Journal of Machine Learning Research 26 (2025) 1-39 Submitted 12/24; Revised 8/25; Published 11/25

Stochastic Interior-Point Methods for Smooth Conic Optimization with Applications

Chuan He CHUAN.HE@LIU.SE
Department of Mathematics
Linköping University
Linköping, SE-581 83, Sweden

Zhanwang Deng DZW_OPT2022@STU.PKU.EDU.CN
Academy for Advanced Interdisciplinary Studies
Peking University
Beijing, 100871, China

Editor: Peter Richtarik

Abstract

Conic optimization plays a crucial role in many machine learning (ML) problems. However, practical algorithms for conic constrained ML problems with large datasets are often limited to specific use cases, as stochastic algorithms for general conic optimization remain underdeveloped. To fill this gap, we introduce a stochastic interior-point method (SIPM) framework for general conic optimization, along with four novel SIPM variants leveraging distinct stochastic gradient estimators. Under mild assumptions, we establish the iteration complexity of our proposed SIPMs, which, up to a polylogarithmic factor, matches the best-known results in stochastic unconstrained optimization. Finally, our numerical experiments on robust linear regression, multi-task relationship learning, and clustering data streams demonstrate the effectiveness and efficiency of our approach.

Keywords: Conic optimization, stochastic interior-point methods, iteration complexity, robust linear regression, multi-task relationship learning, clustering data streams

1. Introduction

Conic optimization covers a broad class of optimization problems with constraints represented by convex cones, including common forms such as linear constraints, second-order cone constraints, and semidefinite constraints. Over the years, this class of optimization problems has found applications across various fields, including control (Fares et al., 2001), energy systems (Zohrizadeh et al., 2020), combinatorial optimization (Wolkowicz et al., 2012), and machine learning (ML) (Sra et al., 2011). In practice, the efficient handling of traditional conic optimization problems has been extensively studied for decades, with interior-point methods (IPMs) standing out due to their ability to effectively and elegantly solve a wide range of conic constrained problems within a unified framework (see the monograph by Nesterov and Nemirovski (1994)).

1.1 Stochastic Optimization with Conic Constraints for Machine Learning

Existing studies on IPMs primarily focus on the deterministic regime, despite the recent widespread applications of stochastic conic optimization in ML. These applications include multi-task relationship learning (Argyriou et al., 2008; Zhang and Yeung, 2010), robust learning with chance constraints (Shivaswamy et al., 2006; Xu et al., 2012), kernel learning (Bach et al., 2004), and clustering data streams (Bidaurrazaga et al., 2021; Peng and Wei, 2007; Sun et al., 2021). Next, we briefly highlight two examples.

Multi-task relationship learning Multi-task learning is an ML paradigm where related tasks are learned together to improve generalization by sharing information (see Zhang and Yang (2021)). In many applications, task correlations are not explicitly available. To learn these correlations from data, a regularization framework is proposed, where the relationships between models for different tasks are controlled by a regularizer defined using the covariance matrix Σ (Argyriou et al., 2008; Zhang and Yeung, 2010):

$$\min_{W\in\mathbb{R}^{p\times d},\Sigma\in\mathbb{R}^{p\times p}} \frac{1}{p}\sum_{i=1}^{p}\frac{1}{m}\sum_{j=1}^{m}\ell(w_i,a_{ij})+\lambda\mathrm{tr}(W^TP(\Sigma)W) \quad \text{s.t.} \quad \Sigma\in\mathbb{S}^p_+, \quad \mathrm{tr}(\Sigma)=1, \quad (1)$$

where $W=[w_1,\ldots,w_p]^T$ denotes the model coefficients, $\mathbb{S}^p_+$ denotes the positive semidefinite cone, $\ell(\cdot,\cdot)$ is the loss function, w_i and $\{a_{ij}\}_{j=1}^m$ are respectively the model weight and the training set for the ith task, $1\le i\le p$, $\lambda>0$ is a tuning parameter, $P:\mathbb{R}^{p\times p}\to\mathbb{R}^{p\times p}$ is a given map that controls the interaction between W and Σ, and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. The constraint $\mathrm{tr}(\Sigma)=1$ in (1) is imposed to control the complexity of Σ, as discussed in Zhang and Yeung (2010).

Robust learning with chance constraints Consider supervised learning with feature-label pairs $\{(a_i,b_i)\}_{i=1}^p$, where the a_i's are assumed to be generated from a distribution $\mathcal{D}$ with expected value $\bar{a}$ and covariance matrix Σ. To mitigate the uncertainty in the a_i's, a chance constraint $\mathbb{P}_{a\sim\mathcal{D}}(|w^T(a-\bar{a})|\ge\theta)\le\eta$ is proposed to be incorporated when performing robust linear regression (Shivaswamy et al., 2006; Xu et al., 2012), where θ and η are the confidence level and desired probability, respectively, and w denotes the model coefficients to be optimized. By approximating this chance constraint with a second-order cone constraint, Shivaswamy et al. (2006) propose the following robust regression problem with conic constraints:

$$\min_{w\in\mathbb{R}^d,\theta,v\ge0} \frac{1}{p}\sum_{i=1}^{p}\phi(w^Ta_i-b_i)+\lambda_1\theta+\lambda_2 v \quad \text{s.t.} \quad (w,v)\in\mathbb{Q}^{d+1}, \quad (\Sigma^{1/2}w,\sqrt{\eta}\theta)\in\mathbb{Q}^{d+1}, \quad (2)$$

where $\phi(\cdot)$ is the loss, $\lambda_1,\lambda_2>0$ are tuning parameters, and $\mathbb{Q}^{d+1}\doteq\{(u,t)\in\mathbb{R}^d\times\mathbb{R}_+ : \|u\|\le t\}$ denotes the second-order cone.

Both examples in (1) and (2) involve nonlinear conic constraints and are typically coupled with large datasets in real applications. However, stochastic algorithms for addressing general conic optimization problems in ML remain largely underdeveloped. Existing algorithmic developments for conic constrained ML problems are limited to specific use cases: for instance, alternating minimization is widely employed in the literature (Argyriou et al., 2008; Zhang and Yeung, 2010) to solve (1), while Shivaswamy et al. (2006) formulates (2) as

a second-order cone program, assuming that ϕ is convex. Nevertheless, these developments do not unify the treatment of conic constraints and often lack convergence guarantees for problems with nonconvex objective functions.

In this paper, we aim for proposing a stochastic interior-point method (SIPM) framework for smooth conic optimization problems, including (1) and (2) as special cases. This class of problems takes the following form:

$$\min_x f(x) \quad \text{s.t.} \quad x \in \Omega \doteq \{x : Ax = b, x \in \mathcal{K}\}, \tag{3}$$

where f is continuously differentiable and possibly nonconvex on Ω, but its derivatives are not accessible. Instead, we rely on stochastic estimators $G(\cdot;\xi)$ of $\nabla f(\cdot)$, where ξ is a random variable with sample space Ξ (see Assumptions 1(c) and 3 for assumptions on $G(\cdot;\xi)$). Here, $A \in \mathbb{R}^{m\times n}$ is of full row rank, $b \in \mathbb{R}^m$, and $\mathcal{K} \subseteq \mathbb{R}^n$ is a closed and pointed convex cone with a nonempty interior. Assume that (3) has at least one optimal solution.

1.2 Our Contributions

In this paper, we propose an SIPM framework (Algorithm 1) for solving problem (3), which, to the best of our knowledge, is the first stochastic algorithmic framework for this problem. Building on Algorithm 1, we introduce four novel SIPM variants by incorporating different stochastic gradient estimators: mini-batch estimators, Polyak momentum, extrapolated Polyak momentum, and recursive momentum. Under mild assumptions, we establish the iteration complexity for these variants. For our SIPMs, we summarize the samples per iteration, iteration complexity, and smoothness assumptions in Table 1. In addition, numerical results demonstrate the practical advantages of our SIPMs over existing methods.

Table 1: Samples per iteration, iteration complexity for finding an ϵ-SSP, and smoothness assumptions.

method	samples per iteration	iteration complexity	smoothness assumption
SIPM-ME	$\widetilde{\mathcal{O}}(\epsilon^{-2})$	$\widetilde{\mathcal{O}}(\epsilon^{-2})$	locally smooth ∇f (Asm. 1(b))
SIPM-PM	1	$\widetilde{\mathcal{O}}(\epsilon^{-4})$	locally smooth ∇f (Asm. 1(b))
SIPM-EM	1	$\widetilde{\mathcal{O}}(\epsilon^{-7/2})$	locally smooth ∇f & $\nabla^2 f$ (Asm. 1(b) & 2(b))
SIPM-RM	1	$\widetilde{\mathcal{O}}(\epsilon^{-3})$	locally average smooth G (Asm. 3)

Our main contributions are highlighted below.

- We propose an SIPM framework (Algorithm 1) for solving problem (3). To the best of our knowledge, this is the first stochastic algorithmic framework for ML problems with general conic constraints. Building upon this framework, we introduce four novel SIPM variants that incorporate different stochastic gradient estimators.
- We establish the iteration complexity of our SIPMs using mini-batch estimations, Polyak momentum, extrapolated Polyak momentum, and recursive momentum under *locally Lipschitz-type* conditions (Assumptions 1(b), 2(b), and 3). All of the iteration complexity results established in this paper are entirely new and match, up to a polylogarithmic factor, the best-known complexity bounds in stochastic unconstrained optimization.

- We conduct numerical experiments (Section 4) to compare our SIPMs with existing methods on robust linear regression, multi-task relationship learning, and clustering data streams. The results demonstrate that our SIPMs achieve solution quality comparable to or better than existing methods, and importantly, they consistently solve ML problems with general conic constraints, unlike existing task-specific approaches.

1.3 Related Work

Interior-point methods In the deterministic regime, IPMs are recognized as a fundamental and widely used algorithmic approach for solving constrained optimization problems, having been extensively studied for decades (Alizadeh, 1995; Byrd et al., 1999; Kim et al., 2007; Nesterov and Nemirovski, 1994; Potra and Wright, 2000; Vanderbei and Shanno, 1999; Wächter and Biegler, 2006; Wright, 1997; Forsgren et al., 2002). In particular, primal-dual IPMs are a popular class of methods that iterate towards to an optimal solution by applying Newton's method to solve a system of equations derived from the perturbed first-order necessary conditions of the optimization problem (Wächter and Biegler, 2006; Wright, 1997). Another early and classical family of IPMs is the affine scaling method, which iteratively improves a feasible point within the relative interior of the feasible region by scaling the search direction to prevent the solution from exiting the boundary (Tseng and Luo, 1992; Dvurechensky and Staudigl, 2024). The SIPMs developed in this paper are more closely aligned with the algorithmic ideas of affine scaling methods. Moreover, IPM-based solvers are widely adopted for large-scale constrained optimization, including Ipopt (Wächter and Biegler, 2006), Knitro (Byrd et al., 1999), and LOQO (Vanderbei and Shanno, 1999) for functional constrained problems, and SDPT3 (Toh et al., 1999), SeDuMi (Sturm, 1999), and Mosek (ApS, 2019) for conic constrained problems.

Studies on SIPMs have only emerged recently. In particular, Badenbroek and de Klerk (2022) and Narayanan (2016) propose randomized IPMs for minimizing a linear function over a convex set. In addition, Curtis et al. (2025) introduces an SIPM for bound-constrained optimization by augmenting the objective function with a log-barrier function, and Curtis et al. (2024) generalizes this approach to solve inequality constrained optimization problems. The optimization problems and algorithmic ideas studied in these previous works differ significantly from those in this paper.

Stochastic first-order methods Recent significant developments in ML, fueled by large-scale data, have made stochastic first-order optimization methods widely popular for driving ML applications. In particular, many stochastic first-order methods have been developed for solving unconstrained and simple constrained problems of the form $\min_{x\in X} f(x)$, where $X \subseteq \mathbb{R}^n$ is a set for which the projection can be computed exactly (Cutkosky and Mehta, 2020; Cutkosky and Orabona, 2019; Fang et al., 2018; Ghadimi and Lan, 2013; Lan, 2020; Tran-Dinh et al., 2022; Wang et al., 2019; Xu and Xu, 2023; Li et al., 2021). Assuming that f has a Lipschitz continuous gradient, iteration complexity of $\mathcal{O}(\epsilon^{-4})$ has been established for the methods in Ghadimi and Lan (2013); Cutkosky and Mehta (2020), in terms of minimizing the stationary measure $\text{dist}(0, \nabla f(x^k) + \mathcal{N}_X(x^k))$, where $\mathcal{N}_X(x^k)$ denotes the normal cone of X at x^k. The iteration complexity can be improved to $\mathcal{O}(\epsilon^{-7/2})$ by using an implicit gradient transport technique, assuming that f has a Lipschitz continuous Hessian Cutkosky and Mehta (2020), and to $\mathcal{O}(\epsilon^{-3})$ by using various variance reduction techniques,

assuming that ∇f has a stochastic estimator satisfying the average smoothness condition Cutkosky and Orabona (2019); Fang et al. (2018); Tran-Dinh et al. (2022); Wang et al. (2019); Xu and Xu (2023); Li et al. (2021). In addition, a number of recent efforts have been devoted to equality constrained optimization in the stochastic regime, including sequential quadratic programming methods Berahas et al. (2023a, 2021, 2023b); Curtis et al. (2021); Fang et al. (2024); Na et al. (2023); Na and Mahoney (2022) and penalty methods Alacaoglu and Wright (2024); Li et al. (2024); Lu et al. (2024); Shi et al. (2025).

2. Notation and Preliminaries

Throughout this paper, let $\mathbb{R}^n$ denote the n-dimensional Euclidean space and $\langle\cdot,\cdot\rangle$ denote the standard inner product. We use $\|\cdot\|$ to denote the Euclidean norm of a vector or the spectral norm of a matrix. For the closed convex cone $\mathcal{K}$, its interior and dual cone are denoted by $\text{int}\mathcal{K}$ and $\mathcal{K}^*$, respectively. Define the perturbed barrier function of problem (3) as:

$$\phi_\mu(x) \doteq f(x) + \mu(f(x) + B(x)) \qquad \forall x \in \Omega^\circ \doteq \{x \in \text{int}\mathcal{K} : Ax = b\} \text{ and } \mu \in (0, 1]. \tag{4}$$

For a finite set $\mathcal{B}$, let $|\mathcal{B}|$ denote its cardinality. For any $t \in \mathbb{R}$, let $[t]_+$ and $[t]_-$ denote its nonnegative and nonpositive parts, respectively (i.e., set to zero if t is negative or positive, respectively). Also, let $\lfloor t \rfloor$ denote the largest integer less than or equal to t. We use the standard big-O notation $\mathcal{O}(\cdot)$ to present the complexity, and $\widetilde{\mathcal{O}}(\cdot)$ to represent the order with a polylogarithmic factor omitted.

For the rest of this section, we review some background on logarithmically homogeneous self-concordant barriers and introduce the approximate optimality conditions.

2.1 Logarithmically Homogeneous Self-Concordant Barrier

Logarithmically homogeneous self-concordant (LHSC) barrier functions play a crucial role in the development of IPMs for conic optimization (see Nesterov and Nemirovski (1994)). in this paper, the design and analysis of SIPMs also heavily rely on the LHSC barrier function. Throughout this paper, assume that $\mathcal{K}$ is equipped with a ϑ-*LHSC* barrier function B, where $\vartheta \geq 1$. Specifically, $B : \text{int}\mathcal{K} \to \mathbb{R}$ satisfies the following conditions: (i) B is convex and three times continuously differentiable in $\text{int}\mathcal{K}$, and moreover, $|\varphi'''(0)| \leq 2(\varphi''(0))^{3/2}$ holds for all $x \in \text{int}\mathcal{K}$ and $u \in \mathbb{R}^n$, where $\varphi(t) = B(x + tu)$; (ii) B is a *barrier function* for $\mathcal{K}$, meaning that $B(x)$ goes to infinity as x approaches the boundary of $\mathcal{K}$; (iii) B satisfies the *logarithmically homogeneous property*: $B(tx) = B(x) - \vartheta \ln t$ for all $x \in \text{int}\mathcal{K}, t > 0$.

For any $x \in \text{int}\mathcal{K}$, the function B induces the following local norms for vectors:

$$\|v\|_x \doteq \left(v^T \nabla^2 B(x) v\right)^{1/2}, \quad \|v\|_x^* \doteq \left(v^T \nabla^2 B(x)^{-1} v\right)^{1/2} \qquad \forall v \in \mathbb{R}^n. \tag{5}$$

The induced local norm for matrices is given by:

$$\|M\|_x^* \doteq \max_{\|v\|_x \leq 1} \|Mv\|_x^* \qquad \forall M \in \mathbb{R}^{n \times n}. \tag{6}$$

2.2 Approximate Optimality Conditions

Since f is nonconvex, finding a global solution to (3) is generally impossible. Instead, we aim to find a point that satisfies approximate optimality conditions, as is common in nonconvex optimization. For deterministic IPMs developed to solve (3), the following approximate optimality conditions are proposed in He and Lu (2023); He et al. (2024):

$$x \in \text{int}\mathcal{K}, \quad Ax = b, \quad \nabla f(x) + A^T\tilde{\lambda} \in \mathcal{K}^*, \quad \|\nabla f(x) + A^T\tilde{\lambda}\|_x^* \le \epsilon, \tag{7}$$

where $\epsilon \in (0,1)$ denotes the tolerance. In addition, stochastic algorithms typically produce solutions that satisfy approximate optimality conditions only in expectation. To facilitate our developments of SIPMs, we next derive an alternative approximate optimality condition for (3) using B, which is a sufficient condition for (7). Its proof is deferred to Section 5.1.

Lemma 1 *Let $\mu > 0$ be given. Suppose that $(x,\lambda) \in \Omega^\circ \times \mathbb{R}^m$ satisfies $\|\nabla\phi_\mu(x) + A^T\lambda\|_x^* \le \mu$, where ϕ_μ and Ω° are given in (4). Then, x also satisfies (7) with $\tilde{\lambda} = \lambda/(1+\mu)$ and any $\epsilon \ge (1+\sqrt{\vartheta})\mu$.*

The above lemma offers an alternative approximate optimality condition for (3). We extend this condition to an expectation form and define an approximate stochastic stationary point for problem (3), which our SIPMs aim to achieve.

Definition 2 *Let $\epsilon \in (0,1)$. We say that $x \in \Omega^\circ$ is an ϵ-stochastic stationary point (ϵ-SSP) of problem (3) if it, together with some $\lambda \in \mathbb{R}^m$, satisfies $\mathbb{E}[\|\nabla\phi_\mu(x) + A^T\lambda\|_x^*] \le \mu$ for some $\mu \le \epsilon/(1+\sqrt{\vartheta})$.*

One can also define an approximate stochastic stationary point for (3) that satisfies (7) with high probability, as follows.

Definition 3 *Let $\epsilon, \delta \in (0,1)$. We say that $x \in \Omega^\circ$ is an (ϵ,δ)-stochastic stationary point ((ϵ,δ)-SSP) of problem (3) if it, together with some $\lambda \in \mathbb{R}^m$, satisfies $\|\nabla\phi_\mu(x) + A^T\lambda\|_x^* \le \mu$ for some $\mu \le \epsilon/(1+\sqrt{\vartheta})$ with probability at least $1-\delta$.*

Using Markov's inequality, we obtain that

$$\mathbb{P}(\|\nabla\phi_\mu(x) + A^T\lambda\|_x^* > \mu) \le \frac{\mathbb{E}[\|\nabla\phi_\mu(x) + A^T\lambda\|_x^*]}{\mu}$$

holds for any $\mu > 0$. Thus, if $x \in \Omega^\circ$, together with $\lambda \in \mathbb{R}^m$, satisfies $\mathbb{E}[\|\nabla\phi_\mu(x) + A^T\lambda\|_x^*] \le \delta\mu$ for some $\mu \le \epsilon/(1+\sqrt{\vartheta})$ and $\epsilon, \delta \in (0,1)$, one can derive that x is an (ϵ,δ)-SSP of problem (3), which leads to the following lemma.

Lemma 4 *Let $\epsilon, \delta \in (0,1)$ be given. Suppose that $(x,\lambda) \in \Omega^\circ \times \mathbb{R}^m$ satisfies $\mathbb{E}[\|\nabla\phi_\mu(x) + A^T\lambda\|_x^*] \le \delta\mu$ for some $\mu \le \epsilon/(1+\sqrt{\vartheta})$, where ϕ_μ and Ω° are given in (4). Then, x is an (ϵ,δ)-SSP of problem (3).*

For the remainder of this paper, we focus on establishing the iteration complexity of our proposed SIPMs for finding an ϵ-SSP of problem (3). While we do not discuss it in detail, by applying Lemma 4, one can also extend our analysis to establish the iteration complexity of our SIPMs for finding an (ϵ,δ)-SSP of problem (3).

3. Stochastic Interior-Point Methods

In this section, we propose an SIPM framework for solving (3) and then analyze the iteration complexity of four SIPM variants. We now make the following additional assumptions that will be used throughout this section.

Assumption 1 (a) *The Slater's condition holds, that is, Ω°, defined in (4), is nonempty. In addition, there exists a finite $\phi_{\rm low}$ such that*

$$\inf_{x\in\Omega^\circ,\mu\in(0,1]}\{f(x)+\mu B(x)\}\geq\phi_{\rm low}. \tag{8}$$

(b) *For any $s_\eta\in(0,1)$, there exists an $L_1>0$ such that*

$$\|\nabla f(y)-\nabla f(x)\|_x^*\leq L_1\|y-x\|_x \qquad \forall x,y\in\Omega^\circ \text{ with } \|y-x\|_x\leq s_\eta. \tag{9}$$

(c) *We have access to a stochastic gradient estimator $G:\Omega^\circ\times\Xi\to\mathbb{R}^n$ that satisfies*

$$\mathbb{E}_\xi[G(x;\xi)]=\nabla f(x),\quad \mathbb{E}_\xi[(\|G(x;\xi)-\nabla f(x)\|_x^*)^2]\leq\sigma^2 \qquad \forall x\in\Omega^\circ \tag{10}$$

for some $\sigma>0$.

Remark 5 (i) *Assumption 1(a) is reasonable. Particularly, the assumption in (8) means that the barrier function $f(x)+\mu B(x)$ is uniformly bounded below whenever the barrier parameter μ is no larger than 1. It usually holds for problems where the barrier method converges, and similar assumptions are also used in He and Lu (2023); He et al. (2024). Otherwise, in case that (8) fails, one can instead solve a perturbed counterpart of (3):*

$$\min_x f(x)+\tau\|x\|^2 \quad \text{s.t.} \quad Ax=b, x\in\mathcal{K} \tag{11}$$

for a sufficiently small $\tau>0$. It can be verified that (8) holds for the perturbed problem (11) with $f(\cdot)$ replaced by $f(\cdot)+\tau\|\cdot\|^2$. Indeed, let f^* be the optimal value of (3). Then, one has

$$\begin{aligned}
&\inf_{x\in\Omega^\circ,\mu\in(0,1]}\{f(x)+\tau\|x\|^2+\mu B(x)\}\\
&\geq f^*+\inf_{\mu\in(0,1]}\Big\{\mu\inf_{x\in\Omega^\circ}\{(\tau/\mu)\|x\|^2+B(x)\}\Big\}\geq f^*+\inf_{\mu\in(0,1]}\Big\{\mu\inf_{x\in\Omega^\circ}\{\tau\|x\|^2+B(x)\}\Big\}\\
&\geq f^*-\Big|\inf_{x\in\Omega^\circ}\{\tau\|x\|^2+B(x)\}\Big|>-\infty,
\end{aligned}$$

where the last inequality is due to the strong convexity of $\tau\|x\|^2+B(x)$. Hence, the assumption in (8) holds for the perturbed problem (11).

(ii) *As a consequence of Assumption 1(a), one can see that the perturbed barrier function defined in (4) is bounded below:*

$$\phi_\mu(x)=(1+\mu)\Big(f(x)+\frac{\mu}{1+\mu}B(x)\Big)\geq(1+\mu)\phi_{\rm low} \qquad \forall\mu\in(0,1],x\in\Omega^\circ. \tag{12}$$

In addition, for notational convenience, we define

$$\Delta(x)\doteq f(x)+[f(x)+B(x)]_+-2[\phi_{\rm low}]_- \qquad \forall x\in\Omega^\circ. \tag{13}$$

(iii) *Assumption 1(b) implies that ∇f is locally Lipschitz continuous on Ω° with respect to local norms. Similar local smoothness assumptions are also used for other barrier methods (He and Lu, 2023; Dvurechensky and Staudigl, 2024; He et al., 2024). As will be shown in Lemma 27(iv), ∇B is locally Lipschitz continuous on Ω° with respect to local norms, even if it is not well defined on the boundary of Ω. In general, we recall from He and Lu (2023) that for any bounded x, $\nabla^2 B(x)^{-1}$ is bounded with respect to the spectral norm. Then, one can show that when Ω is bounded, the locally Lipschitz condition (9) can be implied by the globally Lipschitz condition (see He and Lu (2023, Section 5) for detailed discussions). For convenience, we define*

$$L_\phi \doteq 2L_1 + 1/(1 - s_\eta), \tag{14}$$

which denotes the Lipschitz constant of $\nabla \phi_\mu$ for any $\mu \in (0, 1]$ (see Theorem 28 below).

(iv) *Assumption 1(c) implies that $G(\cdot; \xi)$ is an unbiased estimator of $\nabla f(\cdot)$ and that its variance, with respect to the local norm, is bounded above. As noted in Theorem 5(iii), $\nabla^2 B(x)^{-1}$ is bounded with respect to the spectral norm for any bounded x. Using this and (5), we have that when Ω is bounded, the second relation in (10) holds if the variance of $G(\cdot; \xi)$ with respect to the Euclidean norm is bounded.*

In what follows, we propose an SIPM framework in Algorithm 1 for solving problem (3). Subsequently, we will employ four distinct stochastic estimators to construct $\{\overline{m}^k\}_{k\geq 0}$, with specific schemes provided in (16), (22), (29), and (38), respectively.

We remark that computations involving $\nabla^2 B(x^k)^{-1}$ can be performed efficiently for common nonlinear cones $\mathcal{K}$. For example, when $\mathcal{K}$ is the second-order cone, $\nabla B(x^k)^{-1}$ can be computed analytically (e.g., see Alizadeh and Goldfarb (2003)), and $(A\nabla^2 B(x^k)^{-1}A^T)^{-1}v$ for any $v \in \mathbb{R}^m$ can be efficiently evaluated using Cholesky factorization. When $\mathcal{K}$ is the semidefinite cone, we have that $\nabla^2 B(X^k)^{-1}[V] = X^k V X^k$ holds for any $n \times n$ symmetric matrix V, and that $A\nabla^2 B(X^k)^{-1}A^T$ can be efficiently computed by exploiting the sparsity of A (see Toh et al. (1999) for details).

Algorithm 1 An SIPM framework

Input: starting point $x^0 \in \Omega^\circ$, nonincreasing step sizes $\{\eta_k\}_{k\geq 0} \subset (0, s_\eta]$ with $s_\eta \in (0, 1)$, nonincreasing barrier parameters $\{\mu_k\}_{k\geq 0} \subset (0, 1]$.

for $k = 0, 1, 2, \ldots$ **do**

Construct an estimator $\overline{m}^k$ for $\nabla f(x^k)$, and set $m^k = \overline{m}^k + \mu_k(\overline{m}^k + \nabla B(x^k))$.

Update dual and primal variables as follows:

$$\lambda^k = -(AH_kA^T)^{-1}AH_km^k, \quad x^{k+1} = x^k - \eta_k H_k(m^k + A^T\lambda^k)/\|m^k + A^T\lambda^k\|^*_{x^k}, \tag{15}$$

where $H_k = \nabla^2 B(x^k)^{-1}$.

end for

Our next lemma shows that all iterates $\{x^k\}_{k\geq 0}$ generated by Algorithm 1 are strictly feasible, i.e., $x^k \in \Omega^\circ$ holds for all $k \geq 0$. Its proof is deferred to Section 5.1.

Lemma 6 *Suppose that Assumption 1 holds. Let $\{x^k\}_{k\geq 0}$ be generated by Algorithm 1. Then, $\|x^{k+1} - x^k\|_{x^k} = \eta_k$ and $x^k \in \Omega^\circ$ for all $k \geq 0$, where Ω° is defined in (4).*

In the remainder of this section, we propose four variants of SIPMs in Sections 3.1 to 3.4 and study their iteration complexity. The developments and guarantees of these four SIPM variants are independent of each other.

3.1 An SIPM with Mini-Batch Estimators

We now describe a variant of Algorithm 1, where $\{\overline{m}^k\}_{k\geq 0}$ is constructed using mini-batch estimators:

$$\overline{m}^k = \sum\nolimits_{i\in\mathscr{B}_k} G(x^k;\xi_i^k)/|\mathscr{B}_k| \qquad \forall k\geq 0, \tag{16}$$

where $G(\cdot,\cdot)$ satisfies Assumption 1(c), and the sequence $\{\mathscr{B}_k\}_{k\geq 0}$ denotes the sets of sample indices. We refer this variant as SIPM with mini-batch estimators (SIPM-ME).

The following lemma establishes an upper bound for the estimation error of the mini-batch estimators defined in (16), and its proof is deferred to Section 5.3.

Lemma 7 *Suppose that Assumption 1 holds. Let $\{x^k\}_{k\geq 0}$ be generated by Algorithm 1 with $\{\overline{m}^k\}_{k\geq 0}$ constructed as in (16) and input parameters $\{(\eta_k,\mathscr{B}_k)\}_{k\geq 0}$. Then,*

$$\mathbb{E}_{\{\xi_i^k\}_{i\in\mathscr{B}_k}}[(\|\overline{m}^k-\nabla f(x^k)\|_{x^k}^*)^2] \leq \sigma^2/|\mathscr{B}_k| \qquad \forall k\geq 0, \tag{17}$$

where σ is given in Assumption 1(c).

We next provide an upper bound for the average expected error of the stationary condition across all iterates generated by SIPM-ME. Its proof is relegated to Section 5.3.

Theorem 8 *Suppose that Assumption 1 holds. Let $\{(x^k,\lambda^k)\}_{k\geq 0}$ be the sequence generated by Algorithm 1 with $\{\overline{m}^k\}_{k\geq 0}$ constructed as in (16) and input parameters $\{(\eta_k,\mathscr{B}_k,\mu_k)\}_{k\geq 0}$. Assume that $\{\eta_k\}_{k\geq 0}$ is nonincreasing. Then, for any $K\geq 1$,*

$$\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla\phi_{\mu_k}(x^k)+A^T\lambda^k\|_{x^k}^*] \leq \frac{\Delta(x^0)}{\eta_{K-1}}+\frac{1}{\eta_{K-1}}\sum_{k=0}^{K-1}\eta_k\left(\frac{4\sigma}{|\mathscr{B}_k|^{1/2}}+\frac{L_\phi}{2}\eta_k\right), \tag{18}$$

where $\Delta(\cdot)$, L_ϕ, and σ are given in (13), (14), and Assumption 1(c), respectively.

3.1.1 Hyperparameters and Iteration Complexity

In this subsection, we establish the iteration complexity of SIPM-ME with its input parameters specified $\{(\eta_k,\mathscr{B}_k,\mu_k)\}_{k\geq 0}$ as:

$$\eta_k=\frac{s_\eta}{(k+1)^{1/2}}, \quad |\mathscr{B}_k|=k+1, \quad \mu_k=\max\left\{\frac{1}{(k+1)^{1/2}},\frac{\epsilon}{1+\sqrt{\vartheta}}\right\} \qquad \forall k\geq 0, \tag{19}$$

where $s_\eta\in(0,1)$ is a user-defined maximum length for step sizes in Algorithm 1, and $\epsilon\in(0,1)$ denotes the tolerance. It follows that $\{\eta_k\}_{k\geq 0}\subset(0,s_\eta]$ and $\{\mu_k\}_{k\geq 0}\subset(0,1]$, with both sequences nonincreasing.

The following theorem presents the iteration complexity of SIPM-ME with inputs specified in (19). Its proof is deferred to Section 5.3.

Theorem 9 *Suppose that Assumption 1 holds. Consider Algorithm 1 with $\{\overline{m}^k\}_{k\geq 0}$ constructed as in (16) and $\{(\eta_k, \mathscr{B}_k, \mu_k)\}_{k\geq 0}$ specified as in (19). Let $\kappa(K)$ be uniformly drawn from $\{\lfloor K/2\rfloor, \ldots, K-1\}$, and define*

$$M_{\rm me} \doteq 2\Big(\frac{\Delta(x^0)}{s_\eta} + 8\sigma + s_\eta L_\phi\Big), \tag{20}$$

where $\Delta(\cdot)$ is defined in (13), L_ϕ and σ are given in (14) and Assumption 1(c), respectively, and s_η is an input of Algorithm 1. Then,

$$\begin{aligned}
&\mathbb{E}[\|\nabla\phi_{\mu_{\kappa(K)}}(x^{\kappa(K)}) + A^T\lambda^{\kappa(K)}\|^*_{x^{\kappa(K)}}] \leq \mu_{\kappa(K)} \quad \textit{with} \quad \mu_{\kappa(K)} \leq \epsilon/(1+\sqrt{\vartheta}) \\
&\forall K \geq \max\Big\{2\Big(\frac{1+\sqrt{\vartheta}}{\epsilon}\Big)^2, \Big(\frac{4M_{\rm me}(1+\sqrt{\vartheta})}{\epsilon}\ln\Big(\frac{4M_{\rm me}(1+\sqrt{\vartheta})}{\epsilon}\Big)\Big)^2, 3\Big\}.
\end{aligned} \tag{21}$$

Remark 10 *From Theorem 9, we observe that SIPM-ME returns an ϵ-SSP within $\widetilde{\mathcal{O}}(\epsilon^{-2})$ iterations. In addition, by this and the definition of $|\mathscr{B}_k|$, one can see that the stochastic gradient evaluations per iteration of SIPM-ME is at most $\widetilde{\mathcal{O}}(\epsilon^{-2})$.*

3.2 An SIPM with Polyak Momentum

We now propose a variant of Algorithm 1, in which $\{\overline{m}^k\}_{k\geq 0}$ is constructed using Polyak momentum (Cutkosky and Mehta, 2020) as follows:

$$\gamma_{-1} = 1, \quad \overline{m}^{-1} = 0, \quad \overline{m}^k = (1-\gamma_{k-1})\overline{m}^{k-1} + \gamma_{k-1}G(x^k, \xi^k) \qquad \forall k \geq 0, \tag{22}$$

where $G(\cdot,\cdot)$ satisfies Assumption 1(c), and $\{\gamma_k\}_{k\geq 0} \subset (0,1]$ denotes the sequence of momentum parameters. We refer this variant as SIPM with Polyak momentum (SIPM-PM).

The next lemma provides the recurrence relation for the estimation error of the gradient estimators based on Polyak momentum defined in (22). Its proof is deferred to Section 5.4.

Lemma 11 *Suppose that Assumption 1 holds. Let $\{x^k\}_{k\geq 0}$ be generated by Algorithm 1 with $\{\overline{m}^k\}_{k\geq 0}$ constructed as in (22) and input parameters $\{(\eta_k, \gamma_k)\}_{k\geq 0}$. For all $k \geq 0$, assume that $\gamma_k > \eta_k$ holds, and define $\alpha_k = 1 - (1-\gamma_k)/(1-\eta_k)$. Then,*

$$\begin{aligned}
&\mathbb{E}_{\xi^{k+1}}[(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|^*_{x^{k+1}})^2] \\
&\leq (1-\alpha_k)(\|\overline{m}^k - \nabla f(x^k)\|^*_{x^k})^2 + \frac{L_1^2\eta_k^2}{\alpha_k} + \sigma^2\gamma_k^2 \qquad \forall k \geq 0,
\end{aligned} \tag{23}$$

where L_1 and σ are given in Assumption 1.

We now derive an upper bound for the average expected error of the stationary condition across all iterates generated by SIPM-PM. Its proof is deferred to Section 5.4.

Theorem 12 *Suppose that Assumption 1 holds. Let $\{(x^k, \lambda^k)\}_{k\geq 0}$ be the sequence generated by Algorithm 1 with $\{\overline{m}^k\}_{k\geq 0}$ constructed as in (22) and input parameters $\{(\eta_k, \gamma_k, \mu_k)\}_{k\geq 0}$. Assume that $\{\eta_k\}_{k\geq 0}$ is nonincreasing and also that $\gamma_k > \eta_k$ holds for all $k \geq 0$. Define $\alpha_k = 1 - (1-\gamma_k)/(1-\eta_k)$ for all $k \geq 0$. Then, for all $K \geq 1$,*

$$\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|^*_{x^k}] \leq \frac{\Delta(x^0) + \sigma^2/L_1}{\eta_{K-1}} + \frac{1}{\eta_{K-1}}\sum_{k=0}^{K-1}\left(\frac{L_\phi\eta_k^2}{2} + \frac{5L_1\eta_k^2}{\alpha_k} + \frac{\sigma^2\gamma_k^2}{L_1}\right), \tag{24}$$

where $\Delta(\cdot)$ is defined in (13), L_ϕ is given in (14), and L_1 and σ are given in Assumption 1.

3.2.1 Hyperparameters and Iteration Complexity

In this subsection, we establish the iteration complexity of SIPM-PM with its input parameters $\{(\eta_k, \gamma_k, \mu_k)\}_{k\geq 0}$ specified as:

$$\eta_k = \frac{s_\eta}{(k+1)^{3/4}}, \quad \gamma_k = \frac{1}{(k+1)^{1/2}}, \quad \mu_k = \max\Big\{\frac{1}{(k+1)^{1/4}}, \frac{\epsilon}{1+\sqrt{\vartheta}}\Big\} \qquad \forall k \geq 0, \quad (25)$$

where $s_\eta \in (0,1)$ is a user-defined input of Algorithm 1, and $\epsilon \in (0,1)$ denotes the tolerance. It can be verified that $\{\eta_k\}_{k\geq 0} \subset (0, s_\eta]$ and $\{\mu_k\}_{k\geq 0} \subset (0,1]$, with both sequences nonincreasing. From (25) and $s_\eta \in (0,1)$, we observe that the sequence $\{\alpha_k\}_{k\geq 0}$ defined in Theorems 11 and 12 satisfies:

$$\alpha_k = \frac{(k+1)^{1/4} - s_\eta}{(k+1)^{3/4} - s_\eta} > \frac{(k+1)^{1/4} - s_\eta}{(k+1)^{3/4}} = \frac{1 - s_\eta/(k+1)^{1/4}}{(k+1)^{1/2}} \geq \frac{1-s_\eta}{(k+1)^{1/2}} \qquad \forall k \geq 0. \quad (26)$$

The following theorem presents the iteration complexity of SIPM-PM with its inputs specified in (25). Its proof is relegated to Section 5.4.

Theorem 13 *Suppose that Assumption 1 holds. Consider Algorithm 1 with $\{\overline{m}^k\}_{k\geq 0}$ constructed as in (22) and $\{(\eta_k, \gamma_k, \mu_k)\}_{k\geq 0}$ specified as in (25). Let $\kappa(K)$ be uniformly drawn from $\{\lfloor K/2 \rfloor, \ldots, K-1\}$, and define*

$$M_{\rm pm} \doteq 2\Big(\frac{\Delta(x^0) + \sigma^2/L_1}{s_\eta} + \frac{3 s_\eta L_\phi}{2} + 2\Big(\frac{5 s_\eta L_1}{1-s_\eta} + \frac{\sigma^2}{s_\eta L_1}\Big)\Big), \quad (27)$$

where $\Delta(\cdot)$ is defined in (13), L_ϕ is given in (14), L_1 and σ are given in Assumption 1, and s_η is an input of Algorithm 1. Then,

$$\begin{aligned} &\mathbb{E}[\|\nabla \phi_{\mu_{\kappa(K)}}(x^{\kappa(K)}) + A^T \lambda^{\kappa(K)}\|^*_{x^{\kappa(K)}}] \leq \mu_{\kappa(K)} \quad \textit{with} \quad \mu_{\kappa(K)} \leq \epsilon/(1+\sqrt{\vartheta}) \\ &\forall K \geq \max\Big\{2\Big(\frac{1+\sqrt{\vartheta}}{\epsilon}\Big)^4, \Big(\frac{8 M_{\rm pm}(1+\sqrt{\vartheta})}{\epsilon} \ln\Big(\frac{8 M_{\rm pm}(1+\sqrt{\vartheta})}{\epsilon}\Big)\Big)^4, 3\Big\}. \end{aligned} \quad (28)$$

Remark 14 (i) *From Theorem 13, we see that SIPM-PM returns an ϵ-SSP within $\widetilde{\mathcal{O}}(\epsilon^{-4})$ iterations. This complexity result matches that of stochastic unconstrained optimization with Lipschitz continuous gradient Ghadimi and Lan (2013); Cutkosky and Mehta (2020), up to a polylogarithmic factor.*

(ii) *It is worth mentioning that our analysis implies that only a point selected from a subset of the first K iterates generated by SIPM-PM (and similarly for other variants) is an ϵ-SSP, provided that K is sufficiently large. This guarantee is consistent with those in Cutkosky and Orabona (2019); Cutkosky and Mehta (2020) for stochastic first-order methods with momentum in unconstrained optimization. It would be interesting to develop SIPMs with a stronger guarantee that all x^K are ϵ-SSPs when K is larger than a threshold, which we leave as a direction for future research.*

3.3 An SIPM with Extrapolated Polyak Momentum

We now propose a variant of Algorithm 1, where $\{\overline{m}^k\}_{k\geq 0}$ is constructed based on extrapolated Polyak momentum (Cutkosky and Mehta, 2020) as follows:

$$\gamma_{-1} = 1, \quad x^{-1} = x^0, \quad \overline{m}^{-1} = 0, \tag{29a}$$

$$z^k = x^k + \frac{1-\gamma_{k-1}}{\gamma_{k-1}}(x^k - x^{k-1}), \quad \overline{m}^k = (1-\gamma_{k-1})\overline{m}^{k-1} + \gamma_{k-1}G(z^k, \xi^k) \qquad \forall k \geq 0, \tag{29b}$$

where $G(\cdot,\cdot)$ satisfies Assumption 1(c), and $\{\gamma_k\}_{k\geq 0} \subset (0,1]$ denotes momentum parameters. We refer to this variant as SIPM with extrapolated Polyak momentum (SIPM-EM).

To analyze SIPM-EM, we make the following additional assumption regarding the local Lipschitz continuity of $\nabla^2 f$.

Assumption 2 (a) *The function f is twice continuously differentiable on Ω°.*

(b) *For any $s_\eta \in (0,1)$, there exists an $L_2 > 0$ such that*

$$\|\nabla^2 f(y) - \nabla^2 f(x)\|_x^* \leq L_2\|y-x\|_x \qquad \forall x, y \in \Omega^\circ \text{ with } \|y-x\|_x \leq s_\eta. \tag{30}$$

The following lemma shows that the iterates $\{z^k\}_{k\geq 0}$ generated by SIPM-EM lie in Ω°. Its proof is deferred to Section 5.5.

Lemma 15 *Suppose that Assumptions 1 and 2 hold. Let $\{z^k\}_{k\geq 0}$ be generated by Algorithm 1 with $\{\overline{m}^k\}_{k\geq 0}$ constructed as in (29) and input parameters $\{(\eta_k,\gamma_k)\}_{k\geq 0}$. Assume $\eta_k/\gamma_k \leq s_\eta$ for all $k \geq 0$, where s_η is an input of Algorithm 1. Then, $z^k \in \Omega^\circ$ for all $k \geq 0$, where Ω° is defined in (4).*

The next lemma provides the recurrence relation for the estimation error of the gradient estimators based on extrapolated Polyak momentum defined in (29). Its proof is deferred to Section 5.5.

Lemma 16 *Suppose that Assumptions 1 and 2 hold. Let $\{x^k\}_{k\geq 0}$ be generated by Algorithm 1 with $\{\overline{m}^k\}_{k\geq 0}$ constructed as in (29) and input parameters $\{(\eta_k,\gamma_k)\}_{k\geq 0}$. Assume that $\{\eta_k\}_{k\geq 0}$ is nonincreasing, $\{\gamma_k\}_{k\geq 0} \subset (0,1]$, and that $\eta_k/\gamma_k \leq s_\eta$ holds for all $k \geq 0$, where s_η is an input of Algorithm 1. Define $\alpha_k = 1-(1-\gamma_k)/(1-\eta_k)$ for all $k \geq 0$. Then,*

$$\begin{aligned}\mathbb{E}_{\xi^{k+1}}[(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|_{x^{k+1}}^*)^2] \leq (1-\alpha_k)(\|\overline{m}^k - \nabla f(x^k)\|_{x^k}^*)^2 \\ + \frac{L_2^2\eta_k^4}{(1-\eta_0)^2\gamma_k^2\alpha_k} + \frac{\sigma^2\gamma_k^2}{(1-\eta_0)^2(1-\eta_k/\gamma_k)^2} \qquad \forall k \geq 0,\end{aligned} \tag{31}$$

where σ and L_2 are given in Assumptions 1(c) and 2(b), respectively.

We next derive an upper bound for the average expected error of the stationary condition across all iterates generated by SIPM-EM. Its proof is relegated to Section 5.5.

Theorem 17 *Suppose that Assumptions 1 and 2 hold. Let $\{(x^k, \lambda^k)\}_{k\geq 0}$ be generated by Algorithm 1 with $\{\overline{m}^k\}_{k\geq 0}$ constructed as in (29) and input parameters $\{(\eta_k, \gamma_k, \mu_k)\}_{k\geq 0}$. Assume that $\{\eta_k\}_{k\geq 0}$ is nonincreasing, $\{\gamma_k\}_{k\geq 0} \subset (0, 1]$, and that $\eta_k/\gamma_k \leq s_\eta$ holds for all $k \geq 0$, where s_η is an input of Algorithm 1. Define $\alpha_k = 1 - (1-\gamma_k)/(1-\eta_k)$ for all $k \geq 0$. Let $\{p_k\}_{k\geq 0}$ be a nondecreasing sequence satisfying $(1-\alpha_k)p_{k+1} \leq (1-\alpha_k/2)p_k$ for all $k \geq 0$. Then, for all $K \geq 1$,*

$$\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla \phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^*] \leq \frac{\Delta(x^0) + p_0\sigma^2}{\eta_{K-1}} + \frac{1}{\eta_{K-1}} \sum_{k=0}^{K-1} \left(\frac{L_\phi}{2}\eta_k^2 + \frac{8\eta_k^2}{p_k\alpha_k} + \frac{L_2^2\eta_k^4 p_{k+1}}{(1-\eta_0)^2\gamma_k^2\alpha_k} + \frac{\sigma^2\gamma_k^2 p_{k+1}}{(1-\eta_0)^2(1-\eta_k/\gamma_k)^2} \right), \quad (32)$$

where $\Delta(\cdot)$ and L_ϕ are defined in (13) and (14), respectively, and σ and L_2 are given in Assumptions 1(c) and 2(b), respectively.

3.3.1 Hyperparameters and Iteration Complexity

In this subsection, we establish the iteration complexity of SIPM-EM with its input parameters $\{(\eta_k, \gamma_k, \mu_k)\}_{k\geq 0}$ specified as:

$$\eta_k = \frac{5s_\eta}{7(k+1)^{5/7}}, \quad \gamma_k = \frac{1}{(k+1)^{4/7}}, \quad \mu_k = \max\left\{ \frac{1}{(k+1)^{2/7}}, \frac{\epsilon}{1+\sqrt{\vartheta}} \right\} \qquad \forall k \geq 0, \quad (33)$$

where $s_\eta \in (0, 1)$ is a user-defined input of Algorithm 1, and $\epsilon \in (0, 1)$ denotes the tolerance. It then follows that $\{\eta_k\}_{k\geq 0} \subset (0, s_\eta]$ and $\{\mu_k\}_{k\geq 0} \subset (0, 1]$, with both sequences nonincreasing. We also define

$$p_k = (k+1)^{1/7} \qquad \forall k \geq 0. \quad (34)$$

The next lemma provides some useful properties of the sequences $\{\alpha_k\}_{k\geq 0}$ and $\{p_k\}_{k\geq 0}$ defined in Theorem 17 and (34), respectively. These properties will be used to establish the iteration complexity of SIPM-EM, and the proof is deferred to Section 5.5.

Lemma 18 *Let $\{(\eta_k, \gamma_k)\}_{k\geq 0}$ be defined in (33), and $\{\alpha_k\}_{k\geq 0}$ and $\{p_k\}_{k\geq 0}$ be defined in Theorem 17 and (34), respectively. Then, $\alpha_k \geq (1 - 5s_\eta/7)/(k+1)^{4/7}$ and $(1-\alpha_k)p_{k+1} \leq (1-\alpha_k/2)p_k$ hold for all $k \geq 0$.*

The next theorem establishes the iteration complexity of SIPM-EM with its inputs specified in (33). Its proof is relegated to Section 5.5.

Theorem 19 *Suppose that Assumptions 1 and 2 hold. Consider Algorithm 1 with $\{\overline{m}^k\}_{k\geq 0}$ constructed as in (29) and $\{(\eta_k, \gamma_k, \mu_k)\}_{k\geq 0}$ specified in (33). Let $\kappa(K)$ be uniformly drawn from $\{\lfloor K/2 \rfloor, \ldots, K-1\}$, and define*

$$M_{\text{em}} \doteq \frac{14}{5}\left(\frac{\Delta(x^0) + \sigma^2}{s_\eta} + \frac{40 s_\eta L_\phi}{49} + 2\Big(\frac{200 s_\eta}{7(7-5s_\eta)} + \frac{1250 L_2^2 s_\eta^3}{7(7-5s_\eta)^3} + \frac{2\sigma^2}{s_\eta(1-5s_\eta/7)^4} \Big) \right), \quad (35)$$

where $\Delta(\cdot)$ and L_ϕ are defined in (13) and (14), respectively, σ and L_2 are given in Assumptions 1(c) and 2(b), respectively, and s_η is an input of Algorithm 1. Then,

$$\mathbb{E}[\|\nabla\phi_{\mu_{\kappa(K)}}(x^{\kappa(K)}) + A^T\lambda^{\kappa(K)}\|^*_{x^{\kappa(K)}}] \le \mu_{\kappa(K)} \quad \text{with} \quad \mu_{\kappa(K)} \le \epsilon/(1+\sqrt{\vartheta})$$
$$\forall K \ge \max\Big\{2\Big(\frac{1+\sqrt{\vartheta}}{\epsilon}\Big)^{7/2}, \Big(\frac{7M_{\rm em}(1+\sqrt{\vartheta})}{\epsilon}\ln\Big(\frac{7M_{\rm em}(1+\sqrt{\vartheta})}{\epsilon}\Big)\Big)^{7/2}, 3\Big\}. \tag{36}$$

Remark 20 *From Theorem 19, we observe that SIPM-EM returns an ϵ-SSP within $\widetilde{\mathcal{O}}(\epsilon^{-7/2})$ iterations. This iteration complexity matches that of stochastic unconstrained optimization with Lipschitz continuous Hessian Cutkosky and Mehta (2020), up to a polylogarithmic factor.*

3.4 An SIPM with Recursive Momentum

This subsection incorporates recursive momentum into Algorithm 1. We make the following additional assumptions throughout this subsection.

Assumption 3 *We have access to a stochastic gradient estimator $G : \Omega^\circ \times \Xi \to \mathbb{R}^n$ such that (10) holds for some $\sigma > 0$, and for any $s_\eta \in (0,1)$,*

$$\mathbb{E}_\xi[(\|G(y,\xi) - G(x,\xi)\|^*_x)^2] \le L^2\|y-x\|_x^2 \qquad \forall x, y \in \Omega^\circ \text{ with } \|y-x\|_x \le s_\eta \tag{37}$$

holds for some $L > 0$.

Remark 21 *Assumption 3 can be seen as a local-norm variant of the average smoothness condition (Cutkosky and Orabona, 2019; Fang et al., 2018; Li et al., 2021). It is generally stronger than Assumption 1(b), as implied by the following:*

$$(\|\nabla f(y) - \nabla f(x)\|^*_x)^2 = (\|\mathbb{E}_\xi[G(y,\xi) - G(x,\xi)]\|^*_x)^2 \le \mathbb{E}_\xi[(\|G(y,\xi) - G(x,\xi)\|^*_x)^2],$$

where the equality follows from the unbiasedness of $G(\cdot,\xi)$, and the inequality follows from Jensen's inequality and the convexity of $\|\cdot\|^*_x$.

We next describe a variant of Algorithm 1 with recursive momentum, in which $\{\overline{m}^k\}_{k\ge0}$ is constructed based on recursive momentum (Cutkosky and Orabona, 2019):

$$\gamma_{-1} = 1, \quad x^{-1} = x^0, \quad \overline{m}^{-1} = 0, \tag{38a}$$
$$\overline{m}^k = G(x^k,\xi^k) + (1-\gamma_{k-1})(\overline{m}^{k-1} - G(x^{k-1},\xi^k)) \qquad \forall k \ge 0, \tag{38b}$$

where $G(\cdot,\cdot)$ satisfies Assumption 3, and $\{\gamma_k\}_{k\ge0} \subset (0,1]$ is the sequence of momentum parameters. For ease of reference, we refer to this variant of Algorithm 1 as SIPM with recursive momentum (SIPM-RM).

The next lemma provides the recurrence for the estimation error of the gradient estimators based on the recursive momentum described in (38). Its proof is deferred to Section 5.6.

Lemma 22 *Suppose that Assumptions 1 and 3 hold. Let $\{x^k\}_{k\ge0}$ be all iterates generated by Algorithm 1 with $\{\overline{m}^k\}_{k\ge0}$ updated according to (38) and input parameters $\{(\eta_k,\gamma_k)\}_{k\ge0}$.*

Assume that $\{\eta_k\}_{k\geq 0}$ is nonincreasing. For all $k \geq 0$, assume that $\gamma_k > \eta_k$ and define $\alpha_k = 1 - (1-\gamma_k)/(1-\eta_k)$. Then,

$$\begin{aligned}&\mathbb{E}_{\xi^{k+1}}[(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|^*_{x^{k+1}})^2] \\ &\leq (1-\alpha_k)(\|\overline{m}^k - \nabla f(x^k)\|^*_{x^k})^2 + \frac{3(L_1^2+L^2)\eta_k^2 + 3\sigma^2\gamma_k^2}{(1-\eta_0)^2} \qquad \forall k \geq 0, \end{aligned} \tag{39}$$

where L_1 and σ are given in Assumption 1, and L is given in Assumption 3.

We next provide an upper bound for the average expected error of the stationary condition among all iterates generated by SIPM-RM. Its proof is relegated to Section 5.6.

Theorem 23 *Suppose that Assumptions 1 and 3 hold. Let $\{(x^k, \lambda^k)\}_{k\geq 0}$ be generated by Algorithm 1 with $\{\overline{m}^k\}_{k\geq 0}$ updated according to (38) and input parameters $\{(\eta_k, \gamma_k, \mu_k)\}_{k\geq 0}$. Assume that $\{\eta_k\}_{k\geq 0}$ is nonincreasing and also that $\gamma_k > \eta_k$ for all $k \geq 0$. Define $\alpha_k = 1 - (1-\gamma_k)/(1-\eta_k)$ for all $k \geq 0$. Let $\{p_k\}_{k\geq 0}$ be a nondecreasing sequence satisfying $(1-\alpha_k)p_{k+1} \leq (1-\alpha_k/2)p_k$ for all $k \geq 0$. Then, for all $K \geq 1$,*

$$\begin{aligned}\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla \phi_{\mu_k}(x^k) + A^T\lambda^k\|^*_{x^k}] &\leq \frac{\Delta(x^0) + p_0\sigma^2}{\eta_{K-1}} \\ &\frac{1}{\eta_{K-1}} \sum_{k=0}^{K-1} \Big(\frac{L_\phi}{2}\eta_k^2 + \frac{8\eta_k^2}{p_k\alpha_k} + \frac{3(L_1^2+L^2)\eta_k^2 p_{k+1} + 3\sigma^2\gamma_k^2 p_{k+1}}{(1-\eta_0)^2} \Big), \end{aligned} \tag{40}$$

where $\Delta(\cdot)$ and L_ϕ are defined in (13) and (14), respectively, L_1 and σ are given in Assumption 1, and L is given in Assumption 3.

3.4.1 Hyperparameters and Iteration Complexity

In this subsection, we establish the iteration complexity of SIPM-RM with its input parameters $\{(\eta_k, \gamma_k, \mu_k)\}_{k\geq 0}$ specified as:

$$\eta_k = \frac{s_\eta}{3(k+1)^{2/3}}, \quad \gamma_k = \frac{1}{(k+1)^{2/3}}, \quad \mu_k = \max\Big\{\frac{1}{(k+1)^{1/3}}, \frac{\epsilon}{1+\sqrt{\vartheta}}\Big\} \qquad \forall k \geq 0, \tag{41}$$

where $s_\eta \in (0,1)$ is a user-defined input of Algorithm 1, and $\epsilon \in (0,1)$ denotes the tolerance. It then follows that $\{\eta_k\}_{k\geq 0} \subset (0, s_\eta]$ and $\{\mu_k\}_{k\geq 0} \subset (0,1]$, with both sequences nonincreasing. We also define

$$p_k = (k+1)^{1/3} \qquad k \geq 0. \tag{42}$$

The following lemma provides some useful properties of the sequences $\{\alpha_k\}_{k\geq 0}$ and $\{p_k\}_{k\geq 0}$ defined in Theorem 23 and (42), respectively. Its proof is deferred to Section 5.6.

Lemma 24 *Let $\{(\eta_k, \gamma_k)\}_{k\geq 0}$ be defined in (41), and $\{\alpha_k\}_{k\geq 0}$ and $\{p_k\}_{k\geq 0}$ be defined in Theorem 23 and (42), respectively. Then, $\alpha_k \geq (1-s_\eta/3)/(k+1)^{2/3}$ and $(1-\alpha_k)p_{k+1} \leq (1-\alpha_k/2)p_k$ hold for all $k \geq 0$.*

The following theorem presents the iteration complexity of SIPM-RM with its inputs specified in (41). Its proof is relegated to Section 5.6.

Theorem 25 *Suppose that Assumptions 1 and 3 hold. Consider Algorithm 1 with* $\{\overline{m}^k\}_{k\geq 0}$ *updated using (38) and input parameters* $\{(\eta_k, \gamma_k, \mu_k)\}_{k\geq 0}$ *specified in (41). Let* $\kappa(K)$ *be uniformly drawn from* $\{\lfloor K/2\rfloor, \ldots, K-1\}$*, and define*

$$M_{\rm rm} \doteq 6\Big(\frac{\Delta(x^0)+\sigma^2}{s_\eta} + \frac{2s_\eta L_\phi}{9} + 4\Big(\frac{4s_\eta}{3(3-s_\eta)} + \frac{3(L_1^2+L^2)s_\eta}{(3-s_\eta)^2} + \frac{3\sigma^2}{s_\eta(1-s_\eta/3)^2}\Big)\Big), \quad (43)$$

where $\Delta(\cdot)$ *and* L_ϕ *are defined in (13) and (14), respectively,* L_1 *and* σ *are given in Assumption 1, and* L *is given in Assumption 3. Then,*

$$\mathbb{E}[\|\nabla\phi_{\mu_{\kappa(K)}}(x^{\kappa(K)}) + A^T\lambda^{\kappa(K)}\|^*_{x^{\kappa(K)}}] \leq \mu_{\kappa(K)} \quad \text{with} \quad \mu_{\kappa(K)} \leq \epsilon/(1+\sqrt{\vartheta})$$
$$\forall K \geq \max\Big\{2\Big(\frac{1+\sqrt{\vartheta}}{\epsilon}\Big)^3, \Big(\frac{6M_{\rm rm}(1+\sqrt{\vartheta})}{\epsilon}\ln\Big(\frac{6M_{\rm rm}(1+\sqrt{\vartheta})}{\epsilon}\Big)\Big)^3, 3\Big\}. \quad (44)$$

Remark 26 *From Theorem 25, we see that SIPM-RM returns an* ϵ*-SSP within* $\widetilde{\mathcal{O}}(\epsilon^{-3})$ *iterations. This iteration complexity bound matches the best-known results for stochastic unconstrained optimization under average smoothness condition Cutkosky and Orabona (2019); Fang et al. (2018); Li et al. (2021), up to a polylogarithmic factor.*

4. Numerical Experiments

We now conduct numerical experiments to evaluate the performance of our proposed SIPMs. For SIPM-ME, we consider two versions: SIPM-ME$^+$ with increasing batch sizes and SIPM-ME^1 with a fixed batch size. We compare our SIPMs against a deterministic variant of Algorithm 1 with full-batch gradients (IPM-FG) and other popular methods on robust linear regression (Section 4.1), multi-task relationship learning (Section 4.2), and clustering data streams (Section 4.3).

We evaluate the approximate solutions found by our SIPMs using two measures: relative objective value: $f(x^k)/f(x^0)$; and relative estimated stationary error: $\|m^k+A^T\lambda^k\|^*_{x^k}/\|m^0+A^T\lambda^0\|^*_{x^0}$. All experiments are carried out using Matlab 2024b on a standard PC with 3.20 GHz AMD R7 5800H microprocessor with 16GB of memory. The code to reproduce our numerical results is available at `https://github.com/ChuanH6/SIPM`.

Table 2: The relative objective value and relative estimated stationary error for all methods applied to solve problem (2).

	wine-quality				energy-efficiency			
	$(d,p)=(10,2000)$		$(d,p)=(20,4000)$		$(d,p)=(10,5000)$		$(d,p)=(20,10000)$	
	objective	stationary	objective	stationary	objective	stationary	objective	stationary
SIPM-ME^1	0.2860	1.550e-2	0.3170	2.500e-2	0.8243	3.303e-3	0.8406	6.500e-3
SIPM-ME$^+$	0.2376	3.611e-3	0.2220	3.813e-3	0.8128	5.982e-5	0.8126	5.493e-5
SIPM-PM	0.2307	2.002e-3	0.2372	3.703e-3	0.8131	8.123e-5	0.8133	8.051e-5
SIPM-EM	0.2307	2.002e-3	0.2322	3.702e-3	0.8128	6.954e-6	0.8127	7.527e-6
SIPM-RM	0.2306	1.903e-3	0.2302	4.303e-3	0.8128	2.167e-6	0.8127	2.444e-6
IPM-FG	0.2359	4.652e-3	0.2350	4.503e-3	0.8128	2.974e-6	0.8127	1.453e-5
SALM-RM	0.2583	–	0.2570	–	0.8532	–	0.8241	–

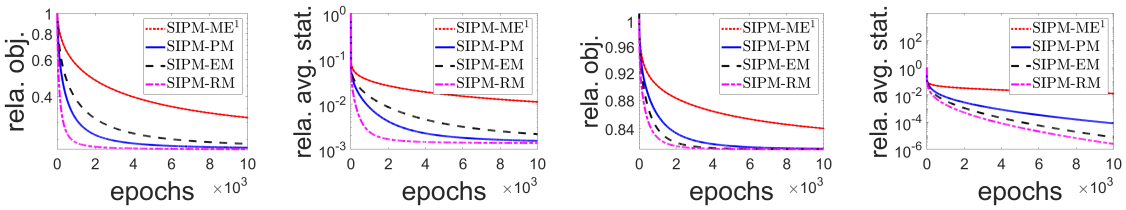

Figure 1: Convergence behavior of the relative objective value and average relative stationary error for each epoch. The first two and last two plots correspond to the 'wine-quality' and 'energy-efficiency' datasets, respectively.

4.1 Robust Linear Regression with Chance Constraints

In this subsection, we consider the second-order cone constrained regression problem in (2), which is derived from robust regression with chance constraints. We let $\phi(t) = t^2/(1+t^2)$ as a robust modification of the quadratic loss (Carmon et al., 2017). We consider two typical regression datasets, namely, 'wine-quality' and 'energy-efficiency', from the UCI repository.[1] Following Shivaswamy et al. (2006), we simulate missing values by randomly deleting 25% of the features in 50% of the training samples, and filling the missing values using linear regression on the remaining features.

We apply our SIPMs, IPM-FG, and a stochastic augmented Lagrangian method with recursive momentum in Alacaoglu and Wright (2024); Lu et al. (2024) (SALM-RM) to solve (2). For SIPMs, we set the barrier function as $B(u,t) = -\ln(t^2 - \|u\|^2)$, associated with the second-order cone $\mathbb{Q}^{d+1} \doteq \{(u,t) \in \mathbb{R}^d \times \mathbb{R}_+ : \|u\| \le t\}$. For SIPM-ME^1, SIPM-PM, SIPM-EM, SIPM-RM, and SALM-RM, we set the maximum number of epochs as $10,000$, and the batch size as 200 and 500 for the 'wine-quality' and 'energy-efficiency' datasets, respectively. For SIPM-ME$^+$, we initialize the batch size as 1 and increase it by 1 per iteration. For a fair comparison, we set the maximum number of iterations for SIPM-ME$^+$ and IPM-FG such that the total number of data points used during training equals that of the other methods in $10,000$ epochs. For all methods, we set the initial point (w^0, v^0, θ^0) to $(0,1,1)$. We set the other hyperparameters—including step sizes, momentum parameters, and barrier parameters—as diminishing sequences of the form $\{(k+1)^{-\alpha}\}_{k\ge 0}$, with the exponent α tuned via grid search to best suit each method in terms of computational performance.

Comparing the relative objective values and the relative estimated stationary errors in Table 2, we see that our SIPM-ME$^+$, SIPM-PM, SIPM-EM, and SIPM-RM yield solutions of similar quality to IPM-FG. In addition, SIPM-ME^1 converges slowly and returns sub-optimal solutions, which corroborates the theoretical results that incorporating momentum facilitates gradient estimation and improves solution quality. We also observe that the solution quality of SALM-RM is generally worse than that of our SIPMs, except SIPM-ME^1. From Figure 1, we observe that SIPM-ME^1 converges much slower than the other three variants, and SIPM-RM is slightly faster than SIPM-PM and SIPM-EM, which corroborates our established iteration complexity for these methods.

1. see `archive.ics.uci.edu/datasets`

4.2 Multi-Task Relationship Learning

In this subsection, we consider the problem of multi-task relationship learning in (1), where $a_{ij} = (p_{ij}, q_{ij}) \in \mathbb{R}^d \times \mathbb{R}$ for all $1 \leq i \leq p$ and $1 \leq j \leq m_i$, $\ell(w_i, a_{ij}) = \phi(w_i^T p_{ij} - q_{ij})$, and $\phi(t) = t^2/(1+t^2)$. We consider five typical regression datasets from the UCI repository: 'wine-quality-red', 'wine-quality-white', 'energy-efficiency', 'air-quality', and 'abalone'. In our multi-task learning, we consider problems with five tasks and ten tasks, where, in the former case, we treat a subset of each of the five datasets as a separate task, and in the latter case, we take two subsets from each dataset, treating each subset as a separate task.

We apply our SIPMs, IPM-FG, and the alternating minimization method in Argyriou et al. (2008) (AM) to solve (1). For SIPMs, we choose the barrier function to be $B(\Sigma) = -\ln(\det(\Sigma))$, associated with the positive semidefinite cone $\mathbb{S}^p_+ \doteq \{\Sigma \succeq 0\}$. For SIPM-ME^1, SIPM-PM, SIPM-EM, and SIPM-RM, we choose the batch size as 200 and 500, and set the maximum number of epochs as 250. For SIPM-ME$^+$, we initialize the batch size as 10 and increase it by 10 per iteration. For a fair comparison, we set the maximum number of iterations for SIPM-ME$^+$, IPM-FG and AM such that the total number of data points used during training equals that of the other methods in 250 epochs. For all methods, we set the initial point W^0 as the all-zero matrix and Σ^0 as a diagonal matrix with diagonal elements equal to $1/p$. We set the other hyperparameters—including step sizes, momentum parameters, and barrier parameters—according to the strategy described in Section 4.1.

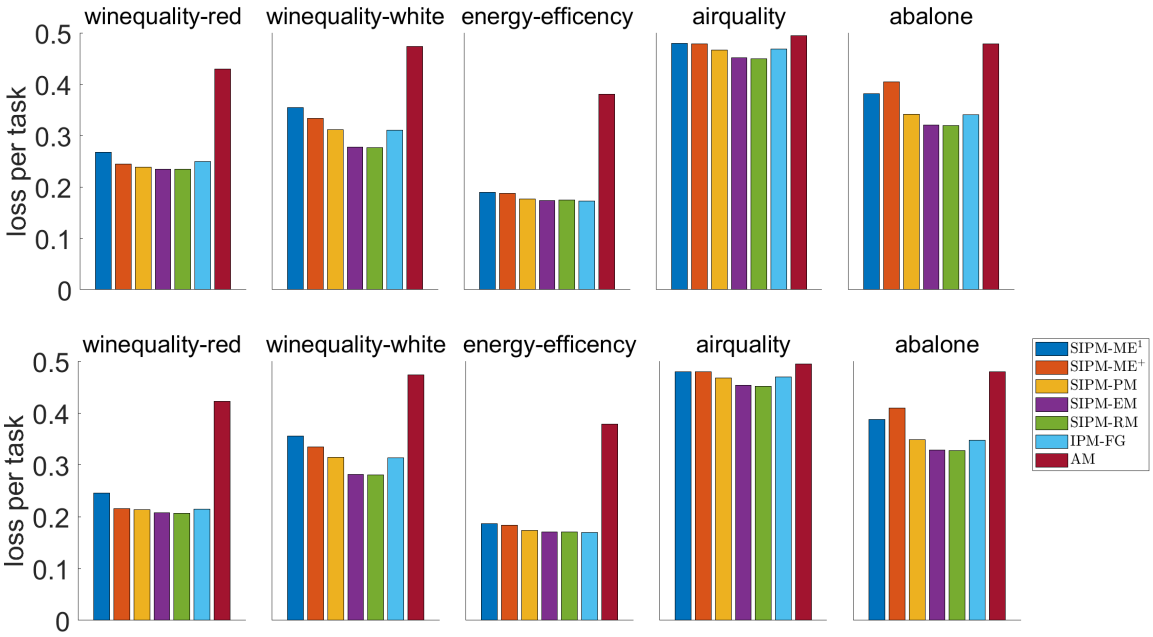

Figure 2: Loss per task for the training (top) and validation (bottom).

From Figure 2, our SIPMs and IPM-FG outperform AM in loss for each task on both training and validation datasets, improving multi-task learning performance. By comparing the relative objective values and the relative estimated stationary errors in Table 3, we observe that our SIPM-ME$^+$, SIPM-PM, SIPM-EM, and SIPM-RM yield solutions of similar

Table 3: The relative objective value and relative estimated stationary error for all methods applied to solve problem (1).

	five tasks				ten tasks			
	$m = 200$		$m = 500$		$m = 200$		$m = 500$	
	objective	stationary	objective	stationary	objective	stationary	objective	stationary
SIPM-ME^1	0.3260	2.985e-2	0.3137	3.002e-2	0.3985	8.985e-3	0.4245	3.587e-3
SIPM-ME$^+$	0.3254	8.763e-3	0.2854	8.733e-3	0.2879	4.803e-3	0.2987	9.564e-3
SIPM-PM	0.3215	8.654e-3	0.2868	1.383e-2	0.2875	5.123e-3	0.2884	1.251e-2
SIPM-EM	0.3203	8.685e-3	0.2850	8.254e-3	0.2865	4.954e-3	0.2764	8.527e-3
SIPM-RM	0.3198	8.534e-3	0.2843	8.234e-3	0.2765	4.547e-3	0.2774	8.436e-3
IPM-FG	0.3243	8.778e-3	0.2975	8.447e-3	0.2909	4.987e-3	0.2894	8.554e-3
AM	0.3535	–	0.3743	–	0.4043	–	0.4253	–

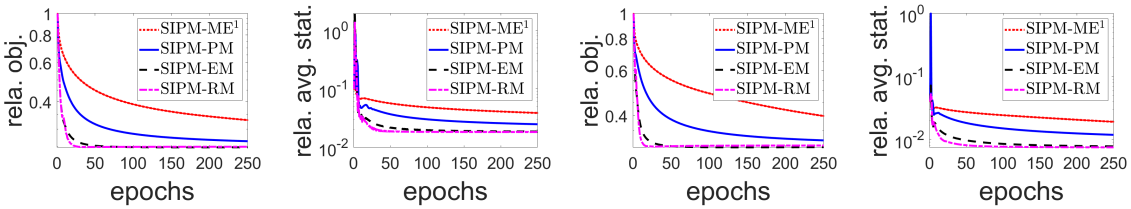

Figure 3: Convergence behavior of the relative objective value and average relative stationary error for each epoch. The first two figures correspond to the problem with five tasks, and the last two correspond to the problem with ten tasks.

quality compared to IPM-FG. In addition, SIPM-ME^1 converges slowly and returns suboptimal solutions, which corroborates the theoretical results that incorporating momentum aids gradient estimation and improves solution quality. We also observe that the solution quality, in terms of relative objective value, of AM is generally worse than that of our SIPM variants. From Figure 3, we observe that SIPM-ME^1 converges more slowly than the other three variants, and SIPM-RM is faster than SIPM-PM and SIPM-EM, which corroborates our established iteration complexity for these methods.

4.3 Clustering Data Streams

In this subsection, we consider the problem of clustering data streams (Bidaurrazaga et al., 2021), which aims at assigning d points to k clusters, with each point having p data observations that arrive continuously. Applying the semidefinite relaxation from Peng and Wei (2007) to this clustering problem results in:

$$\min_{W\in\mathbb{R}^{d\times d}} \frac{1}{p}\sum_{i=1}^{p}\langle A_i, W\rangle + \tau\sum_{i=1}^{d}\ln(\gamma+\lambda_i(W)) \;\text{ s.t. }\; W\in\mathbb{S}^d_+,\;\; We_d = e_d,\;\; \langle I_d, W\rangle = k, \quad (45)$$

where $\{A_i\}_{i=1}^p$ are computed from data streams, $\tau\sum_{i=1}^d \ln(\gamma+\lambda_i(W))$ is a nonconvex regularizer that imposes low rankness, with $\lambda_i(W)$ being the ith largest eigenvalue of W, and τ and γ being tuning parameters (Lu et al., 2014), and e_d and I_d denote the d-dimensional

all-one vector and the $d \times d$ identity matrix, respectively. We consider two typical classification datasets, namely, 'spam-base' and 'cover-type', from the UCI repository. To simulate stream scenarios, we apply $(1+\varepsilon)$-drifts adapted from Bidaurrazaga et al. (2021) to each dataset, where ε is randomly drawn from a standard Gaussian distribution.

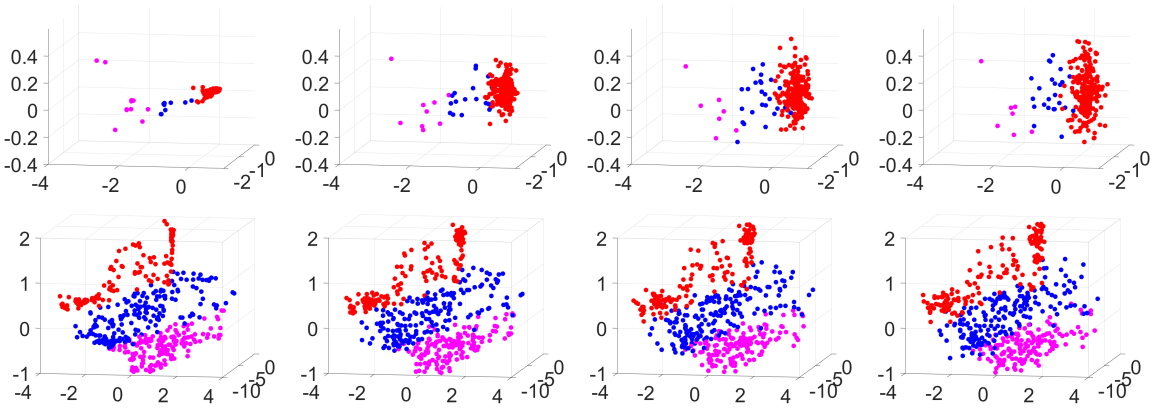

Figure 4: Visualization of the clustering results obtained by solving (45) using our SIPMs at the 1st, 333rd, 666th, and 1000th data observations in the stream (from left to right). The first and second rows display the clustering results for the 'spam-base' and 'cover-type' datasets, respectively.

Table 4: The relative objective value and relative estimated stationary error for all methods applied to solve problem (45).

	spam-base				cover-type			
	$(d,p)=(100,100)$		$(d,p)=(500,100)$		$(d,p)=(100,500)$		$(d,p)=(500,500)$	
	objective	stationary	objective	stationary	objective	stationary	objective	stationary
SIPM-ME^1	-19.14	3.108e-1	-17.21	1.544e-1	0.2909	6.822e-2	0.2302	3.996e-2
SIPM-ME$^+$	-19.15	2.950e-2	-17.24	3.789e-2	0.2908	6.165e-3	0.2299	9.096e-3
SIPM-PM	-19.15	2.952e-2	-17.23	3.790e-2	0.2908	6.165e-3	0.2297	1.108e-2
SIPM-EM	-19.15	2.948e-2	-17.23	3.611e-2	0.2908	6.163e-3	0.2297	9.179e-3
SIPM-RM	-19.15	2.947e-2	-17.24	3.609e-2	0.2908	6.163e-3	0.2296	9.096e-3
IPM-FG	-19.15	2.950e-2	-17.24	3.623e-2	0.2909	6.235e-3	0.2297	9.026e-3

We apply our SIPMs and IPM-FG to solve (45). For SIPMs, we set the barrier function as $B(\Sigma) = -\ln(\det(\Sigma))$, associated with the positive semidefinite cone $\mathbb{S}^p_+ \doteq \{\Sigma \succeq 0\}$. For SIPM-ME^1, SIPM-PM, SIPM-EM, and SIPM-RM, we set the batch size as 10 and 50 for the case when p is 100 and 500, respectively, and set the maximum number of epochs as 100. For SIPM-ME$^+$, we initialize the batch size as 1 and increase it by 1 per iteration. For a fair comparison, we set the maximum number of iterations for SIPM-ME$^+$ and IPM-FG such that the total number of data points used during training equals that of the other methods in 100 epochs. For all methods, we set the initial point W^0 with all diagonal elements equal to k/d and all other elements equal to $(d-k)/(d(d-1))$. We set the other

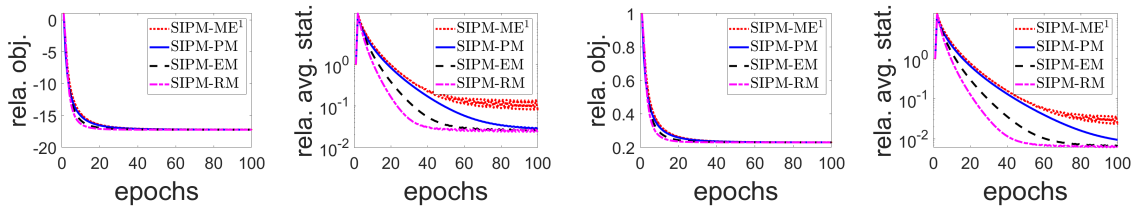

Figure 5: Convergence behavior of the relative objective value and average relative stationary error for each epoch. The first two and last two plots correspond to the 'spam-base' and 'cover-type' datasets, respectively.

hyperparameters—including step sizes, momentum parameters, and barrier parameters—according to the strategy described in Section 4.1.

From Figure 4, we observe that the solutions obtained using our SIPMs on (45) effectively cluster the data streams. By comparing the relative objective values and the relative estimated stationary errors in Table 4, we observe that our proposed SIPM-ME$^+$, SIPM-PM, SIPM-EM, and SIPM-RM yield solutions of similar quality to the deterministic variant IPM-FG. In addition, SIPM-ME^1 converges slowly and returns suboptimal solutions, which corroborates the theoretical results that incorporating momentum greatly facilitates gradient estimation, thereby improving solution quality. From Figure 5, we see that SIPM-ME^1 converges much more slowly than the other three variants, while SIPM-RM is slightly faster than SIPM-PM and SIPM-EM. This observation corroborates the iteration complexity we established for these methods.

5. Proof of the Main Results

In this section, we provide proofs of our main results presented in Sections 2 and 3, which are particularly, Theorems 1, 6, 7, 11, 15, 16, 18, 22 and 24, and Theorems 8, 9, 12, 13, 17, 19, 23 and 25.

We start with the next lemma regarding the properties of the ϑ-LHSC barrier function.

Lemma 27 *Let $x \in \mathrm{int}\mathcal{K}$ and $s_\eta \in (0,1)$ be given. Then the following statements hold for the ϑ-LHSC barrier function B.*

(i) $(\|\nabla B(x)\|_x^*)^2 = -x^T\nabla B(x) = \|x\|_x^2 = \vartheta$.

(ii) $\{y : \|y-x\|_x < 1\} \subset \mathrm{int}\mathcal{K}$ *and* $\{s : \|s+\nabla B(x)\|_x^* \le 1\} \subseteq \mathcal{K}^*$.

(iii) *For any y satisfying $\|y-x\|_x < 1$, the relation $(1-\|y-x\|_x)\|v\|_x^* \le \|v\|_y^* \le (1-\|y-x\|_x)^{-1}\|v\|_x^*$ holds for all $v \in \mathbb{R}^n$.*

(iv) $\|\nabla B(y) - \nabla B(x)\|_x^* \le \|y-x\|_x/(1-s_\eta)$ *holds for all y with $\|y-x\|_x \le s_\eta$.*

Proof The proof of statements (i)-(iii) can be found in Lemma 1 in He and Lu (2023).

We next prove statement (iv). Let y satisfy $\|y-x\|_x \le s_\eta$. Using this and Eq.(2.3) in Nemirovski (2004), we obtain that $|z^T(\nabla B(y) - \nabla B(x))| \le \|y-x\|_x/(1-s_\eta)$ holds

for all z satisfying $\|z\|_x \le 1$. In addition, notice that $\max_{\|z\|_x\le 1} |z^T(\nabla B(y) - \nabla B(x))| = \|\nabla B(y) - \nabla B(x)\|_x^*$. Combining these, we conclude that statement (iv) holds as desired. ■

5.1 Proof of Theorems 1 and 6

In this subsection, we prove Theorems 1 and 6.

Proof of Theorem 1 It suffices to prove that the last two relations in (7) hold. Using $\mu > 0$, $\|\nabla\phi_\mu(x) + A^T\lambda\|_x^* \le \mu$, and the definition of ϕ_μ in (4), we have $\|((1+\mu)\nabla f(x) + A^T\lambda)/\mu + \nabla B(x)\|_x^* \le 1$, which together with Theorem 27(ii) implies that $((1+\mu)\nabla f(x) + A^T\lambda)/\mu \in \mathcal{K}^*$. Thus, we observe that x satisfies the third relation in (7) with $\tilde\lambda = \lambda/(1+\mu)$.

We next show that x satisfies the fourth relation in (7) with $\tilde\lambda = \lambda/(1+\mu)$ and $\epsilon \ge (1+\sqrt{\vartheta})\mu$. Using $\|\nabla\phi_\mu(x) + A^T\lambda\|_x^* \le \mu$ and $\|\nabla B(x)\|_x^* = \sqrt{\vartheta}$ from Theorem 27(i), we have $\mu \ge \|(1+\mu)\nabla f(x) + A^T\lambda\|_x^* - \mu\|\nabla B(x)\|_x^* = \|(1+\mu)\nabla f(x) + A^T\lambda\|_x^* - \mu\sqrt{\vartheta}$. Therefore, it follows that $\|\nabla f(x) + A^T\lambda/(1+\mu)\|_x^* \le (1+\sqrt{\vartheta})\mu/(1+\mu) < (1+\sqrt{\vartheta})\mu$, which implies that x satisfies the fourth relation in (7) with $\tilde\lambda = \lambda/(1+\mu)$ and any $\epsilon \ge (1+\sqrt{\vartheta})\mu$. Hence, the proof is complete. ■

Proof of Theorem 6 Notice from the update of x^{k+1} in (15) that

$$\|x^{k+1} - x^k\|_{x^k} \overset{(15)}{=} \eta_k \frac{\|H_k(m^k + A^T\lambda^k)\|_{x^k}}{\|m^k + A^T\lambda^k\|_{x^k}^*} \overset{(5)}{=} \eta_k \le s_\eta < 1.$$

Hence, $\|x^{k+1} - x^k\|_{x^k} = \eta_k$ for all $k \ge 0$. We now prove $x^k \in \Omega^\circ$ for all $k \ge 0$ by induction. Note from Algorithm 1 that $x^0 \in \Omega^\circ$. Suppose $x^k \in \Omega^\circ$ for some $k \ge 0$. We next prove $x^{k+1} \in \Omega^\circ$. Since $x^k \in \mathrm{int}\mathcal{K}$ and $\|x^{k+1} - x^k\|_{x^k} < 1$, it then follows from Theorem 27(ii) that $x^{k+1} \in \mathrm{int}\mathcal{K}$. In addition, by $Ax^k = b$ and (15), one has that

$$Ax^{k+1} \overset{(15)}{=} Ax^k - \eta_k \frac{AH_k(m^k - A^T(AH_kA^T)^{-1}AH_km^k)}{\|m^k + A^T\lambda^k\|_{x^k}^*} = Ax^k = b.$$

This together with $x^{k+1} \in \mathrm{int}\mathcal{K}$ implies that $x^{k+1} \in \Omega^\circ$, which completes the induction. ■

5.2 Proof of Some Auxiliary Lemmas

In this subsection, we prove several auxiliary lemmas. The following lemma shows that ϕ_μ is locally Lipschitz continuous under Assumption 1(b) and a useful descent inequality holds.

Lemma 28 *Suppose that Assumption 1(b) holds. Let $\mu \in (0,1]$. Then,*

$$\|\nabla\phi_\mu(y) - \nabla\phi_\mu(x)\|_x^* \le L_\phi\|y - x\|_x \qquad \forall x, y \in \Omega^\circ \text{ with } \|y - x\|_x \le s_\eta, \tag{46}$$

$$\phi_\mu(y) \le \phi_\mu(x) + \nabla\phi_\mu(x)^T(y - x) + L_\phi\|y - x\|_x^2/2 \qquad \forall x, y \in \Omega^\circ \text{ with } \|y - x\|_x \le s_\eta, \tag{47}$$

where ϕ_μ and Ω° are defined in (4), and L_ϕ is defined in (14).

Proof Fix any $x, y \in \Omega^\circ$ satisfying $y \in \{y : \|y - x\|_x \le s_\eta\}$. Using the definition of ϕ_μ in (4), $\mu \in (0,1]$, Assumption 1(b), and Theorem 27(iv), we obtain that (46) holds.

We next prove (47). Indeed, one has

$$\begin{aligned}\phi_\mu(y) - \phi_\mu(x) - \nabla\phi_\mu(x)^T(y-x) &= \int_0^1 (\nabla\phi_\mu(x+t(y-x)) - \nabla\phi_\mu(x))^T(y-x)\mathrm{d}t \\ &\overset{(5)}{\leq} \int_0^1 \|\nabla\phi_\mu(x+t(y-x)) - \nabla\phi_\mu(x)\|_x^* \mathrm{d}t \|y-x\|_x \overset{(46)}{\leq} \frac{L_\phi}{2}\|y-x\|_x^2.\end{aligned}$$

Hence, (47) holds as desired. ■

The following lemma establishes a key inequality under Assumption 2, whose proof is identical to that of Lemma 3 in He and Lu (2023) and is therefore omitted here.

Lemma 29 *Suppose that Assumption 2 holds. Then,*

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x)\|_x^* \leq \frac{L_2}{2}\|y-x\|_x^2 \qquad \forall x, y \in \Omega^\circ \text{ with } \|y-x\|_x \leq s_\eta,$$

where Ω° is defined in (4), L_2 is given in Assumption 2(b), and s_η is an input of Algorithm 1.

The following lemma concerns the descent of ϕ_μ for iterates generated by Algorithm 1.

Lemma 30 *Suppose that Assumption 1 holds. Let $\{(x^k, \lambda^k)\}_{k\geq 0}$ be the sequence generated by Algorithm 1 with input parameters $\{(\eta_k, \mu_k)\}$. Then, for all $k \geq 0$, it holds that*

$$\phi_{\mu_k}(x^{k+1}) \leq \phi_{\mu_k}(x^k) - \eta_k\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^* + 4\eta_k\|\nabla f(x^k) - \overline{m}^k\|_{x^k}^* + \frac{L_\phi}{2}\eta_k^2, \quad (48)$$

where ϕ_{μ_k} and L_ϕ are defined in (4) and (14), respectively.

Proof We fix an arbitrary $k \geq 0$. Recall from Theorem 6 that $\|x^{k+1} - x^k\|_{x^k} = \eta_k$ and $x^k, x^{k+1} \in \Omega^\circ$. It then follows that $A(x^{k+1} - x^k) = 0$, and from (15) that

$$\langle m^k + A^T\lambda^k, x^{k+1} - x^k\rangle \overset{(15)}{=} -\eta_k \frac{(m^k + A^T\lambda^k)^T H_k (m^k + A^T\lambda^k)}{\|m^k + A^T\lambda^k\|_{x^k}^*} \overset{(5)}{=} -\eta_k\|m^k + A^T\lambda^k\|_{x^k}^*. \quad (49)$$

In view of these and (47) with $(x, y, \mu, \eta) = (x^k, x^{k+1}, \mu_k, \eta_k)$, one can see that

$$\begin{aligned}\phi_{\mu_k}(x^{k+1}) &\leq \phi_{\mu_k}(x^k) + \langle\nabla\phi_{\mu_k}(x^k), x^{k+1} - x^k\rangle + \frac{L_\phi}{2}\|x^{k+1} - x^k\|_{x^k}^2 \\ &= \phi_{\mu_k}(x^k) + \langle m^k + A^T\lambda^k, x^{k+1} - x^k\rangle + \langle\nabla\phi_{\mu_k}(x^k) - m^k, x^{k+1} - x^k\rangle + \frac{L_\phi}{2}\eta_k^2 \\ &\overset{(49)}{=} \phi_{\mu_k}(x^k) - \eta_k\|m^k + A^T\lambda^k\|_{x^k}^* + \langle\nabla\phi_{\mu_k}(x^k) - m^k, x^{k+1} - x^k\rangle + \frac{L_\phi}{2}\eta_k^2 \\ &\leq \phi_{\mu_k}(x^k) - \eta_k\|m^k + A^T\lambda^k\|_{x^k}^* + \eta_k\|\nabla\phi_{\mu_k}(x^k) - m^k\|_{x^k}^* + \frac{L_\phi}{2}\eta_k^2 \\ &\leq \phi_{\mu_k}(x^k) - \eta_k\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^* + 2\eta_k\|\nabla\phi_{\mu_k}(x^k) - m^k\|_{x^k}^* + \frac{L_\phi}{2}\eta_k^2,\end{aligned}$$

$$= \phi_{\mu_k}(x^k) - \eta_k\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^* + 2\eta_k(1+\mu_k)\|\nabla f(x^k) - \overline{m}^k\|_{x^k}^* + \frac{L_\phi}{2}\eta_k^2,$$

where the first equality is due to $\|x^{k+1} - x^k\|_{x^k} = \eta_k$ and $A(x^{k+1} - x^k) = 0$, the second inequality follows from the Cauchy-Schwarz inequality and $\|x^{k+1} - x^k\|_{x^k} = \eta_k$, the last inequality is due to the triangular inequality, and the last equality is due to the definition of ϕ_μ and $m^k = \overline{m}^k + \mu_k(\overline{m}^k + \nabla B(x^k))$. This together with $\mu_k \le 1$ proves this lemma as desired. ■

We next present a lemma regarding the estimation of the partial sums of series.

Lemma 31 *Let $\zeta(\cdot)$ be a convex univariate function. Then, for any integers a, b satisfying $[a - 1/2, b + 1/2] \subset \mathrm{dom}\zeta$, it holds that $\sum_{p=a}^{b}\zeta(p) \le \int_{a-1/2}^{b+1/2}\zeta(\tau)\mathrm{d}\tau$.*

Proof Since ζ is convex, one has $\sum_{p=a}^{b}\zeta(p) \le \sum_{p=a}^{b}\int_{p-1/2}^{p+1/2}\zeta(\tau)\mathrm{d}\tau = \int_{a-1/2}^{b+1/2}\zeta(\tau)\mathrm{d}\tau$. ■

As a consequence of Theorem 31, we consider $\zeta(\tau) = 1/\tau^\alpha$ for some $\alpha \in (0, \infty]$, where $\tau \in (0, \infty)$. Then, for any positive integers a, b, one has

$$\sum_{p=a}^{b} 1/p^\alpha \le \begin{cases} \ln(b+1/2) - \ln(a-1/2) & \text{if } \alpha = 1, \\ \frac{1}{1-\alpha}((b+1/2)^{1-\alpha} - (a-1/2)^{1-\alpha}) & \text{if } \alpha \in (0,1)\cup(1,+\infty). \end{cases} \tag{50}$$

We next provide an auxiliary lemma that will be used to analyze the complexity involving polylogarithmic terms for our methods.

Lemma 32 *Let $\beta \in (0,1)$ and $u \in (0, 1/e)$ be given. Then, $1/v^\beta \ln v \le 2u/\beta$ holds for all $v \ge (1/u\ln(1/u))^{1/\beta}$.*

Proof Fix any v satisfying $v \ge (1/u\ln(1/u))^{1/\beta}$. It then follows from $u \in (0, 1/e)$ that

$$v \ge (1/u\ln(1/u))^{1/\beta} > e^{1/\beta}. \tag{51}$$

Denote $\phi(v) \doteq 1/v^\beta \ln v$. One can verify that ϕ is decreasing over $(e^{1/\beta}, \infty)$. Using this and (51), we obtain that

$$1/v^\beta \ln v = \phi(v) \le \phi((1/u\ln(1/u))^{1/\beta}) = \frac{u}{\beta}\Big(1 + \frac{\ln\ln(1/u)}{\ln(1/u)}\Big) \le \frac{2u}{\beta},$$

where the last inequality is due to $\ln\ln(1/u) \le \ln(1/u)$ for all $u \in (0, 1/e)$. Hence, the conclusion of this lemma holds as desired. ■

5.3 Proof of the Main Results in Section 3.1

Proof of Theorem 7 Fix any $k \ge 0$. It follows from (10) and (16) that

$$\mathbb{E}_{\{\xi_i^k\}_{i\in\mathscr{B}_k}}[(\|\overline{m}^k - \nabla f(x^k)\|_{x^k}^*)^2] = \frac{1}{|\mathscr{B}_k|^2}\sum_{i\in\mathscr{B}_k}\mathbb{E}_{\xi_i^k}[(\|G(x^k;\xi_i^k) - \nabla f(x^k)\|_{x^k}^*)^2] \le \frac{\sigma^2}{|\mathscr{B}_k|},$$

where the first equality follows from (16) and the first relation in (10), and the last inequality follows from the second relation in (10). Hence, (17) holds as desired. ■

Proof of Theorem 8 By summing (48) over $k = 0, \ldots, K-1$ and taking expectation on $\{\xi_i^k\}_{i\in\mathscr{B}_k, 1\le k\le K-1}$, we have

$$\begin{aligned}
&\sum_{k=0}^{K-1} \eta_k \mathbb{E}[\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^*] \le \phi_{\mu_0}(x^0) - \mathbb{E}[\phi_{\mu_K}(x^K)] + \sum_{k=0}^{K-1}(\mu_{k+1}-\mu_k)\mathbb{E}[f(x^{k+1}) + B(x^{k+1})] \\
&\quad + 4\sum_{k=0}^{K-1} \eta_k \mathbb{E}[\|\nabla f(x^k) - \overline{m}^k\|_{x^k}^*] + \frac{L_\phi}{2}\sum_{k=0}^{K-1}\eta_k^2 \\
&\le \phi_{\mu_0}(x^0) - \mathbb{E}[\phi_{\mu_K}(x^K)] + (\mu_K - \mu_0)\phi_{\text{low}} + 4\sum_{k=0}^{K-1} \eta_k \mathbb{E}[\|\nabla f(x^k) - \overline{m}^k\|_{x^k}^*] + \frac{L_\phi}{2}\sum_{k=0}^{K-1}\eta_k^2 \\
&\le \phi_{\mu_0}(x^0) - \mathbb{E}[\phi_{\mu_K}(x^K)] + (\mu_K - \mu_0)\phi_{\text{low}} + 4\sigma\sum_{k=0}^{K-1} \frac{\eta_k}{|\mathscr{B}_k|^{1/2}} + \frac{L_\phi}{2}\sum_{k=0}^{K-1}\eta_k^2, \\
&\le f(x^0) + \mu_0(f(x^0) + B(x^0)) - (1+\mu_0)\phi_{\text{low}} + 4\sigma\sum_{k=0}^{K-1} \frac{\eta_k}{|\mathscr{B}_k|^{1/2}} + \frac{L_\phi}{2}\sum_{k=0}^{K-1}\eta_k^2 \\
&\le f(x^0) + [f(x^0) + B(x^0)]_+ - 2[\phi_{\text{low}}]_- + 4\sigma\sum_{k=0}^{K-1} \frac{\eta_k}{|\mathscr{B}_k|^{1/2}} + \frac{L_\phi}{2}\sum_{k=0}^{K-1}\eta_k^2,
\end{aligned}$$

where the first inequality follows from (48), the second inequality follows from $\mu_k - \mu_{k+1} \ge 0$ and $f(x^{k+1}) + B(x^{k+1}) \ge \phi_{\text{low}}$ (due to (8) and $x^{k+1} \in \Omega^\circ$), the third inequality follows from $\mathbb{E}[X]^2 \le \mathbb{E}[X^2]$ for any random variable X and (17), the fourth inequality is due to the definitions of ϕ_μ and (12), and the last inequality is due to $\mu_0 \in (0, 1]$. Using the above inequality, (13), and the fact that $\{\eta_k\}_{k\ge 0}$ is nonincreasing, we have

$$\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^*] \le \frac{\Delta(x^0)}{\eta_{K-1}} + \frac{1}{\eta_{K-1}}\sum_{k=0}^{K-1}\eta_k(4\sigma/|\mathscr{B}_k|^{1/2} + L_\phi\eta_k/2).$$

Hence, (18) holds as desired. ■

Proof of Theorem 9 Substituting (19) into (18), we obtain that for all $K \ge 3$,

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^*] \overset{(18)(19)}{\le} \frac{K^{1/2}\Delta(x^0)}{s_\eta} + K^{1/2}\sum_{k=0}^{K-1}\frac{8\sigma + s_\eta L_\phi}{2(k+1)} \\
&\le \frac{K^{1/2}\Delta(x^0)}{s_\eta} + \Big(4\sigma + \frac{s_\eta L_\phi}{2}\Big)K^{1/2}\ln(2K+1) \le \Big(\frac{\Delta(x^0)}{s_\eta} + 8\sigma + s_\eta L_\phi\Big)K^{1/2}\ln K \\
&\overset{(20)}{=} M_{\text{me}}K^{1/2}\ln K/2,
\end{aligned} \tag{52}$$

where the second inequality follows from $\sum_{k=0}^{K-1} 1/(k+1) \le \ln(2K+1)$ due to (50) with $(a, b, \alpha) = (1, K, 1)$, and the last inequality follows from $1 \le \ln K$ and $\ln(2K+1) \le 2\ln K$

given that $K \geq 3$. Since $\kappa(K)$ is uniformly drawn from $\{\lfloor K/2 \rfloor, \ldots, K-1\}$, we have that for all $K \geq 3$,

$$\mathbb{E}[\|\nabla \phi_{\mu_{\kappa(K)}}(x^{\kappa(K)}) + A^T \lambda^{\kappa(K)}\|^*_{x^{\kappa(K)}}] = \frac{1}{K - \lfloor K/2 \rfloor} \sum_{k=\lfloor K/2 \rfloor}^{K-1} \mathbb{E}[\|\nabla \phi_{\mu_k}(x^k) + A^T \lambda^k\|^*_{x^k}]$$

$$\leq \frac{2}{K} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla \phi_{\mu_k}(x^k) + A^T \lambda^k\|^*_{x^k}] \overset{(52)}{\leq} M_{\rm me} K^{-1/2} \ln K. \tag{53}$$

By Theorem 32 with $(\beta, u, v) = (1/2, \epsilon/(4M_{\rm me}(1+\sqrt{\vartheta})), K)$, one can see that

$$K^{-1/2} \ln K \leq \frac{\epsilon}{M_{\rm me}(1+\sqrt{\vartheta})} \qquad \forall K \geq \Big(\frac{4M_{\rm me}(1+\sqrt{\vartheta})}{\epsilon} \ln\Big(\frac{4M_{\rm me}(1+\sqrt{\vartheta})}{\epsilon}\Big)\Big)^2,$$

which together with (53) implies that

$$\begin{aligned} &\mathbb{E}[\|\nabla \phi_{\mu_{\kappa(K)}}(x^{\kappa(K)}) + A^T \lambda^{\kappa(K)}\|^*_{x^{\kappa(K)}}] \leq \frac{\epsilon}{1+\sqrt{\vartheta}} \\ &\forall K \geq \max\Big\{\Big(\frac{4M_{\rm me}(1+\sqrt{\vartheta})}{\epsilon} \ln\Big(\frac{4M_{\rm me}(1+\sqrt{\vartheta})}{\epsilon}\Big)\Big)^2, 3\Big\}. \end{aligned} \tag{54}$$

On the other hand, when $K \geq 2((1+\sqrt{\vartheta})/\epsilon)^2$, by the definition of $\{\mu_k\}_{k\geq 0}$ in (19) and the fact that $\kappa(K)$ is uniformly selected from $\{\lfloor K/2 \rfloor, \ldots, K-1\}$, one has that $\mu_{\kappa(K)} = \mu_{\lfloor K/2 \rfloor} = \epsilon/(1+\sqrt{\vartheta})$. Combining this with (54), we obtain that (21) holds as desired, and the proof of this theorem is complete. ■

5.4 Proof of the Main Results in Section 3.2

Proof of Theorem 11 Fix any $k \geq 0$. Recall from Theorem 6 that $\|x^{k+1} - x^k\|_{x^k} = \eta_k$. Using this and (22), we have that

$$\begin{aligned} &\mathbb{E}_{\xi^{k+1}}[(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|^*_{x^{k+1}})^2] \\ &\overset{(22)}{=} \mathbb{E}_{\xi^{k+1}}[(\|(1-\gamma_k)(\overline{m}^k - \nabla f(x^{k+1})) + \gamma_k(G(x^{k+1}, \xi^{k+1}) - \nabla f(x^{k+1}))\|^*_{x^{k+1}})^2] \\ &\overset{(10)}{\leq} (1-\gamma_k)^2(\|\overline{m}^k - \nabla f(x^{k+1})\|^*_{x^{k+1}})^2 + \sigma^2\gamma_k^2 \\ &\leq (1-\alpha_k)^2(\|\overline{m}^k - \nabla f(x^{k+1})\|^*_{x^k})^2 + \sigma^2\gamma_k^2, \end{aligned} \tag{55}$$

where the last inequality follows from Theorem 27(iii) and the definition of α_k. Also, notice that for all $a > 0$,

$$\begin{aligned} &(1-\alpha_k)^2(\|\overline{m}^k - \nabla f(x^{k+1})\|^*_{x^k})^2 \\ &\leq (1-\alpha_k)^2(1+a)(\|\overline{m}^k - \nabla f(x^k)\|^*_{x^k})^2 + (1-\alpha_k)^2(1+1/a)(\|\nabla f(x^{k+1}) - \nabla f(x^k)\|^*_{x^k})^2 \\ &\leq (1-\alpha_k)^2(1+a)(\|\overline{m}^k - \nabla f(x^k)\|^*_{x^k})^2 + (1-\alpha_k)^2(1+1/a)L_1^2\eta_k^2, \end{aligned}$$

where the first inequality is due to $\|u+v\|^2 \leq (1+a)\|u\|^2 + (1+1/a)\|v\|^2$ for all $u, v \in \mathbb{R}^n$ and $a > 0$, and the last inequality follows from (9) and $\|x^{k+1} - x^k\|_{x^k} = \eta_k$. Letting

$a = \alpha_k/(1-\alpha_k)$ and combining this inequality with (55), we obtain that

$$\mathbb{E}_{\xi^{k+1}}[(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|_{x^{k+1}}^*)^2] \le (1-\alpha_k)(\|\overline{m}^k - \nabla f(x^k)\|_{x^k}^*)^2 + \frac{(1-\alpha_k)^2 L_1^2 \eta_k^2}{\alpha_k} + \sigma^2\gamma_k^2.$$

Since $\gamma_k > \eta_k$, we have $\alpha_k \in (0,1)$. This along with the above inequality implies that (23) holds as desired. ■

Proof of Theorem 12 For convenience, we define the following potentials:

$$P_k \doteq \phi_{\mu_k}(x^k) + (\|\overline{m}^k - \nabla f(x^k)\|_{x^k}^*)^2/L_1 \qquad \forall k \ge 0. \tag{56}$$

Recall from Algorithm 1 that $\{\mu_k\}$ is nonincreasing. By these, (8), (23), and (48), one has that for all $k \ge 0$,

$$\begin{aligned}
\mathbb{E}_{\xi^{k+1}}[P_{k+1}] &\overset{(56)}{=} \mathbb{E}_{\xi^{k+1}}[\phi_{\mu_{k+1}}(x^{k+1}) + (\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|_{x^{k+1}}^*)^2/L_1] \\
&= (\mu_{k+1}-\mu_k)(f(x^{k+1}) + B(x^{k+1})) + \mathbb{E}_{\xi^{k+1}}[\phi_{\mu_k}(x^{k+1}) + (\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|_{x^{k+1}}^*)^2/L_1] \\
&\overset{(8)(23)(48)}{\le} (\mu_{k+1}-\mu_k)\phi_{\rm low} + \phi_{\mu_k}(x^k) - \eta_k\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^* + 4\eta_k\|\overline{m}^k - \nabla f(x^k)\|_{x^k}^* \\
&\quad + \frac{L_\phi}{2}\eta_k^2 + (1-\alpha_k)(\|\overline{m}^k - \nabla f(x^k)\|_{x^k}^*)^2/L_1 + \frac{L_1\eta_k^2}{\alpha_k} + \frac{\sigma^2\gamma_k^2}{L_1}.
\end{aligned} \tag{57}$$

In addition, notice that $4\eta_k\|\overline{m}^k - \nabla f(x^k)\|_{x^k}^* \le 4L_1\eta_k^2/\alpha_k + \alpha_k(\|\overline{m}^k - \nabla f(x^k)\|_{x^k}^*)^2/L_1$, which together with (56) and (57) implies that for all $k \ge 0$,

$$\mathbb{E}_{\xi^{k+1}}[P_{k+1}] \le (\mu_{k+1}-\mu_k)\phi_{\rm low} + P_k - \eta_k\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^* + \frac{L_\phi}{2}\eta_k^2 + \frac{5L_1\eta_k^2}{\alpha_k} + \frac{\sigma^2\gamma_k^2}{L_1}. \tag{58}$$

On the other hand, by (10), (12), (56), $\overline{m}^0 = G(x^0,\xi^0)$, and $\mu_0 \le 1$, one has

$$\begin{aligned}
&\mathbb{E}_{\xi^0}[P_0] = \phi_{\mu_0}(x^0) + \mathbb{E}_{\xi^0}[(\|\overline{m}^0 - \nabla f(x^0)\|_{x^0}^*)^2]/L_1 \le f(x^0) + [f(x^0) + B(x^0)]_+ + \sigma^2/L_1, \\
&\mathbb{E}_{\{\xi^k\}_{k=0}^K}[P_K] = \phi_{\mu_K}(x^K) + \mathbb{E}_{\{\xi^k\}_{k=0}^K}[(\|\overline{m}^K - \nabla f(x^K)\|_{x^K}^*)^2]/L_1 \overset{(12)}{\ge} (1+\mu_K)\phi_{\rm low}.
\end{aligned}$$

By summing (58) over $k = 0,\ldots,K-1$, and using the above two inequalities, (13), and the fact that $\{\eta_k\}_{k\ge0}$ is nonincreasing, we obtain that

$$\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^*] \le \frac{\Delta(x^0) + \sigma^2/L_1}{\eta_{K-1}} + \frac{1}{\eta_{K-1}}\sum_{k=0}^{K-1}\left(\frac{L_\phi}{2}\eta_k^2 + \frac{5L_1\eta_k^2}{\alpha_k} + \frac{\sigma^2\gamma_k^2}{L_1}\right).$$

Hence, the conclusion of this theorem holds. ■

Proof of Theorem 13 Observe from (25) that $\gamma_k > \eta_k$ for all $k \ge 0$. Thus, $\{(\eta_k,\gamma_k)\}_{k\ge0}$ defined as in (25) satisfies the assumption on $\{(\eta_k,\gamma_k)\}_{k\ge0}$ in Theorem 12. Substituting (25) and (26) into (24), we obtain for all $K \ge 3$,

$$\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^*]$$

$$\overset{(24)(25)(26)}{\leq} K^{3/4}\Big(\frac{\Delta(x^0)+\sigma^2/L_1}{s_\eta}+\sum_{k=0}^{K-1}\Big(\frac{s_\eta L_\phi}{2(k+1)^{3/2}}+\frac{5s_\eta L_1/(1-s_\eta)+\sigma^2/(s_\eta L_1)}{k+1}\Big)\Big)$$
$$< K^{3/4}\Big(\frac{\Delta(x^0)+\sigma^2/L_1}{s_\eta}+\frac{3s_\eta L_\phi}{2}+\Big(\frac{5s_\eta L_1}{1-s_\eta}+\frac{\sigma^2}{s_\eta L_1}\Big)\ln(2K+1)\Big)$$
$$\leq \Big(\frac{\Delta(x^0)+\sigma^2/L_1}{s_\eta}+\frac{3s_\eta L_\phi}{2}+2\Big(\frac{5s_\eta L_1}{1-s_\eta}+\frac{\sigma^2}{s_\eta L_1}\Big)\Big)K^{3/4}\ln K \overset{(27)}{=} M_{\rm pm}K^{3/4}\ln K/2, \quad (59)$$

where the second inequality is because $\sum_{k=0}^{K-1}1/(k+1)^{3/2}\leq 2\sqrt{2}<3$ and $\sum_{k=0}^{K-1}1/(k+1)\leq \ln(2K+1)$ due to (50) with $(a,b)=(1,K)$ and $\alpha=3/2,1$, and the last inequality is due to $1\leq \ln K$ and $\ln(2K+1)\leq 2\ln K$ given that $K\geq 3$. Since $\kappa(K)$ is uniformly drawn from $\{\lfloor K/2\rfloor,\ldots,K-1\}$, we have that for all $K\geq 3$,

$$\mathbb{E}[\|\nabla\phi_{\mu_{\kappa(K)}}(x^{\kappa(K)})+A^T\lambda^{\kappa(K)}\|^*_{x^{\kappa(K)}}]=\frac{1}{K-\lfloor K/2\rfloor}\sum_{k=\lfloor K/2\rfloor}^{K-1}\mathbb{E}[\|\nabla\phi_{\mu_k}(x^k)+A^T\lambda^k\|^*_{x^k}]$$
$$\leq \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla\phi_{\mu_k}(x^k)+A^T\lambda^k\|^*_{x^k}]\overset{(59)}{\leq} M_{\rm pm}K^{-1/4}\ln K. \quad (60)$$

By Theorem 32 with $(\beta,u,v)=(1/4,\epsilon/(8M_{\rm pm}(1+\sqrt{\vartheta})),K)$, one can see that

$$K^{-1/4}\ln K\leq \frac{\epsilon}{M_{\rm pm}(1+\sqrt{\vartheta})}\qquad \forall K\geq \Big(\frac{8M_{\rm pm}(1+\sqrt{\vartheta})}{\epsilon}\ln\Big(\frac{8M_{\rm pm}(1+\sqrt{\vartheta})}{\epsilon}\Big)\Big)^4,$$

which together with (60) implies that

$$\mathbb{E}[\|\nabla\phi_{\mu_{\kappa(K)}}(x^{\kappa(K)})+A^T\lambda^{\kappa(K)}\|^*_{x^{\kappa(K)}}]\leq\frac{\epsilon}{1+\sqrt{\vartheta}}$$
$$\forall K\geq \max\Big\{\Big(\frac{8M_{\rm pm}(1+\sqrt{\vartheta})}{\epsilon}\ln\Big(\frac{8M_{\rm pm}(1+\sqrt{\vartheta})}{\epsilon}\Big)\Big)^4,3\Big\}. \quad (61)$$

On the other hand, when $K\geq 2((1+\sqrt{\vartheta})/\epsilon)^4$, by the definition of $\{\mu_k\}_{k\geq 0}$ in (25) and the fact that $\kappa(K)$ is uniformly selected from $\{\lfloor K/2\rfloor,\ldots,K-1\}$, one has that $\mu_{\kappa(K)}=\mu_{\lfloor K/2\rfloor}=\epsilon/(1+\sqrt{\vartheta})$. Combining this with (61), we obtain that (28) holds as desired, and the proof of this theorem is complete. ■

5.5 Proof of the Main Results in Section 3.3

Proof of Theorem 15 Recall from Algorithm 1 and (29a) that $x^0=x^{-1}\in\Omega^\circ$. Let $\eta_{-1}=0$. By these and Theorem 6, we have that $\|x^{k+1}-x^k\|_{x^k}=\eta_k$ and $x^k\in\Omega^\circ$ hold for all $k\geq -1$. Fix an arbitrary $k\geq -1$, we next prove $z^{k+1}\in\Omega^\circ$. In view of $\|x^{k+1}-x^k\|_{x^k}=\eta_k$ and (29b), we observe that $\|z^{k+1}-x^k\|_{x^k}=\eta_k/\gamma_k\leq s_\eta<1$, which together with Theorem 27(ii) implies that $z^{k+1}\in{\rm int}\mathcal{K}$. In addition, by (29b) and $x^{k+1},x^k\in\Omega^\circ$, one has that $Az^{k+1}=Ax^{k+1}+[(1-\gamma_k)/\gamma_k]A(x^{k+1}-x^k)=b$. Hence, $z^{k+1}\in\Omega^\circ$, which completes the proof. ■

Proof of Theorem 16 Fix any $k \geq 0$. Observe from (29b) that $z^{k+1}-x^k = (x^{k+1}-x^k)/\gamma_k$. Recall from Theorem 6 that $\|x^{k+1}-x^k\|_{x^k} = \eta_k$. It then follows that $\|z^{k+1}-x^k\|_{x^k} = \eta_k/\gamma_k$. In addition, using the update of $\overline{m}^{k+1}$ in (29b), we obtain that

$$\begin{aligned}\overline{m}^{k+1} - \nabla f(x^{k+1}) = (1-\gamma_k)(\overline{m}^k - \nabla f(x^k)) + \gamma_k(G(z^{k+1},\xi^{k+1}) - \nabla f(z^{k+1})) \\ - (\nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)(x^{k+1}-x^k)) \\ + \gamma_k(\nabla f(z^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)(z^{k+1}-x^k)). \end{aligned} \tag{62}$$

It then follows that

$$\begin{aligned}&\mathbb{E}_{\xi^{k+1}}[(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|^*_{x^{k+1}})^2] \leq \frac{1}{(1-\eta_k)^2}\mathbb{E}_{\xi^{k+1}}[(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|^*_{x^k})^2] \\ &\overset{(62)}{=} \frac{\gamma_k^2\mathbb{E}_{\xi^{k+1}}[(\|G(z^{k+1},\xi^{k+1}) - \nabla f(z^{k+1})\|^*_{x^k})^2]}{(1-\eta_k)^2} \\ &\quad + \frac{1}{(1-\eta_k)^2}(\|(1-\gamma_k)(\overline{m}^k - \nabla f(x^k)) - (\nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)(x^{k+1}-x^k)) \\ &\qquad + \gamma_k(\nabla f(z^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)(z^{k+1}-x^k))\|^*_{x^k})^2 \\ &\leq \frac{\gamma_k^2\mathbb{E}_{\xi^{k+1}}[(\|G(z^{k+1},\xi^{k+1}) - \nabla f(z^{k+1})\|^*_{z^{k+1}})^2]}{(1-\eta_k)^2(1-\eta_k/\gamma_k)^2} + (1-\alpha_k)^2(1+a)(\|\overline{m}^k - \nabla f(x^k)\|^*_{x^k})^2 \\ &\quad + \frac{2(1+1/a)}{(1-\eta_k)^2}(\|\nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)(x^{k+1}-x^k)\|^*_{x^k})^2 \\ &\quad + \frac{2(1+1/a)\gamma_k^2}{(1-\eta_k)^2}(\|\nabla f(z^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)(z^{k+1}-x^k)\|^*_{x^k})^2, \end{aligned} \tag{63}$$

where the first inequality follows from Theorem 27(iii) and $\|x^{k+1}-x^k\|_{x^k} = \eta_k$, the equality follows from (62) and the first relation in (10), and the second inequality is due to $\|u+v\|^2 \leq (1+a)\|u\|^2+(1+1/a)\|v\|^2$ for all $u, v \in \mathbb{R}^n$ and $a > 0$, Theorem 27(iii), and $\|z^{k+1}-x^k\|_{x^k} = \eta_k/\gamma_k$. In addition, using Theorem 29, $\|x^{k+1}-x^k\|_{x^k} = \eta_k$, and $\|z^{k+1}-x^k\|_{x^k} = \eta_k/\gamma_k \leq s_\eta$, we obtain that

$$\begin{aligned}&\|\nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)(x^{k+1}-x^k)\|^*_{x^k} \leq L_2\|x^{k+1}-x^k\|^2_{x^k}/2 \leq L_2\eta_k^2/2, \\ &\|\nabla f(z^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)(z^{k+1}-x^k)\|^*_{x^k} \leq L_2\|z^{k+1}-x^k\|^2_{x^k}/2 \leq L_2\eta_k^2/(2\gamma_k^2).\end{aligned}$$

In view of (63), the above two inequalities, and the second relation in (10), and letting $a = \alpha_k/(1-\alpha_k)$, we obtain that

$$\begin{aligned}&\mathbb{E}_{\xi^{k+1}}[(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|^*_{x^{k+1}})^2] \\ &\leq \frac{\sigma^2\gamma_k^2}{(1-\eta_k)^2(1-\eta_k/\gamma_k)^2} + (1-\alpha_k)(\|\overline{m}^k - \nabla f(x^k)\|^*_{x^k})^2 + \frac{L_2^2\eta_k^4}{2(1-\eta_k)^2\alpha_k} + \frac{L_2^2\eta_k^4}{2(1-\eta_k)^2\gamma_k^2\alpha_k}.\end{aligned}$$

This along with $\gamma_k \leq 1$ and the fact that $\{\eta_k\}_{k\geq 0}$ is nonincreasing implies that this lemma holds. ■

Proof of Theorem 17 For convenience, we define the following potentials:

$$P_k \doteq \phi_{\mu_k}(x^k) + p_k(\|\overline{m}^k - \nabla f(x^k)\|^*_{x^k})^2 \qquad \forall k \geq 0. \tag{64}$$

Recall from Algorithm 1 that $\{\mu_k\}$ is nonincreasing. By these, (8), (31), and (48), one has that for all $k \geq 0$,

$$\begin{aligned}
&\mathbb{E}_{\xi^{k+1}}[P_{k+1}] \overset{(64)}{=} \mathbb{E}_{\xi^{k+1}}[\phi_{\mu_{k+1}}(x^{k+1}) + p_{k+1}(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|_{x^{k+1}}^*)^2] \\
&= (\mu_{k+1} - \mu_k)(f(x^{k+1}) + B(x^{k+1})) + \mathbb{E}_{\xi^{k+1}}[\phi_{\mu_k}(x^{k+1}) + p_{k+1}(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|_{x^{k+1}}^*)^2] \\
&\overset{(8)(31)(48)}{\leq} (\mu_{k+1} - \mu_k)\phi_{\rm low} + \phi_{\mu_k}(x^k) - \eta_k\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^* + 4\eta_k\|\nabla f(x^k) - \overline{m}^k\|_{x^k}^* + \frac{L_\phi}{2}\eta_k^2 \\
&+ (1-\alpha_k)p_{k+1}(\|\overline{m}^k - \nabla f(x^k)\|_{x^k}^*)^2 + \frac{L_2^2\eta_k^4 p_{k+1}}{(1-\eta_0)^2\gamma_k^2\alpha_k} + \frac{\sigma^2\gamma_k^2 p_{k+1}}{(1-\eta_0)^2(1-\eta_k/\gamma_k)^2} \\
&\leq (\mu_{k+1} - \mu_k)\phi_{\rm low} + \phi_{\mu_k}(x^k) - \eta_k\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^* + 4\eta_k\|\nabla f(x^k) - \overline{m}^k\|_{x^k}^* + \frac{L_\phi}{2}\eta_k^2 \\
&+ (1-\alpha_k/2)p_k(\|\overline{m}^k - \nabla f(x^k)\|_{x^k}^*)^2 + \frac{L_2^2\eta_k^4 p_{k+1}}{(1-\eta_0)^2\gamma_k^2\alpha_k} + \frac{\sigma^2\gamma_k^2 p_{k+1}}{(1-\eta_0)^2(1-\eta_k/\gamma_k)^2},
\end{aligned}$$

where the second inequality follows from $(1-\alpha_k)p_{k+1} \leq (1-\alpha_k/2)p_k$. In addition, note that $4\eta_k\|\nabla f(x^k) - \overline{m}^k\|_{x^k}^* \leq \alpha_k p_k(\|\overline{m}^k - \nabla f(x^k)\|_{x^k}^*)^2/2 + 8\eta_k^2/(p_k\alpha_k)$, which together with (64) and the above inequality implies that for all $k \geq 0$,

$$\begin{aligned}
\mathbb{E}_{\xi^{k+1}}[P_{k+1}] \overset{(64)}{\leq}\ & (\mu_{k+1} - \mu_k)\phi_{\rm low} + P_k - \eta_k\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^* \\
& + \frac{L_\phi}{2}\eta_k^2 + \frac{8\eta_k^2}{p_k\alpha_k} + \frac{L_2^2\eta_k^4 p_{k+1}}{(1-\eta_0)^2\gamma_k^2\alpha_k} + \frac{\sigma^2\gamma_k^2 p_{k+1}}{(1-\eta_0)^2(1-\eta_k/\gamma_k)^2}. \qquad (65)
\end{aligned}$$

On the other hand, by (10), (12), (64), $\overline{m}^0 = G(x^0, \xi^0)$, and the fact that $\mu_0 \leq 1$, one has

$$\begin{aligned}
&\mathbb{E}_{\xi^0}[P_0] = \phi_{\mu_0}(x^0) + p_0\mathbb{E}_{\xi^0}[(\|\overline{m}^0 - \nabla f(x^0)\|_{x^0}^*)^2] \leq f(x^0) + [f(x^0) + B(x^0)]_+ + p_0\sigma^2, \\
&\mathbb{E}_{\{\xi^k\}_{k=0}^K}[P_K] = \phi_{\mu_K}(x^K) + p_K\mathbb{E}_{\{\xi^k\}_{k=0}^K}[(\|\overline{m}^K - \nabla f(x^K)\|_{x^K}^*)^2] \overset{(12)}{\geq} (1+\mu_K)\phi_{\rm low}.
\end{aligned}$$

By summing (65) over $k = 0, \ldots, K-1$, and using the above two inequalities, (13), and the fact that $\{\eta_k\}_{k\geq 0}$ is nonincreasing, we obtain that

$$\begin{aligned}
\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^*] \leq\ & \frac{\Delta(x^0) + p_0\sigma^2}{\eta_{K-1}} \\
& + \frac{1}{\eta_{K-1}}\sum_{k=0}^{K-1}\left(\frac{L_\phi}{2}\eta_k^2 + \frac{8\eta_k^2}{p_k\alpha_k} + \frac{L_2^2\eta_k^4 p_{k+1}}{(1-\eta_0)^2\gamma_k^2\alpha_k} + \frac{\sigma^2\gamma_k^2 p_{k+1}}{(1-\eta_0)^2(1-\eta_k/\gamma_k)^2}\right).
\end{aligned}$$

Hence, the conclusion of this theorem holds as desired. ■

Proof of Theorem 18 It follows from the definition of $\{(\eta_k, \gamma_k, \alpha_k)\}_{k\geq 0}$ that

$$\alpha_k = \frac{(k+1)^{1/7} - 5s_\eta/7}{(k+1)^{5/7} - 5s_\eta/7} > \frac{1 - 5s_\eta/(7(k+1)^{1/7})}{(k+1)^{4/7}} \geq \frac{1 - 5s_\eta/7}{(k+1)^{4/7}} \qquad \forall k \geq 0,$$

where the first inequality is due to $s_\eta \in (0,1)$. We next prove $(1-\alpha_k)p_{k+1} \le (1-\alpha_k/2)p_k$ for all $k \ge 0$. By the definition of $\{(\eta_k,\gamma_k,\alpha_k)\}_{k\ge 0}$, one has for all $k \ge 0$ that

$$\frac{1-\alpha_k/2}{1-\alpha_k} = 1 + \frac{(1-5s_\eta/(7(k+1)^{1/7}))/2}{(k+1)^{4/7}-1} > 1 + \frac{(1-5s_\eta/7)/2}{(k+1)^{4/7}} > 1 + \frac{1}{7(k+1)^{4/7}},$$

where the inequalities are due to $s_\eta \in (0,1)$ and $k \ge 0$. In addition, recall from (34) that $p_{k+1}/p_k = (1+1/(k+1))^{1/7} \le 1 + 1/(7(k+1))$ for all $k \ge 0$, where the second inequality is due to $(1+a)^r \le 1+ar$ for all $a, r \in [0,1]$. Combining the above two inequalities with the fact that $k+1 \ge (k+1)^{4/7}$ for all $k \ge 0$, we obtain that $(1-\alpha_k)p_{k+1} \le (1-\alpha_k/2)p_k$ holds for all $k \ge 0$. Hence, this lemma holds. ■

Proof of Theorem 19 Recall from (33) that $\eta_k/\gamma_k < s_\eta$ for all $k \ge 0$. Thus, $\{(\eta_k,\gamma_k)\}_{k\ge 0}$ defined in (33) satisfies the assumption on $\{(\eta_k,\gamma_k)\}_{k\ge 0}$ in Theorem 17. Notice from Theorem 18 that $\{p_k\}_{k\ge 0}$ defined in (34) and $\{\alpha_k\}_{k\ge 0}$ defined in Theorem 17 satisfy the assumption on $\{(\alpha_k,p_k)\}_{k\ge 0}$ in Theorem 17. Substituting (33), (34), and $\alpha_k \ge (1-5s_\eta/7)/(k+1)^{4/7}$ (see Theorem 18) into (32), we obtain for all $K \ge 3$,

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^*] \\
&\le \frac{\Delta(x^0)+p_0\sigma^2}{\eta_{K-1}} + \frac{1}{\eta_{K-1}}\sum_{k=0}^{K-1}\left(\frac{L_\phi}{2}\eta_k^2 + \frac{8\eta_k^2}{p_k\alpha_k} + \frac{2L_2^2\eta_k^4 p_k}{(1-\eta_0)^2\gamma_k^2\alpha_k} + \frac{2\sigma^2\gamma_k^2 p_k}{(1-\eta_0)^2(1-\eta_k/\gamma_k)^2}\right) \\
&\le \frac{7K^{5/7}}{5}\Bigg(\frac{\Delta(x^0)+\sigma^2}{s_\eta} + \sum_{k=0}^{K-1}\Big(\frac{25s_\eta L_\phi}{98(k+1)^{10/7}} \\
&\qquad + \frac{200s_\eta/(7(7-5s_\eta)) + 1250L_2^2 s_\eta^3/(7(7-5s_\eta)^3) + 2\sigma^2/(s_\eta(1-5s_\eta/7)^4)}{k+1}\Big)\Bigg) \\
&\le \frac{7K^{5/7}}{5}\left(\frac{\Delta(x^0)+\sigma^2}{s_\eta} + \frac{40s_\eta L_\phi}{49} + \Big(\frac{200s_\eta}{7(7-5s_\eta)} + \frac{1250L_2^2 s_\eta^3}{7(7-5s_\eta)^3} + \frac{2\sigma^2}{s_\eta(1-5s_\eta/7)^4}\Big)\ln(2K+1)\right) \\
&\le \frac{7}{5}\left(\frac{\Delta(x^0)+\sigma^2}{s_\eta} + \frac{40s_\eta L_\phi}{49} + 2\Big(\frac{200s_\eta}{7(7-5s_\eta)} + \frac{1250L_2^2 s_\eta^3}{7(7-5s_\eta)^3} + \frac{2\sigma^2}{s_\eta(1-5s_\eta/7)^4}\Big)\right)K^{5/7}\ln K \\
&\overset{(35)}{=} M_{\rm em}K^{5/7}\ln K/2,
\end{aligned} \tag{66}$$

where the first inequality follows from (32) and the fact that $p_{k+1} \le 2p_k$ for any $k \ge 0$, the second inequality is due to (33), (34), and $\alpha_k \ge (1-5s_\eta/7)/(k+1)^{4/7}$ for all $k \ge 0$, the third inequality follows from $\sum_{k=0}^{K-1} 1/(k+1)^{10/7} \le (7/3)2^{3/7} < 16/5$ and $\sum_{k=0}^{K-1} 1/(k+1) \le \ln(2K+1)$ due to (50) with $(a,b) = (1,K)$ and $\alpha = 10/7, 1$, and the last inequality is due to $1 \le \ln K$ and $\ln(2K+1) \le 2\ln K$ given that $K \ge 3$. Since $\kappa(K)$ is uniformly drawn from $\{\lfloor K/2\rfloor, \ldots, K-1\}$, we have that for all $K \ge 3$,

$$\begin{aligned}
\mathbb{E}[\|\nabla\phi_{\mu_{\kappa(K)}}(x^{\kappa(K)}) + A^T\lambda^{\kappa(K)}\|_{x^{\kappa(K)}}^*] &= \frac{1}{K-\lfloor K/2\rfloor}\sum_{k=\lfloor K/2\rfloor}^{K-1}\mathbb{E}[\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^*] \\
&\le \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^*] \overset{(66)}{\le} M_{\rm em}K^{-2/7}\ln K.
\end{aligned} \tag{67}$$

By Theorem 32 with $(\beta, u, v) = (2/7, \epsilon/(7M_{\rm em}(1+\sqrt{\vartheta})), K)$, one can see that

$$K^{-2/7}\ln K \le \frac{\epsilon}{M_{\rm em}(1+\sqrt{\vartheta})} \qquad \forall K \ge \Big(\frac{7M_{\rm em}(1+\sqrt{\vartheta})}{\epsilon}\ln\Big(\frac{7M_{\rm em}(1+\sqrt{\vartheta})}{\epsilon}\Big)\Big)^{7/2},$$

which together with (67) implies that

$$\begin{aligned}&\mathbb{E}[\|\nabla\phi_{\mu_{\kappa(K)}}(x^{\kappa(K)}) + A^T\lambda^{\kappa(K)}\|^*_{x^{\kappa(K)}}] \le \frac{\epsilon}{1+\sqrt{\vartheta}}\\ &\forall K \ge \max\Big\{\Big(\frac{7M_{\rm em}(1+\sqrt{\vartheta})}{\epsilon}\ln\Big(\frac{7M_{\rm em}(1+\sqrt{\vartheta})}{\epsilon}\Big)\Big)^{7/2}, 3\Big\}.\end{aligned} \tag{68}$$

On the other hand, when $K \ge 2((1+\sqrt{\vartheta})/\epsilon)^{7/2}$, by the definition of $\{\mu_k\}_{k\ge0}$ in (33) and the fact that $\kappa(K)$ is uniformly selected from $\{\lfloor K/2\rfloor, \ldots, K-1\}$, one has that $\mu_{\kappa(K)} = \mu_{\lfloor K/2\rfloor} = \epsilon/(1+\sqrt{\vartheta})$. Combining this with (68), we obtain that (36) holds as desired, and the proof of this theorem is complete. ■

5.6 Proof of the Main Results in Section 3.4

Proof of Theorem 22 Fix any $k \ge 0$. Recall from Theorem 6 that $\|x^{k+1}-x^k\|_{x^k} = \eta_k$. Using this and (38), we have that

$$\begin{aligned}&\mathbb{E}_{\xi^{k+1}}[(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|^*_{x^{k+1}})^2] \le \frac{1}{(1-\eta_k)^2}\mathbb{E}_{\xi^{k+1}}[(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|^*_{x^k})^2]\\
&\overset{(38)}{=} \frac{1}{(1-\eta_k)^2}\mathbb{E}_{\xi^{k+1}}[(\|G(x^{k+1},\xi^{k+1}) + (1-\gamma_k)(\overline{m}^k - G(x^k,\xi^{k+1})) - \nabla f(x^{k+1})\|^*_{x^k})^2]\\
&= (1-\alpha_k)^2(\|\overline{m}^k - \nabla f(x^k)\|^*_{x^k})^2\\
&\quad + \frac{1}{(1-\eta_k)^2}\mathbb{E}_{\xi^{k+1}}[(\|G(x^{k+1},\xi^{k+1}) - \nabla f(x^{k+1}) + (1-\gamma_k)(\nabla f(x^k) - G(x^k,\xi^{k+1}))\|^*_{x^k})^2],\end{aligned} \tag{69}$$

where the first inequality is due to Theorem 27(iii) and $\|x^{k+1}-x^k\|_{x^k} = \eta_k$, and the second equality follows from the first relation in (10) and the definition of α_k. Also, observe that

$$\begin{aligned}&\mathbb{E}_{\xi^{k+1}}[(\|G(x^{k+1},\xi^{k+1}) - \nabla f(x^{k+1}) + (1-\gamma_k)(\nabla f(x^k) - G(x^k,\xi^{k+1}))\|^*_{x^k})^2]\\
&\le 3(\|\nabla f(x^{k+1}) - \nabla f(x^k)\|^*_{x^k})^2 + 3\mathbb{E}_{\xi^{k+1}}[(\|G(x^{k+1},\xi^{k+1}) - G(x^k,\xi^{k+1})\|^*_{x^k})^2]\\
&\quad + 3\gamma_k^2\mathbb{E}_{\xi^{k+1}}[(\|\nabla f(x^k) - G(x^k,\xi^{k+1})\|^*_{x^k})^2]\Big)\\
&\le 3(L_1^2+L^2)\|x^{k+1}-x^k\|^2_{x^k} + 3\gamma_k^2\mathbb{E}_{\xi^{k+1}}[(\|\nabla f(x^k) - G(x^k,\xi^{k+1})\|^*_{x^k})^2]\\
&= 3(L_1^2+L^2)\eta_k^2 + 3\gamma_k^2\mathbb{E}_{\xi^{k+1}}[(\|\nabla f(x^k) - G(x^k,\xi^{k+1})\|^*_{x^k})^2] \le 3(L_1^2+L^2)\eta_k^2 + 3\sigma^2\gamma_k^2,\end{aligned}$$

where the first inequality is due to $\|a+b+c\|^2 \le 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$ for all $a, b, c \in \mathbb{R}^n$, the second inequality is due to (9) and (37), the first equality is due to $\|x^{k+1}-x^k\|_{x^k} = \eta_k$, and the last inequality follows from the second relation in (10). Combining (69) with the above inequality, we obtain that

$$\mathbb{E}_{\xi^{k+1}}[(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|^*_{x^{k+1}})^2] \le (1-\alpha_k)^2(\|\overline{m}^k - \nabla f(x^k)\|^*_{x^k})^2 + \frac{3(L_1^2+L^2)\eta_k^2 + 3\sigma^2\gamma_k^2}{(1-\eta_k)^2}.$$

Notice from $\gamma_k > \eta_k$ and the definition of α_k that $\alpha_k \in (0,1)$. Using this, the above relation, and the fact that $\{\eta_k\}_{k\geq 0}$ is nonincreasing, we obtain that this lemma holds as desired. ■

Proof of Theorem 23 For convenience, we construct potentials as

$$P_k = \phi_{\mu_k}(x^k) + p_k(\|\overline{m}^k - \nabla f(x^k)\|_{x^k}^*)^2 \qquad \forall k \geq 0. \tag{70}$$

Recall from Algorithm 1 that $\{\mu_k\}$ is nonincreasing. Using these, (8), (39), and (48), we obtain that for all $k \geq 0$,

$$\begin{aligned}
&\mathbb{E}_{\xi^{k+1}}[P_{k+1}] \overset{(70)}{=} \mathbb{E}_{\xi^{k+1}}[\phi_{\mu_{k+1}}(x^{k+1}) + p_{k+1}(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|_{x^{k+1}}^*)^2] \\
&= (\mu_{k+1} - \mu_k)(f(x^{k+1}) + B(x^{k+1})) + \mathbb{E}_{\xi^{k+1}}[\phi_{\mu_k}(x^{k+1}) + p_{k+1}(\|\overline{m}^{k+1} - \nabla f(x^{k+1})\|_{x^{k+1}}^*)^2] \\
&\overset{(8)(39)(48)}{\leq} (\mu_{k+1} - \mu_k)\phi_{\rm low} + \phi_{\mu_k}(x^k) - \eta_k\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^* + 4\eta_k\|\nabla f(x^k) - \overline{m}^k\|_{x^k}^* + \frac{L_\phi}{2}\eta_k^2 \\
&\quad + (1-\alpha_k)p_{k+1}(\|\nabla f(x^k) - \overline{m}^k\|_{x^k}^*)^2 + \frac{3(L_1^2+L^2)\eta_k^2 p_{k+1} + 3\sigma^2\gamma_k^2 p_{k+1}}{(1-\eta_0)^2} \\
&\leq (\mu_{k+1} - \mu_k)\phi_{\rm low} + \phi_{\mu_k}(x^k) - \eta_k\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^* + 4\eta_k\|\nabla f(x^k) - \overline{m}^k\|_{x^k}^* + \frac{L_\phi}{2}\eta_k^2 \\
&\quad + (1-\alpha_k/2)p_k(\|\nabla f(x^k) - \overline{m}^k\|_{x^k}^*)^2 + \frac{3(L_1^2+L^2)\eta_k^2 p_{k+1} + 3\sigma^2\gamma_k^2 p_{k+1}}{(1-\eta_0)^2},
\end{aligned}$$

where the second inequality follows from $(1-\alpha_k)p_{k+1} \leq (1-\alpha_k/2)p_k$. In addition, notice that $4\eta_k\|\nabla f(x^k) - \overline{m}^k\|_{x^k}^* \leq \alpha_k p_k(\|\nabla f(x^k) - \overline{m}^k\|_{x^k}^*)^2/2 + 8\eta_k^2/(p_k\alpha_k)$, which together with (70) and the above inequality implies that for all $k \geq 0$,

$$\begin{aligned}
\mathbb{E}_{\xi^{k+1}}[P_{k+1}] &\overset{(70)}{\leq} (\mu_{k+1} - \mu_k)\phi_{\rm low} + P_k - \eta_k\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^* \\
&\quad + \frac{L_\phi}{2}\eta_k^2 + \frac{8\eta_k^2}{p_k\alpha_k} + \frac{3(L_1^2+L^2)\eta_k^2 p_{k+1} + 3\sigma^2\gamma_k^2 p_{k+1}}{(1-\eta_0)^2}.
\end{aligned} \tag{71}$$

On the other hand, by (10), (12), (70), $\overline{m}^0 = G(x^0,\xi^0)$, and $\mu_0 \leq 1$, one has

$$\begin{aligned}
&\mathbb{E}_{\xi^0}[P_0] = \phi_{\mu_0}(x^0) + p_0\mathbb{E}[(\|\overline{m}^0 - \nabla f(x^0)\|_{x^0}^*)^2] \leq f(x^0) + [f(x^0) + B(x^0)]_+ + p_0\sigma^2, \\
&\mathbb{E}_{\{\xi^k\}_{k=0}^K}[P_K] = \phi_{\mu_K}(x^K) + p_K\mathbb{E}[(\|\overline{m}^K - \nabla f(x^K)\|_{x^K}^*)^2] \overset{(12)}{\geq} (1+\mu_K)\phi_{\rm low}.
\end{aligned}$$

By summing (71) over $k = 0,\ldots,K-1$, and using the above two inequalities, (13), and the fact that $\{\eta_k\}_{k\geq 0}$ is nonincreasing, we obtain that

$$\begin{aligned}
\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla\phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^*] &\leq \frac{\Delta(x^0) + p_0\sigma^2}{\eta_{K-1}} \\
&\quad + \frac{1}{\eta_{K-1}}\sum_{k=0}^{K-1}\Big(\frac{L_\phi}{2}\eta_k^2 + \frac{8\eta_k^2}{p_k\alpha_k} + \frac{3(L_1^2+L^2)\eta_k^2 p_{k+1} + 3\sigma^2\gamma_k^2 p_{k+1}}{(1-\eta_0)^2}\Big).
\end{aligned}$$

Hence, the conclusion of this theorem holds. ■

Proof of Theorem 24 It follows from the definition of $\{(\eta_k, \gamma_k, \alpha_k)\}_{k\geq 0}$ that

$$\alpha_k = \frac{1 - s_\eta/3}{(k+1)^{2/3} - s_\eta/3} > \frac{1 - s_\eta/3}{(k+1)^{2/3}} \qquad \forall k \geq 0,$$

where the inequality is due to $s_\eta \in (0,1)$. We next prove $(1-\alpha_k)p_{k+1} \leq (1-\alpha_k/2)p_k$ for all $k \geq 0$. By the definition of $\{(\eta_k, \gamma_k, \alpha_k)\}_{k\geq 0}$, one has for all $k \geq 0$ that

$$\frac{1-\alpha_k/2}{1-\alpha_k} = 1 + \frac{(1-s_\eta/3)/2}{(k+1)^{2/3} - 1} > 1 + \frac{1}{3(k+1)^{2/3}},$$

where the first inequality is due to $s_\eta \in (0,1)$ and $k \geq 0$. Also, we recall from (42) that $p_{k+1}/p_k = (1 + 1/(k+1))^{1/3} \leq 1 + 1/(3(k+1))$ for all $k \geq 0$, where the second inequality is due to $(1+a)^r \leq 1 + ar$ for all $a, r \in [0,1]$. Combining the above two inequalities with the fact that $k + 1 \geq (k+1)^{2/3}$, we obtain that $(1-\alpha_k)p_{k+1} \leq (1-\alpha_k/2)p_k$ for all $k \geq 0$. Hence, this lemma holds as desired. ■

Proof of Theorem 25 Recall from (41) that $\eta_k < \gamma_k$ for all $k \geq 0$. Therefore, $\{(\eta_k, \gamma_k)\}_{k\geq 0}$ defined in (41) satisfies the assumption on $\{(\eta_k, \gamma_k)\}_{k\geq 0}$ in Theorem 23. Notice from Theorem 24 that $\{\alpha_k\}_{k\geq 0}$ defined in Theorem 23 and $\{p_k\}_{k\geq 0}$ defined in (42) satisfy the assumption on $\{(\alpha_k, p_k)\}_{k\geq 0}$ in Theorem 23. By substituting (41), (42), and $\alpha_k \geq (1 - s_\eta/3)/(k+1)^{2/3}$ (see Theorem 24) into (40), one can obtain that for all $K \geq 3$,

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}[(\|\nabla \phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^*)^2] \\
&\leq \frac{\Delta(x^0) + p_0\sigma^2}{\eta_{K-1}} + \frac{1}{\eta_{K-1}} \sum_{k=0}^{K-1} \Big(\frac{L_\phi}{2}\eta_k^2 + \frac{8\eta_k^2}{p_k\alpha_k} + \frac{6(L_1^2 + L^2)\eta_k^2 p_k + 6\sigma^2\gamma_k^2 p_k}{(1-\eta_0)^2} \Big) \\
&\leq 3K^{2/3} \Bigg(\frac{\Delta(x^0) + \sigma^2}{s_\eta} \\
&\qquad + \sum_{k=0}^{K-1} \Big(\frac{s_\eta L_\phi}{18(k+1)^{4/3}} + \frac{8s_\eta/(3(3-s_\eta)) + 6(L_1^2 + L^2)s_\eta/(3-s_\eta)^2 + 6\sigma^2/(s_\eta(1-s_\eta/3)^2)}{k+1} \Big) \Bigg) \\
&\leq 3K^{2/3} \Bigg(\frac{\Delta(x^0) + \sigma^2}{s_\eta} + \frac{2s_\eta L_\phi}{9} + 2\Big(\frac{4s_\eta}{3(3-s_\eta)} + \frac{3(L_1^2 + L^2)s_\eta}{(3-s_\eta)^2} + \frac{3\sigma^2}{s_\eta(1-s_\eta/3)^2} \Big) \ln(2K+1) \Bigg) \\
&\leq 3\Bigg(\frac{\Delta(x^0) + \sigma^2}{s_\eta} + \frac{2s_\eta L_\phi}{9} + 4\Big(\frac{4s_\eta}{3(3-s_\eta)} + \frac{3(L_1^2 + L^2)s_\eta}{(3-s_\eta)^2} + \frac{3\sigma^2}{s_\eta(1-s_\eta/3)^2} \Big) \Bigg) K^{2/3} \ln K \\
&\overset{(43)}{=} M_{\rm rm} K^{2/3} \ln K/2,
\end{aligned} \tag{72}$$

where the first inequality is due to (40) and $p_{k+1} \leq 2p_k$ for all $k \geq 0$, the second inequality follows from (41), (42), and $\alpha_k \geq (1 - s_\eta/3)/(k+1)^{2/3}$ for all $k \geq 0$, the third inequality is because $\sum_{k=0}^{K-1} 1/(k+1)^{4/3} \leq 3(2)^{1/3} < 4$ and $\sum_{k=0}^{K} 1/(k+1) \leq \ln(2K+1)$ due to (50) with $(a,b) = (1,K)$ and $\alpha = 4/3, 1$, and the last inequality follows from $1 \leq \ln K$, and $\ln(2K+1) \leq 2\ln K$ given that $K \geq 3$. Since $\kappa(K)$ is uniformly drawn from $\{\lfloor K/2 \rfloor, \ldots, K-1\}$, we obtain that for all $K \geq 3$,

$$\mathbb{E}[\|\nabla \phi_{\mu_{\kappa(K)}}(x^{\kappa(K)}) + A^T\lambda^{\kappa(K)}\|_{x^{\kappa(K)}}^*] = \frac{1}{K - \lfloor K/2 \rfloor} \sum_{k=\lfloor K/2 \rfloor}^{K-1} \mathbb{E}[\|\nabla \phi_{\mu_k}(x^k) + A^T\lambda^k\|_{x^k}^*]$$

$$\leq \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla\phi_{\mu_k}(x^k)+A^T\lambda^k\|_{x^k}^*] \overset{(72)}{\leq} M_{\rm rm}K^{-1/3}\ln K. \tag{73}$$

By Theorem 32 with $(\beta,u,v)=(1/3,\epsilon/(6M_{\rm rm}(1+\sqrt{\vartheta})),K)$, one can see that

$$K^{-1/3}\ln K \leq \frac{\epsilon}{M_{\rm rm}(1+\sqrt{\vartheta})} \qquad \forall K \geq \Big(\frac{6M_{\rm rm}(1+\sqrt{\vartheta})}{\epsilon}\ln\Big(\frac{6M_{\rm rm}(1+\sqrt{\vartheta})}{\epsilon}\Big)\Big)^3,$$

which together with (73) implies that

$$\begin{aligned}&\mathbb{E}[\|\nabla\phi_{\mu_{\kappa(K)}}(x^{\kappa(K)})+A^T\lambda^{\kappa(K)}\|_{x^{\kappa(K)}}^*] \leq \frac{\epsilon}{1+\sqrt{\vartheta}}\\ &\forall K \geq \max\Big\{\Big(\frac{6M_{\rm rm}(1+\sqrt{\vartheta})}{\epsilon}\ln\Big(\frac{6M_{\rm rm}(1+\sqrt{\vartheta})}{\epsilon}\Big)\Big)^3,3\Big\}.\end{aligned} \tag{74}$$

On the other hand, when $K \geq 2((1+\sqrt{\vartheta})/\epsilon)^3$, by the definition of $\{\mu_k\}_{k\geq 0}$ in (41) and the fact that $\kappa(K)$ is uniformly selected from $\{\lfloor K/2\rfloor,\ldots,K-1\}$, one has that $\mu_{\kappa(K)}=\mu_{\lfloor K/2\rfloor}=\epsilon/(1+\sqrt{\vartheta})$. Combining this with (74), we obtain that (44) holds as desired, and the proof of this theorem is complete. ■

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

A. Alacaoglu and S. J. Wright. Complexity of single loop algorithms for nonlinear programming with stochastic objective and constraints. In *AISTATS*, pages 4627–4635, 2024.

F. Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM J. Optim.*, 5(1):13–51, 1995.

F. Alizadeh and D. Goldfarb. Second-order cone programming. *Math. Program.*, 95(1): 3–51, 2003.

M. ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 10.1.0.*, 2019. URL `http://docs.mosek.com/10.1/toolbox/index.html`.

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Mach. Learn.*, 73:243–272, 2008.

F. R. Bach, G. R. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *ICML*, page 6, 2004.

R. Badenbroek and E. de Klerk. Complexity analysis of a sampling-based interior point method for convex optimization. *Math. Oper. Res.*, 47(1):779–811, 2022.

A. S. Berahas, F. E. Curtis, D. Robinson, and B. Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM J. Optimi.*, 31(2):1352–1379, 2021.

A. S. Berahas, F. E. Curtis, M. J. O'Neill, and D. P. Robinson. A stochastic sequential quadratic optimization algorithm for nonlinear-equality-constrained optimization with rank-deficient Jacobians. *Math. Oper. Res.*, 2023a.

A. S. Berahas, J. Shi, Z. Yi, and B. Zhou. Accelerating stochastic sequential quadratic programming for equality constrained optimization using predictive variance reduction. *Comput. Optim. Appl.*, 86(1):79–116, 2023b.

A. Bidaurrazaga, A. Pérez, and M. Capó. K-means for evolving data streams. In *IEEE ICDM*, pages 1006–1011, 2021.

R. H. Byrd, M. E. Hribar, and J. Nocedal. An interior point algorithm for large-scale nonlinear programming. *SIAM J. Optimi.*, 9(4):877–900, 1999.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. "Convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *ICML*, pages 654–663, 2017.

F. E. Curtis, D. P. Robinson, and B. Zhou. Inexact sequential quadratic optimization for minimizing a stochastic objective function subject to deterministic nonlinear equality constraints. *arXiv:2107.03512*, 2021.

F. E. Curtis, X. Jiang, and Q. Wang. Single-loop deterministic and stochastic interior-point algorithms for nonlinearly constrained optimization. *arXiv:2408.16186*, 2024.

F. E. Curtis, V. Kungurtsev, D. P. Robinson, and Q. Wang. A stochastic-gradient-based interior-point algorithm for solving smooth bound-constrained optimization problems. *SIAM J. Optimi.*, 35(2):1030–1059, 2025.

A. Cutkosky and H. Mehta. Momentum improves normalized SGD. In *ICML*, pages 2260–2268, 2020.

A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex SGD. In *NIPS*, volume 32, 2019.

P. Dvurechensky and M. Staudigl. Hessian barrier algorithms for non-convex conic optimization. *Math. Program.*, pages 1–59, 2024.

C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *NIPS*, volume 31, 2018.

Y. Fang, S. Na, M. W. Mahoney, and M. Kolar. Fully stochastic trust-region sequential quadratic programming for equality-constrained optimization problems. *SIAM J. Optimi.*, 34(2):2007–2037, 2024.

B. Fares, P. Apkarian, and D. Noll. An augmented Lagrangian method for a class of LMI-constrained problems in robust control theory. *Int. J. Control*, 74(4):348–360, 2001.

A. Forsgren, P. E. Gill, and M. H. Wright. Interior methods for nonlinear optimization. *SIAM Review*, 44(4):525–597, 2002.

S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optimi.*, 23(4):2341–2368, 2013.

C. He and Z. Lu. A Newton-CG based barrier method for finding a second-order stationary point of nonconvex conic optimization with complexity guarantees. *SIAM J. Optimi.*, 33 (2):1191–1222, 2023.

C. He, H. Huang, and Z. Lu. A Newton-CG based barrier-augmented Lagrangian method for general nonconvex conic optimization. *Comput. Optim. Appl.*, 89(3):843–894, 2024.

S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale ℓ_1-regularized least squares. *IEEE J. Sel. Top. Signal Process.*, 1(4):606–617, 2007.

G. Lan. *First-Order and Stochastic Optimization Methods for Machine Learning*. Springer, 2020.

Z. Li, H. Bao, X. Zhang, and P. Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *ICML*, pages 6286–6295, 2021.

Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu. Stochastic inexact augmented Lagrangian method for nonconvex expectation constrained optimization. *Comput. Optim. Appl.*, 87 (1):117–147, 2024.

C. Lu, J. Tang, S. Yan, and Z. Lin. Generalized nonconvex nonsmooth low-rank minimization. In *IEEE CVPR*, pages 4130–4137, 2014.

Z. Lu, S. Mei, and Y. Xiao. Variance-reduced first-order methods for deterministically constrained stochastic nonconvex optimization with strong convergence guarantees. *arXiv preprint arXiv:2409.09906*, 2024.

S. Na and M. W. Mahoney. Asymptotic convergence rate and statistical inference for stochastic sequential quadratic programming. *arXiv: 2205.13687*, 2022.

S. Na, M. Anitescu, and M. Kolar. An adaptive stochastic sequential quadratic programming with differentiable exact augmented Lagrangians. *Math. Program.*, 199(1):721–791, 2023.

H. Narayanan. Randomized interior point methods for sampling and optimization. *Ann. Appl. Probab.*, 26(1):597–641, 2016.

A. Nemirovski. Interior point polynomial time methods in convex programming. *Lecture Notes*, 2004.

Y. Nesterov and A. Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, Philadelphia, 1994.

J. Peng and Y. Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM J. Optimi.*, 18(1):186–205, 2007.

F. A. Potra and S. J. Wright. Interior-point methods. *J. Comput. Appl. Math.*, 124(1-2): 281–302, 2000.

Q. Shi, X. Wang, and H. Wang. A momentum-based linearized augmented Lagrangian method for nonconvex constrained stochastic optimization. *Math. Oper. Res.*, 2025.

P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola. Second order cone programming approaches for handling missing and uncertain data. *J. Mach. Learn. Res.*, pages 1283–1314, 2006.

S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*. MIT Press, 2011.

J. F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optim. Methods Softw.*, 11(1-4):625–653, 1999.

D. Sun, K.-C. Toh, and Y. Yuan. Convex clustering: Model, theoretical guarantee and efficient algorithm. *J. Mach. Learn. Res.*, 22(9):1–32, 2021.

K.-C. Toh, M. J. Todd, and R. H. Tütüncü. SDPT3—a MATLAB software package for semidefinite programming, version 1.3. *Optim. Methods Softw.*, 11(1-4):545–581, 1999.

Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen. A hybrid stochastic optimization framework for composite nonconvex optimization. *Math. Program.*, 191(2):1005–1071, 2022.

P. Tseng and Z.-Q. Luo. On the convergence of the affine-scaling algorithm. *Math. Program.*, 56(1):301–319, 1992.

R. J. Vanderbei and D. F. Shanno. An interior-point algorithm for nonconvex nonlinear programming. *Comput. Optim. Appl.*, 13:231–252, 1999.

A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Program.*, 106:25–57, 2006.

Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh. Spiderboost and momentum: Faster variance reduction algorithms. In *NIPS*, volume 32, 2019.

H. Wolkowicz, R. Saigal, and L. Vandenberghe. *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*. Springer Science & Business Media, 2012.

S. J. Wright. *Primal-Dual Interior-Point Methods*. SIAM, 1997.

H. Xu, C. Caramanis, and S. Mannor. Optimization under probabilistic envelope constraints. *Oper. Res.*, 60(3):682–699, 2012.

Y. Xu and Y. Xu. Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization. *J. Optim. Theory Appl.*, 196(1):266–297, 2023.

Y. Zhang and Q. Yang. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.*, 34(12):5586–5609, 2021.

Y. Zhang and D.-Y. Yeung. A convex formulation for learning task relationships in multi-task learning. In *UAI*, 2010.

F. Zohrizadeh, C. Josz, M. Jin, R. Madani, J. Lavaei, and S. Sojoudi. Conic relaxations of power system optimization: Theory and algorithms. *Eur. J. Oper. Res.*, 287(2):391–409, 2020.