The importance of better models in stochastic optimization

Hilal Asi[a,1] **and John C. Duchi**[a,b]

[a]Department of Electrical Engineering, Stanford University, Stanford, CA 94305; and [b]Department of Statistics, Stanford University, Stanford, CA 94305

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved October 1, 2019 (received for review May 8, 2019)

Standard stochastic optimization methods are brittle, sensitive to stepsize choice and other algorithmic parameters, and they exhibit instability outside of well-behaved families of objectives. To address these challenges, we investigate models for stochastic optimization and learning problems that exhibit better robustness to problem families and algorithmic parameters. With appropriately accurate models—which we call the APROX family—stochastic methods can be made stable, provably convergent, and asymptotically optimal; even modeling that the objective is nonnegative is sufficient for this stability. We extend these results beyond convexity to weakly convex objectives, which include compositions of convex losses with smooth functions common in modern machine learning. We highlight the importance of robustness and accurate modeling with experimental evaluation of convergence time and algorithm sensitivity.

stochastic optimization | large-scale optimization

A major challenge in stochastic optimization—the algorithmic workhorse for much of modern statistical and machine-learning applications—is in setting algorithm parameters (or hyperparameter tuning). This sensitivity causes multiple issues. It results in thousands to millions of wasted engineer and computational hours. It also leads to a lack of clarity in research and development of algorithms—in claiming that one algorithm is better than another, it is unclear whether this is due to judicious choice of dataset or judicious parameter settings or whether indeed the algorithm does exhibit new desirable behavior. Consequently, in this paper we pursue 2 main thrusts: First, by using models more accurate than the first-order models common in stochastic gradient methods, we develop families of algorithms that are provably more robust to input parameter choices, with several corresponding optimality properties. Second, we argue for a different type of experimental evidence in evaluating stochastic optimization methods, where one jointly evaluates convergence speed and sensitivity of the methods.

The wasted computational and engineering energy is especially pronounced in deep learning, where engineers use models with millions of parameters, requiring days to weeks to train a single model. To get a sense of this energy use, we consider a few recent papers we view as exemplars of this broader trend: In searching for optimal neural network architectures and hyperparameters, the papers (1–3) used approximately 3,150 graphics processing unit (GPU) days, 22,000 GPU days, and 750,000 central processing unit (CPU) days of computation, respectively. To put this in perspective, assuming standard CPU energy use of between 60 and 100 W, the energy (ignoring network interconnect, monitors, etc.) for the paper (3) is roughly between 4 and $6 \cdot 10^{12}$ J. At 10^9 J per tank of gas, this is sufficient to drive 4,000 Toyota Camrys the 380 miles between San Francisco and Los Angeles.

To address these challenges, we develop stochastic optimization procedures that exhibit similar convergence to classical approaches—when the classical approaches have good tuning parameters—while enjoying better robustness, achieving this performance over a range of parameters. We argue too for evaluation of optimization algorithms based not only on convergence time but also on robustness to input choices. Briefly, a fast algorithm that converges for a small range of stepsizes is too brittle; we argue instead for (potentially slightly slower) algorithms that converge for broad ranges of stepsizes and other parameters. Our theory and experiments demonstrate the effectiveness of our methods for applications including phase retrieval, matrix completion, and deep learning.

Problem Setting and Approach

We begin by making our setting concrete. We study the stochastic optimization problem

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

In problem **1**, the set $\mathcal{S}$ is a sample space, $\mathcal{X} \subset \mathbb{R}^n$ is closed convex, and $f(x;s)$ is the loss x suffers on sample s. In this paper, we move beyond convex optimization by considering $\rho(s)$-weakly convex functions f, meaning (cf. refs. 4 and 5) that $f(x;s) + \frac{\rho(s)}{2}\|x\|_2^2$ is convex. We recover convexity when $\rho(s) \le 0$. Examples include linear regression, $f(x;(a,b)) = (\langle a,x\rangle - b)^2$, and phase retrieval, $f(x;(a,b)) = |\langle a,x\rangle^2 - b|$, which is $2\|a\|_2^2$-weakly convex.

Most optimization methods iterate by making an approximation—a model—of the objective at the current iterate, minimizing this model and reapproximating. Stochastic (sub)gradient methods (6, 7) instantiate this approach using a linear approximation; following initial work of our own and others (5, 8, 9), we study the modeling approach in more depth for stochastic optimization. Thus, the APROX algorithms we develop iterate as follows: For $k = 1, 2, \ldots$, we draw a random $S_k \sim P$ and then

Significance

Sensitivity of optimization algorithms to problem and algorithmic parameters leads to tremendous waste in time and energy, especially in applications with millions of parameters, such as deep learning. We address this by developing stochastic optimization methods demonstrably—both by theory and by experimental evidence—more robust, enjoying optimal convergence guarantees for a variety of stochastic optimization problems. Additionally, we highlight the importance of method sensitivity to problem difficulty and algorithmic parameters.

Author contributions: H.A. and J.C.D. designed research, performed research, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Data deposition: Data and code for this work have been deposited in GitHub (https://github.com/HilalAsi/APROX-Robust-Stochastic-Optimization-Algorithms).

[1]To whom correspondence may be addressed. Email: asi@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1908018116/-/DCSupplemental.

First published October 30, 2019.

update the iterate x_k by minimizing a regularized approximation to $f(\cdot; S_k)$, setting

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

We call $f_x(\cdot; s)$ the model of f at x, where f_x satisfies 3 conditions (cf. refs. 5, 8, and 9):

C.i) (Model convexity): The function $y \mapsto f_x(y; s)$ is convex and subdifferentiable.

C.ii) (Weak lower bound): The model f_x satisfies

$$f_x(y; s) \le f(y; s) + \frac{\rho(s)}{2} \|y - x\|_2^2 \text{ for all } y \in \mathcal{X}.$$

C.iii) (Local accuracy): We have $f_x(x; s) = f(x; s)$.

The containment $\partial_y f_x(y; s)|_{y=x} \subset \partial_x f(x; s)$ is immediate from condition C.iii. We provide examples presently.

We show that models slightly more accurate than the first-order model used by the stochastic gradient method—sometimes as simple as recognizing that if f is nonnegative, we should truncate the approximation at zero—achieve substantially better theoretical guarantees and practical performance. While the iterates of gradient methods can (superexponentially) diverge for misspecified stepsizes, our methods guarantee the iterates never diverge. Even more, this stability guarantees convergence and, in convex cases, optimal asymptotic normality of the averaged iterates. Finally, we evaluate the performance of our methods, validating our theoretical findings on convergence and robustness for a range of problems, including matrix completion, phase retrieval, and classification with neural networks. We defer proofs to *SI Appendix*.

In optimization broadly, proximal point methods and their related robust convergence are classical (10–12), and their role in smoothing and Moreau–Yosida regularization is also central in convex and variational analysis (13–15). In signal processing, least-mean squares for adaptive filtering is an important instance of the stochastic proximal point method (16, 17). More recent work in large-scale optimization and machine learning revisits Moreau smoothing and regularization, extending acceleration and stability properties of proximal-point-type methods to finite sum and stochastic problems (18–20).

Notation and Basic Assumptions

For a weakly convex function f, we let $\partial f(x)$ denote its Fréchet subdifferential at the point x, and $f'(x) \in \partial f(x)$ denotes an arbitrary element of the subdifferential. Throughout, we let $x^\star$ denote a minimizer of problem **1** and $\mathcal{X}^\star = \operatorname{argmin}_{x \in \mathcal{X}} F(x)$ denote the optimal set for problem **1**. We let $\mathcal{F}_k := \sigma(S_1, \ldots, S_k)$ denote the σ field generated by the first k random variables S_i. Note that $x_k \in \mathcal{F}_{k-1}$ for all k. Unless stated otherwise, we assume that the function $f(x; s)$ is $\rho(s)$ -weakly convex for each $s \in \mathcal{S}$. Finally, the following assumption implicitly holds throughout.

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 for each $x^\star \in \mathcal{X}^\star$ and selection $f'(x^\star; s) \in \partial f(x^\star; s)$, we have $\mathbb{E}[\|f'(x^\star; S)\|_2^2] \le \sigma^2$.*

Methods

To make our approach more concrete, we identify several models that fit into our framework. These have appeared in refs. 5, 8, and 9, but we believe a self-contained presentation is beneficial. Each one satisfies our conditions C.i to C.iii. The most common model in stochastic optimization is the first-order model.

Stochastic Subgradient Methods. The stochastic subgradient method uses the model

$$f_x(y; s) := f(x; s) + \langle f'(x; s), y - x \rangle. \quad [3]$$

Proximal Point Methods. In the convex setting (8, 20, 21), the stochastic proximal point method uses the model $f_x(y; s) := f(y; s)$; in the weakly convex setting, we regularize and use

$$f_x(y; s) := f(y; s) + \frac{\rho(s)}{2} \|y - x\|_2^2. \quad [4]$$

Other models require less knowledge than proximal model **4** but preserve structural properties in the original function.

Prox-Linear Model. Let the function f have the composite structure $f(x; s) = h(c(x; s); s)$, where $h(\cdot; s)$ is convex and $c(\cdot; s)$ is smooth. The stochastic prox-linear method applies h to a first-order approximation of c, using

$$f_x(y; s) := h(c(x; s) + \nabla c(x; s)^T (y - x); s). \quad [5]$$

In the nonstochastic setting, these models are classical (22), while recent work establishes convergence and convergence rates in restrictive stochastic settings (5, 9). When h is L_h Lipschitz and c has an L_c-Lipschitz gradient, then f is $\rho = L_h \cdot L_c$-weakly convex.

Example 1 (phase retrieval): In phase retrieval (23), we wish to recover an object $x^\star \in \mathbb{C}^n$ from a diffraction pattern $Ax^\star$, where $A \in \mathbb{C}^{m \times n}$, but physical sensor limitations mean we observe only amplitudes $b = |Ax^\star|^2$. A natural objective is

$$\operatorname*{minimize}_{x \in \mathbb{C}^n} \frac{1}{m} \sum_{i=1}^m f(x; (a_i, b_i)), \quad f(x; (a_i, b_i)) = \left| |\langle a_i, x \rangle|^2 - b_i \right|.$$

This is the composition of $h(z) = |z|$ and $c(x; (a_i, b_i)) = |\langle < a_i, x \rangle|^2 - b_i$, so $f(\cdot; (a_i, b_i))$ is $2\|a_i\|_2^2$-weakly convex (24). ◊

Example 2 (matrix completion): In the matrix completion problem (25), which arises (for example) in the design of recommendation systems, we have a matrix $M \in \mathbb{R}^{m \times n}$ with decomposition $M = X_\star Y_\star^T$ for $X_\star \in \mathbb{R}^{m \times r}$ and $Y_\star \in \mathbb{R}^{n \times r}$. Based on the incomplete set of known entries $\Omega \subset [m] \times [n]$, our goal is to recover the matrix M, giving rise to the objective

$$\operatorname*{minimize}_{X \in \mathbb{R}^{m \times r}, Y \in \mathbb{R}^{n \times r}} \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} f(x_i, y_j; M_{ij}),$$

where $f(x, y; z) := |\langle x, y \rangle - z|$ and x_i, y_j are the (i, j) rows of X and Y. This is the composition of $h(z) = |z|$ and $c(x, y, z) = \langle x, y \rangle - z$, so that $f = h \circ c$ is 1-weakly convex. ◊

Truncated Models. The prox-linear model **5** may be challenging to implement for complex compositions (e.g., deep learning). If instead we know a lower bound on f, we may incorporate this into the model

$$f_x(y; s) := \max \left\{ f(x; s) + \langle f'(x; s), y - x \rangle, \inf_{z \in \mathcal{X}} f(z; s) \right\}. \quad [6]$$

In our examples—linear and logistic regression, phase retrieval, and matrix completion (more generally, typical loss functions in machine learning)—we have $\inf_z f(z; s) = 0$. The assumption that we have a lower bound is thus rarely restrictive. This model satisfies the conditions C.i to C.iii, also satisfying the following condition.

C.iv) (Lower optimality): For all $s \in \mathcal{S}$ and $x, y \in \mathcal{X}$,

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

As we show, condition C.iv is sufficient to derive several optimality and stability properties.

Stability and Its Consequences

In our initial study of stability in optimization (8), we defined an algorithm as stable if its iterates remain bounded and then showed several consequences of this in convex optimization (which we review presently). Here, we develop 2 important extensions. First, we show that any model satisfying condition C.iv has stable iterates under mild assumptions, in strong contrast to models (e.g., linear) that fail the condition. Second, we develop an analogous stability theory for weakly convex functions, proving that accurate enough models are stable. In parallel to the convex case, stability suffices for more: It implies convergence (with an asymptotic rate) to stationary points for any model-based method on weakly convex functions. Let us formalize stability (8). A pair $(\mathcal{F}, \mathcal{P})$ is a collection of problems if $\mathcal{P}$ consists of probability measures on a sample space $\mathcal{S}$ and $\mathcal{F}$ of functions $f : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$.

Definition 1. *An algorithm generating iterates x_k according to the model-based update* **2** *is stable in probability for the class of problems $(\mathcal{F}, \mathcal{P})$ if for all $f \in \mathcal{F}$, $P \in \mathcal{P}$ defining $F(x) = \mathbb{E}_P[f(x; S)]$, and $\mathcal{X}^\star = \operatorname{argmin}_{x \in \mathcal{X}} F(x)$,*

$$\sup_k \operatorname{dist}(x_k, \mathcal{X}^\star) < \infty \quad \textit{with probability } 1. \tag{7}$$

Typically, stability **7** requires the standard assumptions

$$\alpha_k > 0, \quad \sum_{k \geq 1} \alpha_k = \infty, \quad \text{and} \quad \sum_{k \geq 1} \alpha_k^2 < \infty. \tag{8}$$

Even under these, models such as the linear model **3** and consequent subgradient method are unstable (ref. 8, section 3). They may even cause superexponential divergence.

Example 3 (divergence): Let $F(x) = e^x + e^{-x}$, $p < \infty$, and $\alpha_0 > 0$, and let α_k satisfy $\alpha_k \geq \alpha_0 k^{-p}$. Let $x_{k+1} = x_k - \alpha_k F'(x_k) = x_k - \alpha_k (e^{x_k} - e^{-x_k})$ be generated by the gradient method. For large x_1, $\log \frac{x_{k+1}}{x_k} \geq 2^k$ for all k. $\diamondsuit$

The Importance of Stability in Stochastic Convex Optimization. To set the stage for what follows, we begin by motivating the importance of stable procedures. Briefly, any stable APROX model converges for any convex function under weak assumptions, which we now elucidate. First, we make an assumption.

Assumption A2. *There exists $\mathsf{G}_{\text{big}} : \mathbb{R}_+ \to [0, \infty)$ such that for all $x \in \mathcal{X}$ and each measurable selection $f'(x; s) \in \partial f(x; s)$,*

$$\mathbb{E}\left[\left\|f'(x; S)\right\|_2^2\right] \leq \mathsf{G}_{\text{big}}(\|x\|_2).$$

Assumption A2 is equivalent to assuming $\mathbb{E}[\left\|f'(x; S)\right\|_2^2]$ is bounded on compact sets; it allows arbitrary growth as long as the subgradients have second moments.

Proposition 1 [Asi and Duchi (8), proposition 1]. *Assume that $f(\cdot; s)$ is convex for each $s \in \mathcal{S}$ and let Assumption A2 hold. Let the iterates x_k be generated by any method satisfying conditions C.i to C.iii and* **[8]**. *On the event $\sup_k \|x_k\| < \infty$, $\sum_k \alpha_k (F(x_k) - F(x^\star)) < \infty$ and $\operatorname{dist}(x_k, \mathcal{X}^\star) \xrightarrow{a.s.} 0$.*

Proposition 1 establishes convergence of stable procedures and also (via Jensen's inequality) provides asymptotic rates of convergence for weighted averages $\sum_k \alpha_k x_k / \sum_k \alpha_k$.

Stability is additionally important when the functions f are smooth: Any stable APROX method achieves asymptotically optimal convergence. In particular, let us assume F is $\mathcal{C}^2$ near $x^\star = \operatorname{argmin}_{\mathcal{X}} F(x)$ with $\nabla^2 F(x^\star) \succ 0$, and the $f(\cdot; s)$ have an $L(s)$ -Lipschitz gradient near $x^\star$ with $\mathbb{E}[L(S)^2] < \infty$.

Proposition 2 [Asi and Duchi (8), theorem 2]. *In addition to the conditions of Proposition 1, let the conditions of the previous paragraph hold. Then $\bar{x}_k = \frac{1}{k} \sum_{i=1}^k x_i$ satisfies*

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{d} \mathsf{N}\left(0, \nabla^2 F(x^\star)^{-1} \operatorname{Cov}(\nabla f(x^\star; S)) \nabla^2 F(x^\star)^{-1}\right).$$

This convergence is optimal for any method (26).

Stability of Lower-Bounded Models for Convex Functions. With these consequences of stability in hand—convergence and asymptotic optimality—it behooves us to provide conditions sufficient to guarantee stability. To that end, we show that lower-bounded models satisfying condition C.iv are stable in probability (*Definition 1*) for functions whose (sub)gradients grow at most polynomially. We begin with an assumption.

Assumption A3. *There exist $C < \infty$, $2 \leq p < \infty$ such that*

$$\mathbb{E}\left[\left\|f'(x; S)\right\|_2^2\right] \leq C(1 + \operatorname{dist}(x, \mathcal{X}^\star)^p), \quad \textit{all } x \in \mathcal{X},$$

and $\mathbb{E}[(f(x^\star; S) - \inf_{z \in \mathcal{X}} f(z; S))^{p/2}] \leq C$ for all $x^\star \in \mathcal{X}^\star$.

The analogous condition (27) for stochastic gradient methods holds for $p = 2$, or quadratic growth, without which the method may diverge. In contrast, *Assumption A3* allows polynomial growth; for example, the function $f(x) = x^4$ is permissible, while the gradient method may exponentially diverge even for stepsizes $\alpha_k = 1/k$. The key consequence of *Assumption A3* is that if it holds, truncated models are stable:

Theorem 1. *Assume the function $f(\cdot; s)$ is convex for each $s \in \mathcal{S}$. Let Assumption A3 hold and $\alpha_k = \alpha_0 k^{-\beta}$ with $\frac{p+2}{p+4} < \beta < 1$. Let x_k be generated by the iteration* **2** *with a model satisfying conditions C.i to C.iv. Then*

$$\sup_{k \in \mathbb{N}} \operatorname{dist}(x_k, \mathcal{X}^\star) < \infty \quad \textit{with probability } 1.$$

Theorem 1 shows that truncated methods enjoy the benefits of stability we outline in *Propositions 1* and *2* above. Thus, these models, whose updates are typically as cheap to compute as a stochastic gradient step (especially in the common case that $\inf_z f(z; s) = 0$) provide substantial advantage over methods using only (sub)gradient approximations.

Stability and Its Consequences for Weakly Convex Functions. We continue our argument that—if possible—it is beneficial to use more accurate models, even in situations beyond convexity, investigating the stability of proximal models (Eq. **4**) for weakly convex functions. Establishing stability in the weakly convex case requires a different approach to the convex case, as the iterates may not make progress toward a fixed optimal set. In this case, to show stability, we require an assumption bounding the size of $f'(x; S)$ relative to the population subgradient F'.

Assumption A4. *There exist $C_1, C_2 < \infty$ such that for all measurable selections $f'(x; s) \in \partial f(x; s)$ and $F'(x) \in \partial F(x)$,*

$$\mathbb{E}\left[\left\|f'(x; S) - F'(x)\right\|_2^2\right] \leq C_1 \left\|F'(x)\right\|_2^2 + C_2.$$

By providing a relative noise condition on f', *Assumption A4* allows for more than the typical class of functions with global Lipschitz properties (cf. ref. 5), such as the phase retrieval and matrix completion objectives (*Examples 1* and *2*). It can allow exponential growth, addressing the challenges in *Example 3*. For example, let $f(x; 1) = e^x$ and $f(x; 2) = e^{-x}$, where S is uniform in $\{1, 2\}$ so that $F(x) = \frac{1}{2}(e^x + e^{-x})$; then $\mathbb{E}[f'(x; S)^2] = 2F'(x)^2 + 1$.

To describe convergence and stability in nonconvex (even nonsmooth) settings, we require appropriate definitions. Finding global minima of nonconvex functions is computationally infeasible (28), so we follow established practice and consider convergence to stationary points via the Moreau envelope (5, 29). To formalize, for $x \in \mathbb{R}^n$ and $\lambda \geq 0$, the Moreau envelope and associated proximal map are

$$F_\lambda(x) := \inf_{y \in \mathcal{X}} \left\{F(y) + \frac{\lambda}{2} \|y - x\|_2^2\right\} \quad \text{and}$$
$$\operatorname{prox}_{F/\lambda}(x) := \operatorname*{argmin}_{y \in \mathcal{X}} \left\{F(y) + \frac{\lambda}{2} \|y - x\|_2^2\right\}.$$

For large λ, the minimizer $x^\lambda := \operatorname{prox}_{F/\lambda}(x)$ is unique whenever F is weakly convex. Adopting the techniques Davis and Drusvyatskiy (5) pioneer for weakly convex problems, we rely on the Moreau envelope's connections to (near) stationarity:

$$\begin{aligned} &\nabla F_\lambda(x) = \lambda(x - x^\lambda), \quad F(x^\lambda) \leq F(x), \\ &\operatorname{dist}(0, \partial F(x^\lambda)) \leq \left\|\nabla F_\lambda(x)\right\|_2. \end{aligned} \tag{9}$$

The 3 properties in **[9]** imply that any nearly stationary point x of F_λ—when $\|\nabla F_\lambda(x)\|_2$ is small—is close to a nearly stationary point x^λ of F. To prove convergence for weakly convex F, then, it suffices to show $\nabla F_\lambda(x_k) \to 0$.

Using full proximal models guarantees convergence.

Theorem 2. *Let Assumption A4 hold, let $\lambda < \infty$ satisfy $\mathbb{E}[\rho(S)] < \lambda$, and assume $\inf_{x \in \mathcal{X}} F(x) > -\infty$ and $\mathbb{E}[\rho(S)^2] < \infty$. Let x_k follow the iteration* **2** *with proximal model* **4** *and stepsizes* **8**. *Then there exists a random variable $G_\lambda < \infty$ satisfying*

$$F_\lambda(x_k) \to G_\lambda \quad \textit{and} \quad \sum_k \alpha_k \left\|\nabla F_\lambda(x_k)\right\|_2^2 < \infty \quad w.p.\ 1.$$

Theorem 2 shows that $F_\lambda(x_k)$ is bounded almost surely. Thus, if F is coercive, meaning $F(x) \uparrow \infty$ as $\|x\| \to \infty$, the Moreau envelope F_λ is coercive, yielding the following.

Corollary 1. *Let the conditions of Theorem 2 hold and let F be coercive. Then*

$$\sup_{k \in \mathbb{N}} \operatorname{dist}(x_k, \mathcal{X}^\star) < \infty \quad \textit{with probability } 1.$$

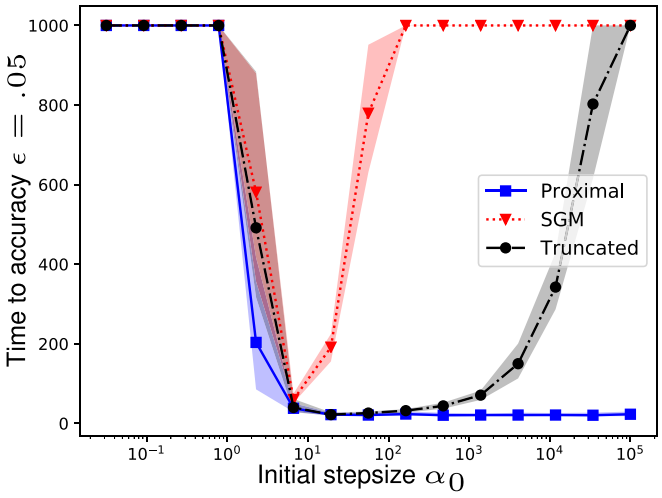

Fig. 1. The number of iterations to achieve ϵ accuracy versus initial stepsize α_0 for phase retrieval with $n = 50$, $m = 1{,}000$. SGM, stochastic gradient method.

In parallel with our development of the convex case, stability is sufficient to develop convergence results for any APROX method, highlighting its importance. Indeed, we can show that stable methods guarantee convergence, although for probability 1 convergence of the iterates, we require a slightly elaborate assumption (cf. refs. 9 and 30), which rules out pathological limits.

Assumption A5 (Weak Sard). *Let $\mathcal{X}_{\text{stat}} = \{x \mid 0 \in \partial F(x)\}$ be the collection of stationary points of F over $\mathcal{X}$. The Lebesgue measure of the image $F(\mathcal{X}_{\text{stat}})$ is zero.*

Under this assumption, APROX methods converge to stationary points whenever the iterates are stable.

Proposition 3. *Let Assumption A2 hold and the iterates x_k be generated by any method satisfying conditions C.i to C.iii and* **[8]**. *Assume that λ is large enough that $\mathbb{E}[\rho(S)] < \lambda$. There exists a finite random variable G_λ such that on the event that $\sup_k \|x_k\|_2 < \infty$, with probability 1 we have*

$$\sum_k \alpha_k \|\nabla F_\lambda(x_k)\|_2^2 < \infty \quad \text{and} \quad F_\lambda(x_k) \to G_\lambda. \qquad \textbf{[10]}$$

Under Assumption A5, then $\operatorname{dist}(x_k, \mathcal{X}_{\text{stat}}) \overset{a.s.}{\to} 0$ *and* $\|\nabla F_\lambda(x_k)\|_2 \overset{a.s.}{\to} 0$.

The condition **10** is enough to develop a conditional ℓ_2-convergence guarantee similar to what stochastic (sub)gradient methods achieve to stationary points for Lipschitz F (5, 31). Indeed, assume $\alpha_k = \alpha_0 k^{-\beta}$ for some $\beta \in (\frac{1}{2}, 1)$ and that the iterates x_k are stable; choose $I_k \in \{1, \ldots, k\}$ with probability $P(I_k = i) = \alpha_i / \sum_{j=1}^k \alpha_j$. Then inequality **10** shows

$$\limsup_k k^{1-\beta} \mathbb{E}\left[\left\|\nabla F_\lambda(x_{I_k})\right\|_2^2 \mid \mathcal{F}_k\right] < \infty \text{ with probability 1.}$$

Fast Convergence for Easy Problems

In many engineering and learning applications, solutions interpolate the data. Consider, for example, signal recovery problems with $b = Ax^\star$ or modern machine-learning applications, where frequently training error is zero (32, 33). We consider such problems here, showing how models that satisfy the lower-bound condition C.iv enjoy linear convergence, extending our earlier results (8) beyond convex optimization.

Definition 2. *Let $F(x) := \mathbb{E}_P[f(x; S)]$. Then F is easy to optimize if for each $x^\star \in \mathcal{X}^\star$ and P almost all $s \in \mathcal{S}$,*

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

For such problems, we can guarantee progress toward minimizers for appropriate f, as the following lemma shows.

Lemma 1. *Let F be easy to optimize (Definition 2). Let x_k be generated by the updates* **2** *using a model satisfying conditions C.i to C.iv. Then for any $x^\star \in \mathcal{X}^\star$,*

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

Lemma 1 allows us to prove fast convergence as long as f grows quickly enough away from $x^\star$; a sufficient condition for us is a sharp growth condition away from the optimal set $\mathcal{X}^\star$. To meld with *Lemma 1*, we consider the following:

Assumption A6 (Expected Sharp Growth). *There exist constants $\lambda_0, \lambda_1 > 0$ such that for $\alpha \in \mathbb{R}_+$, $x \in \mathcal{X}$, and $x^\star \in \mathcal{X}^\star$,*

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

Assumption A6 is tailored to *Lemma 1*, so we discuss a few situations where it holds. One sufficient condition is the small-ball condition that there

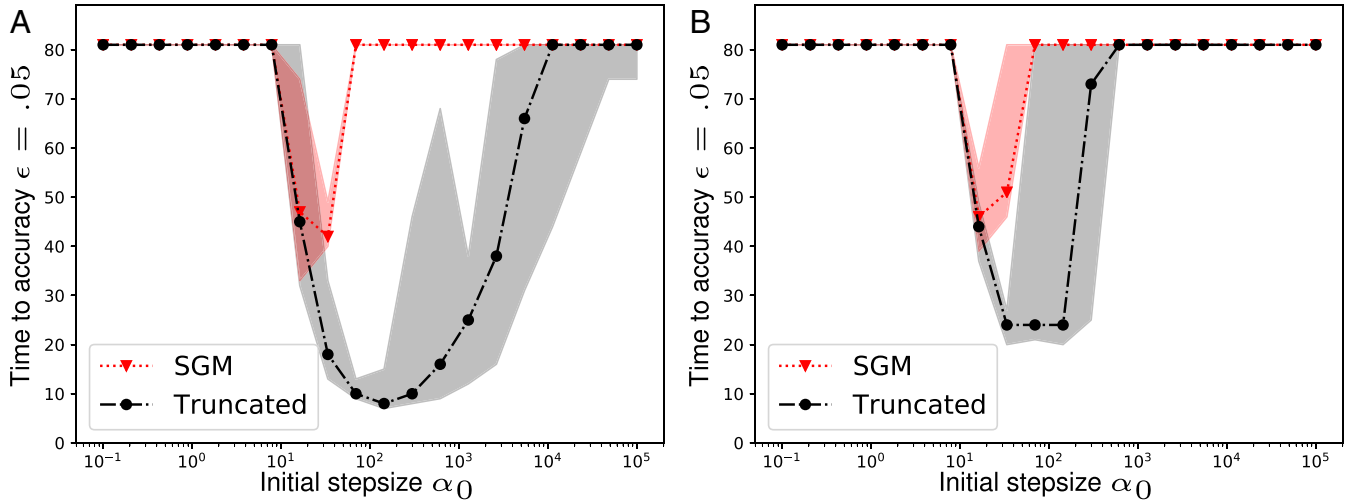

Fig. 2. Number of iterations to achieve ϵ accuracy versus initial stepsize α_0 for matrix completion with $m = 2{,}000$, $n = 2{,}400$, $r = 5$. Shown are estimated ranks (*A*) $\hat{r} = 5$ and (*B*) $\hat{r} = 10$.

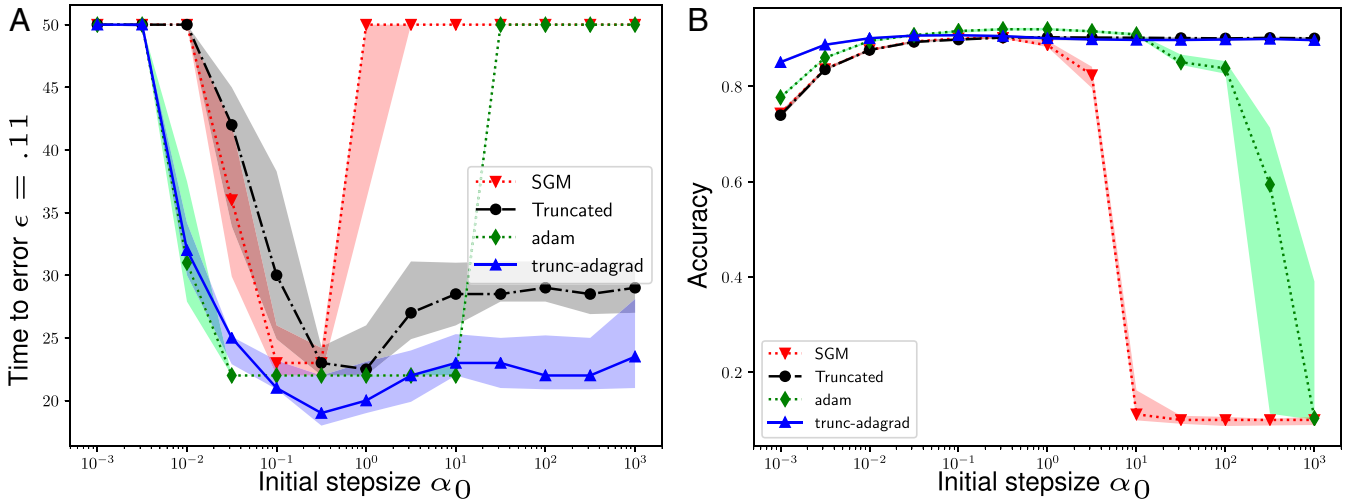

Fig. 3. (*A*) The number of iterations to achieve ϵ test error versus initial stepsize α_0 for CIFAR10. (*B*) The best achieved accuracy after $T = 50$ epochs.

exists C such that $\mathbb{P}(f(x; S) - f(x^\star; S) \geq \epsilon \operatorname{dist}(x, \mathcal{X}^\star)) \geq 1 - C\epsilon$ for $\epsilon > 0$ and $\mathbb{E}[\left\|f'(x; S)\right\|_2^2] \leq C(1 + \operatorname{dist}(x, \mathcal{X}^\star)^2)$. We can be more explicit:

Example 4 (Example 1 continued): Consider the (real-valued) phase retrieval problem with objective $f(x; (a, b)) = |\langle a, x\rangle^2 - b|$. Assume the vectors $a_i \in \mathbb{R}^n$ are drawn from a distribution satisfying the small-ball condition $P(|\langle a_i, u\rangle| \geq \epsilon \|u\|_2) \geq 1 - \epsilon$ for $\epsilon > 0$ and any $u \in \mathbb{R}^n$ and additionally that $\mathbb{E}[\|a_i\|_2^2] \leq M^2 n$ and $\mathbb{E}[\langle a_i, x\rangle^2] \leq M^2 \|x\|_2^2$ for some $M < \infty$. For a sample of size m, *Assumption A3* holds with high probability for the objective $F(x) = \frac{1}{m}\sum_{i=1}^m f(x; (a_i, b_i))$ with $\lambda_0 = c\|x^\star\|_2$, and $\lambda_1 = \frac{c}{16M^4 n}$, for a numerical constant $c > 0$. The full calculation is in *SI Appendix*. ◊

The following proposition is our main result in this section, showing lower-bounded models may enjoy linear convergence.

Proposition 4. *Let Assumption A6 hold and x_k be generated by the stochastic iteration* **2** *using any model satisfying conditions C.i to C.iv, where the stepsizes α_k satisfy $\alpha_k = \alpha_0 k^{-\beta}$ for some $\beta \in (0, 1)$. If $f(\cdot; S_k)$ is $\rho(S_k)$-weakly convex with $\mathbb{E}[\rho(S_k)] = \overline{\rho}$, then for any $m \in \mathbb{N}$ and $\epsilon > 0$, there exists a finite random variable $V_{\infty,m} < \infty$ such that*

$$\frac{\operatorname{dist}(x_k, \mathcal{X}^\star)^2}{(1 - \lambda_1)^k} \cdot 1\left\{\max_{m \leq i \leq k-1} \operatorname{dist}(x_i, \mathcal{X}^\star) \leq \frac{\lambda_0}{(1 + \epsilon)\overline{\rho}}\right\} \xrightarrow{a.s.} V_{\infty,m}.$$

When the functions f are convex, we have $\overline{\rho} = 0$, so that *Proposition 4* guarantees linear convergence for easy problems. In the case that $\overline{\rho} > 0$, the result is conditional: If an APROX method converges to one of the sharp minimizers of f, then this convergence is linear (i.e., geometrically fast). In the case of phase retrieval, we can guarantee convergence:

Example 5 (Example 4 continued): Let $A \in \mathbb{R}^{m \times n}$ be a matrix with rows a_i that satisfy the conditions of *Example 4*. For $F(x) = \frac{1}{m}\left\||Ax|^2 - |Ax^\star|^2\right\|_1$ where $m \gtrsim n$, the truncated model **6** requires overall computation time $O(mn \log \frac{1}{\epsilon})$ to achieve an ϵ-accurate solution to phase retrieval, which is the best-known time complexity. See proof in *SI Appendix*. ◊

Experiments

An important question in the development of any optimization method is its sensitivity to algorithm parameters. Consequently, we conclude by experimentally examining convergence time and robustness of each of our optimization methods. We consider each of the models in this paper: the stochastic gradient method (i.e., the linear model **3**), the proximal model **4**, the prox-linear model **5**, and the (lower) truncated model **6**.

We test both convergence time and, dovetailing with our focus in this paper, robustness to stepsize for several problems: phase retrieval, matrix completion, and 2 classification problems using deep learning. We consider stepsize sequences of the form $\alpha_k = \alpha_0 k^{-\beta}$ and perform K iterations over a wide range of different initial stepsizes α_0. (For brevity, we present results only for the power $\beta = 0.6$; experiments with varied $\beta \in (\frac{1}{2}, 1)$ were similar.) For a fixed accuracy $\epsilon > 0$, we record the number of steps k to achieve $F(x_k) - F(x^\star) \leq \epsilon$, reporting these times (where we terminate each run at iteration K). We perform T experiments for each initial

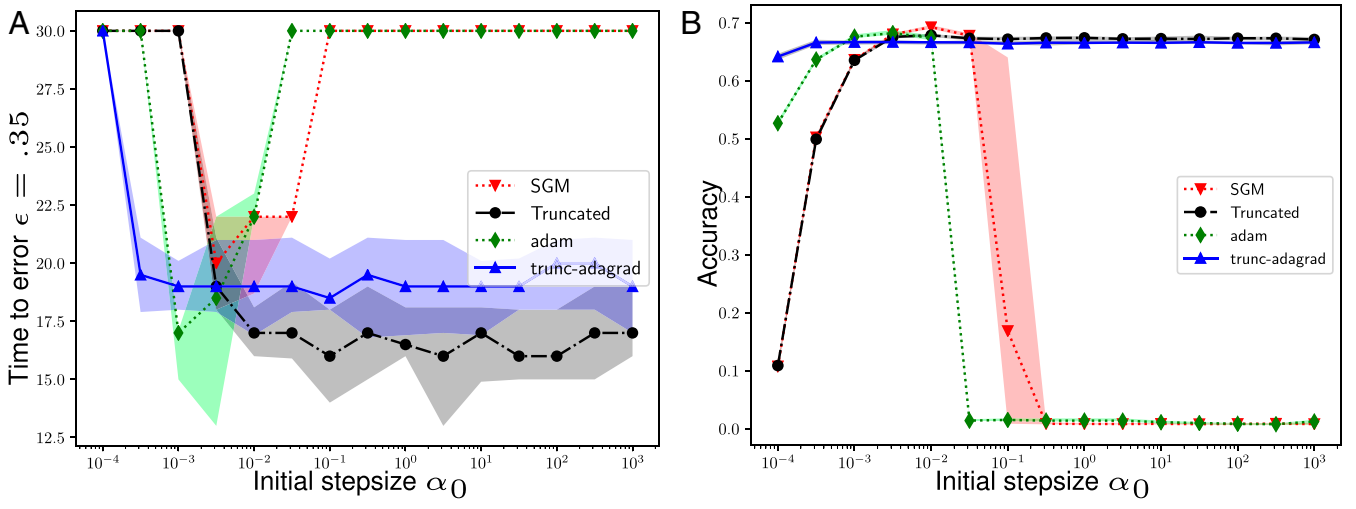

Fig. 4. (*A*) The number of iterations to achieve ϵ test error versus initial stepsize α_0 for the Stanford dogs dataset. (*B*) The best achieved accuracy after $T = 30$ epochs.

stepsize choice, reporting median time to ϵ accuracy and 90% confidence intervals.

Phase Retrieval. We start our experiments with the phase retrieval problem in *Examples 1* and *4*, focusing on the real case for simplicity, where we are given $A \in \mathbb{R}^{m \times n}$ with rows $a_i \in \mathbb{R}^n$ and $b = (Ax^\star)^2 \in \mathbb{R}^m_+$ for some $x^\star \in \mathbb{R}^n$. Our objective is the nonconvex and nonsmooth function

$$F(x) = \frac{1}{m} \sum_{i=1}^{m} \left| \langle a_i, x \rangle^2 - b_i \right|.$$

We sample the entries the vectors a_i and $x^\star$ i.i.d. N(0, I_n).

We present the results in Fig. 1, comparing the stochastic gradient method **3**, the proximal method **4**, and the truncated method **6** (whose updates are identical to the prox-linear model **5** in this case). The plots demonstrate the expected result that the stochastic gradient method has good performance in a narrow range of stepsizes, $\alpha_1 \approx 10$ in this case, while better approximations for APROX yield convergence over a large range of stepsizes. The truncated model **6** exhibits oscillation for large stepsizes, in contrast to the exact model **4**.

Matrix Completion. For our second experiment, we investigate APROX procedures for the matrix completion problem of *Example 2*. In this setting, we are given $M = X_\star Y_\star^T$, for $X_\star \in \mathbb{R}^{m \times r}$ and $Y_\star \in \mathbb{R}^{n \times r}$, and a set of indexes $\Omega \subset [m] \times [n]$. We aim to recover M observing only $\{M_{ij}\}_{i,j \in \Omega}$, so our goal is to

$$\text{minimize } F(X, Y) := \frac{1}{|\Omega|} \sum_{i,j \in \Omega} \left| X_i^T Y_j - M_{i,j} \right|.$$

We optimize over matrices $X \in \mathbb{R}^{m \times \hat{r}}$ and $Y \in \mathbb{R}^{n \times \hat{r}}$, where the estimated rank $\hat{r} \geq r$. We generate $X_\star$ and $Y_\star$ by drawing their entries i.i.d. N(0, 1), choosing Ω uniformly at random of size $|\Omega| = 5(nr + mr)$. We present the timing results in Fig. 2, which tells a similar story to Fig. 1: Better approximations, such as the truncated models (which again yield identical updates to the prox-linear models **5**), are significantly more robust to stepsize specification. The proximal update requires solving a nontrivial quartic, so we omit it.

Neural Networks. As one of our main motivations is to address the extraordinary effort—in computational and engineering hours—spent carefully tuning optimization methods, we would be remiss to avoid experiments on deep neural networks. Therefore, in our last set of experiments, we test the performance of our models for training neural networks for classification tasks over the CIFAR10 dataset (34) and the fine-grained 128-class Stanford dog multiclass recognition task (35). For our CIFAR10 experiment, we use the Resnet18 architecture (36); we replace the rectified linear unit (RELU) activations internal to the architecture with exponentiated linear units (ELUs) (37) so that the loss is of composite form $f = h \circ c$ for h convex and c smooth. For Stanford dogs we use the VGG16 architecture (38) pretrained on Imagenet (39), again substituting ELUs for RELU activations. For this experiment, we also test a modified version of the truncated method, TRUNCADAGRAD, which uses the truncated model in iteration **2** and a diagonally scaled Euclidean distance (40), updating at iteration k by setting x_k to minimize

$$[f(x_k; S_k) + \langle g_k, x - x_k \rangle]_+ + \frac{1}{2\alpha_0} (x - x_k)^T H_k (x - x_k),$$

where $H_k = \text{diag}(\sum_{i=1}^k g_i g_i^T)^{1/2}$ for $g_i = f'(x_i; S_i)$. This update requires no more of standard deep-learning software than computing a gradient (backpropagation) and loss. We also compare to ADAM, the default optimizer in TensorFlow (41).

Figs. 3 and 4 show our results for the CIFAR10 and Stanford dogs datasets, respectively. Fig. 3*A* and 4*A* give the number of iterations required to achieve ϵ test-classification error (on the highest or "top-1" predicted class), while Figs. 3*B* and 4*B* show the maximal accuracy each procedure achieves for a given initial stepsize α_0. The plots demonstrate the sensitivity of the standard stochastic gradient method to stepsize choice, which converges only for a small range of stepsizes, in both experiments. ADAM exhibits better robustness for CIFAR10, while it is extremely sensitive in the second experiment (Fig. 4), converging only for a small range of stepsizes—this difference in sensitivities highlights the importance of robustness. In contrast, our procedures using the truncated model are apparently robust for all large enough stepsizes. Figs. 3*B* and 4*B* show additionally that the maximal accuracy the 2 truncated methods achieve changes only slightly for $\alpha_0 \geq 10^{-1}$, again in strong contrast to the other methods, which achieve their best accuracy only for a small range of stepsizes.

These results reaffirm the insights from our theoretical results and experiments: It is important and possible to develop methods that enjoy good convergence guarantees and are robust to algorithm parameters.

Data Availability. All data discussed in this paper are available at GitHub (https://github.com/HilalAsi/APROX-Robust-Stochastic-Optimization-Algorithms) (42).

ACKNOWLEDGMENTS. H.A. and J.C.D. were supported by National Science Foundation (NSF)-CAREER Award CCF-1553086, Office of Naval Research Young Investigator Program Award N00014-19-2288, and the Stanford DAWN Consortium.

1. E. Real, A. Aggarwal, Y. Huang, Q. V. Le, "Regularized evolution for image classifier architecture search" in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, P. Stone, Ed. (AAAI Press, Palo Alto, CA, 2019), vol. 33, pp. 4780–4789.
2. B. Zoph, Q. V. Le, "Neural architecture search with reinforcement learning" in *Proceedings of the Fifth International Conference on Learning Representations*, Y. Bengio, Y. LeCun, Eds. (ICLR, 2017).
3. J. Collins, J. Sohl-Dickstein, D. Sussillo, Capacity and trainability in recurrent neural networks. arXiv:1611.09913 [stat.ML] (29 November 2016).
4. R. T. Rockafellar, R. J. B. Wets, *Variational Analysis* (Springer, New York, NY, 1998).
5. D. Davis, D. Drusvyatskiy, Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.* **29**, 207–239 (2019).
6. H. Robbins, S. Monro, A stochastic approximation method. *Ann. Math. Stat.* **22**, 400–407 (1951).
7. A. Nemirovski, A. Juditsky, G. Lan, A. Shapiro, Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19**, 1574–1609 (2009).
8. H. Asi, J. C. Duchi, Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM J. Optim.* **29**, 2257–2290 (2019).
9. J. C. Duchi, F. Ruan, Stochastic methods for composite and weakly convex optimization problems. *SIAM J. Optim.* **28**, 3229–3259 (2018).
10. B. Martinet, Regularisation d'inéquations variationelles par approximations succesives. *Revue Francaise d'Informatique et de Recherche Operationelle* **4**, 154–158 (1970).
11. R. T. Rockafellar, Monotone operators and the proximal point algorithm. *SIAM J. Control Optim.* **14**, 877–898 (1976).
12. O. Güler, On the convergence of the proximal point algorithm for convex minimization. *SIAM J. Control Optim.* **29**, 403–419 (1991).
13. H. H. Bauschke, P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* (Springer, 2011), vol. 408.
14. J. Hiriart-Urruty, C. Lemaréchal, *Convex Analysis and Minimization Algorithms I & II* (Springer, New York, NY, 1993).
15. J. F. Bonnans, A. Shapiro, *Perturbation Analysis of Optimization Problems* (Springer, 2000).
16. B. Widrow, M. E. Hoff, "Adaptive switching circuits" in *1960 IRE WESCON Convention Record* (IRE [Institute of Radio Engineers], 1960), pp. 96–104M, Reprinted in Neurocomputing, 1988.
17. A. H. Sayed, *Fundamentals of Adaptive Filtering* (John Wiley & Sons, 2003).
18. S. Shalev-Shwartz, T. Zhang, "Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization" in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing, T. Jebara, Eds. (PMLR, 2014), vol. 32, pp. 64–72.
19. H. Lin, J. Mairal, Z. Harchaoui, Catalyst acceleration for first-order convex optimization: From theory to practice. *J. Mach. Learn. Res.* **18**, 1–54 (2018).
20. A. Patrascu, I. Necoara, Nonasymptotic convergence of stochastic proximal point algorithms for constrained convex optimization. *J. Mach. Learn. Res.* **18**, 1–42 (2018).
21. D. P. Bertsekas, Incremental proximal methods for large scale convex optimization. *Math. Program. Ser. B* **129**, 163–195 (2011).
22. R. Fletcher, A model algorithm for composite nondifferentiable optimization problems. *Math. Program. Study* **17**, 67–76 (1982).
23. Y. Schechtman *et al.*, Phase retrieval with application to optical imaging. *IEEE Signal Process. Mag.* **32**, 87–109 (2015).
24. J. Duchi, F. Ruan, Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval.*Inform. Infer. J. IMA* **8**, 471–529 (2018).
25. E. J. Candes, B. Recht, Exact matrix completion via convex optimization. *Found. Comput. Math.* **9**, 717–772 (2008).
26. J. C. Duchi, F. Ruan, Asymptotic optimality in stochastic optimization. arXiv: 1612.05612 (16 December 2016).
27. B. T. Polyak, A. B. Juditsky, Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30**, 838–855 (1992).
28. A. Nemirovski, D. Yudin, *Problem Complexity and Method Efficiency in Optimization* (Wiley, 1983).
29. D. Drusvyatskiy, A. Lewis, Error bounds, quadratic growth, and linear convergence of proximal methods. *Math. Oper. Res.* **43**, 919–948 (2018).
30. D. Davis, D. Drusvyatskiy, S. Kakade, J. D. Lee, Stochastic subgradient method converges on tame functions (Springer, New York, NY, 2019).

31. S. Ghadimi, G. Lan, Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.* **23**, 2341–2368 (2013).
32. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
33. M. Belkin, D. Hsu, P. Mitra, "Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate" in *Advances in Neural Information Processing Systems*, S. Bengio, Ed. (Curran Associates, Inc., 2018), vol. 31, pp. 2300–2311.
34. A. Krizhevsky, "Learning multiple layers of features from tiny images" (Tech Rep., University of Toronto, Toronto, ON, Canada, 2009).
35. A. Khosla, N. Jayadevaprakash, B. Yao, F. F. Li, "Novel dataset for fine-grained image categorization" in *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, N. Pinto (IEEE, Piscataway, NJ, 2011).
36. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, E. Mortensen, K. Saenko, Eds. (IEEE, Piscataway, NJ, 2016), pp. 770–778.
37. D. A. Clevert, T. Unterthiner, S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)" in *Proceedings of the Fourth International Conference on Learning Representations*, Y. Bengio, Y. LeCun, Eds. (ICLR, 2016).
38. K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition" in *Proceedings of the Third International Conference on Learning Representations*, Y. Bengio, Y. LeCun, Eds. (ICLR, 2015).
39. J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* P. Flynn, E. Mortensen, Eds. (IEEE, Piscataway, NJ, 2009), pp. 248–255.
40. J. C. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12**:2121–2159 (2011).
41. M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning" in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, K. Keeton, T. Roscoe, Eds. (USENIX Association, 2016), pp. 265–283.
42. H. Asi, J. Duchi, APROX: Robust Stochastic Optimization Algorithms. GitHub. https://github.com/HilalAsi/APROX-Robust-Stochastic-Optimization-Algorithms. Deposited 18 October 2019.