STOCHASTIC MODEL-BASED MINIMIZATION OF WEAKLY CONVEX FUNCTIONS*

DAMEK DAVIS† AND DMITRIY DRUSVYATSKIY‡

Abstract. We consider a family of algorithms that successively sample and minimize simple stochastic models of the objective function. We show that under reasonable conditions on approximation quality and regularity of the models, any such algorithm drives a natural stationarity measure to zero at the rate $O(k^{-1/4})$. As a consequence, we obtain the first complexity guarantees for the stochastic proximal point, proximal subgradient, and regularized Gauss-Newton methods for minimizing compositions of convex functions with smooth maps. The guiding principle, underlying the complexity guarantees, is that all algorithms under consideration can be interpreted as approximate descent methods on an implicit smoothing of the problem, given by the Moreau envelope. Specializing to classical circumstances, we obtain the long-sought convergence rate of the stochastic projected gradient method, without batching, for minimizing a smooth function on a closed convex set.

Key words. stochastic, subgradient, proximal, prox-linear, Moreau envelope, weakly convex

AMS subject classifications. 65K05, 65K10, 90C15, 90C30

1. Introduction.

Stochastic optimization plays a central role in the statistical sciences, underlying all aspects of learning from data. The goal of stochastic optimization in data science is to learn a decision rule from a limited data sample, which generalizes well to the entire population. Learning such a decision rule amounts to minimizing the *regularized population risk*:

$$\text{(SO)} \qquad \min_{x\in\mathbb{R}^d} \ \varphi(x) = f(x) + r(x) \qquad \text{where} \qquad f(x) = \mathbb{E}_{\xi\sim P}[f(x,\xi)].$$

Here, ξ encodes the population data, which is assumed to follow some fixed but unknown probability distribution P. The functions f and r play qualitatively different roles. Typically, $f(x,\xi)$ evaluates the loss of the decision rule parametrized by x on a data point ξ. In contrast, the function $r\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ models constraints on the parameters x or encourages x to have some low dimensional structure, such as sparsity or low rank. Within a Bayesian framework, the regularizer r can model prior distributional information on x.

Robbins-Monro's pioneering 1951 work [54] gave the first procedure for solving (SO) in the setting when $f(\cdot,\xi)$ are smooth and strongly convex and $r = 0$, thereby inspiring decades of further research. Among such algorithms, the proximal stochastic (sub)gradient method is the most successful and widely used in practice. This method constructs a sequence of approximations x_t of the minimizer of (SO) by iterating:

$$\text{(SG)} \qquad \left\{ \begin{array}{l} \text{Sample } \xi_t \sim P \\ \text{Set } x_{t+1} = \text{prox}_{\alpha_t r}\left(x_t - \alpha_t \nabla_x f(x_t,\xi_t)\right) \end{array} \right\},$$

*Submitted to the editors March 29, 2018. The results in this paper are a combination of the two arXiv preprints [21, 22].

Funding: Research of Drusvyatskiy was supported by the AFOSR YIP award FA9550-15-1-0237 and by the NSF DMS 1651851 and CCF 1740551 awards.

†School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850, USA (dsd95@cornell.edu, people.orie.cornell.edu/dsd95).

‡Department of Mathematics, University of Washington, Seattle, WA 98195 (ddrusv@uw.edu, www.math.washington.edu/~ddrusv).

where $\alpha_t > 0$ is an appropriate control sequence and $\mathrm{prox}_{\alpha r}(\cdot)$ is the proximal map

$$\mathrm{prox}_{\alpha r}(x) := \operatorname*{argmin}_y \left\{ r(y) + \tfrac{1}{2\alpha}\|y - x\|^2 \right\}.$$

Thus, in each iteration, the method travels from x_t in the direction opposite to a sampled gradient $\nabla_x f(x_t, \xi_t)$, followed by a proximal operation.

Nonsmooth convex problems may be similarly optimized by replacing sample gradients by sample subgradients $v_t \in \partial_x f(x_t, \xi_t)$, where $\partial_x f(x, \xi)$ is the subdifferential in the sense of convex analysis [59]. Even more broadly, when $f(\cdot, \xi)$ is neither smooth nor convex, the symbol $\partial_x f(\cdot, \xi)$ may refer to a generalized subdifferential. The formal definition of the subdifferential appears in Section 2, and is standard in the optimization literature (e.g. [58, Definition 8.3]). The reader should keep in mind that in practice, the functions $f(\cdot, \xi_t)$ are often all differentiable along the iterate sequence $\{x_t\}$. Therefore from the viewpoint of implementation, one always computes the true gradients of the sampled functions, using conventional means. On the other hand, the nonsmoothness cannot be ignored in the analysis, since (i) the gradients do not vary continuously and (ii) the objective function can be nonsmooth at every limit point of the process. We will expand on these two observations shortly.

Performance of stochastic optimization methods is best judged by their *sample complexity* – the number of i.i.d. realizations $\xi_1, \ldots, \xi_N \sim P$ needed to reach a desired accuracy of the decision rule. Classical results [45] stipulate that for convex problems, it suffices to generate $O(\varepsilon^{-2})$ samples to reach functional accuracy ε in expectation, and this complexity is unimprovable without making stronger assumptions. For smooth problems, the stochastic gradient method has sample complexity of $O(\varepsilon^{-4})$ to reach a point with the gradient norm at most ε in expectation [35, 36, 69].

Despite the ubiquity of the stochastic subgradient method in applications, its sample complexity is not yet known for any reasonably wide class of problems beyond those that are smooth or convex. This is somewhat concerning as the stochastic subgradient method is the simplest and most widely-used optimization algorithm for large-scale problems arising in machine learning and is the core optimization subroutine in industry backed software libraries, such as Google's TensorFlow [1].

The purpose of this work is to provide the first sample complexity bounds for a number of popular stochastic algorithms on a reasonably broad class of nonsmooth and nonconvex optimization problems. The problem class we consider captures a variety of important computational tasks in data science, as we illustrate below, while the algorithms we analyze include the proximal stochastic subgradient, proximal point, and regularized Gauss-Newton methods. Before stating the complexity guarantees, we must first explain the "stationarity measure" that we will use to judge the quality of the iterates. It is this stationarity measure that tends to zero at a controlled rate.

The search for stationary points. Convex optimization algorithms are judged by the rate at which they decrease the function value along the iterate sequence. Analysis of smooth optimization algorithms focuses instead on the magnitude of the gradients along the iterates. The situation becomes quite different for problems that are neither smooth nor convex.

As in the smooth setting, the primary goal of nonsmooth nonconvex optimization is the search for stationary points. A point $x \in \mathbb{R}^d$ is called *stationary* for the problem (SO) if the inclusion $0 \in \partial\varphi(x)$ holds. In "primal terms", these are precisely the points where the directional derivative of φ is nonnegative in every direction. Indeed, under

mild conditions on φ, equality holds [58, Proposition 8.32]:

$$\operatorname{dist}(0; \partial\varphi(x)) = -\inf_{v:\, \|v\| \leq 1} \varphi'(x; v).$$

Thus a point x, satisfying $\operatorname{dist}(0; \partial\varphi(x)) \leq \varepsilon$, approximately satisfies first-order necessary conditions for optimality.

An immediate difficulty in analyzing stochastic methods for nonsmooth and nonconvex problems is that it is not a priori clear how to measure the progress of the algorithm. Neither the functional suboptimality gap, $\varphi(x_t) - \min \varphi$, nor the stationarity measure, $\operatorname{dist}(0; \partial\varphi(x_t))$, necessarily tend to zero along the iterate sequence. This difficulty persists even in the simplest setting of minimizing a smooth function on a closed convex set by the stochastic projected gradient method. Indeed, what is missing is a continuous measure of stationarity to monitor, instead of the highly discontinuous function $x \mapsto \operatorname{dist}(0; \partial\varphi(x))$.

Weak convexity and the Moreau envelope. In this work, we focus on a class of problems that naturally admit a continuous measure of stationarity. We say that a function $g\colon \mathbb{R}^d \to \mathbb{R}$ is *ρ-weakly convex* if the assignment $x \mapsto g(x) + \frac{\rho}{2}\|x\|^2$ is convex. The class of weakly convex functions, first introduced in English in [50], is broad. It includes all convex functions and smooth functions with Lipschitz continuous gradient. More generally, any function of the form

$$g(x) = h(c(x)),$$

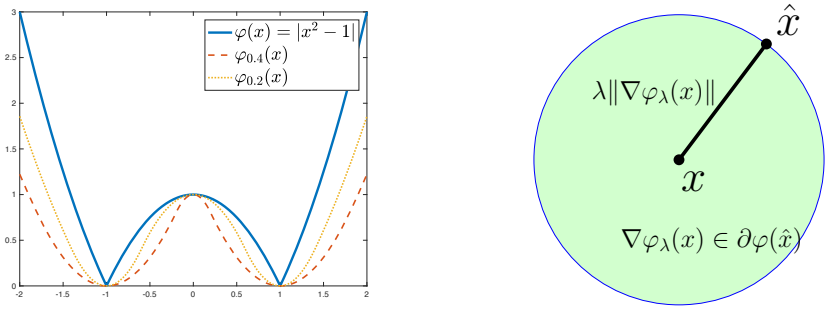
with h convex and Lipschitz and c a smooth map with Lipschitz Jacobian, is weakly convex [29, Lemma 4.2]. Notice that such composite functions need not be smooth nor convex; instead, the composite function class nicely interpolates between the smooth and convex settings. Classical literature highlights the importance of weak convexity in optimization [51, 52, 56], while recent advances in statistical learning and signal processing have further reinvigorated the problem class. Nonlinear least squares, phase retrieval [23, 30, 34], minimization of the Conditional Value-at-Risk [7, 8, 60], graph synchronization [2, 6, 63], covariance estimation [17], and robust principal component analysis [11, 14] directly lead to weakly convex formulations. For a recent discussion on the role of weak convexity in large-scale optimization, see e.g., [26].

It has been known since Nurminskii's work [49, 50] that when the functions $f(\cdot, \xi)$ are ρ-weakly convex and $r = 0$, the stochastic subgradient method on (SO) generates an iterate sequence that subsequentially converges to a stationary point of the problem, almost surely. Nonetheless, the sample complexity of the basic method and of its proximal extension, has remained elusive. Our approach to resolving this open question relies on an elementary observation: weakly convex problems naturally admit a continuous measure of stationarity through implicit smoothing. The key construction we use is the *Moreau envelope* [43]:

$$\varphi_\lambda(x) := \min_y \left\{ \varphi(y) + \tfrac{1}{2\lambda}\|y - x\|^2 \right\},$$

where $\lambda > 0$. Standard results (e.g. [43], [55, Theorem 31.5]) show that as long as φ is ρ-weakly convex and $\lambda < \rho^{-1}$, the envelope φ_λ is C^1-smooth with the gradient given by

$$\nabla\varphi_\lambda(x) = \lambda^{-1}(x - \operatorname{prox}_{\lambda\varphi}(x)).$$



(a) Moreau envelope of $\varphi(x) = |x^2 - 1|$

(b) Approximate stationarity

Fig. 1: An illustration of the Moreau envelope

See Figure 1a for an illustration.

When f is C^1-smooth with β-Lipschitz gradient and there is no regularizer r, the norm $\|\nabla\varphi_{1/\beta}(x)\|$ is proportional to the magnitude of the true gradient $\|\nabla f(x)\|$. More generally, when f is C^1-smooth and r is nonzero, the norm $\|\nabla\varphi_{1/\beta}(x)\|$ is proportional to the size of the proximal gradient step, commonly used to measure convergence in additive composite minimization [47]. See the end of Section 2.2 for a precise statement. In the broader nonsmooth setting, the norm of the gradient $\|\nabla\varphi_\lambda(x)\|$ has an intuitive interpretation in terms of near-stationarity for the target problem $\min_x \varphi(x)$. Namely, the definition of the Moreau envelope directly implies that for any point $x \in \mathbb{R}^d$, the proximal point $\hat{x} := \mathrm{prox}_{\lambda\varphi}(x)$ satisfies

$$\begin{cases} \|\hat{x} - x\| & = \lambda\|\nabla\varphi_\lambda(x)\|, \\ \varphi(\hat{x}) & \leq \varphi(x), \\ \mathrm{dist}(0; \partial\varphi(\hat{x})) & \leq \|\nabla\varphi_\lambda(x)\|. \end{cases}$$

Thus a small gradient $\|\nabla\varphi_\lambda(x)\|$ implies that x is *near* some point $\hat{x}$ that is *nearly stationary* for φ; see Figure 1b. In the language of numerical analysis, one can interpret algorithms that drive the gradient of the Moreau envelope to zero as being "backward-stable". For a longer discussion of the near-stationarity concept, we refer to reader to [26] or [29, Section 4.1].

Contributions. In this paper, we show that as long as the functions $f(\cdot, \xi) + r(\cdot)$ are ρ-weakly convex and mild Lipschitz conditions hold, the proximal stochastic subgradient method will generate a point x satisfying $\mathbb{E}\|\nabla\varphi_{1/(2\rho)}(x)\| \leq \varepsilon$ after at most $O(\varepsilon^{-4})$ iterations. This is perhaps surprising, since neither the Moreau envelope nor the proximal map of φ explicitly appear in the definition of the stochastic proximal subgradient method. This work appears to be the first to recognize the Moreau envelope as a useful potential function for analyzing subgradient methods.

Indeed, we will show that the worst-case complexity $O(\varepsilon^{-4})$ holds for a much wider family of algorithms than the stochastic subgradient method. Setting the stage, recall that the stochastic subgradient method relies on sampling subgradient estimates of f, or equivalently sampling good linear models of the function. More broadly, suppose that f is an arbitrary function (not necessarily written as an expectation), and for every point x we have available a family of "models" $\{f_x(\cdot, \xi)\}_{\xi \sim P}$, indexed by a random element $\xi \sim P$. The oracle concept we use assumes that the only access to f

is by sampling a model $f_x(\cdot, \xi)$ centered around any base point x. Naturally, to make use of such models we must have some control on their approximation quality. We will call the assignment $(x, y, \xi) \mapsto f_x(y, \xi)$ a *stochastic one-sided model* if it satisfies

$$\mathbb{E}_\xi[f_x(x, \xi)] = f(x) \qquad \text{and} \qquad \mathbb{E}_\xi[f_x(y, \xi) - f(y)] \leq \frac{\tau}{2}\|y - x\|_2^2 \qquad \forall x, y, \tag{1.1}$$

Thus in each expectation, each model $f_x(\cdot, \xi)$ should lower bound f up to a quadratic error, while agreeing with f at the basepoint x. See Figure 2 for an illustration.

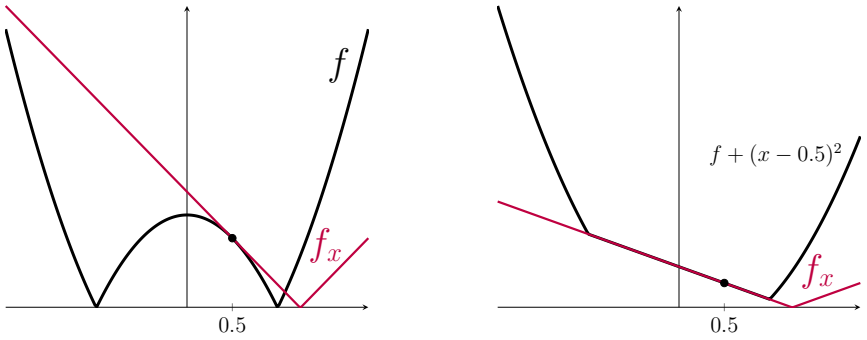


Fig. 2: Illustration of a one-sided model: $f(x) = |x^2 - 1|$, $f_{0.5}(y) = |1.25 - y|$

The methods we consider then simply iterate the steps:

$$\begin{aligned} &\text{Sample } \xi_t \sim P, \\ &\text{Set } x_{t+1} = \operatorname*{argmin}_y \left\{ f_{x_t}(y, \xi_t) + r(y) + \frac{1}{2\alpha_t}\|y - x_t\|^2 \right\}. \end{aligned} \tag{1.2}$$

We will prove that under mild Lipschitz conditions and provided that each function $f_x(\cdot, \xi) + r(\cdot)$ is ρ-weakly convex, Algorithm 1.2 finds a point x with $\mathbb{E}\|\nabla \varphi_{1/2\rho}(x)\| \leq \varepsilon$ after at most $O(\varepsilon^{-4})$ iterations. The main principle underlying the convergence guarantees is interesting in its own right. We will show that Algorithm 1.2 can be interpreted as an approximate descent method on the Moreau envelope:

$$\mathbb{E}[\varphi_\lambda(x_{t+1})] \leq \mathbb{E}[\varphi_\lambda(x_t)] - \alpha_t c_1 \mathbb{E}[\|\nabla \varphi_\lambda(x_t)\|^2] + \alpha_t^2 c_2, \tag{1.3}$$

where λ, c_1, c_2 are problem dependent constants.

When the models $f_x(\cdot, \xi)$ are true under-estimators of f in expectation, meaning that (1.1) holds with $\tau = 0$, and the functions $f_x(\cdot, \xi) + r(\cdot)$ are convex, one expects guarantees that are analogous to the stochastic subgradient method for convex minimization. Indeed, we will show that under these circumstances, Algorithm (1.2) has complexity $O(\varepsilon^{-2})$ in terms of function value. The complexity estimate improves to $O(\frac{1}{\mu\varepsilon})$ when the functions $f_x(\cdot, \xi) + r(\cdot)$ are μ-strongly convex. Though the convexity assumption may appear stringent, it does hold in a number of nonclassical circumstances, such as for minimizing the Condition Value-at Risk (cVaR) of a loss function; see Example 2.6 and Section 4.2 for details.

To crystallize the ideas, consider the setting of stochastic composite optimization, studied recently by Duchi-Ruan [31]:

$$f(x, \xi) = h\big(c(x, \xi), \xi\big),$$

where the functions $h(\cdot,\xi)$ are convex and the maps $c(\cdot,\xi)$ are smooth. Note that in the simplest setting when P is a discrete distribution on $\{1,\dots,m\}$, the problem (SO) reduces to minimizing a regularized empirical average of composite functions:

$$\min_{x\in\mathbb{R}^d} \ \varphi(x) = f(x) + r(x) \qquad \text{where} \qquad f(x) = \frac{1}{m}\sum_{i=1}^{m} h_i(c_i(x))$$

The following three stochastic one-sided models appear naturally:

$$f_x(y,\xi) = f(x) + \langle \nabla c(x,\xi)^T w(x,\xi), y - x\rangle, \tag{1.4}$$

$$f_x(y,\xi) = h\big(c(x,\xi) + \nabla c(x,\xi)(y-x),\xi\big), \tag{1.5}$$

$$f_x(y,\xi) = h\big(c(y,\xi)\big), \tag{1.6}$$

where $w(x,\xi) \in \partial h(c(x,\xi),\xi)$ is a subgradient selection. Each iteration of Algorithm 1.2 with the models (1.4) reduces to the *stochastic proximal subgradient* update, already mentioned previously. When equipped with the models (1.5), the method becomes the *stochastic prox-linear algorithm* — a close variant of Gauss-Newton. Both of these schemes were recently investigated in [31], where the authors showed that almost surely all limit points are stationary for the problem (SO). Algorithm 1.2 equipped with the models (1.6) is the *stochastic proximal-point algorithm*. This scheme was recently considered for convex minimization in [61, 66, 67] and extended to monotone inclusions in [9]. Notice that in contrast to the stochastic proximal subgradient method, the stochastic proximal point and prox-linear algorithms require solving an auxiliary subproblem. The advantage of these two schemes is that the models (1.5) and (1.6) provide much finer approximation quality, in that they are two-sided instead of one-sided. Indeed, empirical evidence [31, Section 4] suggests that the latter two algorithms can perform significantly better and are much more robust to the choice of the sequence α_t. We also observe this phenomenon in our experiments in Section 5.

The outline of the paper is as follows. We begin with Section 2, which records some basic notation and results focusing on weak convexity and the Moreau envelope. This section also presents a number of illustrative applications that will be readily amenable to our algorithmic techniques. We then present three distinct convergence arguments: for the stochastic projected subgradient method in Section 3.1, for the stochastic proximal subgradient method in Section 3.2, and for algorithms based on general stochastic one-sided models in Section 4. Each argument has its own virtue. In particular, our guarantees for the stochastic projected subgradient method place no restriction on the parameters α_t to be used, in contrast to our latter results. The argument for the stochastic proximal subgradient method generalizes verbatim to the setting when f is C^1-smooth and the stochastic gradient estimator has bounded variance, instead of a bounded second moment that we assume elsewhere. Section 4 applies to the most general classes of algorithms including stochastic proximal subgradient, prox-linear, and proximal point methods.

Context and related literature. The convergence guarantees we develop for the proximal stochastic subgradient method are new even in simplified cases. Two such settings are (i) when $f(\cdot,\xi)$ are smooth and r is the indicator function of a closed convex set, and (ii) when f is nonsmooth, we have explicit access to the exact subgradients of f, and $r = 0$.

Analogous convergence guarantees when r is an indicator function of a closed convex set were recently established for a different algorithm in [24]. This scheme

proceeds by directly applying the gradient descent method to the Moreau envelope φ_λ, with each proximal point $\mathrm{prox}_{\lambda\varphi}(x)$ approximately evaluated by a convex subgradient method. In contrast, we show here that the basic stochastic subgradient method in the fully proximal setting, and without any modification or parameter tuning, already satisfies the desired convergence guarantees.

Our work also improves in two fundamental ways on the results in the seminal papers on the stochastic proximal gradient method for smooth functions [35, 36, 69]: first, we allow $f(\cdot, \xi)$ to be nonsmooth and second, even when $f(\cdot, \xi)$ are smooth, we do not require the variance of our stochastic estimator for $\nabla f(x_t)$ to decrease as a function of t. The second contribution removes the well-known "mini-batching" requirements common to [36, 69], while the first significantly expands the class of functions for which the rate of convergence of the stochastic proximal subgradient method is known. It is worthwhile to mention that our techniques rely on weak convexity of the regularizer r, while [69] makes no such assumption.

The results in this paper are orthogonal to the recent line of work on accelerated rates of convergence for smooth nonconvex finite sum minimization problems, e.g., [3, 4, 38, 53]. These works crucially exploit the finite sum structure and/or (higher order) smoothness of the objective functions to push beyond the $O(\varepsilon^{-4})$ complexity. We leave it as an intriguing open question whether such improvement is possible for the nonsmooth weakly convex setting we consider here.

The unifying concept of stochastic one-sided models has not been explicitly used before. The complexity guarantees for the proximal stochastic subgradient, prox-linear, and proximal point methods (Theorem 4.3) for stochastic composite minimization are new and nicely complement the recent paper [31]. There, the authors proved that almost surely all limit points of the first two methods are stationary. For a historical account of the prox-linear method, see e.g., [10, 29, 40] and the references therein. For a systematic study of two-sided models (e.g. (1.5) and (1.6)) in optimization, see [27]. Stochastic compositional problems have also appeared in a parallel line of work beginning with [68]. There, the authors require the entire composite function to be either convex or smooth. We make no such assumptions here.

The convergence rate of Algorithm 1.2 in terms of function values in the convex setting is presented in Theorems 4.1, 4.2, and is intriguing. Even specializing to the proximal stochastic subgradient method, Theorems 4.1 and 4.2 appear to be stronger than the state of the art. Namely, in contrast to previous work [19, 32], the norms of the subgradients of r do not enter the complexity bounds established in Theorem 4.1, while Theorem 4.2 extends the nonuniform averaging technique of [62] for strongly convex minimization to the fully proximal setting.

The observation that Algorithm 1.2 is an approximate descent method on the Moreau envelope (1.3) is tangentially related to the recent work on "inexact first-order oracles" in convex optimization [25, 48] and its partial extensions to nonconvex settings [33]. Expanding on the precise relationship between the techniques is an interesting open question.

2. Basic notation and preliminaries. Throughout, we consider a Euclidean space $\mathbb{R}^d$ endowed with an inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\| = \sqrt{\langle x, x \rangle}$. For any function $\varphi \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, the *domain* and *epigraph* are the sets

$$\operatorname{dom} \varphi = \{x \in \mathbb{R}^d : \varphi(x) < \infty\}, \qquad \operatorname{epi} \varphi = \{(x, r) \in \mathbb{R}^d \times \mathbb{R} : r \geq \varphi(x)\},$$

respectively. We say that $\varphi \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is *closed* if the $\operatorname{epi} \varphi$ is a closed set.

This work focuses on algorithms for minimizing weakly convex functions.[1] A function $\varphi\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is called *ρ-weakly convex* if the assignment $x \mapsto \varphi(x) + \frac{\rho}{2}\|x\|^2$ is a convex function. In this section, we summarize some basic properties of this function class. All results we state in this section are either standard, or follow quickly from analogous results for convex functions. For further details and a historical account, we refer the reader to the short note [26].

2.1. Examples of weakly convex functions. Weakly convex functions are widespread in applications and are typically easy to recognize. One common source is the composite function class:

$$\varphi(x) := h(c(x)), \tag{2.1}$$

where $h\colon \mathbb{R}^m \to \mathbb{R}$ is convex and L-Lipschitz and $c\colon \mathbb{R}^d \to \mathbb{R}^m$ is a C^1-smooth map with β-Lipschitz continuous Jacobian. An easy argument shows that the composite function φ is $L\beta$-weakly convex [29, Lemma 4.2]. Below, we list a few examples to illustrate how widespread this problem class is in large-scale data scientific applications. The examples are here only to set the context; the reader can safely skip this discussion during the initial reading.

EXAMPLE 2.1 (Robust phase retrieval). Phase retrieval is a common computational problem, with applications in diverse areas such as imaging, X-ray crystallography, and speech processing. For simplicity, we will consider the version of the problem over the reals. The (real-valued) phase retrieval problem seeks to determine a point x satisfying the magnitude conditions,

$$|\langle a_i, x\rangle| \approx b_i \quad \text{for } i = 1, \dots, m,$$

where $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$ are given. Whenever gross outliers occur in the measurements b_i, the following robust formulation of the problem is appealing [23, 30, 34]:

$$\min_x \ \frac{1}{m}\sum_{i=1}^m |\langle a_i, x\rangle^2 - b_i^2|.$$

The use of the ℓ_1 penalty promotes strong recovery and stability properties even in the noiseless setting [30, 34]. Numerous other nonconvex approaches to phase retrieval exist, which rely on different problem formulations; for example, [12, 15, 64].

EXAMPLE 2.2 (Covariance matrix estimation). The problem of covariance estimation from quadratic measurements, introduced in [16], is a higher rank variant of phase retrieval. Let $a_1, \dots, a_m \in \mathbb{R}^d$ be measurement vectors. The goal is to recover a low rank decomposition of a covariance matrix $\bar{X}\bar{X}^T$, with $\bar{X} \in \mathbb{R}^{d\times r}$ for a given $0 \le r \le d$, from quadratic measurements

$$b_i \approx a_i^T \bar{X}\bar{X}^T a_i = \mathrm{Tr}(\bar{X}\bar{X}^T a_i a_i^T).$$

Note that we can only recover $\bar{X}$ up to multiplication by an orthogonal matrix. This problem arises in a variety of contexts, such as covariance sketching for data streams and spectrum estimation of stochastic processes. We refer the reader to [16] for details. Supposing that m is even, the authors of [16] show that the following potential function has strong recovery guarantees under usual statistical assumptions:

$$\min_{X\in\mathbb{R}^{d\times r}} \ \frac{1}{m}\sum_{i=1}^m \left|\left\langle XX^T, a_{2i}a_{2i}^T - a_{2i-1}a_{2i-1}^T\right\rangle - (b_{2i} - b_{2i-1})\right|. \tag{2.2}$$

[1] To the best of our knowledge, the class of weakly convex functions was introduced in [50].

EXAMPLE 2.3 (Blind deconvolution and biconvex compressive sensing).
The problem of blind deconvolution seeks to recover a pair of vectors in two low-dimensional structured spaces from their pairwise convolution. This problem occurs in a number of fields, such as astronomy and computer vision [13, 39]. For simplicity focusing on the real-valued case, one appealing formulation of the problem reads

$$\min_{x,y} \ \frac{1}{m}\sum_{i=1}^{m} |\langle u_i, x\rangle\langle v_i, y\rangle - b_i|,$$

where u_i and v_i are known vectors, and b_i are the convolution measurements. More broadly, problems of this form fall within the area of biconvex compressive sensing [41]. Similarly to the previous two examples, the use of the ℓ_1-penalty on the residuals yields strong recovery and stability guarantees under statistical assumptions. Details will appear in a forthcoming paper.

EXAMPLE 2.4 (Sparse dictionary learning). The problem of sparse dictionary learning seeks to find a sparse representation of the input data as a linear combination of basic atoms, which comprise the "dictionary". This technique is routinely used in image and video processing. More formally, given a set of vectors $\{x_1, \ldots, x_m\} \subset \mathbb{R}^d$, we wish to find a matrix $D \in \mathbb{R}^{d\times n}$ and sparse weights $\{r_1, \ldots, r_m\} \subset \mathbb{R}^n$ such that the error $\|x_i - Dr_i\|_2$ is small for all i. The following is a robust variant of the standard relaxation of the problem:

$$\min_{D\in\mathbb{R}^{d\times n},\ r_1\in\mathbb{R}^n} \ \frac{1}{m}\sum_{i=1}^{m} \|x_i - Dr_i\|_2 + \lambda\|r_i\|_1 \qquad \text{subject to} \quad \|D_i\| \le 1 \ \forall i. \tag{2.3}$$

More precisely, typical formulations use the squared norm $\|\cdot\|_2^2$ instead of the norm $\|\cdot\|_2$; see e.g. [42, 65]. When there are outliers in the data (i.e. not all of the data vectors x_i can be sparsely represented), the formulation (2.3) may be more appealing.

EXAMPLE 2.5 (Robust PCA). In robust principal component analysis, one seeks to identify sparse corruptions of a low-rank matrix [11, 14]. One typical example is image deconvolution, where the low-rank structure models the background of an image while the sparse corruption models the foreground. Formally, given a $m \times n$ matrix M, the goal is to find a decomposition $M = L + S$, where L is low rank and S is sparse. A common relaxation of the problem is

$$\min_{U\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{n\times r}} \ \|UV^T - M\|_1,$$

where r is the target rank. As is common, the entrywise ℓ_1 norm encourages a sparse residual $UV^T - M$.

EXAMPLE 2.6 (Conditional Value-at-Risk). As in the introduction, let $f(x, \xi)$ be a loss of a decision rule parametrized by x on a data point ξ, where the population data follows a probability distribution $\xi \sim P$. Rather than minimizing the expectation $f(x) = \mathbb{E}_{\xi\sim P} f(x, \xi)$, one often wishes to minimize the conditional expectation of the random variable $f(x, \cdot)$ over its α-tail, for some fixed $\alpha \in (0, 1)$. This quantity is called the Conditional Value-at-Risk (cVaR) and it has a distinguished history. In particular, it is well known from the seminal work [60] that minimizing cVaR of the

loss function can be formalized as[2]

$$\min_{\gamma\in\mathbb{R},x\in\mathbb{R}^d} \ (1-\alpha)\gamma + \mathbb{E}_{\xi\sim P}[(f(x,\xi)-\gamma)^+],$$

where we use the notation $r^+ = \max\{0, r\}$. If the loss function $f(\cdot,\xi)$ is ρ-weakly convex for a.e. ξ, then the entire objective function is ρ-weakly convex jointly in (γ, x). In particular, this is the case when $f(\cdot,\xi)$ is C^1-smooth with Lipschitz gradient, or when the loss is $f(\cdot,\xi)$ is convex for a.e. ξ. Notice that the terms inside the expectation $(f(\cdot,\xi)-\gamma)^+$ are always nonsmooth, even if the loss function $f(\cdot,\xi)$ is smooth.

2.2. Subdifferential and the Moreau envelope. A key property of convex functions is that any subgradient yields a global affine under-estimator of the function. It is this availability of global under-estimators that enables convergence guarantees for nonsmooth convex optimization. An analogous property is true for weakly convex functions, where the subdifferential is meant in a broader variational analytic sense and the affine under-estimators are replaced by concave quadratic under-estimators. We now formalize this observation.

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

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

We set $\partial\varphi(x)=\emptyset$ for all $x\notin \operatorname{dom}\varphi$. When φ is C^1-smooth, the subdifferential $\partial\varphi(x)$ consists only of the gradient $\{\nabla\varphi(x)\}$, while for convex functions it reduces to the subdifferential in the sense of convex analysis. The following characterization of weak convexity is standard; we provide a short proof for completeness.

LEMMA 2.1 (Subdifferential characterization).
The following are equivalent for any lower-semicontinuous function $\varphi\colon \mathbb{R}^d\to\mathbb{R}\cup\{\infty\}$.

1. *The function f is ρ-weakly convex.*
2. *The approximate secant inequality holds:*

$$\varphi(\lambda x+(1-\lambda)y) \leq \lambda\varphi(x)+(1-\lambda)\varphi(y)+\tfrac{\rho\lambda(1-\lambda)}{2}\|x-y\|^2, \tag{2.4}$$

   for all $x,y\in\mathbb{R}^d$ and $\lambda\in[0,1]$.
3. *The subgradient inequality holds:*

$$\varphi(y)\geq\varphi(x)+\langle v,y-x\rangle-\frac{\rho}{2}\|y-x\|^2, \qquad \forall x,y\in\mathbb{R}^d,\ v\in\partial\varphi(x). \tag{2.5}$$

4. *The subdifferential map is hypomontone:*

$$\langle v-w, x-y\rangle \geq -\rho\|x-y\|^2,$$

   for all $x,y\in\mathbb{R}^d$, $v\in\partial\varphi(x)$, and $w\in\partial\varphi(y)$.

If φ is C^2-smooth, then the four properties above are all equivalent to

$$\nabla^2\varphi(x)\succeq -\rho I \qquad\qquad \forall x\in\mathbb{R}^d.$$

Proof. Algebraic manipulation shows that the usual secant inequality on the function $\varphi+\frac{\rho}{2}\|\cdot\|^2$ is precisely the approximate secant inequality (2.4) on φ. Therefore

[2]We refer the reader to [57, pp. 44] and [8] for a historical account of the cVaR minimization formula, and in particular its interpretation as the "optimized certainty equivalent" introduced in [7].

we deduce the equivalence 1 $\Leftrightarrow$ 2. Suppose now 1 holds and define the function $g(x) = \varphi(x) + \frac{\rho}{2}\|x\|^2$. Note the equality $\partial g(x) = \partial\varphi(x) + \rho x$; see e.g. [58, Exercise 8.8]. Since g is convex, the inequality, $g(y) \geq g(x) + \langle v + \rho x, y - x\rangle$, holds for all $x, y \in \mathbb{R}^d$ and $v \in \partial\varphi(x)$. Algebraic manipulations then immediately imply (2.5), and therefore 3 holds. The implication 3 $\Rightarrow$ 4 follows by adding to (2.5) the analogous inequality with x and y interchanged. Finally suppose that 4 holds. Algebraic manipulations than imply that the subdifferential of $\varphi + \frac{\rho}{2}\|\cdot\|^2$ is a globally monotone map. Applying [58, Theorem 12.17], we conclude that $\varphi + \frac{\rho}{2}\|\cdot\|^2$ is convex and therefore 1 holds. Finally the characterization of weak convexity when φ is C^2-smooth is immediate from the second-order characterization of convexity of the function $\varphi + \frac{\rho}{2}\|\cdot\|^2$. □

For any function $\varphi\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and $\lambda > 0$, the *Moreau envelope* and the *proximal map* are defined by

$$\begin{aligned} \varphi_\lambda(x) &:= \min_y \ \left\{\varphi(y) + \tfrac{1}{2\lambda}\|y - x\|^2\right\}, \\ \operatorname{prox}_{\lambda\varphi}(x) &:= \operatorname*{argmin}_y \ \left\{\varphi(y) + \tfrac{1}{2\lambda}\|y - x\|^2\right\}, \end{aligned} \tag{2.6}$$

respectively. Classically, the Moreau envelope of a convex function is C^1-smooth for any $\lambda > 0$; see [43]. The same is true for weakly convex functions, provided λ is sufficiently small.

LEMMA 2.2. *Consider a ρ-weakly convex function $\varphi\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. Then for any $\lambda \in (0, \rho^{-1})$, the Moreau envelope φ_λ is C^1-smooth with gradient given by*

$$\nabla\varphi_\lambda(x) = \lambda^{-1}(x - \operatorname{prox}_{\lambda\varphi}(x)).$$

See Figure 1a for an illustration.

As mentioned in the introduction, the norm of the gradient $\|\nabla\varphi_\lambda(x)\|$ has an intuitive interpretation in terms of near-stationarity. Namely, the optimality conditions for the minimization problem in (2.6) directly imply that for any point $x \in \mathbb{R}^d$, the proximal point $\hat{x} := \operatorname{prox}_{\lambda\varphi}(x)$ satisfies

$$\begin{cases} \|\hat{x} - x\| & = \lambda\|\nabla\varphi_\lambda(x)\|, \\ \varphi(\hat{x}) & \leq \varphi(x), \\ \operatorname{dist}(0; \partial\varphi(\hat{x})) & \leq \|\nabla\varphi_\lambda(x)\|. \end{cases}$$

Thus a small gradient $\|\nabla\varphi_\lambda(x)\|$ implies that x is *near* some point $\hat{x}$ that is *nearly stationary* for φ; see Figure 1b. All of the convergence guarantees that we present will be in terms of the quantity $\|\nabla\varphi_\lambda(x)\|$.

It is important to keep in mind that in more classical circumstances, the size of the gradient of the Moreau envelope is proportional to more familiar quantities. To illustrate, consider the optimization problem

$$\min_{x \in \mathbb{R}^d} \ \varphi(x) := f(x) + r(x) \tag{2.7}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is C^1-smooth with ρ-Lipschitz gradient and $r\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is closed and convex. Much of the literature [36, 47] focusing on this problem class highlights the role of the *prox-gradient mapping*:

$$\mathcal{G}_\lambda(x) = \lambda^{-1}\left(x - \operatorname{prox}_{\lambda r}(x - \lambda\nabla f(x))\right). \tag{2.8}$$

Indeed, complexity estimates are typically stated in terms of the norm $\|\mathcal{G}_{1/\rho}(x)\|$. On the other hand, one can show that the two stationarity measures, $\|\nabla\varphi_{1/2\rho}(x)\|$ and $\|\mathcal{G}_{1/\rho}(x)\|$, are proportional [28, Theorem 4.5]:

$$\tfrac{1}{4}\|\nabla\varphi_{1/2\rho}(x)\| \le \|\mathcal{G}_{1/\rho}(x)\| \le \tfrac{3}{2}\left(1+\tfrac{1}{\sqrt{2}}\right)\|\nabla\varphi_{1/2\rho}(x)\| \qquad \forall x\in\mathbb{R}^d.$$

Thus when specializing our results to the setting (2.7), all of the convergence guarantees can be immediately translated in terms of the prox-gradient mapping.

3. Proximal stochastic subgradient method. In this section, we analyze the proximal stochastic subgradient method for weakly convex minimization. Throughout, we consider the optimization problem

$$\min_{x\in\mathbb{R}^d} \ \varphi(x) = f(x) + r(x), \tag{3.1}$$

where $r\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is a closed convex function and $f\colon \mathbb{R}^d\to\mathbb{R}$ is a ρ-weakly convex function. We assume that the only access to f is through a stochastic subgradient oracle.

ASSUMPTION A (Stochastic subgradient oracle). Fix a probability space $(\Omega,\mathcal{F},P)$ and equip $\mathbb{R}^d$ with the Borel σ-algebra. We make the following three assumptions:

- (A1) It is possible to generate i.i.d. realizations $\xi_1,\xi_2,\ldots\sim P$.
- (A2) There is an open set U containing $\operatorname{dom} r$ and a measurable mapping $G\colon U\times\Omega\to\mathbb{R}^d$ satisfying $\mathbb{E}_\xi[G(x,\xi)]\in\partial f(x)$ for all $x\in U$.
- (A3) There is a real $L\ge 0$ such that the inequality, $\mathbb{E}_\xi\left[\|G(x,\xi)\|^2\right]\le L^2$, holds for all $x\in\operatorname{dom} r$.

The three assumption (A1), (A2), (A3) are standard in the literature on stochastic subgradient methods: assumptions (A1) and (A2) are identical to assumptions (A1) and (A2) in [44], while Assumption (A3) is the same as the assumption listed in [44, Equation (2.5)]. We will investigate the efficiency of the proximal stochastic subgradient method, described in Algorithm 3.1.

Algorithm 3.1 Proximal stochastic subgradient method

Input: $x_0\in\operatorname{dom} r$, a sequence $\{\alpha_t\}_{t\ge 0}\subset\mathbb{R}_+$, and iteration count T

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

$$\left\{\begin{array}{l}\text{Sample } \xi_t\sim P\\ \text{Set } x_{t+1}=\operatorname{prox}_{\alpha_t r}\left(x_t-\alpha_t G(x_t,\xi_t)\right)\end{array}\right\},$$

Sample $t^*\in\{0,\ldots,T\}$ according to $\mathbb{P}(t^*=t)=\dfrac{\alpha_t}{\sum_{i=0}^{T}\alpha_i}$.

Return x_{t^*}

Henceforth, the symbol $\mathbb{E}_t[\cdot]$ will denote the expectation conditioned on all the realizations $\xi_0,\xi_1,\ldots,\xi_{t-1}$.

3.1. Projected stochastic subgradient method.

Our analysis of Algorithm 3.1 is shorter and more transparent when r is the indicator function of a closed, convex set $\mathcal{X}$. This is not surprising, since projected subgradient

methods are typically much easier to analyze than their proximal extensions (e.g. [19, 32]). Note that (3.1) then reduces to the constrained problem

$$\min_{x\in\mathcal{X}} \; f(x), \tag{3.2}$$

and the proximal map $\mathrm{prox}_{\alpha r}(\cdot)$ becomes the nearest point projection $\mathrm{proj}_{\mathcal{X}}(\cdot)$. Thus throughout Section 3.1, we suppose that Assumptions (A1), (A2), and (A3) hold and that $r(\cdot)$ is the indicator function of a closed convex set $\mathcal{X}$. The following is the main result of this section.

THEOREM 3.1 (Stochastic projected subgradient method). *Let x_{t^*} be the point returned by Algorithm 3.1. Then in terms of any constant $\bar{\rho} > \rho$, the estimate holds:*

$$\mathbb{E}\left[\varphi_{1/\bar{\rho}}(x_{t+1})\right] \leq \mathbb{E}[\varphi_{1/\bar{\rho}}(x_t)] - \frac{\alpha_t(\bar{\rho}-\rho)}{\bar{\rho}}\mathbb{E}\left[\|\nabla\varphi_{1/\bar{\rho}}(x_t)\|^2\right] + \frac{\alpha_t^2\bar{\rho}L^2}{2}, \tag{3.3}$$

and therefore we have

$$\mathbb{E}\left[\|\nabla\varphi_{1/\bar{\rho}}(x_{t^*})\|^2\right] \leq \frac{\bar{\rho}}{\bar{\rho}-\rho}\cdot\frac{(\varphi_{1/\bar{\rho}}(x_0)-\min\varphi)+\frac{\bar{\rho}L^2}{2}\sum_{t=0}^T\alpha_t^2}{\sum_{t=0}^T\alpha_t}. \tag{3.4}$$

In particular, if Algorithm 3.1 uses the constant parameter $\alpha_t = \sqrt{\frac{\Delta}{\rho L^2(T+1)}}$, for some $\Delta \geq \varphi_{1/2\rho}(x_0) - \min\varphi$, then the point x_{t^*} satisfies:

$$\mathbb{E}\left[\|\nabla\varphi_{1/2\rho}(x_{t^*})\|^2\right] \leq \sqrt{\frac{2\rho\Delta L^2}{T+1}}. \tag{3.5}$$

Proof. Let x_t denote the points generates by Algorithm 3.1. For each index t, define $v_t := \mathbb{E}_t[G(x_t,\xi)] \in \partial f(x_t)$ and set $\hat{x}_t := \mathrm{prox}_{\varphi/\bar{\rho}}(x_t)$. We successively deduce

$$\begin{aligned}
\mathbb{E}_t\left[\varphi_{1/\bar{\rho}}(x_{t+1})\right] &\leq \mathbb{E}_t\left[f(\hat{x}_t) + \frac{\bar{\rho}}{2}\|\hat{x}_t - x_{t+1}\|^2\right] && (3.6)\\
&= f(\hat{x}_t) + \frac{\bar{\rho}}{2}\mathbb{E}_t\left[\|\mathrm{proj}_{\mathcal{X}}(x_t - \alpha_t G(x_t,\xi_t)) - \mathrm{proj}_{\mathcal{X}}(\hat{x}_t)\|^2\right]\\
&\leq f(\hat{x}_t) + \frac{\bar{\rho}}{2}\mathbb{E}_t\left[\|(x_t-\hat{x}_t) - \alpha_t G(x_t,\xi_t)\|^2\right] && (3.7)\\
&\leq f(\hat{x}_t) + \frac{\bar{\rho}}{2}\|x_t-\hat{x}_t\|^2 + \bar{\rho}\alpha_t\mathbb{E}_t\left[\langle\hat{x}_t - x_t, G(x_t,\xi_t)\rangle\right] + \frac{\alpha_t^2\bar{\rho}L^2}{2}\\
&\leq \varphi_{1/\bar{\rho}}(x_t) + \bar{\rho}\alpha_t\langle\hat{x}_t - x_t, v_t\rangle + \frac{\alpha_t^2\bar{\rho}L^2}{2}\\
&\leq \varphi_{1/\bar{\rho}}(x_t) + \bar{\rho}\alpha_t\left(f(\hat{x}_t) - f(x_t) + \frac{\rho}{2}\|x_t-\hat{x}_t\|^2\right) + \frac{\alpha_t^2\bar{\rho}L^2}{2}, && (3.8)
\end{aligned}$$

where (3.6) follows directly from the definition of the proximal map, (3.7) uses that the projection $\mathrm{proj}_{\mathcal{X}}(\cdot)$ is 1-Lipschitz, and (3.8) follows from (2.5).

Next, observe that the function $x \mapsto f(x) + \frac{\bar{\rho}}{2}\|x-x_t\|^2$ is strongly convex with parameter $\bar{\rho}-\rho$, and therefore

$$\begin{aligned}
f(x_t) - f(\hat{x}_t) - \frac{\rho}{2}\|x_t-\hat{x}_t\|^2 &= \left(f(x_t) + \frac{\bar{\rho}}{2}\|x_t-x_t\|^2\right) - \left(f(\hat{x}_t) + \frac{\bar{\rho}}{2}\|x_t-\hat{x}_t\|^2\right)\\
&\quad + \frac{\bar{\rho}-\rho}{2}\|x_t-\hat{x}_t\|^2\\
&\geq (\bar{\rho}-\rho)\|x_t-\hat{x}_t\|^2 = \frac{\bar{\rho}-\rho}{\bar{\rho}^2}\|\nabla\varphi_{1/\bar{\rho}}(x_t)\|^2,
\end{aligned}$$

where the last equality follows from Lemma 2.2. Thus we deduce

$$\mathbb{E}_t\left[\varphi_{1/\bar{\rho}}(x_{t+1})\right] \leq \varphi_{1/\bar{\rho}}(x_t) - \frac{\alpha_t(\bar{\rho}-\rho)}{\bar{\rho}}\|\nabla\varphi_{1/\bar{\rho}}(x_t)\|^2 + \frac{\alpha_t^2\bar{\rho}L^2}{2}.$$

Taking expectations of both sides with respect to $\xi_0, \xi_1, \ldots, \xi_{t-1}$, and using the law of total expectation yields the claimed inequality (3.3)

Unfolding the recursion (3.3) yields:

$$\mathbb{E}\left[\varphi_{1/\bar{\rho}}(x_{T+1})\right] \leq \varphi_{1/\bar{\rho}}(x_0) + \frac{\bar{\rho}L^2}{2}\sum_{t=0}^{T}\alpha_t^2 - \frac{\bar{\rho}-\rho}{\bar{\rho}}\cdot\sum_{t=0}^{T}\alpha_t\mathbb{E}\left[\|\nabla\varphi_{1/\bar{\rho}}(x_t)\|^2\right].$$

Lower-bounding the left-hand side by $\min\varphi$ and rearranging, we obtain the bound:

$$\frac{1}{\sum_{t=0}^{T}\alpha_t}\sum_{t=0}^{T}\alpha_t\mathbb{E}\left[\|\nabla\varphi_{1/\bar{\rho}}(x_t)\|^2\right] \leq \frac{\bar{\rho}}{\bar{\rho}-\rho}\cdot\frac{\varphi_{1/\bar{\rho}}(x_0) - \min\varphi + \frac{\bar{\rho}L^2}{2}\sum_{t=0}^{T}\alpha_t^2}{\bar{\rho}\sum_{t=0}^{T}\alpha_t}.$$

Notice that the left-hand-side is precisely the expectation $\mathbb{E}\left[\|\nabla\varphi_{1/\bar{\rho}}(x_{t^*})\|^2\right]$. Thus (3.4) holds, as claimed. Finally, (3.5) follows from (3.4) by setting $\bar{\rho} = 2\rho$ and $\alpha_t = \frac{\gamma}{\sqrt{T+1}}$ for all indices $t = 0, 1, \ldots, T$. ∎

Let us translate the estimate (3.5) into a more convenient complexity bound. In particular, suppose that f is L-Lipschitz and the diameter of $\mathcal{X}$ is bounded by some $D > 0$. Then we may set $\Delta := \min\left\{\rho D^2, DL\right\}$, where the first term follows from the definition of the Moreau envelope and the second follows from Lipschitz continuity. Then the number of subgradient evaluations required to find a point x satisfying $\mathbb{E}\|\nabla\varphi_{1/2\rho}(x)\| \leq \varepsilon$ is at most

$$\left\lceil 16\cdot\frac{(\rho LD)^2\cdot\min\left\{1, \frac{L}{\rho D}\right\}}{\varepsilon^4}\right\rceil. \tag{3.9}$$

Improved complexity under convexity. It is intriguing to ask if the complexity (3.9) can be improved when f is a convex function. The answer, unsurprisingly, is yes. Since f is convex, here and for the rest of the section, we will let the constant $\rho > 0$ be arbitrary. As a first attempt, one may follow the observation of Nesterov [46] for smooth minimization. The idea is that the right-hand-side of the guarantee (3.4) depends on the initial gap $\varphi(x_0) - \min\varphi$. We can make this quantity as small as we wish in expectation by a separate subgradient method. Namely, we may simply run a stochastic subgradient method for T iterations to decrease the expected gap $\varphi(x_0) - \min\varphi$ to $\Delta := LD/\sqrt{T+1}$; see for example [37, Proposition 5.5] for this basic guarantee. Then we run another round of a stochastic subgradient method for T iterations using the optimal choice $\alpha := \sqrt{\frac{\Delta}{\rho L^2(T+1)}}$. A quick computation shows that the resulting two-round scheme will find a point x satisfying $\mathbb{E}\|\nabla\varphi_{1/(2\rho)}(x)\| \leq \varepsilon$ after at most $O(1)\cdot\frac{L^2(\rho D)^{2/3}}{\varepsilon^{8/3}}$ iterations.

By following a completely different technique, introduced by Allen-Zhu [5] for smooth stochastic minimization, this complexity can be even further improved to $\widetilde{O}\left(\frac{(L^2+\rho^2D^2)\log^3(\frac{\rho D}{\varepsilon})}{\varepsilon^2}\right)$ by running logarithmically many rounds of the stochastic subgradient method on quadratically regularized problems. Since this procedure and its analysis is somewhat long and is independent of the rest of the material, we have placed it in an independent arXiv technical report [20].

3.2. Proximal stochastic subgradient method. We next move on to convergence guarantees of Algorithm 3.1 in full generality. An important consequence we discuss at the end of the section is a convergence guarantee for the stochastic proximal gradient method for minimizing a sum of a smooth function and a convex function, where the gradient oracle has bounded variance (instead of bounded second moment). Those not interested in this guarantee can in principle skip to Section 4, which details our most general convergence result for nonsmooth minimization.

Before we proceed, note that for any $x \in U$ and $v \in \partial f(x)$, we have $\|v\| \leq L$. To see this, observe that (A2) and (A3) directly imply that whenever f is differentiable at $x \in U$, we have

$$\|\nabla f(x)\|^2 = \|\mathbb{E}_\xi[G(x,\xi)]\|^2 \leq \mathbb{E}_\xi[\|G(x,\xi)\|^2] \leq L^2.$$

Since at any point x, the subdifferential $\partial f(x)$ is the convex hull of limits of gradients at nearby points [55, Theorem 25.6], the claim follows. We will use this estimate in the proof of Lemma 3.3.

We break up the analysis of Algorithm 3.1 into two lemmas. Henceforth, fix a real $\bar{\rho} > \rho$. Let x_t be the iterates produced by Algorithm 3.1 and let $\xi_t \sim P$ be the i.i.d. realizations used. For each index t, define $v_t := \mathbb{E}_t[G(x_t,\xi)] \in \partial f(x_t)$ and set $\hat{x}_t := \mathrm{prox}_{\varphi/\bar{\rho}}(x_t)$. Observe that by the optimality conditions of the proximal map and the subdifferential sum rule [58, Exercise 10.10], there exists a vector $\hat{v}_t \in \partial f(\hat{x}_t)$ satisfying $\bar{\rho}(x_t - \hat{x}_t) \in \partial r(\hat{x}_t) + \hat{v}_t$. The following lemma realizes $\hat{x}_t$ as a proximal point of r.

LEMMA 3.2. *For each index $t \geq 0$, equality holds:*

$$\hat{x}_t = \mathrm{prox}_{\alpha_t r}\left(\alpha_t\bar{\rho}x_t - \alpha_t\hat{v}_t + (1-\alpha_t\bar{\rho})\hat{x}_t\right).$$

Proof. By the definition of $\hat{v}_t$, we have

$$\begin{aligned}\alpha_t\bar{\rho}(x_t - \hat{x}_t) \in \alpha_t\partial r(\hat{x}_t) + \alpha_t\hat{v}_t &\iff \alpha_t\bar{\rho}x_t - \alpha_t\hat{v}_t + (1-\alpha_t\bar{\rho})\hat{x}_t \in \hat{x}_t + \alpha_t\partial r(\hat{x}_t)\\ &\iff \hat{x}_t = \mathrm{prox}_{\alpha_t r}(\alpha_t\bar{\rho}x_t - \alpha_t\hat{v}_t + (1-\alpha_t\bar{\rho})\hat{x}_t),\end{aligned}$$

where the last equivalence follows from the optimality conditions for the proximal subproblem. This completes the proof. ∎

The next lemma establishes a crucial descent property for the iterates.

LEMMA 3.3. *Suppose $\bar{\rho} \in (\rho, 2\rho]$ and we have $\alpha_t \in (0, 1/\bar{\rho}]$ for all indices $t \geq 0$. Then the inequality holds:*

$$\mathbb{E}_t\|x_{t+1} - \hat{x}_t\|^2 \leq \|x_t - \hat{x}_t\|^2 + 4\alpha_t^2L^2 - 2\alpha_t(\bar{\rho}-\rho)\|x_t - \hat{x}_t\|^2.$$

Proof. Set $\delta := 1 - \alpha_t\bar{\rho}$. We successively deduce

$$\begin{aligned}\mathbb{E}_t\|x_{t+1} - \hat{x}_t\|^2 &= \mathbb{E}_t\|\mathrm{prox}_{\alpha_t r}(x_t - \alpha_t G(x_t,\xi_t)) - \mathrm{prox}_{\alpha_t r}(\alpha_t\bar{\rho}x_t - \alpha_t\hat{v}_t + \delta\hat{x}_t)\|^2\\
&\leq \mathbb{E}_t\|x_t - \alpha_t G(x_t,\xi_t) - (\alpha_t\bar{\rho}x_t - \alpha_t\hat{v}_t + \delta\hat{x}_t)\|^2 \qquad (3.10)\\
&= \mathbb{E}_t\|\delta(x_t - \hat{x}_t) - \alpha_t(G(x_t,\xi_t) - \hat{v}_t)\|^2 \qquad (3.11)\\
&= \delta^2\|x_t - \hat{x}_t\|^2 - 2\delta\alpha_t\mathbb{E}_t\left[\langle x_t - \hat{x}_t, G(x_t,\xi_t) - \hat{v}_t\rangle\right]\\
&\quad + \alpha_t^2\mathbb{E}_t\|G(x_t,\xi_t) - \hat{v}_t\|^2\\
&= \delta^2\|x_t - \hat{x}_t\|^2 - 2\delta\alpha_t\langle x_t - \hat{x}_t, v_t - \hat{v}_t\rangle + 4\alpha_t^2L^2\\
&\leq \delta^2\|x_t - \hat{x}_t\|^2 + 2\delta\alpha_t\rho\|x_t - \hat{x}_t\|^2 + 4\alpha_t^2L^2 \qquad (3.12)\\
&= (1 - (2\alpha_t(\bar{\rho}-\rho) + \alpha_t^2\bar{\rho}(2\rho - \bar{\rho})))\|x_t - \hat{x}_t\|^2 + 4\alpha_t^2L^2,\end{aligned}$$

where the first equation follows from Lemma 3.2, (3.10) uses that $\mathrm{prox}_{\alpha_t r}(\cdot)$ is 1-Lipschitz [58, Proposition 12.19], and (3.12) follows from (4). The result now follows from the assumed inequality $\bar{\rho} \leq 2\rho$. ∎

With Lemma 3.3 proved, we can now establish convergence guarantees of Algorithm 3.1 in full generality.

THEOREM 3.4 (Stochastic proximal subgradient method). *Fix a real $\bar{\rho} \in (\rho, 2\rho]$ and a stepsize sequence $\alpha_t \in (0, 1/\bar{\rho}]$. Then the iterates x_t generated by Algorithm 3.1 satisfy*

$$\mathbb{E}\left[\varphi_{1/\bar{\rho}}(x_{t+1})\right] \leq \mathbb{E}[\varphi_{1/\bar{\rho}}(x_t)] - \frac{\alpha_t(\bar{\rho}-\rho)}{\bar{\rho}}\mathbb{E}\left[\|\nabla\varphi_{1/\bar{\rho}}(x_t)\|^2\right] + \alpha_t^2\bar{\rho}L^2, \tag{3.13}$$

and the point x_{t^*} returned by Algorithm 3.1 satisfies:

$$\mathbb{E}\left[\|\nabla\varphi_{1/\bar{\rho}}(x_{t^*})\|^2\right] \leq \frac{\bar{\rho}}{\bar{\rho}-\rho} \cdot \frac{(\varphi_{1/\bar{\rho}}(x_0) - \min\varphi) + 2\bar{\rho}L^2\sum_{t=0}^T \alpha_t^2}{\sum_{t=0}^T \alpha_t}. \tag{3.14}$$

In particular, if Algorithm 3.1 uses the parameter $\alpha_t = \frac{1}{2}\min\left\{\frac{1}{\rho}, \sqrt{\frac{\Delta}{\rho L^2(T+1)}}\right\}$ for some real $\Delta \geq \varphi_{1/\bar{\rho}}(x_0) - \min\varphi$, then the point x_{t^*} satisfies:

$$\mathbb{E}\left[\|\nabla\varphi_{1/2\rho}(x_{t^*})\|^2\right] \leq 8 \cdot \max\left\{\frac{\rho\Delta}{T+1}, L\sqrt{\frac{\rho\Delta}{T+1}}\right\}. \tag{3.15}$$

Proof. We successively observe

$$\begin{aligned}
\mathbb{E}_t\left[\varphi_{1/\bar{\rho}}(x_{t+1})\right] &\leq \mathbb{E}_t\left[\varphi(\hat{x}_t) + \frac{\bar{\rho}}{2}\|\hat{x}_t - x_{t+1}\|^2\right] \\
&\leq \varphi(\hat{x}_t) + \frac{\bar{\rho}}{2}\left[\|x_t - \hat{x}_t\|^2 + 4\alpha_t^2L^2 - 2\alpha_t(\bar{\rho}-\rho)\|x_t - \hat{x}_t\|^2\right] \\
&= \varphi_{1/\bar{\rho}}(x_t) + \bar{\rho}\left[2\alpha_t^2L^2 - \alpha_t(\bar{\rho}-\rho)\|x_t - \hat{x}_t\|^2\right],
\end{aligned}$$

where the first inequality follows directly from the definition of the proximal map and the second follows from Lemma 3.3. Taking expectations with respect to $\xi_0, \ldots, \xi_{t-1}$ yields the claimed inequality (3.13). The rest of the proof proceeds as in Theorem 3.1. Namely, unfolding the recursion (3.13) yields:

$$\mathbb{E}\left[\varphi_{1/\bar{\rho}}(x_{T+1})\right] \leq \varphi_{1/\bar{\rho}}(x_0) + 2\bar{\rho}L^2\sum_{t=0}^T \alpha_t^2 - \frac{\bar{\rho}-\rho}{\bar{\rho}}\mathbb{E}\sum_{t=0}^T \alpha_t\|x_t - \hat{x}_t\|^2.$$

Lower-bounding the left-hand side by $\min\varphi$ and rearranging, we obtain the bound:

$$\frac{1}{\sum_{t=0}^T \alpha_t}\sum_{t=0}^T \alpha_t\mathbb{E}\left[\|\nabla\varphi_{1/\bar{\rho}}(x_t)\|^2\right] \leq \frac{\bar{\rho}}{\bar{\rho}-\rho} \cdot \frac{(\varphi_{1/\bar{\rho}}(x_0) - \min\varphi) + 2\bar{\rho}L^2\sum_{t=0}^T \alpha_t^2}{\sum_{t=0}^T \alpha_t}. \tag{3.16}$$

Recognizing the left-hand-side as $\mathbb{E}\left[\|x_{t^*} - \hat{x}_{t^*}\|^2\right]$ establishes (3.14).

To establish (3.15), set $\bar{\rho} := 2\rho$ and $\alpha_t := \min\left\{\frac{1}{2\rho}, \sqrt{\frac{\Delta}{4\rho L^2(T+1)}}\right\}$. There are two cases to consider. Supposing first $\sqrt{\frac{\Delta}{4\rho L^2(T+1)}} \leq \frac{1}{2\rho}$, and plugging $\alpha_t = \sqrt{\frac{\Delta}{4\rho L^2(T+1)}}$ into (3.14) yields

$$\mathbb{E}\left[\|\nabla\varphi_{1/2\rho}(x_{t^*})\|^2\right] \leq 8L\sqrt{\frac{\rho\Delta}{T+1}}.$$

Hence (3.15) holds. Suppose instead $\sqrt{\frac{\Delta}{4\rho L^2(T+1)}} \geq \frac{1}{2\rho}$, or equivalently $L^2 \leq \frac{\rho\Delta}{T+1}$. Then plugging $\alpha_t = \frac{1}{2\rho}$ into (3.14) yields the estimate

$$\mathbb{E}\left[\|\nabla\varphi_{1/2\rho}(x_{t^*})\|^2\right] \leq \frac{4\rho\Delta}{T+1} + 4L^2 \leq \frac{8\rho\Delta}{T+1}.$$

Thus (3.15) is proved. ∎

Proximal stochastic gradient for smooth minimization. We next look at the consequences of our results in the setting when f is C^1-smooth with ρ-Lipschitz gradient. Note, that then f is automatically ρ-weakly convex. In this smooth setting, it is common to replace assumption (A3) with the finite variance condition:

$\overline{(A3)}$ There is a real $\sigma \geq 0$ such that the inequality, $\mathbb{E}_\xi\left[\|G(x,\xi) - \nabla f(x)\|^2\right] \leq \sigma^2$, holds for all $x \in \operatorname{dom} r$.

Henceforth, let us therefore assume that f is C^1-smooth with ρ-Lipschitz gradient, and Assumptions (A1), (A2), and $\overline{(A3)}$ hold.

All of the results in Section 3.2 can be easily modified to apply to this setting. In particular, Lemma 3.2 holds verbatim, while Lemma 3.3 extends as follows.

LEMMA 3.5. *Fix a real $\bar{\rho} > \rho$ and a sequence $\alpha_t \in (0, 1/\bar{\rho}]$. Then the inequality holds:*

$$\mathbb{E}_t\|x_{t+1} - \hat{x}_t\|^2 \leq \|x_t - \hat{x}_t\|^2 + \alpha_t^2\sigma^2 - \alpha_t(\bar{\rho} - \rho)\|x_t - \hat{x}_t\|^2.$$

Proof. By the same argument as in Lemma 3.3, we arrive at (3.11) with $\hat{v}_t = \nabla f(\hat{x}_t)$. Set $\delta := 1 - \alpha_t\bar{\rho}$ and $w_t := G(x_t,\xi_t) - \nabla f(x_t)$. Adding and subtracting $\nabla f(x_t)$, we successively deduce

$$\begin{aligned}
\mathbb{E}_t\|x_{t+1} - \hat{x}_t\|^2 &\leq \mathbb{E}_t\|\delta(x_t - \hat{x}_t) - \alpha_t(G(x_t,\xi_t) - \nabla f(\hat{x}_t))\|^2 \\
&= \mathbb{E}_t\|\delta(x_t - \hat{x}_t) - \alpha_t(\nabla f(x_t) - \nabla f(\hat{x}_t)) - \alpha_t w_t\|^2 \\
&= \|\delta(x_t - \hat{x}_t) - \alpha_t(\nabla f(x_t) - \nabla f(\hat{x}_t))\|^2 + \alpha_t^2\mathbb{E}_t\|w_t\|^2 \qquad (3.17)\\
&\leq \delta^2\|x_t - \hat{x}_t\|^2 - 2\delta\alpha_t\langle x_t - \hat{x}_t, \nabla f(x_t) - \nabla f(\hat{x}_t)\rangle \\
&\quad + \alpha_t^2\|\nabla f(x_t) - \nabla f(\hat{x}_t)\|^2 + \alpha_t^2\sigma^2 \qquad (3.18)\\
&\leq (\delta^2 + 2\delta\alpha_t\rho + \rho^2\alpha_t^2)\|x_t - \hat{x}_t\|^2 + \alpha_t^2\sigma^2 \qquad (3.19)\\
&= \|x_t - \hat{x}_t\|^2 + \alpha_t^2\sigma^2 - \alpha_t(\bar{\rho} - \rho)(2 - \alpha_t(\bar{\rho} - \rho))\|x_t - \hat{x}_t\|^2,
\end{aligned}$$

where (3.17) follows from assumption (A2), namely $\mathbb{E}_t G(x_t,\xi_t) = \nabla f(x_t)$, (3.18) follows by expanding the square and using assumption $\overline{(A3)}$, and (3.19) follows from (4) and Lipschitz continuity of ∇f. The assumption $\bar{\rho} \geq \rho$ guarantees $2 - \alpha_t(\bar{\rho} - \rho) \geq 1$. The result follows. ∎

We can now state the convergence guarantees of the proximal stochastic gradient method. The proof is completely analogous to that of Theorem 3.4, with Lemma 3.5 playing the role of Lemma 3.3.

COROLLARY 3.6 (Stochastic prox-gradient method for smooth minimization). *Fix a real $\bar{\rho} > \rho$ and a stepsize sequence $\alpha_t \in (0, 1/\bar{\rho}]$. Then the iterates x_t generated by Algorithm 3.1 satisfy*

$$\mathbb{E}\left[\varphi_{1/\bar{\rho}}(x_{t+1})\right] \leq \mathbb{E}[\varphi_{1/\bar{\rho}}(x_t)] - \frac{\alpha_t(\bar{\rho} - \rho)}{2\bar{\rho}}\mathbb{E}\left[\|\nabla\varphi_{1/\bar{\rho}}(x_t)\|^2\right] + \frac{\alpha_t^2\bar{\rho}\sigma^2}{2}, \qquad (3.20)$$

and the point x_{t^*} returned by Algorithm 3.1 satisfies:

$$\mathbb{E}\left[\|\nabla\varphi_{1/\bar{\rho}}(x_{t^*})\|^2\right] \leq \frac{2\bar{\rho}}{\bar{\rho}-\rho}\cdot\frac{(\varphi_{1/\bar{\rho}}(x_0)-\min\varphi)+\frac{\bar{\rho}\sigma^2}{2}\sum_{t=0}^T\alpha_t^2}{\sum_{t=0}^T\alpha_t}. \tag{3.21}$$

In particular, if Algorithm 3.1 uses the constant parameter $\alpha_t = \min\left\{\frac{1}{2\rho}, \sqrt{\frac{\Delta}{\rho\sigma^2(T+1)}}\right\}$ for some $\Delta \geq \varphi_{1/2\rho}(x_0) - \min\varphi$, then the point x_{t^*} satisfies:

$$\mathbb{E}\left[\|\nabla\varphi_{1/(2\rho)}(x_{t^*})\|^2\right] \leq 8\cdot\max\left\{\frac{2\rho\Delta}{T+1}, \sigma\sqrt{\frac{\rho\Delta}{T+1}}\right\}. \tag{3.22}$$

As mentioned at the end of Section 2.2, it is immediate to translate the complexity estimate in Corollary 3.6 to an analogous estimate in terms of the size of the prox-gradient mapping (2.8), thereby allowing for a direct comparison with previous results.

4. Stochastic model-based minimization. In the previous section, we established the complexity of $O(\varepsilon^{-4})$ for the stochastic proximal subgradient methods. In this section, we show that the complexity $O(\varepsilon^{-4})$ persists for a much wider class of algorithms, including the stochastic proximal point and prox-linear algorithms. Henceforth, we consider the optimization problem

$$\min_{x\in\mathbb{R}^d}\ \varphi(x) := f(x) + r(x), \tag{4.1}$$

where $r\colon \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ is a closed function (not necessarily convex) and $f\colon \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz. We assume that the only access to f is through a *stochastic one-sided model*.

Assumption B (Stochastic one-sided model). Fix a probability space $(\Omega, \mathcal{F}, P)$ and equip $\mathbb{R}^d$ with the Borel σ-algebra. We assume that there exist real $\tau, \eta, \mathsf{L} \in \mathbb{R}$ such that the following four properties hold:

(B1) **(Sampling)** It is possible to generate i.i.d. realizations $\xi_1, \xi_2, \ldots \sim P$.

(B2) **(One-sided accuracy)** There is an open convex set U containing $\operatorname{dom} r$ and a measurable function $(x, y, \xi) \mapsto g_x(y, \xi)$, defined on $U \times U \times \Omega$, satisfying

$$\mathbb{E}_\xi\left[f_x(x,\xi)\right] = f(x) \qquad \forall x \in U,$$

and

$$\mathbb{E}_\xi\left[f_x(y,\xi) - f(y)\right] \leq \frac{\tau}{2}\|y-x\|^2 \qquad \forall x, y \in U.$$

(B3) **(Weak-convexity)** The function $f_x(\cdot,\xi) + r(\cdot)$ is η-weakly convex $\forall x \in U$, a.e. $\xi \in \Omega$.

(B4) **(Lipschitz property)** There exists a measurable function $L\colon \Omega \to \mathbb{R}_+$ satisfying $\sqrt{\mathbb{E}_\xi\left[L(\xi)^2\right]} \leq \mathsf{L}$ and such that

$$f_x(x,\xi) - f_x(y,\xi) \leq L(\xi)\|x-y\|, \tag{4.2}$$

for all $x, y \in U$ and a.e. $\xi \sim P$.

It will be useful for the reader to keep in mind the following lemma, which shows that the objective function φ is itself weakly convex with parameter $\tau+\eta$ and that f is L-Lipschitz continuous on U.

LEMMA 4.1. *The function φ is $(\tau+\eta)$-weakly convex and the inequality holds:*

$$|f(x) - f(y)| \le \mathsf{L}\|x-y\|, \qquad \text{for all } x, y \in U. \tag{4.3}$$

Proof. Fix arbitrary points $x, y \in \operatorname{dom} r$ and a real $\lambda \in [0,1]$, and set $\bar{x} = \lambda x + (1-\lambda)y$. Define the function $f_x(y) := \mathbb{E}_\xi[f_x(y,\xi)]$. Taking into account the equivalence of weak convexity with the approximate secant inequality (2.4), we successively deduce

$$\varphi(\bar{x}) = \mathbb{E}_\xi\left[r(\bar{x}) + f_{\bar{x}}(\bar{x},\xi)\right] \tag{4.4}$$

$$\le \lambda \mathbb{E}_\xi\left[r(x) + f_{\bar{x}}(x,\xi)\right] + (1-\lambda)\mathbb{E}_\xi\left[r(y) + f_{\bar{x}}(y,\xi)\right] + \tfrac{\eta\lambda(1-\lambda)}{2}\|x-y\|^2 \tag{4.5}$$

$$= \lambda(r(x) + f_{\bar{x}}(x)) + (1-\lambda)(r(y) + f_{\bar{x}}(y)) + \tfrac{\eta\lambda(1-\lambda)}{2}\|x-y\|^2$$

$$\le \lambda\varphi(x) + (1-\lambda)\varphi(y) + \tfrac{\tau(\lambda^2(1-\lambda)+\lambda(1-\lambda)^2)}{2}\|x-y\|^2 + \tfrac{\eta\lambda(1-\lambda)}{2}\|x-y\|^2 \tag{4.6}$$

$$= \lambda\varphi(x) + (1-\lambda)\varphi(y) + \tfrac{(\tau+\eta)\lambda(1-\lambda)}{2}\|x-y\|^2,$$

where (4.4) uses (B2), inequality (4.5) uses (B3), and (4.6) uses (B2). Thus φ is $(\tau+\eta)$-weakly convex, as claimed.

Next, taking expectations in (B2) and in (4.2) yields the estimates:

$$f(x) - f_x(y) \le \mathsf{L}\|x-y\| \qquad \text{and} \qquad f_x(y) - f(y) \le \frac{\tau}{2}\|x-y\|^2.$$

Thus for any point $x \in U$, we deduce

$$\limsup_{y\to x} \frac{f(x)-f(y)}{\|x-y\|} \le \limsup_{y\to x} \frac{\mathsf{L}\|x-y\| + \frac{\tau}{2}\|y-x\|^2}{\|x-y\|} = \mathsf{L}.$$

In particular, when f is differentiable at x, setting $y = x - s\nabla f(x)$ with $s \searrow 0$, we deduce $\|\nabla f(x)\| \le \mathsf{L}$. Since f is locally Lipschitz continuous, its Lipschitz constant on U is no greater than $\sup_{y\in U}\{\|\nabla f(y)\| : f \text{ is differentiable at } y\}$.[3] We therefore deduce that f is L-Lipschitz continuous on U, as claimed. □

We can now formalize the algorithm we investigate, as Algorithm 4.1. The reader should note that, in contrast to the previously discussed algorithms, Algorithm 4.1 employs a nondecreasing stepsize β_t, which is inversely proportional to α_t. This notational choice will simplify the analysis and complexity guarantees that follow.

Algorithm 4.1 Stochastic Model Based Minimization

Input: $x_0 \in \mathbb{R}^d$, real $\bar{\rho} > \tau + \eta$, a sequence $\{\beta_t\}_{t\ge 0} \subseteq (\bar{\rho}, \infty)$, and iteration count T

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

$$\left\{ \begin{array}{l} \text{Sample } \xi_t \sim P \\ \text{Set } x_{t+1} = \operatorname*{argmin}_x \left\{ r(x) + f_{x_t}(x,\xi_t) + \dfrac{\beta_t}{2}\|x - x_t\|^2 \right\} \end{array} \right\},$$

Sample $t^* \in \{0, \ldots, T\}$ according to the discrete probability distribution

$$\mathbb{P}(t^* = t) \propto \frac{\bar{\rho} - \tau - \eta}{\beta_t - \eta}.$$

Return x_{t^*}

[3] This follows by combining gradient formula for the Clarke subdifferential [18, Theorem 8.1] with the mean value theorem [18, Theorem 2.4].

4.1. Analysis of the algorithm. Henceforth, let $\{x_t\}_{t\geq 0}$ be the iterates generated by Algorithm 4.1 and let $\{\xi_t\}_{t\geq 0}$ be the corresponding samples used. For each index $t \geq 0$, define the proximal point

$$\hat{x}_t = \mathrm{prox}_{\varphi/\bar{\rho}}(x_t).$$

As in Section 3, we will use the symbol $\mathbb{E}_t[\cdot]$ to denote the expectation conditioned on all the realizations $\xi_0, \xi_1, \ldots, \xi_{t-1}$. The analysis of Algorithm 4.1 relies on the following lemma, which establishes two descent type properties. Estimate (4.7) is in the same spirit as Lemma 3.3 in Section 3. The estimate (4.8), in contrast, will be used at the end of the section to obtain the convergence rate of Algorithm 4.1 in function values under convexity assumptions.

LEMMA 4.2. *In general, for every index $t \geq 0$, we have*

$$\mathbb{E}_t\|\hat{x}_t - x_{t+1}\|^2 \leq \|\hat{x}_t - x_t\|^2 - \frac{\bar{\rho} - \tau - \eta}{\beta_t - \eta}\|\hat{x}_t - x_t\|^2 + \frac{4\mathsf{L}^2}{(\beta_t - \eta)(\beta_t - \bar{\rho})}. \tag{4.7}$$

Moreover, for any point $x \in \operatorname{dom} r$, the inequality holds:

$$\mathbb{E}_t\left[\|x_{t+1} - x\|^2\right] \leq \frac{\beta_t + \tau}{\beta_t - \eta}\|x_t - x\|^2 - \frac{2}{\beta_t - \eta}\mathbb{E}_t[\varphi(x_{t+1}) - \varphi(x)] + \frac{2\mathsf{L}^2}{\beta_t(\beta_t - \eta)}. \tag{4.8}$$

Proof. Recall that the function $x \mapsto r(x) + f_{x_t}(x, \xi_t) + \frac{\beta_t}{2}\|x - x_t\|^2$ is strongly convex with constant $\beta_t - \eta$ and x_{t+1} is its minimizer. Hence for any $x \in \operatorname{dom} r$, the inequality holds:

$$\begin{aligned}\left(r(x) + f_{x_t}(x, \xi_t) + \tfrac{\beta_t}{2}\|x - x_t\|^2\right) \geq \left(r(x_{t+1}) + f_{x_t}(x_{t+1}, \xi_t) + \tfrac{\beta_t}{2}\|x_{t+1} - x_t\|^2\right)\\ + \tfrac{\beta_t - \eta}{2}\|x - x_{t+1}\|^2.\end{aligned}$$

Rearranging and taking expectations we successively deduce

$$\begin{aligned}
&\mathbb{E}_t\left[\frac{\beta_t - \eta}{2}\|x - x_{t+1}\|^2 + \frac{\beta_t}{2}\|x_{t+1} - x_t\|^2 - \frac{\beta_t}{2}\|x - x_t\|^2\right]\\
&\leq \mathbb{E}_t[r(x) + f_{x_t}(x, \xi_t) - r(x_{t+1}) - f_{x_t}(x_{t+1}, \xi_t)]\\
&\leq \mathbb{E}_t[r(x) + f_{x_t}(x, \xi_t) - r(x_{t+1}) - f_{x_t}(x_t, \xi_t) + L(\xi)\|x_{t+1} - x_t\|] \qquad (4.9)\\
&\leq r(x) + \mathbb{E}_\xi[f_{x_t}(x, \xi)] - \mathbb{E}_t[r(x_{t+1})] - \mathbb{E}_\xi[f_{x_t}(x_t, \xi)]\\
&\quad + \sqrt{\mathbb{E}_\xi[L(\xi)^2]} \cdot \sqrt{\mathbb{E}_t[\|x_{t+1} - x_t\|^2]} \qquad (4.10)\\
&\leq r(x) + f(x) - \mathbb{E}_t[r(x_{t+1})] - f(x_t) + \frac{\tau}{2}\|x - x_t\|^2 + \mathsf{L}\sqrt{\mathbb{E}_t[\|x_{t+1} - x_t\|^2]} \qquad (4.11)\\
&= \mathbb{E}_t[r(x) + f(x) - r(x_{t+1}) - f(x_t)] + \frac{\tau}{2}\|x - x_t\|^2 + \mathsf{L}\sqrt{\mathbb{E}_t[\|x_{t+1} - x_t\|^2]}\\
&\leq \mathbb{E}_t[r(x) + f(x) - r(x_{t+1}) - f(x_{t+1})] + \frac{\tau}{2}\|x - x_t\|^2 \qquad (4.12)\\
&\quad + \mathsf{L}\mathbb{E}_t[\|x_{t+1} - x_t\|] + \mathsf{L}\sqrt{\mathbb{E}_t[\|x_{t+1} - x_t\|^2]},
\end{aligned}$$

where (4.9) follows from Assumption (B4), inequality (4.10) follows from Cauchy-Schwartz, inequality (4.11) follows from (B2), (4.12) follows from Lemma 4.1.

Define $\delta := \sqrt{\mathbb{E}_t[\|x_{t+1} - x_t\|^2]}$ and notice $\delta \geq \mathbb{E}_t\|x_t - x_{t+1}\|$. Rearranging (4.12), we immediately deduce

$$\begin{aligned}\mathbb{E}_t\left[\frac{\beta_t - \eta}{2}\|x - x_{t+1}\|^2\right] &\leq \mathbb{E}_t\left[\frac{\beta_t + \tau}{2}\|x^* - x_t\|^2\right] - \frac{\beta_t\delta^2}{2} + 2\mathsf{L}\delta - \mathbb{E}_t[\varphi(x_{t+1}) - \varphi(x)] \\ &\leq \mathbb{E}_t\left[\frac{\beta_t + \tau}{2}\|x - x_t\|^2\right] + \frac{2\mathsf{L}^2}{\beta_t} - \mathbb{E}_t[\varphi(x_{t+1}) - \varphi(x)],\end{aligned}$$

where the last inequality follows by maximizing the right-hand-side in $\delta \in \mathbb{R}$. Dividing through by $\frac{\beta-\eta}{2}$, we arrive at the claimed inequality (4.8).

Next setting $x = \hat{x}_t$ in (4.12) and using the definition of the prox-point, we obtain

$$\begin{aligned}&\mathbb{E}_t\left[\frac{\beta_t - \eta}{2}\|\hat{x}_t - x_{t+1}\|^2 + \frac{\beta_t}{2}\|x_{t+1} - x_t\|^2 - \frac{\beta_t}{2}\|\hat{x}_t - x_t\|^2\right] \\ &\leq \mathbb{E}_t\left[-\frac{\bar{\rho}}{2}\|\hat{x}_t - x_t\|^2 + \frac{\bar{\rho}}{2}\|x_{t+1} - x_t\|^2\right] + \frac{\tau}{2}\|\hat{x}_t - x_t\|^2 + 2\mathsf{L}\delta \\ &= \frac{\tau - \bar{\rho}}{2}\|\hat{x}_t - x_t\|^2 + \frac{\bar{\rho}}{2}\cdot\mathbb{E}_t[\|x_{t+1} - x_t\|^2] + 2\mathsf{L}\delta.\end{aligned}$$

Rearranging, we deduce

$$\begin{aligned}\frac{\beta_t - \eta}{2}\cdot\mathbb{E}_t\|\hat{x}_t - x_{t+1}\|^2 &\leq \frac{\beta_t - \bar{\rho} + \tau}{2}\|\hat{x}_t - x_t\|^2 + \frac{\bar{\rho} - \beta_t}{2}\delta^2 + 2\mathsf{L}\delta \\ &\leq \frac{\beta_t - \bar{\rho} + \tau}{2}\|\hat{x}_t - x_t\|^2 + \frac{2\mathsf{L}^2}{\beta_t - \bar{\rho}},\end{aligned} \tag{4.13}$$

where the last inequality follows by maximizing the right-hand-side of (4.13) in $\delta \in \mathbb{R}$. After multiplying through by $\frac{2}{\beta_t - \eta}$, we arrive at the claimed estimate (4.7). ∎

We can now establish the convergence guarantees of Algorithm 4.1.

THEOREM 4.3 (Convergence rate). *Fix a real $\bar{\rho} > \tau + \eta$ and a sequence $\{\beta_t\}_{t\geq 0} \in (\bar{\rho}, \infty)$. Then the iterates x_t generated by Algorithm 4.1 satisfy*

$$\mathbb{E}\left[\varphi_{1/\bar{\rho}}(x_{t+1})\right] \leq \mathbb{E}[\varphi_{1/\bar{\rho}}(x_t)] - \frac{\bar{\rho} - \tau - \eta}{2\bar{\rho}(\beta_t - \eta)}\mathbb{E}\left[\|\nabla\varphi_{1/\bar{\rho}}(x_t)\|^2\right] + \frac{2\bar{\rho}\mathsf{L}^2}{(\beta_t - \eta)(\beta_t - \bar{\rho})}, \tag{4.14}$$

and the point x_{t^*} returned by Algorithm 3.1 satisfies:

$$\mathbb{E}\|\nabla\varphi_{1/\bar{\rho}}(x_{t^*})\|^2 \leq \frac{\bar{\rho}(\varphi_{1/\bar{\rho}}(x_0) - \min_x\varphi) + 2\bar{\rho}^2\mathsf{L}^2\cdot\sum_{t=0}^T\frac{1}{(\beta_t - \eta)(\beta_t - \bar{\rho})}}{\sum_{t=0}^T\frac{\bar{\rho} - \tau - \eta}{2(\beta_t - \eta)}}. \tag{4.15}$$

In particular, if Algorithm 3.1 uses the constant parameter $\beta_t = \bar{\rho} + \sqrt{\frac{2\bar{\rho}\mathsf{L}^2(T+1)}{\Delta}}$, for some real $\Delta \geq \varphi_{1/\bar{\rho}}(x_0) - \min\varphi$, and sets $\bar{\rho} = 2(\rho + \eta)$, then the point x_{t^*} satisfies:

$$\mathbb{E}\|\nabla\varphi_{1/\bar{\rho}}(x_{t^*})\|^2 \leq \frac{4\bar{\rho}\Delta}{T+1} + 8\mathsf{L}\sqrt{\frac{2\bar{\rho}\Delta}{T+1}}. \tag{4.16}$$

Proof. Using the definition of the Moreau envelope and appealing to the estimate

(4.7) in Lemma 4.2, we deduce

$$\begin{aligned}\mathbb{E}_t[\varphi_{1/\bar{\rho}}(x_{t+1})] &\leq \mathbb{E}_t\left[\varphi(\hat{x}_t) + \frac{\bar{\rho}}{2}\|x_{t+1} - \hat{x}_t\|^2\right] \\ &\leq \varphi(\hat{x}_t) + \frac{\bar{\rho}}{2}\cdot\mathbb{E}_t\left[\|x_{t+1} - \hat{x}_t\|^2\right], \\ &\leq \varphi(\hat{x}_t) + \frac{\bar{\rho}}{2}\left[\|\hat{x}_t - x_t\|^2 - \frac{\bar{\rho}-\tau-\eta}{\beta_t-\eta}\|\hat{x}_t - x_t\|^2 + \frac{4\mathsf{L}^2}{(\beta_t-\eta)(\beta_t-\bar{\rho})}\right] \\ &= \varphi_{1/\bar{\rho}}(x_t) - \frac{\bar{\rho}-\tau-\eta}{2\bar{\rho}(\beta_t-\eta)}\|\nabla\varphi_{1/\bar{\rho}}(x_t)\|^2 + \frac{2\bar{\rho}\mathsf{L}^2}{(\beta_t-\eta)(\beta_t-\bar{\rho})}.\end{aligned}$$

Taking expectations with respect to $\xi_0, \ldots, \xi_{t-1}$ and using the tower rule yields the claimed inequality (4.14). Unfolding the recursion (4.14) yields:

$$\mathbb{E}[\varphi_{1/\bar{\rho}}(x_{t+1})] \leq \varphi_{1/\bar{\rho}}(x_0) - \sum_{t=0}^{T}\left[\frac{\bar{\rho}-\tau-\eta}{2\bar{\rho}(\beta_t-\eta)}\mathbb{E}[\|\nabla\varphi_{1/\bar{\rho}}(x_t)\|^2]\right] + 2\bar{\rho}\mathsf{L}^2\cdot\sum_{t=0}^{T}\tfrac{1}{(\beta_t-\eta)(\beta_t-\bar{\rho})}.$$

Using the inequality $\varphi_{1/\bar{\rho}}(x_{t+1}) \geq \min\varphi$ and rearranging yields

$$\sum_{t=0}^{T}\frac{\bar{\rho}-\tau-\eta}{\beta_t-\eta}\mathbb{E}[\|\varphi_{1/\bar{\rho}}(x_t)\|^2] \leq 2\bar{\rho}(\varphi_{1/\bar{\rho}}(x_0) - \min\varphi) + 4\mathsf{L}^2\bar{\rho}^2\sum_{t=0}^{T}\frac{1}{(\beta_t-\eta)(\beta_t-\bar{\rho})}$$

Dividing through by $\sum_{t=0}^{T}\frac{\bar{\rho}-\tau-\eta}{\beta_t-\eta}$ and recognizing the left side as $\mathbb{E}[\|\varphi_{1/\bar{\rho}}(x_{t^*})\|^2]$ yields (4.15). Setting $\bar{\rho} = 2(\rho+\eta)$ and $\beta_t = \bar{\rho} + \sqrt{\frac{2\bar{\rho}\mathsf{L}^2(T+1)}{\Delta}}$ in (4.15) immediately yields the final guarantee (4.16). ∎

Next we consider the "convex setting", that is when the models $\mathbb{E}_\xi f(\cdot,\xi)$ globally lower abound f, without quadratic error, and the functions $f_x(\cdot,\xi)+r(\cdot)$ are μ-strongly convex. By analogy with the stochastic subgradient method, one would expect that Algorithm 4.1 drives the function gap $\mathbb{E}[\varphi(x_t) - \varphi(\cdot))]$ to zero at the rates $O(\frac{1}{\sqrt{t}})$ and $O(\frac{1}{\mu t})$, in the settings $\mu = 0$ and $\mu > 0$, respectively. The following two theorems establish exactly that. Even when specializing to the stochastic proximal subgradient method, Theorems 4.1 and 4.2 improve on the state of the art. In contrast to previous work [19, 32], the norms of the subgradients of r do not enter the complexity bounds established in Theorem 4.1, while Theorem 4.2 extends the nonuniform averaging technique of [62] for strongly convex minimization to the fully proximal setting.

THEOREM 4.1 (Convergence rate under convexity).
Suppose that $\tau = 0$ and the functions $f_x(\cdot,\xi)+r(\cdot)$ are convex. Let $\{x_t\}$ be the iterates generated by Algorithm 4.1 and set $\alpha_t = \beta_t^{-1}$. Then for all $T > 0$, we have

$$\mathbb{E}\left[\varphi\left(\tfrac{1}{\sum_{t=0}^{T}\alpha_t}\sum_{t=0}^{T}\alpha_t x_{t+1}\right) - \varphi(x^*)\right] \leq \frac{\frac{1}{2}\|x_0 - x^*\|^2 + \mathsf{L}^2\sum_{t=0}^{T}\alpha_t^2}{\sum_{t=0}^{T}\alpha_t}, \tag{4.17}$$

where x^* is any minimizer of φ. In particular, if Algorithm 3.1 uses the constant parameter $\alpha_t = \frac{D}{L\sqrt{2(T+1)}}$, for some real $D > \|x_0 - x^*\|$, then the estimate holds

$$\mathbb{E}\left[\varphi\left(\tfrac{1}{T+1}\sum_{t=1}^{T+1}x_t\right) - \varphi(x^*)\right] \leq \frac{\sqrt{2}LD}{\sqrt{T+1}}. \tag{4.18}$$

Proof. Setting $\eta := 0$ and $x := x^*$ in the estimate (4.8) in Lemma 4.2, and taking expectations of both sides yields

$$2\alpha_t \mathbb{E}[\varphi(x_{t+1}) - \varphi(x^*)] \leq \mathbb{E}\|x_t - x^*\|^2 - \mathbb{E}\left[\|x_{t+1} - x^*\|^2\right] + 2\mathsf{L}^2\alpha_t^2.$$

The estimate (4.17) then follows by summing across $t = 0, \ldots, T$, dividing through by $\sum_{t=0}^T \alpha_t$, and using convexity of φ. The estimate (4.18) is immediate from (4.17). ∎

The following theorem uses the nonuniform averaging technique from [62].

THEOREM 4.2 (Convergence rate under strong convexity). *Suppose that $\tau = 0$ and the functions $f_x(\cdot, \xi) + r(\cdot)$ are μ-strongly convex for some $\mu > 0$. Then for all $T > 0$, the iterates generated by Algorithm 4.1 with $\beta_t = \frac{\mu(t+1)}{2}$ satisfy*

$$\mathbb{E}\left[\varphi\left(\tfrac{2}{(T+2)(T+3)-2}\sum_{t=1}^{T+1}(t+1)x_t\right) - \varphi(x^*)\right] \leq \frac{\mu\|x_0 - x^*\|^2}{(T+2)^2} + \frac{8\mathsf{L}^2}{\mu(T+2)}.$$

where x^* is any minimizer of φ.

Proof. Define $\Delta_t := \frac{1}{2}\mathbb{E}[\|x - x_t\|^2]$. Setting $\eta := -\mu$ and $x := x^*$ in the estimate (4.8) of Lemma 4.2, taking expectations of both sides, and multiplying through by $(\beta_t + \mu)/2$ yields

$$\mathbb{E}[\varphi(x_{t+1}) - \varphi(x^*)] \leq \beta_t\Delta_t - (\beta_t + \mu)\Delta_{t+1} + \frac{\mathsf{L}^2}{\beta_t}.$$

Plugging in $\beta_t := \frac{\mu(t+1)}{2}$, multiplying through by $t+2$, and summing, we get

$$\begin{aligned}\sum_{t=0}^T (t+2)\mathbb{E}[\varphi(x_{t+1}) - \varphi(x^*)] &\leq \sum_{t=0}^T \left(\tfrac{\mu(t+1)(t+2)}{2}\Delta_t - \tfrac{\mu(t+2)(t+3)}{2}\Delta_{t+1}\right) + \sum_{t=0}^T \tfrac{2\mathsf{L}^2(t+2)}{\mu(t+1)} \\ &\leq \mu\Delta_0 + \frac{4\mathsf{L}^2(T+1)}{\mu}\end{aligned}$$

Dividing through by the sum $\sum_{t=0}^T (t+2) = \frac{(T+2)(T+3)}{2} - 1$ and using convexity of φ, we deduce

$$\begin{aligned}\mathbb{E}\left[\varphi\left(\tfrac{2}{(T+2)(T+3)-2}\sum_{t=0}^{T}(t+2)x_{t+1}\right) - \varphi(x^*)\right] &\leq \tfrac{\mu\|x_0 - x^*\|^2}{(T+2)(T+3)-2} + \tfrac{8\mathsf{L}^2(T+1)}{\mu((T+2)(T+3)-2)} \\ &\leq \frac{\mu\|x_0 - x^*\|^2}{(T+2)^2} + \frac{8\mathsf{L}^2}{\mu(T+2)}, \qquad (4.19)\end{aligned}$$

where (4.19) uses the estimate $(T+2)(T+3) - 2 \geq (T+2)^2$. The proof is complete. ∎

4.2. Algorithmic examples. We next look at the consequences of Theorem 4.3 and Theorem 4.1. We begin with the algorithms briefly mentioned in the introduction: stochastic proximal point, prox-linear, and proximal subgradient. In each case, we list the standard assumptions under which the methods are applicable, and then verify properties (B1)-(B4) for some $\tau, \eta, \mathsf{L} \geq 0$. Complexity guarantees for each method then follow immediately from Theorem 4.3. We then describe the problem of minimizing the expectation of a convex monotone composition (e.g. Conditional Value-at-Risk), and describe a natural model-based algorithm for the problem. Convergence guarantees in function values then follow from Theorem 4.1.

Stochastic proximal point. Consider the optimization problem (4.1) under the following assumptions.

(C1) It is possible to generate i.i.d. realizations $\xi_1, \xi_2, \ldots \sim P$.

(C2) There is an open convex set U containing $\operatorname{dom} r$ and a measurable function $(x, y, \xi) \mapsto f_x(y, \xi)$ defined on $U \times U \times \Omega$ satisfying $\mathbb{E}_\xi[f_x(y, \xi)] = f(y)$ for all $x, y \in U$.

(C3) Each function $r(\cdot) + f_x(\cdot, \xi)$ is ρ-weakly convex $\forall x \in U$, a.e. $\xi \in \Omega$.

(C4) There exists a measurable function $L\colon \Omega \to \mathbb{R}_+$ satisfying $\sqrt{\mathbb{E}_\xi\left[L(\xi)^2\right]} \leq \mathsf{L}$ and such that

$$f_x(x, \xi) - f_x(y, \xi) \leq L(\xi)\|x - y\|,$$

for all $x, y \in U$ and a.e. $\xi \in \Omega$.

The stochastic proximal point method is Algorithm 4.1 with the models $f_x(y, \xi)$. It is immediate to see that (B1)-(B4) hold with $\tau = 0$ and $\eta = \rho$.

Stochastic proximal subgradient. We next slightly loosen the assumptions (A1)-(A3) for the proximal stochastic subgradient method, by allowing r to be non-convex, and show how these assumptions imply (B1)-(B4). Consider the optimization problem (4.1), and let us assume that the following properties are true.

(D1) It is possible to generate i.i.d. realizations $\xi_1, \xi_2, \ldots \sim P$.

(D2) The function f is ρ_1-weakly convex and r is ρ_2-weakly convex, for some $\rho_1, \rho_2 \geq 0$.

(D3) There is an open convex set U containing $\operatorname{dom} r$ and a measurable mapping $G\colon U \times \Omega \to \mathbb{R}^d$ satisfying $\mathbb{E}_\xi[G(x, \xi)] \in \partial f(x)$ for all $x \in U$.

(D4) There is a real $\mathsf{L} \geq 0$ such that the inequality, $\mathbb{E}_\xi\left[\|G(x, \xi)\|^2\right] \leq \mathsf{L}^2$, holds for all $x \in U$.

The stochastic subgradient method is Algorithm 4.1 with the linear models

$$f_x(y, \xi) = f(x) + \langle G(x, \xi), y - x\rangle\,.$$

Observe that (B1) and (B3) with $\eta = \rho_2$ are immediate from the definitions; (B2) with $\tau = \rho_1$ follows from the discussion in [22, Section 2]. Assumption (B4) is also immediate from (D4).

Stochastic prox-linear. Consider the optimization problem (4.1) with

$$f(x) = \mathbb{E}_{\xi \sim P}\left[h\big(c(x, \xi), \xi\big)\right].$$

We assume that there exists an open convex set U containing $\operatorname{dom} r$ such that the following properties are true.

(E1) It is possible to generate i.i.d. realizations $\xi_1, \xi_2, \ldots \sim P$.

(E2) The assignments $h\colon \mathbb{R}^m \times \Omega \to \mathbb{R}$ and $c\colon U \times \Omega \to \mathbb{R}^m$ are measurable.

(E3) The function r is ρ-weakly convex, and there exist square integrable functions $\ell, \gamma, M\colon \Omega \to \mathbb{R}$ such that for a.e. $\xi \in \Omega$, the function $z \mapsto h(z, \xi)$ is convex and $\ell(\xi)$-Lipschitz, the map $x \mapsto c(x, \xi)$ is C^1-smooth with $\gamma(\xi)$-Lipschitz Jacobian, and the inequality, $\|\nabla c(x, \xi)\|_{\mathrm{op}} \leq M(\xi)$, holds for all $x \in U$ and a.e. $\xi \in \Omega$.

The stochastic prox-linear method [31] is Algorithm 4.1 with the convex models

$$f_x(y, \xi) = h\big(c(x, \xi) + \nabla c(x, \xi)(y - x), \xi\big).$$

Observe that (B1) and (B3) hold trivially with $\eta = \rho$. Assumption (B2) holds with $\tau = \sqrt{\mathbb{E}_\xi[\ell(\xi)]^2}\sqrt{\mathbb{E}_\xi[\gamma(\xi)^2]}$ by [31, Lemma 3.12]. Assumption (E3) also directly implies (B4) with $\mathsf{L} = \sqrt{\mathbb{E}_\xi[\ell(\xi)]^2}\sqrt{\mathbb{E}_\xi[M(\xi)^2]}$.

Expectation of convex monotone compositions. As an application of Theorem 4.1, suppose we wish to optimize the problem (4.1), where r is convex and f is given by

$$f(x) = \mathbb{E}_\xi[h(c(x,\xi),\xi)],$$

Suppose that $h(\cdot,\xi)\colon \mathbb{R}\to\mathbb{R}$ and $c(\cdot,\xi)\colon \mathbb{R}\to\mathbb{R}$ are convex, and $h(\cdot,\xi)$ is also nondecreasing. Note that we do not assume smoothness of $c(\cdot,x)$ and therefore this problem class does not fall with the composite framework discussed above.

We assume that there exists an open convex set U containing $\operatorname{dom} r$ such that the following properties are true.

- (F1) It is possible to generate i.i.d. realizations $\xi_1,\xi_2,\ldots\sim P$.
- (F2) The assignments $h\colon \mathbb{R}\times\Omega\to\mathbb{R}$ and $c\colon U\times\Omega\to\mathbb{R}^m$ are measurable, the functions $h(\cdot,\xi)$, $c(\cdot,\xi)$, and $r(\cdot)$ are convex, and $h(\cdot,\xi)$ is also nondecreasing.
- (F3) There is a measurable mapping $G\colon U\times\Omega\to\mathbb{R}^d$ satisfying $G(x,\xi)\in\partial_x c(x,\xi)$ for all $x\in U$.
- (F4) There exist square integrable functions $\ell, M\colon \Omega\to\mathbb{R}$ such that for a.e. $\xi\in\Omega$, the function $z\mapsto h(z,\xi)$ is $\ell(\xi)$-Lipschitz and the map $x\mapsto c(x,\xi)$ is $M(\xi)$-Lipschitz for a.e. $\xi\in\Omega$.

One reasonable class of models then reads:

$$f_x(y,\xi) = h(c(x,\xi)+\langle G(x,\xi), y-x\rangle,\xi).$$

Assumption (B1) is immediate from (F1). Assumption ($F2$) directly implies ($B2$) with $\tau=0$ and (B3) with $\eta=0$. Finally (F4) readily implies (B2) with $\mathsf{L} = \sqrt{\mathbb{E}_\xi[\ell(\xi)]^2}\sqrt{\mathbb{E}_\xi[M(\xi)^2]}$. Thus the stochastic model-based algorithm (Algorithm 4.1) enjoys the $O(\frac{1}{\sqrt{t}})$ convergence guarantee in expected function value gap (Theorem 4.1).

As an illustration, consider the Conditional Value-at-Risk problem, discussed in Example 2.6:

$$\min_{\gamma\in\mathbb{R},x\in\mathbb{R}^d}\ (1-\alpha)\gamma + \mathbb{E}_{\xi\sim P}[(g(x,\xi)-\gamma)^+] + r(x),$$

under the assumption that the loss $g(\cdot,\xi)$ is convex. Then given an iterate (x_t,γ_t), the stochastic model-based algorithm would sample $\xi_t\sim P$, choose a subgradient $v_t\in\partial_x g(x_t,\xi_t)$ and perform the simple update

$$(x_{t+1},\gamma_{t+1}) = \operatorname*{argmin}_{\gamma\in\mathbb{R},y\in\mathbb{R}^d}\ (1-\alpha)\gamma + \left[g(x_t,\xi_t)+\langle v_t, y-x_t\rangle-\gamma\right]^+ + r(y) + \frac{\beta_t}{2}(\|y-x_t\|^2+\|\gamma-\gamma_t\|^2).$$

5. Numerical Illustrations. In this section, we illustrate our three running examples (stochastic subgradient, prox-linear, prox-point) on phase retrieval and blind deconvolution problems, outlined in Section 2.1. In particular, our experiments complement the recent paper [31], which performs an extensive numerical study of the stochastic subgradient and prox-linear algorithms on the phase retrieval problem.

Our main goal in this section is to illustrate that the update rules for all three algorithms, have essentially the same computational cost. Indeed, the subproblems for the stochastic prox-point and prox-linear algorithms have a closed form solution. Note that our theoretical guarantees (Theorem 4.3) imply essentially the same worst-case complexity for the stochastic subgradient, prox-linear, and proximal point algorithms. In contrast, our numerical results on both problems clearly show that the latter two algorithms are much better empirically both in terms of speed and robustness to the

choice of stepsize. Intuitively, the reason appears to be that the models used by the latter two algorithms provide much tighter approximation. Indeed, the models are two-sided in the sense that the two-sided error $|\mathbb{E}_{\xi\sim P}[f_x(y,\xi)]-f(y)|$ is upper-bounded by a multiple of the quadratic $\|y-x\|^2$.

5.1. Phase retrieval. The experimental set-up for the phase retrieval problem is as follows. We generate standard Gaussian measurements $a_i \sim N(0, I_{d\times d})$, for $i = 1,\dots,m$; generate the target signal $\bar{x}$ and initial point x_0 uniformly on the unit sphere; and set $b_i = \langle a_i, \bar{x}\rangle^2$ for each $i = 1,\dots,m$. We then apply the three stochastic algorithms to the problem

$$\min_{x\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^m |\langle a_i, x\rangle^2 - b_i|.$$

Each step of the algorithms is trivial to implement. Since the three methods only use one data point at a time, let us define the function

$$g(x) = |\langle a, x\rangle^2 - b|,$$

for a fixed vector $a \in \mathbb{R}^d$ and a real $b \geq 0$.

Stochastic subgradient. The stochastic subgradient method simply needs to evaluate an element of the subdifferential

$$\partial g(x) = 2\langle a, x\rangle a \cdot \left\{ \begin{array}{ll} \operatorname{sign}(\langle a, x\rangle^2 - b), & \text{if } \langle a, x\rangle^2 \neq b \\ [-1, 1], & \text{o.w.} \end{array} \right\}.$$

Stochastic prox-linear. The stochastic prox-linear method needs to solve subproblems of the form

$$\operatorname*{argmin}_{\Delta\in\mathbb{R}^d} |\langle a, x\rangle^2 + 2\langle a, x\rangle\langle a, \Delta\rangle - b| + \frac{1}{2\lambda}\|\Delta\|^2.$$

Then the next iterate is defined to be $x + \Delta$. Setting $\gamma = \lambda(\langle a, x\rangle^2 - b)$ and $\zeta = 2\lambda\langle a, x\rangle a$, we therefore seek to solve the problem

$$\operatorname*{argmin}_{\Delta\in\mathbb{R}^d} |\gamma + \langle\zeta, \Delta\rangle| + \frac{1}{2}\|\Delta\|^2. \tag{5.1}$$

An explicit solution Δ^* to this subproblem follows from a standard Lagrangian calculation, and is recorded for example in [31, Section 4]:

$$\Delta^* = \operatorname{proj}_{[-1,1]}\left(\frac{-\gamma}{\|\zeta\|^2}\right)\zeta. \tag{5.2}$$

Stochastic proximal point. Finally, the stochastic proximal point method requires solving the problem

$$\operatorname*{argmin}_{y} |\langle a, y\rangle^2 - b| + \frac{1}{2\lambda}\|y - x\|^2 \tag{5.3}$$

Let us compute the candidate solutions using first-order optimality conditions:

$$\lambda^{-1}(x - y) \in 2\langle a, y\rangle a \cdot \left\{ \begin{array}{ll} \operatorname{sign}(\langle a, y\rangle^2 - b), & \text{if } \langle a, y\rangle^2 \neq b \\ [-1, 1], & \text{o.w.} \end{array} \right\}. \tag{5.4}$$

An easy computation shows that there are at most four point y that satisfy (5.4):

$$\left\{x - \left(\frac{2\lambda\langle a, x\rangle}{2\lambda\|a\|^2 \pm 1}\right) a,\ x - \left(\frac{\langle a, x\rangle \pm \sqrt{b}}{\|a\|^2}\right) a\right\}.$$

Therefore we may set the next iterate to be the candidate solution y with the lowest function value for the subproblem (5.3).

We perform three sets of experiments corresponding to $(d, m) = (10, 30)$, $(50, 150)$, $(100, 300)$, and record the result in Figure 3. The dashed blue line indicates the initial functional error. In each set of experiments, we use 100 equally spaced step-size parameters β_t^{-1} between 10^{-4} and 1. The figures on the left record the function gap after 100 passes through the data, averaged over 15 rounds. The figures on the right output the number of epochs used by the stochastic prox-linear and proximal point methods to find a point achieving 10^{-4} functional suboptimality, averaged over 15 rounds. It is clear from the figures that the stochastic prox-linear and proximal point algorithms perform much better and are more robust to the choice of the step-size parameter than the stochastic subgradient method.

5.2. Blind deconvolution. We next consider a problem inspired by blind deconvolution and biconvex compressive sensing [41]. The experimental set-up is as follows. We generate Gaussian measurements $u_i \sim N(0, I_{d_1 \times d_1})$ and $v_i \sim N(0, I_{d_2 \times d_2})$, for $i = 1, \ldots, m$; generate the target signal $\bar{x}$ uniformly on the unit sphere; and set $b_i = \langle u_i, \bar{x}\rangle\langle v_i, \bar{x}\rangle$ for each $i = 1, \ldots, m$. The problem formulation reads:

$$\min_{x,y}\ \frac{1}{m}\sum_{i=1}^{m} |\langle u_i, x\rangle\langle v_i, y\rangle - b_i|,$$

Again, since the three methods access one data point at a time, define the function

$$g(x, y) = |\langle u, x\rangle\langle v, y\rangle - b|$$

for some vectors $u \in \mathbb{R}_1^d$ and $v \in \mathbb{R}_2^d$ and real $b \in \mathbb{R}$.

Stochastic subgradient. The stochastic subgradient method, in each iteration, evaluates an element of the subdifferential

$$\partial g(x, y) = (\langle v, y\rangle u, \langle u, x\rangle v) \cdot \left\{ \begin{array}{ll} \operatorname{sign}(\langle u, x\rangle\langle v, y\rangle - b), & \text{if } \langle u, x\rangle\langle v, y\rangle \neq b \\ [-1, 1], & \text{o.w.} \end{array} \right\}.$$

Stochastic prox-linear. The stochastic prox-linear method needs to solve subproblems of the form:

$$\operatorname*{argmin}_{\Delta_1, \Delta_2}\ |\langle u, x\rangle\langle v, y\rangle + \langle v, y\rangle\langle u, \Delta_1\rangle + \langle u, x\rangle\langle v, \Delta_2\rangle - b| + \frac{1}{2\lambda}(\|\Delta_1\|^2 + \|\Delta_2\|^2).$$

Once a solution (Δ_1, Δ_2) is found, the next iterate is $(x + \Delta_1, y + \Delta_2)$. Clearly, we may rewrite the prox-linear subproblem in the form (5.1) under the identification $\Delta = (\Delta_1, \Delta_2)$, $\zeta = \lambda(\langle v, y\rangle u, \langle u, x\rangle v)$, and $\gamma = \lambda(\langle u, x\rangle\langle v, y\rangle - b)$. We may then read off the solution directly from (5.2).

Stochastic proximal point. Finally, the stochastic proximal point method requires solving the problem

$$\operatorname*{argmin}_{x,y}\ |\langle u, x\rangle\langle v, y\rangle - b| + \frac{1}{2\lambda}\|x - x_0\|^2 + \frac{1}{2\lambda}\|y - y_0\|^2. \tag{5.5}$$

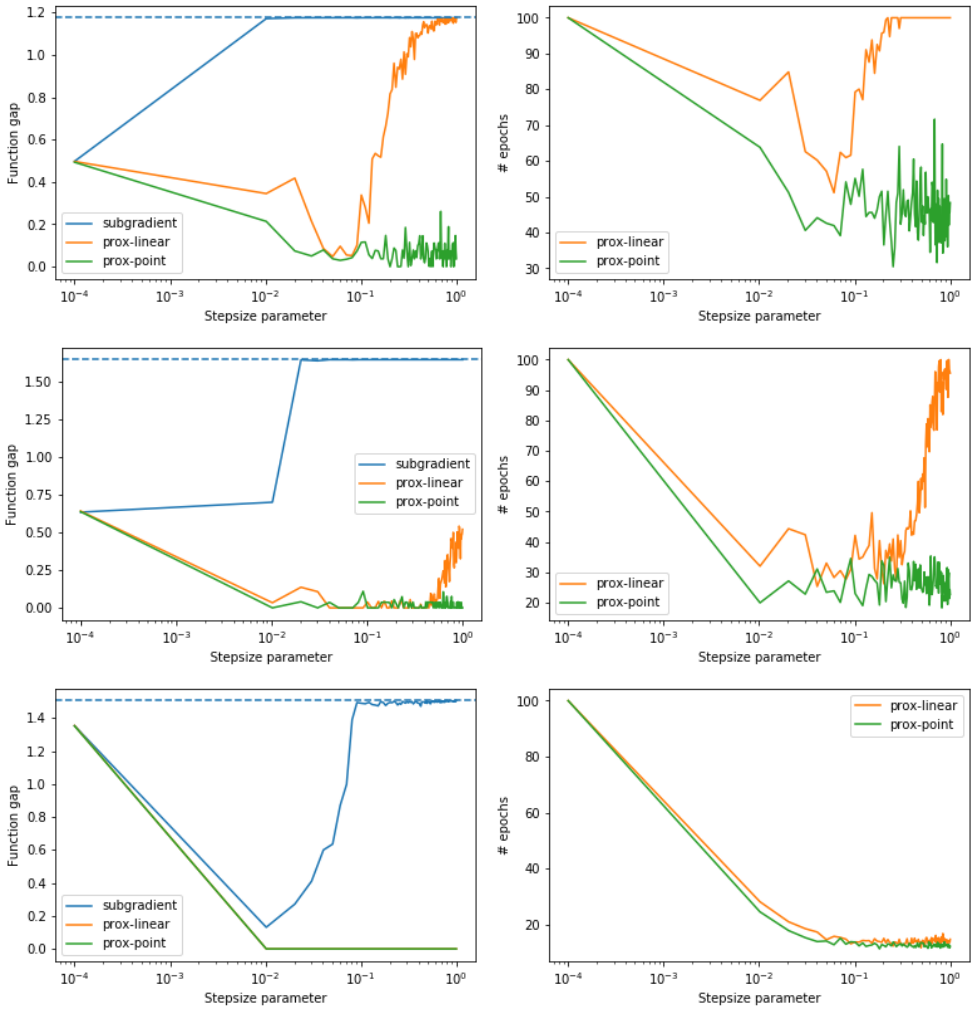


Fig. 3: Bottom to top: $(d, m) = (10, 30), (50, 150), (100, 300)$. The dashed blue line indicates the initial functional error.

Let us enumerate the critical points. Writing out the optimality conditions for (x, y), there are two cases to consider. In the first case $\langle u, x\rangle\langle v, y\rangle \neq b$, it is straightforward to show that the possible critical point have the form

$$\begin{aligned} x &= x_0 - \lambda\left(\frac{\pm\langle v, y_0\rangle - \lambda\|v\|^2\langle u, x_0\rangle}{1-\lambda^2\|u\|^2\|v\|^2}\right)u \\ y &= y_0 - \lambda\left(\frac{\pm\langle u, x_0\rangle - \lambda\|u\|^2\langle v, y_0\rangle}{1-\lambda^2\|u\|^2\|v\|^2}\right)v \end{aligned} \tag{5.6}$$

Indeed, suppose for the moment $\langle u, x\rangle\langle v, y\rangle > b$. Then optimality conditions for (5.5) imply

$$x = x_0 - \lambda\langle v, y\rangle u, \qquad y = y_0 - \lambda\langle u, x\rangle v \tag{5.7}$$

Thus if we determine $\langle v, y\rangle$ and $\langle u, x\rangle$, we will have an explicit formula for (x, y). Taking the dot product of the first equation with u and the second with v yields

$$\lambda\langle v, y\rangle\|u\|^2 = \langle u, x_0\rangle - \langle u, x\rangle, \qquad \lambda\langle u, x\rangle\|v\|^2 = \langle v, y_0\rangle - \langle v, y\rangle.$$

Solving for $\langle v, y\rangle$ and $\langle u, x\rangle$, we get

$$\langle u, x\rangle = \frac{\langle u, x_0\rangle - \lambda\|u\|^2\langle v, y_0\rangle}{1 - \lambda^2\|u\|^2\|v\|^2}, \qquad \langle v, y\rangle = \frac{\langle v, y_0\rangle - \lambda\|v\|^2\langle u, x_0\rangle}{1 - \lambda^2\|u\|^2\|v\|^2}.$$

Combining these expressions with (5.7), we deduce that x and y can be expressed as in (5.6). The setting $\langle u, x\rangle\langle v, y\rangle > b$ is completely analogous.

In the second case, suppose $\langle u, x\rangle\langle v, y\rangle = b$. Then optimality condition for (5.5) imply that there exists γ such that

$$x = x_0 - \gamma\langle v, y\rangle u, \qquad y = y_0 - \gamma\langle u, x\rangle v, \qquad b = \langle u, x\rangle\langle v, y\rangle.$$

We must solve this system of equations for γ, $\eta := \langle u, x\rangle$, and $\delta := \langle v, y\rangle$. Substituting the third equation into the first yields:

$$x = x_0 - \gamma\left(\frac{b}{\eta}\right)u, \qquad y = y_0 - \gamma\eta v, \qquad b = \eta\delta. \tag{5.8}$$

Taking the dot product of the first equation with u and the second with v yields

$$\eta = \langle u, x_0\rangle - \gamma\left(\frac{b}{\eta}\right)\|u\|^2, \qquad \frac{b}{\eta} = \langle v, y_0\rangle - \gamma\eta\|v\|^2.$$

Solving the first equation for γ, we get the expression $\gamma = \frac{\eta\langle u, x_0\rangle - \eta^2}{b\|u\|^2}$. Plugging this formula into the second equation and clearing the denominator, we arrive at the quartic polynomial

$$0 = \eta^4\|v\|^2 - \eta^3\|v\|^2\langle u, x_0\rangle + b\eta\|u\|^2\langle v, y_0\rangle - b^2\|u\|^2.$$

Thus after finding each root η, we may set $\gamma = \frac{\eta\langle u, x_0\rangle - \eta^2}{b\|u\|^2}$, and then obtain an explicit formula for (x, y) using (5.8).

Our numerical experiments are similar to those for phase retrieval. We perform three sets of experiments corresponding to $(d_1, d_2, m) = (10, 10, 30)$, $(50, 50, 200)$, $(100, 100, 400)$, and record the result in Figure 4. The dashed blue line indicates the initial functional error. In each set of experiments, we use 100 equally spaced step-size parameters β_t^{-1} between 10^{-4} and 1. The figures on the left record the function gap after 100 passes through the data, averaged over 10 rounds. The figures on the right output the number of epochs used by the stochastic prox-linear and proximal point methods to find a point achieving 10^{-4} functional suboptimality, averaged over 10 rounds. As in phase retrieval, it is clear from the figure that the stochastic prox-linear and proximal point algorithms perform much better and are more robust to the choice of the step-size parameter than the stochastic subgradient method.

Acknowledgments. The authors thank John Duchi for his careful reading and helpful feedback on the initial version of this manuscript.

REFERENCES

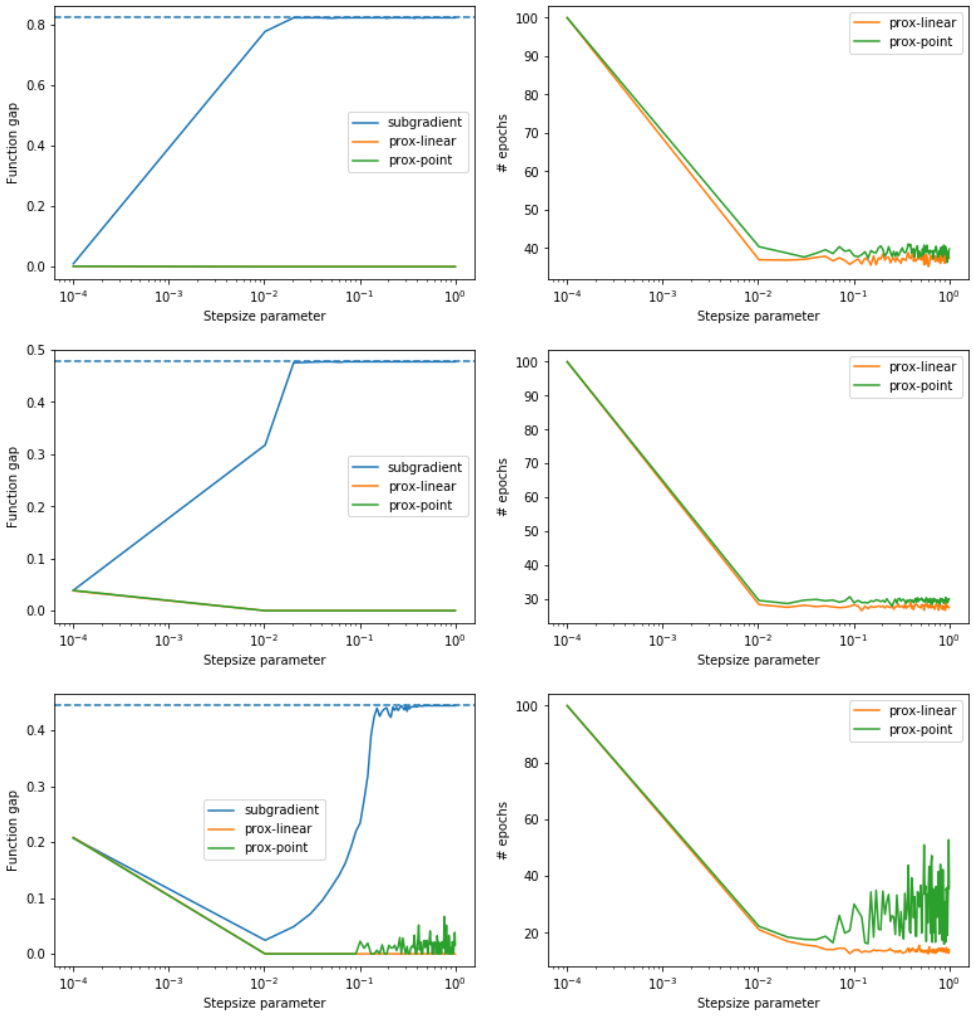


Fig. 4: Bottom to top: $(d_1, d_2, m) = (10, 10, 50), (50, 50, 200), (100, 100, 400)$. The dashed blue line indicates the initial functional error.

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, and et al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015, https://www.tensorflow.org/. Software available from tensorflow.org.

[2] E. Abbe, A. Bandeira, A. Bracher, and A. Singer, *Decoding binary node labels from censored edge measurements: phase transition and efficient recovery*, IEEE Trans. Network Sci. Eng., 1 (2014), pp. 10–22, https://doi.org/10.1109/TNSE.2014.2368716.

[3] Z. Allen-Zhu, *Katyusha: The First Direct Acceleration of Stochastic Gradient Methods*, in STOC, 2017.

[4] Z. Allen-Zhu, *Natasha 2: Faster non-convex optimization than sgd*, arXiv preprint arXiv:1708.08694, (2017).

[5] Z. Allen-Zhu, *How to make gradients small stochastically*, Preprint arXiv:1801.02982 (version 1), (2018).

[6] A. Bandeira, N. Boumal, and V. Voroninski, *On the low-rank approach for semidefinite programs arising in synchronization and community detection*, in Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, June 23-26, 2016, 2016, pp. 361–

382, http://jmlr.org/proceedings/papers/v49/bandeira16.html.

[7] A. Ben-Tal and M. Teboulle, *Expected utility, penalty functions, and duality in stochastic nonlinear programming*, Manage. Sci., 32 (1986), pp. 1445–1466, https://doi.org/10.1287/mnsc.32.11.1445, http://dx.doi.org/10.1287/mnsc.32.11.1445.

[8] A. Ben-Tal and M. Teboulle, *An old-new concept of convex risk measures: the optimized certainty equivalent*, Math. Finance, 17 (2007), pp. 449–476, https://doi.org/10.1111/j.1467-9965.2007.00311.x, https://doi-org.offcampus.lib.washington.edu/10.1111/j.1467-9965.2007.00311.x.

[9] P. Bianchi, *Ergodic convergence of a stochastic proximal point algorithm*, SIAM Journal on Optimization, 26 (2016), pp. 2235–2260.

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

[11] E. Candès, X. Li, Y. Ma, and J. Wright, *Robust principal component analysis?*, J. ACM, 58 (2011), pp. Art. 11, 37, https://doi.org/10.1145/1970392.1970395, http://dx.doi.org/10.1145/1970392.1970395.

[12] E. Candès, X. Li, and M. Soltanolkotabi, *Phase retrieval via Wirtinger flow: theory and algorithms*, IEEE Trans. Inform. Theory, 61 (2015), pp. 1985–2007, https://doi.org/10.1109/TIT.2015.2399924.

[13] T. F. Chan and C.-K. Wong, *Total variation blind deconvolution*, IEEE Transactions on Image Processing, 7 (1998), pp. 370–375, https://doi.org/10.1109/83.661187.

[14] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. Willsky, *Rank-sparsity incoherence for matrix decomposition*, SIAM J. Optim., 21 (2011), pp. 572–596, https://doi.org/10.1137/090761793, http://dx.doi.org/10.1137/090761793.

[15] Y. Chen and E. Candès, *Solving random quadratic systems of equations is nearly as easy as solving linear systems*, Comm. Pure Appl. Math., 70 (2017), pp. 822–883, https://doi-org.offcampus.lib.washington.edu/10.1002/cpa.21638.

[16] Y. Chen, Y. Chi, and A. Goldsmith, *Exact and stable covariance estimation from quadratic sampling via convex programming*, IEEE Transactions on Information Theory, 61 (2015), pp. 4034–4059.

[17] Y. Chen, Y. Chi, and A. J. Goldsmith, *Exact and stable covariance estimation from quadratic sampling via convex programming*, IEEE Trans. Inform. Theory, 61 (2015), pp. 4034–4059, https://doi.org/10.1109/TIT.2015.2429594, https://doi-org.offcampus.lib.washington.edu/10.1109/TIT.2015.2429594.

[18] F. Clarke, Y. Ledyaev, R. Stern, and P. Wolenski, *Nonsmooth Analysis and Control Theory*, Texts in Math. 178, Springer, New York, 1998.

[19] J. Cruz, *On proximal subgradient splitting method for minimizing the sum of two nonsmooth convex functions*, Set-Valued Var. Anal., 25 (2017), pp. 245–263, https://doi.org/10.1007/s11228-016-0376-5.

[20] D. Davis and D. Drusvyatskiy, *Complexity of finding near-stationary points of convex functions stochastically*, arXiv:1802.08556, (2018).

[21] D. Davis and D. Drusvyatskiy, *Stochastic model-based minimization of weakly convex functions*, Preprint arXiv:1803.06523, (2018).

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

[23] D. Davis, D. Drusvyatskiy, and C. Paquette, *The nonsmooth landscape of phase retrieval*, Preprint arXiv:1711.03247, (2017).

[24] D. Davis and B. Grimmer, *Proximally guided stochastic method for nonsmooth, nonconvex problems*, Preprint arXiv:1707.03505, (2017).

[25] O. Devolder, F. Glineur, and Y. Nesterov, *First-order methods of smooth convex optimization with inexact oracle*, Mathematical Programming, 146 (2014), pp. 37–75, https://doi.org/10.1007/s10107-013-0677-5, https://doi.org/10.1007/s10107-013-0677-5.

[26] D. Drusvyatskiy, *The proximal point method revisited*, To appear in SIAG/OPT Views and News, arXiv:1712.06038, (2018).

[27] D. Drusvyatskiy, A. Ioffe, and A. Lewis, *Nonsmooth optimization using taylor-like models: error bounds, convergence, and termination criteria*, Preprint arXiv:1610.03446, (2016).

[28] D. Drusvyatskiy and A. Lewis, *Error bounds, quadratic growth, and linear convergence of proximal methods*, To appear in Math. Oper. Res., arXiv:1602.06661, (2016).

[29] D. Drusvyatskiy and C. Paquette, *Efficiency of minimizing compositions of convex functions and smooth maps*, Mathematical Programming, (2018).

[30] J. Duchi and F. Ruan, *Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval*, Preprint arXiv:1705.02356, (2017).

[31] J. Duchi and F. Ruan, *Stochastic methods for composite optimization problems*, Preprint

arXiv:1703.08570, (2017).
[32] J. Duchi and Y. Singer, *Efficient online and batch learning using forward backward splitting*, J. Mach. Learn. Res., 10 (2009), pp. 2899–2934.
[33] P. Dvurechensky, *Gradient method with inexact oracle for composite non-convex optimization*, arXiv:1703.09180, (2017).
[34] Y. Eldar and S. Mendelson, *Phase retrieval: stability and recovery guarantees*, Appl. Comput. Harmon. Anal., 36 (2014), pp. 473–494, https://doi.org/10.1016/j.acha.2013.08.003, http://dx.doi.org/10.1016/j.acha.2013.08.003.
[35] S. Ghadimi and G. Lan, *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*, SIAM J. Optim., 23 (2013), pp. 2341–2368.
[36] S. Ghadimi, G. Lan, and H. Zhang, *Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization*, Math. Program., 155 (2016), pp. 267–305.
[37] A. Juditsky and A. Nemirovski, *First order methods for nonsmooth convex large-scale optimization, I: General purpose methods*, in Optimization for Machine Learning, S. Sra, S. Nowozin, and S. Write, eds., MIT Press, 2011, ch. 1, pp. 266–290.
[38] L. Lei, C. Ju, J. Chen, and M. I. Jordan, *Non-convex finite-sum optimization via scsg methods*, in Advances in Neural Information Processing Systems, 2017, pp. 2345–2355.
[39] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, *Understanding blind deconvolution algorithms*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 33 (2011), pp. 2354–2367, https://doi.org/10.1109/TPAMI.2011.148.
[40] A. Lewis and S. Wright, *A proximal method for composite minimization*, Math. Program., (2015), pp. 1–46, https://doi.org/10.1007/s10107-015-0943-9, http://dx.doi.org/10.1007/s10107-015-0943-9.
[41] S. Ling and T. Strohmer, *Self-calibration and biconvex compressive sensing*, Inverse Problems, 31 (2015), pp. 115002, 31, https://doi.org/10.1088/0266-5611/31/11/115002, https://doi.org/10.1088/0266-5611/31/11/115002.
[42] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. Bach, *Supervised dictionary learning*, in Advances in Neural Information Processing Systems 21, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds., Curran Associates, Inc., 2009, pp. 1033–1040, http://papers.nips.cc/paper/3448-supervised-dictionary-learning.pdf.
[43] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299, http://www.numdam.org.offcampus.lib.washington.edu/item?id=BSMF_1965__93__273_0.
[44] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.
[45] A. Nemirovsky and D. Yudin, *Problem complexity and method efficiency in optimization*, A Wiley-Interscience Publication, John Wiley & Sons, Inc., New York, 1983.
[46] Y. Nesterov, *How to make the gradients small*, OPTIMA, MPS Newsletter, (2012), pp. 10–11.
[47] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161, https://doi.org/10.1007/s10107-012-0629-5, http://dx.doi.org/10.1007/s10107-012-0629-5.
[48] Y. Nesterov, *Universal gradient methods for convex optimization problems*, Mathematical Programming, 152 (2015), pp. 381–404.
[49] E. Nurminskii, *Minimization of nondifferentiable functions in the presence of noise*, Cybernetics, 10 (1974), pp. 619–621, https://doi.org/10.1007/BF01071541, https://doi.org/10.1007/BF01071541.
[50] E. A. Nurminskii, *The quasigradient method for the solving of the nonlinear programming problems*, Cybernetics, 9 (1973), pp. 145–150, https://doi.org/10.1007/BF01068677, https://doi.org/10.1007/BF01068677.
[51] R. Poliquin and R. Rockafellar, *Amenable functions in optimization*, in Nonsmooth optimization: methods and applications (Erice, 1991), Gordon and Breach, Montreux, 1992, pp. 338–353.
[52] R. Poliquin and R. Rockafellar, *Prox-regular functions in variational analysis*, Trans. Amer. Math. Soc., 348 (1996), pp. 1805–1838.
[53] S. J. Reddi, S. Sra, B. Poczos, and A. J. Smola, *Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization*, in Advances in Neural Information Processing Systems, 2016, pp. 1145–1153.
[54] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statistics, 22 (1951), pp. 400–407.
[55] R. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.
[56] R. Rockafellar, *Favorable classes of Lipschitz-continuous functions in subgradient optimization*, in Progress in nondifferentiable optimization, vol. 8 of IIASA Collaborative Proc. Ser.

CP-82, Int. Inst. Appl. Sys. Anal., Laxenburg, 1982, pp. 125–143.

[57] R. Rockafellar and S. Uryasev, *The fundamental risk quadrangle in risk management, optimization and statistical estimation*, Surveys in Operations Research and Management Science, 18 (2013), pp. 33 – 53, https://doi.org/https://doi.org/10.1016/j.sorms.2013.03.001, http://www.sciencedirect.com/science/article/pii/S1876735413000032.

[58] R. Rockafellar and R.-B. Wets, *Variational Analysis*, Grundlehren der mathematischen Wissenschaften, Vol 317, Springer, Berlin, 1998.

[59] R. T. Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, N.J., 1970.

[60] R. T. Rockafellar and S. Uryasev, *Optimization of conditional value-at-risk*, Journal of Risk, 2 (2000), pp. 21–41.

[61] E. Ryu and S. Boyd, *Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent*, Preprint www.math.ucla.edu/∼eryu/.

[62] M. Schmidt, N. L. Roux, and F. Bach, *A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method*, arXiv:1212.2002, (2013).

[63] A. Singer, *Angular synchronization by eigenvectors and semidefinite programming*, Appl. Comput. Harmon. Anal., 30 (2011), pp. 20–36, https://doi.org/10.1016/j.acha.2010.02.001.

[64] J. Sun, Q. Qu, and J. Wright, *A geometric analysis of phase retrieval*, To appear in Found. Comp. Math., arXiv:1602.06664, (2017).

[65] I. Tosic and P. Frossard, *Dictionary learning*, IEEE Signal Processing Magazine, 28 (2011), pp. 27–38, https://doi.org/10.1109/MSP.2010.939537.

[66] P. Toulis and E. Airoldi, *Asymptotic and finite-sample properties of estimators based on stochastic gradients*, Ann. Statist., 45 (2017), pp. 1694–1727, https://doi.org/10.1214/16-AOS1506, https://doi-org.offcampus.lib.washington.edu/10.1214/16-AOS1506.

[67] P. Toulis, T. Horel, and E. Airoldi, *Stable Robbins-Monro approximations through stochastic proximal updates*, arXiv:1510.00967, (2015).

[68] M. Wang, E. Fang, and H. Liu, *Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions*, Math. Program., 161 (2017), pp. 419–449, https://doi.org/10.1007/s10107-016-1017-3, https://doi.org/10.1007/s10107-016-1017-3.

[69] Y. Xu and W. Yin, *Block stochastic gradient iteration for convex and nonconvex optimization*, SIAM J. Optim., 25 (2015), pp. 1686–1716.