Stochastic model-based minimization of weakly convex functions

Damek Davis* Dmitriy Drusvyatskiy†

Abstract

We consider an algorithm that successively samples and minimizes stochastic models of the objective function. We show that under weak-convexity and Lipschitz conditions, the algorithm drives the expected norm of the gradient of the Moreau envelope to zero at the rate $O(k^{-1/4})$. Our result yields the first complexity guarantees for the stochastic proximal point algorithm on weakly convex problems and for the stochastic prox-linear algorithm for minimizing compositions of convex functions with smooth maps. Our general framework also recovers the recently obtained complexity estimate for the stochastic proximal subgradient method on weakly convex problems.

1 Introduction

Numerous algorithms for minimizing a function g on $\mathbb{R}^d$ can be written in the form:

$$x_{t+1} = \operatorname*{argmin}_y \left\{ g_{x_t}(y) + \frac{\beta_t}{2}\|y - x_t\|^2 \right\}, \tag{1.1}$$

where $g_{x_t}(\cdot)$ is a "simple model" of g formed at the current iterate x_t and $\beta_t > 0$ is a control sequence balancing the fidelity of the model with proximity to x_t. Typical algorithms of this type use models $g_x(\cdot)$ that at the very least satisfy the properties:

$$g_x(x) = g(x) \qquad \text{and} \qquad g_x(y) - g(y) \leq \frac{\tau}{2}\|y - x\|^2 \qquad \forall x, y \in \mathbb{R}^d. \tag{1.2}$$

where $\tau > 0$ is some fixed real number. Thus one requires the models $g_x(\cdot)$ to agree with g at x and to lower-bound g up to a quadratic error relative to the base-point x. To simplify notation, we will call any map $(x, y) \mapsto g_x(y)$ satisfying (1.2) a *one-sided model of* g.

Let us look at an example motivating the rest of our discussion. Consider an optimization problem of the form

$$\min_{x \in \mathbb{R}^d} \; g(x) = h(c(x)), \tag{1.3}$$

*School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850, USA; people.orie.cornell.edu/dsd95/.

†Department of Mathematics, U. Washington, Seattle, WA 98195; www.math.washington.edu/~ddrusv.
Research of Drusvyatskiy was supported by the AFOSR YIP award FA9550-15-1-0237 and by the NSF DMS 1651851 and CCF 1740551 awards.

with h convex and ℓ-Lipschitz and c a smooth map with γ-Lipschitz Jacobian. Such composite problems appear often in computation science. Nonlinear least squares [15, Section 10.3] and exact penalty formulations of nonlinear programs [15, Section 17.2] are classical examples, while notable contemporary instances include robust phase retrieval [6, 10], covariance matrix estimation [3, 5], and matrix factorization problems such as NMF [12, 13].

The *subgradient* and the *prox-linear* methods are two influential model-based algorithms for the problem class (1.3). The subgradient method performs the update (1.1) using the linear model

$$g_x(y) = g(x) + \langle \nabla c(x)^T v, y - x \rangle,$$

for an arbitrary subgradient selection $v \in \partial h(c(x))$. Equivalently, each iteration written in closed form is

$$\text{choose } x_{t+1} \in x_t - \frac{1}{\beta_t} \nabla c(x_t)^T \partial h(c(x_t)).$$

An easy argument shows that the assignment $(x, y) \mapsto g_x(y)$ is indeed a one-sided model for g with $\tau = \ell\gamma$; see e.g. [9, Lemma 4.2].

The prox-linear algorithm, in contrast, uses convex models with stronger approximation guarantees, while possibly incurring a higher per iteration cost. Namely, the prox-linear method performs the update (1.1) using the nonlinear convex models

$$g_x(y) = h\big(c(x) + \nabla c(x)(y - x)\big).$$

Notice in contrast to the subgradient method, the prox-linear algorithm in each iteration requires solving an auxiliary convex subproblem:

$$x_{t+1} = \operatorname*{argmin}_y \left\{ h\big(c(x_t) + \nabla c(x_t)(y - x_t)\big) + \frac{\beta_t}{2}\|y - x_t\|^2 \right\}.$$

The assignment $(x, y) \mapsto g_x(y)$ is again a one-sided model with $\tau = \ell\gamma$. Indeed, the prox-linear model satisfies the much stronger two-sided estimate $|g_x(y) - g(y)| \leq \frac{\tau}{2}\|y - x\|^2$ for all $x, y \in \mathbb{R}^d$; see e.g., [9, Section 4.1]. For a historical account of the prox-linear method, see e.g., [2, 14] and the references therein. For a systematic study of two-sided models in optimization, see [8].

Recent literature [4, 7, 9] has identified the gradient of the Moreau envelope of g as a natural measure of stationarity on which to base comparison of algorithms for the composite problem class (1.3). We will review this notion in detail in Section 2. In terms of this quantity, the subgradient method and the prox-linear algorithm have complexity $O(\varepsilon^{-4})$ and $O(\varepsilon^{-2})$, respectively, under an appropriate choice of the control sequences β_t. The superior iteration complexity of the prox-linear method is not surprising since the prox-linear models satisfy a much stronger approximation guarantee. The caveat is of course that the per iteration cost of the prox-linear method can be higher than that of the subgradient method, since each iteration requires solving an auxiliary convex problem.

Though the outlined model-based techniques are appealingly simple and largely classical, computing exact one-sided models of g is often prohibitively expensive. This is especially so for the huge-scale problems that now routinely appear in practice. Instead, in our current work, we suppose that for every point $x \in \mathbb{R}^d$ there is a family of models $g_x(\cdot, \xi)$, indexed

by a random variable ξ that follows a probability distribution P. We will only require that the models $g_x(\cdot, \xi)$ satisfy (1.2) in expectation. More formally, we will call the assignment $(x, y, \xi) \mapsto g_x(y, \xi)$ a *stochastic one-sided model* of g whenever the expectation $(x, y) \mapsto \mathbb{E}_\xi[g_x(y, \xi)]$ is a (deterministic) one-sided model of g. The algorithm we propose simply iterates the steps

$$\boxed{\begin{array}{l}\text{Sample } \xi_t \sim P, \\ \text{Set } x_{t+1} = \operatorname*{argmin}_y \left\{ g_{x_t}(y, \xi_t) + \dfrac{\beta_t}{2}\|y - x_t\|^2 \right\}.\end{array}} \tag{1.4}$$

The last ingredient we require is that for every ξ and $x \in \mathbb{R}^d$, the models $g_x(\cdot, \xi)$ are ρ-*weakly convex*, by which we mean that the assignment $y \mapsto g_x(y, \xi) + \frac{\rho}{2}\|y\|^2$ is convex. This assumption is indeed natural since we would like the subproblem defining x_{t+1} to be strongly convex, and hence for x_{t+1} to be uniquely determined. Weakly convex functions have recently found a number of interesting applications in large-scale optimization; in particular, the composite function (1.3) is $\ell\gamma$-weakly convex. We will prove that under standard measurability and Lipschitz conditions and with an appropriate choice of the control sequence β_t, the generic algorithm (1.4) has complexity $O(\varepsilon^{-4})$ in expectation.

Let us look at some concrete applications of our result. In the setting that for all $x \in \mathbb{R}^d$, the expectation $\mathbb{E}_\xi[g_x(\cdot, \xi)]$ agrees with $g(\cdot)$, the algorithm (1.4) reverts to the stochastic proximal point method. This algorithm was recently considered for convex minimization in [19] and extended to monotone inclusion problems in [1]. Thus we obtain a new complexity guarantee of $O(\varepsilon^{-4})$ for the stochastic proximal point algorithm on weakly convex problems.

As the next application, we return to the running example (1.3). The recent paper [11] investigated the following stochastic composite optimization problem:

$$\min_{x \in \mathbb{R}^d} \; g(x) = \mathbb{E}_{\xi \sim P}[h\big(c(x, \xi), \xi\big)], \tag{1.5}$$

Define the family of stochastic linear models

$$g_x(y, \xi) = g(x) + \langle \nabla c(x, \xi)^T w(x, \xi), y - x \rangle,$$

where $w(x, \xi) \in \partial h(c(x, \xi), \xi)$ is a measurable subgradient selection. Then each iteration of Algorithm 1.4 reduces to the stochastic subgradient update

$$\left\{ \begin{array}{l} \text{Sample } \xi_t \sim P \\ \text{Choose } x_{t+1} \in x_t - \dfrac{1}{\beta_t} \nabla c(x_t, \xi_t)^T \partial h\big(c(x_t, \xi_t), \xi_t\big) \end{array} \right\}.$$

Under mild technical conditions, the map $(x, y, \xi) \mapsto g_x(y, \xi)$ is indeed a stochastic one-sided model of g, and therefore we again can conclude the $O(\varepsilon^{-4})$ complexity. The same complexity guarantee was recently obtained in [4] for stochastic subgradient methods for an even larger class of weakly convex problems. That being said, the argument in [4] does not extend to the algorithm (1.4) in its full generality. Conversely, the assumptions presented here do not fully capture the noise model of stochastic proximal *gradient* methods developed

in [4]. Consequently, our current techniques are distinct from those in [4], though the idea of using the Moreau envelope as the potential function remains the same.

As the final example, let us look at the stochastic prox-linear method, introduced in [11]. In each iteration, the method performs the update (1.4) using the stochastic models

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

Thus the stochastic prox-linear method iterates the steps

$$\left\{\begin{array}{l}\text{Sample } \xi_t \sim P\\ \text{Set } x_{t+1} = \operatorname*{argmin}_y \ \{h\big(c(x_t,\xi_t) + \nabla c(x_t,\xi_t)(y-x_t),\xi_t\big) + \frac{\beta_t}{2}\|y-x_t\|^2\}\end{array}\right\}.$$

Under mild technical conditions, the map $(x,y,\xi) \mapsto g_x(y,\xi)$ is a stochastic one-sided model of g, and therefore our results immediately yield the $O(\varepsilon^{-4})$ complexity. This complexity guarantee is new, and nicely complements the recent paper [11]. There, the authors proved that almost surely all limit points of the stochastic prox-linear and subgradient methods applied to the problem (1.5) are stationary. Though the complexity of the stochastic subgradient method and of the stochastic prox-linear method are of the same order, empirical evidence [11, Section 4] suggests that the stochastic prox-linear method can perform significantly better. This is not surprising since the prox-linear models satisfy a much stronger approximation guarantee in expectation.

2 The problem set-up and the stationarity measure

Throughout, we consider a Euclidean space $\mathbb{R}^d$ endowed with an inner product $\langle\cdot,\cdot\rangle$ and the induced norm $\|x\| = \sqrt{\langle x,x\rangle}$. We say that a function $f\colon \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ is *ρ-weakly convex* if the assignment $f + \frac{\rho}{2}\|x\|^2$ is convex.[1] A trivial consequence of this definition is that f is ρ-weakly convex if, and only if, the following approximate secant inequality holds:

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y) + \tfrac{\rho\lambda(1-\lambda)}{2}\|x-y\|^2, \tag{2.1}$$

for all $x,y \in \mathbb{R}^d$ and all $\lambda \in [0,1]$. We will use this equivalence in the proof of Lemma 2.1.

Throughout, we consider the optimization problem

$$\min_{x\in\mathbb{R}^d} \ \varphi(x) := g(x) + r(x), \tag{2.2}$$

where $r\colon \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ is a closed function and the only access to $g\colon \mathbb{R}^d \to \mathbb{R}$ is through a *stochastic one-sided model*. Formally, we fix a probability space $(\Omega,\mathcal{F},P)$ and equip $\mathbb{R}^d$ with the Borel σ-algebra. We assume that there exist real $\tau,\eta,L>0$ such that the following four properties hold:

(A1) **(Sampling)** It is possible to generate i.i.d. realizations $\xi_1,\xi_2,\ldots \sim P$.

[1] To the best of our knowledge, the class of weakly convex functions was introduced in [16]. Here we use a slightly more restrictive definition than originally developed there.

(A2) **(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

$$\mathbb{E}_\xi [g_x(x, \xi)] = g(x) \qquad \forall x \in U,$$

and

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

(A3) **(Weak-convexity)** The function $r(\cdot) + g_x(\cdot, \xi)$ is η-weakly convex $\forall x \in U$, a.e. $\xi \in \Omega$.

(A4) **(Lipschitz property)** For all $x, y, z \in U$ and a.e. $\xi \in \Omega$, the inequalities hold:

$$|g(y) - g(z)| \leq L\|y - z\|, \tag{2.3}$$
$$|g_x(y, \xi) - g_x(z, \xi)| \leq L\|y - z\|. \tag{2.4}$$

Remark 1. It is worthwhile to note that if Assumption (A2) is strengthened to a two-sided estimate

$$|\mathbb{E}_\xi [g_x(y, \xi) - g(y)]\,| \leq \frac{\tau}{2}\|y - x\|^2 \qquad \text{for all } x, y \in U,$$

the Lipschitz property of the models (2.4) automatically implies the analogous property for g itself (2.3). To see this, observe that trivially, for any $x \in U$, the function $g_x(y) := \mathbb{E}_{\xi\sim P} [g_x(y, \xi)]$ is L-Lipschitz continuous on U. Therefore, for any $\bar{x} \in U$, we have

$$\limsup_{x,y\to\bar{x}} \frac{|g(y) - g(x)|}{\|y - x\|} \leq \limsup_{x,y\to\bar{x}} \frac{|g_x(y) - g_x(x)| + \frac{\tau}{2}\|y - x\|^2}{\|y - x\|} \leq L,$$

where the last inequality follows from (2.4). Since U is open and convex, we deduce that g is L-Lipschitz continuous on U, as claimed.

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 $\tau + \eta$.

Lemma 2.1. *The function φ is $(\tau + \eta)$-weakly convex.*

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 $g_x(y) := \mathbb{E}_\xi[g_x(y, \xi)]$. Taking into account the equivalence of weak convexity with the approximate secant inequality (2.1), we successively deduce

$$\begin{aligned}
\varphi(\bar{x}) &= \mathbb{E}_\xi [r(\bar{x}) + g_{\bar{x}}(\bar{x}, \xi)] && (2.5)\\
&\leq \lambda\mathbb{E}_\xi [r(x) + g_{\bar{x}}(x, \xi)] + (1 - \lambda)\mathbb{E}_\xi [r(y) + g_{\bar{x}}(y, \xi)] + \tfrac{\eta\lambda(1-\lambda)}{2}\|x - y\|^2 && (2.6)\\
&= \lambda(r(x) + g_{\bar{x}}(x)) + (1 - \lambda)(r(y) + g_{\bar{x}}(y)) + \tfrac{\eta\lambda(1-\lambda)}{2}\|x - y\|^2 \\
&\leq \lambda\varphi(x) + (1 - \lambda)\varphi(y) + \tfrac{\tau(\lambda^2(1-\lambda)+\lambda(1-\lambda)^2)}{2}\|x - y\|^2 + \tfrac{\eta\lambda(1-\lambda)}{2}\|x - y\|^2 && (2.7)\\
&= \lambda\varphi(x) + (1 - \lambda)\varphi(y) + \tfrac{(\tau+\eta)\lambda(1-\lambda)}{2}\|x - y\|^2,
\end{aligned}$$

where (2.5) uses (A2), inequality (2.6) uses (A3), and (2.7) uses (A2). Thus φ is $(\tau + \eta)$-weakly convex, as claimed. □

We can now formalize the algorithm we investigate, as Algorithm 1.

Algorithm 1: Stochastic Model Based Minimization

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

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

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

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

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

Return x_{t^*}

The basic goal of algorithms for nonsmooth and nonconvex optimization, such as those of Algorithm 1, is to find stationary points. These are the points where the directional derivative of the objective function is nonnegative in every direction. Measuring the progress of numerical methods towards stationary points requires a continuous measure of stationarity. Such a continuous measure is readily available for our problem class (2.2). The key construction is the *Moreau envelope*:

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

where $\lambda > 0$. It follows directly from Lemma 2.1 and [17, Theorem 31.5] that as long as $\lambda < (\tau + \eta)^{-1}$, the envelope φ_λ is C^1-smooth with the gradient given by

$$\nabla \varphi_\lambda(x) = \lambda^{-1}(x - \operatorname{prox}_{\lambda\varphi}(x)). \tag{2.8}$$

where $\operatorname{prox}_{\lambda\varphi}(x)$ is the *proximal point*:

$$\operatorname{prox}_{\lambda\varphi}(x) := \operatorname*{argmin}_y \ \{\varphi(y) + \tfrac{1}{2\lambda}\|y - x\|^2\}.$$

It is easy to see that stationary points of φ_λ coincide with those of φ. Moreover, the norm of the gradient $\|\nabla\varphi_\lambda(x)\|$ has an intuitive interpretation in terms of near-stationarity for the target problem (2.2). Namely, the definition of the Moreau envelope directly implies that for any point $x \in \mathbb{R}^d$, the proximal point $\hat{x} := \operatorname{prox}_{\lambda\varphi}(x)$ satisfies

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

where the subdifferential $\partial\varphi$ is meant in the standard variational analytic sense [18, Definition 8.3]. Thus a small gradient $\|\nabla\varphi_\lambda(x)\|$ implies that x is *near* some point $\hat{x}$ that is *nearly stationary* for (2.2); for a more detailed discussion, see [9, Section 4.1]. In summary, the norm of the gradient $\|\nabla\varphi_\lambda(x)\|$ serves as a continuous measure of stationary, and we will judge the performance of Algorithm 1 by the rate at which it drives this quantity to zero.

3 Analysis of the algorithm

Henceforth, let $\{x_t\}_{t\geq 0}$ be the iterates generated by Algorithm 1 and let $\{\xi_t\}_{t\geq 0}$ be the corresponding samples used. For each index $t \geq 0$, define the proximal point

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

To simplify notation, 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 1 crucially relies on the following lemma, which compares the step taken by the algorithm, with the gradient of the Moreau envelope.

Lemma 3.1. *For every index $t \geq 0$, we have*

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

Proof. Recall that the function $x \mapsto r(x) + g_{x_t}(x, \xi_t) + \frac{\beta_t}{2}\|x - x_t\|^2$ 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:

$$\big(r(x) + g_{x_t}(x, \xi_t) + \tfrac{\beta_t}{2}\|x - x_t\|^2\big) \geq \big(r(x_{t+1}) + g_{x_t}(x_{t+1}, \xi_t) + \tfrac{\beta_t}{2}\|x_{t+1} - x_t\|^2\big) + \tfrac{\beta_t - \eta}{2}\|x - x_{t+1}\|^2.$$

Setting $x = \hat{x}_t$, rearranging, and taking expectations we successively deduce

$$\begin{aligned}
&\mathbb{E}_t\left[\frac{\beta_t - \eta}{2}\|\hat{x}_t - x_{t+1}\|^2 + \frac{\beta_t}{2}\|x_{t+1} - x_t\|^2 - \frac{\beta_t}{2}\|\hat{x}_t - x_t\|^2\right] \\
&\leq \mathbb{E}_t[r(\hat{x}_t) + g_{x_t}(\hat{x}_t, \xi_t) - r(x_{t+1}) - g_{x_t}(x_{t+1}, \xi_t)] \\
&\leq \mathbb{E}_t[r(\hat{x}_t) + g_{x_t}(\hat{x}_t, \xi_t) - r(x_{t+1}) - g_{x_t}(x_t, \xi_t) + L\|x_{t+1} - x_t\|] \qquad (3.1) \\
&= r(\hat{x}_t) + \mathbb{E}_\xi[g_{x_t}(\hat{x}_t, \xi)] - \mathbb{E}_t[r(x_{t+1})] - \mathbb{E}_\xi[g_{x_t}(x_t, \xi)] + L \cdot \mathbb{E}_t\|x_{t+1} - x_t\| \\
&\leq r(\hat{x}_t) + g(\hat{x}_t) - \mathbb{E}_t[r(x_{t+1})] - g(x_t) + \frac{\tau}{2}\|\hat{x}_t - x_t\|^2 + L \cdot \mathbb{E}_t\|x_{t+1} - x_t\| \qquad (3.2) \\
&= \mathbb{E}_t[r(\hat{x}_t) + g(\hat{x}_t) - r(x_{t+1}) - g(x_t)] + \frac{\tau}{2}\|\hat{x}_t - x_t\|^2 + L \cdot \mathbb{E}_t\|x_{t+1} - x_t\| \\
&\leq \mathbb{E}_t[r(\hat{x}_t) + g(\hat{x}_t) - r(x_{t+1}) - g(x_{t+1})] + \frac{\tau}{2}\|\hat{x}_t - x_t\|^2 + 2L \cdot \mathbb{E}_t\|x_{t+1} - x_t\| \qquad (3.3) \\
&\leq \mathbb{E}_t\left[-\frac{\hat{\rho}}{2}\|\hat{x}_t - x_t\|^2 + \frac{\hat{\rho}}{2}\|x_{t+1} - x_t\|^2\right] + \frac{\tau}{2}\|\hat{x}_t - x_t\|^2 + 2L \cdot \mathbb{E}_t\|x_{t+1} - x_t\| \qquad (3.4) \\
&= \frac{\tau - \hat{\rho}}{2}\|\hat{x}_t - x_t\|^2 + \frac{\hat{\rho}}{2} \cdot \mathbb{E}_t\|x_{t+1} - x_t\|^2 + 2L \cdot \mathbb{E}_t\|x_{t+1} - x_t\|, \qquad (3.5)
\end{aligned}$$

where (3.1) and (3.3) follow from Assumption (A4), inequality (3.2) follows from (A2), and (3.4) follows directly from the definition of $\hat{x}_t$ as a proximal point,

Define $\gamma := \mathbb{E}_t\|x_t - x_{t+1}\|$ and notice $\gamma^2 \leq \mathbb{E}_t\|x_t - x_{t+1}\|^2$. Rearranging, we deduce

$$\begin{aligned}
\frac{\beta_t - \eta}{2} \cdot \mathbb{E}_t\|\hat{x}_t - x_{t+1}\|^2 &\leq \frac{\beta_t - \hat{\rho} + \tau}{2}\|\hat{x}_t - x_t\|^2 + \frac{\hat{\rho} - \beta_t}{2}\gamma^2 + 2L\gamma \qquad (3.6) \\
&\leq \frac{\beta_t - \hat{\rho} + \tau}{2}\|\hat{x}_t - x_t\|^2 + \frac{2L^2}{\beta_t - \hat{\rho}},
\end{aligned}$$

where the last inequality follows by maximizing the right-hand-side of (3.6) in $\gamma \in \mathbb{R}$. After multiplying through by $\frac{2}{\beta_t - \eta}$, the result follows. □

We can now establish the convergence guarantees of Algorithm 1.

Theorem 3.2 (Convergence rate)**.** *The point x_{t^*} returned by Algorithm 1 satisfies:*

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

In particular, setting $\beta_t = \hat{\rho} + \alpha^{-1}\sqrt{T+1}$ for any real $\alpha > 0$, yields the complexity guarantee

$$\mathbb{E}\|\nabla\varphi_{1/\hat{\rho}}(x_{t^*})\|^2 \leq \frac{2\left(\hat{\rho}(\varphi_{1/\hat{\rho}}(x_0) - \min_x \varphi) + 2\hat{\rho}^2 L^2 \alpha^2\right)}{\hat{\rho}-\tau-\eta} \cdot \left(\frac{\hat{\rho}-\eta}{T+1} + \frac{1}{\alpha\sqrt{T+1}}\right).$$

Proof. Using the definition of the Moreau envelope and appealing to Lemma 3.1, we deduce

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

Using the tower rule for expectations and iterating the recursion yields

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

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

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

Dividing through by $\sum_{t=0}^T \frac{\hat{\rho}-\tau-\eta}{\beta_t-\eta}$ and recognizing the left hand-side as $\mathbb{E}[\|x_{t^*} - \hat{x}_{t^*}\|^2]$, the result follows. □

Let us now look at the consequences of Theorem 3.2 on the three 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 (A1)-(A4) for some $\tau, \eta, L \geq 0$. Complexity guarantees for each method then follow immediately from Theorem 3.2.

Stochastic proximal point Consider the optimization problem (2.2) under the following assumptions.

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

(B2) 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 $\mathbb{E}_\xi[g_x(y, \xi)] = g(y)$ for all $x, y \in U$.

(B3) Each function $r(\cdot) + g_x(\cdot, \xi)$ is ρ-weakly convex $\forall x \in U$, a.e. $\xi \in \Omega$.

(B4) There is a real $L \geq 0$ such that the inequality

$$|g_x(y, \xi) - g_x(z, \xi)| \leq L\|y - z\|,$$

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

The stochastic proximal point method is Algorithm 1 with the models $g_x(y, \xi)$. Taking into account Remark 1, it is immediate to see that (A1)-(A4) hold with $\tau = 0$ and $\eta = \rho$.

Stochastic proximal subgradient Consider the optimization problem (2.2), and let us assume that the following properties are true.

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

(C2) The function g is ρ_1-weakly convex and r is ρ_2-weakly convex, for some $\rho_1, \rho_2 \geq 0$.

(C3) 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 $\mathbb{E}_\xi[G(x, \xi)] \in \partial g(x)$ for all $x \in U$.

(C4) There is a real $L \geq 0$ such that the inequality, $\mathbb{E}_\xi\left[\|G(x, \xi)\|^2\right] \leq L^2$, holds for all $x \in U$.

The stochastic subgradient method is Algorithm 1 with the linear models

$$g_x(y, \xi) = g(x) + \langle G(x, \xi), y - x \rangle .$$

Observe that (A1) and (A3) with $\eta = \rho_2$ are immediate from the definitions; (A2) with $\tau = \rho_1$ follows from the discussion in [4, Section 2]. To see (A4), observe that (C3) and (C4) directly imply that whenever g is differentiable at $x \in U$, we have

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

Since at any x, the subdifferential $\partial g(x)$ is the convex hull of limits of gradients at nearby points [17, Theorem 25.6], the claimed assumption (A4) follows. An analogous guarantee was recently shown in [4] when r is convex, using a different argument.

Stochastic prox-linear Consider the optimization problem (2.2) with

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

We assume that there exists an open convex set U containing $\operatorname{dom} r$, and reals $\ell, \gamma > 0$ such 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 assignments $h\colon \mathbb{R}^m \times \Omega \to \mathbb{R}$ and $c\colon U \times \Omega \to \mathbb{R}^m$ are measurable.

(D3) The function r is ρ-weakly convex, while for a.e. $\xi \in \Omega$, the function $z \mapsto h(z,\xi)$ is convex and ℓ-Lipschitz, and the map $x \mapsto c(x,\xi)$ is C^1-smooth with γ-Lipschitz Jacobian.

(D4) The inequality, $\|\nabla c(x,\xi)\|_{\mathrm{op}} \leq M$, holds for all $x \in U$ and a.e. $\xi \in \Omega$.

The stochastic prox-linear method [11] is Algorithm 1 with the convex models

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

Observe that (A1) and (A3) hold trivially with $\eta = \rho$. Assumption (A2) holds with $\tau = \ell\gamma$ by [11, Lemma 3.12]. Combining (D4) with Remark 1 directly implies (A4) with $L = \ell M$.

References

[1] P. Bianchi. Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization*, 26(4):2235–2260, 2016.

[2] J.V. Burke. Descent methods for composite nondifferentiable optimization problems. *Math. Programming*, 33(3):260–279, 1985.

[3] Y. Chen, Y. Chi, and A.J Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

[4] 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.

[5] D. Davis, D. Drusvyatskiy, and K.J. MacPhee. Subgradient methods for sharp weakly convex functions. *Preprint arXiv:1803.02461*, 2018.

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

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

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

[9] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Preprint arXiv:1605.00125*, 2016.

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

[11] J.C. Duchi and F. Ruan. Stochastic methods for composite optimization problems. *Preprint arXiv:1703.08570*, 2017.

[12] N. Gillis. The why and how of nonnegative matrix factorization. In *Regularization, optimization, kernels, and support vector machines*, Chapman & Hall/CRC Mach. Learn. Pattern Recogn. Ser., pages 257–291. CRC Press, Boca Raton, FL, 2015.

[13] N. Gillis. Introduction to nonnegative matrix factorization. *SIAG/OPT Views and News*, 25(1):7–16, 2017.

[14] A.S. Lewis and S.J. Wright. A proximal method for composite minimization. *Math. Program.*, pages 1–46, 2015.

[15] J. Nocedal and S.J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[16] E. A. Nurminskii. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9(1):145–150, Jan 1973.

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

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

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