Variance-reduced first-order methods for deterministically constrained stochastic nonconvex optimization with strong convergence guarantees

Zhaosong Lu * Sanyou Mei * Yifeng Xiao *

September 15, 2024

Abstract

In this paper, we study a class of deterministically constrained stochastic optimization problems. Existing methods typically aim to find an ϵ-stochastic stationary point, where the expected violations of both the constraints and first-order stationarity are within a prescribed accuracy of ϵ. However, in many practical applications, it is crucial that the constraints be nearly satisfied with certainty, making such an ϵ-stochastic stationary point potentially undesirable due to the risk of significant constraint violations. To address this issue, we propose single-loop variance-reduced stochastic first-order methods, where the stochastic gradient of the stochastic component is computed using either a truncated recursive momentum scheme or a truncated Polyak momentum scheme for variance reduction, while the gradient of the deterministic component is computed exactly. Under the error bound condition with a parameter $\theta \geq 1$ and other suitable assumptions, we establish that the proposed methods achieve a sample complexity and first-order operation complexity of $\tilde{\mathcal{O}}(\epsilon^{-\max\{4,2\theta\}})$ [1] for finding a stronger ϵ-stochastic stationary point, where the constraint violation is within ϵ with *certainty*, and the expected violation of first-order stationarity is within ϵ. To the best of our knowledge, this is the first work to develop methods with provable complexity guarantees for finding an approximate stochastic stationary point of such problems that nearly satisfies all constraints with *certainty*.

Keywords: stochastic optimization, Polyak momentum, recursive momentum, variance reduction, quadratic penalty, sample complexity

Mathematics Subject Classification: 90C15, 90C26, 90C30, 65K05

1 Introduction

In this paper, we consider constrained stochastic nonconvex optimization problems in the form of

$$\begin{aligned} \min_{x\in X} \quad & f(x) := \mathbb{E}[\tilde{f}(x,\xi)] \\ \text{s.t.} \quad & c(x) = 0, \end{aligned} \tag{1}$$

where $X \subseteq \mathbb{R}^n$ is a simple closed convex set,[2] ξ is a random variable with sample space Ξ, $\tilde{f}(\cdot,\xi)$ is continuously differentiable for each $\xi \in \Xi$, and $c : \mathbb{R}^n \to \mathbb{R}^m$ is a deterministic smooth mapping.

*Department of Industrial and Systems Engineering, University of Minnesota, USA (email: zhaosong@umn.edu, mei00035@umn.edu, xiao0414@umn.edu). This work was partially supported by the National Science Foundation Award IIS-2211491 and the Office of Naval Research Award N00014-24-1-2702.

[1] The symbol $\widetilde{\mathcal{O}}(\cdot)$ denotes the asymptotic upper bound that ignores logarithmic factors.

[2] The set X is said to be simple if the projection of any point onto X can be computed exactly.

Problem (1) arises in a variety of important areas, including energy systems [33], healthcare [34], image processing [26], machine learning [9, 20], network optimization [5], optimal control [6], PDE-constrained optimization [31], resource allocation [16], and transportation [27]. More applications can be found, for example, in [7, 8, 19, 23], and references therein.

Numerous stochastic gradient methods have been developed for solving specific instances of problem (1) with $c = 0$ (e.g., see [13, 14, 17, 18, 36, 38, 39]). Notably, when f is Lipschitz smooth (see Assumption 3), the methods in [17, 18] achieve a sample complexity of $\mathcal{O}(\epsilon^{-4})$ for finding an ϵ-stochastic stationary point x satisfying

$$\mathbb{E}[\mathrm{dist}\,(0, \nabla f(x) + \mathcal{N}_X(x))] \leq \epsilon.$$

Furthermore, when $\tilde{f}(\cdot, \xi)$ is Lipschitz smooth on average (see Assumption 2), the methods in [13, 14, 36, 38, 39] improve this sample complexity to $\mathcal{O}(\epsilon^{-3})$ for finding an ϵ-stochastic stationary point.

Additionally, various methods have been proposed for problem (1) with $X = \mathbb{R}^n$ and $c \neq 0$. For instance, [37] developed a stochastic penalty method that applies a stochastic gradient method to solve a sequence of quadratic penalty subproblems. Stochastic sequential quadratic programming (SQP) methods have also been proposed in [2, 3, 4, 11, 12, 15, 28, 29], which modify the classical SQP framework by using stochastic approximations of f and by appropriately selecting step sizes. Under suitable assumptions, these methods ensure the asymptotic convergence of the expected violations of feasibility and first-order stationarity to zero. Moreover, the methods in [10, 29] guarantee almost-sure convergence of these quantities. Besides, the sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-4})$ for finding an ϵ-stochastic stationary point is achieved by methods in [11, 29]. It is worth mentioning that their operation complexity is often higher than the sample complexity, due to the need to solve linear systems. Additionally, these methods may not be applicable to problem (1) when $X \neq \mathbb{R}^n$.

Recently, several methods have been proposed for solving problem (1) with $X \neq \mathbb{R}^n$ and $c \neq 0$. For instance, [35] proposed a momentum-based linearized augmented Lagrangian method for this problem, achieving a sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-5})$ for finding an ϵ-stochastic stationary point that satisfies

$$\mathbb{E}[\|c(x)\|] \leq \epsilon, \qquad \mathbb{E}[\mathrm{dist}\,(0, \nabla f(x) + \nabla c(x)\lambda + \mathcal{N}_X(x))] \leq \epsilon \tag{2}$$

for some Lagrangian multiplier λ. This sample complexity improves to $\widetilde{\mathcal{O}}(\epsilon^{-4})$ when a nearly feasible point of (1) is available. More recently, [1] proposed a stochastic quadratic penalty method that iteratively applies a single stochastic gradient descent step to a sequence of quadratic penalty functions $Q_{\rho_k}(x)$, where ρ_k is a penalty parameter, and Q_ρ is defined as

$$Q_\rho(x) := f(x) + \frac{\rho}{2}\|c(x)\|^2. \tag{3}$$

In this method, the stochastic gradient is computed using the recursive momentum scheme introduced in [13], treating ρ as part of the variables (see Section 2 for more detailed discussions). Under the error bound condition (5) with $\theta = 1$ and other suitable assumptions, this method achieves a sample and first-order operation complexity of $\widetilde{\mathcal{O}}(\epsilon^{-4})$ for finding an ϵ-stochastic stationary point satisfying (2).

In many applications such as energy systems [33], machine learning [9, 20], resource allocation [16], and transportation [27], all or some of the constraints in problem (1) are hard constraints representing imperative requirements. Consequently, any desirable approximate solution must (nearly) satisfy these constraints. As mentioned above, the ϵ-stochastic stationary point x found by existing methods [1, 3, 11, 12, 35, 37] satisfies $\mathbb{E}[\|c(x)\|] \leq \epsilon$, guaranteeing that $\|c(x)\| \leq \delta$ with probability at least $1 - \epsilon/\delta$ for any $\delta \geq \epsilon$. However, it is possible that $\|c(x)\|$ may still be excessively large, leading to significant constraint violations, which is undesirable in applications where practitioners require nearly exact constraint satisfaction.

To address the aforementioned issue, we propose single-loop variance-reduced stochastic first-order methods for solving problem (1), inspired by the framework of [1, Algorithm 2], but with a significantly different approach to constructing the stochastic gradient. Specifically, starting from any initial point $x_0 \in X$, we iteratively solve a sequence of quadratic penalty problems $\min_{x\in X} Q_{\rho_k}(x)$ by performing only a *single* stochastic gradient descent step

$$x_{k+1} = \Pi_X(x_k - \eta_k G_k),$$

where $\eta_k > 0$ is a step size, G_k is a variance-reduced estimator of $\nabla Q_{\rho_k}(x_k)$, and Π_X denotes the projection operator onto the set X. In our methods, G_k is constructed by handling the stochastic part $f(x)$ and the deterministic part $\rho_k\|c(x)\|^2/2$ of $Q_{\rho_k}(x)$ separately. More precisely, $\nabla f(x_k)$ is approximated by a stochastic estimator g_k, obtained via a truncated recursive momentum scheme or a truncated Polyak momentum scheme for $f(x)$, while $\nabla(\rho_k\|c(x)\|^2/2)|_{x=x_k}$ is computed exactly as $\rho_k\nabla c(x_k)c(x_k)$. Combining these two components gives $G_k = g_k + \rho_k\nabla c(x_k)c(x_k)$, which serves as a stochastic estimator of $\nabla Q_{\rho_k}(x_k)$ (see Algorithms 1 and 2 for details). Under the error bound condition (5) with $\theta \geq 1$ and other suitable assumptions, our methods achieve a sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-\max\{4,2\theta\}})$ for finding an ϵ-stochastic stationary point x that satisfies

$$\|c(x)\| \leq \epsilon, \qquad \mathbb{E}[\mathrm{dist}\,(0, \nabla f(x) + \nabla c(x)\lambda + \mathcal{N}_X(x))] \leq \epsilon \tag{4}$$

for some λ. This ϵ-stochastic stationary point nearly satisfies all the constraints with certainty and is stronger than the one found by existing methods. Furthermore, when $\theta = 1$, our methods enjoy the best-known sample complexity and first-order operation complexity, which is however achieved in [1] for finding a weaker ϵ-stochastic stationary point satisfying (4). In addition, for $\theta > 1$, our methods exhibit provable convergence rate, while the convergence of existing methods remains unknown.

The main contributions of our paper are summarized as follows.

- We propose single-loop variance-reduced stochastic first-order methods with a truncate recursive momentum or a truncated Polyak momentum for solving problem (1).
- We show that under the error bound condition (5) with $\theta \geq 1$ and other suitable assumptions, our proposed methods achieve a sample complexity and first-order operation complexity of $\widetilde{\mathcal{O}}(\epsilon^{-\max\{4,2\theta\}})$ for finding an ϵ-stochastic stationary point of problem (1) satisfying (4), which is stronger than the one found by existing methods.

To the best of our knowledge, this is the first work to develop methods with provable complexity guarantees for finding an approximate stochastic stationary point of problem (1) that nearly satisfies all constraints with *certainty*.

The rest of this paper is organized as follows. In Subsection 1.1, we introduce some notation, terminology, and assumption. In Sections 2 and 3, we propose stochastic first-order methods with a truncated recursive momentum or a truncated Polyak momentum for problem (1) and analyze their convergence. We provide the proof of the main results in Section 4. Finally, concluding remarks are given in Section 5.

1.1 Notation, terminology, and assumption

The following notation will be used throughout this paper. Let $\mathbb{R}_{>0}$ denote the set of positive real numbers, and $\mathbb{R}^n$ denote the Euclidean space of dimension n. The standard inner product and Euclidean norm are denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$, respectively. For any $r > 0$, let $\mathscr{B}(r)$ represent the Euclidean ball centered at the origin with radius r, that is, $\mathscr{B}(r) = \{x : \|x\| \leq r\}$. For any $t \in \mathbb{R}$, let $\lceil t \rceil$ denote the least integer greater than or equal to t.

A mapping ϕ is said to be L_ϕ-*Lipschitz continuous* on a set Ω if $\|\phi(x) - \phi(x')\| \leq L_\phi\|x - x'\|$ for all $x, x' \in \Omega$. Also, it is said to be $L_{\nabla\phi}$-*smooth* on Ω if $\|\nabla\phi(x) - \nabla\phi(x')\| \leq L_{\nabla\phi}\|x - x'\|$ for all $x, x' \in \Omega$, where $\nabla\phi$ denotes the transpose of the Jacobian of ϕ. Given a nonempty closed convex set

Ω, $\mathrm{dist}(x, \Omega)$ denotes the Euclidean distance from x to Ω, and $\Pi_\Omega(x)$ denotes the Euclidean projection of x onto Ω. In addition, the normal cone of Ω at any $x \in \Omega$ is denoted by $\mathcal{N}_\Omega(x)$. Finally, we use $\widetilde{\mathcal{O}}(\cdot)$ to denote the asymptotic upper bound that ignores logarithmic factors.

Throughout this paper, we make the following assumptions for problem (1).

Assumption 1. *(i) The optimal value f^* of problem* (1) *and $Q_1^* := \min_{x\in X}\{f(x) + \|c(x)\|^2/2\}$ are finite.*

(ii) f is differentiable and L_f-Lipschitz continuous on X.

(iii) For each $\xi \in \Xi$, $\tilde{f}(\cdot, \xi)$ is differentiable on X and satisfies the following conditions:

$$\mathbb{E}[\nabla \tilde{f}(x, \xi)] = \nabla f(x), \quad \mathbb{E}[\|\nabla \tilde{f}(x, \xi) - \nabla f(x)\|^2] \le \sigma^2 \qquad \forall x \in X$$

for some constant $\sigma \ge 0$.

(iv) The mapping c is L_c-Lipschitz continuous and $L_{\nabla c}$-smooth on X. Additionally, $\|c(x)\| \le C_c$ for all $x \in X$, and there exist constants $\gamma > 0$ and $\theta \ge 1$ such that

$$\mathrm{dist}\left(0, \nabla c(x) c(x) + \mathcal{N}_X(x)\right) \ge \gamma \|c(x)\|^\theta \qquad \forall x \in X. \tag{5}$$

In addition, for notational convenience, we define

$$L := L_c^2 + C_c L_{\nabla c}. \tag{6}$$

It follows from this and Assumption 1 that $\|c(x)\|^2/2$ is L-smooth on X, and

$$\|\nabla f(x)\| \le L_f, \quad \|\nabla c(x)\| \le L_c \qquad \forall x \in X. \tag{7}$$

Before ending this subsection, we make some remarks on Assumption 1.

Remark 1. *(i) The assumption on the finiteness of Q_1^* is generally weaker than the condition $\min_{x\in X} f(x) \ge 0$, which is imposed in related work such as [1]. Moreover, this assumption is quite mild. Specifically, since the optimal value f^* of* (1) *is finite and*

$$\lim_{\rho\to\infty} \min_{x\in X}\{f(x) + \rho\|c(x)\|^2/2\} = f^*,$$

there exists some $\underline{\rho} > 0$ such that $\min_{x\in X}\{f(x)+\rho\|c(x)\|^2/2\}$ and consequently $\min_{x\in X}\{\rho^{-1}f(x)+\|c(x)\|^2/2\}$ are finite for all $\rho \ge \underline{\rho}$. Therefore, if Assumption 1(i) does not hold, one can replace f with $\rho^{-1}f$ for some $\rho \ge \underline{\rho}$, ensuring the resulting problem (1) *satisfies Assumption 1(i).*

(ii) Assumption 1(iii) is standard and implies that $\nabla \tilde{f}(x, \xi)$ is an unbiased estimator of $\nabla f(x)$ with a bounded variance for all $x \in X$.

(iii) Assumption 1(iv) with $\theta = 1$ is commonly used in the literature to develop algorithms for optimization problems involving nonconvex functional constraints (e.g., see [1, 21, 22, 32]). In contrast, our assumption is more general, as it covers a broader range of $\theta \in [1, \infty)$. The error bound condition in Assumption 1(iv) plays a crucial role in designing algorithms that yield nearly feasible solutions to problem (1).

2 A stochastic first-order method with a truncated recursive momentum for problem (1)

In this section, we propose a stochastic first-order method with a truncated recursive momentum for solving problem (1), inspired by the framework of [1, Algorithm 2], but employing a significantly different approach to constructing the stochastic gradient. Moreover, the proposed method exhibits stronger convergence properties compared to existing methods (see Remark 2).

Specifically, starting from any initial point $x_0 \in X$, we approximately solve a sequence of quadratic penalty problems $\min_{x\in X} Q_{\rho_k}(x)$ by performing only a *single* stochastic gradient descent step $x_{k+1} = \Pi_X(x_k - \eta_k G_k)$, where ρ_k is a penalty parameter, $\eta_k > 0$ is a step size, G_k is a variance-reduced estimator of $\nabla Q_{\rho_k}(x_k)$, and Q_{ρ_k} is given in (3). Notice from (3) that $\nabla Q_{\rho_k}(x_k) = \nabla f(x_k) + \rho_k \nabla c(x_k)c(x_k)$. Based on this, we particularly choose $G_k = g_k + \rho_k \nabla c(x_k)c(x_k)$, where g_k is a variance-reduced estimator of $\nabla f(x_k)$, computed recursively as follows:

$$g_k = \Pi_{\mathscr{B}(L_f)}\big(\nabla \tilde{f}(x_k, \xi_k) + (1 - \alpha_{k-1})(g_{k-1} - \nabla \tilde{f}(x_{k-1}, \xi_k))\big) \tag{8}$$

for some $\alpha_{k-1} \in (0, 1]$ and a randomly drawn sample ξ_k. This scheme is a slight modification of the recursive momentum scheme introduced in [13], incorporating a truncation operation via the projection operator $\Pi_{\mathscr{B}(L_f)}$ to ensure the boundedness of $\{g_k\}$, which is crucial for the subsequent analysis. Interestingly, despite this truncation, the modified scheme preserves a variance-reduction property similar to the original scheme in [13] (see Lemma 4).

The proposed stochastic first-order method with a truncated recursive momentum for solving problem (1) is presented in Algorithm 1 below.

Algorithm 1 A stochastic first-order method with a truncated recursive momentum for problem (1)

Input: $x_1 \in X$, $\{\alpha_k\} \subset (0, 1]$, $\{\rho_k\}, \{\eta_k\} \subset \mathbb{R}_{>0}$, and L_f given in Assumption 1.

1: Sample ξ_1 and set $g_1 = \Pi_{\mathscr{B}(L_f)}(\nabla \tilde{f}(x_1, \xi_1))$.
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad G_k = g_k + \rho_k \nabla c(x_k)c(x_k)$.
4: $\quad x_{k+1} = \Pi_X(x_k - \eta_k G_k)$.
5: $\quad$ Sample ξ_{k+1} and set $g_{k+1} = \Pi_{\mathscr{B}(L_f)}\big(\nabla \tilde{f}(x_{k+1}, \xi_{k+1}) + (1 - \alpha_k)(g_k - \nabla \tilde{f}(x_k, \xi_{k+1}))\big)$.
6: **end for**

The parameters $\{\alpha_k\}$, $\{\rho_k\}$ and $\{\eta_k\}$ will be specified in Theorem 1 for Algorithm 1 to achieve a desirable convergence rate. While Algorithm 1 shares a similar framework with [1, Algorithm 2], the construction of the variance-reduced estimator G_k of $\nabla Q_{\rho_k}(x_k)$ differs significantly between the two algorithms. In particular, G_k in [1, Algorithm 2] is obtained by applying the recursive momentum scheme introduced in [13] to the entire function $Q_\rho(x)$, treating ρ as part of the variables, and it is given by

$$G_k = \widetilde{\nabla} Q_{\rho_k}(x_k, \xi_k) + (1 - \alpha_{k-1})(G_{k-1} - \widetilde{\nabla} Q_{\rho_{k-1}}(x_{k-1}, \xi_k)),$$

where $\widetilde{\nabla} Q_\rho(x, \xi) = \nabla \tilde{f}(x, \xi) + \rho \nabla c(x)c(x)$. In contrast, G_k in Algorithm 1 is constructed by handling the stochastic part $f(x)$ and the deterministic part $\rho_k\|c(x)\|^2/2$ of $Q_{\rho_k}(x)$ separately. Specifically, $\nabla f(x_k)$ is approximated by a stochastic estimator g_k, obtained via truncated recursive momentum scheme as given in (8), while $\nabla(\rho_k\|c(x)\|^2/2)|_{x=x_k}$ is computed exactly as $\rho_k \nabla c(x_k)c(x_k)$. Combining these two components gives $G_k = g_k + \rho_k \nabla c(x_k)c(x_k)$ for Algorithm 1.

Due to this significant difference in the choice of G_k, [1, Algorithm 2] and Algorithm 1 exhibit vastly different convergence properties. Specifically, under Assumptions 1 and 2 with $\theta = 1$, [1, Algorithm 2] generate a sequence $\{\tilde{x}_k\}$ satisfying

$$\mathbb{E}[\|c(\tilde{x}_{\iota_k})\|^2] = \widetilde{\mathcal{O}}(k^{-1/2}), \quad \mathbb{E}\left[\operatorname{dist}^2\left(0, \nabla f(\tilde{x}_{\iota_k}) + \nabla c(\tilde{x}_{\iota_k})\tilde{\lambda}_{\iota_k} + \mathcal{N}_X(\tilde{x}_{\iota_k})\right)\right] = \widetilde{\mathcal{O}}(k^{-1/2})$$

for some sequence $\{\tilde{\lambda}_k\}$. In contrast, Algorithm 1 generates a sequence $\{x_k\}$ that satisfies

$$\|c(x_{\iota_k})\|^2 = \widetilde{\mathcal{O}}(k^{-1/2}), \quad \mathbb{E}\left[\mathrm{dist}^2\left(0, \nabla f(x_{\iota_k}) + \nabla c(x_{\iota_k})\lambda_{\iota_k} + \mathcal{N}_X(x_{\iota_k})\right)\right] = \widetilde{\mathcal{O}}(k^{-1/2})$$

for some sequence $\{\lambda_k\}$, where ι_k is uniformly drawn from $\{\lceil k/2\rceil + 1, \ldots, k\}$ for $k \geq 2$ (see Theorem 1 and [1, Theorem 4.2]). Clearly, the sequence $\{x_k\}$ generated by Algorithm 1 exhibits a stronger convergence property, since $\|c(x_{\iota_k})\|^2 = \widetilde{\mathcal{O}}(k^{-1/2})$ implies $\mathbb{E}[\|c(x_{\iota_k})\|^2] = \widetilde{\mathcal{O}}(k^{-1/2})$, while the reverse implication generally does not hold. Moreover, under Assumptions 1 and 2 with $\theta > 1$, the convergence of [1, Algorithm 2] remains unknown, while the sequence $\{x_k\}$ generated by Algorithm 1 satisfies

$$\|c(x_{\iota_k})\|^2 = \widetilde{\mathcal{O}}(k^{-\nu}), \quad \mathbb{E}\left[\mathrm{dist}^2\left(0, \nabla f(x_{\iota_k}) + \nabla c(x_{\iota_k})\lambda_{\iota_k} + \mathcal{N}_X(x_{\iota_k})\right)\right] = \widetilde{\mathcal{O}}(k^{-\nu})$$

with $\nu = \min\{1/2, \theta^{-1}\}$ for some sequence $\{\lambda_k\}$.

Before presenting convergence results for Algorithm 1, we make the following assumption regarding the average smoothness condition for problem (1).

Assumption 2. *The function $\tilde{f}(x,\xi)$ satisfies the average smoothness condition:*

$$\mathbb{E}[\|\nabla \tilde{f}(u,\xi) - \nabla \tilde{f}(v,\xi)\|^2] \leq \bar{L}_{\nabla f}^2 \|u - v\|^2 \qquad \forall u, v \in X.$$

Assumption 2 is commonly imposed in the literature to design algorithms for solving problems of the form $\min_x \mathbb{E}[\tilde{f}(x,\xi)] + P(x)$, where P is either zero or a simple but possibly nonsmooth function (e.g., see [13, 14, 36, 38, 39]). It can be observed that Assumption 2 implies that ∇f is $\bar{L}_{\nabla f}$-smooth on X, that is,

$$\|\nabla f(u) - \nabla f(v)\| \leq \bar{L}_{\nabla f}\|u - v\| \qquad \forall u, v \in X. \tag{9}$$

However, the reverse implication does not hold in general (e.g., see [18]).

We are now ready to present the convergence results for Algorithm 1, with the proof deferred to Subsection 4.1. Specifically, we will present convergence rates for the following two quantities:

$$\|c(x_{\iota_k})\|^2 \ \text{ and } \ \mathbb{E}\left[\mathrm{dist}^2\left(0, \nabla f(x_{\iota_k}) + \rho_{\iota_k - 1}\nabla c(x_{\iota_k})c(x_{\iota_k}) + \mathcal{N}_X(x_{\iota_k})\right)\right], \tag{10}$$

where ι_k is uniformly drawn from $\{\lceil k/2\rceil + 1, \ldots, k\}$. These quantities measure the constraint violation and the expected stationarity violation at x_{ι_k}.

Theorem 1. *Suppose that Assumptions 1 and 2 hold, and $\{x_k\}$ is generated by Algorithm 1. Let L be defined in (6), L_f, $\bar{L}_{\nabla f}$, L_c, C_c, σ, γ, θ and Q_1^* be given in Assumptions 1 and 2, g_1 be given in Algorithm 1, and ι_k be the random variable uniformly generated from $\{\lceil k/2\rceil + 1, \ldots, k\}$ for $k \geq 2$. Then the following statements hold.*

(i) Suppose that $\theta \in [1, 2)$ and its actual value is known. Let ρ_k, η_k and α_k be chosen as

$$\rho_k = k^{\frac{\theta}{4}}, \quad \eta_k = k^{-\frac{1}{2}}, \quad \alpha_k = k^{-\frac{1}{2}}. \tag{11}$$

Then for all $k \geq 2\widetilde{K}_1$, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{dist}^2\left(0, \nabla f(x_{\iota_k}) + \rho_{\iota_k - 1}\nabla c(x_{\iota_k})c(x_{\iota_k}) + \mathcal{N}_X(x_{\iota_k})\right)\right] \\
&\leq \frac{51}{2(k-1)^{\frac{1}{2}}}\Big(f(x_1) + \frac{1}{2}\|c(x_1)\|^2 - Q_1^* + \|g_1 - \nabla f(x_1)\|^2 + \frac{\theta(6-\theta)C_1}{4(2-\theta)} + \frac{1}{2}\big(\widetilde{K}_1^{\frac{\theta}{4}} - 1\big)C_c^2 \\
&\quad + 3\sigma^2(1 + \log k) + (1 + \log \widetilde{K}_1)\big(\bar{L}_{\nabla f} + \widetilde{K}_1^{\frac{\theta}{4}} L + 12\bar{L}_{\nabla f}^2\big)\big(L_f^2 + C_c^2 L_c^2 \widetilde{K}_1^{\frac{\theta}{2}}\big)\Big), \\
&\|c(x_{\iota_k})\|^2 \leq 2\sqrt{2}C_1 k^{-\frac{1}{2}},
\end{aligned}$$

where

$$\widetilde{K}_1 = \left\lceil \max\left\{1, 64\bar{L}_{\nabla f}^2, (48\bar{L}_{\nabla f}^2)^2, (8L)^{\frac{4}{2-\theta}}, \left(2^{2-\frac{\theta}{2}}\gamma^{-2}\right)^{\frac{4}{4-\theta}}\right\}\right\rceil, \tag{12}$$

$$C_1 = \max\left\{1, \widetilde{K}_1^{1/2}C_c^2/2, 2^{2-\theta/2}L_f^2\gamma^{-2}\right\}. \tag{13}$$

(ii) Suppose that $\theta \geq 1$ and its actual value is unknown. Let ρ_k, η_k and α_k be chosen as

$$\rho_k = k^{\frac{1}{2}}, \quad \eta_k = k^{-\frac{1}{2}}/\log(k+2), \quad \alpha_k = k^{-\frac{1}{2}}. \tag{14}$$

Then for all $k \geq 2\widetilde{K}_2$, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{dist}^2\left(0, \nabla f(x_{\iota_k}) + \rho_{\iota_k - 1}\nabla c(x_{\iota_k})c(x_{\iota_k}) + \mathcal{N}_X(x_{\iota_k})\right)\right] \\
&\leq \frac{51\log(k+2)}{2(k-1)^{\frac{1}{2}}}\Big(f(x_1) + \frac{1}{2}\|c(x_1)\|^2 - Q_1^* + \|g_1 - \nabla f(x_1)\|^2 + \frac{1}{2}C_2 k^{\frac{1}{2}-\nu}(1+\log k) \\
&\quad + \frac{1}{2}\big(\widetilde{K}_2^{\frac{1}{2}} - 1\big)C_c^2 + 3\sigma^2(1+\log k) + (1+\log\widetilde{K}_2)\big(\bar{L}_{\nabla f} + \widetilde{K}_2^{\frac{1}{2}}L + 12\bar{L}_{\nabla f}^2\big)\big(L_f^2 + C_c^2 L_c^2 \widetilde{K}_2\big)\Big), \\
&\|c(x_{\iota_k})\|^2 \leq 2^{1+\nu}C_2 k^{-\nu},
\end{aligned}$$

where

$$\widetilde{K}_2 = \left\lceil \max\left\{64\bar{L}_{\nabla f}^2, (48\bar{L}_{\nabla f}^2)^2, e^{8L}, e^{2\theta}, \left(e^{-1}\gamma^{-2}2^{2-\theta/2}\log(e^{2\theta}+2)\right)^{2\theta}\right\}\right\rceil, \tag{15}$$

$$\nu = \min\{1/2, \theta^{-1}\}, \quad C_2 = \max\left\{1, \widetilde{K}_2^{\nu}C_c^2/2, 2^{2-\theta/2}L_f^2\gamma^{-2}\right\}. \tag{16}$$

Remark 2. *(i) As shown in Theorem 1, when $\theta \in [1,2)$ and its actual value is known, the choice of ρ_k, η_k, and α_k in* (11) *ensures an $\widetilde{\mathcal{O}}(k^{-1/2})$ convergence rate for the quantities in* (10)*. Additionally, if ρ_k, η_k, and α_k are chosen according to* (14)*, an $\widetilde{\mathcal{O}}(k^{-\min\{1/2,\theta^{-1}\}})$ convergence rate is guaranteed for the same quantities.*

(ii) For $\theta \in [1,2)$, the choices of ρ_k, η_k, and α_k provided in (11) *and* (14) *ensure the same order of convergence rates for the quantities in* (10)*, regardless of whether the actual value of θ is known. However, the constant $\widetilde{K}_1$ generally depends less on L compared to $\widetilde{K}_2$. Therefore, when $\theta \in [1,2)$ and its actual value is known, the parameters ρ_k, η_k, and α_k specified in* (11) *are typically the better choice.*

(iii) Under Assumptions 1 and 2 with $\theta = 1$, [1, Algorithm 2] can generate a sequence $\{\tilde{x}_k\}$ satisfying

$$\mathbb{E}[\|c(\tilde{x}_{\iota_k})\|^2] = \widetilde{\mathcal{O}}(k^{-1/2}), \quad \mathbb{E}\left[\mathrm{dist}^2\left(0, \nabla f(\tilde{x}_{\iota_k}) + \nabla c(\tilde{x}_{\iota_k})\tilde{\lambda}_{\iota_k} + \mathcal{N}_X(\tilde{x}_{\iota_k})\right)\right] = \widetilde{\mathcal{O}}(k^{-1/2})$$

for some sequence $\{\tilde{\lambda}_k\}$. In contrast, under the same assumptions, Algorithm 1 generates a sequence $\{x_k\}$ that satisfies

$$\|c(x_{\iota_k})\|^2 = \widetilde{\mathcal{O}}(k^{-1/2}), \quad \mathbb{E}\left[\mathrm{dist}^2\left(0, \nabla f(x_{\iota_k}) + \nabla c(x_{\iota_k})\lambda_{\iota_k} + \mathcal{N}_X(x_{\iota_k})\right)\right] = \widetilde{\mathcal{O}}(k^{-1/2})$$

with $\lambda_k = \rho_{k-1}c(x_k)$ for all $k \geq 2$. Clearly, the sequence $\{x_k\}$ generated by Algorithm 1 exhibits a stronger convergence property, since $\|c(x_{\iota_k})\|^2 = \widetilde{\mathcal{O}}(k^{-1/2})$ implies $\mathbb{E}[\|c(x_{\iota_k})\|^2] = \widetilde{\mathcal{O}}(k^{-1/2})$, while the reverse implication generally does not hold. Moreover, under Assumptions 1 and 2 with $\theta > 1$, the convergence of [1, Algorithm 2] remains unknown, while Algorithm 1 enjoys an $\widetilde{\mathcal{O}}(k^{-\min\{1/2,\theta^{-1}\}})$ convergence rate for the quantities in (10)*.*

(iv) To the best of our knowledge, no prior algorithm was developed for problem (1) *that guarantees constraint violations converge to zero with certainty at a provable rate.*

The following result is an immediate consequence of Theorem 1. It provides iteration complexity results for Algorithm 1 to find an ϵ-stochastic stationary point x_{ι_k} of problem (1) satisfying (17) below.

Corollary 1. *Suppose that Assumptions 1 and 2 hold, and $\{x_k\}$ is generated by Algorithm 1. Let θ be given in Assumption 1, and ι_k be the random variable uniformly generated from $\{\lceil k/2\rceil+1,\ldots,k\}$ for $k\geq 2$. Then the following statements hold.*

(i) Suppose that $\theta\in[1,2)$ and its actual value is known. Let ρ_k, η_k and α_k be chosen as in (11). Then for any $\epsilon>0$, there exists some $T_1=\widetilde{\mathcal{O}}(\epsilon^{-4})$ such that

$$\|c(x_{\iota_k})\|\leq\epsilon,\quad \mathbb{E}\left[\operatorname{dist}\left(0,\nabla f(x_{\iota_k})+\rho_{\iota_k-1}\nabla c(x_{\iota_k})c(x_{\iota_k})+\mathcal{N}_X(x_{\iota_k})\right)\right]\leq\epsilon \tag{17}$$

hold for all $k\geq T_1$.

(ii) Suppose that $\theta\geq 1$ and its actual value is unknown. Let ρ_k, η_k and α_k be chosen as in (14). Then for any $\epsilon>0$, there exists some $T_2=\widetilde{\mathcal{O}}\left(\epsilon^{-\max\{4,2\theta\}}\right)$ such that (17) holds for all $k\geq T_2$.

Since Algorithm 1 requires one sample, one gradient evaluation of c, and two gradient evaluations of $\tilde f$ per iteration, its sample complexity and first-order operation complexity[3] are of the same order as its iteration complexity. It follows from Corollary 1 that Algorithm 1 achieves a sample complexity and first-order operation complexity of $\widetilde{\mathcal{O}}\left(\epsilon^{-\max\{4,2\theta\}}\right)$ for finding an ϵ-stochastic stationary point x_{ι_k} for problem (1) that satisfies (17). To the best of our knowledge, no algorithm prior to our work achieved these results except in the case where $\theta=1$. In that case, [1, Algorithm 2] achieves a sample complexity and first-order operation complexity of $\widetilde{\mathcal{O}}(\epsilon^{-4})$ for finding an ϵ-stochastic stationary point $\tilde x_{\iota_k}$ for problem (1) that satisfies:

$$\mathbb{E}[\|c(\tilde x_{\iota_k})\|]\leq\epsilon,\quad \mathbb{E}\left[\operatorname{dist}\left(0,\nabla f(\tilde x_{\iota_k})+\nabla c(\tilde x_{\iota_k})\tilde\lambda_{\iota_k}+\mathcal{N}_X(\tilde x_{\iota_k})\right)\right]\leq\epsilon$$

for some sequence $\tilde\lambda_k$. Although this algorithm achieves the same order of complexity as Algorithm 1, the ϵ-stochastic stationary point it finds is weaker than that obtained by Algorithm 1, since $\mathbb{E}[\|c(\tilde x_{\iota_k})\|]\leq\epsilon$ does not imply $\|c(\tilde x_{\iota_k})\|\leq\epsilon$ in general, whereas the reverse implication always holds.

3 A stochastic first-order method with a truncated Polyak momentum for problem (1)

In this section, we propose a stochastic first-order method with a truncated Polyak momentum for solving problem (1). This method modifies Algorithm 1, with g_k being recursively generated using the following truncated Polyak momentum scheme:

$$g_k=\Pi_{\mathscr{B}(L_f)}\left(\alpha_{k-1}\nabla\tilde f(x_k,\xi_k)+(1-\alpha_{k-1})g_{k-1}\right)$$

for some $\alpha_{k-1}\in(0,1]$ and a randomly drawn sample ξ_k, where L_f is given in Assumption 1. This scheme is a slight modification of the well-known Polyak momentum scheme [17, 30, 40], incorporating a truncation operation via the projection operator $\Pi_{\mathscr{B}(L_f)}$ to ensure the boundedness of the sequence $\{g_k\}$. This boundedness is crucial for our subsequent analysis. Despite the truncation, the modified scheme preserves the variance-reduction property of the original Polyak momentum scheme (see Lemma 10).

The proposed stochastic first-order method with a truncated Polyak momentum is presented in Algorithm 2.

[3]Sample complexity and first-order operation complexity refer to the total number of samples and gradient evaluations of $\tilde f$ used throughout the algorithm, respectively.

Algorithm 2 A stochastic first-order method with a truncated Polyak momentum for (1)

Input: $x_1 \in X$, $\{\alpha_k\} \subset (0,1]$, and $\{\rho_k\}, \{\eta_k\} \subset \mathbb{R}_{>0}$, and L_f given in Assumption 1.

1: Sample ξ_1 and set $g_1 = \Pi_{\mathscr{B}(L_f)}\big(\nabla \tilde{f}(x_1, \xi_1)\big)$.
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad G_k = g_k + \rho_k \nabla c(x_k) c(x_k)$.
4: $\quad x_{k+1} = \Pi_X(x_k - \eta_k G_k)$.
5: $\quad$ Sample ξ_{k+1} and set $g_{k+1} = \Pi_{\mathscr{B}(L_f)}\big((1-\alpha_k) g_k + \alpha_k \nabla \tilde{f}(x_{k+1}, \xi_{k+1})\big)$.
6: **end for**

Algorithm 2 will be shown to achieve the same order of convergence rate as Algorithm 1, but under weaker assumptions (see Theorem 2). This result is somewhat surprising because when $c = 0$, Algorithms 1 and 2 reduce to special cases of [39, Algorithm 1] and [17, Algorithm 1], respectively, where Algorithm 1 achieves a better convergence rate. Additionally, if the set $\{\nabla \tilde{f}(x,\xi) : x \in X, \xi \in \Xi\}$ is bounded, the recursion of g_{k+1} in step 5 of Algorithm 2 can be replaced with $g_{k+1} = (1-\alpha_k) g_k + \alpha_k \nabla \tilde{f}(x_{k+1}, \xi_{k+1})$. This clearly guarantees the boundedness of g_k, and the resulting algorithm enjoys the same rate of convergence as Algorithm 2.

To present the convergence results for Algorithm 2, we make the following assumption regarding the Lipschitz smoothness condition for problem (1).

Assumption 3. *The function f is $L_{\nabla f}$-smooth on X, that is,*

$$\|\nabla f(u) - \nabla f(v)\| \le L_{\nabla f} \|u - v\| \qquad \forall u, v \in X.$$

As remarked in Section 2, the average smoothness condition implies the Lipschitz smoothness condition, but the reverse implication generally does not hold. Therefore, Assumption 3 is weaker than Assumption 2 in general.

We are now ready to present the convergence results for Algorithm 2, with the proof deferred to Subsection 4.2. Specifically, we will establish convergence rates for the quantities introduced in (10).

Theorem 2. *Suppose that Assumptions 1 and 3 hold, and $\{x_k\}$ is generated by Algorithm 2. Let L be defined in (6), L_f, $L_{\nabla f}$, L_c, C_c, σ, γ, θ and Q_1^* be given in Assumptions 1 and 3, g_1 be given in Algorithm 2, and ι_k be the random variable uniformly generated in $\{\lceil k/2 \rceil + 1, \ldots, k\}$ for $k \ge 2$. Then the following statements hold.*

(i) Suppose that $\theta \in [1,2)$ and its actual value is known. Let ρ_k, η_k and α_k be chosen as

$$\rho_k = k^{\frac{\theta}{4}}, \quad \eta_k = k^{-\frac{1}{2}}/\log(k+2), \quad \alpha_k = k^{-\frac{1}{2}}. \tag{18}$$

Then for all $k \ge 2\widetilde{K}_3$, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{dist}^2\left(0, \nabla f(x_{\iota_k}) + \rho_{\iota_k - 1} \nabla c(x_{\iota_k}) c(x_{\iota_k}) + \mathcal{N}_X(x_{\iota_k})\right)\right] \\
&\le \frac{51 \log(k+2)}{2(k-1)^{\frac{1}{2}}} \Bigg(f(x_1) + \frac{1}{2}\|c(x_1)\|^2 - Q_1^* + \|g_1 - \nabla f(x_1)\|^2 + \frac{\theta(6-\theta)}{4(2-\theta)} C_3 + \frac{1}{2}(\widetilde{K}_3^{\frac{\theta}{4}} - 1) C_c^2 \\
&\quad + \sigma^2(1 + \log k) + (1 + \log \widetilde{K}_3)\big(L_{\nabla f} + \widetilde{K}_3^{\frac{\theta}{4}} L + 2\widetilde{K}_3^{\frac{1}{2}} L_{\nabla f}^2\big)\big(L_f^2 + C_c^2 L_c^2 \widetilde{K}_3^{\frac{\theta}{2}}\big)\Bigg),
\end{aligned}$$

$$\|c(x_{\iota_k})\|^2 \le 2C_3 (k/2)^{-\frac{1}{2}},$$

where

$$\widetilde{K}_3 = \left\lceil \max\left\{ e^2, 64 L_{\nabla f}^2, e^{8L_{\nabla f}^2}, (8L)^{\frac{4}{2-\theta}}, \left(\frac{2^{2-\frac{\theta}{2}} \log(e^2+2)}{e\gamma^2}\right)^{\frac{4}{2-\theta}} \right\} \right\rceil, \tag{19}$$

$$C_3 = \max\left\{1, \widetilde{K}_3^{1/2} C_c^2/2, 2^{2-\theta/2} L_f^2 \gamma^{-2}\right\}. \tag{20}$$

(ii) Suppose that $\theta \geq 1$ and its actual value is unknown. Let ρ_k, η_k and α_k be chosen as

$$\rho_k = k^{\frac{1}{2}}, \quad \eta_k = k^{-\frac{1}{2}}/\log(k+2), \quad \alpha_k = k^{-\frac{1}{2}}. \tag{21}$$

Then for all $k \geq 2\widetilde{K}_4$, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{dist}^2\left(0, \nabla f(x_{\iota_k}) + \rho_{\iota_k - 1}\nabla c(x_{\iota_k})c(x_{\iota_k}) + \mathcal{N}_X(x_{\iota_k})\right)\right] \\
&\leq \frac{51\log(k+2)}{2(k-1)^{\frac{1}{2}}}\Bigg(f(x_1) + \frac{1}{2}\|c(x_1)\|^2 - Q_1^* + \|g_1 - \nabla f(x_1)\|^2 + \frac{1}{2}C_4 k^{\frac{1}{2}-\nu}(1+\log k) \\
&\quad + \frac{1}{2}\big(\widetilde{K}_4^{\frac{1}{2}} - 1\big)C_c^2 + \sigma^2(1+\log k) + (1+\log \widetilde{K}_4)\big(L_{\nabla f} + \widetilde{K}_4^{\frac{1}{2}} L + 2\widetilde{K}_4^{\frac{1}{2}} L_{\nabla f}^2\big)\big(L_f^2 + C_c^2 L_c^2 \widetilde{K}_4\big)\Bigg), \\
&\|c(x_{\iota_k})\|^2 \leq 2C_4(k/2)^{-\nu},
\end{aligned}$$

where

$$\widetilde{K}_4 = \left\lceil \max\left\{64L_{\nabla f}^2, e^{8L_{\nabla f}^2}, e^{8L}, e^{2\theta}, \left(\frac{2^{2-\frac{\theta}{2}}\log(e^{2\theta}+2)}{e\gamma^2}\right)^{2\theta}\right\}\right\rceil, \tag{22}$$

$$\nu = \min\{1/2, \theta^{-1}\}, \quad C_4 = \max\left\{1, \widetilde{K}_4^{\nu} C_c^2/2, 2^{2-\theta/2} L_f^2 \gamma^{-2}\right\}. \tag{23}$$

The following result is an immediate consequence of Theorem 2. It provides iteration complexity results for Algorithm 2 to find an ϵ-stochastic stationary point x_{ι_k} of problem (1) that satisfies (17).

Corollary 2. *Suppose that Assumptions 1 and 3 hold, and $\{x_k\}$ is generated by Algorithm 2. Let θ be given in Assumption 1, and ι_k be the random variable uniformly generated from $\{\lceil k/2\rceil + 1, \ldots, k\}$ for $k \geq 2$. Then the following statements hold.*

(i) Suppose that $\theta \in [1, 2)$ and its actual value is known. Let ρ_k, η_k and α_k be chosen as in (18). Then for any $\epsilon > 0$, there exists some $T_3 = \widetilde{\mathcal{O}}\left(\epsilon^{-4}\right)$ such that (17) holds for all $k \geq T_3$.

(ii) Suppose that $\theta \geq 1$ and its actual value is unknown. Let ρ_k, η_k and α_k be chosen as in (21). Then for any $\epsilon > 0$, there exists some $T_4 = \widetilde{\mathcal{O}}\left(\epsilon^{-\max\{4,2\theta\}}\right)$ such that (17) holds for all $k \geq T_4$.

Since Algorithm 2 requires one sample, one gradient evaluation of c, and one gradient evaluation of $\tilde{f}$ per iteration, its sample complexity and first-order operation complexity are of the same order as its iteration complexity. It follows from Corollary 2 that Algorithm 2 achieves both a sample complexity and a first-order operation complexity of $\widetilde{\mathcal{O}}\left(\epsilon^{-\max\{4,2\theta\}}\right)$ to find an ϵ-stochastic stationary point x_{ι_k} for problem (1) that satisfies (17). Although Algorithms 1 and 2 achieve the same order of complexity, Algorithm 2 operates under weaker assumptions, as Assumption 3 is less restrictive than Assumption 2. Additionally, Algorithm 2 requires only one gradient evaluation of $\tilde{f}$ per iteration, while Algorithm 1 requires two.

4 Proof of the main results

In this section we provide a proof of our main results presented in Sections 2 and 3, which are particularly Theorems 1 and 2.

4.1 Proof of the main result in Section 2

In this subsection we first establish several technical lemmas and then use them to prove Theorem 1.

For notational convenience, we define

$$h(x) := \frac{1}{2}\|c(x)\|^2. \tag{24}$$

One can observe from Assumption 1(iv) that h is L-smooth on X, where L is given in (6).

The following lemma establishes a relationship between $h(x_{k+1})$ and $h(x_k)$, which will be used to derive bounds for $\|c(x_k)\|^2$.

Lemma 1. *Suppose that Assumption 1 holds, and x_{k+1} is generated by Algorithm 1 for some $k \geq 1$ with $\rho_k\eta_k \leq (\sqrt{5}-1)/(2L)$. Then we have*

$$h(x_{k+1}) + 2^{\theta-2}\gamma^2\rho_k\eta_k[h(x_{k+1})]^\theta \leq h(x_k) + L_f^2\rho_k^{-1}\eta_k/2,$$

where ρ_k and η_k are given in Algorithm 1, L_f, γ and θ are given in Assumption 1, and L and h are defined in (6) *and* (24)*, respectively.*

Proof. Let G_k be given in Algorithm 1. For convenience, we define

$$\widetilde{G}_k = \rho_k^{-1}G_k, \quad \tilde{\eta}_k = \rho_k\eta_k. \tag{25}$$

It then follows from these, (24), and the expression of x_{k+1} in Algorithm 1 that

$$x_{k+1} = \Pi_X(x_k - \eta_k G_k) = \Pi_X(x_k - \tilde{\eta}_k\widetilde{G}_k), \tag{26}$$

which implies that

$$0 \in x_{k+1} - x_k + \tilde{\eta}_k\widetilde{G}_k + \mathcal{N}_X(x_{k+1}) \quad \Rightarrow \quad \nabla h(x_{k+1}) + \tilde{\eta}_k^{-1}(x_k - x_{k+1}) - \widetilde{G}_k \in \nabla h(x_{k+1}) + \mathcal{N}_X(x_{k+1}). \tag{27}$$

Using this, (24) and Assumption 1(iv), we have

$$\begin{aligned}
2^\theta\gamma^2[h(x_{k+1})]^\theta &= \gamma^2\|c(x_{k+1})\|^{2\theta} \leq \mathrm{dist}^2\left(0, \nabla c(x_{k+1})c(x_{k+1}) + \mathcal{N}_X(x_{k+1})\right) \\
&\overset{(24)}{=} \mathrm{dist}^2\left(0, \nabla h(x_{k+1}) + \mathcal{N}_X(x_{k+1})\right) \overset{(27)}{\leq} \|\nabla h(x_{k+1}) + \tilde{\eta}_k^{-1}(x_k - x_{k+1}) - \widetilde{G}_k\|^2 \\
&\leq 2\|\tilde{\eta}_k^{-1}(x_k - x_{k+1}) + \nabla h(x_k) - \widetilde{G}_k\|^2 + 2\|\nabla h(x_{k+1}) - \nabla h(x_k)\|^2 \\
&= 2\tilde{\eta}_k^{-2}\|x_{k+1} - x_k\|^2 + 4\tilde{\eta}_k^{-1}\langle \widetilde{G}_k - \nabla h(x_k), x_{k+1} - x_k\rangle + 2\|\widetilde{G}_k - \nabla h(x_k)\|^2 \\
&\quad + 2\|\nabla h(x_{k+1}) - \nabla h(x_k)\|^2 \\
&\leq 2(\tilde{\eta}_k^{-2} + L^2)\|x_{k+1} - x_k\|^2 + 4\tilde{\eta}_k^{-1}\langle \widetilde{G}_k - \nabla h(x_k), x_{k+1} - x_k\rangle + 2\|\widetilde{G}_k - \nabla h(x_k)\|^2,
\end{aligned} \tag{28}$$

where the first inequality follows from Assumption 1(iv), the second inequality is due to the convexity of $\|\cdot\|^2$, and the last inequality follows from the L-smoothness of h. In addition, by (26) and $x_k \in X$, one has

$$\langle x_{k+1} - x_k + \tilde{\eta}_k\widetilde{G}_k, x_k - x_{k+1}\rangle \geq 0 \quad \Rightarrow \quad \langle \widetilde{G}_k, x_{k+1} - x_k\rangle \leq -\tilde{\eta}_k^{-1}\|x_{k+1} - x_k\|^2.$$

This together with the L-smoothness of h yields

$$\begin{aligned}
h(x_{k+1}) &\leq h(x_k) + \langle \nabla h(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&= h(x_k) + \langle \widetilde{G}_k, x_{k+1} - x_k\rangle + \langle \nabla h(x_k) - \widetilde{G}_k, x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&\leq h(x_k) - \tilde{\eta}_k^{-1}\|x_{k+1} - x_k\|^2 + \langle \nabla h(x_k) - \widetilde{G}_k, x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2.
\end{aligned}$$

Using this and (28), we obtain that

$$h(x_{k+1})+2^{\theta-2}\gamma^2\tilde{\eta}_k[h(x_{k+1})]^\theta \le h(x_k)+\frac{1}{2}(L^2\tilde{\eta}_k-\tilde{\eta}_k^{-1}+L)\|x_{k+1}-x_k\|^2+\frac{\tilde{\eta}_k}{2}\|\widetilde{G}_k-\nabla h(x_k)\|^2. \tag{29}$$

Observe from (25) and $\rho_k\eta_k \le (\sqrt{5}-1)/(2L)$ that $L\tilde{\eta}_k = L\rho_k\eta_k \le (\sqrt{5}-1)/2$, which implies that

$$L^2\tilde{\eta}_k-\tilde{\eta}_k^{-1}+L=\tilde{\eta}_k^{-1}(L^2\tilde{\eta}_k^2+L\tilde{\eta}_k-1)\le 0. \tag{30}$$

Notice from the expression of g_k in Algorithm 1 that $g_k \in \mathscr{B}(L_f)$ and hence $\|g_k\| \le L_f$. Also, observe from Algorithm 1 and (24) that $G_k = g_k + \rho_k\nabla h(x_k)$. Using these and (25), we have

$$\|\widetilde{G}_k-\nabla h(x_k)\| = \|\rho_k^{-1}G_k-\nabla h(x_k)\| = \|\rho_k^{-1}(g_k+\rho_k\nabla h(x_k))-\nabla h(x_k)\| = \rho_k^{-1}\|g_k\| \le \rho_k^{-1}L_f.$$

It then follows from this, (29) and (30) that

$$h(x_{k+1})+2^{\theta-2}\gamma^2\tilde{\eta}_k[h(x_{k+1})]^\theta \le h(x_k)+L_f^2\tilde{\eta}_k/(2\rho_k^2).$$

This and the definition of $\tilde{\eta}_k$ in (25) imply that the conclusion of this lemma holds. □

The next two lemmas derive bounds for $\|c(x_k)\|^2$ under two different choices of ρ_k, η_k and α_k in Algorithm 1.

Lemma 2. *Let $\widetilde{K}_1$ and C_1 be given in* (12) *and* (13), *respectively. Suppose that Assumption 1 holds with $\theta \in [1,2)$ and $\{x_k\}$ is generated by Algorithm 1 with $\{\rho_k\}$, $\{\eta_k\}$ and $\{\alpha_k\}$ given in* (11). *Then we have $\|c(x_k)\|^2 \le 2C_1k^{-1/2}$ for all $k \ge \widetilde{K}_1$.*

Proof. Let h be defined in (24). To prove this lemma, it is equivalent to show that $h(x_k) \le C_1k^{-1/2}$ for all $k \ge \widetilde{K}_1$. We now prove this by induction. Indeed, notice from Algorithm 1 that $x_{\widetilde{K}_1} \in X$. It then follows from (13), (24) and Assumption 1(iv) that

$$h(x_{\widetilde{K}_1}) \overset{(24)}{=} \frac{1}{2}\|c(x_{\widetilde{K}_1})\|^2 \le \frac{1}{2}C_c^2 \overset{(13)}{\le} C_1\widetilde{K}_1^{-1/2}.$$

Hence, the conclusion holds for $k=\widetilde{K}_1$. Now, suppose for induction that $h(x_k) \le C_1k^{-1/2}$ holds for some $k \ge \widetilde{K}_1$. Recall that $\theta\in[1,2)$ and ρ_k, η_k and $\widetilde{K}_1$ are given in (11) and (12). In view of these, one can observe that

$$\rho_k\eta_k \overset{(11)}{=} k^{\frac{\theta-2}{4}} \le \widetilde{K}_1^{\frac{\theta-2}{4}} \overset{(12)}{\le} \frac{1}{8L} < \frac{\sqrt{5}-1}{2L},$$

and hence Lemma 1 holds for such k. Using Lemma 1 with the choice of ρ_k and η_k given in (11), we obtain that

$$h(x_{k+1})+2^{\theta-2}\gamma^2k^{\frac{\theta-2}{4}}[h(x_{k+1})]^\theta \le h(x_k)+L_f^2k^{-\frac{\theta+2}{4}}/2. \tag{31}$$

Further, let

$$\phi(t)=t+2^{\theta-2}\gamma^2k^{\frac{\theta-2}{4}}t^\theta. \tag{32}$$

Notice from (13) that $C_1\ge 1$. Using this and (32), we have

$$\begin{aligned}
&\phi(C_1(k+1)^{-1/2})-C_1k^{-1/2}-L_f^2k^{-\frac{\theta+2}{4}}/2\\
&\overset{(32)}{=} C_1^\theta 2^{\theta-2}\gamma^2k^{\frac{\theta-2}{4}}(k+1)^{-\frac{\theta}{2}}+C_1(k+1)^{-1/2}-C_1k^{-1/2}-L_f^2k^{-\frac{\theta+2}{4}}/2\\
&\ge C_1^\theta 2^{\theta-2}\gamma^2k^{\frac{\theta-2}{4}}(k+1)^{-\frac{\theta}{2}}-C_1k^{-\frac{3}{2}}/2-L_f^2k^{-\frac{\theta+2}{4}}/2\\
&= k^{-\frac{\theta+2}{4}}\left(C_1^\theta 2^{\theta-2}\gamma^2\left(\frac{k}{k+1}\right)^{\frac{\theta}{2}}-C_1k^{\frac{\theta-4}{4}}/2-L_f^2/2\right)\\
&\ge k^{-\frac{\theta+2}{4}}\left(C_1 2^{\frac{\theta}{2}-2}\gamma^2-C_1k^{\frac{\theta-4}{4}}/2-L_f^2/2\right),
\end{aligned} \tag{33}$$

where the first inequality follows from $(k+1)^{-1/2} - k^{-1/2} \geq -k^{-3/2}/2$ thanks to the convexity of $t^{-1/2}$, and the second inequality is due to $\theta \geq 1$, $C_1 \geq 1$ and $k/(k+1) \geq 1/2$. In addition, it follows from (12), $1 \leq \theta < 2$ and $k \geq \widetilde{K}_1$ that

$$k^{\frac{\theta-4}{4}} \leq \widetilde{K}_1^{\frac{\theta-4}{4}} \leq 2^{\frac{\theta}{2}-2}\gamma^2.$$

By this and (13), one has

$$C_1 2^{\frac{\theta}{2}-2}\gamma^2 - C_1 k^{\frac{\theta-4}{4}}/2 - L_f^2/2 \geq C_1 2^{\frac{\theta}{2}-2}\gamma^2/2 - L_f^2/2 \overset{(13)}{\geq} 0,$$

which together with (33) implies that

$$\phi(C_1(k+1)^{-1/2}) - C_1 k^{-1/2} - L_f^2 k^{-\frac{\theta+2}{4}}/2 \geq 0.$$

Using this, (31), (32) and the induction hypothesis that $h(x_k) \leq C_1 k^{-1/2}$, we obtain that

$$\phi(C_1(k+1)^{-1/2}) \geq C_1 k^{-1/2} + L_f^2 k^{-\frac{\theta+2}{4}}/2 \geq h(x_k) + L_f^2 k^{-\frac{\theta+2}{4}}/2 \overset{(31)(32)}{\geq} \phi(h(x_{k+1})).$$

It then follows from this inequality and the strict monotonicity of ϕ on $[0,\infty)$ that $h(x_{k+1}) \leq C_1(k+1)^{-1/2}$. Hence, the induction is completed and the conclusion of this lemma holds. □

Lemma 3. *Let $\widetilde{K}_2$, ν and C_2 be given in* (15) *and* (16)*, respectively. Suppose that Assumption 1 holds, and $\{x_k\}$ is generated by Algorithm 1 with $\{\rho_k\}$, $\{\eta_k\}$ and $\{\alpha_k\}$ given in* (14)*. Then we have $\|c(x_k)\|^2 \leq 2C_2 k^{-\nu}$ for all $k \geq \widetilde{K}_2$.*

Proof. Let h be defined in (24). To prove this lemma, it is equivalent to show that $h(x_k) \leq C_2 k^{-\nu}$ for all $k \geq \widetilde{K}_2$. We now prove this by induction. Indeed, notice from Algorithm 1 that $x_{\widetilde{K}_2} \in X$. It then follows from (16), (24) and Assumption 1(iv) that

$$h(x_{\widetilde{K}_2}) \overset{(24)}{=} \frac{1}{2}\|c(x_{\widetilde{K}_2})\|^2 \leq \frac{1}{2}C_c^2 \overset{(16)}{\leq} C_2 \widetilde{K}_2^{-\nu}.$$

Hence, the conclusion holds for $k = \widetilde{K}_2$. Now, suppose for induction that $h(x_k) \leq C_2 k^{-\nu}$ holds for some $k \geq \widetilde{K}_2$. Recall that ρ_k, η_k and $\widetilde{K}_2$ are given in (14) and (15). In view of these, one can observe that

$$\rho_k \eta_k \overset{(14)}{=} \frac{1}{\log(k+2)} \leq \frac{1}{\log(\widetilde{K}_2+2)} \overset{(15)}{\leq} \frac{1}{8L} \leq \frac{\sqrt{5}-1}{2L},$$

and hence Lemma 1 holds for such k. Using Lemma 1 with the choice of ρ_k and η_k given in (14), we obtain that

$$h(x_{k+1}) + 2^{\theta-2}\gamma^2[h(x_{k+1})]^\theta/\log(k+2) \leq h(x_k) + L_f^2 k^{-1}/(2\log(k+2)). \tag{34}$$

Further, let

$$\phi(t) = t + 2^{\theta-2}\gamma^2 t^\theta/\log(k+2). \tag{35}$$

Notice from (16) that $\nu = \min\{1/2, 1/\theta\}$ and $C_2 \geq 1$. Using these and (35), we have

$$\begin{aligned}
&\phi(C_2(k+1)^{-\nu}) - C_2 k^{-\nu} - L_f^2 k^{-1}/(2\log(k+2)) \\
&\overset{(35)}{=} C_2^\theta 2^{\theta-2}\gamma^2(k+1)^{-\theta\nu}/\log(k+2) + C_2(k+1)^{-\nu} - C_2k^{-\nu} - L_f^2 k^{-1}/(2\log(k+2)) \\
&\geq C_2^\theta 2^{\theta-2}\gamma^2(k+1)^{-\theta\nu}/\log(k+2) - \nu C_2 k^{-\nu-1} - L_f^2 k^{-1}/(2\log(k+2)) \\
&= \frac{k^{-\theta\nu}}{\log(k+2)}\left(C_2^\theta 2^{\theta-2}\gamma^2\left(\frac{k}{k+1}\right)^{\theta\nu} - \nu C_2 k^{(\theta-1)\nu-1}\log(k+2) - L_f^2 k^{\theta\nu-1}/2\right) \\
&\geq \frac{k^{-\theta\nu}}{\log(k+2)}\left(C_2^\theta 2^{\theta-\theta\nu-2}\gamma^2 - \nu C_2 k^{-\frac{1}{\theta}}\log(k+2) - L_f^2/2\right) \\
&\geq \frac{k^{-\theta\nu}}{\log(k+2)}\left(C_2 2^{\frac{\theta}{2}-2}\gamma^2 - C_2 k^{-\frac{1}{\theta}}\log(k+2)/2 - L_f^2/2\right),
\end{aligned} \tag{36}$$

where the first inequality follows from $(k+1)^{-\nu} - k^{-\nu} \geq -\nu k^{-\nu-1}$ thanks to the convexity of $t^{-\nu}$, the second inequality is due to $\theta \geq 1$, $\nu \leq 1/\theta$ and $k/(k+1) \geq 1/2$, and the last inequality follows from $\theta \geq 1$, $C_2 \geq 1$ and $\nu \leq 1/2$. In addition, one can verify that $t^{-\frac{1}{2\theta}}\log(t+2)$ is decreasing on $[e^{2\theta}, \infty)$. Using this, (15) and $k \geq \widetilde{K}_2 \geq e^{2\theta}$, we obtain that

$$k^{-\frac{1}{2\theta}}\log(k+2) \leq \log(e^{2\theta}+2)/e, \quad k^{-\frac{1}{2\theta}} \leq \widetilde{K}_2^{-\frac{1}{2\theta}} \leq e\gamma^2/(2^{2-\frac{\theta}{2}}\log(e^{2\theta}+2)).$$

Multiplying both sides of these inequalities yields $k^{-1/\theta}\log(k+2) \leq 2^{\theta/2-2}\gamma^2$, which together with (16) implies that

$$C_2 2^{\frac{\theta}{2}-2}\gamma^2 - C_2 k^{-\frac{1}{\theta}}\log(k+2)/2 - L_f^2/2 \geq C_2 2^{\frac{\theta}{2}-2}\gamma^2/2 - L_f^2/2 \overset{(16)}{\geq} 0.$$

Using this, (34), (35), (36), and the induction hypothesis that $h(x_k) \leq C_2 k^{-\nu}$, we obtain that

$$\phi(C_2(k+1)^{-\nu}) \geq C_2k^{-\nu} + L_f^2k^{-1}/(2\log(k+2)) \geq h(x_k) + L_f^2 k^{-1}/(2\log(k+2)) \overset{(34)(35)}{\geq} \phi(h(x_{k+1})).$$

It then follows from this inequality and the strict monotonicity of ϕ on $[0,\infty)$ that $h(x_{k+1}) \leq C_2(k+1)^{-\nu}$. Hence, the induction is completed and the conclusion of this lemma holds. □

The following lemma provides a relationship between $\mathbb{E}\left[\|g_{k+1} - \nabla f(x_{k+1})\|^2\right]$ and $\mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right]$.

Lemma 4. *Suppose that Assumptions 1 and 2 hold, and $\{g_k\}$ and $\{x_k\}$ are generated by Algorithm 1. Then for all $k \geq 1$, we have*

$$\mathbb{E}\left[\|g_{k+1} - \nabla f(x_{k+1})\|^2\right] \leq (1-\alpha_k)^2\mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right] + 6\bar{L}_{\nabla f}^2\mathbb{E}\left[\|x_{k+1} - x_k\|^2\right] + 3\sigma^2\alpha_k^2,$$

where $\{\alpha_k\}$ is given in Algorithm 1, and σ and $\bar{L}_{\nabla f}$ are given in Assumptions 1 and 2, respectively.

Proof. Notice from (7) that $\nabla f(x_{k+1}) \in \mathscr{B}(L_f)$ and hence $\nabla f(x_{k+1}) = \Pi_{\mathscr{B}(L_f)}(\nabla f(x_{k+1}))$. By this, the expression of g_{k+1}, and the nonexpansiveness of the projection operator $\Pi_{\mathscr{B}(L_f)}$, one has

$$\begin{aligned}
&\|g_{k+1} - \nabla f(x_{k+1})\|^2 = \|\Pi_{\mathscr{B}(L_f)}(\nabla\tilde{f}(x_{k+1},\xi_{k+1}) + (1-\alpha_k)(g_k - \nabla\tilde{f}(x_k,\xi_{k+1}))) - \Pi_{\mathscr{B}(L_f)}(\nabla f(x_{k+1}))\|^2 \\
&\leq \|\nabla\tilde{f}(x_{k+1},\xi_{k+1}) + (1-\alpha_k)(g_k - \nabla\tilde{f}(x_k,\xi_{k+1})) - \nabla f(x_{k+1})\|^2 \\
&= \|\nabla\tilde{f}(x_{k+1},\xi_{k+1}) - \nabla f(x_{k+1}) + (1-\alpha_k)(g_k - \nabla f(x_k) + \nabla f(x_k) - \nabla\tilde{f}(x_k,\xi_{k+1}))\|^2 \\
&= \|\nabla\tilde{f}(x_{k+1},\xi_{k+1}) - \nabla f(x_{k+1}) + (1-\alpha_k)(\nabla f(x_k) - \nabla\tilde{f}(x_k,\xi_{k+1}))\|^2 + (1-\alpha_k)^2\|g_k - \nabla f(x_k)\|^2 \\
&\quad + 2(1-\alpha_k)\langle g_k - \nabla f(x_k), \nabla\tilde{f}(x_{k+1},\xi_{k+1}) - \nabla f(x_{k+1})\rangle \\
&\quad + 2(1-\alpha_k)^2\langle g_k - \nabla f(x_k), \nabla f(x_k) - \nabla\tilde{f}(x_k,\xi_{k+1})\rangle.
\end{aligned} \tag{37}$$

Let $\Xi_k = \{\xi_1, \ldots, \xi_k\}$ denote the collection of samples drawn up to iteration $k-1$ in Algorithm 1. It then follows from Assumption 1(iii) that

$$\mathbb{E}[\nabla \tilde{f}(x_{k+1}, \xi_{k+1}) - \nabla f(x_{k+1})|\Xi_k] = 0, \quad \mathbb{E}[\nabla f(x_k) - \nabla \tilde{f}(x_k, \xi_{k+1})|\Xi_k] = 0,$$

which imply that

$$\mathbb{E}[\langle g_k - \nabla f(x_k), \nabla \tilde{f}(x_{k+1}, \xi_{k+1}) - \nabla f(x_{k+1})\rangle|\Xi_k] = \langle g_k - \nabla f(x_k), \mathbb{E}[\nabla \tilde{f}(x_{k+1}, \xi_{k+1}) - \nabla f(x_{k+1})|\Xi_k]\rangle = 0,$$
$$\mathbb{E}[\langle g_k - \nabla f(x_k), \nabla f(x_k) - \nabla \tilde{f}(x_k, \xi_{k+1})\rangle|\Xi_k] = \langle g_k - \nabla f(x_k), \mathbb{E}[\nabla f(x_k) - \nabla \tilde{f}(x_k, \xi_{k+1})|\Xi_k]\rangle = 0.$$

Using these and taking a conditional expectation on both sides of (37), we have

$$\begin{aligned}\mathbb{E}[\|g_{k+1} - \nabla f(x_{k+1})\|^2|\Xi_k] \le& \mathbb{E}[\|\nabla \tilde{f}(x_{k+1}, \xi_{k+1}) - \nabla f(x_{k+1}) + (1-\alpha_k)(\nabla f(x_k) - \nabla \tilde{f}(x_k, \xi_{k+1}))\|^2|\Xi_k] \\ &+ (1-\alpha_k)^2\|g_k - \nabla f(x_k)\|^2. \end{aligned} \tag{38}$$

In addition, it follows from Assumption 1 that

$$\begin{aligned}&\mathbb{E}[\|\nabla \tilde{f}(x_{k+1}, \xi_{k+1}) - \nabla f(x_{k+1}) + (1-\alpha_k)(\nabla f(x_k) - \nabla \tilde{f}(x_k, \xi_{k+1}))\|^2|\Xi_k] \\ &= \mathbb{E}[\|\nabla \tilde{f}(x_{k+1}, \xi_{k+1}) - \nabla \tilde{f}(x_k, \xi_{k+1}) + \nabla f(x_k) - \nabla f(x_{k+1}) - \alpha_k(\nabla f(x_k) - \nabla \tilde{f}(x_k, \xi_{k+1}))\|^2|\Xi_k] \\ &\le 3\mathbb{E}[\|\nabla \tilde{f}(x_{k+1}, \xi_{k+1}) - \nabla \tilde{f}(x_k, \xi_{k+1})\|^2|\Xi_k] + 3\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \\ &\quad + 3\alpha_k^2\mathbb{E}[\|\nabla f(x_k) - \nabla \tilde{f}(x_k, \xi_{k+1})\|^2|\Xi_k] \le 6\bar{L}_{\nabla f}^2\|x_{k+1} - x_k\|^2 + 3\sigma^2\alpha_k^2,\end{aligned}$$

where the first inequality follows from the convexity of $\|\cdot\|^2$, and the last inequality is due to (9) and Assumptions 1(iii) and 2. By this and (38), one has

$$\mathbb{E}[\|g_{k+1} - \nabla f(x_{k+1})\|^2|\Xi_k] \le (1-\alpha_k)^2\|g_k - \nabla f(x_k)\|^2 + 6\bar{L}_{\nabla f}^2\|x_{k+1} - x_k\|^2 + 3\sigma^2\alpha_k^2.$$

The conclusion of this lemma follows from taking expectation on both sides of this inequality. □

The next lemma provides an upper bound on $\mathbb{E}[Q_{\rho_k}(x_k) + \|g_k - \nabla f(x_k)\|^2]$.

Lemma 5. *Suppose that Assumptions 1 and 2 hold, and $\{g_k\}$ and $\{x_k\}$ are generated by Algorithm 1 with $\eta_k \le \alpha_k \le 1$. Then for all $k \ge 1$, we have*

$$\begin{aligned}\mathbb{E}\left[Q_{\rho_k}(x_k) + \|g_k - \nabla f(x_k)\|^2\right] \le& Q_{\rho_1}(x_1) + \|g_1 - \nabla f(x_1)\|^2 + \frac{1}{2}\sum_{i=1}^{k-1}(\rho_{i+1} - \rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] \\ &+ \frac{1}{2}\sum_{i=1}^{k-1}\left(\bar{L}_{\nabla f} + \rho_i L - \eta_i^{-1} + 12\bar{L}_{\nabla f}^2\right)\mathbb{E}\left[\|x_{i+1} - x_i\|^2\right] + 3\sigma^2\sum_{i=1}^{k-1}\alpha_i^2,\end{aligned} \tag{39}$$

where $\{\alpha_k\}$, $\{\rho_k\}$ and $\{\eta_k\}$ are given in Algorithm 1, Q_ρ and L are respectively defined in (3) *and* (6)*, and σ and $\bar{L}_{\nabla f}$ are given in Assumptions 1 and 2, respectively.*

Proof. Observe from Assumptions 1 and 2 and the definition of Q_ρ in (3) that Q_{ρ_k} is $(\bar{L}_{\nabla f} + \rho_k L)$-smooth on X, where L is defined in (6). Notice from Algorithm 1 that $x_k \in X$ and $x_{k+1} = \Pi_X(x_k - \eta_k G_k)$, which imply that

$$\langle x_{k+1} - x_k + \eta_k G_k, x_k - x_{k+1}\rangle \ge 0 \quad \Rightarrow \quad \langle G_k, x_{k+1} - x_k\rangle \le -\eta_k^{-1}\|x_{k+1} - x_k\|^2. \tag{40}$$

Also, notice from Algorithm 1 and (3) that

$$G_k = g_k + \rho_k\nabla c(x_k)c(x_k), \quad \nabla Q_{\rho_k}(x_k) = \nabla f(x_k) + \rho_k\nabla c(x_k)c(x_k),$$

and hence $\nabla Q_{\rho_k}(x_k) - G_k = g_k - \nabla f(x_k)$. In addition, by Young's inequality, one has

$$\langle \nabla Q_{\rho_k}(x_k) - G_k, x_{k+1} - x_k \rangle \le \frac{1}{2\eta_k}\|x_{k+1} - x_k\|^2 + \frac{\eta_k}{2}\|\nabla Q_{\rho_k}(x_k) - G_k\|^2 \tag{41}$$

Using the last two relations, (40), and the $(\bar{L}_{\nabla f} + \rho_k L)$-smoothness of Q_{ρ_k}, we obtain that

$$\begin{aligned}
Q_{\rho_k}(x_{k+1}) &\le Q_{\rho_k}(x_k) + \langle \nabla Q_{\rho_k}(x_k), x_{k+1} - x_k \rangle + \frac{1}{2}(\bar{L}_{\nabla f} + \rho_k L)\|x_{k+1} - x_k\|^2 \\
&= Q_{\rho_k}(x_k) + \langle G_k, x_{k+1} - x_k \rangle + \langle \nabla Q_{\rho_k}(x_k) - G_k, x_{k+1} - x_k \rangle + \frac{1}{2}(\bar{L}_{\nabla f} + \rho_k L)\|x_{k+1} - x_k\|^2 \\
&\overset{(41)}{\le} Q_{\rho_k}(x_k) + \langle G_k, x_{k+1} - x_k \rangle + \frac{1}{2}\left(\bar{L}_{\nabla f} + \rho_k L + \eta_k^{-1}\right)\|x_{k+1} - x_k\|^2 + \frac{\eta_k}{2}\|\nabla Q_{\rho_k}(x_k) - G_k\|^2 \\
&\le Q_{\rho_k}(x_k) + \frac{1}{2}\left(\bar{L}_{\nabla f} + \rho_k L - \eta_k^{-1}\right)\|x_{k+1} - x_k\|^2 + \frac{\eta_k}{2}\|g_k - \nabla f(x_k)\|^2,
\end{aligned}$$

where the first inequality is due to the $(\bar{L}_{\nabla f} + \rho_k L)$-smoothness of Q_{ρ_k}, and the last inequality follows from (40) and the relation $\nabla Q_{\rho_k}(x_k) - G_k = g_k - \nabla f(x_k)$. By this and (3), we further have

$$\begin{aligned}
Q_{\rho_{k+1}}(x_{k+1}) &\le Q_{\rho_k}(x_k) + \frac{1}{2}\left(\bar{L}_{\nabla f} + \rho_k L - \eta_k^{-1}\right)\|x_{k+1} - x_k\|^2 + \frac{\eta_k}{2}\|g_k - \nabla f(x_k)\|^2 \\
&\quad + Q_{\rho_{k+1}}(x_{k+1}) - Q_{\rho_k}(x_{k+1}) \\
&\overset{(3)}{=} Q_{\rho_k}(x_k) + \frac{1}{2}\left(\bar{L}_{\nabla f} + \rho_k L - \eta_k^{-1}\right)\|x_{k+1} - x_k\|^2 + \frac{\eta_k}{2}\|g_k - \nabla f(x_k)\|^2 + \frac{1}{2}(\rho_{k+1} - \rho_k)\|c(x_{k+1})\|^2.
\end{aligned} \tag{42}$$

Recall from the assumption that $0 < \eta_k \le \alpha_k \le 1$, which implies that $(1-\alpha_k)^2 + \eta_k \le 1 - \alpha_k + \eta_k \le 1$. Using this, taking expectation on both sides of (42), and summing the resulting inequality with the inequality in Lemma 4, we obtain that

$$\begin{aligned}
&\mathbb{E}\left[Q_{\rho_{k+1}}(x_{k+1}) + \|g_{k+1} - \nabla f(x_{k+1})\|^2\right] \\
&\le \mathbb{E}\left[Q_{\rho_k}(x_k) + \left((1-\alpha_k)^2 + \eta_k\right)\|g_k - \nabla f(x_k)\|^2\right] + \frac{1}{2}\left(\bar{L}_{\nabla f} + \rho_k L - \eta_k^{-1} + 12\bar{L}_{\nabla f}^2\right)\mathbb{E}\left[\|x_{k+1} - x_k\|^2\right] \\
&\quad - \frac{\eta_k}{2}\mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right] + \frac{1}{2}(\rho_{k+1} - \rho_k)\mathbb{E}\left[\|c(x_{k+1})\|^2\right] + 3\sigma^2\alpha_k^2 \\
&\le \mathbb{E}\left[Q_{\rho_k}(x_k) + \|g_k - \nabla f(x_k)\|^2\right] + \frac{1}{2}\left(\bar{L}_{\nabla f} + \rho_k L - \eta_k^{-1} + 12\bar{L}_{\nabla f}^2\right)\mathbb{E}\left[\|x_{k+1} - x_k\|^2\right] \\
&\quad - \frac{\eta_k}{2}\mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right] + \frac{1}{2}(\rho_{k+1} - \rho_k)\mathbb{E}\left[\|c(x_{k+1})\|^2\right] + 3\sigma^2\alpha_k^2.
\end{aligned} \tag{43}$$

The conclusion of this lemma follows by replacing k with i in the above inequalities and summing them up for all $1 \le i \le k-1$. $\square$

The following lemma provides an upper bound on $\mathrm{dist}^2\left(0, \nabla Q_{\rho_k}(x_{k+1}) + \mathcal{N}_X(x_{k+1})\right)$.

Lemma 6. *Suppose that Assumptions 1 and 2 hold, and $\{g_k\}$ and $\{x_k\}$ are generated by Algorithm 1. Then for all $k \ge 1$, we have*

$$\mathrm{dist}^2\left(0, \nabla Q_{\rho_k}(x_{k+1}) + \mathcal{N}_X(x_{k+1})\right) \le 3\left(\eta_k^{-2} + (\bar{L}_{\nabla f} + \rho_k L)^2\right)\|x_{k+1} - x_k\|^2 + 3\|g_k - \nabla f(x_k)\|^2, \tag{44}$$

where $\{\rho_k\}$ and $\{\eta_k\}$ are given in Algorithm 1, $\bar{L}_{\nabla f}$ is given in Assumption 2, and L and Q_ρ are defined in (6) *and* (3)*, respectively.*

Proof. By the expression of x_{k+1} in Algorithm 1, one has

$$0 \in x_{k+1} - x_k + \eta_k G_k + \mathcal{N}_X(x_{k+1}) \quad \Rightarrow \quad \eta_k^{-1}(x_k - x_{k+1}) - G_k \in \mathcal{N}_X(x_{k+1}). \tag{45}$$

Notice from the definition of Q_ρ in (3) that $\nabla Q_{\rho_k}(x) = \nabla f(x) + \rho_k \nabla c(x) c(x)$, which together with (6), (9) and Assumption 1(iv) implies that Q_{ρ_k} is $(\bar{L}_{\nabla f} + \rho_k L)$-smooth on X. Using (45), the expression of ∇Q_{ρ_k}, and $G_k = g_k + \rho_k \nabla c(x_k) c(x_k)$ (see Algorithm 1), we have

$$\eta_k^{-1}(x_k - x_{k+1}) + \nabla f(x_k) - g_k - \nabla Q_{\rho_k}(x_k) = \eta_k^{-1}(x_k - x_{k+1}) - g_k - \rho_k \nabla c(x_k) c(x_k) \in \mathcal{N}_X(x_{k+1}).$$

By this and the $(\bar{L}_{\nabla f} + \rho_k L)$-smoothness of Q_{ρ_k}, one has

$$\begin{aligned}
&\mathrm{dist}^2\left(0, \nabla Q_{\rho_k}(x_{k+1}) + \mathcal{N}_X(x_{k+1})\right) \leq \|\nabla Q_{\rho_k}(x_{k+1}) + (\eta_k^{-1}(x_k - x_{k+1}) + \nabla f(x_k) - g_k - \nabla Q_{\rho_k}(x_k))\|^2 \\
&\leq 3\left(\|\nabla Q_{\rho_k}(x_{k+1}) - \nabla Q_{\rho_k}(x_k)\|^2 + \eta_k^{-2}\|x_{k+1} - x_k\|^2 + \|g_k - \nabla f(x_k)\|^2\right) \\
&\leq 3\left(\eta_k^{-2} + (\bar{L}_{\nabla f} + \rho_k L)^2\right)\|x_{k+1} - x_k\|^2 + 3\|g_k - \nabla f(x_k)\|^2,
\end{aligned}$$

where the second inequality follows from the convexity of $\|\cdot\|^2$, and the last inequality is due to the $(\bar{L}_{\nabla f} + \rho_k L)$-smoothness of Q_{ρ_k}. Hence, the conclusion of this lemma holds. $\square$

We are now ready to prove the main result in Section 2, which is particularly Theorem 1.

Proof of Theorem 1. (i) It follows from (11), (12) and the assumption $1 \leq \theta < 2$ that for all $i \geq \widetilde{K}_1$,

$$\begin{aligned}
&\bar{L}_{\nabla f} + \rho_i L = \bar{L}_{\nabla f} i^{-\frac{1}{2}} \eta_i^{-1} + L i^{\frac{\theta-2}{4}} \eta_i^{-1} \leq \bar{L}_{\nabla f} \widetilde{K}_1^{-\frac{1}{2}} \eta_i^{-1} + L \widetilde{K}_1^{\frac{\theta-2}{4}} \eta_i^{-1} \leq \eta_i^{-1}/4, \\
&12\bar{L}_{\nabla f}^2 = 12\bar{L}_{\nabla f}^2 i^{-\frac{1}{2}} \eta_i^{-1} \leq 12\bar{L}_{\nabla f}^2 \widetilde{K}_1^{-\frac{1}{2}} \eta_i^{-1} \leq \eta_i^{-1}/4,
\end{aligned} \tag{46}$$

which imply that

$$\bar{L}_{\nabla f} + \rho_i L + 12\bar{L}_{\nabla f}^2 \leq \eta_i^{-1}/2 \qquad \forall i \geq \widetilde{K}_1. \tag{47}$$

In addition, observe from (11) that $\eta_k \leq \alpha_k \leq 1$ for all $k \geq 1$. It then follows from the proof of Lemma 5 that (43) holds. Using (47) and rearranging the terms of (43) with k replaced by i, we obtain that for all $i \geq \widetilde{K}_1$,

$$\begin{aligned}
&\frac{1}{4\eta_i}\mathbb{E}\left[\|x_{i+1} - x_i\|^2\right] + \frac{\eta_i}{2}\mathbb{E}\left[\|g_i - \nabla f(x_i)\|^2\right] \\
&\overset{(43)}{\leq} \mathbb{E}\left[Q_{\rho_i}(x_i) + \|g_i - \nabla f(x_i)\|^2\right] - \mathbb{E}\left[Q_{\rho_{i+1}}(x_{i+1}) + \|g_{i+1} - \nabla f(x_{i+1})\|^2\right] \\
&\quad + \frac{1}{2}\left(\bar{L}_{\nabla f} + \rho_i L + 12\bar{L}_{\nabla f}^2 - \eta_i^{-1}/2\right)\mathbb{E}\left[\|x_{i+1} - x_i\|^2\right] + \frac{1}{2}(\rho_{i+1} - \rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] + 3\sigma^2\alpha_i^2 \\
&\overset{(47)}{\leq} \mathbb{E}\left[Q_{\rho_i}(x_i) + \|g_i - \nabla f(x_i)\|^2\right] - \mathbb{E}\left[Q_{\rho_{i+1}}(x_{i+1}) + \|g_{i+1} - \nabla f(x_{i+1})\|^2\right] \\
&\quad + \frac{1}{2}(\rho_{i+1} - \rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] + 3\sigma^2\alpha_i^2.
\end{aligned} \tag{48}$$

Recall that ι_k is the random variable uniformly generated in $\{\lceil k/2 \rceil + 1, \ldots, k\}$. In addition, observe from (11) that $\eta_i^{-1} < \eta_{k-1}^{-1}$ for all $\lceil k/2 \rceil \leq i \leq k-1$. By these, (3), (11), (39), (44), (46) and (48), one has that for all $k \geq 2\widetilde{K}_1$,

$$
\begin{aligned}
&\mathbb{E}\left[\operatorname{dist}^2\left(0, \nabla f(x_{\iota_k}) + \rho_{\iota_k - 1}\nabla c(x_{\iota_k})c(x_{\iota_k}) + \mathcal{N}_X(x_{\iota_k})\right)\right] = \mathbb{E}\left[\operatorname{dist}^2\left(0, \nabla Q_{\rho_{\iota_k-1}}(x_{\iota_k}) + \mathcal{N}_X(x_{\iota_k})\right)\right] \\
&= \frac{1}{k - \lceil k/2 \rceil}\sum_{i=\lceil k/2\rceil}^{k-1} \mathbb{E}\left[\operatorname{dist}^2\left(0, \nabla Q_{\rho_i}(x_{i+1}) + \mathcal{N}_X(x_{i+1})\right)\right] \\
&\overset{(44)}{\leq} \frac{3}{k - \lceil k/2 \rceil}\sum_{i=\lceil k/2\rceil}^{k-1} \left(\left(\eta_i^{-2} + (\bar{L}_{\nabla f} + \rho_i L)^2\right)\mathbb{E}\left[\|x_{i+1} - x_i\|^2\right] + \mathbb{E}\left[\|g_i - \nabla f(x_i)\|^2\right]\right) \\
&\overset{(46)}{\leq} \frac{3}{k - \lceil k/2 \rceil}\sum_{i=\lceil k/2\rceil}^{k-1} \left((\eta_i^{-2} + \eta_i^{-2}/16)\mathbb{E}\left[\|x_{i+1} - x_i\|^2\right] + \mathbb{E}\left[\|g_i - \nabla f(x_i)\|^2\right]\right) \\
&\leq \frac{51}{8(k-1)}\sum_{i=\lceil k/2\rceil}^{k-1} \left(\eta_i^{-2}\mathbb{E}\left[\|x_{i+1} - x_i\|^2\right] + 2\mathbb{E}\left[\|g_i - \nabla f(x_i)\|^2\right]\right) \\
&\leq \frac{51}{2(k-1)\eta_{k-1}}\sum_{i=\lceil k/2\rceil}^{k-1} \left(\frac{1}{4\eta_i}\mathbb{E}\left[\|x_{i+1} - x_i\|^2\right] + \frac{\eta_i}{2}\mathbb{E}\left[\|g_i - \nabla f(x_i)\|^2\right]\right) \\
&\overset{(48)}{\leq} \frac{51}{2(k-1)\eta_{k-1}}\sum_{i=\lceil k/2\rceil}^{k-1} \Big(\mathbb{E}\left[Q_{\rho_i}(x_i) + \|g_i - \nabla f(x_i)\|^2\right] - \mathbb{E}\left[Q_{\rho_{i+1}}(x_{i+1}) + \|g_{i+1} - \nabla f(x_{i+1})\|^2\right] \\
&\qquad + \frac{1}{2}(\rho_{i+1} - \rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] + 3\sigma^2\alpha_i^2\Big) \\
&= \frac{51}{2(k-1)\eta_{k-1}}\Bigg(\mathbb{E}\left[Q_{\rho_{\lceil k/2\rceil}}(x_{\lceil k/2\rceil}) + \|g_{\lceil k/2\rceil} - \nabla f(x_{\lceil k/2\rceil})\|^2\right] - \mathbb{E}\left[Q_{\rho_k}(x_k) + \|g_k - \nabla f(x_k)\|^2\right] \\
&\qquad + \frac{1}{2}\sum_{i=\lceil k/2\rceil}^{k-1}(\rho_{i+1} - \rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] + 3\sigma^2\sum_{i=\lceil k/2\rceil}^{k-1}\alpha_i^2\Bigg) \\
&\leq \frac{51}{2(k-1)\eta_{k-1}}\Bigg(Q_1(x_1) - Q_1^* + \|g_1 - \nabla f(x_1)\|^2 + \frac{1}{2}\sum_{i=1}^{k-1}(\rho_{i+1} - \rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] + 3\sigma^2\sum_{i=1}^{k-1}\alpha_i^2 \\
&\qquad + \frac{1}{2}\sum_{i=1}^{\lceil k/2\rceil - 1}\left(\bar{L}_{\nabla f} + \rho_i L - \eta_i^{-1} + 12\bar{L}_{\nabla f}^2\right)\mathbb{E}\left[\|x_{i+1} - x_i\|^2\right]\Bigg),
\end{aligned}
\tag{49}
$$

where the first equality is due to (3), the first inequality follows from taking expectation on both sides of (44), the third inequality is due to the fact that $\lceil k/2 \rceil \leq (k+1)/2$, the fourth inequality follows from the relation $\eta_i^{-1} < \eta_{k-1}^{-1}$ for all $\lceil k/2 \rceil \leq i \leq k-1$, and the last inequality follows from (39) with k replaced by $\lceil k/2 \rceil$, $\rho_1 = 1$, and the fact that $Q_{\rho_k}(x_k) \geq Q_1(x_k) \geq Q_1^*$.

We next bound each summation term in (49). Indeed, it follows from (7), Assumption 1(iv), the nonexpansiveness of Π_X, and the expressions of x_{k+1}, g_k and G_k in Algorithm 1 that

$$
\begin{aligned}
\|x_{k+1} - x_k\|^2 &= \|\Pi_X(x_k - \eta_k G_k) - \Pi_X(x_k)\|^2 \leq \eta_k^2\|G_k\|^2 = \eta_k^2\|g_k + \rho_k \nabla c(x_k)c(x_k)\|^2 \\
&\leq 2\eta_k^2\left(\|g_k\|^2 + \rho_k^2\|\nabla c(x_k)c(x_k)\|^2\right) \leq 2\eta_k^2\left(L_f^2 + C_c^2 L_c^2 \rho_k^2\right).
\end{aligned}
\tag{50}
$$

Recall that $1 \leq \theta < 2$ and $\|c(x_i)\| \leq C_c$ for all i. Using these, (11), (47), (50), and Lemma 2, we have that for all $k \geq 2\widetilde{K}_1$,

$$
\sum_{i=1}^{\widetilde{K}_1 - 1}(\rho_{i+1} - \rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] \leq C_c^2 \sum_{i=1}^{\widetilde{K}_1 - 1}\left((i+1)^{\frac{\theta}{4}} - i^{\frac{\theta}{4}}\right) = C_c^2\left(\widetilde{K}_1^{\frac{\theta}{4}} - 1\right),
\tag{51}
$$

$$\sum_{i=\widetilde{K}_1}^{k-1}(\rho_{i+1}-\rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] \le 2C_1\sum_{i=\widetilde{K}_1}^{k-1}((i+1)^{\frac{\theta}{4}}-i^{\frac{\theta}{4}})(i+1)^{-\frac{1}{2}} \tag{52}$$

$$\le \frac{1}{2}C_1\theta\sum_{i=\widetilde{K}_1}^{k-1} i^{\frac{\theta-4}{4}}(i+1)^{-\frac{1}{2}} \le \frac{1}{2}C_1\theta\sum_{i=\widetilde{K}_1}^{k-1} i^{\frac{\theta-6}{4}} \le \frac{C_1\theta(6-\theta)}{2(2-\theta)}, \tag{53}$$

$$\sum_{i=1}^{k-1}\alpha_i^2 \overset{(11)}{=} \sum_{i=1}^{k-1} i^{-1} = 1+\sum_{i=2}^{k-1} i^{-1} \le 1+\int_1^{k-1} t^{-1}dt \le 1+\log k, \tag{54}$$

$$\begin{aligned}
&\sum_{i=1}^{\lceil k/2\rceil-1}\left(\bar{L}_{\nabla f}+\rho_i L-\eta_i^{-1}+12\bar{L}_{\nabla f}^2\right)\mathbb{E}\left[\|x_{i+1}-x_i\|^2\right]\\
&= \sum_{i=1}^{\widetilde{K}_1-1}\left(\bar{L}_{\nabla f}+\rho_i L-\eta_i^{-1}+12\bar{L}_{\nabla f}^2\right)\mathbb{E}\left[\|x_{i+1}-x_i\|^2\right] + \sum_{i=\widetilde{K}_1}^{\lceil k/2\rceil-1}\Big((\bar{L}_{\nabla f}+\rho_i L-\eta_i^{-1}+12\bar{L}_{\nabla f}^2)\\
&\qquad \times \mathbb{E}\left[\|x_{i+1}-x_i\|^2\right]\Big)\\
&\overset{(47)}{\le} \sum_{i=1}^{\widetilde{K}_1-1}\left(\bar{L}_{\nabla f}+\rho_i L-\eta_i^{-1}+12\bar{L}_{\nabla f}^2\right)\mathbb{E}\left[\|x_{i+1}-x_i\|^2\right]\\
&\overset{(50)}{\le} 2\sum_{i=1}^{\widetilde{K}_1-1}\left(\bar{L}_{\nabla f}+\rho_i L+12\bar{L}_{\nabla f}^2\right)\eta_i^2\left(L_f^2+C_c^2L_c^2\rho_i^2\right)\\
&\overset{(11)}{\le} 2(1+\log\widetilde{K}_1)\big(\bar{L}_{\nabla f}+\widetilde{K}_1^{\frac{\theta}{4}}L+12\bar{L}_{\nabla f}^2\big)\big(L_f^2+C_c^2L_c^2\widetilde{K}_1^{\frac{\theta}{2}}\big),
\end{aligned} \tag{55}$$

where the inequality in (51) follows from (11) and $\|c(x_i)\| \le C_c$ for all i, (52) is due to (11) and Lemma 2, the first inequality in (53) follows from $(i+1)^{\theta/4}-i^{\theta/4} \le \theta i^{(\theta-4)/4}/4$ for all $i\ge 1$ thanks to the concavity of $t^{\theta/4}$ with $1\le\theta<2$, the third inequality in (53) is due to

$$\sum_{i=\widetilde{K}_1}^{k-1} i^{(\theta-6)/4} \le \sum_{i=1}^{k-1} i^{(\theta-6)/4} = 1+\sum_{i=2}^{k-1} i^{(\theta-6)/4} \le 1+\int_1^{\infty} t^{(\theta-6)/4}dt = 1+\frac{4}{2-\theta},$$

and the last inequality in (55) follows from the relations $\rho_i \le \widetilde{K}_1^{\theta/4}$ for $1\le i\le \widetilde{K}_1-1$ and $\sum_{i=1}^{\widetilde{K}_1-1}\eta_i^2 = \sum_{i=1}^{\widetilde{K}_1-1} i^{-1} \le 1+\log\widetilde{K}_1$ due to the choice of ρ_i and η_i in (11).

Using (11), (49), (51), (52), (54) and (55), we have

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{dist}^2\left(0,\nabla f(x_{\iota_k})+\rho_{\iota_k-1}\nabla c(x_{\iota_k})c(x_{\iota_k})+\mathcal{N}_X(x_{\iota_k})\right)\right]\\
&\le \frac{51}{2(k-1)^{\frac{1}{2}}}\Bigg(Q_1(x_1)-Q_1^*+\|g_1-\nabla f(x_1)\|^2+\frac{C_1\theta(6-\theta)}{4(2-\theta)}+\frac{1}{2}C_c^2\big(\widetilde{K}_1^{\frac{\theta}{4}}-1\big)+3\sigma^2(1+\log k)\\
&\quad +(1+\log\widetilde{K}_1)\big(\bar{L}_{\nabla f}+\widetilde{K}_1^{\frac{\theta}{4}}L+12\bar{L}_{\nabla f}^2\big)\big(L_f^2+C_c^2L_c^2\widetilde{K}_1^{\frac{\theta}{2}}\big)\Bigg) \qquad \forall k\ge 2\widetilde{K}_1.
\end{aligned}$$

By this, (3), $\iota_k > \lceil k/2\rceil \ge \widetilde{K}_1$ for all $k\ge 2\widetilde{K}_1$, and Lemma 2 with k replaced by ι_k, one can see that statement (i) of Theorem 1 holds.

(ii) It follows from (14) and (15) that for all $i \geq \widetilde{K}_2$,

$$\begin{aligned}\bar{L}_{\nabla f} + \rho_i L &= \bar{L}_{\nabla f} i^{-\frac{1}{2}} \eta_i^{-1}/\log(i+2) + L\eta_i^{-1}/\log(i+2) \\ &\leq \bar{L}_{\nabla f} \widetilde{K}_2^{-\frac{1}{2}} \eta_i^{-1} + L\eta_i^{-1}/\log(\widetilde{K}_2+2) \leq \eta_i^{-1}/4, \\ 12\bar{L}_{\nabla f}^2 &= 12\bar{L}_{\nabla f}^2 i^{-\frac{1}{2}} \eta_i^{-1}/\log(i+2) \leq 12\bar{L}_{\nabla f}^2 \widetilde{K}_2^{-\frac{1}{2}} \eta_i^{-1} \leq \eta_i^{-1}/4,\end{aligned}$$

which imply that

$$\bar{L}_{\nabla f} + \rho_i L + 12\bar{L}_{\nabla f}^2 \leq \eta_i^{-1}/2 \qquad \forall i \geq \widetilde{K}_2. \tag{56}$$

By this, Lemma 3, (14), (50), $\nu \leq 1/2$, and $\|c(x_i)\| \leq C_c$ for all i, one has that for all $k \geq 2\widetilde{K}_2$,

$$\sum_{i=1}^{\widetilde{K}_2-1} (\rho_{i+1} - \rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] \leq C_c^2 \sum_{i=1}^{\widetilde{K}_2-1} \left((i+1)^{\frac{1}{2}} - i^{\frac{1}{2}}\right) = C_c^2\big(\widetilde{K}_2^{\frac{1}{2}} - 1\big), \tag{57}$$

$$\sum_{i=\widetilde{K}_2}^{k-1} (\rho_{i+1} - \rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] \leq 2C_2 \sum_{i=\widetilde{K}_2}^{k-1} ((i+1)^{\frac{1}{2}} - i^{\frac{1}{2}})(i+1)^{-\nu} \tag{58}$$

$$\leq C_2 \sum_{i=\widetilde{K}_2}^{k-1} i^{-\frac{1}{2}}(i+1)^{-\nu} \leq C_2 \sum_{i=\widetilde{K}_2}^{k-1} i^{-1}(i+1)^{\frac{1}{2}-\nu} \leq C_2 k^{\frac{1}{2}-\nu} \sum_{i=\widetilde{K}_2}^{k-1} i^{-1} \leq C_2 k^{\frac{1}{2}-\nu}(1+\log k), \tag{59}$$

$$\begin{aligned}&\sum_{i=1}^{\lceil k/2\rceil-1} \left(\bar{L}_{\nabla f} + \rho_i L - \eta_i^{-1} + 12\bar{L}_{\nabla f}^2\right) \mathbb{E}\left[\|x_{i+1} - x_i\|^2\right] \\ &= \sum_{i=1}^{\widetilde{K}_2-1} \left(\bar{L}_{\nabla f} + \rho_i L - \eta_i^{-1} + 12\bar{L}_{\nabla f}^2\right) \mathbb{E}\left[\|x_{i+1} - x_i\|^2\right] + \sum_{i=\widetilde{K}_2}^{\lceil k/2\rceil-1} \Big(\left(\bar{L}_{\nabla f} + \rho_i L - \eta_i^{-1} + 12\bar{L}_{\nabla f}^2\right) \\ &\qquad \times \mathbb{E}\left[\|x_{i+1} - x_i\|^2\right]\Big) \\ &\overset{(56)}{\leq} \sum_{i=1}^{\widetilde{K}_2-1} \left(\bar{L}_{\nabla f} + \rho_i L - \eta_i^{-1} + 12\bar{L}_{\nabla f}^2\right) \mathbb{E}\left[\|x_{i+1} - x_i\|^2\right] \\ &\overset{(50)}{\leq} 2\sum_{i=1}^{\widetilde{K}_2-1} \left(\bar{L}_{\nabla f} + \rho_i L + 12\bar{L}_{\nabla f}^2\right) \eta_i^2 \left(L_f^2 + C_c^2 L_c^2 \rho_i^2\right) \\ &\overset{(14)}{\leq} (1+\log \widetilde{K}_2)\big(\bar{L}_{\nabla f} + \widetilde{K}_2^{\frac{1}{2}} L + 12\bar{L}_{\nabla f}^2\big)(L_f^2 + C_c^2 L_c^2 \widetilde{K}_2),\end{aligned} \tag{60}$$

where the inequality in (57) follows from (14) and $\|c(x_i)\| \leq C_c$ for all i, (58) is due to (14) and Lemma 3, the first inequality in (59) follows from $(i+1)^{1/2} - i^{1/2} \leq i^{-1/2}/2$ for all $i \geq 1$ thanks to the concavity of $t^{1/2}$, the third inequality in (59) is due to $\nu \leq 1/2$, the last inequality in (59) follows from $\sum_{i=\widetilde{K}_2}^{k-1} i^{-1} \leq 1 + \log k$, and the last inequality in (60) is due to the relations $\rho_i \leq \widetilde{K}_2^{1/2}$ for $1 \leq i \leq \widetilde{K}_2 - 1$ and $\sum_{i=1}^{\widetilde{K}_2-1} \eta_i^2 \leq \sum_{i=1}^{\widetilde{K}_2-1} i^{-1} \leq 1 + \log \widetilde{K}_2$ thanks to the choice of ρ_i and η_i in (14).

In addition, observe from (14) that $\eta_k \leq \alpha_k \leq 1$ for all $k \geq 1$. Using this, (56), and similar arguments as in the proof of statement (i) of this theorem, we can see that (49) holds for all $k \geq 2\widetilde{K}_2$. Also, by (14) and (54), one has $\sum_{i=1}^{k-1} \alpha_i^2 \leq 1 + \log k$ for all $k \geq 1$. Using this, (14), (49), (57), (59) and (60), we have

$$\begin{aligned}&\mathbb{E}\left[\mathrm{dist}^2\left(0, \nabla f(x_{\iota_k}) + \rho_{\iota_k-1}\nabla c(x_{\iota_k})c(x_{\iota_k}) + \mathcal{N}_X(x_{\iota_k})\right)\right] \\ &\leq \frac{51\log(k+2)}{2(k-1)^{\frac{1}{2}}}\Bigg(Q_1(x_1) - Q_1^* + \|g_1 - \nabla f(x_1)\|^2 + \frac{1}{2}C_2 k^{\frac{1}{2}-\nu}(1+\log k) + \frac{1}{2}C_c^2\big(\widetilde{K}_2^{\frac{1}{2}} - 1\big)\end{aligned}$$

$$+ 3\sigma^2(1+\log k) + (1+\log \widetilde{K}_2)\big(\bar{L}_{\nabla f} + \widetilde{K}_2^{\frac{1}{2}} L + 12\bar{L}_{\nabla f}^2\big)\big(L_f^2 + C_c^2 L_c^2 \widetilde{K}_2\big)\Bigg) \qquad \forall k \geq 2\widetilde{K}_2.$$

By this, (3), $\iota_k > \lceil k/2 \rceil \geq \widetilde{K}_2$ for all $k \geq 2\widetilde{K}_2$, and Lemma 3 with k replaced by ι_k, one can see that statement (ii) of Theorem 1 holds. □

4.2 Proof of the main result in Section 3

In this subsection we first establish several technical lemmas and then use them to prove Theorem 2.

The following lemma establishes a relationship between $h(x_{k+1})$ and $h(x_k)$, which will be used to derive bounds for $\|c(x_k)\|^2$, where h is defined in (24).

Lemma 7. *Suppose that Assumption 1 holds, and x_{k+1} is generated by Algorithm 2 for some $k \geq 1$ with $\rho_k\eta_k \leq (\sqrt{5}-1)/(2L)$. Then we have*

$$h(x_{k+1}) + 2^{\theta-2}\gamma^2\rho_k\eta_k[h(x_{k+1})]^\theta \leq h(x_k) + L_f^2\rho_k^{-1}\eta_k/2,$$

where ρ_k and η_k are given in Algorithm 2, L_f, γ and θ are given in Assumption 1, and L and h are defined in (6) *and* (24)*, respectively.*

Proof. The proof of this lemma follows from similar arguments as in the proof of Lemma 1. □

The next two lemmas derive bounds for $\|c(x_k)\|^2$ under two different choices of ρ_k, η_k and α_k in Algorithm 2.

Lemma 8. *Let $\widetilde{K}_3$ and C_3 be given in* (19) *and* (20)*, respectively. Suppose that Assumption 1 holds with $\theta \in [1,2)$ and $\{x_k\}$ is generated by Algorithm 2 with $\{\rho_k\}$, $\{\eta_k\}$ and $\{\alpha_k\}$ given in* (18)*. Then we have $\|c(x_k)\|^2 \leq 2C_3k^{-1/2}$ for all $k \geq \widetilde{K}_3$.*

Proof. Let h be defined in (24). To prove this lemma, it is equivalent to show that $h(x_k) \leq C_3k^{-1/2}$ for all $k \geq \widetilde{K}_3$. We now prove this by induction. Indeed, notice from Algorithm 2 that $x_{\widetilde{K}_3} \in X$. It then follows from (20), (24) and Assumption 1(iv) that

$$h(x_{\widetilde{K}_3}) \stackrel{(24)}{=} \frac{1}{2}\|c(x_{\widetilde{K}_3})\|^2 \leq \frac{1}{2}C_c^2 \stackrel{(20)}{\leq} C_3\widetilde{K}_3^{-1/2}.$$

Hence, the conclusion holds for $k = \widetilde{K}_3$. Now, suppose for induction that $h(x_k) \leq C_3k^{-1/2}$ holds for some $k \geq \widetilde{K}_3$. Recall that $\theta \in [1,2)$ and ρ_k, η_k and $\widetilde{K}_3$ are given in (18) and (19). In view of these, one can observe that

$$\rho_k\eta_k \stackrel{(18)}{=} \frac{k^{\frac{\theta-2}{4}}}{\log(k+2)} \leq k^{\frac{\theta-2}{4}} \leq \widetilde{K}_3^{\frac{\theta-2}{4}} \stackrel{(19)}{\leq} \frac{1}{8L} \leq \frac{\sqrt{5}-1}{2L},$$

and hence Lemma 7 holds for such k. Using Lemma 7 with the choice of ρ_k and η_k given in (18), we obtain that

$$h(x_{k+1}) + 2^{\theta-2}\gamma^2k^{\frac{\theta-2}{4}}[h(x_{k+1})]^\theta/\log(k+2) \leq h(x_k) + L_f^2k^{-\frac{\theta+2}{4}}/(2\log(k+2)). \tag{61}$$

Further, let

$$\phi(t) = t + 2^{\theta-2}\gamma^2k^{\frac{\theta-2}{4}}t^\theta/\log(k+2). \tag{62}$$

Notice from (20) that $C_3 \geq 1$. Using this and (62), we have

$$
\begin{aligned}
&\phi(C_3(k+1)^{-\frac{1}{2}}) - C_3 k^{-\frac{1}{2}} - L_f^2 k^{-\frac{\theta+2}{4}}/(2\log(k+2)) \\
&\overset{(62)}{=} C_3^\theta 2^{\theta-2}\gamma^2 k^{\frac{\theta-2}{4}}(k+1)^{-\frac{\theta}{2}}/\log(k+2) + C_3(k+1)^{-\frac{1}{2}} - C_3 k^{-\frac{1}{2}} - L_f^2 k^{-\frac{\theta+2}{4}}/(2\log(k+2)) \\
&\geq C_3^\theta 2^{\theta-2}\gamma^2 k^{\frac{\theta-2}{4}}(k+1)^{-\frac{\theta}{2}}/\log(k+2) - C_3 k^{-\frac{3}{2}}/2 - L_f^2 k^{-\frac{\theta+2}{4}}/(2\log(k+2)) \\
&= \frac{k^{-\frac{\theta+2}{4}}}{\log(k+2)}\left(C_3^\theta 2^{\theta-2}\gamma^2\left(\frac{k}{k+1}\right)^{\frac{\theta}{2}} - C_3 k^{\frac{\theta-4}{4}}\log(k+2)/2 - L_f^2/2\right) \\
&\geq \frac{k^{-\frac{\theta+2}{4}}}{\log(k+2)}\left(C_3 2^{\frac{\theta}{2}-2}\gamma^2 - C_3 k^{\frac{\theta-4}{4}}\log(k+2)/2 - L_f^2/2\right). \qquad (63)
\end{aligned}
$$

where the first inequality follows from $(k+1)^{-1/2} - k^{-1/2} \geq -k^{-3/2}/2$ thanks to the convexity of $t^{-1/2}$, and the second inequality is due to $\theta \geq 1$, $C_3 \geq 1$ and $k/(k+1) \geq 1/2$. In addition, one can verify that $t^{-1/2}\log(t+2)$ is decreasing on $[e^2,\infty)$. Using this, (19), $1 \leq \theta < 2$ and $k \geq \widetilde{K}_3 \geq e^2$, we obtain that

$$
k^{-\frac{1}{2}}\log(k+2) \leq \log(e^2+2)/e, \quad k^{\frac{\theta-2}{4}} \leq \widetilde{K}_3^{\frac{\theta-2}{4}} \overset{(19)}{\leq} e\gamma^2/(2^{2-\frac{\theta}{2}}\log(e^2+2)).
$$

Multiplying both sides of these two inequalities yields $k^{\frac{\theta-4}{4}}\log(k+2) \leq 2^{\frac{\theta}{2}-2}\gamma^2$, which together with (20) implies that

$$
C_3 2^{\frac{\theta}{2}-2}\gamma^2 - C_3 k^{\frac{\theta-4}{4}}\log(k+2)/2 - L_f^2/2 \geq C_3 2^{\frac{\theta}{2}-2}\gamma^2/2 - L_f^2/2 \overset{(20)}{\geq} 0.
$$

Using this, (61), (62), (63), and the induction hypothesis that $h(x_k) \leq C_3 k^{-1/2}$, we obtain that

$$
\phi(C_3(k+1)^{-1/2}) \geq C_3 k^{-1/2} + \frac{L_f^2 k^{-\frac{\theta+2}{4}}}{2\log(k+2)} \geq h(x_k) + \frac{L_f^2 k^{-\frac{\theta+2}{4}}}{2\log(k+2)} \overset{(61)(62)}{\geq} \phi(h(x_{k+1})).
$$

It then follows from this inequality and the strict monotonicity of ϕ on $[0,\infty)$ that $h(x_{k+1}) \leq C_3(k+1)^{-1/2}$. Hence, the induction is completed and the conclusion of this lemma holds. $\square$

Lemma 9. *Let $\widetilde{K}_4$, ν and C_4 be given in* (22) *and* (23)*, respectively. Suppose that Assumption 1 holds, and $\{x_k\}$ is generated by Algorithm 2 with $\{\rho_k\}$, $\{\eta_k\}$ and $\{\alpha_k\}$ given in* (21)*. Then we have $\|c(x_k)\|^2 \leq 2C_4 k^{-\nu}$ for all $k \geq \widetilde{K}_4$.*

Proof. The proof of this lemma follows from similar arguments as in the proof of Lemma 3 with $\widetilde{K}_2$ and C_2 replaced with $\widetilde{K}_4$ and C_4, respectively. $\square$

The following lemma provides a relationship between $\mathbb{E}\left[\|g_{k+1} - \nabla f(x_{k+1})\|^2\right]$ and $\mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right]$.

Lemma 10. *Suppose that Assumption 1 and 3 hold, and $\{g_k\}$ and $\{x_k\}$ are generated by Algorithm 2. Then for all $k \geq 1$, we have*

$$
\mathbb{E}\left[\|g_{k+1} - \nabla f(x_{k+1})\|^2\right] \leq (1-\alpha_k)\mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right] + L_{\nabla f}^2\alpha_k^{-1}\mathbb{E}\left[\|x_{k+1} - x_k\|^2\right] + \sigma^2\alpha_k^2,
$$

where $\{\alpha_k\}$ is given in Algorithm 2, and σ and $L_{\nabla f}$ are given in Assumptions 1 and 2, respectively.

Proof. Let $\Xi_k = \{\xi_1, \ldots, \xi_k\}$ denote the collection of samples drawn up to iteration $k-1$ in Algorithm 2. It then follows from Assumption 1(iii) that

$$
\mathbb{E}[\nabla\tilde{f}(x_{k+1},\xi_{k+1}) - \nabla f(x_{k+1})|\Xi_k] = 0, \quad \mathbb{E}[\|\nabla\tilde{f}(x_{k+1},\xi_{k+1}) - \nabla f(x_{k+1})\|^2|\Xi_k] \leq \sigma^2.
$$

Also, notice from (7) that $\nabla f(x_{k+1}) \in \mathscr{B}(L_f)$ and hence $\nabla f(x_{k+1}) = \Pi_{\mathscr{B}(L_f)}(\nabla f(x_{k+1}))$. By these, the expression of g_{k+1} in Algorithm 2, and the nonexpansiveness of the projection operator $\Pi_{\mathscr{B}(L_f)}$, one has

$$
\begin{aligned}
&\mathbb{E}[\|g_{k+1} - \nabla f(x_{k+1})\|^2 | \Xi_k] = \mathbb{E}[\|\Pi_{\mathscr{B}(L_f)}\big((1-\alpha_k)g_k + \alpha_k \nabla \tilde{f}(x_{k+1}, \xi_{k+1})\big) - \Pi_{\mathscr{B}(L_f)}(\nabla f(x_{k+1}))\|^2 | \Xi_k] \\
&\le \mathbb{E}[\|(1-\alpha_k)g_k + \alpha_k \nabla \tilde{f}(x_{k+1}, \xi_{k+1}) - \nabla f(x_{k+1})\|^2 | \Xi_k] \\
&= \mathbb{E}[\|(1-\alpha_k)(g_k - \nabla f(x_{k+1})) + \alpha_k(\nabla \tilde{f}(x_{k+1}, \xi_{k+1}) - \nabla f(x_{k+1}))\|^2 | \Xi_k] \\
&= (1-\alpha_k)^2 \|g_k - \nabla f(x_{k+1})\|^2 + \alpha_k^2 \mathbb{E}[\|\nabla \tilde{f}(x_{k+1}, \xi_{k+1}) - \nabla f(x_{k+1})\|^2 | \Xi_k] \\
&\quad + 2\alpha_k(1-\alpha_k)\langle g_k - \nabla f(x_{k+1}), \mathbb{E}[\nabla \tilde{f}(x_{k+1}, \xi_{k+1}) - \nabla f(x_{k+1}) | \Xi_k]\rangle \\
&\le (1-\alpha_k)^2 \|g_k - \nabla f(x_{k+1})\|^2 + \sigma^2 \alpha_k^2.
\end{aligned}
$$

Taking expectation on both sides of this inequality yields

$$
\mathbb{E}[\|g_{k+1} - \nabla f(x_{k+1})\|^2] \le (1-\alpha_k)^2 \mathbb{E}[\|g_k - \nabla f(x_{k+1})\|^2] + \sigma^2 \alpha_k^2. \tag{64}
$$

We divide the remainder of the proof by considering two separate cases: $\alpha_k = 1$ and $0 < \alpha_k < 1$.

Case 1) $\alpha_k = 1$. It follows from this and (64) that $\mathbb{E}[\|g_{k+1} - \nabla f(x_{k+1})\|^2] \le \sigma^2 \alpha_k^2$ and hence the conclusion of this lemma clearly holds.

Case 2) $0 < \alpha_k < 1$. By this, (64) and Assumption 3, one has

$$
\begin{aligned}
&\mathbb{E}[\|g_{k+1} - \nabla f(x_{k+1})\|^2] \overset{(64)}{\le} (1-\alpha_k)^2 \mathbb{E}\left[\|g_k - \nabla f(x_k) + \nabla f(x_k) - \nabla f(x_{k+1})\|^2\right] + \sigma^2 \alpha_k^2 \\
&= (1-\alpha_k)^2 \mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right] + (1-\alpha_k)^2 \mathbb{E}\left[\|\nabla f(x_k) - \nabla f(x_{k+1})\|^2\right] \\
&\quad + 2(1-\alpha_k)^2 \mathbb{E}\left[\langle g_k - \nabla f(x_k), \nabla f(x_k) - \nabla f(x_{k+1})\rangle\right] + \sigma^2 \alpha_k^2 \\
&\le (1-\alpha_k)^2 \mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right] + (1-\alpha_k)^2 \mathbb{E}\left[\|\nabla f(x_k) - \nabla f(x_{k+1})\|^2\right] \\
&\quad + (1-\alpha_k)^2 \left(\frac{\alpha_k}{1-\alpha_k} \mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right] + \frac{1-\alpha_k}{\alpha_k} \mathbb{E}\left[\|\nabla f(x_k) - \nabla f(x_{k+1})\|^2\right]\right) + \sigma^2 \alpha_k^2 \\
&= (1-\alpha_k) \mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right] + (1-\alpha_k)^2 \alpha_k^{-1} \mathbb{E}\left[\|\nabla f(x_k) - \nabla f(x_{k+1})\|^2\right] + \sigma^2 \alpha_k^2 \\
&\le (1-\alpha_k) \mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right] + L_{\nabla f}^2 \alpha_k^{-1} \mathbb{E}\left[\|x_{k+1} - x_k\|^2\right] + \sigma^2 \alpha_k^2,
\end{aligned}
$$

where the second inequality follows from $0 < \alpha_k < 1$ and Young's inequality, and the last inequality is due to Assumption 3 and $0 < \alpha_k < 1$. Hence, the conclusion of this lemma also holds in this case. $\square$

The next lemma provides an upper bound on $\mathbb{E}[Q_{\rho_k}(x_k) + \|g_k - \nabla f(x_k)\|^2]$.

Lemma 11. *Suppose that Assumptions 1 and 3 hold, and $\{g_k\}$ and $\{x_k\}$ are generated by Algorithm 1 with $\eta_k \le \alpha_k \le 1$. Then for all $k \ge 1$, we have*

$$
\begin{aligned}
\mathbb{E}\left[Q_{\rho_k}(x_k) + \|g_k - \nabla f(x_k)\|^2\right] \le & Q_{\rho_1}(x_1) + \|g_1 - \nabla f(x_1)\|^2 + \frac{1}{2}\sum_{i=1}^{k-1}(\rho_{i+1} - \rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] \\
& + \frac{1}{2}\sum_{i=1}^{k-1}\left(L_{\nabla f} + \rho_i L - \eta_i^{-1} + 2L_{\nabla f}^2 \alpha_k^{-1}\right)\mathbb{E}\left[\|x_{i+1} - x_i\|^2\right] + \sigma^2 \sum_{i=1}^{k-1} \alpha_i^2.
\end{aligned}
$$

where $\{\alpha_k\}$, $\{\rho_k\}$ and $\{\eta_k\}$ are given in Algorithm 2, Q_ρ and L are respectively defined in (3) and (6), and σ and $L_{\nabla f}$ are given in Assumptions 1 and 3, respectively.

Proof. Observe from (3), (6), and Assumptions 1 and 3 that Q_{ρ_k} is $(L_{\nabla f}+\rho_k L)$-smooth. By this and similar arguments as for deriving (42), one has that for all $k\geq 1$,

$$Q_{\rho_{k+1}}(x_{k+1}) \leq Q_{\rho_k}(x_k) + \frac{1}{2}\left(L_{\nabla f}+\rho_k L - \eta_k^{-1}\right)\|x_{k+1}-x_k\|^2 + \frac{\eta_k}{2}\|g_k - \nabla f(x_k)\|^2 + \frac{1}{2}(\rho_{k+1}-\rho_k)\|c(x_{k+1})\|^2. \tag{65}$$

Notice from the assumption that $1-\alpha_k+\eta_k\leq 1$. Using this, taking expectation on both sides of (65), and summing the resulting inequality with the inequality in Lemma 10, we obtain that

$$\begin{aligned}
&\mathbb{E}\left[Q_{\rho_{k+1}}(x_{k+1}) + \|g_{k+1}-\nabla f(x_{k+1})\|^2\right]\\
&\leq \mathbb{E}\left[Q_{\rho_k}(x_k) + (1-\alpha_k+\eta_k)\|g_k-\nabla f(x_k)\|^2\right] + \frac{1}{2}\left(L_{\nabla f}+\rho_k L-\eta_k^{-1}+2L_{\nabla f}^2\alpha_k^{-1}\right)\mathbb{E}\left[\|x_{k+1}-x_k\|^2\right]\\
&\quad - \frac{\eta_k}{2}\mathbb{E}\left[\|g_k-\nabla f(x_k)\|^2\right] + \frac{1}{2}(\rho_{k+1}-\rho_k)\mathbb{E}\left[\|c(x_{k+1})\|^2\right] + \sigma^2\alpha_k^2\\
&\leq \mathbb{E}\left[Q_{\rho_k}(x_k) + \|g_k-\nabla f(x_k)\|^2\right] + \frac{1}{2}\left(L_{\nabla f}+\rho_k L-\eta_k^{-1}+2L_{\nabla f}^2\alpha_k^{-1}\right)\mathbb{E}\left[\|x_{k+1}-x_k\|^2\right]\\
&\quad - \frac{\eta_k}{2}\mathbb{E}\left[\|g_k-\nabla f(x_k)\|^2\right] + \frac{1}{2}(\rho_{k+1}-\rho_k)\mathbb{E}\left[\|c(x_{k+1})\|^2\right] + \sigma^2\alpha_k^2.
\end{aligned} \tag{66}$$

The conclusion of this lemma follows by replacing k with i in the above inequalities and summing them up for all $1\leq i\leq k-1$. □

The following lemma provides an upper bound on $\mathrm{dist}^2\left(0,\nabla Q_{\rho_k}(x_{k+1})+\mathcal{N}_X(x_{k+1})\right)$.

Lemma 12. *Suppose that Assumptions 1 and 3 hold, and $\{g_k\}$ and $\{x_k\}$ are generated by Algorithm 2. Then for all $k\geq 1$, we have*

$$\mathrm{dist}^2\left(0,\nabla Q_{\rho_k}(x_{k+1})+\mathcal{N}_X(x_{k+1})\right) \leq 3\left(\eta_k^{-2}+(L_{\nabla f}+\rho_k L)^2\right)\|x_{k+1}-x_k\|^2 + 3\|g_k-\nabla f(x_k)\|^2.$$

where $\{\rho_k\}$ and $\{\eta_k\}$ are given in Algorithm 2, $L_{\nabla f}$ is given in Assumption 3, and L and Q_ρ are defined in (6) *and* (3)*, respectively.*

Proof. Recall from the proof of Lemma 11 that Q_{ρ_k} is $(L_{\nabla f}+\rho_k L)$-smooth. The proof of this lemma follows from this and similar arguments as in the proof of Lemma 6. □

Proof of Theorem 2. (i) It follows from (18), (19) and the assumption $1\leq\theta<2$ that for all $i\geq\widetilde{K}_3$,

$$L_{\nabla f}+\rho_i L = L_{\nabla f} i^{-\frac{1}{2}}\eta_i^{-1}/\log(i+2) + L i^{\frac{\theta-2}{4}}\eta_i^{-1}/\log(i+2) \leq L_{\nabla f}\widetilde{K}_3^{-\frac{1}{2}}\eta_i^{-1} + L\widetilde{K}_3^{\frac{\theta-2}{4}}\eta_i^{-1} \leq \eta_i^{-1}/4, \tag{67}$$

$$2L_{\nabla f}^2\alpha_i^{-1} = 2L_{\nabla f}^2\eta_i^{-1}/\log(i+2) \leq 2L_{\nabla f}^2\eta_i^{-1}/\log(\widetilde{K}_3+2) \leq \eta_i^{-1}/4,$$

which imply that

$$L_{\nabla f}+\rho_i L + 2L_{\nabla f}^2\alpha_i^{-1} \leq \eta_i^{-1}/2 \qquad \forall i\geq\widetilde{K}_3. \tag{68}$$

In addition, observe from (18) that $\eta_k\leq\alpha_k\leq 1$ for all $k\geq 1$. It then follows from the proof of Lemma 11 that (66) holds. By (3), (66), (67), (68), Lemmas 11 and 12, and similar arguments as for deriving (49), one can show that for all $k\geq 2\widetilde{K}_3$,

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{dist}^2\left(0,\nabla f(x_{\iota_k})+\rho_{\iota_k-1}\nabla c(x_{\iota_k})c(x_{\iota_k})+\mathcal{N}_X(x_{\iota_k})\right)\right]\\
&\leq \frac{51}{2(k-1)\eta_{k-1}}\Bigg(Q_1(x_1)-Q_1^* + \|g_1-\nabla f(x_1)\|^2 + \frac{1}{2}\sum_{i=1}^{k-1}(\rho_{i+1}-\rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] + \sigma^2\sum_{i=1}^{k-1}\alpha_i^2\\
&\quad + \frac{1}{2}\sum_{i=1}^{\lceil k/2\rceil-1}\left(L_{\nabla f}+\rho_i L-\eta_i^{-1}+2L_{\nabla f}^2\alpha_i^{-1}\right)\mathbb{E}\left[\|x_{i+1}-x_i\|^2\right]\Bigg).
\end{aligned} \tag{69}$$

Further, one can observe that (50) also holds. Using (18), (50), (68), $1 \le \theta < 2$, $\|c(x_i)\| \le C_c$ for all i, Lemma 8, and similar arguments as for deriving (51), (53), (54) and (55), we can show that for all $k \ge 2\widetilde{K}_3$,

$$
\begin{aligned}
&\sum_{i=\widetilde{K}_3}^{k-1} (\rho_{i+1}-\rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] \le \frac{C_3\theta(6-\theta)}{2(2-\theta)},\\
&\sum_{i=1}^{\widetilde{K}_3-1} (\rho_{i+1}-\rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] \le C_c^2\big(\widetilde{K}_3^{\frac{\theta}{4}}-1\big), \qquad \sum_{i=1}^{k-1}\alpha_i^2 \le 1+\log k,\\
&\sum_{i=1}^{\lceil k/2\rceil-1} \big(L_{\nabla f}+\rho_i L-\eta_i^{-1}+2\alpha_i^{-1}L_{\nabla f}^2\big)\,\mathbb{E}\left[\|x_{i+1}-x_i\|^2\right]\\
&\le 2(1+\log\widetilde{K}_3)\big(L_{\nabla f}+L\widetilde{K}_3^{\frac{\theta}{4}}+2L_{\nabla f}^2\widetilde{K}_3^{\frac{1}{2}}\big)\big(L_f^2+C_c^2L_c^2\widetilde{K}_3^{\frac{\theta}{2}}\big).
\end{aligned}
$$

Using these, (18) and (69), we have

$$
\begin{aligned}
&\mathbb{E}\left[\mathrm{dist}^2\left(0,\nabla f(x_{\iota_k})+\rho_{\iota_k-1}\nabla c(x_{\iota_k})c(x_{\iota_k})+\mathcal{N}_X(x_{\iota_k})\right)\right]\\
&\le \frac{51\log(k+2)}{2(k-1)^{\frac{1}{2}}}\Bigg(Q_1(x_1)-Q_1^*+\|g_1-\nabla f(x_1)\|^2+\frac{C_3\theta(6-\theta)}{4(2-\theta)}+\frac{1}{2}C_c^2\big(\widetilde{K}_3^{\frac{\theta}{4}}-1\big)\\
&\quad +\sigma^2(1+\log k)+(1+\log\widetilde{K}_3)\big(L_{\nabla f}+L\widetilde{K}_3^{\frac{\theta}{4}}+2L_{\nabla f}^2\widetilde{K}_3^{\frac{1}{2}}\big)\big(L_f^2+C_c^2L_c^2\widetilde{K}_3^{\frac{\theta}{2}}\big)\Bigg) \qquad \forall k \ge 2\widetilde{K}_3.
\end{aligned}
$$

By this, (3), $\iota_k > \lceil k/2\rceil \ge \widetilde{K}_1$ for all $k \ge 2\widetilde{K}_1$, and Lemma 8 with k replaced by ι_k, one can see that statement (i) of Theorem 2 holds.

(ii) It follows from (21) and (22) that for all $i \ge \widetilde{K}_4$,

$$
\begin{aligned}
L_{\nabla f}+\rho_i L &= L_{\nabla f}i^{-\frac{1}{2}}\eta_i^{-1}/\log(i+2)+L\eta_i^{-1}/\log(i+2)\\
&\le L_{\nabla f}\widetilde{K}_4^{-\frac{1}{2}}\eta_i^{-1}+L\eta_i^{-1}/\log(\widetilde{K}_4+2)\le \eta_i^{-1}/4,\\
2\alpha_i^{-1}L_{\nabla f}^2 &= 2L_{\nabla f}^2\eta_i^{-1}/\log(i+2)\le 2L_{\nabla f}^2\eta_i^{-1}/\log(\widetilde{K}_4+2)\le \eta_i^{-1}/4,
\end{aligned}
$$

which imply that

$$
L_{\nabla f}+\rho_i L+2L_{\nabla f}^2\alpha_i^{-1}\le \eta_i^{-1}/2 \qquad \forall i \ge \widetilde{K}_4. \tag{70}
$$

By this, Lemma 8, (21), (50), $\nu \le 1/2$, $\|c(x_i)\| \le C_c$ for all i, and similar arguments as for deriving (57), (58) and (60), one can show that for all $k \ge 2\widetilde{K}_4$,

$$
\begin{aligned}
&\sum_{i=\widetilde{K}_4}^{k-1} (\rho_{i+1}-\rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] \le C_4k^{\frac{1}{2}-\nu}(1+\log k),\\
&\sum_{i=1}^{\widetilde{K}_4-1} (\rho_{i+1}-\rho_i)\mathbb{E}\left[\|c(x_{i+1})\|^2\right] \le C_c^2\big(\widetilde{K}_4^{\frac{1}{2}}-1\big),\\
&\sum_{i=1}^{\lceil k/2\rceil-1} \big(L_{\nabla f}+\rho_i L-\eta_i^{-1}+2L_{\nabla f}^2\alpha_i^{-1}\big)\,\mathbb{E}\left[\|x_{i+1}-x_i\|^2\right]\\
&\le 2(1+\log\widetilde{K}_4)\big(L_{\nabla f}+L\widetilde{K}_4^{\frac{1}{2}}+2L_{\nabla f}^2\widetilde{K}_4^{\frac{1}{2}}\big)\big(L_f^2+C_c^2L_c^2\widetilde{K}_4\big).
\end{aligned}
$$

In addition, observe from (21) that $\eta_k \le \alpha_k \le 1$ for all $k \ge 1$. Using this, (70), and similar arguments as in the proof of statement (i) of this theorem, we can see that (69) holds for all $k \ge 2\widetilde{K}_4$. Also, by

(21) and (54), one has $\sum_{i=1}^{k-1} \alpha_i^2 \leq 1 + \log k$ for all $k \geq 1$. Using this, (21), (69), and the above three inequalities, we have

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{dist}^2\left(0, \nabla f(x_{\iota_k}) + \rho_{\iota_k - 1} \nabla c(x_{\iota_k}) c(x_{\iota_k}) + \mathcal{N}_X(x_{\iota_k})\right)\right] \\
&\leq \frac{51 \log(k+2)}{2(k-1)^{\frac{1}{2}}} \Bigg(Q_1(x_1) - Q_1^* + \|g_1 - \nabla f(x_1)\|^2 + \frac{1}{2} C_4 k^{\frac{1}{2}-\nu}(1 + \log k) + \frac{1}{2} C_c^2 \big(\widetilde{K}_4^{\frac{1}{2}} - 1\big) \\
&\quad + \sigma^2 (1 + \log k) + (1 + \log \widetilde{K}_4)\big(L_{\nabla f} + L \widetilde{K}_4^{\frac{1}{2}} + 2 L_{\nabla f}^2 \widetilde{K}_4^{\frac{1}{2}}\big)\big(L_f^2 + C_c^2 L_c^2 \widetilde{K}_4\big) \Bigg) \qquad \forall k \geq 2\widetilde{K}_4.
\end{aligned}$$

By this, (3), $\iota_k > \lceil k/2 \rceil \geq \widetilde{K}_4$ for all $k \geq 2\widetilde{K}_4$, and Lemma 9 with k replaced by ι_k, one can see that statement (ii) of Theorem 2 holds. $\square$

5 Concluding remarks

In this paper, we studied a class of deterministically constrained stochastic optimization problems. Existing methods typically aim to find an ϵ-stochastic stationary point, where the expected violations of both the constraints and first-order stationarity are within a prescribed accuracy of ϵ. However, in many practical applications, it is crucial that the constraints be nearly satisfied with certainty, making such an ϵ-stochastic stationary point potentially undesirable due to the risk of significant constraint violations. To address this issue, we proposed single-loop variance-reduced stochastic first-order methods with provable guarantees on both sample complexity and first-order operation complexity to find a stronger ϵ-stochastic stationary point, where the constraint violation is within ϵ with *certainty*, and the expected violation of first-order stationarity is within ϵ.

For future work, we plan to conduct computational studies on the proposed methods and compare their performance with existing approaches. Additionally, we are extending these methods to deterministically constrained stochastic convex optimization in [25], as well as to stochastic optimization with both stochastic objective functions and stochastic constraints in [24].

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, 2024.

[2] A. S. Berahas, F. E. Curtis, M. J. O'Neill, and D. P. Robinson. A stochastic sequential quadratic optimization algorithm for nonlinear-equality-constrained optimization with rank-deficient Jacobians. *Mathematics of Operations Research*, 2023.

[3] 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.

[4] A. S. Berahas, J. Shi, Z. Yi, and B. Zhou. Accelerating stochastic sequential quadratic programming for equality constrained optimization using predictive variance reduction. *Computational Optimization and Applications*, 86(1):79–116, 2023.

[5] D. Bertsekas. Network optimization: continuous and discrete methods, ser. optimization and neural computation series. athena scientific, 1998.

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

[7] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[8] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212, 2011.

[9] S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022.

[10] F. E. Curtis, X. Jiang, and Q. Wang. Almost-sure convergence of iterates and multipliers in stochastic sequential quadratic optimization. *arXiv preprint arXiv:2308.03687*, 2023.

[11] 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.

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

[13] A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex SGD. *Advances in neural information processing systems*, 32, 2019.

[14] C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

[15] 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.

[16] X. Fontaine, S. Mannor, and V. Perchet. An adaptive stochastic optimization algorithm for resource allocation. In *Algorithmic Learning Theory*, pages 319–363. PMLR, 2020.

[17] Y. Gao, A. Rodomanov, and S. U. Stich. Non-convex stochastic composite optimization with Polyak momentum. *arXiv preprint arXiv:2403.02967*, 2024.

[18] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

[19] Z. Jiang, K. Mukherjee, and S. Sarkar. On consensus-disagreement tradeoff in distributed optimization. In *2018 Annual American Control Conference (ACC)*, pages 571–576. IEEE, 2018.

[20] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[21] Z. Li, P.-Y. 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.

[22] Q. Lin, R. Ma, and Y. Xu. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. *Computational optimization and applications*, 82(1):175–224, 2022.

[23] B. Liu, M. Wang, H. Foroosh, M. Tappen, and M. Pensky. Sparse convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 806–814, 2015.

[24] Z. Lu, S. Mei, and Y. Xiao. Variance-reduced first-order methods for stochastic constrained nonconvex optimization. 2024. Working paper.

[25] Z. Lu and Y. Xiao. Variance-reduced first-order methods for deterministically constrained stochastic convex optimization. 2024. Working paper.

[26] D. R. Luke. Proximal methods for image processing. *Nanoscale Photonic Imaging*, 134:165, 2020.

[27] F. Maggioni, M. Kaut, and L. Bertazzi. Stochastic optimization models for a single-sink transportation problem. *Computational Management Science*, 6:251–267, 2009.

[28] 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.

[29] S. Na and M. W. Mahoney. Asymptotic convergence rate and statistical inference for stochastic sequential quadratic programming. *arXiv: 2205.13687 v1*, 2022.

[30] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[31] 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.

[32] M. F. Sahin, A. Alacaoglu, F. Latorre, V. Cevher, et al. An inexact augmented Lagrangian framework for nonconvex optimization with nonlinear constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

[33] V. Shabazbegian, H. Ameli, M. T. Ameli, and G. Strbac. Stochastic optimization model for coordinated operation of natural gas and electricity networks. *Computers & Chemical Engineering*, 142:107060, 2020.

[34] K. S. Shehadeh and R. Padman. Stochastic optimization approaches for elective surgery scheduling with downstream capacity constraints: Models, challenges, and opportunities. *Computers & Operations Research*, 137:105523, 2022.

[35] Q. Shi, X. Wang, and H. Wang. A momentum-based linearized augmented Lagrangian method for nonconvex constrained stochastic optimization. *Optimization Online*, 2022.

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

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

[38] Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh. Spiderboost and momentum: Faster variance reduction algorithms. *Advances in Neural Information Processing Systems*, 32, 2019.

[39] Y. Xu and Y. Xu. Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization. *Journal of Optimization Theory and Applications*, 196(1):266–297, 2023.

[40] H. Yu, R. Jin, and S. Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193, 2019.