, \ldots, m$. ∎

Given Examples 4, 5, and 6, it is of interest to investigate the aProx family for problems satisfying Definition 4.1. In this case, we show that any model satisfying the local approximation conditions (C.i) and (C.ii) and the additional lower bound condition (C.iii) possesses strong adaptivity and convergence properties. This is in contrast to subgradient methods, which (given a precise step-size choice) can exhibit fast convergence, but are typically nonadaptive.

To highlight the types of models we consider, without loss of generality, we may assume that $\inf_{x\in\mathcal{X}} f(x; s) = 0$, as given an oracle that provides the value $\inf f(\cdot; s)$ we can replace f with $f(\cdot; s) - \inf f(\cdot; s)$. As we discuss in section 2 and Example 3, it is frequently easy to compute the infimum $\inf_z f(z; s)$ for an individual sample s. The results in this section thus apply to the lower-truncated model (8),

$$f_x(y; s) := \left[f(x; s) + \langle f'(x; s), y - x\rangle\right]_+ .$$

The updates for this model are easy to compute when $\mathcal{X} = \mathbb{R}^n$; indeed, in this case, the guarded model (8) yields the update

$$x_{k+1} = x_k - \min\left\{\alpha_k, \frac{f(x_k; S_k)}{\|f'(x_k; S_k)\|_2^2}\right\} f'(x_k; S_k).$$

This update is reminiscent of the classical Polyak subgradient method [47], which chooses "optimal" step sizes in the subgradient method when the value $F(x^\star)$ is known.

In the remainder of this section, we analyze the performance of the aProx family of models on easy problems, showing that in a number of settings, these methods even enjoy linear convergence. The starting point of each of our results is the following lemma, whose proof we defer until section B.1, that shows that if a problem has shared minimizers as in Definition 4.1, iterates of any method satisfying the lower bound condition (C.iii) are guaranteed to make progress toward the optimal set.

Lemma 4.2. *Let F be easy to optimize (Definition* 4.1*). Let x_k be generated by the updates* (4) *using a model satisfying conditions* (C.i)–(C.iii)*. Then for any $x^\star \in \mathcal{X}^\star$,*

$$\begin{aligned}&\frac{1}{2}\|x_{k+1} - x^\star\|_2^2 \\ &\le \frac{1}{2}\|x_k - x^\star\|_2^2 - \frac{1}{2}[f(x_k; S_k) - f(x^\star; S_k)] \min\left\{\alpha_k, \frac{f(x_k; S_k) - f(x^\star; S_k)}{\|f'(x_k; S_k)\|_2^2}\right\}.\end{aligned}$$

In the next two sections, we use this lemma to derive conditions on the growth of F that we can leverage for fast convergence of aProx models. We return to our examples in section 4.3, demonstrating that common problems satisfy the assumptions in sections 4.1 and 4.2.

4.1. Sharp growth with shared minimizers. For our first set of problems, we consider objectives that exhibit *sharp growth* away from the optimal set $\mathcal{X}^\star$; classical and recent optimization literature highlights the importance of such conditions for the convergence of deterministic optimization methods [13, 18]. As we shall see, these conditions are sufficient to guarantee linear convergence of aProx models in stochastic settings. We begin with the following assumption, which we use for its ease of applicability following the structure of the progress guarantee in Lemma 4.2.

Assumption A4 (expected sharp growth). There exist constants $\lambda_0, \lambda_1 > 0$ such that for all $\alpha \in \mathbb{R}_+$, $x \in \mathcal{X}$, and $x^\star \in \mathcal{X}^\star$,

$$\mathbb{E}\left[\min\left\{\alpha[f(x;S) - f(x^\star;S)], \frac{(f(x;S) - f(x^\star;S))^2}{\|f'(x;S)\|_2^2}\right\}\right] \geq \operatorname{dist}(x, \mathcal{X}^\star) \min\left\{\lambda_0 \alpha, \lambda_1 \operatorname{dist}(x, \mathcal{X}^\star)\right\}.$$

While Assumption A4 is somewhat complex, a quick calculation shows that for it to hold, it is sufficient that there exist constants $\lambda > 0$ and $p > 0$ such that

$$\mathbb{P}\big(f(x;S) - f(x^\star;S) \geq \lambda \operatorname{dist}(x, \mathcal{X}^\star)\big) \geq p$$

and $\mathbb{E}[\|f'(x;S)\|_2^2] \leq M^2$ for $x \in \mathcal{X}$. That is, $f(x;S) \geq f(x^\star;S) + \lambda \operatorname{dist}(x, \mathcal{X}^\star)$ with nonzero probability, which is reasonably easy to check (for example, using the Paley–Zygmund inequality and Mendelson's small-ball conditions [40]).

Under these conditions, we have the following fast convergence guarantee for the aProx family of methods; we provide the proof in section B.2.

Proposition 4.3. *Let F be easy to optimize, Assumption* A4 *hold, and x_k be generated by the stochastic iteration* (4) *using any model satisfying conditions* (C.i)–(C.iii)*, where the step sizes satisfy $\alpha_k = \alpha_0 k^{-\beta}$ for some $\beta \in (-\infty, 1)$. Define $K_0 := \left\lfloor (\lambda_0 \alpha_0 / (\lambda_1 \operatorname{dist}(x_1, \mathcal{X}^\star)))^{1/\beta} \right\rfloor$. Then*

$$\mathbb{E}[\operatorname{dist}(x_{k+1}, \mathcal{X}^\star)^2] \leq \begin{cases} \exp\left(-\lambda_1 \min\{k, K_0\} - \frac{\lambda_0}{\operatorname{dist}(x_1, \mathcal{X}^\star)} \sum_{i=K_0+1}^{k} \alpha_i\right) \operatorname{dist}(x_1, \mathcal{X}^\star)^2 & \text{if } \beta \geq 0, \\ \exp\left(-\lambda_1 [k - K_0]_+ - \frac{\lambda_0}{\operatorname{dist}(x_1, \mathcal{X}^\star)} \sum_{i=1}^{k \wedge K_0} \alpha_i\right) \operatorname{dist}(x_1, \mathcal{X}^\star)^2 & \text{if } \beta < 0 \end{cases}$$

and with probability 1*, we have the linear convergence*

$$\limsup_{k \to \infty} \frac{\operatorname{dist}(x_k, \mathcal{X}^\star)^2}{(1 - \lambda_1)^k} < \infty.$$

Without careful step-size choices, even nonstochastic subgradient methods do not achieve such convergence guarantees. Indeed, consider the simple objective $F(x) = |x|$, which certainly satisfies the sharp growth conditions, and apply the subgradient method $x_{k+1} = x_k - \alpha_k \operatorname{sign}(x_k)$ (where we take $\operatorname{sign}(0) = +1$). Then the convergence can be no faster than $O(\alpha_k)$; this is the typical jamming behavior of subgradient

methods. In contrast, Proposition 4.3 shows that by leveraging the knowledge that $\inf_x F(x) = 0$, we achieve linear convergence.

4.2. Quadratic growth with shared minimizers. As an alternative to sharp growth conditions, we also consider optimization problems that exhibit quadratic growth—strong convexity—away from their optima. As is the case for the sharpness conditions in section 4.1, strong convexity conditions play an important role in the analysis and implementation of methods for convex optimization [25, 26, 47, 44, 18] as well as stochastic optimization problems [24, 21]. It is thus of interest to develop an understanding of the behavior of the aProx family of methods under strong convexity conditions, so that we make the following assumption (which is slightly weaker than strong convexity).

Assumption A5 (quadratic growth with shared minimizers). There exist constants $\lambda_0, \lambda_1 > 0$ such that for all $x \in \mathcal{X}$ and $\alpha > 0$,

$$\mathbb{E}\left[(f(x;S) - f(x^\star;S)) \min\left\{\alpha, \frac{f(x;S) - f(x^\star;S)}{\|f'(x;S)\|_2^2}\right\}\right] \geq \min\{\lambda_0\alpha, \lambda_1\} \operatorname{dist}(x, \mathcal{X}^\star)^2.$$

Let us give more intuitive conditions sufficient for A5 to hold. Suppose the standard conditions that ∇F is L-Lipschitz and that F has quadratic growth: $F(x) - F(x^\star) \geq c_0 \operatorname{dist}(x, \mathcal{X}^\star)^2$. In addition, assume there exist constants $0 < c, C < \infty$, $p > 0$ such that

$$\mathbb{P}\left(\|\nabla f(x;S)\|_2^2 \leq C\|\nabla F(x)\|_2^2 \text{ and } f(x;S) - f(x^\star;S) \geq c(F(x) - F(x^\star))\right) \geq p > 0.$$

The Lipschitz condition implies $F(x^\star) \leq F(y) \leq F(x) + \langle \nabla F(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2$, and setting $y = x - \frac{1}{L}\nabla F(x)$ gives

$$F(x) - F(x^\star) \geq \frac{1}{2L}\|\nabla F(x)\|_2^2 \quad \text{or} \quad \frac{2L}{\|\nabla F(x)\|_2^2} \geq \frac{1}{F(x) - F(x^\star)}.$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left[(f(x;S) - f(x^\star;S)) \min\left\{\alpha, \frac{f(x;S) - f(x^\star;S)}{\|f'(x;S)\|_2^2}\right\}\right] \\
&\quad \geq cp\,(F(x) - F(x^\star)) \min\left\{\alpha, \frac{F(x) - F(x^\star)}{C\|\nabla F(x)\|_2^2}\right\} \\
&\quad \geq pcc_0 \min\left\{\alpha, \frac{1}{2CL}\right\} \operatorname{dist}(x, \mathcal{X}^\star)^2.
\end{aligned}$$

Under the quadratic growth assumption (Assumption A5), whenever the problem is easy (Definition 4.1) Lemma 4.2 implies the following proposition, which gives nearly linear convergence of the aProx family whenever the lower bound condition (C.iii) holds.

Proposition 4.4. *Let Assumption* A5 *hold and x_k be generated by the stochastic iteration* (4) *by any model satisfying conditions* (C.i)–(C.iii)*, where the step sizes satisfy $\alpha_k = \alpha_0 k^{-\beta}$ for some $\beta \in (-\infty, \infty)$. Define $K_0 = \lfloor(\lambda_0\alpha_0/\lambda_1)^{1/\beta}\rfloor$. If $\beta \geq 0$,*

then

$$\mathbb{E}[\operatorname{dist}(x_{k+1}, \mathcal{X}^\star)^2] \le \exp\left(-\lambda_1 \min\{k, K_0\} - \lambda_0 \sum_{i=K_0+1}^{k} \alpha_i\right) \operatorname{dist}(x_1, \mathcal{X}^\star)^2,$$

while if $\beta < 0$*, then*

$$\mathbb{E}[\operatorname{dist}(x_{k+1}, \mathcal{X}^\star)^2] \le \exp\left(-\lambda_1 \left[k - K_0\right]_+ - \lambda_0 \sum_{i=1}^{K_0 \wedge k} \alpha_i\right) \operatorname{dist}(x_1, \mathcal{X}^\star)^2.$$

Proof. Under Assumption A5, the distance recursion for $D_k := \operatorname{dist}(x_k, \mathcal{X}^\star)$ in Lemma 4.2 then becomes $D_k \le D_{k-1}$ and

$$\mathbb{E}[D_{k+1}^2 \mid \mathcal{F}_{k-1}] \le \max\{1 - \lambda_0 \alpha_k, 1 - \lambda_1\} D_k^2.$$

The claim follows by algebraic manipulations and the fact that $1 - t \le e^{-t}$ for all t. □

Under similar strong convexity assumptions, Schmidt and Le Roux [53] and Ma, Bassily, and Belkin [38] show that stochastic gradient methods can achieve linear or near-linear convergence for easy convex optimization problems. As is typical in the analysis of stochastic gradient methods, however, this requires precise step-size choices that reflect typically unknown constants, such as global Lipschitz conditions and the strong convexity parameter λ_0. In contrast, the aProx family of methods is adaptive to the easiness of the problem, achieving optimal asymptotic behavior (as in section 3) while providing strong finite-sample guarantees and nearly linear convergence (Proposition 4.4) when problems satisfy strong growth conditions.

4.3. Examples of easy problems with strong growth. We now return to Examples 4–6, which fit our framework of easy (Definition 4.1) problems with shared minimizers, exhibiting quantitative growth conditions for each.

4.3.1. Overdetermined linear systems and Kaczmarz algorithms. Kaczmarz algorithms [56, 36, 42, 41] for overdetermined linear systems, as in Example 4, are effective, solving feasible systems $Ax = b$ (where $A \in \mathbb{R}^{m \times n}$, $m \ge n$) using careful stochastic gradient steps on the objective $\|Ax - b\|_2^2$ to achieve fast convergence. We consider instead the mean absolute error $F(x) := \frac{1}{m} \|Ax - b\|_1$, which typically satisfies the sharp growth condition (Assumption A4), so that the aProx method (4) using the truncated model (8) achieves linear convergence. Specialized Kaczmarz algorithms typically achieve slightly better convergence rates [56, 42, 41], but in this case, we have the additional benefit that the aProx methods are adaptive: they still converge outside of linear systems.

We provide conditions sufficient to demonstrate Assumption A4. Let the vectors $a_i \in \mathbb{R}^n$ be drawn independently from a distribution with $\|a_i\|_2 \le M$, and assume for small $c > 0$ there exists $p_c > 0$ such that $\mathbb{P}(|\langle a_i, v\rangle| \ge c \|v\|_2) \ge p_c$ for $v \in \mathbb{R}^n$. Letting $b_i = \langle a_i, x^\star\rangle$ and $f_i(x) = |\langle a_i, x\rangle - b_i|$, we have the following lemma; see the long version [1] for a proof.

Lemma 4.5. *Let the preceding conditions on the vectors* a_i *hold. There exists a numerical constant* $C < \infty$ *such that for* $c > 0$ *and* $t \ge 0$*, if we define*

$$\lambda_0 := c\left(p_c - C\sqrt{\frac{n+t}{m}}\right) \quad \text{and} \quad \lambda_1 := \frac{c^2}{M^2}\left(p_c - C\sqrt{\frac{n+t}{m}}\right),$$

then with probability at least $1 - e^{-t}$ *over the randomness in the* a_i*, simultaneously for all* x*,*

$$\frac{1}{m}\sum_{i=1}^{m} \min\left\{\alpha[f_i(x) - f_i(x^\star)], \frac{(f_i(x) - f_i(x^\star))^2}{\|f_i'(x)\|_2^2}\right\} \geq \|x - x^\star\|_2 \min\left\{\lambda_0\alpha, \lambda_1 \|x - x^\star\|_2\right\}.$$

That is, with high probability over the choice of $A \in \mathbb{R}^{m\times n}$, Assumption A4 holds with parameters λ_0 and λ_1. As a more concrete example, suppose the vectors a_i are uniform on $\sqrt{n}\cdot\mathbb{S}^{n-1}$, the sphere of radius $\sqrt{n}$. Then

$$\mathbb{P}\left(|\langle a_i, v\rangle| \geq \frac{1}{2}\|v\|_2\right) \geq \frac{1}{2},$$

so that $\lambda_1 \gtrsim 1/n$ with high probability, and Proposition 4.3 implies that

$$\limsup_k \|x_k - x^\star\|_2^2/(1 - C/n)^k < \infty.$$

Roughly, then, $O(1)\cdot n\log\frac{1}{\epsilon}$ iterations of the APROX update (4) are sufficient to achieve ϵ-accuracy in the solution of $Ax = b$, each of which requires time $O(n)$, yielding a total operation count of $O(1)\cdot n^2\log\frac{1}{\epsilon}$. This is a less precise version of the bound Strohmer and Vershynin [56, section 2.1] attain for Kaczmarz methods on well-conditioned problems.

4.3.2. Finding a point in the intersection of convex sets. We return now to Example 5, building connections with randomized projection algorithms [3, 36, 37]. Let $C_1, C_2, \ldots, C_m$ be closed convex sets, where $\mathcal{X}^\star := \bigcap_{i=1}^m C_i$ is nonempty. The conditioning of the problem of finding a point in this intersection is related to the ratio $\operatorname{dist}(x, \mathcal{X}^\star)/\max_i \operatorname{dist}(x, C_i)$ (cf. [36, 37]). As a simple special case, if the sets C_i are all half-spaces of the form $C_i = \{x \mid \langle a_i, x\rangle \leq b_i\}$, then the Hoffman error bound [27] implies that $\|[Ax - b]_+\|_\infty \geq c\operatorname{dist}(x, \mathcal{X}^\star)$ for a constant $c > 0$; this result also holds in infinite dimensions under a constraint qualification [28]. Abstracting away the particulars of the sets C_i, consider $F(x) = \frac{1}{m}\sum_{i=1}^m \operatorname{dist}(x, C_i)$, and assume the analog of the Hoffman bound that for some $\lambda > 0$,

$$\operatorname{dist}(x, \mathcal{X}^\star) \geq \max_i \operatorname{dist}(x, C_i) \geq \lambda \operatorname{dist}(x, \mathcal{X}^\star),$$

where the first inequality always holds and λ is a condition number [36]. For example, for half-spaces $C_i = \{x \mid \langle a_i, x\rangle \leq b_i\}$, $i = 1, 2$, with $\|a_i\| = 1$, $\lambda = \sqrt{(1 + \langle a_1, a_2\rangle)/2}$ suffices.

To understand the growth properties of $F(x) = \frac{1}{m}\sum_{i=1}^m \operatorname{dist}(x, C_i)$, recall that $\operatorname{dist}(\cdot, C_i)$ is 1-Lipschitz and subdifferentiable [25, Example VI.3.3], so that the component functions satisfy $\|\partial \operatorname{dist}(x, C_i)\| \leq 1$. Then letting I be uniform in $\{1, \ldots, m\}$, we obtain for any $\alpha > 0$ that

$$\begin{aligned}
\mathbb{E}[\operatorname{dist}(x, C_I)\min\{\alpha, \operatorname{dist}(x, C_I)\}] &= \frac{1}{m}\sum_{i=1}^m \operatorname{dist}(x, C_i)\min\{\alpha, \operatorname{dist}(x, C_i)\} \\
&\geq \frac{1}{m}\max_i \operatorname{dist}(x, C_i)\min\left\{\alpha, \max_i \operatorname{dist}(x, C_i)\right\} \\
&\geq \operatorname{dist}(x, \mathcal{X}^\star)\min\left\{\frac{\alpha\lambda}{m}, \frac{\lambda}{m}\operatorname{dist}(x, \mathcal{X}^\star)\right\}.
\end{aligned}$$

In particular, Assumption A4 holds with constant $\lambda_1 = \lambda/m$, and thus Proposition 4.3 implies the convergence $\limsup_k \operatorname{dist}(x_k, \mathcal{X}^\star)^2/(1-\lambda/m)^k < \infty$. Roughly $\frac{m}{\lambda}\log\frac{1}{\epsilon}$ steps are sufficient to find a point x_k that is within distance ϵ of the set $\mathcal{X}^\star$ using the aProx family.

4.3.3. Interpolation problems. Finally, we consider statistical machine learning problems in which one interpolates the data; see Example 6. Belkin and collaborators [4, 5, 38] suggest that in a number of statistical machine learning problems, it is possible to achieve zero error on a training sample while still achieving optimal convergence rates for the population objective. To give a simple example, we study underparameterized least-squares. Our data comes in m pairs $(a_i, b_i) \in \mathbb{R}^n \times \mathbb{R}$, $f_i(x) = \frac{1}{2}(\langle a_i, x\rangle - b_i)^2$, where $n > m$. The minimum norm interpolant $x^\star := \operatorname{argmin}_x\{\|x\|_2 \mid Ax = b\}$, or equivalently, the minimizer

$$x^\star = \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{\sum_{i=1}^m \ell(\langle a_i, x\rangle - b_i) \;\middle|\; x \in \operatorname{span}\{a_1, \ldots, a_m\}\right\}$$

for any loss $\ell : \mathbb{R} \to \mathbb{R}_+$ uniquely minimized at 0, possesses certain statistical optimality properties while exhibiting strong empirical prediction performance; see [4, 5, 38] and [64, section 5]. In our case, the truncated models (8) guarantee that the iterates of the aProx family lie in the span of the vectors a_i. With our choice f_i above, if we let $M := \max_i \|a_i\|_2$ then

$$\frac{1}{m}\sum_{i=1}^m f_i(x)\min\left\{\alpha, \frac{f_i(x)}{\|\nabla f_i(x)\|_2^2}\right\} \geq \frac{1}{2m}\|A(x-x^\star)\|_2^2 \min\left\{\alpha, \frac{1}{M^2}\right\}.$$

Let $U\Sigma V^T = A$ be the singular value decomposition of A, so

$$\|A(x-x^\star)\|_2^2 = \sum_{i=1}^m \sigma_i(A)^2\langle v_i, x-x^\star\rangle^2 \geq \sigma_m(A)^2\left\|VV^T(x-x^\star)\right\|_2^2.$$

As $x^\star \in \operatorname{span}\{a_i\}$, whenever $x \in \operatorname{span}\{a_i\}$, we have $\left\|VV^T(x-x^\star)\right\|_2 = \|x-x^\star\|_2$, and Assumption A5 holds for such x: if I is uniform on $\{1, \ldots, m\}$ then

$$\mathbb{E}\left[(f_I(x) - f_I(x^\star))\min\left\{\alpha, \frac{f_I(x)-f_I(x^\star)}{\|\nabla f_I(x)\|_2^2}\right\}\right] \geq \frac{\sigma_m(A)^2}{m}\min\left\{\alpha, M^{-2}\right\}\|x-x^\star\|_2^2.$$

Random matrices $A \in \mathbb{R}^{m\times n}$ with independent rows typically satisfy an inequality of the form $\sigma_m(A) \gtrsim \mathbb{E}[\|a\|_2^2]^{1/2}(1-\sqrt{m/n})$ with high probability [62]. Assuming that $M^2 \lesssim \mathbb{E}[\|a\|_2^2]$, we see that in this case, Assumption A5 then holds with constants $\lambda_0 \gtrsim \frac{1}{m}\mathbb{E}[\|a\|_2^2]$ and $\lambda_1 \gtrsim \frac{1}{m}$.

Summarizing, under typical scenarios, Proposition 4.4 guarantees the following corollary.

Corollary 4.6. *Consider the underdetermined least-squares problem, where the matrix A has rows with constant norm $\|a_i\|_2 = \sqrt{n}$ and $\sigma_m(A) \geq \sqrt{m}$, and let x_k be generated by the iteration* (4) *with the truncated model* (8) *and step sizes $\alpha_k = \alpha_0 k^{-\beta}$ for some $\beta > 0$. Then there exists a constant C depending on n, m, α_0, and β such that for all $k \in \mathbb{N}$,*

$$\mathbb{E}[\|x_k - x^\star\|_2^2] \leq C\max\left\{\exp\left(-\frac{k}{m}\right), \exp\left(-\frac{\alpha_0}{1-\beta}k^{1-\beta}\right)\right\}\|x_1 - x^\star\|_2^2.$$

Corollary 4.6 shows that the iterates exhibit nearly linear convergence. With a more careful step-size choice, gradient methods achieve linear convergence [53, 38], but these recommended step sizes cause nonconvergence except on the narrow set of "easy" problems considered. In our scenario, however, aProx methods achieve these convergence rates while enjoying convergence in other problems.

5. Nonasymptotic convergence results. For our final set of theoretical results, we provide two propositions on the nonasymptotic convergence of aProx methods, describing their behavior on Lipschitz objectives, and then showing that for any functions exhibiting a type of strong convexity, the stochastic proximal point method achieves strong nonasymptotic guarantees.

The first result, providing convergence for Lipschitzian objectives, is essentially known [15, Theorem 4.1], and it generalizes the standard results in most treatments of stochastic convex optimization, where one makes some type of Lipschitzian assumption on the random functions f (e.g., [66, 43, 8]).

Assumption A6. There exists $M^2 < \infty$ such that $\mathbb{E}[\|f'(x;S)\|_2^2] \le M^2$ for each $x \in \mathcal{X}$.

Under this Lipschitzian assumption, the following extension of the well-known results on convergence of the stochastic gradient method [66, 43] (also generalizing Bertsekas [8]) holds.

Proposition 5.1. *Let Assumptions* A1 *and* A6 *hold, and let the iterates x_k be generated by algorithm* (4) *by any model satisfying conditions* (C.i) *and* (C.ii). *Define $\overline{x}_k = (\sum_{i=1}^k \alpha_i)^{-1} \sum_{i=1}^k \alpha_i x_i$. Then*

$$\mathbb{E}[F(\overline{x}_k)] - F(x^\star) \le \frac{1}{2\sum_{i=1}^k \alpha_i} \|x_0 - x^\star\|_2^2 + \frac{M^2}{2\sum_{i=1}^k \alpha_i} \sum_{i=1}^k \alpha_i^2.$$

If $\mathcal{X}$ is compact with $R := \sup_{x \in \mathcal{X}} \|x - x^\star\|_2$, then the average $\overline{x}_k := \frac{1}{k}\sum_{i=1}^k x_i$ satisfies

$$\mathbb{E}[F(\overline{x}_k)] - F(x^\star) \le \frac{R^2}{2k\alpha_k} + \frac{M^2}{2k} \sum_{i=1}^k \alpha_i.$$

The proof is a more or less standard application of Lemma 3.10 (see [1, Appendix C]); the only contribution beyond Davis and Drusvyatskiy [15] is the claim under compact $\mathcal{X}$.

We present one final theoretical result, showing how the stochastic proximal point method (the full model (7)) achieves reasonable nonasymptotic convergence guarantees even when the functions f may be non-Lipschitz and nonsmooth, as long as they obey a (restricted) strong convexity condition. The convergence guarantees we present here are impossible with standard stochastic gradient methods, which can diverge under the assumptions we consider.

Assumption A7 (restricted strong convexity). For each $s \in \mathcal{S}$, the functions $f(\cdot; s)$ are strongly convex over $\mathcal{X}$ with respect to the matrix $\Sigma(s) \succeq 0$, that is,

$$f(y;s) \ge f(x;s) + \langle f'(x;s), y - x\rangle + \frac{1}{2}(x-y)^T \Sigma(s)(x-y) \quad \text{for } x, y \in \mathcal{X},$$

for all $f'(x;s) \in \partial f(x;s)$. The matrix Σ satisfies $\mathbb{E}[\Sigma(S)] \succeq \lambda_{\min} I_{n\times n}$, where $\lambda_{\min} > 0$.

We analyze the stochastic proximal point method under Assumption A7, and begin with a technical lemma, which provides a guarantee on the one-step progress of the method. We present the proof of Lemma 5.2 in Appendix C.

LEMMA 5.2. *Let Assumption* A7 *hold and the iterates* x_k *be generated by the stochastic proximal point method* (3). *Define*

$$\overline{\Sigma}_k := \mathbb{E}\left[\frac{1}{1 + 2\alpha_k \lambda_{\max}(\Sigma(S))}\Sigma(S)\right].$$

Then

$$\frac{1}{2}\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2 \mid \mathcal{F}_{k-1}\right] \le \frac{1}{2}(x_k - x^\star)^T\left(I - \alpha_k\overline{\Sigma}_k\right)(x_k - x^\star) + \alpha_k^2\mathbb{E}\left[\|f'(x^\star; S)\|_2^2\right].$$

Lemma 5.2 guarantees that the proximal-point method makes progress whenever the restricted strong convexity conditions hold, irrespective of smoothness of the objective functions. First, we have $0 \prec \overline{\Sigma}_0 \preceq \overline{\Sigma}_k$ for all $k \in \mathbb{N}$, and defining $\lambda_k := \lambda_{\min}(\overline{\Sigma}_k) > 0$, we have $\lambda_k \uparrow \lambda_\infty := \lambda_{\min}(\mathbb{E}[\Sigma(S)])$ when $\alpha_k \downarrow 0$. Under the conditions of Lemma 5.2 and Assumption A1, we thus have

$$\begin{aligned}\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2 \mid \mathcal{F}_{k-1}\right] &\le (1 - \alpha_k\lambda_k)\|x_k - x^\star\|_2^2 + \alpha_k^2\sigma^2 \\ &\le (1 - \alpha_k\lambda_0)\|x_k - x^\star\|_2^2 + \alpha_k^2\sigma^2.\end{aligned}$$

Applying this inequality recursively gives

$$\mathbb{E}[\|x_{k+1} - x^\star\|_2^2] \le \prod_{i=1}^k(1 - \alpha_i\lambda_i)\|x_1 - x^\star\|_2^2 + \sum_{i=1}^k \alpha_i^2 \prod_{j=i+1}^k (1 - \alpha_j\lambda_j)\sigma^2,$$

where we note that $\alpha_j\lambda_j < 1$ for all j. An inductive argument [48] implies the next proposition.

PROPOSITION 5.3. *Let Assumptions* A1 *and* A7 *hold, and let* x_k *be generated by the stochastic proximal point method* (3) *with step sizes satisfying* $\alpha_k = \alpha_0 k^{-\beta}$ *for some* $\beta \in (0, 1)$. *Then for a numerical constant* $C < \infty$,

$$\mathbb{E}[\|x_{k+1} - x^\star\|_2^2] \le \exp\left(-\lambda_0\sum_{i=1}^k \alpha_i\right)\|x_1 - x^\star\|_2^2 + C \cdot \frac{\sigma^2}{\lambda_0}\alpha_k \cdot \log k.$$

An asymptotic argument [48] gives convergence $\mathbb{E}[\|x_k - x^\star\|_2^2] \lesssim \frac{\sigma^2}{\lambda_\infty}\alpha_k$ for large k. For stochastic proximal point methods, if the objective is strongly convex, choosing $\alpha_k = C/k$ for a large constant C yields asymptotic convergence bounds on $\|x_k - x^\star\|_2^2$ of the form $\sigma^2/\lambda_\infty k$. In comparison to convergence results available for stochastic gradient methods [47, 43, 2], Proposition 5.3 holds whenever the subgradient $f'(x^\star; S)$ has finite second moment, and we require this only at the point $x^\star$; moreover, it holds no matter the step-size sequence.

6. Experiments. The final component of this paper is an empirical evaluation of the APROX methods. Our goal in the experiments is to evaluate the relative merits of different approximate models in the iteration (4), and accordingly, we consider the four approximations below.

(i) Stochastic gradient method (SGM): uses the linear model (6).
(ii) Proximal: uses the full model (7).
(iii) Truncated: uses the lower truncated model (8).
(iv) Bundle: uses the bundle (cutting plane) model (9) with two lines, that is, with $i = 1$.

We present several experiments, each comparing different aspects of the aProx models. Throughout, we consider step-size sequences of the form $\alpha_k = \alpha_0 k^{-\beta}$, where $\beta \in (1/2, 1)$. We wish to evaluate the robustness and stability of each of the models (6)–(9) for different problems, investigating both well- and poorly conditioned instances, as well as problems satisfying the "easy" conditions of Definition 4.1. Let us provide some guidance toward expected results. Roughly, for problems with globally Lipschitz gradients (such as linear regression problems with noise), we expect the methods to have fairly similar performance—stochastic gradient, proximal point, truncated, and bundle models are all asymptotically normal with optimal covariance. However, as problems become either (i) easier—closer to satisfying Definition 4.1—or (ii) more poorly conditioned or harder because of large or unbounded Lipschitz constants, we expect stochastic gradient methods to (i) converge more slowly or (ii) be substantially more sensitive to step-size choices.

Within each of our experiments, we run each model-based iteration (4) for K total iterations across multiple initial step sizes α_0. For a fixed accuracy $\epsilon > 0$, we record the number of steps k required to achieve $F(x_k) - F(x^\star) \le \epsilon$, reporting these times (where we terminate each run at iteration K). We perform T experiments for each initial step-size choice, reporting the median of the time-to-ϵ-accuracy; the shaded areas in each plot cover the 5th to 95th percentile of convergence times (90% coverage sets). Finally, while we restrict our experiments to the single-sample batch size setting (i.e., no "mini-batches"), we note that using multiple samples to decrease the variance of (sub)gradients does not necessarily improve the robustness of standard stochastic subgradient methods. Indeed, Examples 1 and 2 show that the deterministic gradient method can be sensitive to step-size choice. In experiments we do not include because of the additional space they require, we verify this intuition—stochastic gradient methods remain similarly sensitive to step-size choice even using large batch sizes. Additionally, it is possible to consider the iteration k at which the averaged iterate $\overline{x}_k$ achieves $F(\overline{x}_k) - F(x^\star) \le \epsilon$; this does not change the qualitative aspects of our plots in any way.

6.1. Linear regression. In our linear regression experiments, we let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, and $F(x) = \frac{1}{2m} \|Ax - b\|_2^2$, where in each individual experiment we generate $x^\star \sim \mathsf{N}(0, I_n) \in \mathbb{R}^n$ and set $b = Ax^\star + \sigma v$ for $v \sim \mathsf{N}(0, I_m)$. We choose σ differently depending on the experiment, setting $\sigma = 0$ in noiseless experiments and $\sigma = \frac{1}{2}$ otherwise. We generate A as $A = QD$, where $Q \in \mathbb{R}^{m \times n}$ has uniformly random orthogonal columns, and $D = \mathrm{diag}(1, 1 + (\kappa - 1)/(n - 1), \ldots, \kappa)$ is a diagonal matrix with linearly spaced entries between 1 and a desired condition number $\kappa \ge 1$.

In Figure 2, we plot the results of our experiments on well-conditioned problems, which use matrices A with condition number $\kappa(A) = 1$, while Figure 3 shows identical results except that we use condition number $\kappa(A) = 15$.[1] Plot (a) of each figure demonstrates the results for the noiseless setting with $\sigma = 0$. In Figure 2(a), we see the expected result that the stochastic gradient method has good performance for a precise range of step sizes in $[10^{-1}, 1]$, while the approximations of the proximal

[1] See online version for color figures.

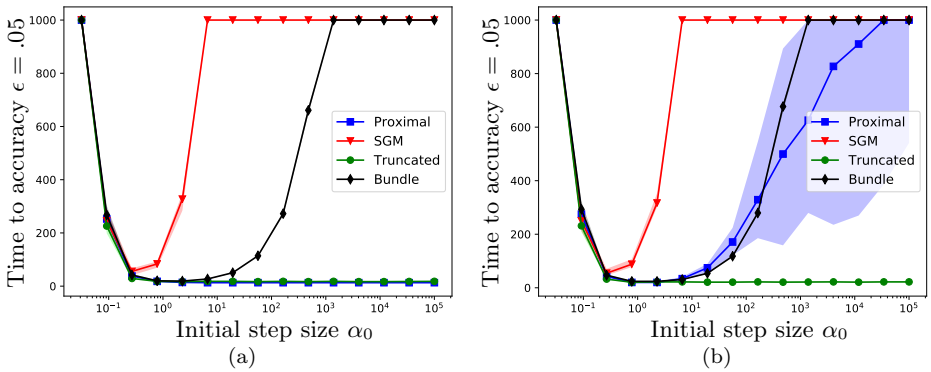

FIG. 2. *The number of iterations to achieve ϵ-accuracy versus initial step size α_0 for linear regression with $m = 1000$, $n = 40$, and condition number $\kappa(A) = 1$. (a) The noiseless setting with $\sigma = 0$. (b) The noisy setting with $\sigma = \frac{1}{2}$.*

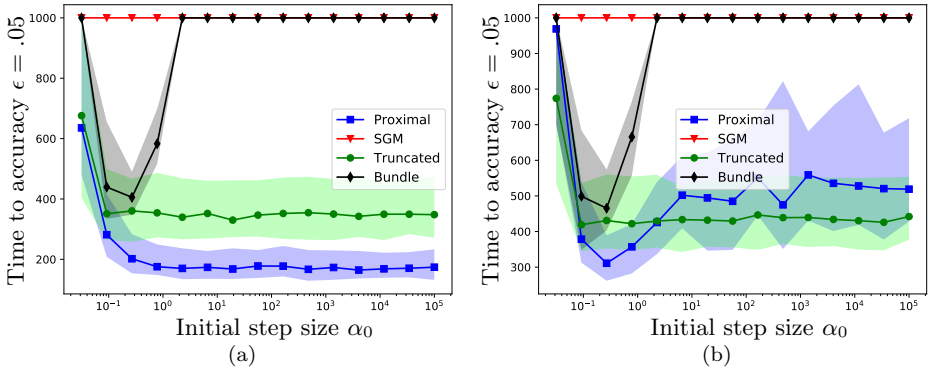

FIG. 3. *The number of iterations to achieve ϵ-accuracy versus initial step size α_0 for linear regression with $m = 1000$, $n = 40$, and condition number $\kappa(A) = 15$. (a) The noiseless setting with $\sigma = 0$. (b) The noisy setting with $\sigma = \frac{1}{2}$.*

point (7) and the truncated (8) models yield better convergence over a range of step sizes with six orders of magnitude. The bundle model (9) shows somewhat more robustness than SGM, but for large step sizes also exhibits some oscillation. In the noisy cases, plot (b) in each figure, the results are similar, except that the full proximal model is somewhat less robust to step-size choice; roughly, in the stochastic proximal point (SPPM) case, the model trusts the instantaneous function *too* much, and overfits to the noise at each iteration.

Figure 3 tells a similar story to Figure 2, except that the stochastic gradient method is essentially not convergent: in no experiment did it ever achieve accuracy below $\epsilon = .05$ in the noisy or noiseless settings. This problem is not in any real sense challenging: a condition number of $\kappa(A) = 15$ is not particularly poorly conditioned [60], yet stochastic gradient methods exhibit very poor behavior. These plots suggest that the reliance on stochastic gradient methods in much of the statistical and machine learning literature may be misplaced.

6.2. Logistic regression. We now turn to classification experiments, beginning with a logistic regression experiment. In logistic regression, widely used for fitting models for binary classification in statistics and machine learning [23], we have data

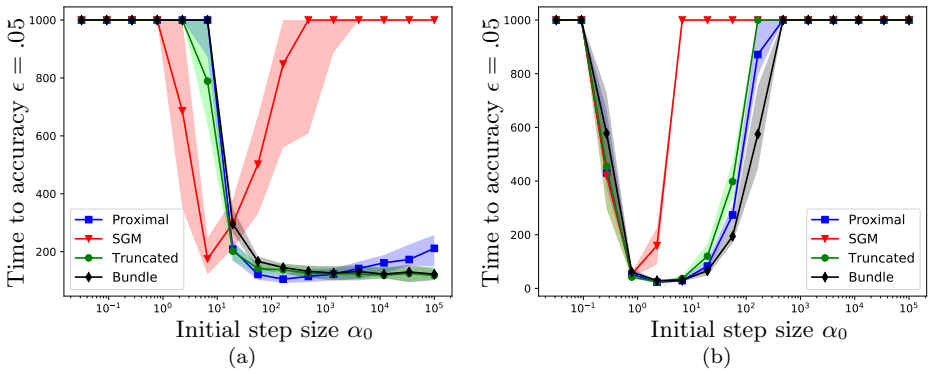

FIG. 4. *The number of iterations to achieve ϵ-accuracy as a function of the initial step size α_0 for logistic regression with $n = 40$, $m = 1000$, and condition number $\kappa(A) = 1$. (a) Noiseless experiment.* (b) *Labels flipped with probability $p = 0.01$.*

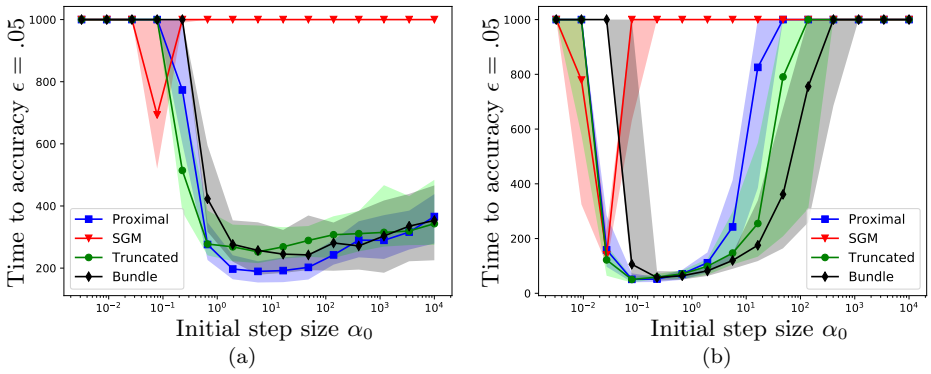

FIG. 5. *The number of iterations to achieve ϵ-accuracy as a function of the initial step size α_0 for logistic regression with parameters $n = 40$, $m = 1000$, and condition number $\kappa(A) = 15$. (a) Noiseless experiment.* (b) *Labels flipped with probability $p = .01$.*

pairs $(a_i, b_i) \in \mathbb{R}^n \times \{\pm 1\}$, and we wish to minimize

$$F(x) := \frac{1}{m} \sum_{i=1}^{m} f(x; (a_i, b_i)), \quad \text{where} \quad f(x; (a, b)) = \log\left(1 + e^{-b\langle a, x\rangle}\right).$$

We generate the data as follows: we sample $a_i \overset{\text{iid}}{\sim} \mathsf{N}(0, I_n)$ and $u^\star \sim \mathsf{N}(0, I_n)$, labeling $b_i = \text{sign}(\langle a_i, u^\star \rangle)$; in the noisy setting we flip each label's sign independently with probability p.

We present the results of this experiment in Figures 4 and 5, including plots for both the noiseless (perfectly separated) and noisy cases (plots (a) and (b) in each figure, respectively), where Figure 5 displays results when the condition number of the data matrix A is $\kappa(A) = 15$. These plots demonstrate similar results to those in the preceding sections, though there are a few differences. First, in the noiseless setting, there is no optimizer $x^\star$, as the optimal value is $\lim_{t \to \infty} F(tu^\star) = 0$, yet still we see the benefits of the more accurate models in Figures 4(a) and 5(a). Moreover, the truncated and proximal models exhibit a wider range of convergent step sizes than the simple stochastic gradient method does even in the noisy case.

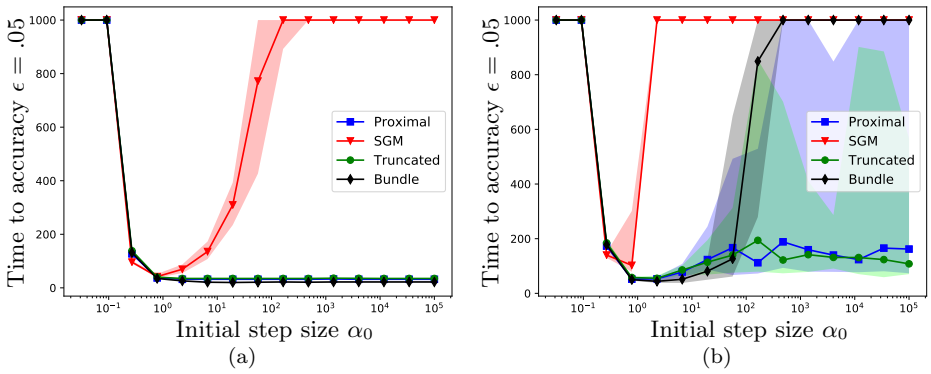

FIG. 6. *The number of iterations to achieve ϵ-accuracy as a function of the initial step size α_0 for multiclass hinge loss with parameters $n = 15$, $m = 2000$, $K = 10$ and* (a) *label randomization probability $p = 0$ and* (b) *label randomization probability $p = 0.01$.*

6.3. Multiclass hinge loss. In our second classification experiment, we focus on a somewhat more complex multiclass setting, using the multiclass hinge loss [14]. In this setting, we receive m vectors $a_i \in \mathbb{R}^n$ and a correct label $\ell_i \in [K]$ for each i, where K is the number of classes. We wish to find a classifier, represented as a collection of K vectors $X = [x_1 \ \cdots \ x_K] \in \mathbb{R}^{n \times K}$, that minimizes

$$F(X) = \frac{1}{m} \sum_{i=1}^{m} \max_{j \neq \ell_i} \left[1 + \langle a_i, x_j - x_{\ell_i} \rangle\right]_+ .$$

In the case that the data is separable with a positive margin, meaning that there exists $X^\star$ such that $\langle a_i, x_{\ell_i} \rangle \geq 1 + \langle a_i, x_j \rangle$ for all $j \neq \ell_i$, this problem is equivalent to finding a point in the intersection of half-spaces (recall section 4.3.2). Accordingly, we expect to see fast convergence for the truncated and proximal-point models ((8) and (7), respectively) for large step sizes.

To generate the data, we draw vectors $a_i \overset{\text{iid}}{\sim} \mathsf{N}(0, I_n)$, then generate an "optimal" classifier $U^\star \in \mathbb{R}^{n \times K}$ with i.i.d. $\mathsf{N}(0, 1)$ entries. In the nonnoisy setting, we set $\ell_i = \operatorname{argmax}_j \langle u_j^\star, a_i \rangle$, while in the noisy setting, for each $i \in \{1, \ldots, m\}$ we resample a value ℓ_i uniformly at random with probability p. We present the results of this experiment in Figure 6. The experiment reinforces the conclusions of the previous experiments: better models (see (4)) in the APROX family are significantly more robust to the step-size values and achieve generally faster convergence than more naive subgradient methods.

6.4. Poisson regression. For our final experiment, we consider a Poisson regression problem (Example 3), for which classical results on stochastic approximation do not apply because of the exponential objective. In this case, we model counts $b_i \in \mathbb{N}$ as coming from a distribution $p(b \mid a, x) = \exp(-e^{\langle a, x \rangle}) \exp(b \langle a, x \rangle)/b!$, giving loss $f(x; (a, b)) = \exp(\langle a, x \rangle) - b \langle a, x \rangle$. We generate the data by first drawing $u \sim \sqrt{n} \cdot \mathsf{Uni}(\mathbb{S}^{n-1})$, then drawing $a_i \overset{\text{iid}}{\sim} \mathsf{N}(0, (1/n) I_n)$ and $b_i \sim \mathsf{Poisson}(e^{\langle a_i, u \rangle})$. In this experiment, the proximal update has no closed form, as it involves minimizing a quadratic plus exponential term, but the minimizer of the proximal update lies on $\{x_k + ta \mid t \in \mathbb{R}\}$, yielding a one-dimensional convex optimization problem. As Example 3 shows, however, it is simple to implement the truncated model (8) by computing $\inf_x f(x; (a, b)) = \log(b!) + b - b \log b$.

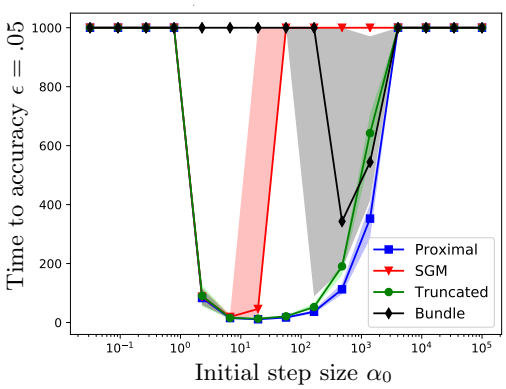

FIG. 7. *The number of iterations to achieve ϵ-accuracy as a function of the initial step size α_0 for Poisson regression with parameters $n = 40$ and $m = 1000$.*

We present the results in Figure 7. It is surprising to us that the stochastic gradient method converges at all on this problem, but with low-noise scenarios and small enough step sizes, it seems that SGM does not leave a region around zero and so is convergent. Importantly, the truncated model (8), in spite of its substantially easier calculation, enjoys convergence nearly as robust as that of the stochastic proximal point method.

Appendix A. Proof of Theorem 3.11. The proof of the theorem proceeds in a series of lemmas, each of which requires some work. Roughly, the outline is as follows: first, we develop a recursion for the iterates x_{k+1}, which parallels a noisy gradient recursion, except that the errors implicitly depend on the next iterate. This allows a decomposition (see (14)) of $\frac{1}{k}\sum_{i=1}^k (x_i - x^\star)$ into a leading term, which is obviously asymptotically normal, and several error terms. Two of these errors are standard stochastic approximation errors (similar, for example, to those in Polyak and Juditsky [48]), though we require care in showing they are negligible (see Lemmas A.1, A.2, and A.3). The last error term involves subgradients of the models $f_{x_k}(\cdot; S_k)$ *at the point* x_{k+1}, causing an implicit and potentially nonsmooth dependence in the errors. To address this, we provide a gradient comparison result (Lemma A.4), based on Davis, Drusvyatskiy, and Paquette [16], which shows that even if the method generating the iterates x_k uses a nonsmooth approximation to $f(\cdot; S_k)$, near x_k the subgradients of the model approximate ∇f. This allows us to adapt the results of Polyak and Juditsky [48] on asymptotic optimality of stochastic gradient methods.

Let $\Delta_k = x_k - x^\star$, and for $H = \nabla^2 F(x^\star)$, define the remainder $R(x) = \nabla F(x) - H(x - x^\star)$, so that $\|R(x)\| = O(\|x - x^\star\|^2)$ as $x \to x^\star$. We perform an expansion to rewrite the implicit iteration $x_{k+1} = x_k - \alpha_k f'_{x_k}(x_{k+1}; S_k)$. Letting $\varepsilon_k := f'_{x_k}(x_{k+1}; S_k) - f'(x_k; S_k)$ and defining the localized (sub)gradient errors

$$\zeta_k = \zeta(x_k, S_k) := (\nabla f'(x_k; S_k) - \nabla f(x^\star; S_k)) - (F'(x_k) - \nabla F(x^\star)), \tag{12}$$

we obtain

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k f'_{x_k}(x_{k+1}; S_k) \\ &= x_k - \alpha_k \Big[H(x_k - x^\star) + R(x_k) + f'(x_k; S_k) - F'(x_k) + \varepsilon_k\Big] \\ &= x_k - \alpha_k H(x_k - x^\star) - \alpha_k \nabla f(x^\star; S_k) - \alpha_k \left[R(x_k) + \zeta_k + \varepsilon_k\right]. \end{aligned} \tag{13}$$

Subtracting $x^\star$ to use $\Delta_k = x_k - x^\star$, the implicit iteration (13) becomes

$$\Delta_{k+1} = (I - \alpha_k H)\Delta_k - \alpha_k \nabla f(x^\star; S_k) - \alpha_k[R(x_k) + \zeta_k + \varepsilon_k].$$

Defining the matrices

$$B_i^k := \alpha_i \sum_{j=i}^{k} \prod_{l=i+1}^{j} (I - \alpha_l H) \quad \text{and} \quad A_i^k := B_i^k - H^{-1},$$

Polyak and Juditsky [48, Lemma 2] show that $\overline{\Delta}_k = \frac{1}{k}\sum_{i=1}^k \Delta_i$ satisfies the equality

$$\begin{aligned} \sqrt{k}\overline{\Delta}_k = &\frac{1}{\sqrt{k}} \sum_{i=1}^{k} H^{-1} \nabla f(x^\star; S_k) \\ &+ \frac{1}{\sqrt{k}} \sum_{i=1}^{k} A_i^k \nabla f(x^\star; S_i) + \frac{1}{\sqrt{k}} \sum_{i=1}^{k} B_i^k [R(x_i) + \zeta_i + \varepsilon_i] + O(1/\sqrt{k}) \end{aligned} \tag{14}$$

and additionally $\sup_{i,k} \|B_i^k\| < \infty$ and $\lim_k \frac{1}{k}\sum_{i=1}^k \|A_i^k\| = 0$. Evidently, equality (14) implies the theorem as soon as we show that each of the terms except $k^{-1/2}H^{-1}\sum_{i=1}^k \nabla f(x^\star; S_k)$ are $o_P(1)$. We thus bound each of the terms in (14) in turn, which gives Theorem 3.11.

LEMMA A.1. *Assume the conditions of Theorem* 3.11 *hold. Then* $\|\Delta_k\| \overset{a.s.}{\to} 0$, $\sum_{k=1}^\infty \alpha_k(F(x_k) - F(x^\star)) < \infty$ *and* $\sum_{k=1}^\infty \alpha_k \|\Delta_k\|^2 < \infty$.

LEMMA A.2. *Under the conditions of Theorem* 3.11, $\frac{1}{\sqrt{k}}\sum_{i=1}^k \|R(x_i)\| \overset{a.s.}{\to} 0$.

LEMMA A.3. *Assume the conditions of Theorem* 3.11 *hold. Then* $\frac{1}{\sqrt{k}}\sum_{i=1}^k \zeta_i \overset{a.s.}{\to} 0$ *and* $\frac{1}{\sqrt{k}}\sum_{i=1}^k A_i^k \zeta_i \overset{a.s.}{\to} 0$.

Polyak and Juditsky [48] prove similar versions of these lemmas, so we omit their proofs. (See [1, Appendix A] for complete proofs.)

Controlling the implicit modeling errors $\varepsilon_k = f'_{x_k}(x_{k+1}; S_k) - \nabla f(x_k; S_k)$ is the most important challenge, and for this, we use the following gradient comparison lemma.

LEMMA A.4 (Davis, Drusvyatskiy, and Paquette [16, Theorem 6.1]). *Let* f *and* h *be convex and subdifferentiable on* $\mathbb{R}^n$ *and* $\epsilon > 0$, $r < \infty$. *Assume that on the set* $\{x \mid \|x - x^\star\| \le \epsilon\}$, *the function* f *has* L*-Lipschitz gradient. Assume additionally that* $f \ge h$ *and at the point* x_0, $h'(x_0) \in \partial f(x_0)$. *Then for any* x *and* $h'(x) \in \partial h(x)$, *if* $\|x - x^\star\| \le \epsilon/4$ *and* $\|x_0 - x^\star\| \le \epsilon/4$, *then*

$$\|\nabla f(x) - h'(x)\| \le 2L \|x - x_0\|.$$

Key to the application of Lemma A.4 is that individual iterates move very little.

LEMMA A.5. *Assume conditions* (C.i) *and* (C.ii) *hold. Then* $\|x_k - x_{k+1}\| \le \alpha_k \|f'(x_k; S_k)\|$ *for some* $f'(x_k; S_k) \in \partial f(x_k; S_k)$.

Proof. Let $g_{k+1} \in \partial f_{x_k}(x_{k+1}; S_k)$ satisfy $\langle \alpha_k g_{k+1} + (x_{k+1} - x_k), y - x_{k+1}\rangle \ge 0$ for all $y \in \mathcal{X}$. Then, using $\langle g_k - g_{k+1}, x_k - x_{k+1}\rangle \ge 0$ for any $g_k \in \partial f_{x_k}(x_k; S_k) \subset \partial f(x_k; S_k)$ and setting $y = x_k$, we have $\alpha_k\langle g_k, x_k - x_{k+1}\rangle \ge \alpha_k\langle g_{k+1}, x_k - x_{k+1}\rangle \ge \|x_k - x_{k+1}\|^2$. Cauchy–Schwarz then gives the result. ∎

With Lemmas A.4 and A.5 in place, we can control the final error terms in the expansion (14).

LEMMA A.6. *Under the conditions of Theorem* 3.11, $\frac{1}{\sqrt{k}} \sum_{i=1}^k \|\varepsilon_i\| \stackrel{a.s.}{\to} 0$.

Proof. In the implicit iteration (13), if $x_k, x_{k+1} \in \{x : \|x - x^\star\| \le \epsilon/4\}$, then Lemma A.4 shows that $\varepsilon_k = f'_{x_k}(x_{k+1}; S_k) - \nabla f(x_k; S_k)$ satisfies

$$\begin{aligned} \|\varepsilon_k\| \le 2L(S_k) \|x_k - x_{k+1}\| &\stackrel{(*)}{\le} 2\alpha_k L(S_k) \|\nabla f(x_k; S_k)\| \\ &\le \alpha_k L(S_k)^2 + \alpha_k \|\nabla f(x_k; S_k)\|^2, \end{aligned} \tag{15}$$

where inequality $(*)$ follows by the single step guarantee in Lemma A.5.

We show that each of the two terms in inequality (15) has small sum. We have $\mathbb{E}[L^2(S)] < \infty$ and $\mathbb{E}[\sum_k \alpha_k k^{-1/2} L^2(S_k)] < \infty$, so the Kronecker lemma implies

$$k^{-1/2} \sum_{i=1}^k \alpha_i L^2(S_i) \stackrel{\text{a.s.}}{\to} 0.$$

For bounding the second term, let $\epsilon > 0$ be such that $\sum_k \alpha_k k^{-1/2+\epsilon} < \infty$, which must exist as $\alpha_k = \alpha_0 k^{-\beta}$ for $\beta \in (1/2, 1)$. Define

$$Z_k := \frac{1}{k^{1/2-\epsilon}} \sum_{i=1}^k \alpha_i \|f'(x_i; S_i)\|^2,$$

which is adapted to $\mathcal{F}_k$. Then

$$\begin{aligned} \mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] &\le \frac{k^{1/2-\epsilon}}{(k+1)^{1/2-\epsilon}} Z_k + (k+1)^{-1/2+\epsilon} \alpha_{k+1} \mathbb{E}[\|f'(x_{k+1}; S_{k+1})\|^2 \mid \mathcal{F}_k] \\ &\le Z_k + (k+1)^{-1/2+\epsilon} \alpha_{k+1} \mathsf{G}_{\text{big}}(\|\Delta_{k+1}\|), \end{aligned}$$

the last inequality following by Assumption A2. On the event that $\sup_k \|\Delta_k\| < \infty$, we have $\sum_k k^{-1/2+\epsilon} \alpha_k \mathsf{G}_{\text{big}}(\|\Delta_k\|) < \infty$, so that the Robbins–Siegmund lemma (Lemma 3.4) applies, and thus $Z_k \stackrel{\text{a.s.}}{\to} Z_\infty$ for some finite random variable Z_∞. Thus $k^{-\epsilon} Z_k = \frac{1}{\sqrt{k}} \sum_{i=1}^k \alpha_i \|f'(x_i; S_i)\|^2 \stackrel{\text{a.s.}}{\to} 0$.

The definition (4) of the iteration for x_{k+1} implies $f'_{x_k}(x_{k+1}; S_k) = \alpha_k^{-1}(x_k - x_{k+1})$, and Lemma A.5 gives $\|f'_{x_k}(x_{k+1}; S_k)\| = \alpha_k^{-1} \|x_k - x_{k+1}\| \le \|\partial f(x_k; S_k)\|$. Thus, we obtain

$$\begin{aligned} \frac{1}{\sqrt{k}} \sum_{i=1}^k \|\varepsilon_i\| \le \frac{2}{\sqrt{k}} \sum_{i=1}^k 1\left\{\|x_i - x^\star\| \ge \epsilon/4 \text{ or } \|x_{i+1} - x^\star\| \ge \epsilon/4\right\} \|f'(x_i; S_i)\| \\ + \frac{1}{\sqrt{k}} \sum_{i=1}^k (\alpha_i L(S_i)^2 + \alpha_i \|f'(x_i; S_i)\|^2), \end{aligned}$$

where the second term is a consequence of inequality (15). Because $x_k \stackrel{\text{a.s.}}{\to} x^\star$ (from Lemma A.1), both of these terms converge to zero with probability 1. ∎

Appendix B. Proofs of fast convergence on easy problems.

B.1. Proof of Lemma 4.2. First, we assume without loss of generality that $f(x^\star; s) = 0$ for all $x^\star \in \mathcal{X}^\star$, as we may replace f with $f - \inf f$. By Lemma 3.7, the update (4) satisfies

$$\frac{1}{2}\left\|x_{k+1} - x^\star\right\|_2^2 \le \frac{1}{2}\left\|x_k - x^\star\right\|_2^2 + \alpha_k\left[f_{x_k}(x^\star; S_k) - f_{x_k}(x_{k+1}; S_k)\right] - \frac{1}{2}\left\|x_{k+1} - x_k\right\|_2^2.$$

For shorthand, let $g_k = f'(x_k; S_k)$ and $f_k = f(x_k; S_k)$. As $f_{x_k}(x^\star; S_k) \le f(x^\star; S_k) = 0$, and by condition (C.iii) we have $f_{x_k}(x_{k+1}; S_k) \ge [f_k + \langle g_k, x_{k+1} - x_k\rangle]_+$, we have

$$\left\|x_{k+1} - x^\star\right\|_2^2 \le \left\|x_k - x^\star\right\|_2^2 + 2\alpha_k\left[f(x^\star; S_k) - [f_k + \langle g_k, x_{k+1} - x_k\rangle]_+\right] - \left\|x_{k+1} - x_k\right\|_2^2. \tag{16}$$

If we let $\widetilde{x}_{k+1}$ denote the unconstrained minimizer

$$\begin{aligned}\widetilde{x}_{k+1} &= \operatorname*{argmin}_x\left\{[f_k + \langle g_k, x - x_k\rangle]_+ + \frac{1}{2\alpha_k}\left\|x - x_k\right\|_2^2\right\} \\ &= x_k - \lambda_k g_k \quad \text{for } \lambda_k := \min\left\{\alpha_k, \frac{f_k}{\left\|g_k\right\|_2^2}\right\},\end{aligned}$$

then because $x_{k+1} \in \mathcal{X}$ we have

$$-\alpha_k f_{x_k}(x_{k+1}; S_k) - \frac{1}{2}\left\|x_{k+1} - x_k\right\|_2^2 \le -\alpha_k f_{x_k}(\widetilde{x}_{k+1}; S_k) - \frac{1}{2}\left\|\widetilde{x}_{k+1} - x_k\right\|_2^2.$$

By inspection, the guarded step size λ_k guarantees that $[f_k + \langle g_k, \widetilde{x}_{k+1} - x_k\rangle]_+ = f_k - \lambda_k\left\|g_k\right\|_2^2$, and thus inequality (16) (setting $f(x^\star; S_k) = 0$) yields

$$\begin{aligned}\left\|x_{k+1} - x^\star\right\|_2^2 &\le \left\|x_k - x^\star\right\|_2^2 - 2\alpha_k f_{x_k}(\widetilde{x}_{k+1}; S_k) - \left\|\widetilde{x}_{k+1} - x_k\right\|_2^2 \\ &\le \left\|x_k - x^\star\right\|_2^2 - 2\lambda_k f_k + \lambda_k^2\left\|g_k\right\|_2^2.\end{aligned}$$

We have two possible cases: whether $f_k/\left\|g_k\right\|_2^2 \lesseqgtr \alpha_k$. In the case that $f_k/\left\|g_k\right\|_2^2 \le \alpha_k$, we have $\lambda_k = f_k/\left\|g_k\right\|_2^2$ and so $-2\lambda_k f_k + \lambda_k^2\left\|g_k\right\|_2^2 = -f_k^2/\left\|g_k\right\|_2^2$. In the alternative case that $f_k/\left\|g_k\right\|_2^2 > \alpha_k$, we have $\lambda_k = \alpha_k$ and $\alpha_k^2\left\|g_k\right\|_2^2 \le \alpha_k f_k$. Combining these cases gives

$$\left\|x_{k+1} - x^\star\right\|_2^2 \le \left\|x_k - x^\star\right\|_2^2 - \min\left\{\alpha_k f_k, \frac{f_k^2}{\left\|g_k\right\|_2^2}\right\}.$$

B.2. Proof of Proposition 4.3. We adopt a bit of shorthand notation. Let $D_k = \operatorname{dist}(x_k, \mathcal{X}^\star)$, so $D_k \in \mathcal{F}_{k-1}$. Then Lemma 4.2 implies that under Assumption A4 we have

$$\begin{aligned}\mathbb{E}[D_{k+1}^2 \mid \mathcal{F}_{k-1}] &\le D_k^2 - \min\left\{\lambda_0\alpha_k D_k, \lambda_1 D_k^2\right\} = \max\left\{(1-\lambda_1), (1 - \lambda_0\alpha_k/D_k)\right\} D_k^2 \\ &\le \max\left\{(1-\lambda_1), (1 - \lambda_0\alpha_k/D_1)\right\} D_k^2, \end{aligned} \tag{17}$$

where we have used that $D_1 \ge D_k$ for all $k \ge 1$ by Lemma 4.2. Inequality (17) immediately implies that if $\beta \ge 0$, then

$$K_0 = \sup\{k \in \mathbb{N} \mid \lambda_0\alpha_k > \lambda_1 D_1\} = \left\lfloor\left(\frac{\lambda_0\alpha_0}{\lambda_1 D_1}\right)^{1/\beta}\right\rfloor$$

is the index for which $k \geq K_0$ implies $\lambda_0 \alpha_k / D_1 \leq \lambda_1$. This gives the first result of the proposition, $\mathbb{E}[D_{k+1}^2] \leq \exp(-\lambda_1 \min\{k, K_0\} - \frac{\lambda_0}{D_1} \sum_{i=K_0+1}^{k} \alpha_i) D_1^2$ if $\beta \geq 0$. For $\beta < 0$, the same choice of K_0 gives $\mathbb{E}[D_{k+1}^2] \leq \exp(-\lambda_1 [k - K_0]_+ - \frac{\lambda_0}{D_1} \sum_{i=1}^{k \wedge K_0} \alpha_i) D_1^2$.

For the second result, we prove only the case that $\beta > 0$, as the other case is similar. Note that $\sum_{i=1}^{k} \alpha_i \gtrsim k^{1-\beta}$, so for any $\epsilon > 0$ we have for constants $0 < c, C < \infty$ depending on α_0, β, λ_0, and λ_1, that

$$\sum_{k=1}^{\infty} \mathbb{P}(D_k > \epsilon \alpha_k) \leq \frac{1}{\epsilon} \sum_{k=1}^{\infty} \exp\left(C - ck^{1-\beta} + \log \frac{1}{\alpha_k}\right) < \infty.$$

The Borel–Cantelli lemma implies that $D_k / \alpha_k \xrightarrow{\text{a.s.}} 0$. The first inequality (17) implies that if $V_k = D_{k+1}^2 / (1 - \lambda_1)^{k+1}$, then

$$\mathbb{E}[V_k \mid \mathcal{F}_{k-1}] \leq \frac{D_k^2}{(1-\lambda)^k} \max\left\{1, \frac{1 - \alpha_k / D_k}{1 - \lambda_1}\right\} \leq \left(1 + \left[\frac{1 - \alpha_k / D_k}{1 - \lambda_1}\right]_+\right) V_{k-1}.$$

As $[(1 - \alpha_k / D_k)/(1 - \lambda_1)]_+ = 0$ eventually holds with probability 1 and is $\mathcal{F}_{k-1}$-measurable, the Robbins–Siegmund lemma (Lemma 3.4) implies that $V_k \xrightarrow{\text{a.s.}} V_\infty$ for some $V_\infty \in \mathbb{R}_+$.

Appendix C. Proof of Lemma 5.2. Let $\Sigma_k = \Sigma(S_k)$, so that for all $g_k \in \partial f(x_{k+1}; S_k)$ and $y \in \mathcal{X}$ we have

$$f(y; S_k) \geq f(x_{k+1}; S_k) + \langle g_k, y - x_{k+1} \rangle + \frac{1}{2}(y - x_{k+1})^T \Sigma_k (y - x_{k+1}).$$

Using this inequality in place of the last step of the proof of Lemma 3.7 yields

$$\begin{aligned} (18) \qquad & \frac{1}{2} \|x_{k+1} - x^\star\|_2^2 + \frac{\alpha_k}{2} (x_{k+1} - x^\star)^T \Sigma_k (x_{k+1} - x^\star) \\ & \leq \frac{1}{2} \|x_k - x^\star\|_2^2 - \alpha_k [f(x_{k+1}; S_k) - f(x^\star; S_k)] - \frac{1}{2} \|x_k - x_{k+1}\|_2^2. \end{aligned}$$

Using that

$$\begin{aligned} & \frac{1}{2}(x_{k+1} - x^\star)^T \Sigma_k (x_{k+1} - x^\star) \\ & = \frac{1}{2}(x_k - x^\star)^T \Sigma_k (x_k - x^\star) + \frac{1}{2}(x_{k+1} - x_k)^T \Sigma_k (x_{k+1} - x_k) \\ & \qquad + (x_{k+1} - x_k)^T \Sigma_k (x_k - x^\star) \\ & \geq \frac{1 - \eta}{2}(x_k - x^\star)^T \Sigma_k (x_k - x^\star) + \frac{\eta - 1}{2\eta}(x_{k+1} - x_k)^T \Sigma_k (x_{k+1} - x_k) \end{aligned}$$

for all $\eta > 0$, where the inequality follows from Young's inequality, (18) then implies

$$\begin{aligned} \frac{1}{2} \|x_{k+1} - x^\star\|_2^2 \leq & \frac{1}{2}(x_k - x^\star)^T \left(I - \alpha_k (1 - \eta) \Sigma_k\right) (x_k - x^\star) \\ & - \alpha_k [f(x_{k+1}; S_k) - f(x^\star; S_k)] \\ & - \frac{1}{2}(x_k - x_{k+1})^T \left(I + \frac{\alpha_k (\eta - 1)}{\eta} \Sigma_k\right) (x_k - x_{k+1}) \end{aligned}$$

for all $\eta > 0$. The choice $\eta_k = \frac{2\alpha_k \lambda_{\max}(\Sigma_k)}{1+2\alpha_k \lambda_{\max}(\Sigma_k)} \in (0,1)$ is sufficient to guarantee $I + \frac{\alpha_k(\eta_k - 1)}{\eta_k}\Sigma_k \succeq \frac{1}{2} I$. Substituting this choice of η_k above, we obtain that

$$\frac{1}{2}\left\|x_{k+1} - x^\star\right\|_2^2 \leq \frac{1}{2}(x_k - x^\star)^T \left(I - \alpha_k(1-\eta_k)\Sigma_k\right)(x_k - x^\star) - \alpha_k\left[f(x_{k+1}; S_k) - f(x^\star; S_k)\right] - \frac{1}{4}\left\|x_k - x_{k+1}\right\|_2^2.$$

Applying Lemma 3.6 with $y = x^\star$, $x_1 = x_{k+1}$, $x_0 = x_k$, $\beta = 2\alpha_k$, and $g(\cdot) = f(\cdot; S_k)$ implies

$$\frac{1}{2}\left\|x_{k+1} - x^\star\right\|_2^2 \leq (x_k - x^\star)^T \left(I - \alpha_k(1-\eta_k)\Sigma_k\right)(x_k - x^\star) + 2\alpha_k \langle f'(x^\star; S_k), x^\star - x_k\rangle + \alpha_k^2 \left\|f'(x^\star; S_k)\right\|_2^2.$$

Taking expectations and using that $\mathbb{E}[\langle f'(x^\star; S), x^\star - x\rangle] \leq 0$ for all x (as in the proof of Theorem 3.2) gives the desired result.

Acknowledgments. We thank two anonymous reviewers for careful reading and feedback on earlier versions of this paper. In particular, both suggested strengthenings of Proposition 3.8 that we have implemented.

REFERENCES

[1] H. Asi and J. C. Duchi, *Stochastic (Approximate) Proximal Point Methods: Convergence, Optimality, and Adaptivity*, preprint, https://arxiv.org/abs/1810.05633, 2018.
[2] F. Bach and E. Moulines, *Non-asymptotic analysis of stochastic approximation algorithms for machine learning*, in Adv. Neural Inf. Process. Syst. 25, Curran Associates, Red Hook, NY, 2011, pp. 451–459,
[3] H. Bauschke and J. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.
[4] M. Belkin, D. Hsu, and P. Mitra, *Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate*, in Adv. Neural. Inf. Process. Syst. 31, Curran Associates, Red Hook, NY, 2018, pp. 2300–2311.
[5] M. Belkin, A. Rakhlin, and A. B. Tsybakov, *Does data interpolation contradict statistical optimality?*, in Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. 89, 2019, pp. 1611–1619; available at http://proceedings.mlr.press/v89/belkin19a.html.
[6] D. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.
[7] D. P. Bertsekas, *Stochastic optimization problems with nondifferentiable cost functionals*, J. Optim. Theory Appl., 12 (1973), pp. 218–231.
[8] D. P. Bertsekas, *Incremental proximal methods for large scale convex optimization*, Math. Program., 129 (2011), pp. 163–195.
[9] P. Bianchi, *Ergodic convergence of a stochastic proximal point algorithm*, SIAM J. Optim., 26 (2016), pp. 2235–2260.
[10] L. Bottou and O. Bousquet, *The tradeoffs of large scale learning*, in Adv. Neural Inf. Process. Syst. 20, Curran Associates, Red Hook, NY, 2007, pp. 161–168.
[11] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.
[12] J. Burke, *Descent methods for composite nondifferentiable optimization problems*, Math. Program., 33 (1985), pp. 260–279.
[13] J. Burke and M. Ferris, *A Gauss–Newton method for convex composite optimization*, Math. Program., 71 (1995), pp. 179–194.
[14] K. Crammer and Y. Singer, *On the algorithmic implementation of multiclass kernel-based vector machines*, J. Mach. Learn. Res., 2 (2001), pp. 265–292.
[15] D. Davis and D. Drusvyatskiy, *Stochastic model-based minimization of weakly convex functions*, SIAM J. Optim., 29 (2019), pp. 207–239.
[16] D. Davis, D. Drusvyatskiy, and C. Paquette, *The Nonsmooth Landscape of Phase Retrieval*, preprint, https://arxiv.org/abs/1711.03247, 2017.

[17] J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Y. Ng, *Large scale distributed deep networks*, in Adv. Neural Inf. Process. Syst. 25, Curran Associates, Red Hook, NY, 2012, pp. 1223–1231.
[18] D. Drusvyatskiy and A. Lewis, *Error bounds, quadratic growth, and linear convergence of proximal methods*, Math. Oper. Res., 43 (2018), pp. 919–948.
[19] J. C. Duchi and F. Ruan, *Stochastic methods for composite and weakly convex optimization problems*, SIAM J. Optim., 28 (2018), pp. 3229–3259.
[20] J. C. Duchi and F. Ruan, *Asymptotic optimality in stochastic optimization*, Ann. Statist., to appear.
[21] S. Ghadimi and G. Lan, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization,* I: *A generic algorithmic framework*, SIAM J. Optim., 22 (2012), pp. 1469–1492.
[22] S. Ghadimi and G. Lan, *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*, SIAM J. Optim., 23 (2013), pp. 2341–2368.
[23] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, New York, 2009.
[24] E. Hazan and S. Kale, *An Optimal Algorithm for Stochastic Strongly Convex Optimization*, preprint, http://arxiv.org/abs/1006.2425, 2011.
[25] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I: *Fundamentals*, Grundlehren Math. Wiss. 305, Springer, New York, 1993.
[26] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* II: *Advanced Theory and Bundle Methods*, Grundlehren Math. Wiss. 306, Springer, New York, 1993.
[27] A. J. Hoffman, *On approximate solutions of systems of linear inequalities*, J. Res. Natl. Inst. Stand., 49 (1952), pp. 263–265.
[28] H. Hu and Q. Wang, *On approximate solutions of infinite systems of linear inequalities*, Linear Algebra Appl., 114–115 (1989), pp. 429–438.
[29] N. Karampatziakis and J. Langford, *Online importance weight aware updates*, in Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, AUAI Press, Arlington, VA, 2011, pp. 392–399.
[30] J. E. Kelley, *The cutting-plane method for solving convex programs*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 703–712.
[31] B. Kulis and P. Bartlett, *Implicit online learning*, in Proceedings of the 27th International Conference on Machine Learning, Omnipress, Madison, WI, 2010, pp. 575–582.
[32] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed., Springer, New York, 2003.
[33] G. Lan, *An optimal method for stochastic composite optimization*, Math. Program., 133 (2012), pp. 365–397.
[34] L. Le Cam and G. L. Yang, *Asymptotics in Statistics: Some Basic Concepts*, Springer, New York, 2000.
[35] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature, 521 (2015), pp. 436–444.
[36] D. Leventhal and A. S. Lewis, *Randomized methods for linear constraints: Convergence rates and conditioning*, Math. Oper. Res., 35 (2010), pp. 641–654.
[37] A. S. Lewis, D. R. Luke, and J. Malick, *Local linear convergence for alternating and averaged nonconvex projections*, Found. Comput. Math., 9 (2009), pp. 485–513.
[38] S. Ma, R. Bassily, and M. Belkin, *The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning*, in Proceedings of the 35th International Conference on Machine Learning, Proc. Mach. Learn. Res. 80, 2018, pp. 3325–3334; available at http://proceedings.mlr.press/v80/.
[39] P. McCullagh and J. Nelder, *Generalized Linear Models*, Chapman and Hall, London, 1989.
[40] S. Mendelson, *Learning without concentration*, J. ACM, 62 (2015), Article no. 21.
[41] D. Needell and J. Tropp, *Paved with good intentions: Analysis of a randomized block Kaczmarz method*, Linear Algebra Appl., 441 (2014), pp. 199–221.
[42] D. Needell, R. Ward, and N. Srebro, *Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm*, in Adv. Neural Inf. Process. Syst. 27, Curran Associates, Red Hook, NY, 2014, pp. 1017–1025.
[43] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.
[44] Y. Nesterov, *Introductory Lectures on Convex Optimization*, Kluwer Academic, Norwell, MA, 2004.
[45] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 2006.

[46] A. Patrascu and I. Necoara, *Nonasymptotic Convergence of Stochastic Proximal Point Algorithms for Constrained Convex Optimization*, preprint, https://arxiv.org/abs/1706.06297, 2017.
[47] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.
[48] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855.
[49] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statist., 22 (1951), pp. 400–407.
[50] 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.
[51] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[52] E. Ryu and S. Boyd, *Stochastic Proximal Iteration: A Non-Asymptotic Improvement upon Stochastic Gradient Descent*, preprint, http://www.math.ucla.edu/~eryu/papers/spi.pdf, 2014.
[53] M. Schmidt and N. Le Roux, *Fast Convergence of Stochastic Gradient Descent under a Strong Growth Condition*, preprint, https://arxiv.org/abs/1308.6370, 2013.
[54] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, *PEGASOS: Primal estimated sub-gradient solver for SVM*, Math. Program., 127 (2011), pp. 3–30.
[55] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, MPS-SIAM Ser. Optim., SIAM, Philadelphia, PA, 2009.
[56] T. Strohmer and R. Vershynin, *A randomized Kaczmarz algorithm with exponential convergence*, J. Fourier Anal. Appl., 15 (2009), pp. 262–278.
[57] C. H. Teo, S. Vishwanthan, A. J. Smola, and Q. V. Le, *Bundle methods for regularized risk minimization*, J. Mach. Learn. Res., 11 (2010), pp. 311–365.
[58] P. Toulis and E. Airoldi, *Asymptotic and finite-sample properties of estimators based on stochastic gradients*, Ann. Statist., 45 (2017), pp. 1694–1727.
[59] P. Toulis, D. Tran, and E. Airoldi, *Towards stability and optimality in stochastic gradient descent*, in Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. 51, 2016, pp. 1290–1298; available at http://proceedings.mlr.press/v51/.
[60] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.
[61] A. W. van der Vaart, *Asymptotic Statistics*, Camb. Ser. Stat. Probab. Math., Cambridge University Press, Cambridge, 1998.
[62] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, in Compressed Sensing: Theory and Applications, Cambridge University Press, Cambridge, 2012, pp. 210–268.
[63] B. Widrow and M. E. Hoff, *Adaptive Switching Circuits*, in 1960 IRE WESCON Convention Record, Part 4, IRE, New York, 1960, pp. 96–104.
[64] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning requires rethinking generalization*, in Proceedings of the Fifth International Conference on Learning Representations, 2017; available at https://openreview.net/forum?id=Sy8gdB9xx.
[65] T. Zhang, *Solving large scale linear prediction problems using stochastic gradient descent algorithms*, in Proceedings of the Twenty-First International Conference on Machine Learning, ACM Press, New York, 2004, doi: 10.1145/1015330.1015332.
[66] M. Zinkevich, *Online convex programming and generalized infinitesimal gradient ascent*, in Proceedings of the Twentieth International Conference on Machine Learning, AAAI Press, Palo Alto, CA, 2003, pp. 928–935.