An Exact Penalty Method for Stochastic Equality-Constrained Optimization

Yawen Cui* Qiankun Shi† Xiao Wang‡ Xiantao Xiao§

November 6, 2025

Abstract

In this paper, we study a penalty method for stochastic equality-constrained optimization, where both the objective and constraints are expressed in general expectation form. We introduce a novel adaptive strategy for updating the penalty parameter, guided by iteration progress to balance reductions in the penalty function with improvements in constraint violation, while each penalty subproblem is approximately solved using a truncated stochastic prox-linear algorithm. Under certain conditions, the proposed penalty method terminates after finitely many iterations, rendering its exactness: an approximate stationary point of the penalty function corresponds to an approximate KKT point of the original problem. We establish oracle complexity for finding a stochastic ϵ-KKT point, requiring $\mathcal{O}(\epsilon^{-3})$ stochastic gradient evaluations for the objective and constraints, along with $\mathcal{O}(\epsilon^{-5})$ stochastic function evaluations for the constraints. We further develop variants for problems with stochastic objective and deterministic constraints and for problems with finite-sum objective and constraints, each with matching complexity analyses. Finally, we report numerical results demonstrating the proposed method's performance on a test problem.

1 Introduction

In this paper we consider the nonconvex constrained optimization problem

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) := \mathbb{E}_\xi[F(x;\xi)] \\ \text{s.t.} \quad & c(x) := \mathbb{E}_\xi[C(x;\xi)] = \mathbf{0}, \end{aligned} \tag{1.1}$$

where ξ is a random variable in the probability space Ξ and independent of x, and $\mathbb{E}_\xi$ refers to the expectation taken with respect to ξ. Here $F : \mathbb{R}^n \times \Xi \to \mathbb{R}$ and $C : \mathbb{R}^n \times \Xi \to \mathbb{R}^m$ are continuously differentiable with respect to x but possibly nonconvex. Throughout the paper, we assume that the feasible set of (1.1) is nonempty. Problem (1.1) arises in many application areas, such as optimal control [3], fairness-constrained optimization problems [11], PDE-constrained optimization [16, 29] and chance constrained programs [24, 28].

Problems as (1.1) but with deterministic constraints, also known as semi-stochastic problems, have been studied in recent years. In this context, stochastic sequential quadratic programming (SQP) algorithms have attracted increasing attention as an effective approach for such problems. Berahas et al. [2] introduce a stochastic SQP algorithm that uses Lipschitz constants (or their estimates in practice) for stepsize selection, and establish convergence in expectation under standard assumptions. In [9], Curtis et al. investigate an adaptive stochastic SQP algorithm for nonlinear problems with equality and inequality constraints, proving its convergence in expectation. Na et al. [25] propose a stochastic SQP (StoSQP) framework that minimizes

*Dalian University of Technology, Dalian, China. (ywcui@mail.dlut.edu.cn)
†Sun Yat-sen University, Guangzhou, China. (shiqk@mail2.sysu.edu.cn)
‡Corresponding author. Sun Yat-sen University, Guangzhou, China. (wangx936@mail.sysu.edu.cn)
§Dalian University of Technology, Dalian, China. (xtxiao@dlut.edu.cn)

a differentiable exact augmented Lagrangian (AL) function and incorporates either a fixed or stochastic line search for stepsize selection. Almost sure convergence is established for both non-adaptive and adaptive variants. In a subsequent work [26], the authors propose a fully online StoSQP method for statistical inference in nonlinearly constrained stochastic optimization problems, establishing an almost sure convergence rate and iteration complexity. A worst-case complexity bound for such problems is established by Curtis et al. [8]. In addition, several works [14, 15] have investigated trust-region-based SQP methods for solving constrained stochastic optimization problems. Another related class of approaches are proximal point methods by Boob et al. [4,5] for inequality-constrained optimization. Their main idea is to transform the original problem into a sequence of convex subproblems with proximal terms. Boob et al. [4] present an inexact constrained proximal point framework that incorporates a novel constraint extrapolation (ConEx) method for solving each convex subproblem. For semi-stochastic problems, they establish a complexity of $\mathcal{O}(\epsilon^{-4})$ under the strong feasibility condition. In [5], a level constrained stochastic proximal gradient method is proposed, with an increasing constraint level scheme that guarantees the subproblem feasibility. Under the Mangasarian-Fromovitz constraint qualification (MFCQ), the algorithm achieves an $\mathcal{O}(\epsilon^{-4})$ oracle complexity for computing an ϵ-KKT point in semi-stochastic case.

Meanwhile, penalty methods provide an alternative strategy for semi-stochastic problems and have been extensively studied in recent literature. Wang et al. [34] present a penalty framework in which an ℓ_2 penalty function is approximately minimized at each iteration using only stochastic first-order or zeroth-order information. They provide worst-case guarantees on the oracle complexity required to reach an ϵ-stochastic critical point. A stochastic primal-dual (SPD) method is proposed by Jin and Wang [18] to address nonconvex optimization problems with a large number of inequality constraints. SPD updates variables by minimizing a linearized AL function constructed from stochastic gradients of the objective and partial information from a randomly chosen subset of constraints, reducing the per-iteration cost. Subsequently, they introduce a stochastic nested primal-dual (STEP) method [19] targeting constrained optimization problems with objective functions formed by two expectation terms. Under a nonsingularity condition, they investigate the iteration and oracle complexities of STEP to reach an ϵ-stationary point. Based on the linearized AL function, Shi et al. [32] adopt a recursive momentum technique to design a stochastic algorithm that achieves oracle complexities of $\mathcal{O}(\epsilon^{-4})$ for reaching both ϵ-stationary and ϵ-KKT points. The complexities can be further improved to $\mathcal{O}(\epsilon^{-3})$ if the initial point is approximately feasible. Subsequently, Lu et al. [23] propose a stochastic first-order method with truncated recursive momentum, achieving a similar complexity of $\tilde{\mathcal{O}}(\epsilon^{-3})$ when using increasing penalty parameters, while the feasibility is guaranteed in the deterministic sense rather than in expectation. In [1], Alacaoglu and Wright employ variance reduction techniques and obtain a complexity bound of $\tilde{\mathcal{O}}(\epsilon^{-4})$ in semi-stochastic case. Notably, in both [23] and [1], the logarithmic factors in the complexity bounds can be removed by an appropriate parameter choice depending on the maximum iteration number, which requires a near-feasible initial point. The aforementioned methods typically require overly large penalty parameters which may bring practical issues. In [33], Wang analyzes inexact cubic-regularized primal-dual methods for finding second-order stationary points and establishes the corresponding complexity bounds. Recent work by Zuo et al. [39] proposes a single-loop adaptive stochastic linearized AL method, which updates penalty parameters dynamically according to the behavior of the iterates. By leveraging a recursive momentum strategy and clipped stochastic gradients to reduce variance, a high-probability oracle complexity analysis for attaining an ϵ-KKT point is established.

In recent years, there are several works for problem with both objective and constraints being stochastic, i.e. (1.1), also known as fully-stochastic problems. The approach in [4] for fully-stochastic inequality-constrained problems also employs an inexact constrained proximal point method with ConEx (ICPPC). Under the strong feasibility assumption, its oracle complexity of order $\mathcal{O}(\epsilon^{-6})$ for obtaining an approximate KKT point is established. Additionally, a stochastic SQP method [31] is proposed, enriching the existing framework for handling such problems. Under a strong LICQ condition, the method achieves an oracle complexity of order $\mathcal{O}(\epsilon^{-8})$ for stationarity and $\mathcal{O}(\epsilon^{-4})$ for feasibility. Li et al. [20] build upon the standard inexact AL method framework and propose stochastic inexact AL methods (Stoc-iALM) by integrating momentum-based variance-reduced proximal stochastic gradient techniques within their subroutines. The sample complexity in [20] is established as order $\mathcal{O}(\epsilon^{-5})$ for finding an ϵ-KKT point under a certain non-

singularity (NSC) condition. Another recent work [1] develops single-loop momentum-based algorithms for such problems. In particular, this approach attains a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-5})$ to find a point satisfying ϵ-KKT conditions, with the logarithmic factors removable through appropriate parameter tuning based on the final iterate and a near-feasible point. In the work by Cui et al. [7], a single-loop two-phase stochastic momentum-based method (TStoM) is proposed. The first phase minimizes an infeasibility measure to obtain a near-feasible point in expectation, which initializes the primal-dual method in the second phase. The oracle complexity of TStoM to achieve an ϵ-stationary point is of order $\mathcal{O}(\epsilon^{-6})$, and improves to $\mathcal{O}(\epsilon^{-5})$ to reach an ϵ-KKT point under the NSC condition. Despite this, the above augmented Lagrangian type methods generally rely on a predetermined penalty parameter that must be sufficiently large.

Research has also focused on stochastic optimization problems like (1.1) with specific nonconvex structures, such as problems with weakly convex functions and problems in finite-sum forms. A recent work by Yang et al. [37] proposes a single-loop stochastic algorithm based on a hinge-based penalty method for nonconvex, nonsmooth constrained optimization with weakly convex objective and constraint functions, achieving an $\mathcal{O}(\epsilon^{-6})$ oracle complexity for finding a nearly ϵ-KKT point under a regularity condition. In [22], Liu and Xu present an exact penalty model for nonconvex nonsmooth stochastic optimization with inequality constraints in expectation, which is solved by a single-loop SPIDER-type stochastic subgradient method. Assuming a (uniform) Slater-type constraint qualification condition and weak convexity of the constraint function, the method achieves a sample complexity of order $\mathcal{O}(\epsilon^{-4})$ for evaluations of both the objective and constraint function subgradients, and $\mathcal{O}(\epsilon^{-6})$ for evaluations of the constraint function value to produce an (ϵ, ϵ)-KKT point. When the constraints enjoy a finite-sum structure, the corresponding complexities become $\mathcal{O}(\epsilon^{-4})$ and $\mathcal{O}(N + \sqrt{N}\epsilon^{-4})$, respectively. Another two related works investigate convex optimization problems where both the objective and the constraints take a finite-sum structure. Lin et al. [21] propose an affine-minorized feasible level-set method that ensures a feasible solution path and attains an absolutely ϵ-optimal solution. Yan and Xu [36] propose an adaptive primal-dual stochastic gradient method based on the Lagrangian function and establish its convergence rate. However, these studies are limited to the convex setting, and analysis for finite-sum problems in nonconvex setting is still scarce.

1.1 Contributions

In this paper, we study an exact penalty method for nonconvex stochastic equality-constrained optimization. This method adopts a double-loop algorithm framework in which the penalty parameter is adaptively updated in outer iterations based on a novel strategy that balances the decrease of the penalty function with improvements in feasibility. In the inner iterations, the penalty function with fixed penalty parameter is approximately minimized by using a truncated stochastic prox-linear algorithm. Under certain constraint qualification conditions we prove that the method terminates in a finite number of outer iterations, resulting in the exactness in the sense that the approximate solution of a penalty subproblem is an approximate KKT point of the original problem. Within this algorithm framework we also propose variant methods for stochastic equality-constrained optimization problems in different settings and analyze the corresponding oracle complexity.

- **Fully-stochastic case.** Both objective and constraint functions are stochastic, as in (1.1). In this case, we establish oracle complexity of order $\mathcal{O}(\epsilon^{-3})$ for the stochastic gradient evaluations of the objective and constraint functions, and of order $\mathcal{O}(\epsilon^{-5})$ for the evaluations of the stochastic constraint function values. To the best of our knowledge, these bounds achieve state-of-the-art performance, with detailed comparisons provided in Table 1.

- **Semi-stochastic case.** Only objective is stochastic. Exact constraint information can be used in the proposed method. The corresponding oracle complexity regarding the stochastic objective gradient computations is in order $\mathcal{O}(\epsilon^{-3})$. As shown in Table 2, our method matches the best-known complexity results to date. Moreover, since we adopt an adaptive way to update the penalty parameter, our method does not require a nearly feasible initial point as [32].

- **Finite-sum case.** Both objective and constraint functions are in finite-sum forms. In this case, we

Algorithm	Problem	f, c	Loop	Assumptions	OGC	CGC	CFC
Algorithm 1 [31]	$\min_{x\in\mathbb{R}^n} f(x)$ s.t. $c(x)=0$	sm, ncx	single	strong LICQ	$\mathcal{O}(\epsilon^{-8})$		
Algorithm 2 [1]	$\min_{x\in X} f(x)$ s.t. $c(x)=0$	sm, ncx	single	non-singularity	$\tilde{\mathcal{O}}(\epsilon^{-5})$		
TStoM [7]	$\min_{x\in X} f(x)+h(x)$ s.t. $c(x)=0$	sm, ncx	single	non-singularity	$\mathcal{O}(\epsilon^{-5})$		
ICPPC [4]	$\min_{x\in X} f(x)+h(x)$ s.t. $c_{\mathcal{I}}(x)+h_{\mathcal{I}}(x)\le 0$	sm, ncx	double	strong feasibility	$\mathcal{O}(\epsilon^{-6})$		
Stoc-iALM [20]	$\min_{x\in\mathbb{R}^n} f(x)+h(x)$ s.t. $c(x)=0$	sm, ncx	double	non-singularity	$\mathcal{O}(\epsilon^{-5})$		
Algorithm 1 [37]	$\min_{x\in\mathbb{R}^n} f(x)$ s.t. $c(x)\le 0$	wc	single	regularity condition	$\mathcal{O}(\epsilon^{-6})$		
3S-Econ [22]	$\min_{x\in X} f(x)$ s.t. $c(x)\le 0$	wc	single	(uniform) Slater-type CQ	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-6})$
Ours	$\min_{x\in\mathbb{R}^n} f(x)$ s.t. $c(x)=0$	sm, ncx	double	strong LICQ	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-5})$

Table 1: Comparison of algorithms for fully-stochastic optimization problems, where $f(x)=\mathbb{E}_\xi[F(x;\xi)]$, $c(x)=\mathbb{E}_\xi[C(x;\xi)]$ and $h(x), h_i, i\in\mathcal{I}$ are convex but possibly nonsmooth, and $X\subseteq\mathbb{R}^n$ is a closed convex set. The objective gradient complexity (OGC) and constraint gradient complexity (CGC) are characterized regarding the total number of evaluations to the stochastic (sub)gradients of the objective and constraint functions, respectively, and the constraint function value complexity (CFC) represents the number of evaluations of the stochastic constraint function. For f and c, sm means smooth, ncx means nonconvex, and wc means weakly convex. The Assumptions column lists the main constraint qualification assumptions. The strong LICQ corresponds to Assumption 4.1 in this paper. The non-singularity condition refers to Assumption 3 in [20]. The (uniform) Slater-type CQ refers to Assumption 2 in [22]. The regularity condition refers to (4) in [37, Theorem 4.2].

adopt a different variance reduction technique to compute stochastic oracles, and the corresponding oracle complexity regarding component function information evaluations is of order $\mathcal{O}(N+N^{4/5}\epsilon^{-2})$.

1.2 Notation and preliminaries

Without any specification, $\|\cdot\|$ denotes the Euclidean norm. We use $\mathbb{E}[\cdot|x_k]$ to denote the expectation conditioned on x_k. The notation $\mathbb{E}[\cdot]$ refers to the full expectation over all random variables generated during an algorithmic process.

In general nonconvex constrained optimization, it is generally challenging to locate a global or even a local minimizer. The main research stream thus focuses on more tractable solutions, the KKT points. Under certain constraint qualification condition, a local minimizer of (1.1) is also a KKT point [27]. In this paper, we will study algorithms for (1.1) in pursuit of an ϵ-KKT point or a stochastic ϵ-KKT point, which are defined as follows.

DEFINITION **1.1.** *Given $\epsilon>0$, we call $x\in\mathbb{R}^n$ an ϵ-KKT point of (1.1), if there exists $\lambda\in\mathbb{R}^m$ such that*

$$\|\nabla f(x)+\nabla c(x)\lambda\|\le\epsilon \ \text{ and } \ \|c(x)\|\le\epsilon,$$

where $\nabla c(x)=(\nabla c_1(x),\dots,\nabla c_m(x))$. We call $x\in\mathbb{R}^n$ a stochastic ϵ-KKT point of (1.1), if

$$\mathbb{E}[\|\nabla f(x)+\nabla c(x)\lambda\|^2]\le\epsilon^2 \ \text{ and } \ \mathbb{E}[\|c(x)\|]\le\epsilon.$$

Algorithm	Problem	Stationary measure	Loop	Assumptions	Complexity
Algorithm 2 [1]	$\min\limits_{x\in X} f(x)$ s.t. $c_i(x)=0,\ i\in\mathcal{E}$	$\mathbb{E}[\mathrm{d}(\nabla f(x)+\nabla c(x)\lambda, -\mathcal{N}_X(x))]\le\epsilon,$ $\mathbb{E}[\|c(x)\|]\le\epsilon$	single	non-singularity	$\tilde{\mathcal{O}}(\epsilon^{-4})$
LCSPG [5]	$\min\limits_{x\in\mathbb{R}^n} f(x)+h(x)$ s.t. $c_i(x)+h_i(x)\le 0,$ $i\in\mathcal{I}$	$\mathbb{E}[\mathrm{d}^2(\nabla f(x)+\partial h(x)+\sum\limits_{i\in\mathcal{I}}\lambda_i(\nabla c_i(x)+\partial h_i(x)),0)]\le\epsilon^2,$ $\mathbb{E}[\sum\limits_{i\in\mathcal{I}}\lambda_i\lvert c_i(x)+h_i(x)\rvert]\le\epsilon^2$, where x is feasible	double	uniform MFCQ	$\mathcal{O}(\epsilon^{-4})$
Algorithm 1 [23]	$\min\limits_{x\in X} f(x)$ s.t. $c_i(x)=0,\ i\in\mathcal{E}$	$\mathbb{E}[\mathrm{d}(\nabla f(x)+\nabla c(x)\lambda, -\mathcal{N}_X(x))]\le\epsilon,$ $\|c(x)\|\le\epsilon$	single	non-singularity condition	$\tilde{\mathcal{O}}(\epsilon^{-3})$
CoSTA [17]	$\min\limits_{x\in\mathbb{R}^n} f(x)+h(x)$ s.t. $c_i(x)\le 0,\ i\in\mathcal{I}$ $g_j(x)\le 0,\ j\in\mathcal{J}$	$\mathbb{E}[\|\nabla f(x)+\nabla g(x)\lambda+\nabla c(x)\nu+\partial h(x)\|]\le\sqrt{\epsilon},$ $\mathbb{E}[\lambda^\top g(x)]\ge-\epsilon,\ \mathbb{E}[\nu^\top c(x)]\ge-\epsilon$	single	parameterized MFCQ	$\tilde{\mathcal{O}}(\epsilon^{-3})$
MLALM [32]	$\min\limits_{x\in X} f(x)+h(x)$ s.t. $c_i(x)\le 0,\ i\in\mathcal{I}$ $c_i(x)=0,\ i\in\mathcal{E}$	$\mathbb{E}[\mathrm{d}^2(\nabla f(x)+\partial h(x)+\sum\limits_{i\in\mathcal{E}\cup\mathcal{I}}\lambda_i\nabla c_i(x), -\mathcal{N}_X(x))]\le\epsilon^2,$ $\mathbb{E}[\|c_{\mathcal{E}}(x)\|^2+\|[c_{\mathcal{I}}(x)]_+\|^2]\le\epsilon^2,\ \mathbb{E}[\sum\limits_{i\in\mathcal{I}}\lambda_i\lvert c_i(x)\rvert]\le\epsilon$	single	extended variant of MFCQ, initial near-feasibility	$\mathcal{O}(\epsilon^{-3})$
Ours	$\min\limits_{x\in\mathbb{R}^n} f(x)$ s.t. $c_i(x)=0,\ i\in\mathcal{E}$	$\mathbb{E}[\|\nabla f(x)+\nabla c(x)\lambda\|^2]\le\epsilon^2,$ $\mathbb{E}[\|c(x)\|]\le\epsilon$	double	strong LICQ	$\mathcal{O}(\epsilon^{-3})$

Table 2: Comparison of algorithms for problems in semi-stochastic setting, where $f(x)=\mathbb{E}_\xi[F(x;\xi)]$ is nonconvex and smooth. In addition, $X\subseteq\mathbb{R}^n$ is a closed convex set, $c_i, i\in\mathcal{E}\cup\mathcal{I}$ are smooth but possibly nonconvex; h and $h_i, i\in\mathcal{I}$ are convex but possibly nonsmooth, and $g_j, j\in\mathcal{J}$ are smooth and convex. The uniform MFCQ refers to Assumption 3 in [5]. The extended variant of MFCQ refers to Assumption 5 in [32]. The non-singularity condition refers to Assumption 1 (iv) in [23]. The parameterized MFCQ refers to Assumption A7 in [17] and the strong LICQ corresponds to Assumption 5.1 in this paper.

The assumptions imposed throughout the rest of this paper are presented here.

Assumption 1.1. *Let $X\subset\mathbb{R}^n$ be an open convex set containing all the iterates generated by the algorithm. The objective function value of problem (1.1) is lower bounded on X by a finite constant C^*. Moreover, there exist positive constants G, M such that $\|\nabla f(x)\|\le G$, $\|c(x)\|\le M$ for all $x\in X$.*

Remark 1.1. *This boundedness assumption is essential for maintaining the stability of the iteration sequence within the framework of stochastic constrained optimization. Similar assumptions have also been adopted in prior studies [2, 9, 31, 33] to ensure desirable theoretical properties.*

Assumption 1.2. *Function $F(\cdot;\xi)$ is differentiable for almost every $\xi\in\Xi$. There exists $\sigma_f>0$ such that for any $x\in\mathbb{R}^n$,*

$$\mathbb{E}_\xi[\nabla F(x;\xi)]=\nabla f(x),\quad \mathbb{E}_\xi[\|\nabla F(x;\xi)-\nabla f(x)\|^2]\le\sigma_f^2.$$

There exists $L_f>0$ such that for any $x,y\in\mathbb{R}^n$,

$$\mathbb{E}_\xi[\|\nabla F(x;\xi)-\nabla F(y;\xi)\|^2]\le L_f^2\|x-y\|^2.$$

Assumption 1.3. *Function $C(\cdot;\xi)$ is differentiable for almost every $\xi\in\Xi$. There exist $\sigma_J,\sigma_c>0$ such that for any $x\in\mathbb{R}^n$,*

$$\begin{aligned}&\mathbb{E}_\xi[\nabla C(x;\xi)]=\nabla c(x),\quad &&\mathbb{E}_\xi[\|\nabla C(x;\xi)-\nabla c(x)\|^2]\le\sigma_J^2,\\ &\mathbb{E}_\xi[C(x;\xi)]=c(x),\quad &&\mathbb{E}_\xi[\|C(x;\xi)-c(x)\|^2]\le\sigma_c^2.\end{aligned}$$

There exist $L_J, L_c>0$ such that for any $x,y\in\mathbb{R}^n$,

$$\mathbb{E}_\xi[\|\nabla C(x;\xi)-\nabla C(y;\xi)\|^2]\le L_J^2\|x-y\|^2,\quad \mathbb{E}_\xi[\|C(x;\xi)-C(y;\xi)\|^2]\le L_c^2\|x-y\|^2.$$

By Jensen's inequality, Assumptions 1.2 and 1.3 imply that

$$\|\nabla f(x)-\nabla f(y)\|\le L_f\|x-y\|,\ \|\nabla c(x)-\nabla c(y)\|\le L_J\|x-y\|,\ \|c(x)-c(y)\|\le L_c\|x-y\|,\tag{1.2}$$

and hence $\|\nabla c(x)\|\le L_c$.

1.3 Outline

The remainder of this paper is outlined as follows. In Section 2, we propose an adaptive penalty method based on stochastic approximations to address problem (1.1). The resulting framework requires solving a sequence of nonconvex composite penalty subproblems, for which we develop a tailored algorithm in Section 3. Section 4 is devoted to the complexity analysis of the overall penalty method. Section 5 presents the adaptations of the method to semi-stochastic, finite-sum, and deterministic settings, along with their oracle complexity analysis. In Section 6, preliminary numerical test results are reported. Finally, we draw conclusions in Section 7.

2 An adaptive penalty method for stochastic equality-constrained optimization

Penalty methods are among the most well-known approaches for solving constrained optimization problems. By incorporating penalties for constraint violations into the objective function, these methods transform the original constrained problem into a sequence of unconstrained subproblems. However, existing penalty methods for (1.1) typically rely on a preset, fixed penalty parameter throughout the entire optimization process [1, 7, 20, 22], without leveraging the progress made during algorithm iterations. In this paper, we will propose an adaptive penalty method based on stochastic approximations for solving (1.1). Our method adopts a double-loop algorithm framework. In outer iterations we dynamically updates the penalty parameter through an adaptive rule, while in inner iterations we solve the penalty subproblem with fixed ρ:

$$\min_{x\in\mathbb{R}^n} \quad \Phi_\rho(x) := f(x) + \rho\|c(x)\| \tag{2.1}$$

by calling a subproblem solver. This section will focus on the adaptive update of the penalty parameter and provide a comprehensive overview of the proposed penalty method. Discussions of the subproblem solver will be deferred to the next section.

In the closely related work [34], an adaptive penalty method for stochastic optimization with deterministic constraints, i.e. (1.1) in semi-stochastic setting, is studied. At current iterate x_k, [34] determines a penalty parameter $\rho := \rho_k \geq \rho_{k-1} + \tau$ such that

$$l(x_k) - \min_{\|s\|\leq 1} l(x_k + s) \geq \rho\zeta\theta(x_k), \tag{2.2}$$

where $\zeta \in (0, 1)$ and

$$\theta(x_k) := \|c(x_k)\| - \min_{\|s\|\leq 1} \|c(x_k) + \nabla c(x_k)^\top s\|,$$
$$l(x_k + s) := f(x_k) + \hat{\nabla} f_k^\top s + \rho\|c(x_k) + J(x_k)s\|.$$

Here, $\hat{\nabla} f_k$ represents stochastic estimates of $\nabla f(x_k)$. The term $\theta(x_k)$ quantifies the potential improvement in constraint violation around x_k, while $l(x_k+s)$ serves as a stochastic approximation to the penalty function at $x_k + s$, i.e., $f(x_k + s) + \rho\|c(x_k + s)\|$. However, the criterion in (2.2) requires solving two ball-constrained global optimization subproblems, which significantly increases the computational burden in practice. We also observe that in theory it actually unnecessary to globally solve these subproblems to satisfy (2.2). Instead, it suffices to ensure that the approximate reduction in the penalty function is adequate relative to the improvement in constraint violation near the current iterate. This observation motivates us to propose new strategies for updating the penalty parameter. Furthermore, we aim to design a strategy such that, under certain conditions, the penalty parameter is updated only a finite number of times. This would allow the penalty function to eventually become *exact*, effectively transforming the original equality-constrained optimization problem (1.1) into an equivalent unconstrained penalty problem.

As mentioned previously, a key to propose an effective penalty parameter update strategy is to guarantee the potential reduction of the penalty function and the improvement of constraint violation. This

is easy to realize, if both the objective function and constraint violation can be reduced, at least along a certain direction. However, in stochastic settings such a direction can only approximately identified by using stochastic estimates $\hat{\nabla} f_k$, $\hat{\nabla} c_k$, and $\hat{c}_k$ which are approximate to the true values $\nabla f(x_k)$, $\nabla c(x_k)$, and $c(x_k)$, respectively. With those stochastic estimates, we define the direction as

$$d_k := -u_k + \alpha v_k, \tag{2.3}$$

where $\alpha \in (0,1)$, v_k solves

$$\min_{v \in \mathbb{R}^n} \ \frac{1}{2}\|v\|^2 \quad \text{s.t. } v \in \operatorname*{argmin}_{w \in \mathbb{R}^n} \frac{1}{2}\|\hat{c}_k + \hat{\nabla} c_k^\top w\|^2$$

and u_k solves

$$\min_{u \in \mathrm{Null}(\hat{\nabla} c_k^\top)} \frac{1}{2}\|u - \hat{\nabla} f_k\|^2.$$

For the component v_k, its primary role is to induce a potential reduction of the constraint violation. It is observed that the computation of v_k can be formulated as an unconstrained least-squares problem. To ensure uniqueness, we specifically select the minimum-norm solution, which can be expressed as

$$v_k = -(\hat{\nabla} c_k^\top)^\dagger \hat{c}_k$$

by Moore-Penrose pseudoinverse. For the component u_k, it is constructed to decrease the objective function while remaining in the tangent space of the constraint manifold at x_k. Note that u_k and v_k are orthogonal. The computation of u_k is analytically tractable and given by

$$u_k = \hat{\nabla} f_k - (\hat{\nabla} c_k^\top)^\dagger \hat{\nabla} c_k^\top \hat{\nabla} f_k.$$

As such, $(-u_k)$ approximates the direction of steepest descent for the objective function that remains tangent to the constraint surface.

We now define

$$\begin{aligned} \theta_k &= \|\hat{c}_k\| - \|\hat{c}_k + \gamma \hat{\nabla} c_k^\top d_k\|, \\ \phi_k &= \rho_{k-1}\theta_k - \gamma \hat{\nabla} f_k^\top d_k - \frac{\gamma}{2}\|d_k\|^2 \quad \text{with } \gamma > 0. \end{aligned} \tag{2.4}$$

Given steering parameter $\zeta \in (0,1)$, we set the condition

$$\phi_k \geq \rho_{k-1}\zeta\theta_k \tag{2.5}$$

as a stopping criterion for our algorithm. It is worthy to note that when $\hat{c}_k = 0$, we have $v_k = 0$ and $\theta_k = 0$ from $d_k = -u_k \in \mathrm{Null}(\hat{\nabla} c_k^\top)$. It then implies that

$$\begin{aligned} \phi_k &= \gamma\big(\hat{\nabla} f_k^\top u_k - \frac{1}{2}\|u_k\|^2\big) \geq \gamma(\hat{\nabla} f_k - u_k)^\top u_k \\ &= \gamma((\hat{\nabla} c_k^\top)^\dagger \hat{\nabla} c_k^\top \hat{\nabla} f_k)^\top (\hat{\nabla} f_k - (\hat{\nabla} c_k^\top)^\dagger \hat{\nabla} c_k^\top \hat{\nabla} f_k) \\ &= \gamma(P_k \hat{\nabla} f_k)^\top (\hat{\nabla} f_k - P_k \hat{\nabla} f_k) \\ &= \gamma \hat{\nabla} f_k^\top P_k^\top (I - P_k)\hat{\nabla} f_k = 0 \\ &= \rho_{k-1}\zeta\theta_k, \end{aligned}$$

where $P_k := (\hat{\nabla} c_k^\top)^\dagger \hat{\nabla} c_k^\top$ satisfies $P_k^\top = P_k$ and $P_k(I - P_k) = 0$. Hence, when $\hat{c}_k = 0$, (2.5) holds naturally. Conversely, if (2.5) fails to hold, we must have $\hat{c}_k \neq 0$. Therefore, if (2.5) is not satisfied, we update the penalty parameter by computing

$$\rho_k = \max\{\beta\rho_{k-1}, \hat{\rho}_{k-1}\} \quad \text{with} \quad \hat{\rho}_{k-1} := \frac{\hat{\nabla} f_k^\top d_k + \frac{1}{2}\|d_k\|^2}{\alpha(1-\zeta)\|\hat{c}_k\|} \ \text{ and } \ \beta > 1. \tag{2.6}$$

Algorithm 2.1

Input: Given steering parameter $\zeta \in (0,1)$, initial iterate x_1, stochastic estimates $\hat{\nabla} f_1$, $\hat{\nabla} c_1$ and $\hat{c}_1$, initial penalty parameter $\rho_0 \geq 1$, parameters $\alpha \in (0,1)$, $\beta > 1$ and $\gamma > 0$, positive integers T and τ.
Output: x_R.

1: **for** $k = 1, 2, \ldots$ **do**
2: **Step (a):** If $k > 1$ and (2.5) is satisfied, terminate the algorithm and return x_R with $R = k$; otherwise, compute ρ_k through (2.6) with d_k defined by (2.3).
3: **Step (b):** Solve subproblem (2.1) with $\rho = \rho_k$ by TSPA (to be introduced in Section 3) to generate x_{k+1} together with $\hat{\nabla} f_{k+1}, \hat{\nabla} c_{k+1}$ and $\hat{c}_{k+1}$, i.e.,

$$(x_{k+1}, \hat{\nabla} f_{k+1}, \hat{\nabla} c_{k+1}, \hat{c}_{k+1}) = \text{TSPA}(x_k, \rho_k, \gamma, T, \tau).$$

4: **end for**

Then we solve the subproblem (2.1) with $\rho = \rho_k$ by applying a subproblem solver to generate the next iterate x_{k+1} in the inner loop. Building upon this, we present the whole algorithm framework of the adaptive penalty method for (1.1) in Algorithm 2.1.

If we consider a deterministic variant of Algorithm 2.1, where all function information used in this algorithm is replaced by exact values and each subproblem is solved exactly (i.e. $x_{k+1} \in \arg\min \Phi_{\rho_k}(x)$), we can prove that, under certain conditions, the penalty function $\Phi_{\rho_k}(x)$ becomes exact when ρ_k is no less than a certain threshold, as can be seen in Appendix A. This means that, in this case, if x_{k+1} is a stationary point of (2.1) with $\rho = \rho_k$, it is also a KKT point of the original problem. This observation motivates the development of our adaptive penalty method for fully-stochastic problem (1.1). In subsequent analysis, we will show that the penalty function remains "exact" in stochastic settings, in the sense that an approximate stationary point of the penalty function is also an approximate KKT point of the original problem. This result also enables us to derive the oracle complexity for associated algorithms accordingly.

3 A truncated stochastic prox-linear algorithm for penalty subproblems

At each iteration of Algorithm 2.1 and after the penalty parameter ρ has been adaptively determined, our remaining task is to solve the penalty subproblem in the form of (2.1). Note that each penalty subproblem is a stochastic composite optimization problem. For this type of problems, stochastic prox-linear approaches have been proposed and extensively studied in the literature (e.g., [12,35,38]). When targeting at (2.1), these approaches iteratively minimize a local approximation of the objective function, by constructing a quadratic approximation to $f(x)$ and a linear approximation to $c(x)$ based on stochastic oracles. More specifically, at current iterate x^i, with stochastic oracles $\hat{\nabla} f^i, \hat{\nabla} c^i$ and $\hat{c}^i$, stochastic prox-linear methods solve the following proximal subproblem to generate the next iterate x^{i+1}:

$$x^{i+1} = \operatorname*{argmin}_x \left\{ (\hat{\nabla} f^i)^\top (x - x^i) + \rho \left\| \hat{c}_i + (\hat{\nabla} c^i)^\top (x - x^i) \right\| + \frac{1}{2\gamma} \|x - x^i\|^2 \right\}. \tag{3.1}$$

To compute stochastic oracles, we employ a variance reduction technique to mitigate possibly large stochastic variances. Motivated by the boundedness assumptions on $\|\nabla c(x)\|$, $\|\nabla f(x)\|$ and $\|c(x)\|$, we apply a truncation technique to ensure the boundedness of the stochastic estimators. This is pivotal in facilitating the convergence and complexity analysis developed in this paper. Specifically, given $\eta > 0$, let $\mathbb{B}(\eta)$ denote the Euclidean ball centered at the origin with radius η, i.e., $\mathbb{B}(\eta) = \{x \in \mathbb{R}^n : \|x\| \leq \eta\}$, and let Π be the projection operator. The stochastic estimators $\hat{\nabla} f^i$, $\hat{c}^i$ and $\hat{\nabla} c^i$ are constructed as follows. Let τ

be a positive integer. If $i \bmod \tau = 0$, we compute

$$\hat{\nabla} f^i = \Pi_{\mathbb{B}(G)}\Big(\frac{1}{|\mathcal{A}_i|}\sum_{\xi\in\mathcal{A}_i}\nabla F(x^i;\xi)\Big),\ \hat{c}^i = \Pi_{\mathbb{B}(M)}\Big(\frac{1}{|\mathcal{B}_i|}\sum_{\xi\in\mathcal{B}_i} C(x^i;\xi)\Big),\ \hat{\nabla} c^i = \Pi_{\mathbb{B}(L_c)}\Big(\frac{1}{|\mathcal{S}_i|}\sum_{\xi\in\mathcal{S}_i}\nabla C(x^i;\xi)\Big), \tag{3.2}$$

while if $i \bmod \tau \neq 0$, we compute

$$\begin{aligned}
\hat{\nabla} f^i &= \Pi_{\mathbb{B}(G)}\Big(\hat{\nabla} f^{i-1} + \frac{1}{|\mathcal{A}_i|}\sum_{\xi\in\mathcal{A}_i}\big(\nabla F(x^i;\xi) - \nabla F(x^{i-1};\xi)\big)\Big),\\
\hat{c}^i &= \Pi_{\mathbb{B}(M)}\Big(\hat{c}^{i-1} + \frac{1}{|\mathcal{B}_i|}\sum_{\xi\in\mathcal{B}_i}\big(C(x^i;\xi) - C(x^{i-1};\xi)\big)\Big),\\
\hat{\nabla} c^i &= \Pi_{\mathbb{B}(L_c)}\Big(\hat{\nabla} c^{i-1} + \frac{1}{|\mathcal{S}_i|}\sum_{\xi\in\mathcal{S}_i}\big(\nabla C(x^i;\xi) - \nabla C(x^{i-1};\xi)\big)\Big).
\end{aligned} \tag{3.3}$$

Here, $\mathcal{A}_i$, $\mathcal{B}_i$ and $\mathcal{S}_i$, $i \geq 0$ are sets of samples randomly and independently generated from Ξ.

Building upon discussions above, we now present the truncated stochastic prox-linear algorithmic framework in Algorithm 3.1.

Algorithm 3.1 TSPA$(\bar{x}, \rho, \gamma, T, \tau)$: **T**runcated **S**tochastic **P**rox-linear **A**lgorithm for (2.1)

1: **Set** $x^0 = \bar{x}$.
2: **for** $i = 0, \ldots, T\tau - 1$ **do**
3: **if** $i \bmod \tau = 0$ **then**
4: Compute $\hat{\nabla} f^i$, $\hat{c}^i$ and $\hat{\nabla} c^i$ through (3.2).
5: **else**
6: Compute $\hat{\nabla} f^i$, $\hat{c}^i$ and $\hat{\nabla} c^i$ through (3.3).
7: **end if**
8: Compute x^{i+1} through (3.1).
9: **end for**

Output: Return x^r, $\hat{\nabla} f^r, \hat{c}^r$ and $\hat{\nabla} c^r$ with r uniformly at random picked from $i = 0, \ldots, T\tau - 1$.

Let $\{x^i\}_{i=0,\ldots,T\tau-1}$ be generated by Algorithm 3.1. The following lemma provides upper bounds on variances of stochastic oracles $\hat{\nabla} f^i, \hat{c}^i, \hat{\nabla} c^i$.

LEMMA **3.1.** *Suppose Assumptions 1.2- 1.3 hold, then for $i = 0, \ldots, T\tau - 1$,*

$$\mathbb{E}[\|\hat{\nabla} f^i - \nabla f(x^i)\|^2] \leq \mathbb{E}[\|\hat{\nabla} f^{\lfloor \frac{i}{\tau}\rfloor\tau} - \nabla f(x^{\lfloor \frac{i}{\tau}\rfloor\tau})\|^2] + \sum_{p=\lfloor \frac{i}{\tau}\rfloor\tau+1}^{i}\frac{L_f^2}{|\mathcal{A}_p|}\mathbb{E}[\|x^p - x^{p-1}\|^2], \tag{3.4}$$

$$\mathbb{E}[\|\hat{c}^i - c(x^i)\|^2] \leq \mathbb{E}[\|\hat{c}^{\lfloor \frac{i}{\tau}\rfloor\tau} - c(x^{\lfloor \frac{i}{\tau}\rfloor\tau})\|^2] + \sum_{p=\lfloor \frac{i}{\tau}\rfloor\tau+1}^{i}\frac{L_c^2}{|\mathcal{B}_p|}\mathbb{E}[\|x^p - x^{p-1}\|^2], \tag{3.5}$$

$$\mathbb{E}[\|\hat{\nabla} c^i - \nabla c(x^i)\|^2] \leq \mathbb{E}[\|\hat{\nabla} c^{\lfloor \frac{i}{\tau}\rfloor\tau} - \nabla c(x^{\lfloor \frac{i}{\tau}\rfloor\tau})\|^2] + \sum_{p=\lfloor \frac{i}{\tau}\rfloor\tau+1}^{i}\frac{L_J^2}{|\mathcal{S}_p|}\mathbb{E}[\|x^p - x^{p-1}\|^2], \tag{3.6}$$

where $\sum_{i=1}^{0}(\cdot) := 0$.

Proof. Due to the structure of Algorithm 3.1, we only present the analysis for the first τ iterations, i.e., $i = 0, \ldots, \tau - 1$, while for the remaining iterations it can be analyzed similarly. When $i = 0$, the conclusion

holds naturally. For any $i = 1, \ldots, \tau-1$, according to the definition of $\hat{\nabla} f^i$ in (3.3) and the nonexpansiveness of the projection operator $\Pi_{\mathbb{B}(G)}$, it holds that

$$\begin{aligned}\|\hat{\nabla} f^i - \nabla f(x^i)\|^2 &= \big\|\Pi_{\mathbb{B}(G)}\big(\hat{\nabla} f^{i-1} + \frac{1}{|\mathcal{A}_i|}\sum_{\xi\in\mathcal{A}_i}(\nabla F(x^i;\xi) - \nabla F(x^{i-1};\xi))\big) - \Pi_{\mathbb{B}(G)}(\nabla f(x^i))\big\|^2 \\ &\le \big\|\hat{\nabla} f^{i-1} + \frac{1}{|\mathcal{A}_i|}\sum_{\xi\in\mathcal{A}_i}\big(\nabla F(x^i;\xi) - \nabla F(x^{i-1};\xi)\big) - \nabla f(x^i)\big\|^2. \end{aligned} \tag{3.7}$$

Taking expectation on the right-hand side of the above inequality conditioned on x^i yields

$$\begin{aligned}&\mathbb{E}[\|\hat{\nabla} f^{i-1} + \frac{1}{|\mathcal{A}_i|}\sum_{\xi\in\mathcal{A}_i}(\nabla F(x^i;\xi) - \nabla F(x^{i-1};\xi)) - \nabla f(x^i)\|^2|x^i] \\ &= \mathbb{E}[\|\hat{\nabla} f^{i-1} - \nabla f(x^{i-1}) + \frac{1}{|\mathcal{A}_i|}\sum_{\xi\in\mathcal{A}_i}(\nabla F(x^i;\xi) - \nabla F(x^{i-1};\xi)) - (\nabla f(x^i) - \nabla f(x^{i-1}))\|^2|x^i] \\ &= \mathbb{E}[\|\hat{\nabla} f^{i-1} - \nabla f(x^{i-1})\|^2|x^i] + \mathbb{E}[\|\frac{1}{|\mathcal{A}_i|}\sum_{\xi\in\mathcal{A}_i}(\nabla F(x^i;\xi) - \nabla F(x^{i-1};\xi)) - (\nabla f(x^i) - \nabla f(x^{i-1}))\|^2|x^i] \\ &\le \mathbb{E}[\|\hat{\nabla} f^{i-1} - \nabla f(x^{i-1})\|^2|x^i] + \frac{1}{|\mathcal{A}_i|^2}\sum_{\xi\in\mathcal{A}_i}\mathbb{E}[\|\nabla F(x^i;\xi) - \nabla F(x^{i-1};\xi)\|^2|x^i] \\ &\le \mathbb{E}[\|\hat{\nabla} f^{i-1} - \nabla f(x^{i-1})\|^2|x^i] + \frac{L_f^2}{|\mathcal{A}_i|}\|x^i - x^{i-1}\|^2,\end{aligned}$$

where the second equality uses the relations

$$\begin{aligned}\mathbb{E}[\langle\frac{1}{|\mathcal{A}_i|}\sum_{\xi\in\mathcal{A}_i}\nabla F(x^i;\xi) - \nabla f(x^i), \hat{\nabla} f^{i-1} - \nabla f(x^{i-1})\rangle|x^i] = 0, \\ \mathbb{E}[\langle\frac{1}{|\mathcal{A}_i|}\sum_{\xi\in\mathcal{A}_i}\nabla F(x^{i-1};\xi) - \nabla f(x^{i-1}), \nabla f(x^{i-1}) - \hat{\nabla} f^{i-1}\rangle|x^i] = 0,\end{aligned}$$

the first inequality comes from

$$2\mathbb{E}[\langle\frac{1}{|\mathcal{A}_i|}\sum_{\xi\in\mathcal{A}_i}\big(\nabla F(x^i;\xi) - \nabla F(x^{i-1};\xi)\big), \nabla f(x^i) - \nabla f(x^{i-1})\rangle|x^i] = 2\|\nabla f(x^i) - \nabla f(x^{i-1})\|^2,$$

and the last inequality is due to Assumption 1.2. Then by taking full expectation on (3.7) we obtain

$$\begin{aligned}\mathbb{E}[\|\hat{\nabla} f^i - \nabla f(x^i)\|^2] &\le \mathbb{E}[\|\hat{\nabla} f^{i-1} - \nabla f(x^{i-1})\|^2] + \frac{L_f^2}{|\mathcal{A}_i|}\mathbb{E}[\|x^i - x^{i-1}\|^2] \\ &\le \mathbb{E}[\|\hat{\nabla} f^0 - \nabla f(x^0)\|^2] + \sum_{p=1}^{i}\frac{L_f^2}{|\mathcal{A}_p|}\mathbb{E}[\|x^p - x^{p-1}\|^2].\end{aligned}$$

Same arguments apply to the remaining iterations, thus we obtain (3.4). Similarly, the upper bounds for stochastic variance of $\hat{c}^i$ and $\hat{\nabla} c^i$ can be established as in (3.5) and (3.6), respectively. □

We now introduce the generalized gradient mapping $\mathcal{G} : \mathbb{R}^n \to \mathbb{R}^n$, by defining

$$\mathcal{G}(x^i) := \frac{1}{\gamma}(x^i - x^{i+1}) \quad \text{and} \quad \bar{\mathcal{G}}(x^i) := \frac{1}{\gamma}(x^i - \bar{x}^i), \quad i = 0, \ldots, T\tau - 1, \tag{3.8}$$

where

$$\bar{x}^i = \underset{x}{\operatorname{argmin}} \left\{ \nabla f(x^i)^\top (x - x^i) + \rho \left\| c(x^i) + \nabla c(x^i)^\top (x - x^i) \right\| + \frac{1}{2\gamma} \|x - x^i\|^2 \right\}. \tag{3.9}$$

Given $\epsilon > 0$, we call x^i an *ϵ-stationary point* of (2.1) if $\|\bar{\mathcal{G}}(x^i)\| \leq \epsilon$.

As previously noted, the true value $\bar{\mathcal{G}}(x^i)$ is not available during the iteration process. Hence, to characterize the behavior of the iterates, it is necessary to establish a relationship between $\bar{\mathcal{G}}(x^i)$ and its stochastic approximation $\mathcal{G}(x^i)$. The proof is inspired by [38], so we state the results here and defer the proof to Appendix B.

LEMMA **3.2.** *Suppose that Assumptions 1.1-1.3 hold. Then for each iterate generated by Algorithm 3.1, it holds that*

$$\begin{aligned} &\left(\gamma - \gamma^2(\rho L_J + L_f)\right) \left\|\bar{\mathcal{G}}(x^i)\right\|^2 \\ &\leq \left(2\gamma + \gamma^2(\rho L_J + L_f)\right) \left\|\mathcal{G}(x^i)\right\|^2 + 2\rho\big(2\|\hat{c}^i - c(x^i)\| + \frac{1}{L_J}\left\|\hat{\nabla} c^i - \nabla c(x^i)\right\|^2\big) + \frac{2}{L_f}\left\|\hat{\nabla} f^i - \nabla f(x^i)\right\|^2 \end{aligned} \tag{3.10}$$

and

$$\begin{aligned} &\left(\gamma - \gamma^2(\rho L_J + L_f)\right) \left\|\mathcal{G}(x^i)\right\|^2 \\ &\leq \left(2\gamma + \gamma^2(\rho L_J + L_f)\right) \left\|\bar{\mathcal{G}}(x^i)\right\|^2 + 2\rho\big(2\|\hat{c}^i - c(x^i)\| + \frac{1}{L_J}\left\|\hat{\nabla} c^i - \nabla c(x^i)\right\|^2\big) + \frac{2}{L_f}\left\|\hat{\nabla} f^i - \nabla f(x^i)\right\|^2. \end{aligned} \tag{3.11}$$

The following lemma demonstrates a descent property of Φ_ρ, which plays a central role in subsequent analysis. The proof of this lemma is deferred to Appendix B,

LEMMA **3.3.** *Assume that Assumption 1.1 and the conditions in (1.2) hold, then for $i = 0, \ldots, T\tau - 1$,*

$$\begin{aligned} \Phi_\rho(x^{i+1}) \leq \Phi_\rho(x^i) - \left(\frac{\gamma}{2} - \gamma^2(\rho L_J + L_f)\right) \|\mathcal{G}(x^i)\|^2 + 2\rho \left\|\hat{c}^i - c(x^i)\right\| + \frac{1}{2L_f}\|\hat{\nabla} f^i - \nabla f(x^i)\|^2 \\ + \frac{\rho}{2L_J}\left\|\hat{\nabla} c^i - \nabla c(x^i)\right\|^2. \end{aligned} \tag{3.12}$$

We provide upper bounds on the expected generalized gradients at the output x^r of Algorithm 3.1. Detailed proof is presented in Appendix B.

LEMMA **3.4.** *Assume that Assumptions 1.1-1.3 hold and $1 - 4\gamma(\rho L_J + L_f) > 0$. And for any given $\tilde{\epsilon} > 0$, suppose that $\delta = \frac{\tilde{\epsilon}^2}{4\rho}$ and*

$$\begin{aligned} &|\mathcal{A}_i| = \left\lceil \frac{5\sigma_f^2}{2L_f\tilde{\epsilon}^2} \right\rceil, \quad |\mathcal{B}_i| = \left\lceil \frac{36\rho^2\sigma_c^2}{\tilde{\epsilon}^4} \right\rceil, \quad |\mathcal{S}_i| = \left\lceil \frac{5\rho\sigma_J^2}{2L_J\tilde{\epsilon}^2} \right\rceil, \quad \textit{for } i \bmod \tau = 0, \\ &|\mathcal{A}_i| = \lceil 60\tau\gamma L_f \rceil, \quad |\mathcal{B}_i| = \left\lceil \frac{72\gamma\rho\tau L_c^2}{\delta} \right\rceil, \quad |\mathcal{S}_i| = \lceil 60\tau\gamma\rho L_J \rceil, \quad \textit{for } i \bmod \tau \neq 0. \end{aligned}$$

Then it holds that

$$\mathbb{E}\left[\|\bar{\mathcal{G}}(x^r)\|^2\right] \leq \frac{24}{\gamma}\left(\frac{\Phi_\rho(x^0) - C^*}{T\tau} + 4\tilde{\epsilon}^2\right) \quad \textit{and} \quad \mathbb{E}[\|\mathcal{G}(x^r)\|^2] \leq \frac{1}{\Gamma\gamma}\left(\frac{\Phi_\rho(x^0) - C^*}{T\tau} + 4\tilde{\epsilon}^2\right), \tag{3.13}$$

where $\Gamma = \frac{1}{4} - \gamma(\rho L_J + L_f)$.

With specified parameters we can derive the oracle complexity of Algorithm 3.1.

THEOREM **3.1.** *Under Assumptions 1.1-1.3 and given $\epsilon \in (0,1)$, $\rho > 0$, suppose that*

$$\gamma := \frac{1}{8(\rho L_J + L_f)}, \quad T = \tau = \lceil C_1^{1/2} \rho^{1/2} \gamma^{-1/2} \epsilon^{-1} \rceil,$$

where $C_1 := \max\{24, \Gamma^{-1}\}(\frac{\Phi_\rho(x_0) - C^*}{\rho} + 4)$ and that the batch sizes are set following Lemma 3.4 with $\tilde{\epsilon} = C_1^{-1/2} \gamma^{1/2} \epsilon$. Then Algorithm 3.1 returns a point x^r satisfying

$$\mathbb{E}[\|\bar{\mathcal{G}}(x^r)\|^2] \le \epsilon^2 \quad \textit{and} \quad \mathbb{E}[\|\mathcal{G}(x^r)\|^2] \le \epsilon^2, \tag{3.14}$$

and the oracle complexity in terms of evaluations of stochastic objective gradient, stochastic constraint gradient, and stochastic constraint function value are of order $\mathcal{O}(\rho^2 \epsilon^{-3})$, $\mathcal{O}(\rho^3 \epsilon^{-3})$, and $\mathcal{O}(\rho^5 \epsilon^{-5})$, respectively.

Proof. It follows from the parameter setting in Lemma 3.4 that

$$\begin{aligned} &|\mathcal{A}_i| \equiv A = \mathcal{O}(\tilde{\epsilon}^{-2}), \ \ |\mathcal{B}_i| \equiv B = \mathcal{O}(\rho^2 \tilde{\epsilon}^{-4}), \ \ |\mathcal{S}_i| \equiv S = \mathcal{O}(\rho \tilde{\epsilon}^{-2}), \ \text{for } i \bmod \tau = 0, \\ &|\mathcal{A}_i| \equiv a = \mathcal{O}(\tau\gamma), \ \ |\mathcal{B}_i| \equiv b = \mathcal{O}(\tau\gamma\rho^2 \tilde{\epsilon}^{-2}), \ \ |\mathcal{S}_i| \equiv s = \mathcal{O}(\tau\gamma\rho), \ \text{for } i \bmod \tau \neq 0, \end{aligned}$$

which guarantees $\mathbb{E}[\|\bar{\mathcal{G}}(x^r)\|^2] \le \epsilon^2$ and $\mathbb{E}[\|\mathcal{G}(x^r)\|^2] \le \epsilon^2$ by $\tilde{\epsilon} = C_1^{-1/2} \gamma^{1/2} \epsilon$. Meanwhile, the total number of stochastic gradient evaluations of f and c are

$$\begin{aligned} TA + T\tau a &= \mathcal{O}(\rho\epsilon^{-1}) \cdot \mathcal{O}(\rho\epsilon^{-2}) + \mathcal{O}(\rho^2\epsilon^{-2}) \cdot \mathcal{O}(\epsilon^{-1}) = \mathcal{O}(\rho^2\epsilon^{-3}), \\ TS + T\tau s &= \mathcal{O}(\rho\epsilon^{-1}) \cdot \mathcal{O}(\rho^2\epsilon^{-2}) + \mathcal{O}(\rho^2\epsilon^{-2}) \cdot \mathcal{O}(\rho\epsilon^{-1}) = \mathcal{O}(\rho^3\epsilon^{-3}), \end{aligned}$$

respectively, and the total number of stochastic function evaluations of constraints is

$$TB + T\tau b = \mathcal{O}(\rho\epsilon^{-1}) \cdot \mathcal{O}(\rho^4\epsilon^{-4}) + \mathcal{O}(\rho^2\epsilon^{-2}) \cdot \mathcal{O}(\rho^3\epsilon^{-3}) = \mathcal{O}(\rho^5\epsilon^{-5}).$$

This completes the proof. □

We can also bound the estimator errors $\mathbb{E}[\|\hat{\nabla} f^r - \nabla f(x^r)\|^2]$, $\mathbb{E}[\|\hat{\nabla} c^r - \nabla c(x^r)\|^2]$ and $\mathbb{E}[\|\hat{c}^r - c(x^r)\|^2]$ at the output x^r. This property is essential to establish the oracle complexity of the whole penalty method, as presented in Algorithm 2.1.

Corollary 3.2. *For any given $\epsilon > 0$ and under the same conditions as Theorem 3.1, it holds that*

$$\mathbb{E}[\|\hat{\nabla} f^r - \nabla f(x^r)\|^2] \le Q_1 \epsilon^2, \quad \mathbb{E}[\|\hat{c}^r - c(x^r)\|^2] \le Q_2 \epsilon^4, \quad \mathbb{E}[\|\hat{\nabla} c^r - \nabla c(x^r)\|^2] \le Q_3 \epsilon^2, \tag{3.15}$$

where $Q_1 = \frac{(24+C_1)\gamma L_f}{60 C_1}$, $Q_2 = \frac{(8+C_1)\gamma^2}{288 C_1^2 \rho^2}$ and $Q_3 = \frac{(24+C_1)\gamma L_J}{60 C_1 \rho}$.

Proof. One can deduce from (3.4) that for any $i = 0, \ldots, T\tau - 1$,

$$\begin{aligned} \mathbb{E}[\|\hat{\nabla} f^i - \nabla f(x^i)\|^2] &\le \mathbb{E}[\|\hat{\nabla} f^{\lfloor \frac{i}{\tau} \rfloor \tau} - \nabla f(x^{\lfloor \frac{i}{\tau} \rfloor \tau})\|^2] + \sum_{p=\lfloor \frac{i}{\tau} \rfloor \tau + 1}^{i} \frac{L_f^2}{|\mathcal{A}_p|} \mathbb{E}[\|x^p - x^{p-1}\|^2] \\ &\le \frac{\sigma_f^2}{A} + \frac{L_f^2}{a} \sum_{p=\lfloor \frac{i}{\tau} \rfloor \tau + 1}^{i} \mathbb{E}[\|x^p - x^{p-1}\|^2], \end{aligned}$$

where $A = |\mathcal{A}_i|$ for $i \bmod \tau = 0$ and $a = |\mathcal{A}_i|$ for $i \bmod \tau \neq 0$. Summing the above inequality over $i = 0, \ldots, T\tau - 1$ and dividing it by $T\tau$, we obtain

$$\frac{1}{T\tau} \sum_{i=0}^{T\tau-1} \mathbb{E}[\|\hat{\nabla} f^i - \nabla f(x^i)\|^2] \le \frac{\sigma_f^2}{A} + \frac{\tau L_f^2}{a T \tau} \sum_{i=0}^{T\tau-1} \mathbb{E}[\|x^{i+1} - x^i\|^2].$$

Based on (3.14) and batch size settings in Lemma 3.4, we can further deduce that

$$\mathbb{E}[\|\hat{\nabla} f^r - \nabla f(x^r)\|^2] \leq \frac{\sigma_f^2}{A} + \frac{\tau\gamma^2 L_f^2}{a}\mathbb{E}[\|\mathcal{G}(x^r)\|^2] \leq \frac{2\gamma L_f}{5C_1}\epsilon^2 + \frac{\gamma L_f}{60}\epsilon^2.$$

Similarly, we obtain from (3.5) and (3.6) that

$$\mathbb{E}[\|\hat{c}^r - c(x^r)\|^2] \leq \frac{\sigma_c^2}{B} + \frac{\tau\gamma^2 L_c^2}{b}\mathbb{E}[\|\mathcal{G}(x^r)\|^2] \leq \frac{\gamma^2}{36C_1^2\rho^2}\epsilon^4 + \frac{\gamma^2}{288C_1\rho^2}\epsilon^4,$$
$$\mathbb{E}[\|\hat{\nabla} c^r - \nabla c(x^r)\|^2] \leq \frac{\sigma_J^2}{S} + \frac{\tau\gamma^2 L_J^2}{s}\mathbb{E}[\|\mathcal{G}(x^r)\|^2] \leq \frac{2\gamma L_J}{5C_1\rho}\epsilon^2 + \frac{\gamma L_J}{60\rho}\epsilon^2.$$

Hence, from the setting of Q_1, Q_2 and Q_3 we derive the conclusion. □

Meanwhile, the output of Algorithm 3.1 enjoys the following property, which forms the basis to analyze the optimality of the output of the penalty method in next section.

Corollary 3.3. *For any given $\epsilon > 0$ and under the same conditions as Theorem 3.1, there exists $\lambda^r \in \mathbb{R}^m$ such that*

$$\mathbb{E}[\|\nabla f(x^r) + \nabla c(x^r)\lambda^r\|^2] \leq \epsilon^2. \tag{3.16}$$

Proof. Following the definition of $\bar{x}^r$ in (3.9), and according to the first-order optimality conditions for (3.9), there exists $v^r \in \partial\|c(x^r) + \nabla c(x^r)^\top(\bar{x}^r - x^r)\|$ such that

$$\nabla f(x^r) + \rho\nabla c(x^r)v^r + \frac{1}{\gamma}(\bar{x}^r - x^r) = 0,$$

which yields that $\nabla f(x^r) + \rho\nabla c(x^r)v^r = \bar{\mathcal{G}}(x^r)$. Then by (3.14) we obtain

$$\mathbb{E}[\|\nabla f(x^r) + \rho\nabla c(x^r)v^r\|^2] = \mathbb{E}[\|\bar{\mathcal{G}}(x^r)\|^2] \leq \epsilon^2.$$

Hence, with $\lambda^r = \rho v^r$ the conclusion can be derived. □

4 Complexity analysis for the penalty method

In this section, we will analyze the oracle complexity of the penalty method, Algorithm 2.1, for finding a stochastic ϵ-KKT point of (1.1). Let $\{x_k\}$ be generated by Algorithm 2.1. We first analyze the behavior of the penalty parameter ρ during the iteration process. To facilitate subsequent analysis, in addition to Assumptions 1.1-1.3 we impose another assumption, which is also used in [31].

Assumption 4.1. *For any $k \geq 1$, the minimum singular values of $\hat{\nabla} c_k$ are uniformly bounded below by $\bar{\nu} > 0$.*

Note that this assumption is only required to hold at outer iterates. Under Assumption 4.1, it follows from (2.3) that

$$d_k = -u_k + \alpha v_k = -\hat{\nabla} f_k + \hat{\nabla} c_k D_k \hat{\nabla} c_k^\top \hat{\nabla} f_k - \alpha\hat{\nabla} c_k D_k \hat{c}_k, \tag{4.1}$$

where $D_k = (\hat{\nabla} c_k^\top \hat{\nabla} c_k)^{-1}$ and $\|D_k\| \leq 1/\bar{\nu}^2$. Recall the definitions of θ_k and ϕ_k in (2.4). It follows from (4.1) and Assumption 4.1 that

$$\theta_k = \gamma\alpha\|\hat{c}_k\| \quad \text{and} \quad \phi_k = \rho_{k-1}\theta_k - \gamma\hat{\nabla} f_k^\top d_k - \frac{\gamma}{2}\|d_k\|^2.$$

Then when $\hat{c}_k \neq 0$, (2.5) holds if

$$\rho_{k-1} \geq \frac{\hat{\nabla} f_k^\top d_k + \frac{1}{2}\|d_k\|^2}{\alpha(1-\zeta)\|\hat{c}_k\|}. \tag{4.2}$$

By the construction of the stochastic estimators in (3.2)-(3.3), we have $\|\hat{\nabla} f_k\| \leq G$, $\|\hat{\nabla} c_k\| \leq L_c$ and $\|\hat{c}_k\| \leq M$, which further implies that

$$\begin{aligned}\hat{\nabla} f_k^\top d_k + \frac{1}{2}\|d_k\|^2 &\leq (\hat{\nabla} f_k + d_k)^\top d_k = (\hat{\nabla} c_k D_k \hat{\nabla} c_k^\top \hat{\nabla} f_k - \alpha \hat{\nabla} c_k D_k \hat{c}_k)^\top d_k \\ &= (D_k \hat{\nabla} c_k^\top \hat{\nabla} f_k - \alpha D_k \hat{c}_k)^\top \hat{\nabla} c_k^\top d_k \\ &= \alpha^2 (\hat{\nabla} c_k D_k \hat{c}_k)^\top \hat{\nabla} c_k D_k \hat{c}_k - \alpha \hat{\nabla} f_k^\top \hat{\nabla} c_k D_k \hat{c}_k \\ &\leq \alpha^2 |\hat{c}_k^\top D_k \hat{c}_k| + \alpha |\hat{\nabla} f_k^\top \hat{\nabla} c_k D_k \hat{c}_k| \\ &\leq \alpha \left(\frac{1}{\bar{\nu}^2} \alpha M + \frac{1}{\bar{\nu}^2} G L_c \right) \|\hat{c}_k\| := \alpha M_1 \|\hat{c}_k\|,\end{aligned}$$

where $M_1 = \frac{1}{\bar{\nu}^2}\alpha M + \frac{1}{\bar{\nu}^2} G L_c$. Thus, it derives that

$$\frac{\hat{\nabla} f_k^\top d_k + \frac{1}{2}\|d_k\|^2}{\alpha(1-\zeta)\|\hat{c}_k\|} \leq \frac{M_1}{1-\zeta} =: \rho_{\max}.$$

Following the update scheme of the penalty parameter given in (2.6), within at most K_1 times of update, where $K_1 = \lceil \log_\beta(\frac{M_1}{\rho_0(1-\zeta)}) \rceil$, the penalty parameter will satisfy (4.2). Hence, (2.5) can be satisfied within K_1 outer iterations during process of Algorithm 2.1, and then the algorithm terminates.

Let x_R denote the output of Algorithm 2.1 upon termination. Then, (2.5) holds with $k = R$. Note that x_R is the output of the inner iteration process in Algorithm 3.1 when solving (2.1) with $\rho = \rho_{R-1}$. With a slight abuse of notation, we denote this inner iteration output by x^r, i.e., $x^r = x_R$. We assume that parameters used in Algorithm 3.1 are set following Theorem 3.1. Then by Corollary 3.3, (3.16) naturally holds at x_R. On the other hand, from the iterate update scheme of Algorithm 3.1, x^{r+1} is computed by

$$x^{r+1} := \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ (\hat{\nabla} f^r)^\top (x - x^r) + \rho_{R-1} \|\hat{c}^r + (\hat{\nabla} c^r)^\top (x - x^r)\| + \frac{1}{2\gamma}\|x - x^r\|^2 \right\}.$$

The optimality of x^{r+1} implies that

$$\begin{aligned}&(\hat{\nabla} f^r)^\top (x^{r+1} - x^r) + \rho_{R-1} \|\hat{c}^r + (\hat{\nabla} c^r)^\top (x^{r+1} - x^r)\| + \frac{1}{2\gamma}\|x^{r+1} - x^r\|^2 \\ &\leq (\hat{\nabla} f^r)^\top (\gamma d_R) + \rho_{R-1} \|\hat{c}^r + (\hat{\nabla} c^r)^\top (\gamma d_R)\| + \frac{1}{2\gamma}\|\gamma d_R\|^2.\end{aligned}$$

It further derives from (2.5) with $k = R$, $\hat{\nabla} f^r = \hat{\nabla} f_R, \hat{c}^r = \hat{c}_R$ and $\hat{\nabla} c^r = \hat{\nabla} c_R$ that

$$\begin{aligned}&\rho_{R-1}\|\hat{c}^r\| - (\hat{\nabla} f^r)^\top (x^{r+1} - x^r) - \rho_{R-1} \|\hat{c}^r + (\hat{\nabla} c^r)^\top (x^{r+1} - x^r)\| - \frac{1}{2\gamma}\|x^{r+1} - x^r\|^2 \\ &\geq \rho_{R-1}\|\hat{c}^r\| - (\hat{\nabla} f^r)^\top (\gamma d_R) - \rho_{R-1} \|\hat{c}^r + (\hat{\nabla} c^r)^\top (\gamma d_R)\| - \frac{1}{2\gamma}\|\gamma d_R\|^2 = \phi_{R-1} \\ &\geq \rho_{R-1}\zeta\theta_R = \alpha\gamma\rho_{R-1}\zeta\|\hat{c}^r\|.\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}&\alpha\gamma\rho_{R-1}\zeta\|\hat{c}^r\| \\ &\leq \rho_{R-1}\|\hat{c}^r\| - (\hat{\nabla} f^r)^\top (x^{r+1} - x^r) - \rho_{R-1} \|\hat{c}^r + (\hat{\nabla} c^r)^\top (x^{r+1} - x^r)\| - \frac{1}{2\gamma}\|x^{r+1} - x^r\|^2 \\ &\leq \rho_{R-1}\|\hat{c}^r\| - (\hat{\nabla} f^r)^\top (x^{r+1} - x^r) - \rho_{R-1} \|\hat{c}^r\| + \rho_{R-1} \|(\hat{\nabla} c^r)^\top (x^{r+1} - x^r)\| - \frac{1}{2\gamma}\|x^{r+1} - x^r\|^2 \\ &\leq \|\hat{\nabla} f^r\| \|x^{r+1} - x^r\| + \rho_{R-1} \|\hat{\nabla} c^r\| \|x^{r+1} - x^r\|.\end{aligned}$$

For given $\epsilon > 0$, following the same parameter settings as Theorem 3.1 and taking the expectations on both sides of above inequality conditioned on ρ_{R-1}, we obtain that

$$\begin{aligned}\alpha\gamma\rho_{R-1}\zeta\mathbb{E}[\|\hat{c}^r\||\rho_{R-1}] &\leq (\mathbb{E}[\|\hat{\nabla}f^r\|^2|\rho_{R-1}])^{1/2}\cdot(\mathbb{E}[\|x^{r+1}-x^r\|^2|\rho_{R-1}])^{1/2}\\ &\quad+\rho_{R-1}(\mathbb{E}[\|\hat{\nabla}c^r\|^2|\rho_{R-1}])^{1/2}\cdot(\mathbb{E}[\|x^{r+1}-x^r\|^2|\rho_{R-1}])^{1/2}\\ &\leq G\gamma\epsilon+\rho_{R-1}L_c\gamma\epsilon,\end{aligned}$$

where the second inequality comes from Corollary 3.2 and (3.14). It implies that

$$\mathbb{E}[\|\hat{c}^r\||\rho_{R-1}] \leq \left(\frac{G}{\alpha\zeta\rho_0}+\frac{L_c}{\alpha\zeta}\right)\epsilon := M_2\epsilon.$$

Then it can be shown from (3.15) that

$$\mathbb{E}[\|c(x^r)\|] \leq \mathbb{E}[\|\hat{c}^r\|]+\mathbb{E}[\|c(x^r)-\hat{c}^r\|] \leq (M_2+Q_2^{1/2}\epsilon)\epsilon. \tag{4.3}$$

Notice that if we replace the ϵ in the parameter settings of Theorem 3.1 with $\epsilon/(M_2+Q_2^{1/2})$, it derives $\mathbb{E}[\|c(x^r)\|] \leq \epsilon$ from (4.3), which together with (3.16) ensures that x^r, i.e. x_R is a stochastic ϵ-KKT point of (1.1). To summarize, when Algorithm 2.1 terminates, it outputs a stochastic ϵ-KKT point of (1.1). Recall that Algorithm 2.1 must terminate within K_1 iterations. As a result, the overall oracle complexity of Algorithm 2.1 is primarily determined by the complexity of the inner subproblem solver, i.e., Algorithm 3.1.

We summarize the above analysis to give the following theorem, characterizing the oracle complexity of Algorithm 2.1 to reach a stochastic ϵ-KKT point of (1.1).

THEOREM **4.1.** *Suppose that Assumptions 1.1-4.1 hold. Let x_R be the output of Algorithm 2.1 with parameter settings of Algorithm 3.1 following Theorem 3.1 except with ϵ being replaced by $\frac{\epsilon}{\max\{1,M_2+Q_2^{1/2}\}}$. Then x_R is a stochastic ϵ-KKT point of* (1.1)*. Moreover, the corresponding oracle complexity for the objective gradient, constraint gradient, and constraint function value evaluations is of order $\mathcal{O}(\epsilon^{-3})$, $\mathcal{O}(\epsilon^{-3})$, and $\mathcal{O}(\epsilon^{-5})$, respectively.*

5 Extensions to semi-stochastic, finite-sum and deterministic cases

In this section, we consider to apply the framework in Algorithm 2.1 to solve more variants of problem (1.1), including the semi-stochastic, finite-sum, and deterministic variants. By assuming appropriate values of σ_f, σ_J and σ_c in Assumptions 1.2-1.3, we can recover the semi-stochastic and deterministic cases directly. Specifically, if we set $\sigma_J = \sigma_c = 0$ but allow $\sigma_f > 0$, it turns to a semi-stochastic setting in which the constraint information is exact while the objective information is noisy. Further imposing $\sigma_f = 0$ leads to the deterministic setting, in which both the objective and constraint information are available exactly. As a special case of (1.1), the finite-sum setting arises when both objective function and constraint functions are given as the averages of a finite number of component functions.

5.1 Semi-stochastic case

In the semi-stochastic case, the stochasticity arises solely from the objective function, while exact constraint gradient and constraint function values can be accessed. The analysis for the semi-stochastic case largely follows that of the fully-stochastic case, while the only adjustment required is to replace all stochastic information of the constraint with its exact values in Algorithm 3.1. We therefore omit the detailed derivation and directly present the main results.

Under Assumptions 1.1-1.2, for fixed $\rho > 0$ and given $\tilde{\epsilon} > 0$, let $\hat{\nabla}f$ be computed through (3.2)-(3.3). Let $\gamma := 1/(8(\rho L_J + L_f))$ and batch sizes $|\mathcal{A}_i|, i = 0, \ldots, T\tau - 1$ follow the same settings as Lemma 3.4. Then in analogy to the analysis in Section 3, it holds that

$$\mathbb{E}[\|\bar{\mathcal{G}}(x^r)\|^2] \leq \frac{24}{\gamma}\left(\frac{\Phi(x^0)-C^*}{T\tau}+\tilde{\epsilon}^2\right) \quad \text{and} \quad \mathbb{E}[\|\mathcal{G}(x^r)\|^2] \leq \frac{1}{\Gamma\gamma}\left(\frac{\Phi_\rho(x^0)-C^*}{T\tau}+\tilde{\epsilon}^2\right).$$

Using the same parameter settings as Theorem 3.1, we can obtain that the output x^r of Algorithm 3.1 in semi-stochastic setting satisfies

$$\mathbb{E}[\|\bar{\mathcal{G}}(x^r)\|^2] \leq \epsilon^2 \quad \text{and} \quad \mathbb{E}[\|\mathcal{G}(x^r)\|^2] \leq \epsilon^2,$$

with an oracle complexity of order $\mathcal{O}(\rho^2\epsilon^{-3})$ regarding evaluations of stochastic objective gradients.

To obtain a stochastic ϵ-KKT point in the semi-stochastic setting, a strong LICQ assumption needs to be imposed. This constraint qualification has also been used in [2, 8, 9].

Assumption 5.1. *For any $k \geq 1$, the singular values of $\nabla c(x_k)^\top$ are uniformly bounded below by $\nu > 0$.*

This assumption corresponds to the deterministic case of Assumption 4.1. Under Assumption 5.1 and following similar analysis in Section 4, we can guarantee that Algorithm 2.1 with exact constraint information being used, terminates finitely in the semi-stochastic case and returns a stochastic ϵ-KKT point of (1.1). Accordingly, the total number of stochastic objective gradient evaluations required by Algorithm 2.1 in semi-stochastic setting is

$$TA + T\tau a = \mathcal{O}(\epsilon^{-1}) \cdot \mathcal{O}(\epsilon^{-2}) + \mathcal{O}(\epsilon^{-2}) \cdot \mathcal{O}(\epsilon^{-1}) = \mathcal{O}(\epsilon^{-3}),$$

while the complexities associated with the gradient and function evaluations of the constraints are of the same order as the inner loop iteration complexity, i.e., $\mathcal{O}(\epsilon^{-2})$. Consequently, the overall oracle complexity regarding stochastic objective gradient evaluations and constraint information evaluations in the semi-stochastic case is of order $\mathcal{O}(\epsilon^{-3})$.

Remark 5.1. *This result yields an oracle complexity bound that aligns with that in [32], but without the need of a nearly feasible initial point.*

5.2 Finite-sum case

In the finite-sum setting, we consider the problem formulated as

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) := \frac{1}{N}\sum_{j=1}^{N} f_j(x) \\ \text{s.t.} \quad & c(x) := \frac{1}{N}\sum_{j=1}^{N} c_j(x) = \mathbf{0}. \end{aligned} \tag{5.1}$$

This can be regarded as a special case of problem (1.1), where ξ follows the uniform distribution over a finite sample set $\{\xi_1, \xi_2, \ldots, \xi_N\}$ with $F(x, \xi_j)$ and $C(x, \xi_j)$ being defined as $f_j(x)$ and $c_j(x)$, respectively. In this section, we assume that each component function satisfies the following assumption.

Assumption 5.2. *For each $j = 1, \ldots, N$, functions f_j and c_j have Lipschitz continuous gradients with constant L_f^j and L_J^j, respectively, and function c_j is Lipschitz continuous with constant L_c^j.*

To solve problem (5.1), we only need to modify the computation of stochastic approximations in Algorithm 3.1 to adjust to the finite-sum structure. In this setting, we adopt a different variance reduction approach to compute stochastic estimate oracles. A similar approach with truncation is also used in [35, 38]. More specifically, we construct the estimators $\hat{\nabla} f^i$, $\hat{c}^i$ and $\hat{\nabla} c^i$ by

$$\begin{aligned} \hat{\nabla} f^i &= \Pi_{\mathbb{B}(G)}\Big(\frac{1}{|\mathcal{A}_i|}\sum_{j\in\mathcal{A}_i}\Big(\nabla f_j(x^i) - \nabla f_j(x^{\lfloor \frac{i}{\tau}\rfloor\tau})\Big) + \nabla f(x^{\lfloor \frac{i}{\tau}\rfloor\tau})\Big), \\ \hat{c}^i &= \Pi_{\mathbb{B}(M)}\Big(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i}\Big(c_j(x^i) - c_j(x^{\lfloor \frac{i}{\tau}\rfloor\tau}) - \nabla c_j(x^{\lfloor \frac{i}{\tau}\rfloor\tau})^\top(x^i - x^{\lfloor \frac{i}{\tau}\rfloor\tau})\Big) \\ &\qquad + c(x^{\lfloor \frac{i}{\tau}\rfloor\tau}) + \nabla c(x^{\lfloor \frac{i}{\tau}\rfloor\tau})^\top(x^i - x^{\lfloor \frac{i}{\tau}\rfloor\tau})\Big), \\ \hat{\nabla} c^i &= \Pi_{\mathbb{B}(L_c)}\Big(\frac{1}{|\mathcal{S}_i|}\sum_{j\in\mathcal{S}_i}\Big(\nabla c_j(x^i) - \nabla c_j(x^{\lfloor \frac{i}{\tau}\rfloor\tau})\Big) + \nabla c(x^{\lfloor \frac{i}{\tau}\rfloor\tau})\Big) \quad \text{for } i \bmod \tau \neq 0, \end{aligned} \tag{5.2}$$

where $\mathcal{A}_i,\ \mathcal{B}_i,\ \mathcal{S}_i \subseteq \{1,\dots,N\}$ are randomly and independently picked index sets and

$$\hat{\nabla} f^i = \frac{1}{N}\sum_{j=1}^{N} \nabla f_j(x^i), \quad \hat{c}^i = \frac{1}{N}\sum_{j=1}^{N} c_j(x^i), \quad \hat{\nabla} c^i = \frac{1}{N}\sum_{j=1}^{N} \nabla c_j(x^i) \quad \text{for } i \bmod \tau = 0. \tag{5.3}$$

Combining the nonexpansiveness of the projection operator and [38, Lemma 3], we can bound the approximation errors associated with these estimators.

LEMMA **5.1.** *Suppose Assumption 5.2 holds and $\hat{\nabla} f^i$, $\hat{c}_i$ and $\hat{\nabla} c^i$ are constructed through (5.2)-(5.3). Then it holds that*

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\nabla} f^i - \nabla f(x^i)\|^2 \mid x^i\right] &\le \frac{L_f^2}{|\mathcal{A}_i|}\|x^i - x^{\lfloor \frac{i}{\tau} \rfloor \tau}\|^2, \\
\mathbb{E}\left[\|\hat{c}^i - c(x^i)\| \mid x^i\right] &\le \frac{L_J}{2\sqrt{|\mathcal{B}_i|}}\|x^i - x^{\lfloor \frac{i}{\tau} \rfloor \tau}\|^2, \\
\mathbb{E}\left[\|\hat{\nabla} c^i - \nabla c(x^i)\|^2 \mid x^i\right] &\le \frac{L_J^2}{|\mathcal{S}_i|}\|x^i - x^{\lfloor \frac{i}{\tau} \rfloor \tau}\|^2 \quad \textit{for } i = 0,\dots,T\tau - 1,
\end{aligned}$$

where $\mathbb{E}[\cdot \mid x^i]$ denotes conditional expectation with respect to the random indices in $\mathcal{A}_i$, $\mathcal{B}_i$, and $\mathcal{S}_i$, and $L_f := \sqrt{\frac{1}{N}\sum_{j=1}^{N}(L_f^j)^2}$ and $L_J := \sqrt{\frac{1}{N}\sum_{j=1}^{N}(L_J^j)^2}$.

For the expected generalized gradients at the output of Algorithm 3.1 in the finite-sum setting, we can also provide upper bounds on the expected generalized gradients at the output. The detailed proof is presented in Appendix C.

Proposition 5.1. *Suppose that Assumptions 1.1, 4.1 and 5.2 hold. For fixed $\rho > 0$ and given $\epsilon > 0$, let $\gamma := 1/\left(8(\rho L_J + L_f)\right)$, $\tau = \lceil 0.5N^{1/5} - 1 \rceil$, and the estimates $\hat{\nabla} f^i$, $\hat{c}^i$ and $\hat{\nabla} c^i$ in Algorithm 3.1 be given in (5.2)-(5.3), with the batch sizes set as*

$$|\mathcal{A}_i| \equiv a = \left\lceil 2N^{2/5} \right\rceil,\ |\mathcal{B}_i| \equiv b = \left\lceil 4N^{4/5} \right\rceil,\ |\mathcal{S}_i| \equiv s = \left\lceil 2N^{2/5} \right\rceil \quad \textit{for } i \bmod \tau \neq 0.$$

Then the output x^r satisfies

$$\mathbb{E}\left[\|\bar{\mathcal{G}}(x^r)\|^2\right] \le \frac{45(\Phi_\rho(x^0) - C^*)}{T\tau\gamma} \quad \textit{and} \quad \mathbb{E}\left[\|\mathcal{G}(x^r)\|^2\right] \le \frac{82(\Phi_\rho(x^0) - C^*)}{T\tau\gamma}. \tag{5.4}$$

Proposition 5.2. *Under Assumptions 1.1, 4.1 and 5.2 and the parameter settings in Proposition 5.1 with $T = \max\{1, \lceil \frac{82(\Phi_\rho(x^0) - C^*)}{\tau\gamma\epsilon^2} \rceil\}$, the output of Algorithm 3.1 in the finite-sum setting satisfies*

$$\mathbb{E}\left[\|\bar{\mathcal{G}}(x^r)\|^2\right] \le \epsilon^2 \quad \textit{and} \quad \mathbb{E}\left[\|\mathcal{G}(x^r)\|^2\right] \le \epsilon^2,$$

and the oracle complexity of evaluating the component gradients of f and c, as well as the component constraint function values, is of order $\mathcal{O}(N + \rho^2 N^{4/5}\epsilon^{-2})$.

Proof. By the conditions of this theorem, we have $T = O(1 + \rho^2 N^{-1/5}\epsilon^{-2})$ and $\tau = O(N^{1/5})$. Consequently, the number of gradient evaluations of the component objective functions and component constraint functions are $TN + T\tau a$ and $TN + T\tau s$, respectively, and both are of order of $\mathcal{O}(N + \rho^2 N^{4/5}\epsilon^{-2})$. And the total number of evaluations of component constraint function values is $TN + T\tau b$ which is of order $\mathcal{O}(N + \rho^2 N^{4/5}\epsilon^{-2})$. The proof is completed. □

The following corollary provides the upper bounds on the variance of the estimators constructed using (5.2)–(5.3) at the output of Algorithm 3.1 in finite-sum setting. The detailed proof is presented in Appendix C.

Proposition 5.3. *For any given $\epsilon > 0$ and under the same conditions as Proposition 5.2, it holds that*

$$\mathbb{E}[\|\hat{\nabla} f^r - \nabla f(x^r)\|^2] \leq Q_4\epsilon^2, \quad \mathbb{E}[\|\hat{\nabla} c^r - \nabla c(x^r)\|^2] \leq Q_4\epsilon^2, \quad \mathbb{E}[\|\hat{c}^r - c(x^r)\|] \leq Q_4\epsilon^2, \tag{5.5}$$

with $Q_4 := \max\{\frac{\gamma^2 L_f^2}{16}, \frac{\gamma^2 L_J^2}{16}, \frac{\gamma^2 L_J}{32}\}$.

In analogy to the analysis presented in Section 4, the outer loop of Algorithm 2.1 in the finite-sum case terminate within a finite number of iterations, producing a stochastic ϵ-KKT point under Assumption 4.1. Leveraging this fact, we can establish the oracle complexity of Algorithm 2.1 with stochastic oracles computed through (5.2)–(5.3). The detailed proof is omitted here to avoid unnecessary repetition.

THEOREM **5.1.** *Suppose that Assumptions 1.1, 4.1 and 5.2 hold. Algorithm 2.1 in the finite-sum setting with the same parameter settings as proposition 5.2 returns a stochastic ϵ-KKT point of* (5.1)*. Moreover, the associated oracle complexity regarding the evaluation of component objective gradients, component constraint gradients, and component constraint function values is of order $\mathcal{O}(N + N^{4/5}\epsilon^{-2})$.*

5.3 Deterministic case

In the deterministic setting, all the exact function information is available. And the update rule (3.1) turns to

$$x^{i+1} = \underset{x}{\operatorname{argmin}} \left\{ \nabla f(x^i)^\top (x - x^i) + \rho \|c(x^i) + \nabla c(x^i)^\top (x - x^i)\| + \frac{1}{2\gamma}\|x - x^i\|^2 \right\}.$$

And meanwhile, the proof of Lemma 3.3 corresponding to the deterministic case can be significantly simplified. It can be derived from (3.12) that

$$\Phi_\rho(x^{i+1}) \leq \Phi_\rho(x^i) - \left(\frac{1}{2\gamma} - (\rho L_J + L_f)\right) \|x^{i+1} - x^i\|^2.$$

Summing the above inequalities over $i = 0, \ldots, T\tau - 1$, we have

$$\begin{aligned} \|\bar{\mathcal{G}}(x^{r_0})\|^2 \sum_{i=0}^{T\tau-1} (\frac{\gamma}{2} - \gamma^2(\rho L_J + L_f)) &\leq \sum_{i=0}^{T\tau-1} (\frac{\gamma}{2} - \gamma^2(\rho L_J + L_f)) \|\bar{\mathcal{G}}(x^i)\|^2 \\ &\leq \Phi_\rho(x^0) - \Phi_\rho(x^{T\tau}) \leq \Phi_\rho(x^0) - C^*, \end{aligned}$$

where $x^{r_0} \in \{x^i\}_{i=0,\ldots,T\tau-1}$ is chosen such that $r_0 = \arg\min_i \|\bar{\mathcal{G}}(x^i)\|^2$. Given that $\gamma := 1/\left(8(\rho L_J + L_f)\right)$, we have $\frac{\gamma}{2} - \gamma^2(\rho L_J + L_f) > \frac{\gamma}{4}$, implying that

$$\|\bar{\mathcal{G}}(x^{r_0})\|^2 \leq \frac{4(\Phi_\rho(x^0) - C^*)}{T\tau\gamma}.$$

Therefore, to achieve $\|\bar{\mathcal{G}}(x^{r_0})\| \leq \epsilon$ for a given $\epsilon > 0$, it suffices to require the total number of iterations $T\tau = \mathcal{O}(\rho^2\epsilon^{-2})$, thus the iteration complexity of Algorithm 3.1 in deterministic setting is of order $\mathcal{O}(\rho^2\epsilon^{-2})$. Unlike Section 4, in the deterministic setting exact information of both objective and constraints are available, allowing a direct derivation of the finite termination of the outer iterations based on the corresponding argument. Moreover, the reasoning in Theorem 4.1 extends naturally to the deterministic case, proving that under (1.2) and Assumption 5.1, the final output x_R of Algorithm 2.1 in deterministic setting is an ϵ-KKT point of (1.1). Consequently, Algorithm 2.1 in the deterministic setting can find an ϵ-KKT point with iteration complexity of order $\mathcal{O}(\epsilon^{-2})$, which matches the best-known result for deterministic nonconvex constrained optimization [6, 8, 10, 30].

6 Numerical Results

In this section, we consider a constrained binary classification problem [31], where the objective is to minimize the logistic loss function subject to an expected linear equality constraint together with an additional spherical constraint, formulated as

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) = \frac{1}{N}\sum_{i=1}^{N} \log\Big(1 + e^{-y_i(X_i^\top x)}\Big) \\ \text{s.t.} \quad & \mathbb{E}[Ax - a] = 0,\ \|x\|_2^2 = 1. \end{aligned}$$

Here, $X_i \in \mathbb{R}^n$, $i \in [N]$ denote the feature vectors and $y_i \in \{-1, 1\}$, $i \in [N]$ are corresponding labels with $[N] = \{1, \ldots, N\}$. The random matrix $A \in \mathbb{R}^{10\times n}$ is generated based on a fixed baseline matrix A_0, where each element at ith row and jth column of A satisfies $A^{(i,j)} \sim \mathcal{N}(A_0^{(i,j)}, \frac{10^{-3}}{n})$ with $A_0^{(i,j)} \sim \mathcal{N}(1, 100)$ for any $(i, j) \in [10] \times [n]$. Similarly, the random vector $a \in \mathbb{R}^{10}$ is drawn from $\mathcal{N}(a_0, 10^{-3}I)$, with a_0 being a fixed baseline vector whose elements are independently follow the distribution of $\mathcal{N}(1, 100)$. In practice, the expectation in the constraint is approximated by a sample-average over 1000 independent samples of (A, a) generated from the above distributions. We test this problem on the *bank-marketing* dataset from the UCI repository [13] with $n = 81$ and the *loan* dataset from LendingClub with $n = 250$ used in [20], where the feature vectors X and corresponding labels y are taken from these datasets, comparing the performance of the three algorithms TStoM [7], Stoc-iALM [20] and the extension of Algorithm 2 in [1, Section 4.1] (shortened as SLQPM below).

For both datasets, we initialize $x_1 \in \mathbb{R}^n$ as a standard Gaussian vector rescaled to satisfy $\|x_1\| = 0.01$. For the *bank-marketing* dataset, we set the maximum number of samples to 2×10^4 and $\rho_0 = 1$, $\beta = 1.2$, $\alpha = 0.8$, $\zeta = 0.8$, $\gamma = 0.001$ for Algorithm 2.1. Figure 1 presents the performances of Algorithm 2.1 in comparison with three other algorithms on the *bank-marketing* dataset. Note that all reported results are averaged over 5 independent runs of each algorithm, and the solid lines depict the mean values, while the shadow area indicates the standard deviation. It can be observed that Algorithm 2.1 outperforms the other algorithms in terms of objective function reduction, achieving a faster and more pronounced decrease. Regarding constraint violation, all four algorithms exhibit similar performance, but TStoM shows slightly better results. For the KKT-residual, Algorithm 2.1 demonstrates a clear advantage over the other methods. These results suggest that Algorithm 2.1 performs favorably compared with the other algorithms on the *bank-marketing* dataset. We then run Algorithm 2.1 with the same parameter settings as before, and compare the performance on the *loan* dataset, as shown in Figure 2. The results in Figure 2 indicate that Algorithm 2.1 shows a marked advantage with respect to objective function reduction and KKT residual, while in terms of constraint violation all algorithms behave similarly, with TStoM slightly ahead of Algorithm 2.1. It is noteworthy that, in our experiments, Algorithm 2.1 maintains stable performance even when its parameters are adjusted, while the competing methods exhibit greater sensitivity to such changes. Furthermore, when applied to a different dataset with identical parameter settings, Algorithm 2.1 exhibits robust performance, indicating its general applicability.

7 Conclusion

In this work we focus on nonconvex stochastic equality-constrained optimization. We begin by studying problems where both the objective and constraint functions are in general expectation forms and propose an exact penalty algorithm. In each iteration, the penalty parameter is adaptively updated, followed by the application of a truncated stochastic prox-linear algorithm to solve the corresponding penalized subproblem. Under certain constraint qualification assumptions, we analyze the oracle complexity of the subproblem algorithm. We also prove the exactness of the penalty method, ensuring the finite termination of the method, and establish the oracle complexity bounds regarding the stochastic gradient evaluations of both the objective and constraint functions, as well as the evaluations of the stochastic constraint function values, respectively. Furthermore, we propose variant of methods to apply for stochastic equality-constrained optimization problems

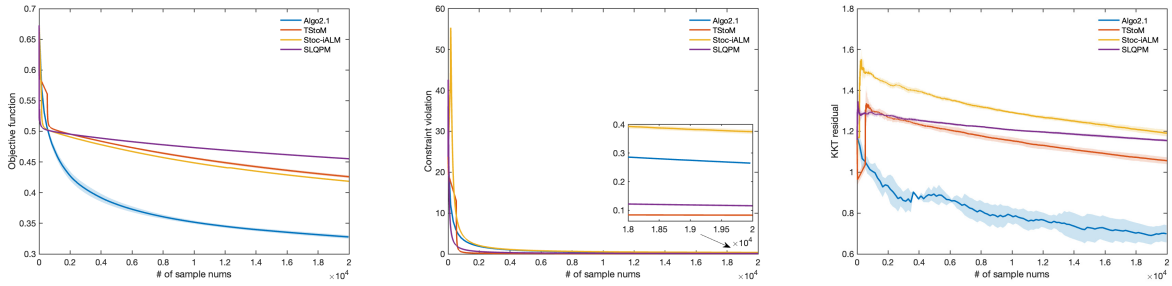

Figure 1: Comparison of Algorithm 2.1, TStoM, Stoc-iALM and SLQPM on the *bank-marketing* dataset

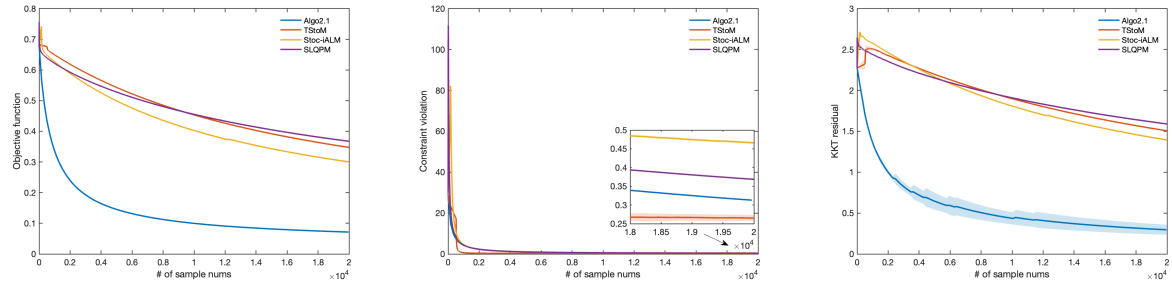

Figure 2: Comparison of Algorithm 2.1, TStoM, Stoc-iALM and SLQPM on the *loan* dataset

in semi-stochastic and finite-sum settings and for deterministic equality-constrained optimization, respectively. For each case, we present the corresponding oracle complexity analysis. Finally, we provide numerical results on a benchmark test problem to illustrate the performance of the proposed method.

Appendix

A Exactness of penalization

In this appendix, let us consider a variant of Algorithm 2.1 in the deterministic setting, where each subproblem is solved exactly, namely

$$x_k \in \operatorname{argmin} \Phi_{\rho_{k-1}}(x). \tag{A.1}$$

With a little abuse of notations, we still use ϕ_k and θ_k with definitions given by (2.4) based on exact information of the objective and constraints. When $\phi_k \geq \rho_{k-1}\zeta\theta_k$, the algorithm terminates and returns x_k; otherwise, we update the penalty parameter through

$$\rho_k = \max\{\beta\rho_{k-1}, \hat{\rho}_{k-1}\} \quad \text{with} \quad \hat{\rho}_{k-1} := \frac{\nabla f(x_k)^\top d_k + \frac{1}{2}\|d_k\|^2}{\alpha(1-\zeta)\|c(x_k)\|} \quad \text{and} \quad \beta > 1,$$

where

$$d_k = -u_k + \alpha v_k$$

with $\alpha \in (0,1)$, $u_k := \nabla f(x_k) - (\nabla c(x_k)^\top)^\dagger \nabla c(x_k)^\top \nabla f(x_k)$ and $v_k := -(\nabla c(x_k)^\top)^\dagger c(x_k)$.

In the following, we will verify that the above iteration process terminates in a finite number of iterations, and returns a KKT point of (1.1), provided that singular values of $\nabla c(x_k)^\top$ for all $k \geq 1$ are uniformly bounded below by $\nu > 0$. More specifically, under above assumption we can guarantee that $\nabla c(x_k)^\top \nabla c(x_k)$

is positive definite and $D_k := (\nabla c(x_k)^\top \nabla c(x_k))^{-1}$ satisfies $\|D_k\| \le 1/\nu^2$. Recall that when $c(x_k) = 0$, $v_k = 0$ and (2.5) holds naturally. When $c(x_k) \neq 0$, (2.5) holds whenever

$$\rho_{k-1} \ge \frac{\nabla f(x_k)^\top d_k + \frac{1}{2}\|d_k\|^2}{\alpha(1-\zeta)\|c(x_k)\|}.$$

Then it follows from the definition of d_k and Assumption 1.1 as well as (1.2) that

$$\begin{aligned}
\nabla f(x_k)^\top d_k + \frac{1}{2}\|d_k\|^2 \le & (\nabla f(x_k) + d_k)^\top d_k = ((\nabla c(x_k)^\top)^\dagger \nabla c(x_k)^\top \nabla f(x_k) - \alpha(\nabla c(x_k)^\top)^\dagger c(x_k))^\top d_k \\
= & (\nabla c(x_k) D_k \nabla c(x_k)^\top \nabla f(x_k) - \alpha \nabla c(x_k) D_k c(x_k))^\top d_k \\
= & (D_k \nabla c(x_k)^\top \nabla f(x_k) - \alpha D_k c(x_k))^\top \nabla c(x_k)^\top d_k \\
= & \alpha^2 (\nabla c(x_k) D_k c(x_k))^\top \nabla c(x_k) D_k c(x_k) - \alpha \nabla f(x_k)^\top \nabla c(x_k) D_k c(x_k) \\
\le & \alpha^2 |c(x_k)^\top D_k c(x_k)| + \alpha |\nabla f(x_k)^\top \nabla c(x_k) D_k c(x_k)| \\
\le & \alpha (\frac{1}{\nu^2}\alpha M + \frac{1}{\nu^2} G L_c)\|c(x_k)\| := \alpha M_1 \|c(x_k)\|,
\end{aligned}$$

where the fourth equality is due to $\nabla c(x_k)^\top u_k = 0$. It follows that $\phi_k \ge \rho_{k-1}\zeta\theta_k$ is satisfied if ρ_{k-1} is greater than $\frac{M_1}{1-\zeta}$, and then the algorithm terminates with output x_k. Then we can prove that the equivalence between the KKT points of problem (2.1) and that of (1.1) in the deterministic setting, stated in the following lemma, indicating the exactness of the penalty method.

THEOREM **A.1.** *Suppose that Assumptions 1.1, 5.1 and* (1.2) *hold. If* $\rho_{k-1} > \frac{M_1}{1-\zeta}$, x_k, *defined by* (A.1), *is also a KKT point of problem (1.1).*

Proof. From the optimality condition for (A.1), there exists $\omega_k \in \partial\|c(x_k)\|$ such that

$$\nabla f(x_k) + \rho_{k-1}\nabla c(x_k)\omega_k = 0, \tag{A.2}$$

and

$$0 = \arg\min_d \left\{ \nabla f(x_k)^\top d + \rho_{k-1}\|c(x_k) + \gamma \nabla c(x_k)^\top d\| + \frac{1}{2\gamma}\|d\|^2 \right\}.$$

According to the definition of ϕ_k, we have

$$\phi_k \le \rho_{k-1}\|c(x_k)\| - \min_s \left\{ \nabla f(x_k)^\top s + \rho_{k-1}\|c(x_k) + \nabla c(x_k)^\top s\| + \frac{1}{2\gamma}\|s\|^2 \right\} = 0.$$

As discussed above, if $\rho_{k-1} > \frac{M_1}{1-\zeta}$, we have $\phi_k \ge \rho_{k-1}\zeta\theta_k$, implying $c(x_k) = 0$, thus x_k is a feasible point of (1.1). Hence, with $\lambda_k = \rho_{k-1}\omega_k$ and by (A.2), x_k is a KKT point of (1.1). □

B Proofs for Section 3

In this appendix, we give the detailed proofs of Lemma 3.2, Lemma 3.3 and Lemma 3.4.

Proof of Lemma 3.2. For notational convenience, we define

$$H(x; x^i) = \rho\|c(x^i) + \nabla c(x^i)^\top (x - x^i)\| \quad \text{and} \quad \hat{H}(x; x^i) = \rho\|\hat{c}^i + (\hat{\nabla} c^i)^\top (x - x^i)\|$$

for $i = 0, \ldots, T\tau - 1$. It follows from the convexity of the function $\|\cdot\|_2$ that functions

$$H(x; x^i) + \nabla f(x^i)^\top (x - x^i) + \frac{1}{2\gamma}\|x - x^i\|^2 \text{ and } \hat{H}(x; x^i) + (\hat{\nabla} f^i)^\top (x - x^i) + \frac{1}{2\gamma}\|x - x^i\|^2$$

are $\frac{1}{\gamma}$-strongly convex. Given that $\bar{x}^i$ and x^{i+1} are minimizers of above two functions, respectively, it holds that

$$H(\bar{x}^i;x^i)+\nabla f(x^i)^\top(\bar{x}^i-x^i)+\frac{1}{2\gamma}\|\bar{x}^i-x^i\|^2\le H(x^{i+1};x^i)+\nabla f(x^i)^\top(x^{i+1}-x^i)+\frac{1}{2\gamma}\|x^{i+1}-x^i\|^2 - \frac{1}{2\gamma}\|x^{i+1}-\bar{x}^i\|^2,$$

and

$$\hat{H}(x^{i+1};x^i)+(\hat{\nabla}f^i)^\top(x^{i+1}-x^i)+\frac{1}{2\gamma}\|x^{i+1}-x^i\|^2\le \hat{H}(\bar{x}^i;x^i)+(\hat{\nabla}f^i)^\top(\bar{x}^i-x^i)+\frac{1}{2\gamma}\|\bar{x}^i-x^i\|^2 - \frac{1}{2\gamma}\|\bar{x}^i-x^{i+1}\|^2.$$

Summing the two inequalities above and rearranging terms leads to

$$\begin{aligned}\frac{1}{\gamma}\|\bar{x}^i-x^{i+1}\|^2 \le{}& H(x^{i+1};x^i)+\nabla f(x^i)^\top(x^{i+1}-x^i)-\hat{H}(x^{i+1};x^i)-(\hat{\nabla}f^i)^\top(x^{i+1}-x^i)\\ &+\hat{H}(\bar{x}^i;x^i)+(\hat{\nabla}f^i)^\top(\bar{x}^i-x^i)-H(\bar{x}^i;x^i)-\nabla f(x^i)^\top(\bar{x}^i-x^i).\end{aligned}\tag{B.1}$$

Note that it follows from the Lipschitz continuity of $\|\cdot\|$ and Young's inequality that

$$\begin{aligned}\left|H(x^{i+1};x^i)-\hat{H}(x^{i+1};x^i)\right| &= \rho\left|\|c(x^i)+\nabla c(x^i)^\top(x^{i+1}-x^i)\|-\|\hat{c}^i+(\hat{\nabla}c^i)^\top(x^{i+1}-x^i)\|\right|\\ &\le \rho\|c(x^i)-\hat{c}^i+\left(\nabla c(x^i)-\hat{\nabla}c^i\right)^\top(x^{i+1}-x^i)\|\\ &\le \rho\|c(x^i)-\hat{c}^i\|+\rho\left\|\nabla c(x^i)-\hat{\nabla}c^i\right\|\cdot\|x^{i+1}-x^i\|\\ &\le \rho\left(\|c(x^i)-\hat{c}^i\|+\frac{1}{2L_J}\|\nabla c(x^i)-\hat{\nabla}c^i\|^2+\frac{L_J}{2}\|x^{i+1}-x^i\|^2\right).\end{aligned}$$

Similarly, we can also obtain

$$\left|H(\bar{x}^i;x^i)-\hat{H}(\bar{x}^i;x^i)\right|\le\rho\left(\|\hat{c}^i-c(x^i)\|+\frac{1}{2L_J}\|\hat{\nabla}c^i-\nabla c(x^i)\|^2+\frac{L_J}{2}\|\bar{x}^i-x^i\|^2\right).$$

On the other hand, Young's inequality implies

$$\begin{aligned}\nabla f(x^i)^\top(\bar{x}^i-x^i)-(\hat{\nabla}f^i)^\top(\bar{x}^i-x^i) &= (\nabla f(x^i)-\hat{\nabla}f^i)^\top(\bar{x}^i-x^i)\\ &\le \frac{1}{2L_f}\|\hat{\nabla}f^i-\nabla f(x^i)\|^2+\frac{L_f}{2}\|\bar{x}^i-x^i\|^2\end{aligned}$$

and

$$\nabla f(x^i)^\top(x^{i+1}-x^i)-(\hat{\nabla}f^i)^\top(x^{i+1}-x^i)\le\frac{1}{2L_f}\|\hat{\nabla}f^i-\nabla f(x^i)\|^2+\frac{L_f}{2}\|x^{i+1}-x^i\|^2.$$

Plugging above relations into (B.1) derives

$$\begin{aligned}\frac{1}{\gamma}\|\bar{x}^i-x^{i+1}\|^2\le{}&\rho\left(2\|\hat{c}^i-c(x^i)\|+\frac{1}{L_J}\|\hat{\nabla}c^i-\nabla c(x^i)\|^2\right)+\frac{1}{L_f}\|\hat{\nabla}f^i-\nabla f(x^i)\|^2\\ &+\frac{\rho L_J+L_f}{2}(\|\bar{x}^i-x^i\|^2+\|x^{i+1}-x^i\|^2).\end{aligned}\tag{B.2}$$

Combining (B.2) with the fact that $\|a+b\|^2\le 2\|a\|^2+2\|b\|^2$, we have

$$\frac{1}{\gamma}\|\bar{x}^i-x^i\|^2\le\frac{2}{\gamma}\|x^{i+1}-x^i\|^2+\frac{2}{\gamma}\|\bar{x}^i-x^{i+1}\|^2$$

$$\leq \frac{2}{\gamma}\|x^{i+1}-x^i\|^2 + 2\rho\left(2\|\hat{c}^i - c(x^i)\| + \frac{1}{L_J}\|\hat{\nabla}c^i - \nabla c(x^i)\|^2\right) + \frac{2}{L_f}\|\hat{\nabla}f^i - \nabla f(x^i)\|^2$$
$$+ (\rho L_J + L_f)(\|\bar{x}^i - x^i\|^2 + \|x^{i+1}-x^i\|^2), \tag{B.3}$$

which gives

$$\left(\frac{1}{\gamma} - (\rho L_J + L_f)\right)\|\bar{x}^i - x^i\|^2$$
$$\leq \left(\frac{2}{\gamma} + (\rho L_J + L_f)\right)\|x^{i+1}-x^i\|^2 + 2\rho\left(2\|\hat{c}^i - c(x^i)\| + \frac{1}{L_J}\|\hat{\nabla}c^i - \nabla c(x^i)\|^2\right) + \frac{2}{L_f}\|\hat{\nabla}f^i - \nabla f(x^i)\|^2.$$

It follows from definitions of $\mathcal{G}$ and $\bar{\mathcal{G}}$ in (3.8) that

$$\left(\gamma - \gamma^2(\rho L_J + L_f)\right)\left\|\bar{\mathcal{G}}(x^i)\right\|^2$$
$$\leq \left(2\gamma + \gamma^2(\rho L_J + L_f)\right)\left\|\mathcal{G}(x^i)\right\|^2 + 2\rho\left(2\|\hat{c}^i - c(x^i)\| + \frac{1}{L_J}\|\hat{\nabla}c^i - \nabla c(x^i)\|^2\right) + \frac{2}{L_f}\|\hat{\nabla}f^i - \nabla f(x^i)\|^2,$$

which yields (3.10). It is also worth noting that, while (B.3) focuses on bounding $\|\bar{x}^i - x^i\|^2$, a similar bound can be derived for $\|x^{i+1}-x^i\|^2$, namely,

$$\frac{1}{\gamma}\|x^{i+1}-x^i\|^2 \leq \frac{2}{\gamma}\|x^{i+1}-\bar{x}^i\|^2 + \frac{2}{\gamma}\|\bar{x}^i - x^i\|^2$$
$$\leq \frac{2}{\gamma}\left\|\bar{x}^i - x^i\right\|^2 + 2\rho\left(2\|\hat{c}^i - c(x^i)\| + \frac{1}{L_J}\|\hat{\nabla}c^i - \nabla c(x^i)\|^2\right) + \frac{2}{L_f}\|\hat{\nabla}f^i - \nabla f(x^i)\|^2$$
$$+ (\rho L_J + L_f)(\|\bar{x}^i - x^i\|^2 + \|x^{i+1}-x^i\|^2),$$

which leads to (3.11) directly. This proof is completed. □

The following is the proof of Lemma 3.3.

Proof of Lemma 3.3. By the Lipschitz continuity of ∇c and ∇f, we have

$$\|c(x)\| \leq \|c(y) + \nabla c(y)^\top(x-y)\| + \|c(x) - c(y) - \nabla c(y)^\top(x-y)\|$$
$$\leq \|c(y) + \nabla c(y)^\top(x-y)\| + \frac{L_J}{2}\|x-y\|^2,$$

and

$$f(x) \leq f(y) + \nabla f(y)^\top(x-y) + \frac{L_f}{2}\|x-y\|^2.$$

Referring to the definition of Φ_ρ in (2.1) leads to

$$\Phi_\rho(x^{i+1})$$
$$\leq f(x^i) + \nabla f(x^i)^\top(x^{i+1}-x^i) + \rho\|c(x^i) + \nabla c(x^i)^\top(x^{i+1}-x^i)\| + \frac{(\rho L_J + L_f)}{2}\|x^{i+1}-x^i\|^2$$
$$= (\hat{\nabla}f^i)^\top(x^{i+1}-x^i) + \rho\|\hat{c}^i + (\hat{\nabla}c^i)^\top(x^{i+1}-x^i)\| + \frac{1}{2\gamma}\|x^{i+1}-x^i\|^2 - \left(\frac{1}{2\gamma} - \frac{(\rho L_J + L_f)}{2}\right)\|x^{i+1}-x^i\|^2$$
$$+ \underbrace{\rho\|c(x^i) + \nabla c(x^i)^\top(x^{i+1}-x^i)\| - \rho\|\hat{c}^i + \nabla c(x^i)^\top(x^{i+1}-x^i)\|}_{Z_1}$$
$$+ \underbrace{\rho\|\hat{c}^i + \nabla c(x^i)^\top(x^{i+1}-x^i)\| - \rho\|\hat{c}^i + (\hat{\nabla}c^i)^\top(x^{i+1}-x^i)\|}_{Z_2}$$

$$+ f(x^i) + \underbrace{(\nabla f(x^i) - (\hat{\nabla} f^i))^\top (x^{i+1} - x^i)}_{Z_3}. \tag{B.4}$$

By optimality conditions for (3.1), we have

$$(\hat{\nabla} f^i)^\top (x^{i+1} - x^i) + \rho \|\hat{c}^i + (\hat{\nabla} c^i)^\top (x^{i+1} - x^i)\| + \frac{1}{2\gamma}\|x^{i+1} - x^i\|^2 \le \rho \|\hat{c}^i\|.$$

Plugging this inequality into (B.4) gives

$$\begin{aligned}
\Phi_\rho(x^{i+1}) &\le \rho\|\hat{c}^i\| - \left(\frac{1}{2\gamma} - \frac{(\rho L_J + L_f)}{2}\right)\|x^{i+1} - x^i\|^2 + f(x^i) + Z_1 + Z_2 + Z_3 \\
&= f(x^i) + \rho\|c(x^i)\| - \left(\frac{1}{2\gamma} - \frac{(\rho L_J + L_f)}{2}\right)\|x^{i+1} - x^i\|^2 + Z_1 + Z_2 + Z_3 + \rho(\|\hat{c}^i\| - \|c(x^i)\|) \\
&= \Phi_\rho(x^i) - \left(\frac{1}{2\gamma} - \frac{(\rho L_J + L_f)}{2}\right)\|x^{i+1} - x^i\|^2 + Z_1 + Z_2 + Z_3 + \rho(\|\hat{c}^i\| - \|c(x^i)\|).
\end{aligned} \tag{B.5}$$

Note that it is easy to obtain

$$\begin{aligned}
Z_1 &\le \rho\|\hat{c}^i - c(x^i)\|, \\
Z_2 &\le \rho\left\|(\hat{\nabla} c^i - \nabla c(x^i))^\top (x^{i+1} - x^i)\right\| \le \rho\|\hat{\nabla} c^i - \nabla c(x^i)\| \cdot \|x^{i+1} - x^i\| \\
&\le \frac{\rho}{2L_J}\|\hat{\nabla} c^i - \nabla c(x^i)\|^2 + \frac{\rho L_J}{2}\|x^{i+1} - x^i\|^2, \\
Z_3 &\le \frac{1}{2L_f}\|\nabla f(x^i) - (\hat{\nabla} f^i)\|^2 + \frac{L_f}{2}\|x^{i+1} - x^i\|^2.
\end{aligned}$$

Substituting these bounds into (B.5) yields

$$\begin{aligned}
\Phi_\rho(x^{i+1}) &\le \Phi_\rho(x^i) - \left(\frac{1}{2\gamma} - \frac{(\rho L_J + L_f)}{2}\right)\|x^{i+1} - x^i\|^2 + 2\rho\|\hat{c}^i - c(x^i)\| + \frac{1}{2L_f}\|\nabla f(x^i) - (\hat{\nabla} f^i)\|^2 \\
&\quad + \frac{L_f}{2}\|x^{i+1} - x^i\|^2 + \frac{\rho}{2L_J}\|\hat{\nabla} c^i - \nabla c(x^i)\|^2 + \frac{\rho L_J}{2}\|x^{i+1} - x^i\|^2 \\
&= \Phi_\rho(x^i) - \left(\frac{1}{2\gamma} - (\rho L_J + L_f)\right)\|x^{i+1} - x^i\|^2 + 2\rho\|\hat{c}^i - c(x^i)\| + \frac{1}{2L_f}\|\nabla f(x^i) - (\hat{\nabla} f^i)\|^2 \\
&\quad + \frac{\rho}{2L_J}\left\|\hat{\nabla} c^i - \nabla c(x^i)\right\|^2.
\end{aligned}$$

After substituting $\mathcal{G}(x^i) = -\frac{1}{\gamma}(x^{i+1} - x^i)$, we derive the desired result. □

Building on the above analysis, we now prove Lemma 3.4.

Proof of Lemma 3.4. Due to the structure of Algorithm 3.1, we first focus on $i = 0, \ldots, \tau - 1$ and then extend to the whole iteration process. Specifically, we set

$$|\mathcal{A}_0| = A,\ |\mathcal{B}_0| = B,\ |\mathcal{S}_0| = S$$

and

$$|\mathcal{A}_i| \equiv a,\ |\mathcal{B}_i| \equiv b,\ |\mathcal{S}_i| \equiv s, \quad \text{for } i = 1, \ldots, \tau - 1.$$

It follows from (3.2), Assumptions 1.2-1.3 and the nonexpansiveness of the projection operator that

$$\mathbb{E}[\|\hat{\nabla} f^0 - \nabla f(x^0)\|^2] \le \frac{\sigma_f^2}{A}, \quad \mathbb{E}[\|\hat{c}^0 - c(x^0)\|^2] \le \frac{\sigma_c^2}{B}, \quad \mathbb{E}[\|\hat{\nabla} c^0 - \nabla c(x^0)\|^2] \le \frac{\sigma_J^2}{S}.$$

Note that from Lemma 3.1 and Young's inequality we can obtain

$$\begin{aligned}\mathbb{E}[\|\hat{c}^i - c(x^i)\|] &\leq \sqrt{\mathbb{E}[\|\hat{c}^i - c(x^i)\|^2]} \\ &\leq \frac{\sigma_c}{\sqrt{B}} + \sqrt{\frac{L_c^2}{b}\sum_{p=1}^{i}\mathbb{E}[\|x^p - x^{p-1}\|^2]} \\ &\leq \frac{\sigma_c}{\sqrt{B}} + \frac{\delta}{2} + \frac{L_c^2}{2b\delta}\sum_{p=1}^{i}\mathbb{E}[\|x^p - x^{p-1}\|^2] \text{ for all } i = 1, \ldots, \tau - 1,\end{aligned}$$

where $\delta = \frac{\tilde{\epsilon}^2}{4\rho}$. Then taking expectations on both sides of (3.12) and applying (3.4) and (3.6) yield

$$\begin{aligned}&\mathbb{E}[\Phi_\rho(x^{i+1})] \\ &\leq \mathbb{E}[\Phi_\rho(x^i)] - \left(\frac{1}{2\gamma} - (\rho L_J + L_f)\right)\mathbb{E}[\|x^{i+1} - x^i\|^2] + 2\rho\mathbb{E}[\|\hat{c}^i - c(x^i)\|] \\ &\quad + \frac{1}{2L_f}\mathbb{E}[\|\hat{\nabla} f^i - \nabla f(x^i)\|^2] + \frac{\rho}{2L_J}\mathbb{E}[\|\hat{\nabla} c^i - \nabla c(x^i)\|^2] \\ &\leq \mathbb{E}[\Phi_\rho(x^i)] - \left(\frac{1}{2\gamma} - (\rho L_J + L_f)\right)\mathbb{E}[\|x^{i+1} - x^i\|^2] + \frac{2\rho\sigma_c}{\sqrt{B}} + \rho\delta + \frac{\rho L_c^2}{b\delta}\sum_{p=1}^{i}\mathbb{E}[\|x^p - x^{p-1}\|^2] \\ &\quad + \frac{\sigma_f^2}{2L_f A} + \frac{L_f}{2a}\sum_{p=1}^{i}\mathbb{E}[\|x^p - x^{p-1}\|^2] + \frac{\rho\sigma_J^2}{2L_J S} + \frac{\rho L_J}{2s}\sum_{p=1}^{i}\mathbb{E}[\|x^p - x^{p-1}\|^2] \\ &= \mathbb{E}[\Phi_\rho(x^i)] - \left(\frac{1}{2\gamma} - (\rho L_J + L_f)\right)\mathbb{E}[\|x^{i+1} - x^i\|^2] + \frac{2\rho\sigma_c}{\sqrt{B}} + \frac{\sigma_f^2}{2L_f A} + \frac{\rho\sigma_J^2}{2L_J S} + \rho\delta \\ &\quad + \left(\frac{\rho L_c^2}{b\delta} + \frac{L_f}{2a} + \frac{\rho L_J}{2s}\right)\sum_{p=1}^{i}\mathbb{E}[\|x^p - x^{p-1}\|^2]. \end{aligned} \tag{B.6}$$

Recall from (3.10) and $1 - 4\gamma(\rho L_J + L_f) > 0$ that

$$\begin{aligned}&\frac{3\gamma}{4}\mathbb{E}[\|\bar{\mathcal{G}}(x^i)\|^2] \\ &\leq \left(\gamma - \gamma^2(\rho L_J + L_f)\right)\mathbb{E}[\|\bar{\mathcal{G}}(x^i)\|^2] \\ &\leq \left(\frac{2}{\gamma} + (\rho L_J + L_f)\right)\mathbb{E}[\|x^{i+1} - x^i\|^2] + 2\rho\left(2\mathbb{E}[\|\hat{c}^i - c(x^i)\|] + \frac{1}{L_J}\mathbb{E}[\|\hat{\nabla} c^i - \nabla c(x^i)\|^2]\right) \\ &\quad + \frac{2}{L_f}\mathbb{E}[\|\hat{\nabla} f^i - \nabla f(x^i)\|^2] \\ &\leq \left(\frac{2}{\gamma} + (\rho L_J + L_f)\right)\mathbb{E}[\|x^{i+1} - x^i\|^2] + \frac{4\rho\sigma_c}{\sqrt{B}} + 2\rho\delta + \frac{2\rho L_c^2}{b\delta}\sum_{p=1}^{i}\mathbb{E}[\|x^p - x^{p-1}\|^2] \\ &\quad + \frac{2\sigma_f^2}{L_f A} + \frac{2L_f}{a}\sum_{p=1}^{i}\mathbb{E}[\|x^p - x^{p-1}\|^2] + \frac{2\rho\sigma_J^2}{L_J S} + \frac{2\rho L_J}{s}\sum_{p=1}^{i}\mathbb{E}[\|x^p - x^{p-1}\|^2] \\ &= \left(\frac{2}{\gamma} + (\rho L_J + L_f)\right)\mathbb{E}[\|x^{i+1} - x^i\|^2] + \left(\frac{2\rho L_c^2}{b\delta} + \frac{2L_f}{a} + \frac{2\rho L_J}{s}\right)\sum_{p=1}^{i}\mathbb{E}[\|x^p - x^{p-1}\|^2] \\ &\quad + \frac{4\rho\sigma_c}{\sqrt{B}} + \frac{2\sigma_f^2}{L_f A} + \frac{2\rho\sigma_J^2}{L_J S} + 2\rho\delta. \end{aligned} \tag{B.7}$$

Given that $1-4\gamma(\rho L_J+L_f)>0$, we have $\frac{1}{18}\le\frac{1}{4}\cdot\frac{1-2\gamma(\rho L_J+L_f)}{2+\gamma(\rho L_J+L_f)}\le 1$. Then multiplying both sides of (B.7) by $\frac{1}{4}\cdot\frac{1-2\gamma(\rho L_J+L_f)}{2+\gamma(\rho L_J+L_f)}$ yields

$$\frac{\gamma}{24}\mathbb{E}\left[\|\bar{\mathcal{G}}(x^i)\|^2\right]\le\left(\frac{1}{4\gamma}-\frac{(\rho L_J+L_f)}{2}\right)\mathbb{E}[\|x^{i+1}-x^i\|^2]+\left(\frac{2\rho L_c^2}{b\delta}+\frac{2L_f}{a}+\frac{2\rho L_J}{s}\right)\sum_{p=1}^{i}\mathbb{E}[\|x^p-x^{p-1}\|^2]$$
$$+\frac{4\rho\sigma_c}{\sqrt{B}}+\frac{2\sigma_f^2}{L_fA}+\frac{2\rho\sigma_J^2}{L_JS}+2\rho\delta.$$

Adding the above inequality to (B.6) gives

$$\frac{\gamma}{24}\mathbb{E}\left[\|\bar{\mathcal{G}}(x^i)\|^2\right]\le\mathbb{E}\left[\Phi_\rho(x^i)\right]-\mathbb{E}\left[\Phi_\rho(x^{i+1})\right]-\left(\frac{1}{4\gamma}-\frac{(\rho L_J+L_f)}{2}\right)\mathbb{E}[\|x^{i+1}-x^i\|^2]$$
$$+\left(\frac{3\rho L_c^2}{b\delta}+\frac{5L_f}{2a}+\frac{5\rho L_J}{2s}\right)\sum_{p=1}^{i}\mathbb{E}[\|x^p-x^{p-1}\|^2]+\frac{6\rho\sigma_c}{\sqrt{B}}+\frac{5\rho\sigma_J^2}{2L_JS}+\frac{5\sigma_f^2}{2L_fA}+3\rho\delta \quad \text{(B.8)}$$

for any $i=1,\dots,\tau-1$. Note that (B.8) works for $i=0$ as well, through direct calculations. Then by summing (B.8) over $i=0,\dots,\tau-1$, we obtain

$$\frac{\gamma}{24}\sum_{i=0}^{\tau-1}\mathbb{E}\left[\|\bar{\mathcal{G}}(x^i)\|^2\right]\le\mathbb{E}\left[\Phi_\rho(x^0)\right]-\mathbb{E}\left[\Phi_\rho(x^\tau)\right]+\left(\frac{6\rho\sigma_c}{\sqrt{B}}+\frac{5\rho\sigma_J^2}{2L_JS}+\frac{5\sigma_f^2}{2L_fA}+3\rho\delta\right)\tau$$
$$-\left(\frac{1}{8\gamma}-\frac{3\tau\rho L_c^2}{b\delta}-\frac{5\tau L_f}{2a}-\frac{5\tau\rho L_J}{2s}\right)\sum_{p=1}^{\tau}\mathbb{E}\left[\|x^p-x^{p-1}\|^2\right].$$

With the parameter settings $\delta=\frac{\tilde{\epsilon}^2}{4\rho}$, $A=\frac{5\sigma_f^2}{2L_f\tilde{\epsilon}^2}$, $B=\frac{36\rho^2\sigma_c^2}{\tilde{\epsilon}^4}$, $S=\frac{5\rho\sigma_J^2}{2L_J\tilde{\epsilon}^2}$, $a=60\tau\gamma L_f$, $b=\frac{72\gamma\tau\rho L_c^2}{\delta}$, and $s=60\tau\gamma\rho L_J$, it holds that

$$\frac{1}{8\gamma}-\frac{3\tau\rho L_c^2}{b\delta}-\frac{5\tau L_f}{2a}-\frac{5\tau\rho L_J}{2s}\ge 0\quad\text{and}\quad\frac{6\rho\sigma_c}{\sqrt{B}}+\frac{5\rho\sigma_c^2}{2L_JS}+\frac{5\sigma_f^2}{2L_fA}+3\rho\delta\le 4\tilde{\epsilon}^2.$$

This leads to

$$\frac{\gamma}{24}\sum_{i=0}^{\tau-1}\mathbb{E}[\|\bar{\mathcal{G}}(x^i)\|^2]\le\mathbb{E}\left[\Phi_\rho(x^0)\right]-\mathbb{E}\left[\Phi_\rho(x^\tau)\right]+4\tau\tilde{\epsilon}^2.$$

Due to the structure of the algorithm we can easily extend above inequality to

$$\frac{\gamma}{24}\sum_{i=0}^{T\tau-1}\mathbb{E}[\|\bar{\mathcal{G}}(x^i)\|^2]\le\mathbb{E}\left[\Phi_\rho(x^0)\right]-\mathbb{E}\left[\Phi_\rho(x^{T\tau})\right]+4T\tau\tilde{\epsilon}^2.$$

Then dividing this inequality by $T\tau$ and using $\Phi_\rho(x^{T\tau})=f(x^{T\tau})+\rho\|c(x^{T\tau})\|\ge C^*$, we derive

$$\frac{1}{T\tau}\sum_{i=0}^{T\tau-1}\mathbb{E}[\|\bar{\mathcal{G}}(x^i)\|^2]\le\frac{24}{\gamma}\left(\frac{\Phi_\rho(x^0)-C^*}{T\tau}+4\tilde{\epsilon}^2\right).$$

Then since x^r is randomly chosen from $\{x^i\}_{i=0,\dots,T\tau-1}$, we derive the first conclusion in (3.13).

On the other hand, according to (B.6), we know that for any $i=0,\dots,\tau-1$,

$$\mathbb{E}[\Phi_\rho(x^{i+1})]\le\mathbb{E}[\Phi_\rho(x^i)]-\left(\frac{1}{2\gamma}-(\rho L_J+L_f)\right)\mathbb{E}[\|x^{i+1}-x^i\|^2]+\frac{2\rho\sigma_c}{\sqrt{B}}+\frac{\rho\sigma_J^2}{2L_JS}+\frac{\sigma_f^2}{2L_fA}+\rho\delta$$

$$+\left(\frac{\rho L_c^2}{b\delta}+\frac{L_f}{2a}+\frac{\rho L_J}{2s}\right)\sum_{p=1}^{i}\mathbb{E}[\|x^p-x^{p-1}\|^2].$$

Rearranging the terms in the above inequality gives

$$\begin{aligned}&\left(\frac{1}{4\gamma}-(\rho L_J+L_f)\right)\mathbb{E}[\|x^{i+1}-x^i\|^2]\\&\leq \mathbb{E}[\Phi_\rho(x^i)]-\mathbb{E}[\Phi_\rho(x^{i+1})]-\frac{1}{4\gamma}\mathbb{E}[\|x^{i+1}-x^i\|^2]+\frac{2\rho\sigma_c}{\sqrt{B}}+\frac{\rho\sigma_J^2}{2L_JS}+\frac{\sigma_f^2}{2L_fA}+\rho\delta\\&\quad+\left(\frac{\rho L_c^2}{b\delta}+\frac{L_f}{2a}+\frac{\rho L_J}{2s}\right)\sum_{p=1}^{i}\mathbb{E}[\|x^p-x^{p-1}\|^2].\end{aligned}$$

Summing up the above inequality over $i=0,\ldots,\tau-1$ gives

$$\begin{aligned}&\left(\frac{1}{4\gamma}-(\rho L_J+L_f)\right)\sum_{i=0}^{\tau-1}\mathbb{E}[\|x^{i+1}-x^i\|^2]\\&\leq \mathbb{E}[\Phi_\rho(x^0)]-\mathbb{E}[\Phi_\rho(x^\tau)]-\left(\frac{1}{4\gamma}-\frac{\tau\rho L_c^2}{b\delta}-\frac{\tau L_f}{2a}-\frac{\tau\rho L_J}{2s}\right)\sum_{i=0}^{\tau-1}\mathbb{E}[\|x^{i+1}-x^i\|^2]\\&\quad+\left(\frac{2\rho\sigma_c}{\sqrt{B}}+\frac{\rho\sigma_J^2}{2L_JS}+\frac{\sigma_f^2}{2L_fA}+\rho\delta\right)\tau.\end{aligned}$$

With the aforementioned parameter settings, i.e. $\delta=\frac{\tilde{\epsilon}^2}{4\rho}$, $A=\frac{5\sigma_f^2}{2L_f\tilde{\epsilon}^2}$, $B=\frac{36\rho^2\sigma_c^2}{\tilde{\epsilon}^4}$, $S=\frac{5\rho\sigma_J^2}{2L_J\tilde{\epsilon}^2}$, $a=60\tau\gamma L_f$, $b=\frac{72\gamma\tau\rho L_c^2}{\delta}$, and $s=60\tau\gamma\rho L_J$, we have

$$\frac{1}{4\gamma}-\frac{\tau\rho L_c^2}{b\delta}-\frac{\tau L_f}{2a}-\frac{\tau\rho L_J}{2s}>0,\quad\text{and}\quad\frac{2\rho\sigma_c}{\sqrt{B}}+\frac{\rho\sigma_J^2}{2L_JS}+\frac{\sigma_f^2}{2L_fA}+\rho\delta<4\tilde{\epsilon}^2.$$

Since $\Gamma=\frac{1}{4}-\gamma(\rho L_J+L_f)>0$, it gives

$$\Gamma\gamma\sum_{i=0}^{\tau-1}\mathbb{E}[\|\mathcal{G}(x^i)\|^2]\leq\mathbb{E}[\Phi_\rho(x^0)]-\mathbb{E}[\Phi_\rho(x^\tau)]+4\tau\tilde{\epsilon}^2.$$

Then considering $i=0,\ldots,T\tau-1$, we derive

$$\frac{1}{T\tau}\sum_{i=0}^{T\tau-1}\mathbb{E}[\|\mathcal{G}(x^i)\|^2]\leq\frac{1}{\Gamma\gamma}\left(\frac{\Phi_\rho(x^0)-C^*}{T\tau}+4\tilde{\epsilon}^2\right).$$

We thus obtain the second conclusion in (3.13). □

C Proofs for Subsection 5.2

In this appendix, we provide detailed proofs of Proposition 5.1 and Proposition 5.3.

Proof of Proposition 5.1. To simplify the convergence analysis, we adopt the following notations:

$$|\mathcal{A}_t^j|=|\mathcal{A}_{j+(t-1)\tau}|=a,\quad|\mathcal{B}_t^j|=|\mathcal{B}_{j+(t-1)\tau}|=b,\quad|\mathcal{S}_t^j|=|\mathcal{S}_{j+(t-1)\tau}|=s$$

for $j = 1, \ldots, \tau - 1; t = 1, \ldots, T$. We also define

$$x_t^j = x^{j+(t-1)\tau}, \quad \Psi_\rho(x_t^j) = \mathbb{E}[\Phi_\rho(x_t^j) + \kappa^j \|\hat{g}_t^j\|^2]$$

for $j = 0, \ldots, \tau - 1; t = 1, \ldots, T$, where $\hat{g}_t^j := \frac{1}{\gamma}(x_t^0 - x_t^j)$ and κ^j is defined as

$$\kappa^\tau = 0, \quad \kappa^j = \kappa^{j+1}\left(1 + \frac{1}{\tau}\right) + \frac{\gamma}{5a} + \frac{\gamma}{3\sqrt{b}} + \frac{\gamma}{5s}.$$

By taking expectations on both sides of (3.12) and applying Lemma 5.1, we obtain

$$\begin{aligned}
&\mathbb{E}[\Phi_\rho(x_t^{j+1})] \\
&\leq \mathbb{E}[\Phi_\rho(x_t^j)] - \left(\frac{1}{2\gamma} - (\rho L_J + L_f)\right)\mathbb{E}[\|x_t^{j+1} - x_t^j\|^2] + 2\rho\mathbb{E}[\|\hat{c}_t^j - c(x_t^j)\|] \\
&\quad + \frac{1}{2L_f}\mathbb{E}[\|\nabla f(x_t^j) - \hat{\nabla} f_t^j\|^2] + \frac{\rho}{2L_J}\mathbb{E}[\|\hat{\nabla} c_t^j - \nabla c(x_t^j)\|^2] \\
&= \mathbb{E}[\Phi_\rho(x_t^j)] - \left(\frac{1}{2\gamma} - (\rho L_J + L_f)\right)\mathbb{E}[\|x_t^{j+1} - x_t^j\|^2] + \left(\frac{L_f}{2a} + \frac{\rho L_J}{\sqrt{b}} + \frac{\rho L_J}{2s}\right)\mathbb{E}[\|x_t^j - x_t^0\|^2]
\end{aligned}$$

for $j = 0, \ldots, \tau - 1$ and $t = 1, \ldots, T$. By the definition of $\hat{g}_t^j$ and $\mathcal{G}(x_t^j) := \frac{1}{\gamma}(x_t^j - x_t^{j+1})$, we have $\hat{g}_t^j = \sum_{p=0}^{j-1} \mathcal{G}(x_t^p)$. Moreover, it holds that

$$\mathbb{E}[\|\hat{g}_t^{j+1}\|^2] = \mathbb{E}[\|\hat{g}_t^j + \mathcal{G}(x_t^j)\|^2] \leq \left(1 + \frac{1}{\tau}\right)\mathbb{E}[\|\hat{g}_t^j\|^2] + (1 + \tau)\mathbb{E}[\|\mathcal{G}(x_t^j)\|^2].$$

Hence, we obtain

$$\begin{aligned}
&\mathbb{E}[\Phi_\rho(x_t^{j+1}) + \kappa^{j+1}\|\hat{g}_t^{j+1}\|^2] \\
&\leq \mathbb{E}[\Phi_\rho(x_t^j)] - \left(\frac{\gamma}{2} - \gamma^2(\rho L_J + L_f) - \kappa^{j+1}(1 + \tau)\right)\mathbb{E}[\|\mathcal{G}(x_t^j)\|^2] \\
&\quad + \left(\frac{\gamma^2 L_f}{2a} + \frac{\gamma^2 \rho L_J}{\sqrt{b}} + \frac{\gamma^2 \rho L_J}{2s} + \kappa^{j+1}(1 + \frac{1}{\tau})\right)\mathbb{E}[\|\hat{g}_t^j\|^2] \\
&\leq \mathbb{E}[\Phi_\rho(x_t^j)] - \left(\frac{\gamma}{4} - \kappa^{j+1}(1 + \tau)\right)\mathbb{E}[\|\mathcal{G}(x_t^j)\|^2] + \left(\frac{\gamma}{4}\left(\frac{1}{2a} + \frac{1}{\sqrt{b}} + \frac{1}{2s}\right) + \kappa^{j+1}(1 + \frac{1}{\tau})\right)\mathbb{E}[\|\hat{g}_t^j\|^2], \quad \text{(C.1)}
\end{aligned}$$

where the last inequality follows from $1 - 4\gamma(\rho L_J + L_f) > 0$ by the definition of γ. Recall from (3.10) and Lemma 5.1 that

$$\begin{aligned}
&\left(\gamma - \gamma^2(\rho L_J + L_f)\right)\mathbb{E}[\|\bar{\mathcal{G}}(x_t^j)\|^2] \\
&\leq \left(2\gamma + \gamma^2(\rho L_J + L_f)\right)\mathbb{E}[\|\mathcal{G}(x_t^j)\|^2] + 2\rho\left(2\mathbb{E}[\|\hat{c}_t^j - c(x_t^j)\|] + \frac{1}{L_J}\mathbb{E}[\|\hat{\nabla} c_t^j - \nabla c(x_t^j)\|^2]\right) \\
&\quad + \frac{2}{L_f}\mathbb{E}[\|\hat{\nabla} f_t^j - \nabla f(x_t^j)\|^2] \\
&\leq \left(2\gamma + \gamma^2(\rho L_J + L_f)\right)\mathbb{E}[\|\mathcal{G}(x_t^j)\|^2] + \left(\frac{2L_f}{a} + \frac{2\rho L_J}{\sqrt{b}} + \frac{2\rho L_J}{s}\right)\mathbb{E}[\|x_t^j - x_t^0\|^2].
\end{aligned}$$

With $1 - 4\gamma(\rho L_J + L_f) > 0$, we have

$$\begin{aligned}
\frac{3\gamma}{4}\mathbb{E}\left[\|\bar{\mathcal{G}}(x_t^j)\|^2\right] &\leq \frac{9\gamma}{4}\mathbb{E}[\|\mathcal{G}(x_t^j)\|^2] + 2\gamma^2\left(\frac{L_f}{a} + \frac{\rho L_J}{\sqrt{b}} + \frac{\rho L_J}{s}\right)\mathbb{E}[\|\hat{g}_t^j\|^2] \\
&\leq \frac{9\gamma}{4}\mathbb{E}[\|\mathcal{G}(x_t^j)\|^2] + \frac{\gamma}{2}\left(\frac{1}{a} + \frac{1}{\sqrt{b}} + \frac{1}{s}\right)\mathbb{E}[\|\hat{g}_t^j\|^2].
\end{aligned}$$

Multiplying both sides of above inequality by $\frac{4}{9\gamma}(\frac{\gamma}{4}-\kappa^{j+1}(1+\tau))$, which is positive due to the settings of $\tau,\ a,\ b$ and s, and adding it to (C.1), we have

$$\begin{aligned}
&\mathbb{E}[\Phi_\rho(x_t^{j+1})+\kappa^{j+1}\|\hat{g}_t^{j+1}\|^2] \\
&\leq \mathbb{E}[\Phi_\rho(x_t^j)]+\left(\frac{\gamma}{5a}+\frac{\gamma}{3\sqrt{b}}+\frac{\gamma}{5s}+\kappa^{j+1}(1+\frac{1}{\tau})\right)\mathbb{E}[\|\hat{g}_t^j\|^2]-\frac{1}{3}\left(\frac{\gamma}{4}-\kappa^{j+1}(1+\tau)\right)\mathbb{E}\left[\|\bar{\mathcal{G}}(x_t^j)\|^2\right].
\end{aligned}$$

Then it gives

$$\frac{1}{3}\left(\frac{\gamma}{4}-\kappa^{j+1}(1+\tau)\right)\mathbb{E}\left[\|\bar{\mathcal{G}}(x_t^j)\|^2\right]\leq\Psi_\rho(x_t^j)-\Psi_\rho(x_t^{j+1}).$$

Denoting $\omega:=\min_{0\leq j\leq\tau-1}\frac{1}{3}(\frac{\gamma}{4}-\kappa^{j+1}(1+\tau))$ and summing up the above inequality over j from 0 to $\tau-1$, we can obtain

$$\omega\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\bar{\mathcal{G}}(x_t^j)\|^2\right]\leq\Psi_\rho(x_t^0)-\Psi_\rho(x_t^\tau)=\Phi_\rho(x_t^0)-\Phi_\rho(x_t^\tau), \tag{C.2}$$

where the equality is due to that $\hat{g}_t^0=0$ and $\kappa^\tau=0$. Recalling the definition of κ^j, we know that

$$\kappa^j+Z=(\kappa^{j+1}+Z)\left(1+\frac{1}{\tau}\right),\quad\text{where}\quad Z=\tau\gamma\left(\frac{1}{5a}+\frac{1}{3\sqrt{b}}+\frac{1}{5s}\right).$$

It further implies that

$$\kappa^j=(\kappa^\tau+Z)\left(1+\frac{1}{\tau}\right)^{\tau-j}-Z\leq Z\left(1+\frac{1}{\tau}\right)^\tau-Z\leq Ze-Z=Z(e-1), \tag{C.3}$$

where the first equality is obtained by recursively expanding the given recurrence relation, starting from the terminal condition $\kappa^\tau=0$, and the last inequality comes from $(1+1/\tau)^\tau\leq e$. Hence, we have

$$\begin{aligned}
\omega&=\min_{0\leq j\leq\tau-1}\frac{1}{3}\left(\frac{\gamma}{4}-\kappa^{j+1}(1+\tau)\right) \\
&\geq\frac{1}{3}\left(\frac{\gamma}{4}-Z(e-1)(1+\tau)\right)=\frac{\gamma}{3}\left(\frac{1}{4}-\tau\left(\frac{1}{5a}+\frac{1}{3\sqrt{b}}+\frac{1}{5s}\right)(e-1)(1+\tau)\right) \\
&\geq\frac{\gamma}{3}\left(\frac{1}{4}-\left(\frac{1}{5a}+\frac{1}{3\sqrt{b}}+\frac{1}{5s}\right)2(1+\tau)^2\right).
\end{aligned}$$

With the parameter settings $\tau=\lceil\frac{1}{2}N^{1/5}-1\rceil$, $a=\lceil 2N^{2/5}\rceil$, $b=\lceil 4N^{4/5}\rceil$, and $s=\lceil 2N^{2/5}\rceil$, it holds that $\omega\geq\frac{\gamma}{45}$. Then making a summation of (C.2) over $t=1,\ldots,T$ and dividing it by $T\tau$ derives

$$\frac{1}{T\tau}\sum_{t=1}^{T}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\bar{\mathcal{G}}(x_t^j)\|^2\right]\leq\frac{\Phi_\rho(x_1^0)-\Phi_\rho(x_T^\tau)}{T\tau\omega}\leq\frac{45(\Phi_\rho(x_1^0)-C^*)}{T\tau\gamma}.$$

Similarly, recall from (3.11) and Lemma 5.1 that

$$\begin{aligned}
&\left(\gamma-\gamma^2(\rho L_J+L_f)\right)\mathbb{E}[\|\mathcal{G}(x_t^j)\|^2] \\
&\leq\left(2\gamma+\gamma^2(\rho L_J+L_f)\right)\mathbb{E}[\|\bar{\mathcal{G}}(x_t^j)\|^2]+2\rho\left(2\mathbb{E}[\|\hat{c}_t^j-c(x_t^j)\|]+\frac{1}{L_J}\mathbb{E}[\|\hat{\nabla}c_t^j-\nabla c(x_t^j)\|^2]\right) \\
&\quad+\frac{2}{L_f}\mathbb{E}[\|\hat{\nabla}f_t^j-\nabla f(x_t^j)\|^2] \\
&\leq\left(2\gamma+\gamma^2(\rho L_J+L_f)\right)\mathbb{E}[\|\bar{\mathcal{G}}(x_t^j)\|^2]+\left(\frac{2L_f}{a}+\frac{2\rho L_J}{\sqrt{b}}+\frac{2\rho L_J}{s}\right)\mathbb{E}[\|x_t^j-x_t^0\|^2].
\end{aligned}$$

It follows from $1-4\gamma(\rho L_J+L_f)>0$ that

$$\begin{aligned}\frac{3\gamma}{4}\mathbb{E}\left[\|\mathcal{G}(x_t^j)\|^2\right] &\leq \frac{9\gamma}{4}\mathbb{E}[\|\bar{\mathcal{G}}(x_t^j)\|^2]+2\gamma^2\left(\frac{L_f}{a}+\frac{\rho L_J}{\sqrt{b}}+\frac{\rho L_J}{s}\right)\mathbb{E}[\|\hat{g}_t^j\|^2]\\ &\leq \frac{9\gamma}{4}\mathbb{E}[\|\bar{\mathcal{G}}(x_t^j)\|^2]+\frac{\gamma}{2}\left(\frac{1}{a}+\frac{1}{\sqrt{b}}+\frac{1}{s}\right)\mathbb{E}[\|\hat{g}_t^j\|^2].\end{aligned}$$

Multiplying both sides of above inequality by $\frac{1}{9}$ and adding it to (C.1) yields

$$\begin{aligned}&\mathbb{E}[\Phi_\rho(x_t^{j+1})+\kappa^{j+1}\|\hat{g}_t^{j+1}\|^2]\\ &\leq \mathbb{E}[\Phi_\rho(x_t^j)]+\left(\frac{\gamma}{5a}+\frac{\gamma}{3\sqrt{b}}+\frac{\gamma}{5s}+\kappa^{j+1}(1+\frac{1}{\tau})\right)\mathbb{E}[\|\hat{g}_t^j\|^2]+\frac{\gamma}{4}\mathbb{E}\left[\|\bar{\mathcal{G}}(x_t^j)\|^2\right]\\ &\quad-\left(\frac{\gamma}{3}-\kappa^{j+1}(1+\tau)\right)\mathbb{E}\left[\|\mathcal{G}(x_t^j)\|^2\right].\end{aligned}$$

Then it gives

$$\left(\frac{\gamma}{3}-\kappa^{j+1}(1+\tau)\right)\mathbb{E}\left[\|\mathcal{G}(x_t^j)\|^2\right]\leq\Psi_\rho(x_t^j)-\Psi_\rho(x_t^{j+1})+\frac{\gamma}{4}\mathbb{E}\left[\|\bar{\mathcal{G}}(x_t^j)\|^2\right].$$

Denoting $\bar{\omega}:=\min_{0\leq j\leq\tau-1}(\frac{\gamma}{3}-\kappa^{j+1}(1+\tau))$ and summing up the above inequality over j from 0 to $\tau-1$, we can obtain

$$\begin{aligned}\bar{\omega}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\mathcal{G}(x_t^j)\|^2\right] &\leq \Psi_\rho(x_t^0)-\Psi_\rho(x_t^\tau)+\frac{\gamma}{4}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\bar{\mathcal{G}}(x_t^j)\|^2\right]\\ &=\Phi_\rho(x_t^0)-\Phi_\rho(x_t^\tau)+\frac{\gamma}{4}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\bar{\mathcal{G}}(x_t^j)\|^2\right],\end{aligned}\tag{C.4}$$

where the equality is due to the fact that $\hat{g}_t^0=0$ and $\kappa^\tau=0$. It is easy to know from (C.3) that

$$\begin{aligned}\bar{\omega}&=\min_{0\leq j\leq\tau-1}\left(\frac{\gamma}{3}-\kappa^{j+1}(1+\tau)\right)\\ &\geq\frac{\gamma}{3}-Z(e-1)(1+\tau)=\gamma\left(\frac{1}{3}-\tau\left(\frac{1}{5a}+\frac{1}{3\sqrt{b}}+\frac{1}{5s}\right)(e-1)(1+\tau)\right)\\ &\geq\gamma\left(\frac{1}{3}-\left(\frac{1}{5a}+\frac{1}{3\sqrt{b}}+\frac{1}{5s}\right)2(1+\tau)^2\right).\end{aligned}$$

With the parameter settings $\tau=\lceil\frac{1}{2}N^{1/5}-1\rceil$, $a=\lceil 2N^{2/5}\rceil$, $b=\lceil 4N^{4/5}\rceil$, and $s=\lceil 2N^{2/5}\rceil$, it holds that $\bar{\omega}\geq\frac{3\gamma}{20}$. Then making a summation of (C.4) over $t=1,\ldots,T$, and dividing it by $T\tau$, we derive

$$\frac{1}{T\tau}\sum_{t=1}^{T}\sum_{j=0}^{\tau-1}\mathbb{E}\left[\|\bar{\mathcal{G}}(x_t^j)\|^2\right]\leq\frac{\Phi_\rho(x_1^0)-\Phi_\rho(x_T^\tau)}{T\tau\bar{\omega}}+\frac{45(\Phi_\rho(x_1^0)-C^*)}{4T\tau\bar{\omega}}\leq\frac{82(\Phi_\rho(x_1^0)-C^*)}{T\tau\gamma}.$$

In summary, since x^r is randomly chosen from $\{x_t^j\}_{t=1,\ldots,T}^{j=0,\ldots,\tau-1}$ and noting that $x_1^0=x^0$, we obtain (5.4). This finishes the proof. □

Proof of Proposition 5.3. Using the same notations as in the proof of Proposition 5.1, we can derive

$$\|\hat{g}_t^j\|^2=\left\|\sum_{p=0}^{j-1}\mathcal{G}(x_t^p)\right\|^2\leq j\sum_{p=0}^{j-1}\|\mathcal{G}(x_t^p)\|^2.$$

Taking the expectation of both sides in the above inequality and sum-averaging over the $T\tau$ iterations yields

$$\frac{1}{T\tau}\sum_{t=1}^{T}\sum_{j=0}^{\tau-1}\mathbb{E}[\|\hat{g}_t^j\|^2] \le \frac{1}{T\tau}\sum_{t=1}^{T}\sum_{j=0}^{\tau-1} j\sum_{p=0}^{j-1}\mathbb{E}[\|\mathcal{G}(x_t^p)\|^2] = \frac{1}{T\tau}\sum_{t=1}^{T}\sum_{p=0}^{\tau-1}\mathbb{E}[\|\mathcal{G}(x_t^p)\|^2]\sum_{j=p+1}^{\tau-1} j$$
$$\le \frac{1}{T\tau}\sum_{t=1}^{T}\sum_{p=0}^{\tau-1}\mathbb{E}[\|\mathcal{G}(x_t^p)\|^2]\cdot\frac{\tau(\tau-1)}{2}.$$

Together with the parameter settings in Proposition 5.1, it follows from Lemma 5.1 and Proposition 5.2 that

$$\mathbb{E}[\|\hat{\nabla} f^r - \nabla f(x^r)\|^2] = \frac{1}{T\tau}\sum_{t=1}^{T}\sum_{j=0}^{\tau-1}\mathbb{E}[\|\hat{\nabla} f_t^j - \nabla f(x_t^j)\|^2] \le \frac{\gamma^2 L_f^2}{a}\cdot\frac{1}{T\tau}\sum_{t=1}^{T}\sum_{j=0}^{\tau-1}\mathbb{E}[\|\hat{g}_t^j\|^2] \le \frac{\gamma^2 L_f^2}{16}\epsilon^2;$$
$$\mathbb{E}[\|\hat{c}^r - c(x^r)\|] = \frac{1}{T\tau}\sum_{t=1}^{T}\sum_{j=0}^{\tau-1}\mathbb{E}[\|\hat{c}_t^j - c(x_t^j)\|] \le \frac{\gamma^2 L_J}{2\sqrt{b}}\cdot\frac{1}{T\tau}\sum_{t=1}^{T}\sum_{j=0}^{\tau-1}\mathbb{E}[\|\hat{g}_t^j\|^2] \le \frac{\gamma^2 L_J}{32}\epsilon^2;$$
$$\mathbb{E}[\|\hat{\nabla} c^r - \nabla c(x^r)\|^2] = \frac{1}{T\tau}\sum_{t=1}^{T}\sum_{j=0}^{\tau-1}\mathbb{E}[\|\hat{\nabla} c_t^j - \nabla c(x_t^j)\|^2] \le \frac{\gamma^2 L_J^2}{s}\cdot\frac{1}{T\tau}\sum_{t=1}^{T}\sum_{j=0}^{\tau-1}\mathbb{E}[\|\hat{g}_t^j\|^2] \le \frac{\gamma^2 L_J^2}{16}\epsilon^2,$$

which derives (5.5). □

References

[1] A. Alacaoglu and S. J. Wright. Complexity of single loop algorithms for nonlinear programming with stochastic objective and constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 4627–4635. PMLR, 2024.

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

[3] J. T. Betts. *Practical methods for optimal control and estimation using nonlinear programming*. SIAM, 2010.

[4] D. Boob, Q. Deng, and G. Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, 197(1):215–279, 2023.

[5] D. Boob, Q. Deng, and G. Lan. Level constrained first order methods for function constrained optimization. *Mathematical Programming*, 209(1):1–61, 2025.

[6] C. Cartis, N. I. Gould, and P. L. Toint. Corrigendum: On the complexity of finding first-order critical points in constrained nonlinear optimization. *Mathematical Programming*, 161(1):611–626, 2017.

[7] Y. Cui, X. Wang, and X. Xiao. A two-phase stochastic momentum-based algorithm for nonconvex expectation-constrained optimization. *Journal of Scientific Computing*, 104(1):16, 2025.

[8] F. E. Curtis, M. J. O'Neill, and D. P. Robinson. Worst-case complexity of an sqp method for nonlinear equality constrained stochastic optimization. *Mathematical Programming*, 205(1):431–483, 2024.

[9] F. E. Curtis, D. P. Robinson, and B. Zhou. Sequential quadratic optimization for stochastic optimization with deterministic nonlinear inequality and equality constraints. *SIAM Journal on Optimization*, 34(4):3592–3622, 2024.

[10] Y. Diouane, M. Gollier, and D. Orban. A nonsmooth exact penalty method for equality-constrained optimization: complexity and implementation. *arXiv preprint arXiv:2410.02188*, 2024.

[11] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems*, 31, 2018.

[12] D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

[13] D. Dua and C. Graff. UCI machine learning repository, 2017.

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

[15] Y. Fang, S. Na, M. W. Mahoney, and M. Kolar. Trust-region sequential quadratic programming for stochastic optimization with random models. *arXiv preprint arXiv:2409.15734*, 2024.

[16] D. B. Gahururu. *PDE-Constrained Equilibrium Problems under Uncertainty: Existence, Optimality Conditions and Regularization*. PhD thesis, Philipps-Universität Marburg, 2021.

[17] B. M. Idrees, L. Arora, and K. Rajawat. Constrained stochastic recursive momentum successive convex approximation. *IEEE Transactions on Signal Processing*, 73, 2025.

[18] L. Jin and X. Wang. A stochastic primal-dual method for a class of nonconvex constrained optimization. *Computational Optimization and Applications*, 83(1):143–180, 2022.

[19] L. Jin and X. Wang. Stochastic nested primal-dual method for nonconvex constrained composition optimization. *Mathematics of Computation*, 94(351):305–358, 2025.

[20] Z. Li, P. Chen, S. Liu, S. Lu, and Y. Xu. Stochastic inexact augmented lagrangian method for nonconvex expectation constrained optimization. *Computational Optimization and Applications*, 87(1):117–147, 2024.

[21] Q. Lin, R. Ma, and T. Yang. Level-set methods for finite-sum constrained convex optimization. In *International Conference on Machine Learning*, pages 3112–3121. PMLR, 2018.

[22] W. Liu and Y. Xu. A single-loop spider-type stochastic subgradient method for expectation-constrained nonconvex nonsmooth optimization. *arXiv preprint arXiv:2501.19214*, 2025.

[23] 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.

[24] B. L. Miller and H. M. Wagner. Chance constrained programming with joint constraints. *Operations Research*, 13(6):930–945, 1965.

[25] S. Na, M. Anitescu, and M. Kolar. An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. *Mathematical Programming*, 199(1):721–791, 2023.

[26] S. Na and M. Mahoney. Statistical inference of constrained stochastic optimization via sketched sequential quadratic programming. *Journal of Machine Learning Research*, 26(33):1–75, 2025.

[27] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, 2006.

[28] A. Prékopa. On probabilistic constrained programming. In *Proceedings of the Princeton Symposium on Mathematical Programming*, volume 113, page 138. Princeton, NJ, 1970.

[29] T. Rees, H. S. Dollar, and A. J. Wathen. Optimal solvers for pde-constrained optimization. *SIAM Journal on Scientific Computing*, 32(1):271–298, 2010.

[30] S. Schechtman, D. Tiapkin, M. Muehlebach, and E. Moulines. Orthogonal directions constrained gradient method: from non-linear equality constraints to stiefel manifold. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1228–1258. PMLR, 2023.

[31] H. Shen, Y. Zeng, and B. Zhou. Sequential quadratic optimization for solving expectation equality constrained stochastic optimization problems. *arXiv preprint arXiv:2503.09490*, 2025.

[32] Q. Shi, X. Wang, and H. Wang. A momentum-based linearized augmented lagrangian method for nonconvex constrained stochastic optimization. *Mathematics of Operations Research*, 2025.

[33] X. Wang. Complexity analysis of inexact cubic-regularized primal-dual methods for finding second-order stationary points. *Mathematics of Computation*, 94(356):2961–3008, 2025.

[34] X. Wang, S. Ma, and Y. Yuan. Penalty methods with stochastic approximation for stochastic nonlinear programming. *Mathematics of Computation*, 86(306):1793–1820, 2017.

[35] Y. Wu and B. Grimmer. Some unified theory for variance reduced prox-linear methods. *arXiv preprint arXiv:2412.15008*, 2024.

[36] Y. Yan and Y. Xu. Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs. *Mathematical Programming Computation*, 14(2):319–363, 2022.

[37] M. Yang, G. Li, Q. Hu, Q. Lin, and T. Yang. Single-loop algorithms for stochastic non-convex optimization with weakly-convex constraints. *arXiv preprint arXiv:2504.15243*, 2025.

[38] J. Zhang and L. Xiao. Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization. *Mathematical Programming*, 195:649–691, 2022.

[39] S. Zuo, X. Wang, and H. Wang. An adaptive single-loop stochastic penalty method for nonconvex constrained stochastic optimization. *https://optimization-online.org/?p=30110*, 2025.