Mathematical Programming
https://doi.org/10.1007/s10107-023-02003-w

FULLLENGTHPAPER

Series A

Stochastic algorithms with geometric step decay converge linearly on sharp functions

Damek Davis[1] Dmitriy Drusvyatskiy[2] Vasileios Charisopoulos[1]

Received: 29 July 2019 / Accepted: 7 July 2023

Abstract

Stochastic (sub)gradient methods require step size schedule tuning to perform well in practice. Classical tuning strategies decay the step size polynomially and lead to optimal sublinear rates on (strongly) convex problems. An alternative schedule, popular in nonconvex optimization, is called *geometric step decay* and proceeds by halving the step size after every few epochs. In recent work, geometric step decay was shown to improve exponentially upon classical sublinear rates for the class of *sharp* convex functions. In this work, we ask whether geometric step decay similarly improves stochastic algorithms for the class of sharp weakly convex problems. Such losses feature in modern statistical recovery problems and lead to a new challenge not present in the convex setting: the region of convergence is local, so one must bound the probability of escape. Our main result shows that for a large class of stochastic, sharp, nonsmooth, and nonconvex problems a geometric step decay schedule endows well-known algorithms with a local linear (or nearly linear) rate of convergence to global minimizers. This guarantee applies to the stochastic projected subgradient, proximal point, and prox-linear algorithms. As an application of our main result, we analyze two statistical recovery tasks—phase retrieval and blind deconvolution—and match the best known guarantees under Gaussian measurement models and establish new guarantees under heavy-tailed distributions.

Research of Drusvyatskiy was supported by the NSF DMS 1651851 and CCF 1740551 awards. Research of Damek Davis supported by an Alfred P. Sloan research fellowship and NSF-DMS award 2047637.

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

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

Vasileios Charisopoulos
https://people.orie.cornell.edu/vc333/

[1] School of ORIE, Cornell University, Ithaca, NY 14850, USA

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

Published online: 05 September 2023

Mathematics Subject Classification 65K05 · 65K10 · 90C15 · 90C30

1 Introduction

Stochastic (sub)gradient methods form the algorithmic core of much of modern statistical and machine learning. Such algorithms are typically sensitive to algorithm parameters, and require extensive step size tuning to achieve adequate performance. Classical tuning strategies decay the step size polynomially and lead to optimal sublinear rates of convergence on convex and strongly convex problems

$$\min_{x \in X} \; f(x) = \mathsf{E}_z[f(x, z)], \tag{SO}$$

where the loss functions $f(\cdot, z)$ are convex and X is a closed convex set [45]. An alternative schedule, called *geometric step decay*, decreases the step size geometrically by halving it after every few epochs. In recent work [66], geometric step decay was shown to improve exponentially upon classical sublinear rates under the sharpness assumption:

$$f(x) - \min_X f \geq \mu \cdot \operatorname{dist}(x, X^*) \qquad \forall x \in X, \tag{1.1}$$

where $\mu > 0$ is some constant and $X^* = \operatorname{argmin}_{x \in X} f(x)$ is the solution set. This result complements early works of Goffin [27] and Shor [62], which show that deterministic subgradient methods converge linearly on sharp convex functions if their step sizes decay geometrically. The work [66] also reveals a departure from the smooth strongly convex setting, where deterministic linear rates degrade to sublinear rates when the gradient is corrupted by noise [45].

Beyond the convex setting, sharp problems appear often in nonconvex statistical recovery problems, for example, in robust matrix sensing [40], phase retrieval [18, 20], blind deconvolution [8], and quadratic/bilinear sensing and matrix completion [7]. For such problems, sharpness is surprisingly common and corresponds to strong identifiability of the statistical model. In such settings, sharpness endows standard deterministic first-order algorithms with rapid local convergence guarantees, enabling the recovery of signals at optimal or near-optimal sample complexity in a variety of nonconvex problems. Despite this, we do not know whether stochastic algorithms equipped with geometric step decay—or any other step size schedule—linearly converge on sharp nonconvex problems.[1] Such algorithms, if available, could pave the way for new sample efficient strategies for these and other statistical recovery problems.

The main result of this work shows that for a large class of stochastic, sharp, nonsmooth, and nonconvex problems

a geometric step decay schedule endows well-known algorithms

[1] There are some exceptions for highly-structured problems, such as interpolation problems where all terms in the expectation share the same minimizer [3, 63].

with a local nearly linear[2] rate of convergence to minimizers.

This guarantee takes hold as soon as the algorithms are initialized in a small neighborhood around the set of minimizers and applies, for example, to the stochastic projected subgradient, proximal point, and prox-linear algorithms. Beyond sharp growth (1.1), we also analyze losses that grow sharply away from some closed set S, which is strictly larger than X^*. Such sets S are akin to "active manifolds" in the sense of Lewis [38] and Wright [64]. For example, the loss $f(x, y) = x^2 + |y|$ is not sharp relative to its minimizer, but is sharp relative to the x-axis. For these problems, our algorithms converge nearly linearly to the set S. Finally, we illustrate the result with two statistical recovery problems: phase retrieval and blind deconvolution. For these recovery tasks, our results match the best known computational and sample complexity guarantees under Gaussian measurement models and establish new guarantees under heavy-tailed distributions.

Related work

Our paper is closely related to a number of influential techniques in stochastic, convex, and nonlinear optimization. We now survey these related topics.

Stochastic model-based methods. In this work, we use algorithms that iteratively sample and minimize simple stochastic convex models of the loss function. Throughout, we call these methods *model-based algorithms*. Such algorithms include the stochastic projected subgradient, prox-linear, and proximal point methods. Stochastic model-based algorithms are known to converge globally to stationary points at a sublinear rate on a large class of nonsmooth and nonconvex problems [11, 19]. Some model-based algorithms also possess superior stability properties and can be less sensitive to step size choice than traditional stochastic subgradient methods [3, 4].

Geometrically decaying learning rate in deterministic optimization. polynomially decaying step-sizes are common in stochastic optimization [35, 53, 55]. In contrast, we develop algorithms with step sizes that decay geometrically. Geometrically decaying step sizes were first analyzed in convex optimization by Shor [62, Thm 2.7, Sec. 2.3] and Goffin [27]. This step size schedule is also closely related to the step size rules of Eremin [21] and Polyak [52]. Similar schedules are known to accelerate convex subgradient methods under Hölder growth as shown in [32, 67]. Geometrically decaying step sizes for deterministic subgradient methods under weak convexity were systematically studied in [12]. The method of this work succeeds under similar assumptions on the population objective as in [12]: sharpness (A2), Lipschitz continuity (A5), and weak convexity (A3), (A4). In contrast to [12] which analyzes a deterministic subgradient method (based on linear approximations of the objective), the method of this work requires only stochastic estimates of the objective and converges under a wider family of *nonlinear* approximations of the objective (see Example 2.1). Surprisingly,

[2] Here, the term "nearly linear" signifies that the *oracle / sample complexity* of the method is $\tilde{O}(\log^3(1/\varepsilon))$ – i.e., it depends polylogarithmically on $1/\varepsilon$, in contrast to standard "linear" convergence where in the oracle complexity scales as $O(\log(1/\varepsilon))$.

the method only incurs an additional cost of $\log^2(1/\varepsilon)$ (stochastic) oracle evaluations, which arises due to our restart scheme. It is an intriguing open question whether one can remove this additional dependence on $\log^2(1/\varepsilon)$.

Geometric step decay in stochastic optimization. The geometric step decay schedule is common in practice: for example, see Krizhevsky et. al. [33] and He et al. [29]. It is a standard option in the popular deep learning libraries, such as Pytorch [49] and TensorFlow [1]. Geometric step decay has been analyzed in a number of recent papers in stochastic convex optimization, including [5, 25, 26, 34, 66, 68]. Among these papers, the work [66] relates most to ours. There, the authors propose two geometric step decay strategies that converge linearly on convex functions that are sharp and have bounded stochastic subgradients. These algorithms either use a moving ball constraint or follow a proximal-point type procedure. We follow the latter strategy, too, at least to obtain high-probability guarantees. The paper [66] differs from our work in that they assume convexity and a uniform bound on the stochastic subgradients. In contrast, we do not assume convexity and only assume that stochastic subgradients have a finite second moment. We are aware of only one subgradient method for stochastic nonconvex problems that converges linearly [3] under favorable assumptions. In [3], the authors develop a "clipped" subgradient method, which resembles a safeguarded stochastic Polyak step. In contrast to our work, their algorithms converge linearly only under "perfect interpolation," meaning that all terms in the expectation share a minimizer. Formally, the interpolation condition therein stipulates that, almost surely over z,

$$\inf_x f(x, z) = f(x_*, z), \quad \text{for any } x_* \in \arg\min_x \mathbb{E}[f(x, z)].$$

We do not make this assumption here. Indeed, the statistical recovery problems from Sect. 4 do not satisfy the interpolation condition in the presence of corruptions.

Restarts in deterministic optimization. Restart techniques have a long history in nonlinear programming, such as for conjugate gradient and limited memory quasi-Newton methods. They have also been used more recently to improve the complexity of algorithms in deterministic convex optimization. For example, restart schemes can accelerate sublinear convergence rates for convex problems that satisfy growth conditions as shown by Nesterov [47] and Nemirovskii and Nesterov [44], Renegar [54], O'Donoghue and Candes [48], Roulet and d'Aspremont [58], Freund and Lu [24], and Fercoq and Qu [22, 23] and others. In the nonconvex setting, stochastic restart methods are challenging to analyze, since the region of linear convergence is local. To overcome this challenge, one must bound the probability that the iterates leave this region. One of our main technical contributions is a technique for bounding this probability.

Finite sums. For finite sums, stochastic algorithms that converge linearly are more common. For example, for finite sums that are sharp and convex, Bertsekas and Nedić [43] prove that an incremental Polyak-type algorithm converges linearly. For finite sums that are smooth and strongly convex, variance reduced methods, such as SAG [59], SAGA [14], SDCA [60], SVRG [31], MISO/Finito [15, 41], SMART [10]

and their proximal extensions converge linearly. The algorithms we develop here do not assume a finite sum structure.

Verifying sharpness. Sharp growth is a central assumption in this work. This property is surprisingly common in statistical recovery problems. For example, sharp growth has been established for robust matrix sensing [40], phase retrieval [18, 20], blind deconvolution [8], quadratic and bilinear sensing and matrix completion [7] problems. Consequently, the results of this paper apply in these settings.

Notation

We will mostly follow standard notation used in convex analysis and stochastic optimization. Throughout, the symbol $\mathbb{R}^d$ will denote a d-dimensional Euclidean space with the inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\| = \sqrt{\langle x, x \rangle}$. We denote the open ball of radius $\varepsilon > 0$ around a point $x \in \mathbb{R}^d$ by the symbol $B_\varepsilon(x)$; moreover, we write $\mathbb{S}^{d-1}$ for the unit sphere in $\mathbb{R}^d$. Given an event A, we write A^c for its complement. We also let 1_A denote the indicator function of A, meaning a function that evaluates to one on A and zero off it. For any set $Q \subset \mathbb{R}^d$, the *distance function* and the *projection map* are defined by

$$\operatorname{dist}(x, Q) := \inf_{y \in Q} \|y - x\| \quad \text{and} \quad \operatorname{proj}_Q(x) := \operatorname{argmin}_{y \in Q} \|y - x\|,$$

respectively. Consider a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ and a point x, with $f(x)$ finite. The *Fréchet subdifferential* of f at x, denoted by $\partial f(x)$, consists of all vectors $v \in \mathbb{R}^d$ satisfying

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

A function f is called ρ-*weakly convex* on an open convex set U if the perturbed function $f + \frac{\rho}{2}\|\cdot\|^2$ is convex on U. The subgradients of such functions automatically satisfy the uniform approximation property (see [56]):

$$f(y) \geq f(x) + \langle v, y - x \rangle - \frac{\rho}{2}\|y - x\|^2 \quad \text{for all } x, y \in U,\ v \in \partial f(x).$$

2 Algorithms, assumptions, and main results

In this section, we formalize our target problem and introduce algorithms to solve it. We then outline our main results. The complete theorem statements and proofs appear in Sect. 3. Throughout, we consider the minimization problem

$$\min_{x \in X} f(x). \tag{2.1}$$

for some function $f : \mathbb{R}^d \to \mathbb{R}$ and a closed convex set $X \subseteq \mathbb{R}^d$. We define the set $X^* := \operatorname{argmin}_{x \in X} f(x)$ and assume it to be nonempty. We also fix a probability space $(\Omega, \mathcal{F}, P)$ and equip $\mathbb{R}^d$ with the Borel σ-algebra and make the following assumption.

(A1) **(Sampling)** It is possible to generate i.i.d. realizations $z_1, z_2, \ldots \sim P$.

The algorithms we develop rely on a stochastic oracle model for Problem (SO) that was recently introduced in [11]. These algorithms assume access to a family of functions $f_x(\cdot, z)$—called stochastic models—indexed by basepoints $x \in \mathbb{R}^d$ and random elements $z \sim P$. Given these models, the generic *stochastic model-based algorithm* of [11] iterates the steps:

$$\begin{aligned} &\text{Sample } z_k \sim P; \\ &\text{Set } y_{k+1} = \operatorname{argmin}_{y \in X} \; f_{y_k}(y, z_k) + \frac{1}{2\alpha} \|y - y_k\|^2 . \end{aligned} \tag{2.2}$$

In this work, model-based algorithms form the core of the following restart strategy: given inner and outer loop sizes K and T, respectively, as well as initial stepsize $\alpha_0 > 0$, perform:

$$\begin{aligned} &\text{For } t = 0, \ldots, T-1: \\ &\quad \text{Initialize } y_0 = x_t, \alpha = 2^{-t}\alpha_0, \text{ and run } K \text{ iterations of (2.2);} \\ &\quad \text{Sample } x_{t+1} \text{ uniformly from } y_0, \ldots, y_K. \end{aligned} \tag{2.3}$$

This restart strategy is common in machine learning practice and is called *geometric step decay*. Restart schemes date back to the fundamental work of Nesterov [47] and Nemirovskii and Nesterov [44] and more recently appear in [22, 23, 25, 48, 54, 58, 66]. These strategies often improve the convergence guarantees of the algorithm they restart under growth assumptions, for example, by boosting an algorithm that converges sublinearly to one that converges linearly. In this work, we will show that restart scheme (2.3) similarly improves (2.2) for a large class of weakly convex stochastic optimization problems.

2.1 Assumptions

In this section, we formalize our assumptions on sharp growth of (2.1) as well as on accuracy, regularity, and Lipschitz continuity of the models.

Sharp Growth

We assume that $f(\cdot)$ grows sharply as x moves in the direction normal to a closed set S.

(A2) **(Sharpness)** There exists a constant $\mu > 0$ and a closed set $S \subseteq X$ satisfying $X^* \subseteq S$ such that the following bound holds:

$$f(x) - \inf_{z \in \operatorname{proj}_S(x)} f(z) \geq \mu \cdot \operatorname{dist}(x, S) \quad \forall x \in X. \tag{2.4}$$

This property generalizes the classical sharp growth condition (1.1), where $S = X^*$. The setting $S = X^*$ is well-studied in nonlinear programming and often underlies rapid convergence guarantees for deterministic local search algorithms. Beyond the classical setting, S could be a sublevel set of f or an "active manifold" in the sense of Lewis [38]; see Sect. D. When $S = X^*$, we design stochastic algorithms that do not necessarily converge to global minimizers, but instead nearly linearly converge to S with high probability. Finally, we note that when $S = X^*$ the results of this paper remain valid if (2.4) only holds in a neighborhood of the solution set.[3]

Accuracy and Regularity

We assume that the models are convex and under-approximate f up to quadratic error.

(A3) **(One-sided accuracy)** There exists $\eta > 0$ and an open convex set U containing X and a measurable function $(x, y, z) \to f_x(y, z)$, defined on $U \times U \times$, satisfying

$$\mathsf{E}_z[f_x(x, z)] = f(x) \qquad \forall x \in U,$$

and

$$\mathsf{E}_z[f_x(y, z) - f(y)] \le \frac{\eta}{2}\|y - x\|^2 \qquad \forall x, y \in U.$$

(A4) **(Convex models)** The function $f_x(\cdot, z)$ is convex $\forall x \in U$ and a.e. $z \in$.

Models satisfying (A3) and (A4), and their algorithmic implications, were analyzed in [11, Assumption B]; a closely related family of models was investigated in [3, 4]. Assumptions (A3) and (A4) imply that f is η-weakly convex on U, meaning that the assignment $x \to f(x) + \frac{\eta}{2}\|x\|^2$ is convex [11, Lemma 4.1]. While we assume (A3) throughout the paper, we show in Remark 2 that our results hold under an even weaker assumption. For certain losses, models that satisfy (A3) and (A4) are easy to construct, as the following example shows.

Example 2.1 (Convex Composite Class) Stochastic convex composite losses take the form

$$f(x) = \mathsf{E}_z f(x, z) \quad \text{with} \quad f(x, z) = h(c(x, z), z),$$

where $h(\cdot, z)$ are Lipschitz and convex and the nonlinear maps $c(\cdot, z)$ are C^1-smooth with Lipschitz Jacobian. Such losses appear often in data science and signal processing (see [7, 16, 19] and references therein). For this problem class, natural models include

(subgradient) $f_x(y, z) = f(x, z) + \nabla c(x, z)^* v, y - x$ for any $v \in \partial h(c(x, z), z)$.
(prox-linear) $f_x(y, z) = h(c(x, z) + \nabla c(x, z)(y - x), z)$.
(proximal point) $f_x(y, z) = f(y, z) + \frac{\eta}{2}\|x - y\|^2$,

[3] In particular, in the tube T_2 described in Assumption (A5).

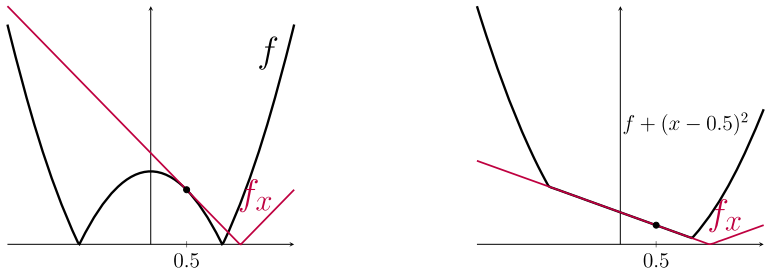

Fig. 1 Illustration of one-sided models: $f(x) = |x^2 - 1|$, $f_{0.5}(y) = |1.25 - y|$ (figure adapted from [11])

where η is large enough to guarantee the proximal model is convex.[4] If a lower bound $\ell(z)$ on $\inf f(\cdot, z)$ is known, one can also choose a clipped model

(clipped) $$\tilde{f}_x(y, z) = \max\{f_x(y, z), \ell(z)\},$$

for any of the models $f_x(\cdot, z)$ above, as was suggested by [3]. Intuitively, models that better approximate $f(x, z)$ are likely to perform better in practice; see [3] for theoretical evidence supporting this claim. Fig. 1 contains a graphical illustration of one-sided models.

Lipschitz continuity

We assume that the models are Lipschitz on a tube surrounding S.

(A5) **(Lipschitz property)** Define the tube

$$T_\gamma := \left\{ x \in X \mid \operatorname{dist}(x, S) \le \frac{\gamma \mu}{\eta} \right\} \qquad \forall \gamma > 0.$$

We assume that there exists a measurable function $L: \mathbb{R}^d \times \to \mathbb{R}_+$ such that

$$\min_{v \in \partial f_x(x,z)} \|v\| \le L(x, z) \tag{2.5}$$

for all $x \in T_2$ and a.e. $z \in$. Moreover, we assume there exists $L > 0$ such that

$$\sup_{x \in T_2} \sqrt{\mathbb{E}_z\, L(x, z)^2} \le L.$$

This Lipschitz property is local and differs from the global assumption of [11, Assumption B4]. The property holds only in T_2 since our algorithms will be initialized in this tube and will never leave it with high probability.[5] The local nature of this property is crucial to signal recovery applications, for example, blind deconvolution and phase retrieval. In these problems, global Lipschitz continuity does not hold; see Sect. 4.

[4] One could choose the product of the Lipschitz constants of h and ∇c.

[5] The set T_2 is a natural region to initialize in since it is provably the widest tube around the solution set containing no extraneous stationary points; see [12, Lemma 2.1] and the discussion following [8, Lemma 3.1].

Remark 1 (Finite-sums) While our results are stated in terms of the streaming model, they are also applicable to the finite-sum setting, where our loss function is in the form

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} f(x, i).$$

Indeed, with $\mathsf{P} = \mathrm{Unif}(\{1, \ldots, m\}) \Rightarrow \mathsf{P}(z_k = i) = \frac{1}{m}$, for any $k \in \mathbb{N}$, $i \in [m]$, we recover

$$\mathsf{E}_z[f(x, z_k)] = \frac{1}{m} \sum_{i=1}^{m} f(x, i) = f(x).$$

2.2 Algorithms and results

Stochastic model-based algorithms (Algorithm 1) iteratively sample and minimize stochastic convex models of the loss function.[6] When equipped with models satisfying (A3), (A4) and a global Lipschitz condition, these algorithms converge to stationary points of (2.1) at a sublinear rate [11, 19]. In this section, we show that such sublinear rates can be improved to local linear (or nearly linear) rates by using the restart strategy outlined in (2.2)–(2.3). We introduce two such strategies that succeed with probability $1 - \delta$, for any $\delta > 0$. The first (Algorithm 2) allows for arbitrary sets S in Assumption (A2), but its sample complexity and initialization region scale poorly in δ. The second (Algorithm 5) assumes $S = X^*$, but has much better dependence on δ.

Algorithm 1: MBA$(y_0, \alpha, K, \texttt{is_conv})$

Input: $y_0 \in \mathbb{R}^d$, $\alpha \geq 0$, iteration count K, flag $\texttt{is\_conv} \in \{\texttt{true}, \texttt{false}\}$

Step $k = 0, \ldots, K$:

$$\left\{ \begin{array}{l} \text{Sample } z_k \sim \mathsf{P} \\ \text{Set } y_{k+1} = \operatorname{argmin}_{y \in X} f_{y_k}(y, z_k) + \frac{1}{2\alpha} \|y - y_k\|^2 \end{array} \right\}.$$

If `is_conv`

  Return $\frac{1}{K+1} \sum_{k=1}^{K+1} y_k$;

Else

  Return y_{K^*}, where $K^* \in \{0, \ldots, K\}$ is selected uniformly at random;

[6] Note that Algorithm 1 can be implemented using $O(d)$ memory by maintaining a running average (in the convex case) or by terminating the iteration after $K^* \sim \mathrm{Unif}\{0, \ldots, K\}$ steps (in the nonconvex case).

Algorithm 2: RMBA$(x_0, \alpha_0, K, T,$ `is_conv`$)$

Input: $x_0 \in \mathbb{R}^d$, $\alpha_0 \geq 0$, counters $K \in \mathbb{N}$, $T \in \mathbb{N}$, flag `is_conv` $\in \{$`true`, `false`$\}$
Step $t = 0, \ldots, T-1$:

$$\alpha_t := 2^{-t}\alpha_0$$
$$x_{t+1} := \text{MBA}(x_t, \alpha_t, K,$$ `is_conv`$).$$

Return x_T

Given assumptions (A1)–(A5), the following theorem shows that the first restart strategy (Algorithm 2—Restarted Model Based Algorithm (RMBA)) converges nearly linearly to S with high probability.

Theorem 2.1 (Informal) *Fix a target accuracy* $\varepsilon > 0$, *failure probability* $\delta \in (0, \frac{1}{4})$, *and a point* $x_0 \in T_{\gamma\sqrt{\delta}}$ *for some* $\gamma \in (0, 1)$. *Then with appropriate parameter settings, the point* $x = \text{RMBA}(x_0, \alpha_0, K, T,$ `false`$)$ *will satisfy* $\text{dist}(x, S) < \varepsilon$ *with probability at least* $1 - 4\delta$. *Moreover, the number of samples* $z_i \sim P$ *generated by the algorithm is at most*

$$O\left(\frac{L}{\delta\mu}\right)^2 \log^3\left(\frac{\gamma\,\mu/\eta}{\varepsilon}\right).$$

Theorem 2.1 has interesting consequences not only for convergence to global minimizers, but also for "active manifold identification." For example, when $S = X^*$, Theorem 2.1 shows that with constant probability, Algorithm 2 converges nearly linearly to the true solution set. When $S = X^*$ and is instead an "active manifold" in the sense of Lewis [38], Algorithm 2 nearly linearly converges to the active manifold. In our numerical evaluation, we illustrate this phenomenon for a sparse logistic regression problem. We empirically observe that the method converges linearly to the support of the solution, even though the overall convergence to the true solution may be sublinear.

In our numerical experiments, we find that Algorithm 2 succeeds for a wide variety of parameter settings, which may not necessarily satisfy the conditions set forth by the theory. Theorem 2.1, on the other hand, only guarantees Algorithm 2 succeeds with high probability when we greatly increase its sample complexity and initialize it close to S. We would like to boost Algorithm 2 into a new algorithm whose sample complexity and initialization requirements scale only polylogarithmicaly in $1/\delta$. As a first attempt, we discuss the following two probabilistic techniques, both of which have limitations:

(Markov) One approach is to call Algorithm 2 multiple times for a moderately small value δ and pick out the "best" iterate from the batch. This approach is flawed since, even in the convex setting, there is no procedure to test which iterate is "best" without increasing sample complexity.
(Ensemble) An alternative approach is based on a well-known resampling trick, which applies when $S = \{\bar{x}\}$ is a singleton set [45, p. 243], [30], [63, Algorithm 1]: Run m trials of Algorithm 2 with any fixed $\delta < 1/4$, and denote the returned

points by $\{x_i\}_{i=1}^m$. Then with high probability, the majority of the points $\{x_i\}_{i=1}^m$ will be close to $\bar{x}$. Finally, to find an estimate near $\bar{x}$, choose any point that has at least $m/2$ other points close to it.

The ensemble technique is promising, but it requires S to be a singleton. This limits its applicability since many low-rank recovery problems (e.g. blind deconvolution, matrix completion, robust PCA) have uncountably many solutions. We overcome this issue by embedding Algorithm 2 and the ensemble method within a proximal-point method. At each stage of this algorithm, we run multiple copies of a stochastic-model based method on a quadratically regularized problem that has a unique solution. Among those copies, we use the ensemble technique to pick out a "successful" iterate. We summarize the resulting nested procedure in Algorithms 3–5: Algorithm 3 (Proximal Model-Based Algorithm—PMBA) is a generic model-based algorithm applied on a quadratically regularized problem; Algorithm 4 (Ensemble PMBA—EPMBA) calls Algorithm 3 as suggested by the ensemble technique; finally, Algorithm 5 (Restarted PMBA (RPMBA)) updates the regularization term, in the style of a proximal point method.

Algorithm 3: PMBA(y_0, ρ, α, K)

Input: $y_0 \in \mathbb{R}^d$, proximal parameter $\rho > \eta$, scalar $\alpha > 0$, and iteration count K
Step $k = 0, \ldots, K$:

$$\left\lfloor \begin{array}{l} \text{Sample } z_k \sim P \\ \text{Set } y_{k+1} = \operatorname{argmin}_{y \in X} \; f_{y_k}(y, z_k) + \frac{1}{2\alpha} \|y - y_k\|^2 + \frac{\rho}{2} \|y - y_0\|^2 \end{array} \right\rfloor,$$

Sample $K^* \in \{0, \ldots, K\}$ uniformly at random.
Return y_{K^*}

Algorithm 4: EPMBA$(y_0, \rho, \alpha, K, m, \quad)$

Input: $y_0 \in \mathbb{R}^d$, proximal parameter $\rho > \eta$, scalar $\sigma > 0$, iteration count K, trial count m, relative error tolerance
Step $j = 1, \ldots, m$:
Set $y_j = \text{PMBA}(y_0, \rho, \alpha, K)$.
Step $j = 1, \ldots, m$:
if $|B_2(y_j) \cap \{y_i\}_{i=1}^m| > \frac{m}{2}$
Return y_j

We will establish the following guarantee. In the theorem, we assume $S = X^*$.

Theorem 2.2 (Informal) *Fix a target accuracy $\varepsilon > 0$, failure probability $\delta \in (0, 1)$, and a point $x_0 \in T_\gamma$ for some $\gamma \in (0, \frac{1}{4})$. Then with appropriate parameter settings, the point $x = \text{RPMBA}(x_0, \rho_0, \alpha_0, K, \varepsilon_0, M, T)$ will satisfy $\operatorname{dist}(x, X^*) < \varepsilon$ with probability at least $1 - \delta$. Moreover the total number of samples $z_i \sim P$ generated by the algorithm is at most*

Algorithm 5: RPMBA$(x_0, \rho_0, \alpha_0, K, \ _0, M, T)$

Input: $x_0 \in \mathbb{R}^d$, proximal parameter $\rho_0 > \eta$, initial accuracy $\ _0 > 0$, stepsize $\alpha_0 > 0$, counts $K, M, T \in \mathbb{N}$

Step $t = 0, \ldots, T-1$:

$$\begin{aligned} \rho_t &:= 2^t \rho_0 \\ \ _t &:= 2^{-t} \ _0 \\ \alpha_t &:= 2^{-t} \alpha_0 \\ x_{t+1} &:= \mathrm{EPMBA}(x_t, \rho_t, \alpha_t, K, M, \ _t). \end{aligned}$$

Return x_{T+1}

$$O\left(\left(\frac{L}{\mu}\right)^2 \log\left(\frac{\gamma\mu/\eta}{\varepsilon}\right) \cdot \log\left(\frac{\log(\frac{\gamma\mu/\eta}{\varepsilon})}{\delta}\right)\right).$$

Theorem 2.2 resolves the initialization and sample complexity issues of Theorem 2.1. Incidentally, its claimed sample complexity also depends more favorably on ε and on the problem parameters μ and η. Theorem 2.2 is new in the weakly convex setting and also improves on prior work by Xu et al. [66] for convex problems. There, the results require stochastic subgradients to be almost surely bounded, hence, sub-Gaussian. In contrast, Theorem 2.2 guarantees that $\mathrm{dist}(x, X^*) < \varepsilon$ with high-probability assuming only the local second moment bound (A5).

3 Proofs of main results

In this section, we establish high-probability nearly linear convergence guarantees for Algorithm 2 and linear convergence guarantees for Algorithm 5 —the main contributions of this work. Throughout this section, we assume that Assumptions (A1)–(A5) hold.

3.1 Warm-up: convex setting

We begin with a short proof of nearly linear convergence for Algorithm 2 in the convex setting. We use this simplified case to explain the general proof strategy and point out the difficulty of extending the argument to the weakly convex setting. Since we restrict ourselves to the convex setting, throughout this section (Sect. 3.1) we suppose:

- Assumption (A3) holds with $\eta = 0$ and Assumption (A2) holds with $S = X^*$.
- The models $f_x(\cdot, z)$ are $L(z)$-Lipschitz on $\mathbb{R}^d$ for all x, where $L : \to \mathbb{R}$ is a measurable function satisfying $\mathbb{E}_z \overline{L(z)^2} \leq L$.

In particular, the tube T_2 is the entire space $T_2 = \mathbb{R}^d$, which alleviates the main difficulty of the weakly convex setting. The proof of convergence relies on the following known sublinear convergence guarantee for Algorithm 1.

Theorem 3.1 ([11, Theorem 4.1]). *Fix an initial point* $y_0 \in \mathbb{R}^d$ *and let* $\alpha = \frac{\sqrt{C}}{K+1}$ *for some* $C > 0$. *Then for any index* $K \in \mathbb{N}$, *the point* $y = \mathrm{MBA}(y_0, \alpha, K, \mathtt{true})$ *satisfies*

$$\mathbb{E}\left[f(y) - \min_X f\right] \leq \frac{\frac{1}{2}\mathrm{dist}^2(y_0, X^*) + C^2 L^2}{C\sqrt{K+1}}. \tag{3.1}$$

The proof of nearly linear convergence now follows by iteratively applying Theorem 3.1 with a carefully chosen parameter $C > 0$. The key idea of the proof is much the same as in the deterministic setting [44, 47, 58]. The proof proceeds by induction on the outer iteration counter t. At the start of each inner iteration, we choose C to minimize the ratio in Equation (3.1), taking into account an inductive estimate on the initial square error $\mathrm{dist}^2(y_0, X^*)$. We then run the inner loop until the estimate decreases by a fixed fraction. This strategy differs from deterministic setting in only one way: since the output of the inner loop is random, we extract a bound on the initial distance using Markov's inequality.

Theorem 3.2 (Nearly linear convergence under convexity). *Fix an initial point* $x_0 \in \mathbb{R}^d$, *real* $\varepsilon > 0, \delta \in (0, 1)$, *and an upper bound* $R_0 \geq \mathrm{dist}(x_0, X^*)$. *Define parameters*

$$T := \log_2 \frac{R_0}{\varepsilon}, \qquad K := 8 \cdot T^2 \cdot \left(\frac{L}{\delta\mu}\right)^2, \qquad \alpha_0 = \sqrt{\frac{R_0^2}{2L^2(K+1)}}.$$

Then with probability at least $1 - \delta$, *the point* $x_T = \mathrm{RMBA}(x_0, \alpha_0, K, T, \mathtt{true})$ *satisfies* $\mathrm{dist}(x_T, X^*) \leq \varepsilon$. *Moreover, the total number of samples* $z_k \sim P$ *generated by the algorithm is bounded by*

$$TK \leq 8\left(\frac{L}{\delta\mu}\right)^2 \log_2^3 \frac{R_0}{\varepsilon}.$$

Proof In what follows, set $C_t = \frac{R_0}{L \cdot 2^{t+\frac{1}{2}}}$ and note the equality $\alpha_t = \frac{\sqrt{C_t}}{K+1}$ for every index t in Algorithm 2. Let E_t denote the event that $\mathrm{dist}(x_t, S) \leq 2^{-t} \cdot R_0$. We wish to show the inequality

$$\mathbb{P}(E_{t+1}) \geq \mathbb{P}(E_t) - \frac{2^{3/2} L}{\mu\sqrt{K+1}}, \tag{3.2}$$

for all $t \in \{0, \cdots, T\}$. To that end, observe

$$\mathbb{P}(E_{t+1}) \geq \mathbb{P}(E_{t+1} \mid E_t)\mathbb{P}(E_t). \tag{3.3}$$

To lower bound the right-hand-side, observe by Markov's inequality the estimate

$$\mathbb{P}(E_{t+1}^c \mid E_t) \leq \frac{\mathbb{E}\left[\mathrm{dist}(x_{t+1}, X^*) \mid E_t\right]}{2^{-(t+1)} R_0} = \frac{\mathbb{E}\left[\mathrm{dist}(x_{t+1}, X^*) 1_{E_t}\right]}{2^{-(t+1)} R_0 \mathbb{P}(E_t)}. \tag{3.4}$$

Combining assumption (A2) and Theorem 3.1, we deduce

$$\mathbb{E}\ \mathrm{dist}(x_{t+1}, X^*)1_{E_t} \le \mathbb{E}\ \mu^{-1}(f(x_{t+1}) - \min_X f)1_{E_t} \le \frac{\frac{1}{2}\mathbb{E}\ \mathrm{dist}^2(x_t, X^*)1_{E_t} + C_t^2 L^2}{\mu C_t \sqrt{K+1}} \le \frac{2^{-(1+2t)} R_0^2 + C_t^2 L^2}{\mu C_t \sqrt{K+1}} = \frac{2^{\frac{1}{2}-t} R_0 L}{\mu \sqrt{K+1}}. \tag{3.5}$$

Therefore, combining (3.3), (3.4), and (3.5) we arrive at the claimed estimate

$$\mathbb{P}(E_{t+1}) \ge \left(1 - \frac{1}{\mathbb{P}(E_t)} \cdot \frac{2^{3/2} L}{\mu \sqrt{K+1}}\right) \mathbb{P}(E_t) \ge \mathbb{P}(E_t) - \frac{2^{3/2} L}{\mu \sqrt{K+1}}.$$

Iterating (3.2) and using the definition of T and K, we conclude that with probability

$$\mathbb{P}(E_T) \ge 1 - \frac{2^{3/2} T L}{\mu \sqrt{K+1}} \ge 1 - \delta,$$

the estimate

$$\mathrm{dist}(x_T, X^*) \le 2^{-T} R_0 \le \varepsilon,$$

holds as claimed. This completes the proof.

3.2 Weakly convex setting

We now present the convergence guarantees for Algorithm 2 in the weakly convex setting under Assumptions (A1)–(A5). The proof of nearly linear convergence proceeds by inductively applying the following Lemma, which is similar to Lemma 3.1. Compared to the convex setting, the weakly convex setting presents a new challenge: the region of nearly linear convergence, T_γ, is local. The iterates of Algorithm 2 must therefore be shown to never leave T_γ with high probability. We show this through a stopping time argument in the proof of the following Lemma (see Sect. A.1).

Lemma 3.3 *Fix real numbers $\delta \in (0, 1)$, $\gamma \in (0, 2)$, $K \in \mathbb{N}$, and $\alpha > 0$. Let y_0 be a random vector and let B denote the event $\{y_0 \in T_\gamma \sqrt{\delta}\}$. Define*

$$y_{K^*} = \mathrm{MBA}(y_0, \alpha, K, \mathtt{false}).$$

Then for any $\varepsilon > 0$, the estimate $\mathrm{dist}(y_{K^*}, S) \le \varepsilon$ holds with probability at least

$$\mathbb{P}(B) - \delta - \left(\frac{\eta}{\gamma\mu}\right)^2 K L^2 \alpha^2 - \frac{1}{\varepsilon} \cdot \frac{\delta \left(\frac{\gamma\mu}{\eta}\right)^2 + (K+1) L^2 \alpha^2}{(2-\gamma)\mu(K+1)\alpha}.$$

The proof of nearly linear convergence of Algorithm 2 in the weakly convex setting now follows by inductively applying Lemma 3.3.

Theorem 3.4 (Nearly linear convergence under weak convexity) *Fix real numbers* $\varepsilon > 0$, $\delta_2 \in (0, 1)$, $\gamma \in (0, 2)$. *Let* R_0 *denote the initial distance estimate satisfying* $\operatorname{dist}(x_0, S) \le R_0 \le \frac{\gamma\mu}{\eta}$. *Furthermore, define algorithm parameters*

$$T := \log_2 \frac{R_0}{\varepsilon}, \quad K := \frac{16}{(2-\gamma)^2} \cdot T^2 \cdot \left(\frac{L}{\delta_2\mu}\right)^2, \quad \alpha_0 = \sqrt{\frac{R_0^2}{L^2(K+1)}}.$$

Then with probability at least $1 - (8/3)R_0^2 \left(\frac{\eta}{\gamma\mu}\right)^2 - \delta_2$, *the point* $x_T = \mathrm{RMBA}(x_0, \alpha_0, K, T, \mathtt{false})$ *satisfies* $\operatorname{dist}(x_T, S) \le \varepsilon$. *Moreover, the total number of samples* $z_k \sim P$ *generated by the algorithm is bounded by*

$$TK \le \frac{16}{(2-\gamma)^2} \left(\frac{L}{\delta_2\mu}\right)^2 \log_2 \left(\frac{R_0}{\varepsilon}\right)^3.$$

Proof For all t, let E_t be the event $\{\operatorname{dist}(x_t, S) \le 2^{-t} \cdot R_0\}$. In addition, define $\delta_1 := R_0^2 \left(\frac{\eta}{\gamma\mu}\right)^2$. We claim that the inequality

$$P(E_{t+1}) \ge P(E_t) - 2\delta_1 2^{-2t} - \frac{4L}{(2-\gamma)\mu\sqrt{K+1}},$$

holds for all $t \in \{0, 1 \dots, T\}$. To see this, apply Lemma 3.3 with $y_0 = x_t$, $\varepsilon = 2^{-(t+1)}R_0$, $\delta := \delta_1 2^{-2t}$, $\alpha = 2^{-t}\alpha_0$, thereby yielding

$$\begin{aligned} P(E_{t+1}) &\ge P(E_t) - \delta_1 2^{-2t} - \left(\frac{\eta}{\gamma\mu}\right)^2 KL^2\alpha_t^2 - \frac{1}{\varepsilon} \cdot \frac{\delta\left(\frac{\gamma\mu}{\eta}\right)^2 + (K+1)L^2\alpha_t^2}{(2-\gamma)\mu(K+1)\alpha_t} \\ &\ge P(E_t) - 2\delta_1 2^{-2t} - \frac{4L}{(2-\gamma)\mu\sqrt{K+1}}, \end{aligned}$$

where the last inequality follows from the definitions of α_0 and R_0. Iterating the inequality, we conclude

$$P(E_T) \ge 1 - 2\delta_1 \sum_{i=0}^{T-1} 2^{-2i} - \frac{4LT}{(2-\gamma)\mu\sqrt{K+1}} \ge 1 - (8/3)\delta_1 - \delta_2. \qquad (3.6)$$

This completes the proof.

Observe that the probability of success $1 - (8/3)R_0^2 \left(\frac{\eta}{\gamma\mu}\right)^2 - \delta_2$ in Theorem 3.4 depends both on the initialization quality R_0 and on δ_2. Moreover, δ_2 also appears inversely in the sample complexity $O\left(\frac{L^2}{\delta_2^2\mu^2}\right)$. In the next section, we introduce an algorithm with probability of success independent of R_0 and with sample complexity that depends only logarithmically on its success probability.

We close this section with the following remark, which shows that the results of Theorem 3.4 extend beyond the weakly convex setting.

Remark 2 (Beyond weakly convex problems). Assumptions (A3) and (A4) imply that f is η-weakly convex on U, meaning that the assignment $x \mapsto f(x) + \frac{\eta}{2}\|x\|^2$ is convex [11, Lemma 4.1]. Revisiting the proof of Lemma 3.3, however, we see that (A3) may be replaced by the following weaker assumption:

(A3) **(Two-point accuracy)** There exists $\eta > 0$ and an open convex set U containing X and a measurable function $(x, y, z) \to f_x(y, z)$, defined on $U \times U \times$, satisfying

$$\mathsf{E}_z[f_x(x, z)] = f(x) \qquad \forall x \in U,$$

and

$$\mathsf{E}_z[f_x(y, z) - f(y)] \le \frac{\eta}{2}\|y - x\|^2 \qquad \forall x \in U, y \in \operatorname{argmin}_{z \in \operatorname{proj}_S(x)} f(z).$$

In the case $S = X^*$, this assumption requires the model to touch the function at x and to lower bound it, up to quadratic error, at its nearest minimizer. This condition does not imply that f is weakly convex.

3.3 Convergence with high probability

In this section, we show that Algorithm 5 succeeds with high probability. Throughout this section (Sect. 3.3), we impose Assumptions (A1)–(A5) with $S = X^*$.

The following lemma guarantees that with appropriate step size, the proximal point of the problem (2.1) at $y \in T_\gamma$ lies in $\operatorname{proj}_{X^*}(y)$. We present the proof in Sect. B.2.

Lemma 3.5 *Fix $\gamma \in (0, 2)$, $\rho > \eta$, and a point $y \in T_\gamma$. Then the proximal subproblem*

$$\min_{x \in X} \; f(x) + \frac{\rho}{2}\|x - y\|^2 \tag{3.7}$$

is strongly convex and therefore has a unique minimizer $\bar{y}$. Moreover, if $\rho < \frac{2-\gamma}{2\gamma}\eta$, then the inclusion $\bar{y} \in \operatorname{proj}_{X^*}(y)$ holds.

Lemma 3.5 shows that, unlike Algorithm 1, we can expect the output of Algorithm 3 to be near the minimizer $\bar{y}_0 \in X^*$ of the proximal subproblem $f(y) + \frac{\rho}{2}\|y - y_0\|^2$, at least with constant probability. This lemma underlies the validity of Lemma 3.6. We present its proof in Sect. B.3.

Lemma 3.6 *Fix real numbers $\delta \in (0, 1)$, $\gamma \in (0, 2)$, $\alpha > 0$ $K \in \mathbb{N}$, and ρ satisfying $\eta < \rho < \frac{2-\gamma\sqrt{\delta}}{2\gamma\sqrt{\delta}}\eta$. Choose any point $y_0 \in T_{\gamma\sqrt{\delta}}$ and set $\bar{y}_0 := \operatorname{argmin}_{x \in X} f(x) + \frac{\rho}{2}\|x - y_0\|^2$. Define*

$$y_{K^*} = \mathrm{PMBA}(y_0, \rho, \alpha, K).$$

Then for all $\varepsilon > 0$, *with probability at least*

$$P(\varepsilon) := 1 - \delta - \left(\frac{\eta}{\gamma\mu}\right)^2 K\mathrm{L}^2\alpha^2 - \frac{1}{\varepsilon^2}\cdot\frac{(K+1)\mathrm{L}^2\alpha + (\alpha^{-1}+\eta)\cdot\delta\left(\frac{\gamma\mu}{\eta}\right)^2}{(\rho-\eta)(K+1)}$$

we have $\|y_{K^*} - \bar{y}_0\| \leq \varepsilon$.

The next lemma shows that we can boost the success probability of the inner loop arbitrarily high at only a logarithmic cost. This is an immediate application of Lemma 3.6 and the ensemble technique described in Sect. 2.2, which is formally stated in Lemma B.1.

Corollary 3.7 *Assume the setting of Lemma 3.6 and suppose* $P(\varepsilon/3) \geq 2/3$. *Then for any* $\delta > 0$, *in the regime* $M > 48\log(1/\delta)$, *the point*

$$\hat{y} = \mathrm{EPMBA}(y_0, \rho, \alpha, K, M, \varepsilon/3)$$

satisfies

$$\mathrm{P}(\|\hat{y} - \bar{y}_0\| \leq \varepsilon) \geq 1-\delta.$$

Finally, we are ready to establish high probability convergence guarantees of Algorithm 5.

Theorem 3.8 (Linear convergence with high probability) *Fix constants* $\gamma \in (0, 2)$, $\varepsilon > 0$, *and* $\delta \in (0, 1)$. *Let* R_0 *denote the initial distance estimate satisfying* $\mathrm{dist}(x_0, X^*) \leq R_0 \leq \frac{\gamma\mu}{4\eta}$. *Furthermore, define algorithm parameters*

$$\rho_0 = \frac{\mu}{2R_0}, \qquad \varepsilon_0 = \frac{R_0}{3}, \qquad \alpha_0 = \sqrt{\frac{R_0^2}{\mathrm{L}^2(K+1)}},$$

and

$$M = \left\lceil 48\log(T/\delta)\right\rceil, \qquad T = \log_2\left(\frac{R_0}{\varepsilon}\right), \qquad K = \left\lceil\frac{864\mathrm{L}}{\mu}\right\rceil.$$

Then with probability at least $1-\delta$, *the point*

$$x_T = \mathrm{RPMBA}(x_0, \rho_0, \alpha_0, K, \varepsilon_0, M, T)$$

satisfies $\mathrm{dist}(x_T, X^*) \leq \varepsilon$. *Moreover, the total number of samples* $z_k \sim \mathrm{P}$ *generated by the algorithm is bounded by*

$$KTM \leq \left\lceil\frac{864\mathrm{L}}{\mu}\right\rceil \cdot \left\lceil\log_2\left(\frac{R_0}{\varepsilon}\right)\right\rceil \cdot \left\lceil 48\log\left(\frac{\log_2\frac{R_0}{\varepsilon}}{\delta}\right)\right\rceil.$$

Proof For all t, let E_t be the event $\{\mathrm{dist}(x_t, X^*) \le 2^{-t} \cdot R_0\}$. Our goal is to show for all $t \in \{0, \dots, T\}$ the estimate $\mathbb{P}(E_t) \ge 1 - t\delta/T$ holds.

We proceed by induction. The base case follows since $\mathbb{P}(E_0) = 1$ by definition of R_0. Now suppose that the claimed estimate $\mathbb{P}(E_t) \ge 1 - t\delta/T$ is true for index t. We will show it remains true with t replaced by $t+1$. We will apply Lemma 3.6 conditionally with $y_0 = x_t$ and error tolerance ε_t in the event E_t. To this end, we define $\delta_1 := R_0^2 \left(\frac{\eta}{\gamma\mu}\right)^2$ and set $\rho = \rho_t, \delta := \delta_1 2^{-2t}, \alpha = \alpha_t$. Before we apply the Lemma, we verify that ρ meets the conditions of Lemma 3.6, namely that $\eta < \rho < \frac{2-\gamma\sqrt{\delta_1 2^{-t}}}{2\gamma\sqrt{\delta_1 2^{-t}}}\eta$. Indeed, given that $\rho = \frac{2^t}{2\gamma\sqrt{\delta_1}} \cdot \eta$ (by definition of δ_1), the bounds follow immediately from the restrictions $\delta_1 < 1/16$ and $\gamma \le 2$. In particular, it is straightforward to verify the bound

$$\rho - \eta \ge \frac{\eta 2^{t-2}}{\gamma\sqrt{\delta_1}} = \frac{2^t\mu}{4R_0}. \tag{3.8}$$

Now Lemma 3.6 yields that the random vector $y_{K^*} = \mathrm{PMBA}(x_t, \rho_t, \alpha_t, K_t)$ satisfies

$$\begin{aligned}
&\mathbb{P}\left(\|y_{K^*} - \bar{y}_0\| \le \varepsilon_t \mid E_t, y_0 = x_t\right) \\
&\ge 1 - \delta_1 2^{-2t} - \left(\frac{\eta}{\gamma\mu}\right)^2 K\mathsf{L}^2\alpha^2 - \frac{1}{t^2} \cdot \frac{(K+1)\mathsf{L}^2\alpha + (\alpha^{-1} + \eta) \cdot \delta \left(\frac{\gamma\mu}{\eta}\right)^2}{(\rho - \eta)(K+1)} \\
&\ge 1 - 2\delta_1 2^{-2t} - \frac{9}{2^{-2(t+1)}R_0^2} \cdot \frac{\mathsf{L}2^{-t}R_0}{(\rho-\eta)\sqrt{K+1}} - \frac{36\eta}{(\rho-\eta)(K+1)} \\
&\ge 1 - 2\delta_1 2^{-2t} - \frac{144(\mathsf{L}/\mu)}{\sqrt{K+1}} - \frac{144\gamma\sqrt{\delta_1 2^{-2t}}}{K+1} \ge 2/3,
\end{aligned}$$

where the second inequality follows from (3.8), while the third inequality uses the definition of K and the bound $\delta_1 < 1/16$ and $\mathsf{L} \ge \mu$. Therefore, since $M \ge 48\log(T/\delta)$, we may apply Corollary 3.7 (conditionally) to deduce

$$\mathbb{P}\left(\|x_{t+1} - \bar{y}_0\| \le 3\varepsilon_t \mid E_t, y_0 = x_t\right) \ge 1 - \delta/T.$$

Consequently,

$$\begin{aligned}
\mathbb{P}\left\{\mathrm{dist}(x_{t+1}, X^*) \le 2^{-(t+1)}R_0\right\} &\ge \mathbb{P}\left\{\mathrm{dist}(x_{t+1}, X^*) \le 2^{-(t+1)}R_0 \mid E_t\right\}\mathbb{P}(E_t) \\
&\ge \mathbb{E}_{y_0}\left[\mathbb{P}\left\{\|x_{t+1} - \bar{y}_0\| \le 2^{-(t+1)}R_0 \mid E_t, y_0 = x_t\right\}\right]\mathbb{P}(E_t) \\
&\ge (1 - \delta/T)(1 - t\delta/T) \\
&\ge 1 - (t+1)\delta/T,
\end{aligned}$$

as desired. This completes the proof.

4 Consequences for statistical recovery problems

Recent work has shown that a variety of statistical recovery problems are both sharp and weakly convex. Prominent examples include robust matrix sensing [40], phase retrieval [18], blind deconvolution [8], quadratic and bilinear sensing and matrix completion [7]. In this section, we briefly comment on how our current work leads to linearly convergent streaming algorithms for robust phase retrieval and blind deconvolution problems.

4.1 Robust Phase retrieval

Phase retrieval is a common task in computational science, with numerous applications including imaging, X-ray crystallography, and speech processing. In this section, we consider the real counterpart of this problem. For details and a historical account of the phase retrieval problem, see for example [6, 18, 28, 61]. Throughout this section, we fix a signal $\bar{x} \in \mathbb{R}^d$ and consider the following measurement model.

Assumption A (Robust Phase Retrieval) Consider random $a \in \mathbb{R}^d$, $\xi \in \mathbb{R}$, and $u \in \{0, 1\}$ and the measurement model

$$b = \langle a, \bar{x} \rangle^2 + u \cdot \xi .$$

We make the following assumptions on the random data.

1. The variable u is independent of ξ and a. The failure probability p_{fail} satisfies
$$p_{\text{fail}} := \mathbb{P}(u = 0) < 1/2.$$
2. The first absolute moment of ξ is finite, $\mathbb{E}[|\xi|] < \infty$.
3. There exist constants $\tilde{\eta}, \tilde{\mu}, \tilde{L} > 0$ such that for all $v, w \in \mathbb{S}^{d-1}$, we have
$$\tilde{\mu} \le \mathbb{E}[|\langle a, v\rangle \langle a, w\rangle|], \qquad \sqrt{\mathbb{E}[\langle a, v\rangle^2 \|a\|^2]} \le \tilde{L}, \qquad \mathbb{E}[\langle a, v\rangle^2]^2 \le \tilde{\eta}.$$

Based on the above assumptions, the following theorem develops three models for the robust phase retrieval problem. We defer the proof to Sect. C.1.

Theorem 4.1 (Phase retrieval parameters). *Consider the population data $\mathcal{F} := (a, u, \xi)$ and form the optimization problem*

$$\min_x \; f(x) = \mathbb{E}_z[f(x, z)] \quad \text{where} \quad f(x, z) := |\langle a, x\rangle^2 - b|.$$

Then the sharpness property (A2) holds with $S = X^* = \{\pm \bar{x}\}$ *and* $\mu = (1 - 2p_{\text{fail}})\tilde{\mu}\|\bar{x}\|$. *Moreover, given a measurable selection* $G(x, z) \in \partial f(x, z)$, *the models*

(subgradient) $f_x^s(y, z) = f(x, z) + \langle G(x, z), y - x\rangle$,
(clipped subgradient) $f_x^{cl}(y, z) = \max\{f(x, z) + \langle G(x, z), y - x\rangle, 0\}$,

(prox-linear) $f_x^{pl}(y, z) = |\langle a, x\rangle^2 - b + 2\langle a, x\rangle\langle a, y - x\rangle|$

satisfy Assumptions (A1)–(A5) with

$$\eta = 2\eta, \quad L(x, z) = 2|\langle a, x\rangle| \quad a \quad \text{and } \tilde{L} \leq 2\tilde{L}\|x\| + \frac{(1 - 2p_{\text{fail}})\mu}{\eta}.$$

With this theorem in hand, we deduce that on the phase retrieval problem, Algorithm 5 with subgradient, clipped subgradient, and prox-linear models converges linearly to X^* with high probability, whenever the method is initialized within constant relative error of the optimal solution.

Theorem 4.2 *Fix constants $\gamma \in (0, 2)$, $\varepsilon > 0$, and $\delta \in (0, 1)$. Consider the subgradient, clipped subgradient, and prox-linear oracles developed in Theorem 4.1. Suppose we are given a point x_0 satisfying*

$$\text{dist}(x_0, \{\pm x\}) \leq \gamma \frac{(1 - 2p_{\text{fail}})\mu}{8\eta} \cdot \|x\|.$$

Set parameters $\rho_0, \alpha_0, K, \ldots_0, M, T$ as in Theorem 3.8. In addition, define the iterate $x_T = \text{RPMBA}(x_0, \rho_0, \alpha_0, K, \ldots_0, M, T)$. Then with probability $1 - \delta$, we have $\text{dist}(x_T, \{\pm\bar{x}\}) \leq \varepsilon\|x\|$ after

$$O\left(\left(\frac{\tilde{L} + \frac{(1-2p_{\text{fail}})\mu}{2\eta}}{\mu(1 - 2p_{\text{fail}})}\right)^2 \log\left(\frac{(1 - 2p_{\text{fail}})\tilde{\mu}/\eta}{\varepsilon}\right) \log\left(\frac{\log\frac{(1-2p_{\text{fail}})\tilde{\mu}/\eta}{\varepsilon}}{\delta}\right)\right)$$

stochastic subgradient, stochastic clipped subgradient, or stochastic prox-linear iterations.

We now examine Theorem 4.2 in the setting where the measurement vectors a follow a Gaussian distribution. We note, however, that the results of this section extend far beyond the Gaussian setting to heavy tailed distributions.

Example 4.1 (Gaussian setting) Let us analyze the population setting where $a \sim N(0, I_{d\times d})$. In this case, it is straightforward to show by direct computation that

$$\tilde{\mu} \geq 1; \qquad \eta = 1; \qquad \tilde{L} \leq \sqrt{d}.$$

Consequently, if $x_0 \in \mathbb{R}^d$ has error $\text{dist}(x_0, \{\pm\bar{x}\}) \leq c(1 - 2p_{\text{fail}})\cdot\|x\|$ for some numerical constant c, then with probability $1 - \delta$, Algorithm 5 will produce a point x_T satisfying $\text{dist}(x_T, \{\pm x\}) \leq \varepsilon\|x\|$ using only

$$O\left(\frac{d}{(1 - 2p_{\text{fail}})^2} \log\frac{1}{\varepsilon} \log\frac{\log\frac{1}{\varepsilon}}{\delta}\right)$$

samples. We note that the spectral initialization of Duchi and Ruan [18, Proposition 3] produces such a point x_0 with sample complexity $O(d(1 - 2p_{\text{fail}})^{-2})$ with

high probability. Therefore, when taken together, combining this spectral initialization with Algorithm 5 produces a point x_T satisfying $\operatorname{dist}(x_T, \{\pm \bar{x}\}) \leq \varepsilon \|\bar{x}\|$ with $O\left(\frac{d}{(1-2p_{\text{fail}})^2} \log \frac{1}{\varepsilon} \log^5 d \log^5 \log \frac{1}{\varepsilon} / \delta^6\right)$ samples, which is the best known sample complexity for Gaussian robust phase retrieval, up to logarithmic factors. We note that by leveraging standard concentration results, it is possible to prove similar results for empirical average minimization $\min_x \frac{1}{m} \sum_{i=1}^m f(x, z_i)$, provided z_i are i.i.d samples of z and the number of samples satisfies $m \gtrsim d(1-2p_{\text{fail}})^{-2}$.

4.2 Robust blind deconvolution

We next apply the proposed algorithms to the blind deconvolution problem. For a detailed discussion of the the problem, see for example the papers [2, 39]. Henceforth, fix integers $d_1, d_2 \in \mathbb{N}$ and an underlying signal $(\bar{x}, \bar{y}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. Define the quantity

$$D := \|\bar{x}\| \|\bar{y}\|.$$

Without loss of generality, we will assume $\|\bar{x}\| = \|\bar{y}\|$. We consider the following measurement model:

Assumption B (Robust Blind Deconvolution). Consider random $a \in \mathbb{R}^{d_1}$, $r \in \mathbb{R}^{d_2}$, $\xi \in \mathbb{R}$, and $u \in \{0, 1\}$ and the measurement model

$$b = |\langle a, \bar{x} \rangle \langle r, \bar{y} \rangle + u \cdot \xi|.$$

We make the following assumptions on the random data.

1. The variable u is independent of ξ, a, and r. The failure probability p_{fail} satisfies
$$p_{\text{fail}} := \mathbb{P}(u = 0) < 1/2.$$
2. We have $\mathbb{E}[|\xi|] < \infty$.
3. There exists constants $\tilde{\eta}, \tilde{\mu}, \tilde{L} > 0$ such that for all $M \in \mathbb{R}^{d_1 \times d_2}$ with $\|M\|_F = 1$ and $\operatorname{rank}(M) \leq 2$, we have
$$\tilde{\mu} \leq \mathbb{E}[|\langle a, Mr \rangle|] \leq \tilde{\eta}$$
4. There exists constants $\tilde{L} > 0$ such that for all $v \in S^{d_1-1}$, $w \in S^{d_2-1}$, we have
$$\sqrt{\mathbb{E}\left(|\langle a, v \rangle \langle r, w \rangle|\right)^2} \leq \tilde{L}.$$

Based on the above assumptions, the following theorem develops three models for the robust blind deconvolution problem. We defer the proof to Sect. C.2.

Theorem 4.3 (Blind deconvolution parameters). *Fix a real $\nu > 1$ and define the set:*

$$X = \{(x, y) \in \mathbb{R}^{d_1+d_2} : \|x\| \|y\| \leq \nu D, \|x\| \leq \nu D\}.$$

Consider the population data $z = (a, u, \xi)$ *and the form the optimization problem*

$$\min_{x,y} f(x, y) := \mathbb{E}\,[f((x, y), z)] \quad \text{where} \quad f((x, y), z) := |\,, \quad x \quad r, y - \quad b|.$$

Then the optimal solution set is $X^* = \{(\alpha \bar{x}, (1/\alpha)\,\bar{y}) \mid (1/\nu) \le |\alpha| \le \nu\}$ *and* f *satisfies the sharpness assumption (A2) with* $S = X^*$ *and*

$$\mu = \frac{\mu(1 - 2\,p_{\text{fail}})\sqrt{D}}{2\sqrt{2}(\nu + 1)}.$$

Moreover, given a measurable selection $G((x, y), z) \in \partial f((x, y), z)$*, the models*

(subgradient) $f^{s}_{(x,y)}((\hat{x}, \hat{y}), z) = f((x, y), z) + G((x, y), z), (\hat{x}, \hat{y}) - (x, y)$,

(clipped subgradient)

$$f^{cl}_{(x,y)}((\hat{x}, \hat{y}), z) = \max\{f((x, y), z) + G((x, y), z), (\hat{x}, \hat{y}) - (x, y), \ 0\},$$

(prox-linear)

$$f^{pl}_{(x,y)}((\hat{x}, \hat{y}), z) = |\,, \quad x \quad r, y - (\,, \bar{} \quad x \quad r, \bar{y} + \quad u \cdot \xi) + \,, \quad x \quad r, \hat{y} - y + \quad r, y\,, \hat{} \quad x - x|$$

satisfy Assumptions (A1)–(A5) with

$$\eta = \eta, \quad L((x, y), z) = |\,, \quad x| \quad r \qquad r, y| + | \qquad \text{and} \quad \mathsf{L} = \nu\tilde{\mathsf{L}}\sqrt{D}.$$

With this theorem in hand, we deduce that on the blind deconvolution problem, Algorithm 5 with subgradient, clipped subgradient, and prox-linear models converges linearly to X^* with high probability, whenever the method is initialized within constant relative error of the solution set.

Theorem 4.4 *Fix constants* $\gamma \in (0, 2), \varepsilon > 0$, *and* $\delta \in (0, 1)$*. Consider the subgradient, clipped subgradient, and prox-linear oracles developed in Theorem 4.1. Suppose we are given a pair* $(x_0, y_0) \in \mathbb{R}^{d_1 + d_2}$ *satisfying*

$$\text{dist}((x_0, y_0), X^*) \le \gamma\,\frac{\mu(1 - 2\,p_{\text{fail}})\sqrt{D}}{8\sqrt{2}(\nu + 1)\eta}.$$

Set parameters $\rho_0, \alpha_0, K, {}_0, M, T$ *as in Theorem 3.8. In addition, we define the iterate* $(x_T, y_T) = \text{RPMBA}((x_0, y_0), \rho_0, \alpha_0, K, {}_0, M, T)$*. Then with probability* $1 - \delta$ *, we have* $\text{dist}((x_T, y_T), X) \le \varepsilon\sqrt{D}$ *after*

$$O\left(\left(\frac{\nu^2\tilde{\mathsf{L}}}{\mu(1 - 2\,p_{\text{fail}})}\right)^2 \log\frac{(1 - 2\,p_{\text{fail}})\tilde{\mu}/\eta}{\varepsilon} \log\left(\frac{\log\frac{(1-2\,p_{\text{fail}})\tilde{\mu}/\eta}{\varepsilon}}{\delta}\right)\right)$$

stochastic subgradient, stochastic clipped subgradient, or stochastic prox-linear iterations.

We now examine Theorem 4.4 in the setting where the measurement vectors a_i, r_i follow a Gaussian distribution. We note, however, that the results of this section extend far beyond the Gaussian setting to heavy tailed distributions.

Example 4.2 (Gaussian setting). Let us analyze the population setting where $(a_i, r_i) \sim N(0, I_{(d_1+d_2)\times(d_1+d_2)})$. In this case, one can show by direct computation that

$$\tilde{\mu} = 1; \quad \eta = 1; \quad L = \sqrt{d_1 + d_2}.$$

Consequently, if $(x_0, y_0) \in \mathbb{R}^{d_1+d_2}$ has error $\operatorname{dist}((x_0, y_0), X^*) \le c(1-2p_{\text{fail}})\cdot\sqrt{D}/\nu$, for some numerical constant $c > 0$, then with probability $1-\delta$, Algorithm 5 will produce a pair (x_T, y_T) satisfying $\operatorname{dist}((x_T, y_T), X^*) \le \varepsilon\sqrt{D}$ using only

$$O\left(\frac{\nu^2(d_1+d_2)}{(1-2p_{\text{fail}})^2}\log\frac{1}{\varepsilon}\log\frac{\log\frac{1}{\varepsilon}}{\delta}\right)$$

samples. We note that the spectral initialization of Charisopoulos et al. [8, Theorem 5.4 and Corollary 5.5] can produce such a pair (x_0, y_0) with sample complexity $O(\nu^2(d_1+d_2)(1-2p_{\text{fail}})^{-2})$ with high probability with $\nu \le \sqrt{3}$. Therefore, when taken together, combining this spectral initialization with Algorithm 5 produces a pair (x_T, y_T) satisfying $\operatorname{dist}((x_T, y_T), X^*) \le \varepsilon\|x\|$ with $O\left(\frac{d_1+d_2}{(1-2p_{\text{fail}})^2}\log\frac{1}{\varepsilon}\log\log\frac{1}{\varepsilon}/\delta\right)$ samples, which is the best known sample complexity for Gaussian robust blind deconvolution, up to logarithmic factors. We note by leveraging standard concentration results, it is possible to prove similar results for empirical average minimization $\min_{(x,y)\in X}\frac{1}{m}\sum_{i=1}^{m} f((x, y), z_i)$, provided z_i are i.i.d samples of z and the number of samples satisfies $m \gtrsim (d_1+d_2)(1-2p_{\text{fail}})^{-2}$.

5 Numerical Experiments

We now evaluate how Algorithm 2 performs both on the statistical recovery problems of Sect. 4 and on a sparse logistic regression problem. We test the convergence behavior, sensitivity to step size, and convergence to an active manifold. While testing the algorithms, we found that Algorithms 2 and 5 perform similarly, despite Algorithm 5 having superior theoretical guarantees. Thus, we do not evaluate Algorithm 5. The problems of Sect. 4 are both convex composite losses of the form in Example 2.1. For these problems, we therefore implement all four models from Example 2.1, using the closed-form solutions developed in [11, Section 5]. For the sparse logistic regression problem, we implement the stochastic proximal gradient method and measure convergence to the support set of the optimal solution. We provide a reference implementation [9] of the methods in `Julia`.

5.1 Convergence behavior

In this section, we demonstrate that Algorithm 2 converges nearly linearly on the Gaussian robust phase retrieval and blind deconvolution problems of Sect. 4 for a particular dimension, noise distribution, corruption frequency, and initialization quality. In phase retrieval, we set $d = 100$ and in blind deconvolution, we set $d = d_1 = d_2 = d$. The measurements are corrupted independently with probability p_{fail}: for phase retrieval, the corruption obeys $\xi = |g|, \ g \sim N(0, 100)$, while for blind deconvolution, it obeys $\xi \sim N(0, 100)$. The algorithms are all randomly initialized at a fixed distance $R_0 > 0$ from the ground truth. The ground truth is normalized in all cases. We use Examples 4.1 and 4.2 to estimate L, η, and μ, and we set $\gamma = 1$, $R_0 = 0.25$, $\delta_2 = \frac{1}{\sqrt{10}}$, and target accuracy $\varepsilon = 10^{-5}$ to obtain T, K and α_0 parameters as in Theorem 3.4

Figures 2 and 3 depict the convergence behavior of Algorithm 2 on robust phase retrieval and blind deconvolution problems in finite sample and streaming settings, respectively. In these plots, solid lines with markers show the mean behavior over 10 runs, while the transparent overlays show one sample standard deviation above and below the mean. In the finite-sample instances, we use $m = 8 \cdot d$ measurements and corrupt a fixed fraction p_{fail} with large magnitude sparse noise; see Fig. 2. In the streaming instances, we draw a new i.i.d. sample at each iteration and corrupt it independently with probability p_{fail}; see Fig. 3. In both figures, we plot in red the rate guaranteed by Theorem 3.4 and observe that the algorithms behave consistently with these guarantees. In presence of noise, the algorithms all converge nearly linearly at the rate predicted by Theorem 3.4, while in the noiseless case, all except the subgradient method converge to an exact solution (modulo numerical accuracy) within far fewer iterations.

5.1.1 Experiments on large-scale problems

We next demonstrate the performance of Algorithm 2 using the subgradient model on large-scale synthetic instances of phase retrieval and blind deconvolution in the finite sample setting with $d = 512 \times \sqrt{512}$. At each step, we sample a random subset of measurements of size $S \in \{32, \sqrt{d}, d\}$. The measurement matrices are sampled from the randomized Hadamard ensemble which consists of k vertically stacked $d \times d$ Hadamard matrices composed with random binary masks:

$$\begin{bmatrix} H_d S_1 \\ H_d S_2 \\ \vdots \\ H_d S_k \end{bmatrix}, \quad \text{where } S_i = \mathbf{diag}\left(\zeta_{i,j}\right_{j=1}^{d}), \ \zeta_{i,j} \sim \text{Unif}(\{\pm 1\})$$

and H_d is the $d \times d$ Walsh-Hadamard matrix. We fix $k = 8$, $p_{\text{fail}} = 0$, $\gamma = 1$ and $\delta_2 = 1/3$ and initialize our algorithm randomly at distance $R_0 = 0.25$ from the ground truth. We estimate L, η and μ using Examples 4.1 and 4.2, adjusted to take into account the batch size S.

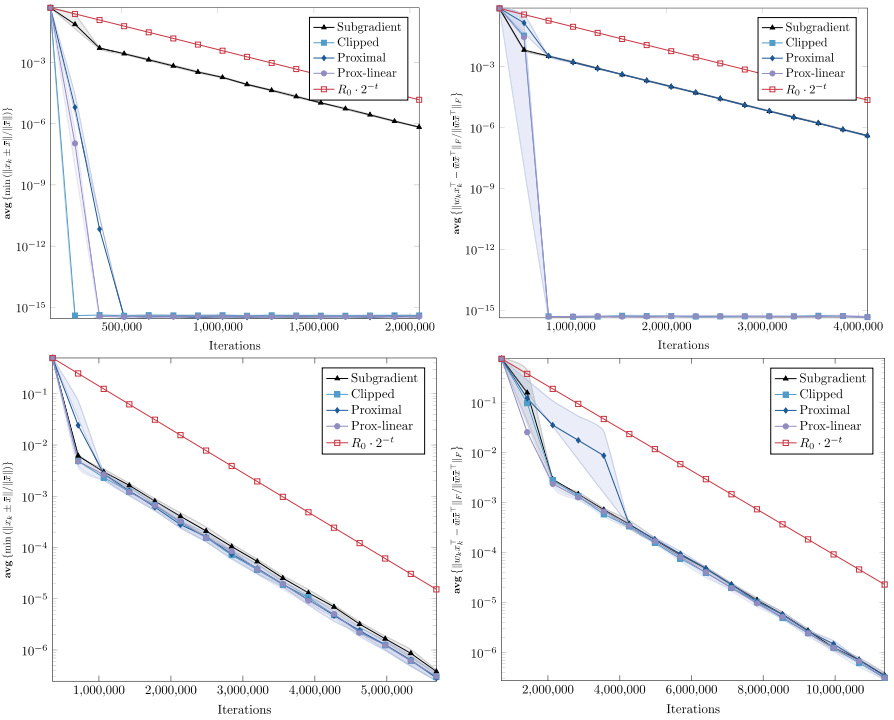

Fig. 2 Convergence behavior for $d = 100$ with finite sample size $m = 8 \cdot d$. Phase Retrieval (left column), Blind Deconvolution (right column), $p_{\text{fail}} = 0{\cdot}0$ (top row), $p_{\text{fail}} = 0{\cdot}2$ (bottom row). Average over 10 runs

We plot the distance from the solution as a function of the number of passes over the dataset in Fig. 4. The plots verify that Algorithm 2 converges nearly linearly to a solution at a dimension-independent rate. Moreover, they indicate that using a larger batch size S does not lead to a reduction over the number of passes required to reach a fixed accuracy.

5.2 Sensitivity to step size

We next explore how Algorithm 2 performs when α_0 is misspecified. Throughout, we scale α_0 by $\lambda := 2^p$ for integers p between -10 and 10. We run 25 trials of the algorithm and for each model and scalar λ, we report two different metrics:

- We report the sample mean and standard deviation of the number of "inner" loop iterations, or samples, needed to reach accuracy $\varepsilon = 10^{-5}$. We use the parameters of Sect. 5.1 to cap the number of total iterations by

$$\frac{16}{(2-\gamma)^2}\left(\frac{L}{\delta_2\mu}\right)^2 \log_2\left(\frac{R_0}{\varepsilon}\right)^3$$

as Theorem 3.4 prescribes. This number is depicted as a dotted line in the figure.

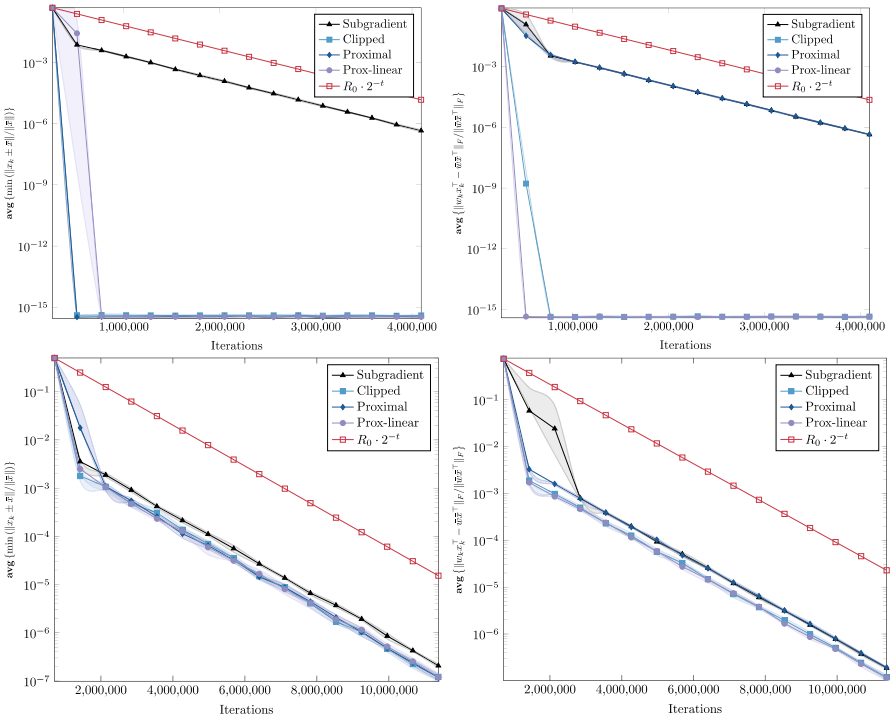

Fig. 3 Convergence behavior for $d = 100$ with streaming data. Phase Retrieval (left column), Blind Deconvolution (right column), $p_{\text{fail}} = 0{\cdot}0$ (top row), $p_{\text{fail}} = 0{\cdot}2$ (bottom row). Average over 10 runs

- We report the sample mean and standard deviation of the distance of the final iterate to the solution set.

Figures 5 and 6 show the results for phase retrieval and blind deconvolution problems with $d = 100$ and $p_{\text{fail}} \in \{0{\cdot}0, 0{\cdot}2\}$. In these plots, solid lines with markers show the mean behavior over 25 runs, while the transparent overlays show one sample standard deviation above and below the mean. The plots show that Algorithm 2 continues to perform as predicted by Theorem 3.4 even if α_0 is misspecified by a few orders of magnitude.

The prox-linear, proximal point, and clipped models perform similarly in all plots. As reported in [4], the prox-linear and clipped methods produce the same iterates. The iterates produced by the stochastic proximal point method and stochastic prox-linear are not identical, but they are practically indistinguishable. This is due to two factors: the proximal and prox-linear models agree up to an error that increases quadratically as we move from the basepoint, and the proximal subproblems force iterates to remain near the basepoint. Running the proximal point method for a much larger stepsize produces different iterates than the prox-linear method, though then the method fails to converge within the specified level of accuracy.

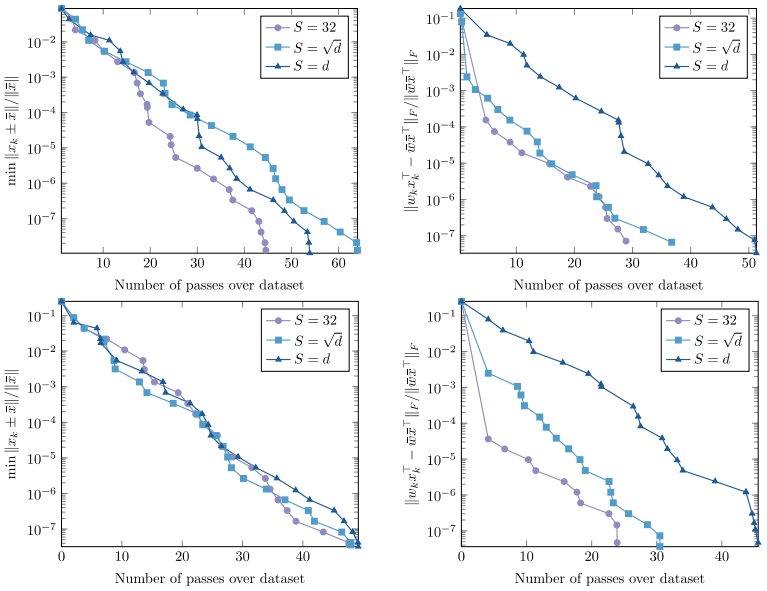

Fig. 4 Convergence behavior of Algorithm 2 using the subgradient model with $d = 128 \times 128$ (top row) and $d = 512 \times 512$ (bottom row) and finite sample size $m = 8d$. Phase retrieval (left column) and blind deconvolution (right column)

5.3 Activity identification

In this section, we demonstrate that Algorithm 2 nearly linearly converges to the active set of nonzero components in a sparse logistic regression problem. We model our experiment on [37, Section 6.2]. There, the authors find a sparse classifier for distinguishing between digits 6 and 7 from the MNIST dataset of handwritten digits [36]. In this problem, we are given a set of $N = 12183$ samples $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$, representing 28×28 dimensional images of digits and their labels, and we seek a target vector $z := (w, b) \in \mathbb{R}^{d+1}$, so that w is sparse and $\operatorname{sign}(w, x_i) + b) = \operatorname{sign}(y_i)$ for most i. To find z, we minimize the function

$$\min_z \frac{1}{N} \sum_{i=1}^{N} f(z; i) + \tau \quad w_1,$$

where each component is a logistic loss:

$$f(z; i) \equiv f(w, b; i) := \log(1 + \exp(-y_i(w, x_i + b))). \tag{5.1}$$

We let $(\bar{w}, \tilde{b})$ denote the minimizer of the logistic loss, which we find using the standard proximal gradient algorithm. Given $\bar{w}$, we denote its support set by $S := \{ i \in [d] : |\bar{w}_i| > \}$, where accounts for the numerical precision of the system.

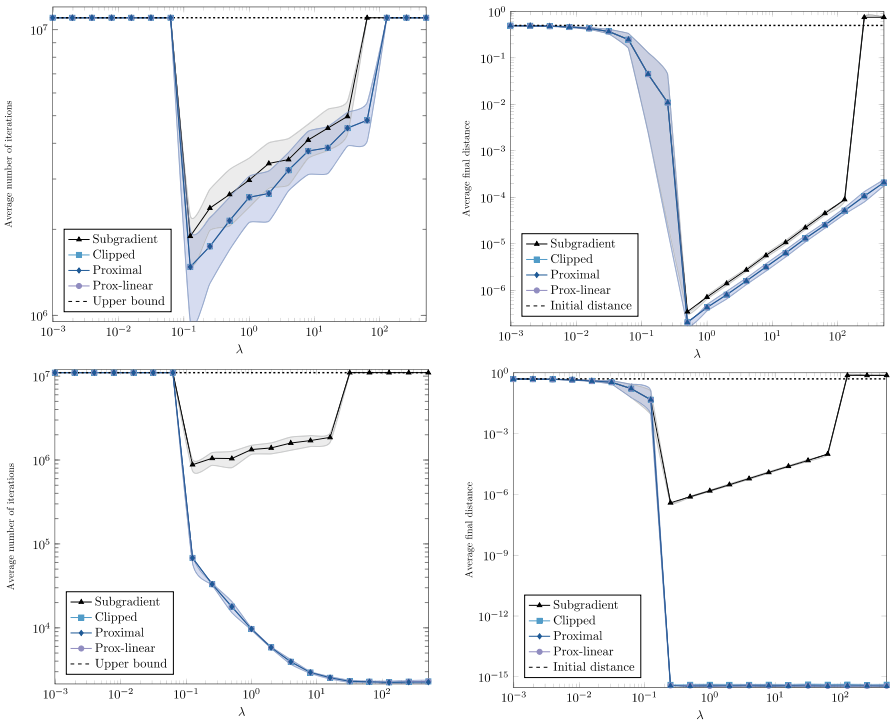

Fig. 5 Sensitivity to step size for the phase retrieval problem with $d = 100$, $p_{\text{fail}} = 0.2$ (top row), $p_{\text{fail}} = 0$ (bottom row). Left: average number of iterations to achieve distance 10^{-5}. Right: average final distance with a fixed computational budget

Our goal is to converge nearly linearly to the set

$$S = \{(w, b) \in \mathbb{R}^{d+1} \mid w_{S^c} = 0\}.$$

To that end, we will apply Algorithm 2 with the stochastic proximal gradient model:

$$f_z((y, a), i) = f(z, i) + \nabla f(z, i), (y, a) - z + \tau \|y\|_1.$$

Algorithm 2 equipped with this model results in the standard stochastic proximal gradient method. To apply the algorithm, we set parameters using Theorem 3.4. We set $\gamma = 1$, $\delta_2 = 1/\sqrt{10}$, $\varepsilon = 10^{-5}$. We initialize $w_0 = 0$, $b_0 = 0$ and set $R_0 = \|(w, b) - (w_0, b_0)\|$. We estimate L by the formula $L := \frac{1}{m} \sum_{i=1}^{m} \|x_i\|_2^2$. Finally, we estimate μ by grid search over (τ, p) using the formula: $\mu = \tau \cdot \sqrt{d} \cdot 2^{-p}$.

5.3.1 Evaluation

We compare the performance of Algorithm 2 with the Regularized Dual Averaging method (RDA) [46, 65], which was shown to have favorable manifold identification

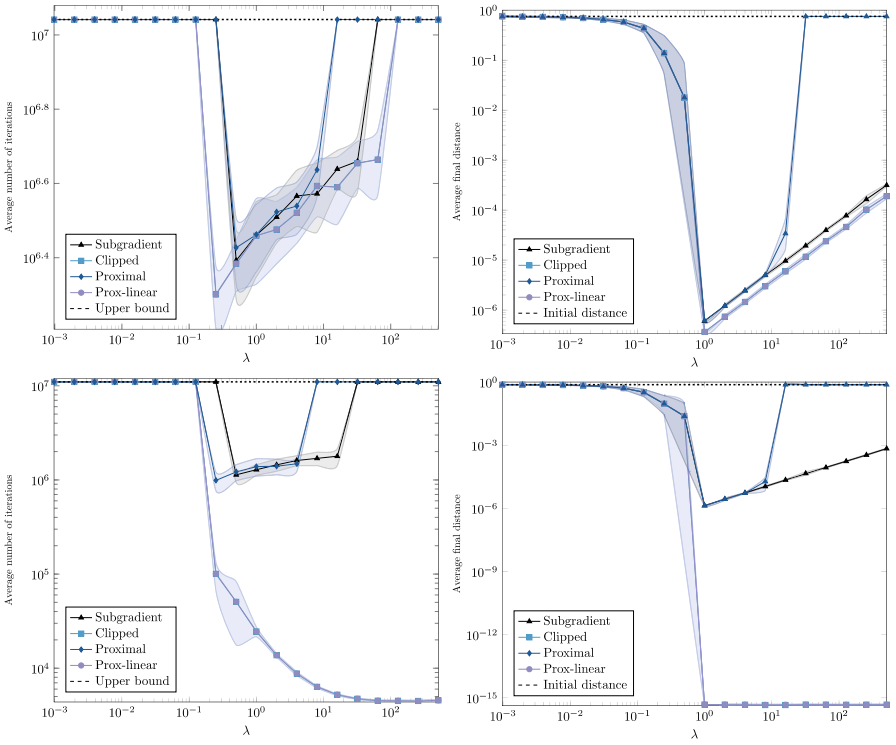

Fig. 6 Sensitivity to step size for the blind convolution problem with $d = 100$, $p_{\text{fail}} = 0{\cdot}2$ (top row), $p_{\text{fail}} = 0$ (bottom row). Left: average number of iterations to achieve distance 10^{-5}. Right: average final distance with a fixed computational budget

properties in [37]. In our setting, the latter method solves the following subproblem:

$$(w_{t+1}, b_{t+1}) \in \operatorname{argmin}_{w,b} \left\langle \bar{g}_t, \begin{pmatrix} w \\ b \end{pmatrix} \right\rangle + \tau \left\| \begin{pmatrix} w \\ b \end{pmatrix} \right\|_1 + \frac{\gamma}{2\sqrt{t}} \left\| \begin{pmatrix} w \\ b \end{pmatrix} \right\|_2^2, \qquad (5.2)$$

where $\bar{g}_t$ in (5.2) is the running average over all stochastic gradients $g_k := \nabla f(z; i_k)$ sampled up to step t, and γ is a tunable parameter which is again determined by a simple grid search. Following the discussion in [37], RDA is initialized at $(w_0, b_0) = 0$; therefore we choose the same initial point (w_0, b_0) for both methods. In addition to RDA, we also performed a comparison with the standard stochastic proximal gradient method, equipped with a range of polynomially decaying step sizes of the form

$$\lambda_k := ck^{-p}, \quad p \in \{1/2, 2/3, 3/4, 1\}.$$

We found that the stochastic proximal gradient method performed comparably with RDA in all metrics, and therefore chose to omit it below.

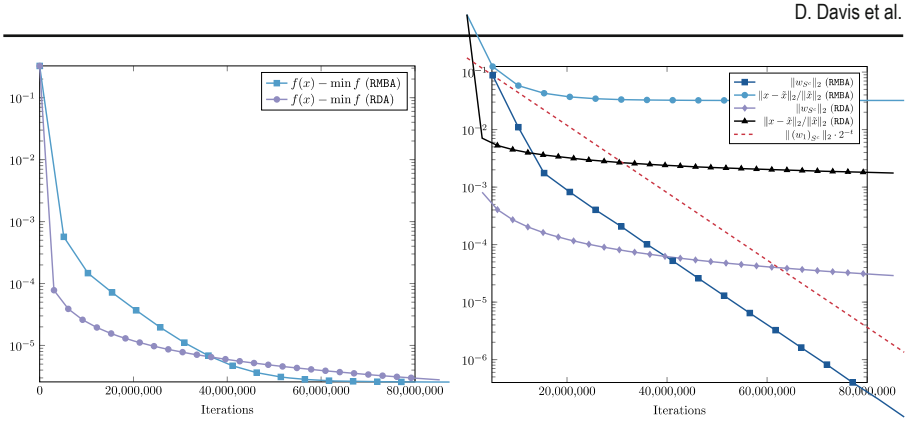

Fig. 7 Performance of RMBA vs. RDA on the sparse logistic regression problem. Left: function value gap. Right: distance to the support set and the solution found by the proximal gradient method

The convergence plots in Fig. 7 confirm that the iterates of Algorithm 2 converge to the set S at a nearly linear rate, while the function values converge at a sublinear rate. In contrast, the iterates generated by RDA converge sublinearly in both metrics.

A Proofs from Sect. 3.2

We will need the following elementary lemma.

Lemma A.1 *Suppose Assumptions (A4) and (A5) hold. Fix an arbitrary* $\gamma \in (0, 2/L)$ *and consider two points* $y \in T_\gamma$ *and* $x \in \mathbb{R}^d$. *Then the estimate holds:*

$$f_\gamma(y, z) \leq f_\gamma(x, z) + L\|y, z\| \, \|x - y\|.$$

Proof Let $v \in \partial f_\gamma(y, z)$ be the minimal norm subgradient. Then, by definition $f_\gamma(y, z) \leq f_\gamma(x, z) + \langle v, y - x \rangle$. Applying (2.5) and Cauchy-Schwarz completes the proof.

A.1 Proof of Lemma 3.3

Throughout the proof, we suppose that Assumptions (A1)–(A5) hold. We let $\mathcal{F}_k$ denote the σ-algebra generated by the history of the algorithm up to iteration k and define the shorthand for conditional expectation $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_k]$. Define the stopping time

$$\tau := \min\{k \geq 0 \mid y_k \notin T_\gamma\},$$

and the sequence of events

$$B := \{y_0 \in T_{\gamma\sqrt{\delta}}\} \quad \text{and} \quad A_k := \{\tau > k\} \cap B.$$

Define also for all indices k, the quantity:

$$D_k := \operatorname{dist}(y_k, S).$$

Recall that our goal is to lower bound $\mathsf{P}(D_{K^*} \leq \varepsilon)$. To this end, we successively compute

$$\begin{aligned}
\mathsf{P}(D_{K^*} \leq \varepsilon) &\geq \mathsf{P}(D_{K^*} \leq \varepsilon \text{ and } A_K) \\
&= \mathsf{P}(D_{K^*} \leq \varepsilon \mid A_K)\,\mathsf{P}(A_K) \\
&= (1 - \mathsf{P}(D_{K^*} \geq \varepsilon \mid A_K))\,\mathsf{P}(A_K) \\
&\geq 1 - \frac{\mathsf{E}[D_{K^*} \mid A_K]}{\varepsilon}\ \mathsf{P}(A_K) \qquad \text{(A.1)} \\
&= \mathsf{P}(A_K) - \frac{\mathsf{E}\, D_{K^*} 1_{A_K}}{\varepsilon},
\end{aligned}$$

where (A.1) follows from Markov's inequality. The result will now follow immediately from the following two propositions, which establish an upper bound on $\mathsf{E}\, D_{K^*} 1_{A_K}$ and a lower bound on $\mathsf{P}(A_K)$, respectively. We note that the first proposition is a quick modification of [11, Lemma 4.2]. We include a proof for completeness.

Proposition A.2 *The following bounds hold:*

$$\mathsf{E}_k\left[D_{k+1}^2 1_{A_k}\right] \leq D_k^2 1_{A_k} + L^2\alpha^2 - (2-\gamma)\mu\alpha\, D_k 1_{A_k}, \qquad \text{(A.2)}$$

$$\mathsf{E}\, D_{K^*} 1_{A_K} \leq \frac{\delta\left(\frac{\gamma\mu}{\eta}\right)^2 + (K+1)\alpha^2 L^2}{(2-\gamma)\mu(K+1)\alpha}. \qquad \text{(A.3)}$$

Proof The loss function $y \to f_{y_k}(y, z_k) + \frac{1}{2\alpha}\|y - y_k\|^2$ is strongly convex on X with constant $1/\alpha$ and y_{k+1} is its minimizer. Hence for any $y \in X$, the inequality holds:

$$\left[f_{y_k}(y, z_k) + \tfrac{1}{2\alpha}\|y - y_k\|^2\right] \geq f_{y_k}(y_{k+1}, z_k) + \tfrac{1}{2\alpha}\|y_{k+1} - y_k\|^2 + \tfrac{1}{2\alpha}\|y - y_{k+1}\|^2.$$

Rearranging and taking expectations, we successively deduce that provided $y_k \in T_\gamma$, we have

$$\begin{aligned}
&\frac{1}{2\alpha} \cdot \mathsf{E}_k\left[\|y - y_{k+1}\|^2 + \|y_{k+1} - y_k\|^2 - \|y - y_k\|^2\right] \\
&\leq \mathsf{E}_k\left[f_{y_k}(y, z_k) - f_{y_k}(y_{k+1}, z_k)\right] \\
&\leq \mathsf{E}_k\left[f_{y_k}(y, z_k) - f_{y_k}(y_k, z_k) + L(y_k, z)\,\|y_{k+1} - y_k\|\right] \qquad \text{(A.4)} \\
&\leq f(y) - f(y_k) + \frac{\eta}{2}\|y - y_k\|^2 + \sup_{w \in T_2} \sqrt{\mathsf{E}_z[L(w, z)^2]} \cdot \sqrt{\mathsf{E}_k[\|y_{k+1} - y_k\|^2]}, \qquad \text{(A.5)}
\end{aligned}$$

where (A.4) follows from Lemma A.1 while inequality (A.5) follows from Cauchy-Schwarz and Assumption (A3).

Define $c := \mathsf{E}_k[\|y_{k+1} - y_k\|^2]$ and notice $c \geq \mathsf{E}_k \|y_k - y_{k+1}\|$. Then, rearranging (A.5), we immediately deduce that if $y_k \in T_\gamma$, we have

$$\begin{aligned}
\frac{1}{2\alpha}\mathsf{E}_k \inf_{y\in \mathrm{projs}(y_k)} \|y - y_{k+1}\|^2 &\leq \frac{1}{2\alpha} \inf_{y\in \mathrm{projs}(y_k)} \mathsf{E}_k \|y - y_{k+1}\|^2 \\
&\leq \frac{\alpha^{-1}+\eta}{2} D_k^2 - \frac{c^2}{2\alpha} + Lc - (f(y_k) - \inf_{y\in \mathrm{projs}(y_k)} f(y)) \\
&\leq \frac{\alpha^{-1}+\eta}{2} D_k^2 - \frac{c^2}{2\alpha} + Lc - \mu D_k \\
&\leq \frac{\alpha^{-1}+\eta}{2} D_k^2 + \frac{\alpha L^2}{2} - \mu D_k.
\end{aligned}$$

where the third inequality follows from assumption (A2), and the fourth inequality follows by maximizing the right-hand-side in $c \in \mathbb{R}$. Then, dividing through by $\frac{1}{2\alpha}$ and multiplying by 1_{A_k}, we arrive at

$$\begin{aligned}
\mathsf{E}_k\left[D_{k+1}^2 1_{A_{k+1}}\right] \leq \mathsf{E}_k\left[D_{k+1}^2 1_{A_k}\right] &\leq (1+\alpha\eta) D_k^2 + \alpha^2 L^2 - 2\mu\alpha D_k 1_{A_k} \\
&\leq D_k^2 + \alpha^2 L^2 - \alpha(2\mu - \eta D_k) D_k 1_{A_k} \\
&\leq D_k^2 1_{A_k} + \alpha^2 L^2 - \alpha(2-\gamma)\mu D_k 1_{A_k},
\end{aligned}$$

where the first inequality follows since $A_{k+1} \subseteq A_k$, the second inequality follows since A_k is $\mathcal{F}_k$ measurable, and the fourth inequality follows since on the event A_k, we have $y_k \in T_\gamma$. This completes the proof of (A.2). Next, applying the law of total expectation, we obtain

$$\mathsf{E}\left[D_{k+1}^2 1_{A_{k+1}}\right] \leq \mathsf{E}\left[D_k^2 1_{A_k}\right] + \alpha^2 L^2 - (2-\gamma)\alpha\mu \mathsf{E}\left[D_k 1_{A_k}\right].$$

Iterating the inequality and rearranging, we deduce

$$\sum_{k=0}^{K} (2-\gamma)\alpha\mu \mathsf{E}\left[D_k 1_{A_k}\right] \leq (K+1)\alpha^2 L^2 + \mathsf{E}\left[D_0^2 1_{A_0}\right]$$

Dividing through by $(K+1)(2-\gamma)\alpha\mu$, we recognize the left-hand side as $\mathsf{E}\left[D_{K^*} 1_{A_{K^*}}\right]$, and therefore

$$\mathsf{E}\left[D_{K^*} 1_{A_{K^*}}\right] \leq \frac{\delta\left(\frac{\gamma\mu}{\eta}\right)^2 + (K+1)\alpha^2 L^2}{(2-\gamma)(K+1)\mu\alpha}.$$

Finally, note $\mathsf{E}\left[D_{K^*} 1_{A_K}\right] \leq \mathsf{E}\left[D_{K^*} 1_{A_{K^*}}\right]$ since $A_K \subseteq A_{K^*}$. This completes the proof of the proposition.

Now we will estimate the probability of the event A_K.

Proposition A.3 *The estimate holds:*

$$\mathrm{P}(A_K) \geq \mathrm{P}(B) - \left(\delta + \frac{\eta}{\gamma\mu}\right)^2 K\alpha^2\mathrm{L}^2 .$$

Proof Observe the decomposition

$$\mathrm{P}(B) = \mathrm{P}(B \text{ and } \tau \leq K) + \mathrm{P}(B \text{ and } \tau > K) = \mathrm{P}(B \text{ and } \tau \leq K) + \mathrm{P}(A_K),$$

and therefore

$$\mathrm{P}(A_K) \geq \mathrm{P}(B) - \mathrm{P}(B \text{ and } \tau \leq K). \tag{A.6}$$

We aim to upper bound $\mathrm{P}(B \text{ and } \tau \leq K)$. To this end, let $Y_k = D_{k\wedge\tau}^2 1_B$ denote the stopped process. We successively compute

$$\mathrm{P}(B \text{ and } \tau \leq K) = \mathrm{P}\left(Y_\tau 1_{\tau\leq K} > \left(\frac{\gamma\mu}{\eta}\right)^2\right) \leq \frac{\mathrm{E}\left[Y_\tau 1_{\tau\leq K}\right]}{\left(\frac{\gamma\mu}{\eta}\right)^2}, \tag{A.7}$$

where the last estimate uses Markov's inequality. Next, observe

$$\mathrm{E}\left[Y_\tau 1_{\tau\leq K}\right] \leq \mathrm{E}\left[Y_K 1_{\tau>K}\right] + \mathrm{E}\left[Y_K 1_{\tau\leq K}\right] = \mathrm{E}\left[Y_K\right]. \tag{A.8}$$

We next upper bound $\mathrm{E}[Y_K]$. To this end, observe

$$\begin{aligned}
\mathrm{E}_k\left[Y_{k+1}\right] &= \underbrace{\mathrm{E}_k\left[Y_{k+1}1_{\tau\leq k}\right]}_{1} + \underbrace{\mathrm{E}_k\left[Y_{k+1}1_{\tau>k}\right]}_{2} \\
&= Y_k 1_{\tau\leq k} + \mathrm{E}_k\left[D_{k+1}^2 1_{A_k}\right] \\
&\leq Y_k 1_{\tau\leq k} + D_k^2 1_{A_k} + \alpha^2\mathrm{L}^2 - (2-\gamma)\alpha\mu D_k 1_{A_k} \qquad \text{(A.9)} \\
&\leq Y_k 1_{\tau\leq k} + D_{k\wedge\tau}^2 1_{\tau>k} 1_B + \alpha^2\mathrm{L}^2 = Y_k + \alpha^2\mathrm{L}^2,
\end{aligned}$$

where (A.9) follows from (A.2). We now use the law of total expectation to iterate the above inequality:

$$\mathrm{E}\left[Y_K\right] \leq \mathrm{E}\left[Y_0\right] + K\alpha^2\mathrm{L}^2 \leq D_0^2 1_B + K\alpha^2\mathrm{L}^2 \leq \delta\left(\frac{\gamma\mu}{\eta}\right)^2 + K\alpha^2\mathrm{L}^2. \tag{A.10}$$

Combining the estimates (A.6), (A.7), (A.8), and (A.10) completes the proof.

B Proofs from Sect. 3.3

B.1 The ensemble method

Lemma B.1 (Ensemble method). *Let* $\{x_i\}_{i=1}^m$ *be i.i.d. realizations of a random vector* $x \in \mathbb{R}^d$. *Suppose that the estimate holds:*

$$\mathbb{P}(\|x - \bar{x}\| \leq \varepsilon) \geq p$$

where $p \in (\frac{1}{2}, 1)$ *and* $\varepsilon > 0$ *are some real numbers and* $\bar{x} \in \mathbb{R}^d$ *is a vector. Then with probability at least* $1 - \exp(-\frac{1}{2p}m(p - \frac{1}{2})^2)$, *there exists an index* i^* *satisfying*

$$|B_{2\varepsilon}(x_{i^*}) \cap \{x_i\}_{i=1}^m| > \frac{m}{2} \tag{B.1}$$

and for any index i^* *satisfying* (B.1) *it must be that* $\|x_{i^*} - \bar{x}\| \leq 3\varepsilon$.

Proof By Chernoff's bound, with probability at least $1 - \exp(-\frac{1}{2p}m(p - \frac{1}{2})^2)$, the estimate holds:

$$|\{i : \|x_i - \bar{x}\| \leq \varepsilon\}| > \frac{m}{2}$$

In particular, there exists an index i^* satisfying (B.1) Fix such an index i^*. Clearly, there must exist another index j satisfying $x_j \in B_\varepsilon(\bar{x}) \cap B_{2\varepsilon}(x_{i^*})$. We therefore conlude $\|x_{i^*} - \bar{x}\| \leq \|x_{i^*} - x_j\| + \|x_j - \bar{x}\| \leq 3\varepsilon$. This completes the proof.

B.2 Proof of Lemma 3.5

Fix $\gamma \in (0, 2)$ and a point $y \in T_\gamma$. Recall that f is η-weakly convex on an open convex set containing X [11, Lemma 4.1]. Consequently, the proximal subproblem (3.7) is $(\rho - \eta)$-strongly convex. Before proving the remaining portion of the lemma, we first show that subgradients of the extended valued function $f + \delta_X$ are bounded below. This was essentially already observed in [12, Lemma 2.1]. We provide a quick proof for completeness.

Lemma B.2 *The estimate:*

$$\operatorname{dist}(0, \partial f(x) + N_X(x)) \geq \left(1 - \frac{\gamma}{2}\right)\mu \quad \text{holds for all } x \in T_\gamma \backslash X^*.$$

Proof Fix any $x \in T_\gamma \backslash X^*$ and $v \in \partial f(x) + N_X(x)$, and let $\bar{x} \in \operatorname{proj}_{X^*}(x)$. We successively compute

$$\mu \cdot \operatorname{dist}(x, X^*) \leq f(x) - \inf_X f \leq \langle v, x - \bar{x} \rangle + \frac{\eta}{2}\operatorname{dist}(x, X^*)^2,$$

where the first inequality follows from sharpness (A2), and the second from weak convexity of f and convexity of X. Rearranging and using the Cauchy-Schwarz inequality, we deduce

$$\mu - \frac{\eta}{2}\operatorname{dist}(x, X^*) \quad \operatorname{dist}(x, X^*) \leq \operatorname{dist}(x, X^*) \qquad v \qquad .$$

Dividing both sides by $\operatorname{dist}(x, X^*)$ yields the result.

Now, let $\bar{y}$ be any minimizer of the proximal problem (3.7) and suppose $\rho < \frac{2-\gamma}{2\gamma}\eta$. Clearly, to establish Lemma 3.5, it suffices to argue the inclusion $\bar{y} \in X^*$. To this end, choose any $\hat{y} \in \operatorname{proj}_{X^*}(y)$. Observe

$$\frac{\rho}{2}\|y - \bar{y}\|^2 \leq f(\hat{y}) - f(\bar{y}) + \frac{\rho}{2}\|y - \bar{y}\|^2 \leq \frac{\rho}{2}\operatorname{dist}^2(y, X^*) \leq \frac{\rho\gamma^2\mu^2}{2\eta^2},$$

where the first inequality follows from the definition of $\bar{y}$, the second uses the definition of $\hat{y}$, and the third follows from the assumption $y \in T_\gamma$. Thus, we deduce $\|y - \bar{y}\| \leq \frac{\gamma\mu}{\eta}$. Consequently, using sharpness we conclude

$$\mu\operatorname{dist}(\bar{y}, X^*) \leq f(\bar{y}) - f(\hat{y}) \leq \frac{\rho}{2}\|y - \bar{y}\|^2 \leq \frac{\rho}{2}\frac{\gamma^2\mu^2}{\eta^2} \leq \frac{\gamma\mu^2}{2\eta}$$

where the last inequality follows from the assumption $\rho < \frac{2-\gamma}{2\gamma}\eta$. Consequently, $\bar{y}$ lies in the tube T_γ. Now, define $v := \rho(y - \bar{y}) \in \partial f(\bar{y}) + N_X(\bar{y})$. Appealing to Lemma B.2, we deduce in the case $\bar{y} \notin X^*$, the contradiction $(1 - \frac{\gamma}{2})\mu/\rho \leq \|v\|/\rho = \|y - \bar{y}\| \leq \frac{\gamma\mu}{\eta}$. Therefore, $\bar{y}$ lies in X^*, as we had to show.

B.3 Proof of Lemma 3.6

As in the proof of Lemma 3.3, we let $\mathcal{F}_k$ denote the σ-algebra generated by the history of the algorithm up to iteration k and define the shorthand for conditional expectation $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_k]$. Define the stopping time

$$\tau := \min\{k \mid y_k \notin T_\gamma\},$$

the sequence of events $A_k := \{\tau > k\}$, and the quantities

$$D_k := \|y_k - \bar{y}_0\| \qquad \text{and} \qquad E_k := \frac{\alpha^{-1} + \eta}{2}D_k^2 + \frac{\rho}{2}\|y_k - y_0\|^2.$$

Note that by Lemma 3.5, the inclusion $\bar{y}_0 \in \operatorname{proj}_{X^*}(y_0)$ holds. We will use this observation throughout.

We begin with the estimate

$$\mathbb{P}\{D_{K^*}^2 \leq \varepsilon^2\} \geq \mathbb{P}\{D_{K^*}^2 \leq \varepsilon^2 \mid A_K\}\,\mathbb{P}(A_K)$$

$$
\begin{aligned}
&= 1 - \mathsf{P}\left(D_{K^*}^2 \geq \varepsilon^2 \mid A_K\right)\mathsf{P}(A_K) \\
&\geq 1 - \frac{\mathsf{E}\left(D_{K^*}^2 \mid A_K\right)}{\varepsilon^2}\mathsf{P}(A_K) \\
&= 1 - \frac{\mathsf{E}\left(D_{K^*}^2 1_{A_K}\right)}{\varepsilon^2 \mathsf{P}(A_K)}\mathsf{P}(A_K) \\
&= \mathsf{P}(A_K) - \frac{\mathsf{E}\left(D_{K^*}^2 1_{A_K}\right)}{\varepsilon^2}
\end{aligned}
$$

The result will now follow immediately from the following two propositions, which establish an upper bound on $\mathsf{E}\left(D_{K^*}^2 1_{A_K}\right)$ and a lower bound on $\mathsf{P}(A_K)$, respectively. We note that the first proposition is a quick modification of [11, Lemma 4.2]. We include a proof for completeness.

Proposition B.3 *Define* $\nu := \rho - \eta$ *. Then the following bounds hold:*

$$
\mathsf{E}_k\left(E_{k+1} 1_{A_k}\right) \leq E_k 1_{A_k} - \frac{\nu}{2} D_k^2 1_{A_k} + \frac{L^2 \alpha}{2}. \tag{B.2}
$$

$$
\mathsf{E}\left(D_{K^*}^2 1_{A_K}\right) \leq \frac{(K+1)L^2\alpha + (\alpha^{-1} + \eta) \cdot \delta \left(\frac{\gamma \mu}{\eta}\right)^2}{\nu(K+1)}. \tag{B.3}
$$

Proof Define the function $g(y) := f(y) + \frac{\rho}{2}\|y - y_0\|^2$ and notice that g is strongly convex with parameter ν. Observe also that the loss function $y \mapsto f_{y_k}(y, z_k) + \frac{1}{2\alpha}\|y - y_k\|^2 + \frac{\rho}{2}\|y - y_0\|^2$ is strongly convex on X with constant $\alpha^{-1} + \rho$ and y_{k+1} is its minimizer. Hence for any $y \in X$, the inequality holds:

$$
\begin{aligned}
&f_{y_k}(y, z_k) + \tfrac{1}{2\alpha}\|y - y_k\|^2 + \tfrac{\rho}{2}\|y - y_0\|^2 \\
&\geq f_{y_k}(y_{k+1}, z_k) + \tfrac{1}{2\alpha}\|y_{k+1} - y_k\|^2 + \tfrac{\rho}{2}\|y_{k+1} - y_0\|^2 \\
&\quad + \tfrac{\alpha^{-1}+\rho}{2}\|y - y_{k+1}\|^2.
\end{aligned}
$$

Rearranging and taking expectations we successively deduce that if $y_k \in T_\gamma$, then

$$
\begin{aligned}
&\mathsf{E}_k\left(\frac{\alpha^{-1} + \rho}{2}\|y - y_{k+1}\|^2 + \frac{1}{2\alpha}\|y_{k+1} - y_k\|^2 - \frac{1}{2\alpha}\|y - y_k\|^2\right) \\
&\leq \mathsf{E}_k\left(f_{y_k}(y, z_k) + \frac{\rho}{2}\|y - y_0\|^2 - \left(f_{y_k}(y_{k+1}, z_k) + \frac{\rho}{2}\|y_{k+1} - y_0\|^2\right)\right) \\
&\leq \mathsf{E}_k\left(f_{y_k}(y, z_k) + \frac{\rho}{2}\|y - y_0\|^2 - \left(f_{y_k}(y_k, z_k) + \frac{\rho}{2}\|y_{k+1} - y_0\|^2\right)\right) \\
&\qquad + \mathsf{E}_k\left(L(y_k, z_k)\|y_{k+1} - y_k\|\right) \\
&\leq \mathsf{E}_k\left(f(y) + \frac{\rho}{2}\|y - y_0\|^2 - \left(f(y_k) + \frac{\rho}{2}\|y_k - y_0\|^2\right)\right)
\end{aligned} \tag{B.4}
$$

$$
\begin{aligned}
&+ \mathsf{E}_k\left[\frac{\rho}{2}\|y_k - y_0\|^2 - \frac{\rho}{2}\|y_{k+1} - y_0\|^2\right] \\
&+ \mathsf{E}_k\left[\frac{\eta}{2}\|y - y_k\|^2 + L(y_k, z_k)\|y_{k+1} - y_k\|^2\right] \\
&\leq g(y) - g(y_k) + \frac{\eta}{2}\|y - y_k\|^2 + \sup_{w \in T_2}\sqrt{\mathsf{E}_z[L(w,z)^2]}\cdot\sqrt{\mathsf{E}_k[\|y_{k+1} - x_k\|^2]} \\
&+ \mathsf{E}_k\left[\frac{\rho}{2}\|y_k - y_0\|^2 - \frac{\rho}{2}\|y_{k+1} - y_0\|^2\right] \qquad \text{(B.5)}
\end{aligned}
$$

where (B.4) follows from Lemma A.1, while inequality (B.5) follows from Cauchy-Schwarz and Assumption (A3).

Define $c := \sup_{w \in T_2}\sqrt{\mathsf{E}_k[\|y_{k+1} - y_k\|^2]}$ and notice $c \geq \mathsf{E}_k\|y_k - y_{k+1}\|$. Thus, letting $y = \bar{y}_0$ and rearranging (B.5), we immediately deduce that if $y_k \in T_\gamma$, we have

$$
\begin{aligned}
&\mathsf{E}_k\left[\frac{\alpha^{-1}+\eta}{2}D_{k+1}^2 + \frac{\rho}{2}\|y_{k+1} - y_0\|^2\right] \\
&\leq \frac{\alpha^{-1}+\eta}{2}D_k^2 + \frac{\rho}{2}\|y_k - y_0\|^2 - \frac{c^2}{2\alpha} + Lc - (g(y_k) - g(\bar{y}_0)) \\
&\leq \frac{\alpha^{-1}+\eta}{2}D_k^2 + \frac{\rho}{2}\|y_k - y_0\|^2 + \frac{L^2\alpha}{2} - \frac{\nu}{2}D_k^2,
\end{aligned}
$$

where the second inequality follows from strong convexity of g and by maximizing the right-hand-side in $c \in \mathbb{R}$. Thus, multiplying through by 1_{A_k}, we deduce that

$$
\mathsf{E}_k\left[E_{k+1}1_{A_k}\right] \leq E_k 1_{A_k} - \frac{\nu}{2}D_k^2 1_{A_k} + \frac{L^2\alpha}{2}.
$$

which proves (B.2). Iterating (B.2), using the tower rule, and rearranging, we deduce

$$
\frac{\nu}{2}\sum_{k=0}^{K}\mathsf{E}\left[D_k^2 1_{A_k}\right] \leq (K+1)\frac{L^2\alpha}{2} + E_0 \leq (K+1)\frac{L^2\alpha}{2} + \frac{\alpha^{-1}+\eta}{2}\cdot\delta\left(\frac{\gamma\mu}{\eta}\right)^2,
$$

where the last inequality follows from Lemma 3.5 and the assumption $y_0 \in T_{\gamma\sqrt{\delta}}$. Dividing through by $\frac{\nu}{2}(K+1)$, we deduce

$$
\mathsf{E}\left[D_{K^*}^2 1_{A_{K^*}}\right] \leq \frac{(K+1)L^2\alpha + (\alpha^{-1}+\eta)\cdot\delta\left(\frac{\gamma\mu}{\eta}\right)^2}{\nu(K+1)}.
$$

Finally, note $\mathsf{E}\left[D_{K^*}^2 1_{A_K}\right] \leq \mathsf{E}\left[D_{K^*}^2 1_{A_{K^*}}\right]$ since $A_K \subseteq A_{K^*}$. This completes the proof of the proposition.

Now we estimate the probability of the event A_K.

Proposition B.4 *The estimate holds:*

$$\mathrm{P}(A_K) \geq 1 - \delta - \left(\frac{\eta}{\gamma\mu}\right)^2 K\mathsf{L}^2\alpha^2.$$

Proof Define $r := 2(\alpha^{-1} + \eta)^{-1}$, let $\overline{Y}_k = \mathrm{dist}^2(y_{k\wedge\tau}, X^*)$, and let $Y_k = r\,E_{k\wedge\tau}$ denote the stopped process. We now estimate

$$\mathrm{P}(\tau \leq K) = \mathrm{P}\left(\overline{Y}_\tau 1_{\tau\leq K} > \left(\frac{\gamma\mu}{\eta}\right)^2\right) \leq \frac{\mathrm{E}\,\overline{Y}_\tau 1_{\tau\leq K}}{\left(\frac{\gamma\mu}{\eta}\right)^2}. \tag{B.6}$$

Next we upper bound the right-hand-side:

$$\mathrm{E}\,\overline{Y}_\tau 1_{\tau\leq K} \leq \mathrm{E}\,Y_\tau 1_{\tau\leq K} \leq \mathrm{E}\,[Y_K 1_{\tau > K}] + \mathrm{E}\,Y_K 1_{\tau\leq K} = \mathrm{E}\,[Y_K], \tag{B.7}$$

where the first inequality follows from the bound $\overline{Y}_k \leq Y_k$. Next, observe

$$\begin{aligned}
\mathrm{E}_k\,Y_{k+1} &= \mathrm{E}_k\,Y_{k+1}1_{\tau\leq k} + \mathrm{E}_k\,Y_{k+1}1_{\tau > k} \\
&= Y_k 1_{\tau\leq k} + \mathrm{E}_k\,r\,E_{k+1}1_{A_k} \\
&\leq Y_k 1_{\tau\leq k} + r\,E_k 1_{A_k} - \frac{\nu}{2}D_k^2 1_{A_k} + \frac{\mathsf{L}^2\alpha}{2} \\
&\leq Y_k 1_{\tau\leq k} + r\,E_k 1_{\tau > k} + r\frac{\mathsf{L}^2\alpha}{2} = Y_k + \frac{\mathsf{L}^2\alpha}{(\alpha^{-1}+\eta)},
\end{aligned}$$

where the first inequality follows from (B.2). We now use the law of total expectation to iterate the above inequality:

$$\mathrm{E}\,[Y_K] \leq \mathrm{E}\,[Y_0] + \frac{K\mathsf{L}^2\alpha}{(\alpha^{-1}+\eta)} \leq D_0^2 + K\mathsf{L}^2\alpha^2 \leq \delta\left(\frac{\gamma\mu}{\eta}\right)^2 + K\mathsf{L}^2\alpha^2, \tag{B.8}$$

where the second inequality follows from the equality $r\,E_0 = D_0^2 = \mathrm{dist}^2(y_0, X^*)$. Combining (B.6), (B.7), and (B.8) completes the proof.

C Proofs from Sect. 4

C.1 Proof of Theorem 4.1

The equality $X^* = \{\pm\bar{x}\}$ and sharpness follows along similar lines as in [18, Proposition 4] and [13, Lemma B.8]. We sketch a quick argument for completeness. Fix $x \in \mathbb{R}^d$ throughout the proof. Let $\hat{f}(x, z) = |a, x\rangle^2 - a, \bar{x}\rangle^2|$ denote the "outlier-free" loss function and set $\hat{f}(x) := \mathrm{E}[\hat{f}(x, z)]$. Setting $v := \frac{x-\bar{x}}{\|x-\bar{x}\|}$ and $w := \frac{x+\bar{x}}{\|x+\bar{x}\|}$,

we have

$$\hat{f}(x,z) = |\langle a, x-\bar{x}\rangle\langle a, x+\bar{x}\rangle| = |x-\bar{x}||x+\bar{x}||\langle a, v\rangle\langle a, w\rangle|.$$

Therefore, we deduce

$$\hat{f}(x) := \mathbb{E}_{z}\hat{f}(x,z)^{2} \geq \tilde{\mu}\,|x-\bar{x}|\,|x+\bar{x}| \geq \mu\,|x|\,\tilde{\operatorname{dist}}(x,\{\pm\bar{x}\}).$$

Now, using this bound, we find that

$$\begin{aligned} f(x) - f(\pm\bar{x}) &= (1-p_{\text{fail}})(\hat{f}(x) - \hat{f}(\bar{x})) + p_{\text{fail}}\mathbb{E}_{a,\xi}\,|\langle a, x\rangle^2 - \langle a, \bar{x}\rangle^2 - \xi| - |\xi| \\ &\geq (1-p_{\text{fail}})\hat{f}(x) - p_{\text{fail}}\mathbb{E}_{a}\,|\langle a, x\rangle^2 - \langle a, \bar{x}\rangle^2| \\ &= (1-2p_{\text{fail}})\hat{f}(x) \geq (1-2p_{\text{fail}})\tilde{\mu}\,|x|\,\operatorname{dist}(x,\{\pm\bar{x}\}). \end{aligned}$$

In particular, we deduce the equality $X^* = \{\pm\bar{x}\}$ and the sharpness estimate (A2) with $\mu = (1-2p_{\text{fail}})\tilde{\mu}\,|\bar{x}|$. Now we estimate the parameters of the models.

We begin with an estimate of η. To that end, fix $y \in \mathbb{R}^d$. Then, using the expansion $\langle a, y\rangle^2 = \langle a, x\rangle^2 + 2\langle a, x\rangle\langle a, y-x\rangle + \langle a, y-x\rangle^2$, we find that for any $z \in$, we have

$$\begin{aligned} f(y,z) &= |\langle a, y\rangle^2 - (\langle a, \bar{x}\rangle^2 + u\cdot\xi)| \\ &= |\langle a, x\rangle^2 + 2\langle a, x\rangle\langle a, y-x\rangle + \langle a, y-x\rangle^2 - (\langle a, \bar{x}\rangle^2 + u\cdot\xi)| \\ &\geq f_x^{pl}(y,z) - \langle a, y-x\rangle^2. \end{aligned}$$

We use this inequality to estimate η for each of the models. Let us analyze each of the models in turn:

(**prox-linear**) We have, $\mathbb{E}_{z} f_x^{pl}(y,z)^{2} \leq \mathbb{E}_{z}\, f(y,z) + \langle a, y-x\rangle^2 \leq f(y) + \tilde{\eta}\,|y-x|^2$.

(**subgradient**) By inspection, we have $G(x,z) \in \partial f_x^{pl}(x,z)$. Thus, we have

$$f_x^s(y,z) = f_x^{pl}(x,z) + \langle G(x,z), y-x\rangle \leq f_x^{pl}(y,z) \leq f(y,z) + \langle a, y-x\rangle^2,$$

and consequently, $\mathbb{E}_{z} f_x^s(y,z) \leq f(y) + \tilde{\eta}\,|y-x|^2$.

(**clipped Subgradient**) As before, we have

$$\begin{aligned} \max\{f_x^s(y,z), 0\} &= \max\{f_x^{pl}(x,z) + \langle G(x,z), y-x\rangle, 0\} \\ &\leq \max\{f_x^{pl}(y,z), 0\} \leq f(y,z) + \langle a, y-x\rangle^2, \end{aligned}$$

and consequently, we have $\mathbb{E}_{z} f_x^{cl}(y,z) \leq f(y) + \tilde{\eta}\,|y-x|^2$.

Therefore in all three cases, we have $\eta = 2\tilde{\eta}$.

Now we analyze $L(x,z)$. Any subgradient of any of the models evaluated at a point $x \in \mathbb{R}^d$ is of the form $v = 2s\langle a, x\rangle a$ for some $s \in [-1,1]$. Consequently, in all three cases, we have $\min_{v\in\partial f_x(x,z)} |v| \leq 2|\langle a, x\rangle|\,\|a\| \leq L(x,z)$, as desired.

C.2 Proof of Theorem 4.1

Throughout the proof, let $(x, y), (\hat{x}, \hat{y}) \in X$. Let $\hat{f}((x, y), z) = |\langle x\, r, y - \bar{x}\, r, \bar{y}\rangle|$ denote the outlier free objective and notice that with $M = \frac{xy^T - \bar{x}\bar{y}^T}{\|xy^T - \bar{x}\bar{y}^T\|_F}$, we have

$$\hat{f}((x, y), z) = |\langle (xy^T - \bar{x}\bar{y}^T)r\rangle| = \|xy^T - \bar{x}\bar{y}^T\|_F |\langle Mr\rangle|.$$

Now taking into account [8, Proposition 4.2], we have

$$\|xy^T - \bar{x}\bar{y}^T\|_F \geq \frac{\sqrt{D}}{\sqrt{2}\sqrt{2(\nu + 1)}} \operatorname{dist}((x, y), X^*).$$

Thus, it follows that

$$\begin{aligned}\hat{f}(x, y) := \mathbb{E}_{\xi}\left[\hat{f}((x, y), z)^2\right] &= \|xy^T - \bar{x}\bar{y}^T\|_F \mathbb{E}_{\xi}[|\langle a, Ma\rangle|] \\ &\geq \frac{p\sqrt{D}}{\sqrt{2}\sqrt{2(\nu + 1)}} \operatorname{dist}((x, y), X^*).\end{aligned}$$

Finally, using this bound, we find that for any $\alpha > 0$ that

$$\begin{aligned}f(x, y) - f(\alpha\bar{x}, (1/\alpha)\bar{y}) &= (1 - p_{\text{fail}})(\hat{f}(x, y) - \hat{f}(\alpha\bar{x}, (1/\alpha)\bar{y})) \\ &\quad + p_{\text{fail}}\mathbb{E}_{\nu, r, \xi}[|\langle x\, r, y - \bar{x}\, r, \bar{y} - \xi\rangle| - |\langle \xi\rangle|] \\ &\geq (1 - p_{\text{fail}})\hat{f}(x, y) - p_{\text{fail}}\mathbb{E}_{\nu, r}[|\langle x\, r, y - \bar{x}\, r, \bar{y}\rangle|] \\ &= (1 - 2p_{\text{fail}})\hat{f}(x, y) \geq \frac{p(1 - 2p_{\text{fail}})\sqrt{D}}{\sqrt{2}\sqrt{2(\nu + 1)}} \operatorname{dist}((x, y), X^*).\end{aligned}$$

This proves sharpness. Now we estimate the parameters of the models.

Let us begin with η. To that end, we observe that with $M = \frac{(y - \hat{y})(x - \hat{x})^T}{\|y - \hat{y}\|\|x - \hat{x}\|^T}$, we have

$$\begin{aligned}|f((\hat{x}, \hat{y}), z) - f^{pl}_{(x,y)}((\hat{x}, \hat{y}), z)| &\leq |\langle \hat{x}\hat{y}^T - xy^T - x(\hat{y} - y) - y(\hat{x} - x)^T)r\rangle| \\ &= |\langle ((y - \hat{y})(x - \hat{x})^T)r\rangle| \\ &= |\langle Mr\rangle| \|y - \hat{y}\| \|x - \hat{x}\| \\ &\leq \frac{|\langle Mr\rangle|}{2}\left(\|y - \hat{y}\|^2 + \|x - \hat{x}\|^2\right).\end{aligned}$$

We use this inequality to establish the weak quadratic approximation property for each of the models. Let us analyze each of the models in turn:

(Prox-linear) Taking expectations, we have

$$\mathbb{E}_{\xi}\left[f^{pl}_{(x,y)}((\hat{x}, \hat{y}), z)^2\right] \leq f(\hat{x}, \hat{y}) + \frac{\eta}{2}\left(\|y - \hat{y}\|^2 + \|x - \hat{x}\|^2\right).$$

(Subgradient) By inspection, the inclusion holds $G((x, y), z) \in \partial f^{pl}_{(x,y)}((x, y), z)$. Therefore,

$$f^{\delta}_{(x,y)}((\hat{x}, \hat{y}), z) = f^{pl}_{(x,y)}((x, y), z) + \langle G((x, y), z), (\hat{x}, \hat{y}) - (x, y) \rangle \leq f^{pl}_{(x,y)}((\hat{x}, \hat{y}), z),$$

and consequently $\mathbb{E}\left[f^{\delta}_{(x,y)}((\hat{x}, \hat{y}), z)\right]^2 \leq \mathbb{E}\left[f^{pl}_{(x,y)}((\hat{x}, \hat{y}), z)\right]^2 \leq f(\hat{x}, \hat{y}) + \frac{\eta}{2}\left\|y - \hat{y}\right\|^2 + \left\|x - \hat{x}\right\|^2$.

(Clipped Subgradient) As before, we have

$$\max\{f^{\delta}_{(x,y)}((\hat{x}, \hat{y}), z), 0\} = \max\{f^{pl}_{(x,y)}((x, y), z) + \langle G((x, y), z), (\hat{x}, \hat{y}) - (x, y)\rangle, 0\} \leq f^{pl}_{(x,y)}((\hat{x}, \hat{y}), z),$$

and consequently $\mathbb{E}\left[f^{cl}_{(x,y)}((\hat{x}, \hat{y}), z)\right]^2 \leq \mathbb{E}\left[f^{pl}_{(x,y)}((\hat{x}, \hat{y}), z)\right]^2 \leq f(\hat{x}, \hat{y}) + \frac{\eta}{2}\left\|y - \hat{y}\right\|^2 + \left\|x - \hat{x}\right\|^2$.

Therefore in all three cases, we have $\eta = \tilde{\eta}$.

Now we analyze $L((x, y), z)$. Any subgradient of any of the models evaluated at the point (x, y) is of the form $v = (r, y, x\, r)s$ for some $s \in [-1, 1]$. Therefore, $\min_{v \in \partial f_x(x,z)} \langle v, (x, r, y, r)\rangle \leq -\|r, y\|^2 \leq L((x, y), z)$, as desired.

Finally, the bound on L follows since

$$L^2 \leq \sup_{(x,y)\in T_2} \mathbb{E}\left[L((x, y), z)\right]^2 \leq \sup_{(v,w)\in S^{d_1-1}\times S^{d_2-1}} v^2 D\mathbb{E}\left[L((v, w), z)\right]^2 \leq v^2 D\tilde{L}^2,$$

as desired.

D Sharpness and identifiability

In this section, we explain that local sharp growth of a function f relative to a set S is equivalent to S being an "active manifold" for f locally around its minimizer. This equivalence is in essence well-known, though we have been unable to find a formal statement. To illustrate on a simple example, consider the function $f(x, y) = x^2 + |y|$ and the set $S = \mathbb{R} \times \{0\}$. Notice that f satisfies two geometric properties. On one hand, f grows sharply (at least linearly) as one moves away from S. On the other hand, S is "active" or "identifiable" in the sense that the subgradients of f are uniformly bounded away from zero outside of S. We will see that these two geometric properties are essentially equivalent. To formalize the notion of an "active set", we follow the work [17], which expands on the earlier papers of Lewis [38] and Wright [64].

Throughout, we use the standard definitions and notation of variational analysis, as set out in the monographs [42, 50, 57]. Namely, consider a function $f\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ and

a point x, with $f(x)$ finite. The *Fréchet subdifferential*, denoted $\partial f(x)$, consists of all vectors $v \in \mathbb{R}^d$ satisfying

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

The *limiting subdifferential*, denoted $\partial_L f(x)$, consists of all vectors $v \in \mathbb{R}^d$ for which there exist sequences $x_i \in \mathbb{R}^d$ and $v_i \in \partial f(x_i)$ satisfying $(x_i, f(x_i), v_i) \to (x, f(x), v)$. Following [51], we say that f is *prox-regular at $\bar{x}$ for $\bar{v} \in \partial_L f(\bar{x})$* if there exist real $\rho > 0$ such that the estimate

$$f(y) \geq f(x) + \langle v, y - x \rangle - \frac{\rho}{2}\|y - x\|^2,$$

holds for any $x, y \in \mathbb{R}^d$ and $v \in \partial_L f(x)$ satisfying $\max\{\|y - \bar{x}\|, \|x - \bar{x}\|, \|v - \bar{v}\|, |f(x) - f(\bar{x})|\} < \varepsilon$. In particular, weakly convex functions are prox-regular.

The following is the formal definition of an identifiable (or active) manifold.

Definition D.1 (Identifiable manifold) Consider a closed function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. We call a set S an *identifiable manifold at $\bar{x}$ for $\bar{v} \in \partial f(\bar{x})$* if the following properties hold.

1. **(smoothness)** The set S is a C^2-smooth manifold around $\bar{x}$ and the restriction f_S is C^2-smooth near $\bar{x}$.
2. **(finite identification)** For any sequences $(x_i, f(x_i), v_i) \to (\bar{x}, f(\bar{x}), \bar{v})$ with $v_i \in \partial_L f(x_i)$, the points x_i must all lie in S for all sufficiently large indices i.

Let us first observe that under a very mild condition on the function f, identifiability at a critical points implies local sharp growth.

Theorem D.2 (Identification implies sharpness) *Consider a closed function $f: \mathbb{R}^d \to \overline{\mathbb{R}}$ and suppose that a closed set S is an identifiable manifold at $\bar{x}$ for $0 \in \partial f(\bar{x})$. Then there exist real $\mu > 0$ satisfying*

$$f(x) \geq f(\mathrm{proj}_S(x)) + \mu \cdot \mathrm{dist}(x, S) \quad \text{for all } x \in B_\varepsilon(\bar{x}).$$

Proof First, we record an immediate consequence of [17, Proposition 10.12]. Namely, there exists $\varepsilon > 0$ satisfying the following. For all $z \in B_\varepsilon(\bar{x}) \cap S$ and $v \in \partial_L f(z) \cap B_\varepsilon(0)$, the inclusion holds:

$$v + \left(S^{d-1} \cap N_S(z)\right) \subset \partial f(z).$$

Next recall that since S is a C^2-smooth manifold, every point x near $\bar{x}$ admits a unique nearest-point projection onto S, characterized by the inclusion $x - \mathrm{proj}_S(x) \in N_S(\mathrm{proj}_S(x))$. For any point x near $\bar{x}$, set $\hat{x} = \mathrm{proj}_S(x)$. Using [17, Proposition 10.11], we deduce that f is prox-regular at $\bar{x}$ for $\bar{v} = 0$. Consequently, there exist $\varepsilon, \gamma, \rho > 0$ such that

$$f(x) \geq f(\hat{x}) + \langle v, x - \hat{x} \rangle - \frac{\rho}{2}\|x - \hat{x}\|^2 \tag{D.1}$$

for any $x \in B_\varepsilon(\bar{x})$ and $v \in B_\gamma(0) \cap \partial_L f(\hat{x})$. Notice that since the subdifferential of f is inner-semicontinuous relative to S at $\bar{x}$ for $\bar{v} = 0$ [17, Proposition 10.2], decreasing we may ensure that $B_\gamma(0) \cap \partial_L f(\hat{x})$ is nonempty for all $x \in B_\varepsilon(\bar{x})$. We therefore deduce for every $x \in B_\varepsilon(\bar{x})$ the estimate:

$$\begin{aligned} f(x) &\geq f(\hat{x}) + v + \frac{x-\hat{x}}{\|x-\hat{x}\|}, x-\hat{x} - \frac{\rho}{2}\|x-\hat{x}\|^2 \\ &\geq f(\hat{x}) + \cdot \operatorname{dist}_S(x) - \operatorname{dist}_S(x) - \frac{\rho}{2}\operatorname{dist}_S^2(x) \\ &= f(\hat{x}) + - \gamma - \frac{\rho}{2} \operatorname{dist}_S(x). \end{aligned}$$

Decreasing γ and , if necessary, completes the proof.

We next prove the converse, namely that a function always grows sharply away from its identifiable manifolds.

Theorem D.3 (Sharpness implies identification). *Consider a closed function $f: \mathbb{R}^d \to \overline{\mathbb{R}}$ that is prox-regular at a point $\bar{x}$ for $0 \in \partial_L f(\bar{x})$. Suppose that there is a closed set S containing $\bar{x}$ and real $\mu > 0$ satisfying*

$$f(x) \geq \min_{z \in \operatorname{proj}_S(x)} f(z) + \mu \cdot \operatorname{dist}(x, S) \quad \text{for all } x \in B_\varepsilon(\bar{x}).$$

Then for any sequences $(x_i, f(x_i), v_i) \to (\bar{x}, f(\bar{x}), \bar{v})$ with $v_i \in \partial_L f(x_i)$, the points x_i must all lie in S for all sufficiently large indices i.

Proof Let $\mu > 0$ be the constants in the assumptions of the theorem. From the definition of prox-regularity, we deduce that there exist real $\rho > 0$ such that the estimate

$$f(y) \geq f(x) + v, \ y - x - \frac{\rho}{2}\|y - x\|^2,$$

holds for any $x, y \in \mathbb{R}^d$ and $v \in \partial_L f(x)$ satisfying $\max\{\|x - \bar{x}\|, |f(x) - f(\bar{x})|\} < $. Shrinking , we may ensure $0 < < \min\{, \frac{\mu}{\rho}\}$. Consider now any point $x \in B_\varepsilon(\bar{x}) \setminus S$ with $|f(x) - f(\bar{x})| <$ and $\operatorname{dist}(0, \partial_L f(x)) <$. We will show that the estimate $\operatorname{dist}(0, \partial_L f(x)) \geq \frac{\mu}{2}$ holds, thereby completing the proof. To verify this estimate, let $v \in \partial f(x)$ have minimal norm and let $\hat{x} \in \operatorname{proj}_S(x)$ achieve $\min_{z \in \operatorname{proj}_S(x)} f(z)$. We then deduce

$$\mu \cdot \operatorname{dist}(x, S) \leq f(x) - f(\hat{x}) \leq v, \ x - \hat{x} + \frac{\rho}{2}\operatorname{dist}^2(x, S).$$

Using the Cauchy-Schwarz inequality, we therefore conclude $v \geq \frac{\mu}{2}$. The result follows.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M. et al.: Tensorflow: large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org 1(2) (2015)
2. Ahmed, A., Recht, B., Romberg, J.: Blind deconvolution using convex programming. IEEE Trans. Inf. Theory **60**(3), 1711–1732 (2014)
3. Asi, H., Duchi, J.C.: Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. arXiv preprint arXiv:1810.05633 (2018)
4. Asi, H., Duchi, J.C.: The importance of better models in stochastic optimization. arXiv preprint arXiv:1903.08619 (2019)
5. Aybat, N.S., Fallah, A., Gurbuzbalaban, M., Ozdaglar, A.: A universally optimal multistage accelerated stochastic gradient method. arXiv preprint arXiv:1901.08022 (2019)
6. Candès, E.J., Strohmer, T., Voroninski, V.: PhaseLift: exact and stable signal recovery from magnitude measurements via convex programming. Comm. Pure Appl. Math. **66**(8), 1241–1274 (2013)
7. Charisopoulos, V., Chen, Y., Davis, D., Díaz, M., Ding, L., Drusvyatskiy, D.: Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. arXiv preprint arXiv:1904.10020 (2019)
8. Charisopoulos, V., Davis, D., Díaz, M., Drusvyatskiy, D.: Composite optimization for robust blind deconvolution. arXiv:1901.01624 (2019)
9. COR-OPT. Geometric step decay: reference implementation. https://github.com/COR-OPT/GeomStepDecay (2019)
10. Davis, D.: SMART: The stochastic monotone aggregated root-finding algorithm. arXiv preprint arXiv:1601.00698 (2016)
11. Davis, D., Drusvyatskiy, D.: Stochastic model-based minimization of weakly convex functions. SIAM J. Optim. **29**(1), 207–239 (2019)
12. Davis, D., Drusvyatskiy, D., MacPhee, K.J., Paquette, C.: Subgradient methods for sharp weakly convex functions. J. Optim. Theory Appl. **179**(3), 962–982 (2018)
13. Davis, D., Drusvyatskiy, D., Paquette, C.: The nonsmooth landscape of phase retrieval. To appear in IMA J. Numer. Anal., arXiv:1711.03247 (2017)
14. Defazio, A., Bach, F., Lacoste-Julien, S.: SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 27, pp. 1646–1654. Curran Associates Inc, New York (2014)
15. Defazio, A., Domke, J., Caetano, T.S.: Finito: a faster, permutable incremental gradient method for big data problems. In: Xing, E.P., Jebara, T. (eds). Proceedings of the 31st International Conference on Machine Learning, volume 32 of Proceedings of Machine Learning Research, pp. 1125–1133. Bejing, China (2014). PMLR
16. Drusvyatskiy, D.: The proximal point method revisited. SIAG/OPT Views and News, 26(1) (2018)
17. Drusvyatskiy, D., Lewis, A.S.: Optimality, identifiablity, and sensitivity. arXiv preprint arXiv:1207.6628 (2012)
18. Duchi, J.C., Ruan, F.: Solving (most) of a set of quadratic equalities: composite optimization for robust phase retrieval. IMA J. Inf. Inference (2018). https://doi.org/10.1093/imaiai/iay015
19. Duchi, J.C., Ruan, F.: Stochastic methods for composite and weakly convex optimization problems. SIAM J. Optim. **28**(4), 3229–3259 (2018)
20. Eldar, Y.C., Mendelson, S.: Phase retrieval: stability and recovery guarantees. Appl. Comput. Harmon. Anal. **36**(3), 473–494 (2014)
21. Eremin, I.I.: The relaxation method of solving systems of inequalities with convex functions on the left-hand side. Dokl. Akad. Nauk SSSR **160**, 994–996 (1965)
22. Fercoq, O., Qu, Z.: Restarting accelerated gradient methods with a rough strong convexity estimate. arXiv preprint arXiv:1609.07358 (2016)
23. Fercoq, O., Qu, Z.: Adaptive restart of accelerated gradient methods under local quadratic growth condition. arXiv preprint arXiv:1709.02300 (2017)
24. Freund, R.M., Haihao, Lu.: New computational guarantees for solving convex optimization problems with first order methods, via a function growth condition measure. Math. Program. **170**(2), 445–477 (2018)

25. Ge, R., Kakade, S.M., Kidambi, R., Netrapalli, P.: The step decay schedule: a near optimal, geometrically decaying learning rate procedure. arXiv preprint arXiv:1904.12838 (2019)
26. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. SIAM J. Optim. **23**(4), 2061–2089 (2013)
27. Goffin, J.L.: On convergence rates of subgradient optimization methods. Math. Program. **13**(3), 329–347 (1977)
28. Goldstein, T., Studer, C.: Phasemax: convex phase retrieval via basis pursuit. IEEE Trans. Inf. Theory **64**(4), 2675–2689 (2018)
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
30. Hsu, D., Sabato, S.: Loss minimization and parameter estimation with heavy tails. J. Mach. Learn. Res. **17**(1), 543–582 (2016)
31. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS'13, pp. 315–323, Curran Associates Inc, USA (2013)
32. Johnstone, P.R., Moulin, P.R.: Faster subgradient methods for functions with hölderian growth. Math. Program. **180**, 417–450 (2019)
33. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
34. Kulunchakov, A., Mairal, J.: A generic acceleration framework for stochastic composite optimization. arXiv preprint arXiv:1906.01164 (2019)
35. Kushner, H.J., Yin, G.G.: Stochastic Approximation and Recursive Algorithms and Applications, volume 35 of Applications of Mathematics (New York). Stochastic Modelling and Applied Probability, 2nd edn. Springer-Verlag, New York (2003)
36. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 Nov (1998)
37. Lee, S., Wright, S.J.: Manifold identification in dual averaging for regularized stochastic online learning. J. Mach. Learn. Res. **13**(JUN), 1705–1744 (2012)
38. Lewis, A.S.: Active sets, nonsmoothness, and sensitivity. SIAM J. Optim. **13**(3), 702–725 (2002)
39. Li, X., Ling, S., Strohmer, T., Wei, K.: Rapid, robust, and reliable blind deconvolution via nonconvex optimization. Appl. Comput. Harmon. Anal. **47**(3), 893–934 (2018)
40. Li, Y., Ma, C., Chen, Y., Chi, Y.: Nonconvex matrix factorization from rank-one measurements. arXiv:1802.06286, 2018
41. Mairal, J.: Incremental majorization-minimization optimization with application to large-scale machine learning. SIAM J. Optim. **25**(2), 829–855 (2015)
42. Mordukhovich, B.S.: Variational Analysis and Generalized Differentiation I: Basic Theory. Grundlehren der mathematischen Wissenschaften, vol. 330. Springer, Berlin (2006)
43. Nedić, A., Bertsekas, D.: Convergence rate of incremental subgradient algorithms. In: Stochastic Optimization: Algorithms and Applications, pp. 223–264. Springer (2001)
44. Nemirovskii, A.S., Nesterov, Yu.E.: Optimal methods of smooth convex minimization. USSR Comput. Math. Math. Phys. **25**(2), 21–30 (1985)
45. Nemirovsky, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. A Wiley-Interscience Publication. John Wiley & Sons Inc, New York (1983). Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics
46. Nesterov, Y.: Primal-dual subgradient methods for convex problems. Math. Program. **120**(1), 221–259 (2009)
47. Nesterov, Y.: A method for solving the convex programming problem with convergence rate $\phi(1/k^2)$. Dokl. Akad. Nauk SSSR **269**(3), 543–547 (1983)
48. O'Donoghue, B., Candès, E.: Adaptive restart for accelerated gradient schemes. Found. Comput. Math. **15**(3), 715–732 (2015)
49. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS-W (2017)
50. Penot, J.-P.: Calculus Without Derivatives. Graduate Texts in Mathematics, vol. 266. Springer, New York (2013)
51. Poliquin, R.A., Rockafellar, R.T.: Prox-regular functions in variational analysis. Trans. Am. Math. Soc. **348**, 1805–1838 (1996)

52. Poljak, B.T.: Minimization of nonsmooth functionals. Ž. Vyčisl. Mat. i Mat. Fiz. **9**, 509–521 (1969)
53. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM J. Control Optim. **30**(4), 838–855 (1992)
54. Renegar, J., Grimmer, B.: A simple nearly-optimal restart scheme for speeding-up first order methods. arXiv preprint arXiv:1803.00151 (2018)
55. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**, 400–407 (1951)
56. Rockafellar, R.T.: Favorable classes of Lipschitz-continuous functions in subgradient optimization. In: Progress in Nondifferentiable Optimization, Volume 8 of IIASA Collaborative Proc. Ser. CP-82, pp. 125–143. Int. Inst. Appl. Sys. Anal., Laxenburg (1982)
57. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Grundlehren der mathematischen Wissenschaften, vol. 317. Springer, Berlin (1998)
58. Roulet, V., d'Aspremont, A.: Sharpness, restart and acceleration. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 1119–1129. Curran Associates Inc, New York (2017)
59. Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Math. Program. **162**(1–2), 83–112 (2017)
60. Shalev-Shwartz, S., Zhang, T.: Proximal stochastic dual coordinate ascent. arXiv:1211.2717 (2012)
61. Shechtman, Y., Eldar, Y.C., Cohen, O., Chapman, H.N., Miao, J., Segev, M.: Phase retrieval with application to optical imaging: a contemporary overview. IEEE Signal Process. Mag. **32**(3), 87–109 (2015)
62. Shor, N.Z.: Minimization Methods for Non-differentiable Functions, vol. 3. Springer Science & Business Media, New York (2012)
63. Tan, Y.S., Vershynin, R.: Phase retrieval via randomized Kaczmarz: Theoretical guarantees. Inf. Inference J. IMA **8**(1), 97–123 (2018)
64. Wright, S.J.: Identifiable surfaces in constrained optimization. SIAM J. Control Optim. **31**(4), 1063–1079 (1993)
65. Xiao, L.: Dual averaging methods for regularized stochastic learning and online optimization. J. Mach. Learn. Res. **11**(OCT), 2543–2596 (2010)
66. Xu, Y., Lin, Q., Yang, T.: Accelerated stochastic subgradient methods under local error bound condition. arXiv preprint arXiv:1607.01027 (2016)
67. Yang, T., Lin, Q.: RSG: beating subgradient method without smoothness and strong convexity. J. Mach. Learn. Res. **19**(1), 236–268 (2018)
68. Yang, T., Yan, Y., Yuan, Z., Jin, R.: Why does stagewise training accelerate convergence of testing error over SGD? arXiv preprint arXiv:1812.03934 (2018)