Contents

1 Introduction, notations and algorithm **2**
1.1 Introduction 2
1.2 Algorithm and Notations 3

2 Analysis **4**
2.1 Assumptions 4
2.2 Comparison with other works 5
2.3 Worse-case complexity 5
2.4 Proof for unifomrly boundness of $\{M_{\rho c}^K\}_K$ 22

A Transforming (77) into LPs. **28**

A Quadratic Penalty Method for Nonconvex Stochastic Constrained Optimization

Lingjun Guo

1 Introduction, notations and algorithm

1.1 Introduction

Consider an equality- and inequality-constrained problem with $m \in \mathbb{N}$ constraints. Let the set $\mathcal{E} \subseteq [m]$ be the index set for equality constraints, and the set $\mathcal{I} \subseteq [m]$ be the index set for the inequality constraints. Suppose $\mathcal{E} \cap \mathcal{I} = \emptyset$ and $\mathcal{E} \cup \mathcal{I} = [m]$. We aim to solve the following problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \ & f(x) = \mathbb{E}_{\xi^f}[\overline{f}(x, \xi^f)] \\ \text{s.t. } & c_i(x) = \mathbb{E}_{\xi^c}[\bar{c}_i(x, \xi^c)] = 0 \text{ for all } i \in \mathcal{E}, \text{ and} \\ & c_j(x) = \mathbb{E}_{\xi^c}[\bar{c}_j(x, \xi^c)] \leq 0 \text{ for all } j \in \mathcal{I}. \end{aligned} \tag{1}$$

By stacking the scalar functions c_i for all $i \in [m]$ together, one obtains a vector function $c : \mathbb{R}^n \to \mathbb{R}^m$ where its ith element $[c]_i = c_i$. Suppose the functions $f : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^m$ are smooth functions with Lipchitz continuous gradient. Samples (ξ^f, ξ^c) are draw from two σ-algebra Ξ^f and Ξ^c, equipped with probability measures, respectively. Moreover, define a function $\mathcal{P} : \mathbb{R}^m \to \mathbb{R}^m$ where its ith element, for all $i \in [m]$, is defined by

$$[\mathcal{P}(c)]_i = \begin{cases} [c]_i & \text{, if } i \in \mathcal{E} \\ \max\{0, [c]_i\} & \text{, if } i \in \mathcal{I}. \end{cases} \tag{2}$$

For any constant $\rho \in \mathbb{R}$, the quadratic merit function at $x \in \mathbb{R}^n$ is defined as

$$m_\rho(x) := f(x) + \frac{\rho}{2} \|\mathcal{P}(c(x))\|^2.$$

For $i \in \mathcal{E}$, from the definition in (2) and the chain rule, one has the derivative

$$\tfrac{d([\mathcal{P}(c(x))]_i)^2}{dx} = \tfrac{d([c]_i)^2}{dx} = 2(\nabla [c(x)]_i)[c(x)]_i.$$

For $j \in \mathcal{I}$, from the chain rule and the fact that $\frac{d(\max\{0,[c]_j\})^2}{d[c]_j}\big|_{[c]_j=0} = 0$, one has

$$\frac{d([\mathcal{P}(c(x))]_j)^2}{dx} = \frac{d(\max\{0,[c]_j\})^2}{dx} = \begin{cases} \mathbf{0_n} & \text{, if } [c(x)]_j \leq 0, \\ 2(\nabla[c(x)]_j)[c(x)]_j & \text{, if } [c(x)]_j > 0. \end{cases}$$

As a result, one has the derivative of the function $\mathcal{P}(c)$ with respect to x to be

$$\begin{aligned} \nabla(\|\mathcal{P}(c(x))\|^2) &= \nabla(\sum_{i\in[m]} \|[\mathcal{P}(c(x))]_i\|^2) \\ &= \sum_{i\in\mathcal{E}} \nabla(\|[\mathcal{P}(c(x))]_i\|^2) + \sum_{j\in\mathcal{I}} \nabla(\|[\mathcal{P}(c(x))]_j\|^2) \\ &= \sum_{i\in\mathcal{E}} 2(\nabla[c(x)]_i)[c(x)]_i + \sum_{j\in\mathcal{I},[c(x)]_j>0} 2(\nabla[c(x)]_j)[c(x)]_j \\ &= 2\nabla c(x)^T \mathcal{P}(c(x)). \end{aligned}$$

Hence, the gradient of ∇m_ρ at x is

$$\nabla m_\rho(x) = \nabla f(x) + \rho \nabla c(x)^T \mathcal{P}(c(x)).$$

Given scalars $(r_\eta, r_\rho, r_\omega) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \times \mathbb{R}$, scalars $(\eta_1, \rho_1, \omega_1) \in \mathbb{R}^3_{>0}$, and define sequences $\{\eta_k\}_{k\in\mathbb{N}}$, $\{\rho_k\}_{k\in\mathbb{N}}$, $\{\omega_{k+1}\}_{k\in\mathbb{N}}$ as

$$\eta_k = \eta_1 k^{-r_\eta}, \ \rho_k = \rho_1 k^{r_\rho} \text{ and } \omega_{k+1} = \omega_1 k^{r_\omega}. \tag{3}$$

Definition 1. Let $x_{\xi_x} \in \mathbb{R}^n$ be a random variable depending on ξ_x. Then x_{ξ_x} is called an ϵ-stochastic-stationary point of the constrained problem (1), if there exits a random variable $\lambda_{\xi_\lambda} \in \mathbb{R}^m$ such that

$$\mathbb{E}_{(\xi_x,\xi_\lambda)}[\|\nabla f(x_{\xi_x}) + \nabla c(x_{\xi_x})^T \lambda_{\xi_\lambda}\| + \|\mathcal{P}(c(x_{\xi_x}))\| + \|\lambda_{\xi_\lambda}^T \mathcal{P}(c(x_{\xi_x}))\|] \leq \epsilon. \tag{4}$$

1.2 Algorithm and Notations

An algorithm for solving (1) is given below.

Algorithm 1 Partial Variance Reduced Penalty Method for (1).

Input: initial point x_1, maximum iteration K, sequences $\{\rho_k, \eta_k, \omega_k\}_{k\in\mathbb{N}}$ given in (3).

1: Draw samples $\mathcal{S}_1 = (\xi_1^f, \xi_1^J, \xi_1^c)$. Let $C_1 = \bar{c}(x_1, \xi_1^c)$ and compute

$$\bar{d}_1 = \nabla \bar{f}(x_1, \xi_1^f) + \rho_1 [\nabla \bar{c}(x_1, \xi_1^J)]^T \mathcal{P}(C_1).$$

2: **for** $k = 1, \cdots, K-1$ **do**

3: Compute $x_{k+1} = x_k - \eta_k \bar{d}_k$.

4: Draw samples $\mathcal{S}_{k+1} = (\xi_{k+1}^f, \xi_{k+1}^J, \xi_{k+1}^c)$.

5: Compute

$$C_{k+1} = \bar{c}(x_{k+1}, \xi_{k+1}^c) + (1-\omega_{k+1})(C_k - \bar{c}(x_k, \xi_{k+1}^c))$$

6: Compute $\bar{d}_{k+1} = \nabla \bar{f}(x_{k+1}, \xi_{k+1}^f) + \rho_{k+1} [\nabla \bar{c}(x_{k+1}, \xi_{k+1}^J)]^T \mathcal{P}(C_{k+1})$.

7: **end for**

Output: $x_{\hat{K}}$ where $\hat{K}$ is uniformly random from $\{1, \cdots, K\}$.

Notations: Let $\mathbb{N}$ be the set of natural numbers. For any $m \in \mathbb{N}$, define the set $[m] = \{1, \cdots, m\}$. For any scalar pair $(m, n) \in \mathbb{N}^2$, let $\mathbb{R}^n$ represents a n-dimensional vector space and $\mathbb{R}^{n\times m}$ represents a n-by-m-dimensional matrix space. Define $\mathbb{R}_{>0} = \{x \in \mathbb{R} : x > 0\}$ and define the set $\mathbb{R}_{\geq 0}$ in a similar way. Define $\mathbf{0}_m \in \mathbb{R}^m$ be a m-dimential zero vector. For any m-dimensional vector v, let $[v]_j$ represents the jth element of v for any $j \in [m]$. We let $\|\cdot\|$ denote the vector/matrix two-norm on finite vector/matrix spaces, respectively. For any $x \in \mathbb{R}$, we let $|x|$ represent the absolute value of x. For a vector function $c : \mathbb{R}^n \to \mathbb{R}^m$, define the Jacobian of c to be the function $\nabla c : \mathbb{R}^n \to \mathbb{R}^{m\times n}$. For any $k \in \mathbb{N}$, we define the set $\mathcal{S}_k = \{\xi_k^f, \xi_k^J, \xi_k^c\} \in \Xi^f \times \Xi^c \times \Xi^c$ to be the set of samples drawn at the kth iteration. Define the set $[\mathcal{S}_k] = \cup_{l=1}^k \mathcal{S}_l$ to be the total samples drawn up to (and including) the kth iteration. For any two nonnegative sequences $\{\mu_k\}_{k\in\mathbb{N}} \subseteq \mathbb{R}_{\geq 0}$ and $\{\nu_k\}_{k\in\mathbb{N}} \subseteq \mathbb{R}_{\geq 0}$, we say $\mu_k = \mathcal{O}(\nu_k)$ if there exists a scalar pair $(K, M) \in \mathbb{N} \times \mathbb{R}_{>0}$ such that $\mu_k \leq M\nu_k$ for all $k \geq K$. We say $\mu_k = \tilde{\mathcal{O}}(\nu_k)$ if such scalar pair (K, M) ensures $\mu_k \leq M(\log k)\nu_k$ for all $k \geq K$. Let $x_1 \in \mathbb{R}^n$ be any initial point for Algorithm 1, and unless mentioned specifically, we assume all the expectations used in the paper are conditioned on x_1, that is $\mathbb{E}[\cdot] := \mathbb{E}[\cdot|x_1]$.

2 Analysis

2.1 Assumptions

We make the following assumptions.

Assumption 1. There exist positive constant pair $(L_{\nabla f}, L_{\nabla c}) \in \mathbb{R}^2_{>0}$ such that the objective

function f has $L_{\nabla f}$-Lipschitz gradient, and the constraint function c has $L_{\nabla c}$-Lipschitz gradient.

Assumption 2. There exist constants $(\kappa_{\nabla f}, \kappa_c) \in \mathbb{R}^2_{>0}$ such that for any $x \in \mathbb{R}^n$, it holds $\|\nabla f(x)\| \leq \kappa_{\nabla f}$ and $\|c(x)\| \leq \kappa_c$.

Assumption 3. For any $x \in \mathbb{R}^n$, the following holds

$$\nabla f(x) = \mathbb{E}_{\xi^f}[\nabla \bar{f}(x, \xi^f)],\ \nabla c(x) = \mathbb{E}_{\xi^J}[\nabla \bar{c}(x, \xi^J)] \text{ and } c(x) = \mathbb{E}_{\xi^c}[\bar{c}(x, \xi^c)].$$

Moreover, there exist constants $(\delta_{\nabla f}, \delta_c, \delta_{\nabla c}) \in \mathbb{R}^3_{>0}$ such that for any $x \in \mathbb{R}^n$, it holds

$$\mathbb{E}_{\xi^f}\left[\|\nabla \bar{f}(x, \xi^f) - \nabla f(x)\|^2\right] \leq \delta^2_{\nabla f},$$
$$\mathbb{E}_{\xi^J}\left[\|\nabla \bar{c}(x, \xi^J) - \nabla c(x)\|^2\right] \leq \delta^2_{\nabla c} \text{ and } \mathbb{E}_{\xi^c}\left[\|\bar{c}(x, \xi^c) - c(x)\|^2\right] \leq \delta^2_c.$$

Assumption 4. There exists a positive scalar $L_{\bar{c}} > 0$ such that for any $(x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n$, it holds

$$\mathbb{E}_{\xi^c}\left[\|\bar{c}(x_1, \xi^c) - \bar{c}(x_2, \xi^c)\|^2\right] \leq L^2_{\bar{c}}\|x_1 - x_2\|^2.$$

Assumption 5. There exists a positive scalar $\sigma_{\nabla c} > 0$ such that for any $x \in \mathbb{R}^n$, one has $\left\|[\nabla c(x)]^T \mathcal{P}(c(x))\right\| \geq \sigma_{\nabla c}\|\mathcal{P}(c(x))\|$.

Remark 1. When there are only equality constraints, i.e., $[m] = \mathcal{E}$, one can observe that Assumption 5 is weaker than the strong LICQ for equality constraints. When there are only inequality constraints, i.e., $[m] = \mathcal{I}$, then Assumption 5 can be implied by the following three assumptions:

1. For any $\epsilon > 0$, there exists a scalar $\sigma(\epsilon) > 0$ such that for any $x \in \{x : \|\mathcal{P}(c(x))\| \geq \epsilon\}$, it holds
$$\|[\nabla c(x)]^T \mathcal{P}(c(x))\| = \tfrac{1}{2}\|\nabla(\|\mathcal{P}(c(x))\|^2)\| \geq \sigma(\epsilon).$$
(Infeasible points are non-stationary points of the function $\|\mathcal{P} \circ c\|^2$!)
2. Strong LICQ at active constraints.
3. x is bounded.

2.2 Comparison with other works

2.3 Worse-case complexity

Lemma 1. Under Assumptions 1 – 3, let $\{x_k\}_{k\in\mathbb{N}}$ be the iterates generated by Algorithm 1 when $K = \infty$. In addition, suppose the parameters (η_1, ρ_1) satisfy

$$\rho_1 \geq 1 \geq \eta_1. \tag{5}$$

	stochastic constraint	convexity	CQ	x_1	ϵ as prior	complexity
[1]	equality	**nonconvex**	Assumption 5	$x_1 \in \mathcal{X}$	**No**	$\tilde{\mathcal{O}}(\epsilon^{-5})$
[2]	equality	**nonconvex**	Strong LICQ	$\boldsymbol{x_1 \in \mathbb{R}^n}$	Yes	$\tilde{\mathcal{O}}(\epsilon^{-5})$
[3]	inequality	**nonconvex**	Assumption 5	$[c]_j(x_1) \leq 0$ $\forall j \in \mathcal{I}$	Yes	$\tilde{\mathcal{O}}(\epsilon^{-7})$
[4]	inequality	weakly-convex	Slater-type CQ	$x_1 \in \mathcal{X}$	Yes	$\tilde{\mathcal{O}}(\epsilon^{-6})$
Ours	**equality, inequality**	**nonconvex**	Assumption 5	$\boldsymbol{x_1 \in \mathbb{R}^n}$	**No**	$\boldsymbol{\tilde{\mathcal{O}}(\epsilon^{-6})}$

Table 1: Comparison with other algorithms. CQ represents constraint qualification, and LICQ represents linear independence constraint qualification. $\mathcal{X}$ represents a closed convex set. ϵ represents the output error on the right-hand side of Definition 1 for inequality case, while removing the $\|\lambda_{\xi_\lambda}^T \mathcal{P}(c(x_\xi))\|$ term for equality case. Requiring ϵ as a prior means that one needs to choose various algorithm parameters depending on ϵ before running. The complexity represents the sample complexity. All the above algorithms require the Assumption 3.

Moreover, define scalars

$$\kappa_{\nabla \bar{c}} = \sqrt{\delta_{\nabla c}^2 + L_{\bar{c}}^2},\ L_{\|\mathcal{P}\circ c\|^2} = 2L_{\bar{c}}^2 + 2\kappa_c L_{\nabla c}, \text{ and } L_{\nabla m_1} = L_{\nabla f} + L_{\|\mathcal{P}\circ c\|^2}. \tag{6}$$

For any $k \in \mathbb{N}$, define

$$\tilde{\delta}_{C_k} = \mathbb{E}_{[\mathcal{S}_k]}[\|C_k - c(x_k)\|^2]. \tag{7}$$

Then, one has the following.

(1). For any $x \in \mathbb{R}^n$, it holds $\|\nabla c(x)\| \leq L_{\bar{c}}$ and $\|\mathcal{P}(c(x))\|^2 \leq \kappa_c^2$. Moreover, the function $\mathcal{P}$ is 1-Lipschitz continuous, the function $\|\mathcal{P} \circ c\|^2$ has $L_{\|\mathcal{P}\circ c\|^2}$-Lipschitz continous gradient, and the function m_{ρ_k} has an $L_{\nabla m_1}\rho_k$-Lipschitz continous gradient.

(2). For any $k \in \mathbb{N}$, it holds

$$\mathbb{E}_{\mathcal{S}_k}[\|\nabla \bar{c}(x_k, \xi_k^J)\|^2 \big| [\mathcal{S}_{k-1}]] \leq \kappa_{\nabla \bar{c}}^2, \text{ and} \tag{8}$$

$$\begin{aligned}\mathbb{E}_{[\mathcal{S}_k]}\left[\|\bar{d}_k\|^2\right] \leq\ & 2\mathbb{E}_{\mathcal{S}_k}[\|\nabla m_{\rho_k}(x_k)\|^2] + 6\delta_{\nabla f}^2 + 6\kappa_{\nabla \bar{c}}^2 \rho_k^2 \tilde{\delta}_{C_k} \\ & + 6\delta_{\nabla c}^2 \rho_k^2 \mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2].\end{aligned} \tag{9}$$

Proof. We first prove item (1). For any $(x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n$, from the Assumption 3, convexity of vector squared two-norm, Jensen's inequality and Assumption 4, one has

$$\begin{aligned}\|c(x_1) - c(x_2)\|^2 &= \|\mathbb{E}_{\xi^c}[\bar{c}(x_1, \xi^c) - \bar{c}(x_2, \xi^c)]\|^2 \\ &\leq \mathbb{E}_{\xi^c}\|[\bar{c}(x_1, \xi^c) - \bar{c}(x_2, \xi^c)\|^2] \\ &\leq \mathbb{E}_{\xi^c}[L_{\bar{c}}^2 \|x_1 - x_2\|^2] \\ &= L_{\bar{c}}^2 \|x_1 - x_2\|^2,\end{aligned}$$

which shows that the function c is $L_{\bar{c}}$-Lipshitz continous. Since c is differentiable by Assumption 1, we have that for any $x \in \mathbb{R}^n$, it holds $\|\nabla c(x)\| \leq L_{\bar{c}}$. Next, we show that $\mathcal{P}$

is 1-Lipschitz continuous. For any $(c_1, c_2) \in \mathbb{R}^m \times \mathbb{R}^m$, from the definition of $\mathcal{P}$ in (2), one has

$$\begin{aligned}\|\mathcal{P}(c_1) - \mathcal{P}(c_2)\|^2 &= \sum_{i\in[m]} ([\mathcal{P}(c_1)]_i - [\mathcal{P}(c_2)]_i)^2 \\ &= \sum_{i\in\mathcal{E}} ([c_1]_i - [c_2]_i)^2 + \sum_{j\in\mathcal{I}} (\max\{0, [c_1]_j\} - \max\{0, [c_2]_j\})^2.\end{aligned} \tag{10}$$

Observe that for the last term on the right-hand side of the last line of (10), for any $j \in \mathcal{I}$, one has

$$\begin{aligned}(\max\{0, [c_1]_j\} - \max\{0, [c_2]_j\})^2 &= \begin{cases} ([c_1]_j - [c_2]_j)^2, & \text{if } [c_1]_j \geq 0, [c_2]_j \geq 0 \\ ([c_2]_j)^2, & \text{if } [c_1]_j \leq 0, [c_2]_j \geq 0 \\ ([c_1]_j)^2, & \text{if } [c_1]_j \geq 0, [c_2]_j \leq 0 \\ 0, & \text{if } [c_1]_j \leq 0, [c_2]_j \leq 0 \end{cases} \\ &\leq ([c_1]_j - [c_2]_j)^2.\end{aligned} \tag{11}$$

Combining (10) and (11), one now has

$$\|\mathcal{P}(c_1) - \mathcal{P}(c_2)\|^2 \leq \sum_{i\in[m]} ([c_1]_i - [c_2]_i)^2 = \|c_1 - c_2\|^2, \tag{12}$$

i.e., the function $\mathcal{P}$ is 1-Lipschitz continous. Moreover, from (12) and Assumption 2, for any $x \in \mathbb{R}^n$, one has

$$\|\mathcal{P}(c(x))\|^2 = \|\mathcal{P}(c(x)) - \mathcal{P}(\mathbf{0_m})\|^2 \leq \|c(x) - \mathbf{0_m}\|^2 = \|c(x)\|^2 \leq \kappa_c^2. \tag{13}$$

Next, we give a Lipschitz constant for $\nabla(\|\mathcal{P} \circ c\|^2)$. For any $(x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n$, by adding, subtracting a term, the triangle inequality and submultiplicity, one has

$$\begin{aligned}&\nabla\left(\|\mathcal{P}(c(x_1))\|^2\right) - \nabla\left(\|\mathcal{P}(c(x_2))\|^2\right) \\ &= 2\|[\nabla c(x_1)]^T \mathcal{P}(c(x_1)) - [\nabla c(x_2)]^T \mathcal{P}(c(x_2))\| \\ &= 2\|[\nabla c(x_1)]^T \mathcal{P}(c(x_1)) - [\nabla c(x_1)]^T \mathcal{P}(c(x_2)) \\ &\qquad + [\nabla c(x_1)]^T \mathcal{P}(c(x_2)) - [\nabla c(x_2)]^T \mathcal{P}(c(x_2))\| \\ &\leq 2\|[\nabla c(x_1)]\|\|\mathcal{P}(c(x_1)) - \mathcal{P}(c(x_2))\| \\ &\qquad + 2\|\nabla c(x_1) - \nabla c(x_2)\|\|\mathcal{P}(c(x_2))\|.\end{aligned} \tag{14}$$

From that $\mathcal{P}$ is 1-Lipschitz continous, c is $L_{\bar{c}}$-Lipschitz continous, Assumption 2 and (13), one has

$$\begin{aligned}&2\|[\nabla c(x_1)]\|\|\mathcal{P}(c(x_1)) - \mathcal{P}(c(x_2))\| + 2\|\nabla c(x_1) - \nabla c(x_2)\|\|\mathcal{P}(c(x_2))\| \\ &\leq 2L_{\bar{c}}\|c(x_1) - c(x_2)\| + 2L_{\nabla c}\|x_1 - x_2\|\kappa_c \\ &\leq 2L_{\bar{c}}^2\|x_1 - x_2\| + 2\kappa_c L_{\nabla c}\|x_1 - x_2\| \\ &= L_{\|\mathcal{P}\circ c\|^2}\|x_1 - x_2\|.\end{aligned} \tag{15}$$

On the other hand, from the additivity of Lipschitz constants and $\rho_k \geq 1$, one has that the function ∇m_{ρ_k} has a Lipschitz constant which is at most

$$L_{\nabla f} + \tfrac{1}{2}\rho_k L_{\|\mathcal{P}\circ c\|^2} \leq L_{\nabla f} + \rho_k L_{\|\mathcal{P}\circ c\|^2} \leq \rho_k (L_{\nabla f} + L_{\|\mathcal{P}\circ c\|^2}) = L_{\nabla m_1}\rho_k.$$

Second, we prove item (2). By adding and subtracting the term $\nabla c(x_k)$, the expansion of quadratic functions. With Assumption 2 and 3, one has

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_k}[\|\nabla \bar{c}(x_k, \xi_k^J)\|^2 \big| [\mathcal{S}_{k-1}]] \\
&= \mathbb{E}_{\mathcal{S}_k}[\|\nabla \bar{c}(x_k, \xi_k^J) - \nabla c(x_k)\|^2 \big| [\mathcal{S}_{k-1}]] + \mathbb{E}_{\mathcal{S}_k}[\|\nabla c(x_k)\|^2 \big| [\mathcal{S}_{k-1}]] \\
&\qquad + 2\mathbb{E}_{\mathcal{S}_k}[(\nabla \bar{c}(x_k, \xi_k^J) - \nabla c(x_k))^T \nabla c(x_k) \big| [\mathcal{S}_{k-1}]] \\
&\leq \delta_{\nabla c}^2 + \kappa_{\nabla c}^2 + 0 = \kappa_{\nabla \bar{c}}^2.
\end{aligned}$$

Hence, (8) is proved. Before proceeding to show (9), we prove a useful inequality. For any $l \in \mathbb{N}$ number of vectors $\nu_i \in \mathbb{R}^n$ where $i \in [l]$. From the convexity of the squared norm and Jensen's inequality, one has

$$\|\frac{1}{l}\sum_{i=1}^{l} \nu_i\|^2 \leq \frac{1}{l}\sum_{i=1}^{l} \|\nu_i\|^2 \iff \|\sum_{i=1}^{l} \nu_i\|^2 \leq l \sum_{i=1}^{l} \|\nu_i\|^2 . \tag{16}$$

By adding and subtracting a term, (16) with $l = 2$, the linearity of expectation, one has

$$\begin{aligned}
\mathbb{E}_{[\mathcal{S}_k]}\left[\|\bar{d}_k\|^2\right] &= \mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k) + \left(\bar{d}_k - \nabla m_{\rho_k}(x_k)\right)\|^2] \\
&\leq 2\mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\|^2] + 2\mathbb{E}_{[\mathcal{S}_k]}[\|\bar{d}_k - \nabla m_{\rho_k}(x_k)\|^2].
\end{aligned} \tag{17}$$

For the second term in the line (17). Again, by adding and subtracting a term, (16) where $l = 3$ and the linearity of expectation, one has

$$\begin{aligned}
&2\mathbb{E}_{[\mathcal{S}_k]}[\|\bar{d}_k - \nabla m_{\rho_k}(x_k)\|^2] \\
&= 2\mathbb{E}_{[\mathcal{S}_k]}[\|(\nabla \overline{f}(x_k; \xi_k^f) - \nabla f(x_k)) + \rho_k \left([\nabla \bar{c}(x_k, \xi_k^J)]^T(\mathcal{P}(C_k) - \mathcal{P}(c(x_k))\right) \\
&\qquad + \rho_k \left(\nabla \bar{c}(x_k, \xi_k^J) - \nabla c(x_k))^T \mathcal{P}(c(x_k))\right)\|^2] \\
&\leq 6\mathbb{E}_{[\mathcal{S}_k]}[\|(\nabla \overline{f}(x_k; \xi_k^f) - \nabla f(x_k))\|^2] + 6\rho_k^2 \mathbb{E}_{[\mathcal{S}_k]}[\|[\nabla \bar{c}(x_k, \xi_k^J)]^T(\mathcal{P}(C_k) - \mathcal{P}(c(x_k))\|^2] \\
&\qquad + 6\rho_k^2 \mathbb{E}_{[\mathcal{S}_k]}[(\nabla \bar{c}(x_k, \xi_k^J) - \nabla c(x_k))^T \mathcal{P}(c(x_k))\|^2].
\end{aligned} \tag{18}$$

Next, we bound the three terms on the right-hand side of the last inequality in (18). From Assumption (3) and conditional expectation, one can conclude that

$$\begin{aligned}
&6\mathbb{E}_{[\mathcal{S}_k]}[\|(\nabla \overline{f}(x_k; \xi_k^f) - \nabla f(x_k))\|^2] \\
&= 6\mathbb{E}_{[\mathcal{S}_{k-1}]}[\mathbb{E}_{\mathcal{S}_k}[\|(\nabla \overline{f}(x_k; \xi_k^f) - \nabla f(x_k))\|^2 \big| [\mathcal{S}_{k-1}]]] \leq 6\delta_{\nabla f}^2.
\end{aligned} \tag{19}$$

Note that the iterate x_k only depends on the initial point x_1 and samples $[\mathcal{S}_{k-1}]$. Hence, when conditioned on x_1 and $[\mathcal{S}_{k-1}]$, the random variables $\nabla\bar{c}(x_k,\xi_k^J)$ and C_k are independent since since they depend on two i.i.d. samples, ξ_k^J and ξ_k^c, respectively. From submultiplicity, conditional expectation, and the aforementioned independence. Also from the Lipschitz constants for $\mathcal{P}$ and (8), one has

$$\begin{aligned}
&6\rho_k^2\mathbb{E}_{[\mathcal{S}_k]}[\|[\nabla\bar{c}(x_k,\xi_k^J)]^T(\mathcal{P}(C_k)-\mathcal{P}(c(x_k))\|^2]\\
&\leq 6\rho_k^2\mathbb{E}_{[\mathcal{S}_{k-1}]}[\mathbb{E}_{\mathcal{S}_k}[\|\nabla\bar{c}(x_k,\xi_k^J)\|^2\|\mathcal{P}(C_k)-\mathcal{P}(c(x_k))\|^2\big|[\mathcal{S}_{k-1}]]]\\
&= 6\rho_k^2\mathbb{E}_{[\mathcal{S}_{k-1}]}[\mathbb{E}_{\mathcal{S}_k}[\|\nabla\bar{c}(x_k,\xi_k^J)\|^2\big|[\mathcal{S}_{k-1}]]\mathbb{E}_{\mathcal{S}_k}[\|\mathcal{P}(C_k)-\mathcal{P}(c(x_k))\|^2]\big|[\mathcal{S}_{k-1}]]\\
&\leq 6\rho_k^2\mathbb{E}_{[\mathcal{S}_{k-1}]}[\kappa_{\nabla\bar{c}}^2\mathbb{E}_{\mathcal{S}_k}[\|C_k-c(x_k)\|^2\big|[\mathcal{S}_{k-1}]]]\\
&= 6\kappa_{\nabla\bar{c}}^2\rho_k^2\tilde{\delta}_{C_k}.
\end{aligned}\tag{20}$$

Similar to the independence argument for (20), for the last term on the right-hand side of the last inequality in (18), one has

$$\begin{aligned}
&6\rho_k^2\mathbb{E}_{[\mathcal{S}_k]}[(\nabla\bar{c}(x_k,\xi_k^J)-\nabla c(x_k))^T\mathcal{P}(c(x_k))\|^2]\\
&= 6\rho_k^2\mathbb{E}_{[\mathcal{S}_{k-1}]}[\mathbb{E}_{\mathcal{S}_k}[\|\nabla\bar{c}(x_k,\xi_k^J)-\nabla c(x_k)\|^2\|\mathcal{P}(c(x_k))\|^2\big|[\mathcal{S}_{k-1}]]]\\
&\leq 6\rho_k^2\mathbb{E}_{[\mathcal{S}_{k-1}]}[\mathbb{E}_{\mathcal{S}_k}[\delta_{\nabla c}^2\|\mathcal{P}(c(x_k))\|^2\big|[\mathcal{S}_{k-1}]]]\\
&= 6\delta_{\nabla c}^2\rho_k^2\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2].
\end{aligned}\tag{21}$$

Combining (17)–(21), one has

$$\mathbb{E}_{[\mathcal{S}_k]}\left[\|\bar{d}_k\|^2\right] \leq 2\mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\|^2]+6\delta_{\nabla f}^2+6\kappa_{\nabla\bar{c}}^2\rho_k^2\tilde{\delta}_{C_k}+6\delta_{\nabla c}^2\rho_k^2\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2].$$

□

Lemma 2. Under the Assumptions 1 – 4, if the nonnegative sequence $\{\omega_k\}_{k\in\mathbb{N}}$ satisfies

$$\omega_{k+1}\leq\tfrac{1}{2}\text{ and }\omega_{k+1}\geq 72L_{\bar{c}}^2\kappa_{\nabla\bar{c}}^2\eta_k^2\rho_k^2.\tag{22}$$

Then, for any $k\in\mathbb{N}$, one has

$$\begin{aligned}
\tilde{\delta}_{C_{k+1}} \leq{}& (1-\omega_{k+1})\,\tilde{\delta}_{C_k}+3\omega_{k+1}^2\delta_c^2+36\delta_{\nabla f}^2L_{\bar{c}}^2\eta_k^2\\
&+12L_{\bar{c}}^2\eta_k^2\mathbb{E}_{\mathcal{S}_k}[\|\nabla m_{\rho_k}(x_k)\|^2]+36L_{\bar{c}}^2\delta_{\nabla c}^2\eta_k^2\rho_k^2\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2].
\end{aligned}\tag{23}$$

Further, suppose

$$\max\{-1,-2r_\eta\}\leq r_\omega\leq 0.\tag{24}$$

Let scalars $(r_{\tilde{\eta}}, K, M_{\rho c}^K, \tilde{\eta}_1, M_{\tilde{\delta}}^K) \in \mathbb{R}_{\leq 0} \times \mathbb{N} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ satisfy

$$\begin{aligned} r_{\tilde{\eta}} &= \max\{r_\omega, -r_\eta\},\ K \geq k_{\tilde{\delta}} := \left\lceil \sqrt[r_\omega+1]{\tfrac{2(r_\omega - 2r_{\tilde{\eta}})}{\omega_1}} \right\rceil, \\ M_{\rho c}^K &\geq \max_{\forall k \in [K]} \{1, \mathbb{E}_{[\mathcal{S}_k]}[\rho_k^2 \|\mathcal{P}(c(x_k))\|^2]\}, \\ \tilde{\eta}_1^2 &\geq 3\delta_c^2\omega_1^2 + (36\delta_{\nabla f}^2 L_{\bar{c}}^2 + 24L_{\bar{c}}^2\kappa_{\nabla f}^2 + 24L_{\bar{c}}^4 + 36L_{\bar{c}}^2\delta_{\nabla c}^2)\eta_1^2, \text{ and} \\ M_{\tilde{\delta}}^K &\geq \max\{\tilde{\delta}_{C_{k_{\tilde{\delta}}}} k_{\tilde{\delta}}^{r_\omega - 2r_{\tilde{\eta}}}, \tfrac{2\tilde{\eta}_1^2}{\omega_1}\} M_{\rho c}^K. \end{aligned} \tag{25}$$

Then, one has

$$\tilde{\delta}_{C_k} \leq M_{\tilde{\delta}}^K k^{-r_\omega + 2r_{\tilde{\eta}}} \text{ for all } k \in [k_{\tilde{\delta}}, K]. \tag{26}$$

Proof. First, one verifies that $k_{\tilde{\delta}}$ is well-defined when (24) is given. Indeed, from (24), one has

$$r_\omega + 1 \geq 0 \text{ and } r_\omega - 2r_{\tilde{\eta}} = \min\{-r_\omega, r_\omega + 2r_\eta\} \geq 0.$$

Hence, $k_{\tilde{\delta}}$ is well-defined. Second, for simplicity of arguments, in this lemma we use a random vector $v_{k+1} \in \mathbb{R}^m$ given by

$$v_{k+1} := (1 - \omega_{k+1})(c(x_k) - \bar{c}(x_k, \xi_{k+1}^c)) - c(x_{k+1}) + \bar{c}(x_{k+1}, \xi_{k+1}^c).$$

Recall the variance reduction formula $C_{k+1} = \bar{c}(x_{k+1}, \xi_{k+1}^c) + (1 - \omega_{k+1})(C_k - \bar{c}(x_k, \xi_{k+1}^c))$ and by expanding a quadratic function, one has

$$\begin{aligned} &\mathbb{E}_{[\mathcal{S}_{k+1}]}[\|C_{k+1} - c(x_{k+1})\|^2] \\ &= \mathbb{E}_{[\mathcal{S}_{k+1}]}[\|\bar{c}(x_{k+1}, \xi_{k+1}^c) + (1 - \omega_{k+1})(C_k - \bar{c}(x_k, \xi_{k+1}^c)) - c(x_{k+1})\|^2] \\ &= \mathbb{E}_{[\mathcal{S}_{k+1}]}[\|(1 - \omega_{k+1})(C_k - c(x_k)) + v_{k+1}\|^2] \\ &= (1 - \omega_{k+1})^2 \mathbb{E}_{[\mathcal{S}_{k+1}]}[\|C_k - c(x_k)\|^2] + \mathbb{E}_{[\mathcal{S}_{k+1}]}[\|v_{k+1}\|^2] \\ &\quad + 2(1 - \omega_{k+1})\mathbb{E}_{[\mathcal{S}_{k+1}]}[(C_k - c(x_k))^T v_{k+1}]. \end{aligned} \tag{27}$$

Next, we bound the last two terms on the right-hand side of the last equality of (27). For the last term, from the definition of v_{k+1}, the linearity of expectation and conditional

expectation, one has

$$\begin{aligned}
&\mathbb{E}_{[\mathcal{S}_{k+1}]}[(C_k - c(x_k))^T v_{k+1}] \\
&= (1-\omega_{k+1})\mathbb{E}_{[\mathcal{S}_{k+1}]}[(C_k - c(x_k))^T (c(x_k) - \bar{c}(x_k, \xi_{k+1}^c)] \\
&\quad + \mathbb{E}_{[\mathcal{S}_{k+1}]}[(C_k - c(x_k))^T (\bar{c}(x_{k+1}, \xi_{k+1}^c) - c(x_{k+1})] \\
&= (1-\omega_{k+1})\mathbb{E}_{[\mathcal{S}_k]}[\mathbb{E}_{\mathcal{S}_{k+1}}[(C_k - c(x_k))^T (c(x_k) - \bar{c}(x_k, \xi_{k+1}^c)] \big| [\mathcal{S}_k]] \\
&\quad + \mathbb{E}_{[\mathcal{S}_k]}[\mathbb{E}_{\mathcal{S}_{k+1}}[[(C_k - c(x_k))^T (\bar{c}(x_{k+1}, \xi_{k+1}^c) - c(x_{k+1}) \big| [\mathcal{S}_k]] \\
&= (1-\omega_{k+1})\mathbb{E}_{[\mathcal{S}_k]}[(C_k - c(x_k))^T (c(x_k) - c(x_k)] \big| [\mathcal{S}_k]] \\
&\quad + \mathbb{E}_{[\mathcal{S}_k]}[[(C_k - c(x_k))^T (c(x_{k+1}) - c(x_{k+1}) \big| [\mathcal{S}_k]] \\
&= 0.
\end{aligned} \tag{28}$$

For the second-to-last term on the right-hand side of the last equality of (27), by rearranging terms, (16) with $l = 3$, Assumption 4 and Lemma 1, one has

$$\begin{aligned}
\mathbb{E}_{[\mathcal{S}_{k+1}]}[\|v_{k+1}\|^2] &= \mathbb{E}_{[\mathcal{S}_{k+1}]}[\|(c(x_k) - c(x_{k+1})) - \omega_{k+1}(c(x_k) - \bar{c}(x_k, \xi_{k+1}^c)) \\
&\qquad - (\bar{c}(x_k, \xi_{k+1}^c) - \bar{c}(x_{k+1}, \xi_{k+1}^c))\|^2] \\
&\le 3\mathbb{E}_{[\mathcal{S}_{k+1}]}[\|c(x_k) - c(x_{k+1})\|^2] \\
&\quad + 3\omega_{k+1}^2 \mathbb{E}_{[\mathcal{S}_{k+1}]}[\|c(x_k) - \bar{c}(x_k, \xi_{k+1}^c)\|^2] \\
&\quad + 3\mathbb{E}_{[\mathcal{S}_{k+1}]}[\|\bar{c}(x_k, \xi_{k+1}^c)) - \bar{c}(x_{k+1}, \xi_{k+1}^c)\|^2] \\
&\le 3L_{\bar{c}}^2 \mathbb{E}_{[\mathcal{S}_{k+1}]}[\|x_k - x_{k+1}\|^2] \\
&\quad + 3\omega_{k+1}^2 \mathbb{E}_{[\mathcal{S}_k]}[\mathbb{E}_{\mathcal{S}_{k+1}}[\|c(x_k) - \bar{c}(x_k, \xi_{k+1}^c)\|^2 \big| [\mathcal{S}_k]] \\
&\quad + 3L_{\bar{c}}^2 \mathbb{E}_{[\mathcal{S}_{k+1}]}[\|x_k - x_{k+1}\|^2] \\
&\le 6L_{\bar{c}}^2 \eta_k^2 \mathbb{E}_{[\mathcal{S}_{k+1}]}[\|\bar{d}_k\|^2] + 3\omega_{k+1}^2 \delta_c^2.
\end{aligned} \tag{29}$$

Combining (27), (28), (29) and (9), one has

$$\begin{aligned}
\tilde{\delta}_{C_{k+1}} &= \mathbb{E}_{[\mathcal{S}_{k+1}]}[\|C_{k+1} - c(x_{k+1})\|^2] \\
&\le (1-\omega_{k+1})^2 \mathbb{E}_{[\mathcal{S}_{k+1}]}[\|C_k - c(x_k)\|^2] + 6L_{\bar{c}}^2 \eta_k^2 \mathbb{E}_{[\mathcal{S}_{k+1}]}[\|\bar{d}_k\|^2] + 3\omega_{k+1}^2 \delta_c^2 \\
&\le (1-\omega_{k+1})^2 \tilde{\delta}_{C_k} + 3\omega_{k+1}^2 \delta_c^2 + 36\delta_{\nabla f}^2 L_{\bar{c}}^2 \eta_k^2 + 36L_{\bar{c}}^2 \kappa_{\nabla \bar{c}}^2 \eta_k^2 \rho_k^2 \tilde{\delta}_{C_k} \\
&\quad + 12L_{\bar{c}}^2 \eta_k^2 \mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\|^2] + 36L_{\bar{c}}^2 \delta_{\nabla c}^2 \eta_k^2 \rho_k^2 \mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2] \\
&= \left((1-\omega_{k+1})^2 + 36L_{\bar{c}}^2 \kappa_{\nabla \bar{c}}^2 \eta_k^2 \rho_k^2\right) \tilde{\delta}_{C_k} + 3\omega_{k+1}^2 \delta_c^2 + 36\delta_{\nabla f}^2 L_{\bar{c}}^2 \eta_k^2 \\
&\quad + 12L_{\bar{c}}^2 \eta_k^2 \mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\|^2] + 36L_{\bar{c}}^2 \delta_{\nabla c}^2 \eta_k^2 \rho_k^2 \mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2].
\end{aligned} \tag{30}$$

Lastly, from (22), one has

$$\begin{aligned}
(1-\omega_{k+1})^2 + 36L_{\bar{c}}^2 \kappa_{\nabla \bar{c}}^2 \eta_k^2 \rho_k^2 &= 1 - 2\omega_{k+1} + \omega_{k+1}^2 + 36L_{\bar{c}}^2 \kappa_{\nabla \bar{c}}^2 \eta_k^2 \rho_k^2 \\
&\le 1 - \omega_{k+1}.
\end{aligned}$$

Combining with (30), one has (23) as desired.

Next, one aims to prove (26). Before that, one derives a useful inequality. That is, for any $(\kappa, y) \in (0,\infty)\times(0,2)$, it holds

$$(1+y)^{-\kappa} \geq 1-\kappa y. \tag{31}$$

To see this, consider the scalar function $g(y) = (1+y)^{-\kappa}+\kappa y$ in the domain $(0,2)$. The derivative of g is $g'(y) = \kappa - \kappa(1+y)^{-\kappa-1}$, which is an increasing function in $(0,2)$ and gives that $g'(y) \geq g'(0) = \kappa - \kappa(1+0)^{-\kappa-1} = 0$ for all $y \in (0,2)$. Hence, we have that the function g is nondecreasing in $(0,2)$, i.e., $g(y) = (1+y)^{-\kappa}+\kappa y \geq g(0) = 1$ for all $x \in (0,2)$. Rearranging terms gives (31).

Combining submultiplicity, inequality (16) with $l=2$, and Assumption 2, one has

$$\begin{aligned}\|\nabla m_{\rho_k}(x_k)\|^2 &\leq \|\nabla f(x_k) + \rho_k[\nabla c(x_k)]^T\mathcal{P}(c(x_k))\|^2\\ &\leq 2\|\nabla f(x_k)\|^2 + 2\rho_k^2\|[\nabla c(x_k)]^T\mathcal{P}(c(x_k))\|^2\\ &\leq 2\kappa_{\nabla f}^2 + 2L_{\bar{c}}^2\rho_k^2\|\mathcal{P}(c(x_k))\|^2.\end{aligned} \tag{32}$$

Combining (32) with (23) – (25), one has the following for all $k \in [k_{\tilde{\delta}}, K]$

$$\begin{aligned}\tilde{\delta}_{C_{k+1}} &\leq (1-\omega_{k+1})\,\tilde{\delta}_{C_k} + 3\omega_{k+1}^2\delta_c^2\\ &\quad + (36\delta_{\nabla f}^2 L_{\bar{c}}^2 + 12L_{\bar{c}}^2(2\kappa_{\nabla f}^2 + 2L_{\bar{c}}^2 M_{\rho c}^K) + 36L_{\bar{c}}^2\delta_{\nabla c}^2 M_{\rho c}^K)\eta_k^2\\ &= (1-\omega_{k+1})\,\tilde{\delta}_{C_k} + 3\delta_c^2\omega_1^2 k^{2r_\omega}\\ &\quad + (36\delta_{\nabla f}^2 L_{\bar{c}}^2 + 12L_{\bar{c}}^2(2\kappa_{\nabla f}^2 + 2L_{\bar{c}}^2 M_{\rho c}^K) + 36L_{\bar{c}}^2\delta_{\nabla c}^2 M_{\rho c}^K)\eta_1^2 k^{-2r_\eta}\\ &\leq (1-\omega_{k+1})\,\tilde{\delta}_{C_k} + 3\delta_c^2\omega_1^2 M_{\rho c}^K k^{2r_\omega}\\ &\quad + (36\delta_{\nabla f}^2 L_{\bar{c}}^2 M_{\rho c}^K + 12L_{\bar{c}}^2(2\kappa_{\nabla f}^2 M_{\rho c}^K + 2L_{\bar{c}}^2 M_{\rho c}^K) + 36L_{\bar{c}}^2\delta_{\nabla c}^2 M_{\rho c}^K)\eta_1^2 k^{-2r_\eta}\\ &\leq (1-\omega_{k+1})\,\tilde{\delta}_{C_k} + \tilde{\eta}_1^2 M_{\rho c}^K k^{2r_{\tilde{\eta}}},\end{aligned} \tag{33}$$

where the second-to-last inequality uses $M_{\rho c}^K \geq 1$. Next, one proves (26) by induction, with the base case $k = k_{\tilde{\delta}}$. For the base case, from the requirement of $M_{\tilde{\delta}}$ in (25), one has

$$\tilde{\delta}_{C_k} = \tilde{\delta}_{C_{k_{\tilde{\delta}}}} \leq \tilde{\delta}_{C_{k_{\tilde{\delta}}}} M_{\rho c}^K \leq \max\{\tilde{\delta}_{C_{k_{\tilde{\delta}}}} k_{\tilde{\delta}}^{r_\omega - 2r_{\tilde{\eta}}}, \tfrac{2\tilde{\eta}_1^2}{\omega_1}\} M_{\rho c}^K k^{-r_\omega+2r_{\tilde{\eta}}} = M_{\tilde{\delta}}^K k^{-r_\omega+2r_{\tilde{\eta}}}.$$

Hence, the base case holds. By induction, one assumes that there exists a $k \geq k_{\tilde{\delta}}$ such that (26) holds. Then, consider the value of $\tilde{\delta}_{k+1}$. Combining (26) and that $M_{\tilde{\delta}}^K \geq \frac{2\tilde{\eta}_1^2 M_{\rho c}^K}{\omega_1}$ in (25), one has

$$\begin{aligned}\tilde{\delta}_{k+1} &\leq (1-\omega_1 k^{r_\omega})\,\tilde{\delta}_k + \tilde{\eta}_1^2 M_{\rho c}^K k^{2r_{\tilde{\eta}}}\\ &\leq (1-\omega_1 k^{r_\omega})\, M_{\tilde{\delta}}^K k^{-r_\omega+2r_{\tilde{\eta}}} + \tilde{\eta}_1^2 M_{\rho c}^K k^{2r_{\tilde{\eta}}}\\ &= M_{\tilde{\delta}}^K k^{-r_\omega+2r_{\tilde{\eta}}} - \left(\omega_1 - \frac{\tilde{\eta}_1^2 M_{\rho c}^K}{M_{\tilde{\delta}}^K}\right) M_{\tilde{\delta}}^K k^{2r_{\tilde{\eta}}}\\ &\leq M_{\tilde{\delta}}^K k^{-r_\omega+2r_{\tilde{\eta}}} - \tfrac{1}{2}\omega_1 M_{\tilde{\delta}}^K k^{2r_{\tilde{\eta}}}.\end{aligned} \tag{34}$$

Recall that it is assumed $r_\omega - 2r_{\tilde{\eta}} \geq 0$, and we have that $\frac{1}{k} \in (0, 2)$ for all $k \geq 1$. Combining with (31), one has

$$\begin{aligned} M_{\tilde{\delta}}^K (k+1)^{-r_\omega + 2r_{\tilde{\eta}}} &= M_{\tilde{\delta}}^K k^{-r_\omega + 2r_{\tilde{\eta}}} \left(1 + \tfrac{1}{k}\right)^{-r_\omega + 2r_{\tilde{\eta}}} \\ &\geq M_{\tilde{\delta}}^K k^{-r_\omega + 2r_{\tilde{\eta}}} \left(1 + \tfrac{-r_\omega + 2r_{\tilde{\eta}}}{k}\right) \\ &= M_{\tilde{\delta}}^K k^{-r_\omega + 2r_{\tilde{\eta}}} - (r_\omega - 2r_{\tilde{\eta}}) M_{\tilde{\delta}}^K k^{-r_\omega + 2r_{\tilde{\eta}} - 1}. \end{aligned} \tag{35}$$

On the other hand, since $k \geq k_{\tilde{\delta}}$ and $r_\omega + 1 \geq 0$, one has

$$\begin{aligned} k \geq \left\lceil \sqrt[r_\omega+1]{\tfrac{2(r_\omega - 2r_{\tilde{\eta}})}{\omega_1}} \right\rceil &\iff \tfrac{\omega_1}{2(r_\omega - 2r_{\tilde{\eta}})} \geq k^{-r_\omega - 1} \\ &\iff \tfrac{1}{2}\omega_1 M_{\tilde{\delta}}^K k^{2r_{\tilde{\eta}}} \geq (r_\omega - 2r_{\tilde{\eta}}) M_{\tilde{\delta}}^K k^{-r_\omega + 2r_{\tilde{\eta}} - 1}. \end{aligned} \tag{36}$$

Combining (34) – (36), one has

$$\tilde{\delta}_{k+1} \leq M_{\tilde{\delta}}^K (k+1)^{-r_\omega + 2r_{\tilde{\eta}}}.$$

The induction is complete. □

Remark 2. We have a few observations for this lemma.

1. The result in (23) is *not* tight when there is no noise on c, due to the analysis strategy used in (29). That is, instead of having the following bound for $\mathbb{E}_{[\mathcal{S}_{k+1}]}[\|v_{k+1}\|^2]$

$$\begin{aligned} &\mathbb{E}_{[\mathcal{S}_{k+1}]}[\|v_{k+1}\|^2] \\ =&\mathbb{E}_{[\mathcal{S}_{k+1}]}[\|(c(x_k) - c(x_{k+1})) - \omega_{k+1}(c(x_k) - \bar{c}(x_k, \xi_{k+1}^c)) - (\bar{c}(x_k, \xi_{k+1}^c) - \bar{c}(x_{k+1}, \xi_{k+1}^c))\|^2] \\ \leq&3\mathbb{E}_{[\mathcal{S}_{k+1}]}[\|c(x_k) - \bar{c}(x_k, \xi_{k+1}^c))\|^2] + 3\omega_{k+1}^2 \mathbb{E}_{[\mathcal{S}_{k+1}]}[\|c(x_k) - \bar{c}(x_k, \xi_{k+1}^c)\|^2] \\ &\quad + 3\mathbb{E}_{[\mathcal{S}_{k+1}]}[\|c(x_{k+1}) - \bar{c}(x_{k+1}, \xi_{k+1}^c)\|^2] \\ \leq&6\delta_c^2 + 3\delta_c^2, \text{ which is zero when } \delta_c = 0. \end{aligned} \tag{37}$$

   It uses what is shown in (29). The analysis strategy used in (29) is standard in variance-reduction type methods, and indeed it gives better bounds when $\delta_c \neq 0$.

2. It is not a conventional approach to obtain (26) by giving a constant bound for $\mathbb{E}_{[\mathcal{S}_k]}[\rho_k^2 \|\mathcal{P}(c(x_k))\|^2]$, $\mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\|^2]$ and applying it later. However, we remark that such handling here does not lead to a relaxed result as long as $\delta_{\nabla c} \neq 0$ or $\delta_{\nabla f} \neq 0$. One can observe that when $\delta_{\nabla c} \neq 0$ or $\delta_{\nabla f} \neq 0$, the right-hand side of (23) is around

$$(1 - \omega_{k+1})\, \tilde{\delta}_{C_k} + 3\omega_{k+1}^2 \delta_c^2 + \mathcal{O}(\eta_k^2) + 12 L_{\bar{c}}^2 \eta_k^2 \mathbb{E}_{\mathcal{S}_k}[\|\nabla m_{\rho_k}(x_k)\|^2]. \tag{38}$$

Hence, bounding $\mathbb{E}_{\mathcal{S}_k}[\|\nabla m_{\rho_k}(x_k)\|^2]$ by a constant does not give a worse result. However, when $\delta_{\nabla c}=0$ and $\delta_{\nabla f}=0$, i.e., the right-hand side of (23) is

$$(1-\omega_{k+1})\,\tilde{\delta}_{C_k}+3\omega_{k+1}^2\delta_c^2+12L_{\bar{c}}^2\eta_k^2\mathbb{E}_{\mathcal{S}_k}[\|\nabla m_{\rho_k}(x_k)\|^2],$$

bounding $\mathbb{E}_{\mathcal{S}_k}[\|\nabla m_{\rho_k}(x_k)\|^2]$ by a constant will give a worse result. One should use the conventional *potential function* approach to derive a better complexity result, which is explained after Lemma 4.

3. From the above point, one can draw a conclusion that when $\delta_{\nabla c}\neq 0$, $\delta_{\nabla c}\neq 0$ and $\delta_{\nabla f}\neq 0$, performing variance reduction separately to $(c,\nabla f)$ or to $(c,\nabla c)$ does not give a better complexity.

Lemma 3. Under the same assumptions as assumed in Lemma 1. In addition, suppose the sequencs $\{\eta_k\}_{k\in\mathbb{N}},\{\rho_k\}_{k\in\mathbb{N}}$ satisfy

$$\rho_k\eta_k\leq\min\{\tfrac{1}{4L_{\nabla m_1}},\tfrac{L_{\bar{c}}^2}{3L_{\nabla m_1}\kappa_{\nabla\bar{c}}^2}\}\text{ for all }k\in\mathbb{N}. \tag{39}$$

Then, one has

$$\begin{aligned}&\mathbb{E}_{[\mathcal{S}_{k+1}]}\left[m_{\rho_{k+1}}(x_{k+1})\right]\\ \leq\;&\mathbb{E}_{[\mathcal{S}_k]}[m_{\rho_k}(x_k)]-\tfrac{1}{2}\eta_k\mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\|^2]+3L_{\nabla m_1}\delta_{\nabla f}^2\eta_k^2\rho_k\\ &+3L_{\bar{c}}^2\eta_k\rho_k^2\tilde{\delta}_{C_k}+(2^\beta\beta\tfrac{\rho_k}{k}+3L_{\nabla m_1}\delta_{\nabla c}^2\eta_k^2\rho_k^3)\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2].\end{aligned} \tag{40}$$

Proof. We first explain that (39) is compatible with (5). Note that (5) requires $\eta_k\leq 1\leq\rho_k$. One can first choose any $\rho_1\geq 1$, then fix ρ_1, and choose an η_1 that is positive but no larger than the real number one and no larger than the right-hand side of (39) divided by ρ_1.

Recall the update $x_{k+1}=x_k-\eta_k\bar{d}_k$ and the function m_{ρ_k} has $L_{\nabla m_1}\rho_k$-Lipschitz continous gradient. Taking the total expectation on both sides of the Taylor's expansion of the merit function m_{ρ_k} at x_k, one has

$$\begin{aligned}&\mathbb{E}_{[\mathcal{S}_k]}[m_{\rho_k}(x_{k+1})]\\ =\;&\mathbb{E}_{[\mathcal{S}_k]}[m_{\rho_k}\left(x_k-\eta_k\bar{d}_k\right)]\\ \leq\;&\mathbb{E}_{[\mathcal{S}_k]}[m_{\rho_k}(x_k)]-\eta_k\mathbb{E}_{[\mathcal{S}_k]}[[\nabla m_{\rho_k}(x_k)]^T\bar{d}_k]+\tfrac{L_{\nabla m_1}\rho_k\eta_k^2}{2}\mathbb{E}_{[\mathcal{S}_k]}[\left\|\bar{d}_k\right\|^2].\end{aligned} \tag{41}$$

Recall that $\bar{d}_k=\nabla\overline{f}(x_k,\xi_k^f)+\rho_k[\nabla\bar{c}(x_k,\xi_k^J)]^T\mathcal{P}(C_k)$. And from a similar argument for (20), one has the random vectors $\nabla\overline{f}(x_k,\xi_k^f),\nabla\bar{c}(x_k,\xi_k^J),\mathcal{P}(C_k)$ being independent when conditioned on x_1 and $[\mathcal{S}_{k-1}]$. From conditional expectation, the aforementioned independence and Assumption 3, one has

$$\begin{aligned}&-\eta_k\mathbb{E}_{[\mathcal{S}_k]}[[\nabla m_{\rho_k}(x_k)]^T\bar{d}_k]\\ =&-\eta_k\mathbb{E}_{[\mathcal{S}_k]}[[\nabla m_{\rho_k}(x_k)]^T(\nabla\overline{f}(x_k,\xi_k^f)+\rho_k[\nabla\bar{c}(x_k,\xi_k^J)]^T\mathcal{P}(C_k))]\\ =&-\eta_k\mathbb{E}_{[\mathcal{S}_{k-1}]}[\mathbb{E}_{\mathcal{S}_k}[[\nabla m_{\rho_k}(x_k)]^T(\nabla\overline{f}(x_k,\xi_k^f)+\rho_k[\nabla\bar{c}(x_k,\xi_k^J)]^T\mathcal{P}(C_k))\big|[\mathcal{S}_{k-1}]]]\\ =&-\eta_k\mathbb{E}_{[\mathcal{S}_k]}[[\nabla m_{\rho_k}(x_k)]^T(\nabla f(x_k)+\rho_k[\nabla c(x_k)]^T\mathcal{P}(C_k))].\end{aligned} \tag{42}$$

By adding and subtracting a term $\rho_k[\nabla c(x_k)]^T c(x_k)$ on the last line of (42), one has

$$\begin{aligned}
&-\eta_k \mathbb{E}_{[\mathcal{S}_k]}[[\nabla m_{\rho_k}(x_k)]^T \bar{d}_k] \\
=&-\eta_k \mathbb{E}_{[\mathcal{S}_k]}[[\nabla m_{\rho_k}(x_k)]^T (\nabla f(x_k) + \rho_k[\nabla c(x_k)]^T \mathcal{P}(c(x_k)))] \\
&\quad -\eta_k \mathbb{E}_{[\mathcal{S}_k]}[[\nabla m_{\rho_k}(x_k)]^T (\rho_k[\nabla c(x_k)]^T (\mathcal{P}(C_k) - \mathcal{P}(c(x_k))))] \\
=&-\eta_k \mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\|^2] - \eta_k \rho_k \mathbb{E}_{[\mathcal{S}_k]}[[\nabla m_{\rho_k}(x_k)]^T [\nabla c(x_k)]^T (\mathcal{P}(C_k) - \mathcal{P}(c(x_k)))].
\end{aligned} \tag{43}$$

From submultiplicity, $\|\nabla c(x_k)\| \leq L_{\bar{c}}$ and $\mathcal{P}$ is 1-Lipschit continous shown in Lemma 1. Combining with Young's inequality, for the last term on the last line of (43), one has

$$\begin{aligned}
&-\eta_k \rho_k \mathbb{E}_{[\mathcal{S}_k]}[\nabla m_{\rho_k}(x_k)]^T [\nabla c(x_k)]^T (\mathcal{P}(C_k) - \mathcal{P}(c(x_k)))] \\
\leq\ & \eta_k \rho_k \mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\| \|\nabla c(x_k)\| \|\mathcal{P}(C_k) - \mathcal{P}(c(x_k))\|] \\
\leq\ & \eta_k \rho_k L_{\bar{c}} \mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\| \|C_k - c(x_k)\|] \\
\leq\ & \tfrac{1}{4}\eta_k \mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\|^2] + 2L_{\bar{c}}^2 \eta_k \rho_k^2 \mathbb{E}_{[\mathcal{S}_k]}[\|C_k - c(x_k)\|^2].
\end{aligned} \tag{44}$$

On the other hand, from the definition of the merit function, one has

$$\mathbb{E}_{[\mathcal{S}_k]}[m_{\rho_{k+1}}(x_{k+1})] = \mathbb{E}_{[\mathcal{S}_k]}\left[m_{\rho_k}(x_{k+1})\right] + \tfrac{1}{2}(\rho_{k+1} - \rho_k)\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2]. \tag{45}$$

Next, we give an upper bound for $\rho_{k+1} - \rho_k$. Consider an auxiliary function $g : \mathbb{R}_{>0} \to \mathbb{R}$ defined as $g(y) = y^\beta$. From the mean value theorem, there exists a scalar $\tilde{k} \in [k, k+1]$ such that

$$(k+1)^\beta - k^\beta = \frac{(k+1)^\beta - k^\beta}{(k+1) - k} = \left.\frac{dg}{dy}\right|_{y=\tilde{k}} = \beta \tilde{k}^{\beta-1}.$$

When $0 < \beta \leq 1$, from the fact that the function $y^{\beta-1}$ is nonincreasing, one has

$$\beta \tilde{k}^{\beta-1} \leq \beta k^{\beta-1} \leq 2^{\beta+1} \beta k^{\beta-1}.$$

When $\beta > 1$, from the fact that the function $y^{\beta-1}$ is nondecreasing and that $k+1 \leq 2k$ for all $k \in \mathbb{N}$, one has

$$\beta \tilde{k}^{\beta-1} \leq \beta (k+1)^{\beta-1} \leq \beta (2k)^{\beta-1} \leq 2^{\beta+1} \beta k^{\beta-1}.$$

Combining both cases for the value of β, we have $(k+1)^\beta - k^\beta \leq 2^{\beta+1} \beta k^{\beta-1}$ and

$$\tfrac{1}{2}(\rho_{k+1} - \rho_k) = \tfrac{1}{2}\rho_1 \left((k+1)^\beta - k^\beta\right) \leq 2^\beta \beta k^{\beta-1} \rho_1 = 2^\beta \beta \tfrac{\rho_k}{k}. \tag{46}$$

Combining (41) – (46) and the bound for $\mathbb{E}_{[\mathcal{S}_k]}[\|\bar{d}_k\|^2]$ in Lemma 1, one has

$$\begin{aligned}
&\mathbb{E}_{[\mathcal{S}_k]}\left[m_{\rho_{k+1}}(x_{k+1})\right] \\
&= \mathbb{E}_{[\mathcal{S}_k]}[m_{\rho_k}(x_{k+1})] + \tfrac{1}{2}(\rho_{k+1}-\rho_k)\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2] \\
&\leq \mathbb{E}_{[\mathcal{S}_k]}[m_{\rho_k}(x_k)] - \tfrac{3}{4}\eta_k\mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\|^2] + 2L_{\bar{c}}^2\eta_k\rho_k^2\tilde{\delta}_{C_k} \\
&\qquad + \tfrac{L_{\nabla m_1}\rho_k\eta_k^2}{2}\mathbb{E}_{[\mathcal{S}_k]}[\|\bar{d}_k\|^2] + \tfrac{1}{2}(\rho_{k+1}-\rho_k)\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2] \\
&\leq \mathbb{E}_{[\mathcal{S}_k]}[m_{\rho_k}(x_k)] - \tfrac{3}{4}\eta_k\mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\|^2] + 2L_{\bar{c}}^2\eta_k\rho_k^2\tilde{\delta}_{C_k} \\
&\qquad + 2^\beta\beta\tfrac{\rho_k}{k}\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2] + \tfrac{L_{\nabla m_1}\rho_k\eta_k^2}{2}(2\mathbb{E}_{\mathcal{S}_k}[\|\nabla m_{\rho_k}(x_k)\|^2] \\
&\qquad + 6\delta_{\nabla f}^2 + 6\kappa_{\nabla\bar{c}}^2\rho_k^2\tilde{\delta}_{C_k} + 6\delta_{\nabla c}^2\rho_k^2\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2]) \\
&= \mathbb{E}_{[\mathcal{S}_k]}[m_{\rho_k}(x_k)] - (\tfrac{3}{4} - L_{\nabla m_1}\rho_k\eta_k)\eta_k\mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\|^2] \\
&\qquad + 3L_{\nabla m_1}\delta_{\nabla f}^2\eta_k^2\rho_k + (2L_{\bar{c}}^2 + 3L_{\nabla m_1}\kappa_{\nabla\bar{c}}^2\eta_k\rho_k)\eta_k\rho_k^2\tilde{\delta}_{C_k} \\
&\qquad + (2^\beta\beta\tfrac{\rho_k}{k} + 3L_{\nabla m_1}\delta_{\nabla c}^2\eta_k^2\rho_k^3)\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2]).
\end{aligned} \tag{47}$$

From (39), one has

$$\begin{aligned}
&\tfrac{3}{4} - L_{\nabla m_1}\rho_k\eta_k \geq \tfrac{1}{2} \\
&2L_{\bar{c}}^2 + 3L_{\nabla m_1}\kappa_{\nabla\bar{c}}^2\eta_k\rho_k \leq 3L_{\bar{c}}^2
\end{aligned} \tag{48}$$

Since the iterate x_{k+1} only depends on the initial point x_1 and samples $[\mathcal{S}_k]$, one has $\mathbb{E}_{[\mathcal{S}_k]}\left[m_{\rho_{k+1}}(x_{k+1})\right] = \mathbb{E}_{[\mathcal{S}_{k+1}]}\left[m_{\rho_{k+1}}(x_{k+1})\right]$. Combining with (47) and (82), one has (40) as desired.

□

Next, we are going to show that there exists a choice for the sequences $\{\eta_k\}_k, \{\rho_k\}_k, \{\omega_k\}_k$ such that one is able to obtain the *lower-bound* worst-case complexity. To focus our attention on arguing the *lower-bound* worst-case complexity, we make a heruistic regarding $\mathbb{E}_{[\mathcal{S}_k]}[\|c(x_k)\|^2]$.

Lemma 4. Under Assumptions 1 – 4. In addition, suppose the scalars $(\eta_1, \rho_1, \omega_1) \in \mathbb{R}^3_{>0}$ satisfy

$$\begin{cases}
\rho_1 \geq 1,\ \eta_1 \leq 1,\ \omega_1 = 72L_{\bar{c}}^2\kappa_{\nabla\bar{c}}^2\eta_1^2\rho_1^2 \text{ and} \\
\eta_1\rho_1 \leq \min\{\frac{1}{4L_{\nabla m_1}}, \frac{L_{\bar{c}}^2}{3L_{\nabla m_1}\kappa_{\nabla\bar{c}}^2}, \frac{1}{12L_{\bar{c}}\kappa_{\nabla\bar{c}}}\}.
\end{cases} \tag{49}$$

And assume there exists a scalar $M'_{\rho c} > 0$ such that

$$M_{\rho c}^K \leq M'_{\rho c} \text{ for all } K \in \mathbb{N}. \tag{50}$$

Then, the scalars

$$(r_\eta, r_\rho, r_\omega) = (\tfrac{2}{3}, \tfrac{1}{6}, -\tfrac{2}{3}) \tag{51}$$

is the optimal choice such that inequalities (5), (22), (24), (39) hold, and the point $x_{\hat{K}}$ is an ϵ-stochastic stationary point of (1) with the lower bound worse-case sample complexity, when $\{\mathbb{E}_{[\mathcal{S}_k]}[m_{\rho_k}(x_k)]\}_{k\in\mathbb{N}}$ is bounded below. Specifically, in at most $\mathcal{O}(\epsilon^{-6})$ number of iterations and with access of at most $\mathcal{O}(\epsilon^{-6})$ single datapoints, $x_{\hat{K}}$ will satisfy

$$\mathbb{E}_{[\mathcal{S}_K]}[\|\nabla f(x_{\hat{K}}) + \nabla c(x_{\hat{K}})^T \lambda_{\hat{K}}\| + \|\mathcal{P}(c(x_{\hat{K}}))\| + \|\lambda_{\hat{K}}^T \mathcal{P}(c(x_{\hat{K}}))\|] \leq \epsilon,$$

where $\lambda_{\hat{K}} = \rho_{\hat{K}} \mathcal{P}(c(x_{\hat{K}})) \in \mathbb{R}^m$.

Proof. One first verifies or gives conditions under which the sequences $\{\eta_k\}_{k\in\mathbb{N}}$, $\{\rho_k\}_{k\in\mathbb{N}}$, $\{\omega_{k+1}\}_{k\in\mathbb{N}}$ satisfy (49) will give the inequalities (5), (22), (24) and (39). Recall that those inequalities are

$$\rho_1 \geq 1 \geq \eta_1,\ \omega_{k+1} \geq 72 L_{\bar{c}}^2 \kappa_{\nabla\bar{c}}^2 \eta_k^2 \rho_k^2,\ \omega_{k+1} \leq \tfrac{1}{2}, \tag{52}$$

$$\eta_k \rho_k \leq \min\{\tfrac{1}{4L_{\nabla m_1}}, \tfrac{L_{\bar{c}}^2}{3L_{\nabla m_1}\kappa_{\nabla\bar{c}}^2}\} \text{ and } \max\{-1, -2r_\eta\} \leq r_\omega \leq 0. \tag{53}$$

One can observe that the first two inequalities in the first line of (49) give the first inequality in (52). The second inequality in (52) is

$$\omega_{k+1} = \omega_1 k^{r_\omega} \geq 72 L_{\bar{c}}^2 \kappa_{\nabla\bar{c}}^2 \eta_k^2 \rho_k^2 = 72 L_{\bar{c}}^2 \kappa_{\nabla\bar{c}}^2 \eta_1^2 \rho_1^2 k^{-2r_\eta + 2r_\rho}, \tag{54}$$

which, when $k = 1$, requires $\omega_2 = \omega_1 \geq 72 L_{\bar{c}}^2 \kappa_{\nabla\bar{c}}^2 \eta_1^2 \rho_1^2$, and is satisfied by the last equation in the first line of (49). In order for (54) to hold for large k, one requires

$$r_\omega \geq -2r_\eta + 2r_\rho, \tag{55}$$

and one can verify that with the last equation in the first line of (49) and the inequality (55), one has (54) hold for all k. To ensure the third inequality in (52) hold for large k, one at least require

$$r_\omega \leq 0, \tag{56}$$

and one can verify that with (49), it holds for all k that

$$\omega_{k+1} \leq \omega_1 = 72 L_{\bar{c}}^2 \kappa_{\nabla\bar{c}}^2 \eta_1^2 \rho_1^2 \leq \tfrac{1}{2}.$$

Via similar arguments, one has that in order for the first inequality in (53) to hold, one requires

$$r_\eta + r_\rho \leq 0, \tag{57}$$

which with (49), it holds for all k. The last inequality in (53) gives a requirement for $(r_\eta, r_\rho, r_\omega)$ explicitly. Combining (55) – (57) and the last inequality in (53), the inequalities (52) and (53) hold for large k requires

$$r_\omega \leq 0, r_\omega \geq -2r_\eta + 2r_\rho, -r_\eta + r_\rho \leq 0, \text{ and } \max\{-1, -2r_\eta\} \leq r_\omega \leq 0, \tag{58}$$

which, combined with (49), ensures (52) and (53) to hold for all k.

Suppose (58) is satisfied for scalars $(r_\eta, r_\rho, r_\omega)$ used here, and in fact, one can verify that (51) satisfies (58). As a result, the assumptions and requirements for Lemma 1 – Lemma 3 are assumed and required here. Next, one applies Lemma 3. From the fact that the seqeunce $\{\eta_k\}_{k\in\mathbb{N}}$ is nonincreasing and that the output $x_{\hat{K}}$ is uniformally chosen from $\{x_k\}_{k=1}^K$, one has

$$\begin{aligned}\sum_{k=1}^K \tfrac{1}{2}\eta_k \mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)]\|^2] &\geq \tfrac{1}{2}\eta_K \sum_{k=1}^K \mathbb{E}_{[\mathcal{S}_K]}\left[\|\nabla m_{\rho_k}(x_k)\|^2\right] \\ &= \tfrac{1}{2}K\eta_K\left(\frac{1}{K}\sum_{k=1}^K \mathbb{E}_{[\mathcal{S}_{K-1}]}[\|\nabla m_{\rho_k}(x_k)\|^2]\right) \\ &= \tfrac{1}{2}\eta_1 K^{1-r_\eta}\mathbb{E}_{\hat{K}}[\|\nabla m_{\rho_{\hat{K}}}(x_{\hat{K}})\|^2].\end{aligned} \tag{59}$$

Combining (50) wit the definition of $M_{\rho c}^K$ in (25), one has

$$\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2] \leq \tfrac{M'_{\rho c}}{\rho_k^2} \text{ for all } k \in \mathbb{N}. \tag{60}$$

Combining (60) with the inequality (40) in Lemma 3. Summing from $k = 1$ to K, rearranging terms, combining with (59) and multiplying $\frac{2K^{r_\eta-1}}{\eta_1}$ one both sides, one has

$$\begin{aligned}&\mathbb{E}_{\hat{K}}[\|\nabla m_{\rho_{\hat{K}}}(x_{\hat{K}})\|^2] \\ \leq&(\mathbb{E}_{[\mathcal{S}_1]}\left[m_{\rho_1}(x_1)\right] - \mathbb{E}_{[\mathcal{S}_{K+1}]}\left[m_{\rho_{K+1}}(x_{K+1})\right])\tfrac{2}{\eta_1}K^{r_\eta-1} &(61)\\ &+ \tfrac{6L_{\nabla m_1}\delta_{\nabla f}^2}{\eta_1}K^{r_\eta-1}\sum_{k=1}^K \eta_k^2\rho_k + \tfrac{6L_{\tilde{c}}^2}{\eta_1}K^{r_\eta-1}\sum_{k=1}^K \eta_k\rho_k^2\tilde{\delta}_{C_k} &(62)\\ &+ \tfrac{2^{\beta+1}\beta M_{\rho c}}{\eta_1}K^{r_\eta-1}\sum_{k=1}^K \tfrac{1}{k\rho_k} + \tfrac{6L_{\nabla m_1}\delta_{\nabla c}^2 M'_{\rho c}}{\eta_1}K^{r_\eta-1}\sum_{k=1}^K \eta_k^2\rho_k. &(63)\end{aligned}$$

For the line (61). Since the initial point x_1 is chosen before the algorithm runs, one has that the value $\mathbb{E}_{[\mathcal{S}_1]}\left[m_{\rho_1}(x_1)\right]$ is given, hence bounded. As a result, when the sequence $\{\mathbb{E}_{[\mathcal{S}_{k+1}]}\left[m_{\rho_{k+1}}(x_{k+1})\right]\}_{k\in\mathbb{N}}$ is lower bounded, one has that the value

$$(\mathbb{E}_{[\mathcal{S}_K]}\left[m_{\rho_0}(x_0)\right] - \mathbb{E}_{[\mathcal{S}_{K+1}]}\left[m_{\rho_{K+1}}(x_{K+1})\right]) \text{ is uniformly upper bounded for all } K. \tag{64}$$

On the other hand, note that $\rho_k = \rho_1 k^{r_\rho}$ with $r_\rho \geq 0$. For the second-to-last term in (63), one has that the term $\frac{1}{k\rho_k}$ has an order that is no larger than -1. Hence, one has the summation

$$\sum_{k=1}^K \tfrac{1}{k\rho_k} \leq \tfrac{1}{\rho_1}\sum_{k=1}^K \tfrac{1}{k} \overset{K\gg 1}{=} \mathcal{O}(\log K) = \tilde{\mathcal{O}}(1). \tag{65}$$

Next, one bounds the summation $\sum_{k=1}^{K}\eta_k^2\rho_k$ which appears in the first term in (62) and the last term in (63). One has $\sum_{k=1}^{K}\eta_k^2\rho_k = \eta_1^2\rho_1\sum_{k=1}^{K}k^{-2r_\eta+r_\rho}$ and the following two cases. When $-2r_\eta+r_\rho\geq 0$, one has

$$\sum_{k=1}^{K}k^{-2r_\eta+r_\rho}\leq\int_{k=1}^{K}k^{-2r_\eta+r_\rho}dk=\tfrac{1}{1-2r_\eta+r_\rho}(K^{1-2r_\eta+r_\rho}-1)\overset{K\gg 1}{=}\mathcal{O}(K^{1-2r_\eta+r_\rho}),\tag{66}$$

where the last equality uses $1-2r_\eta+r_\rho>0$. When $-2r_\eta+r_\rho\leq 0$, one has

$$\sum_{k=1}^{K}k^{-2r_\eta+r_\rho}=1+\sum_{k=2}^{K}k^{-2r_\eta+r_\rho}\leq 1+\int_{k=1}^{K}k^{-2r_\eta+r_\rho}dk.\tag{67}$$

And further

$$\begin{aligned}
&1+\int_{k=1}^{K}k^{-2r_\eta+r_\rho}dk\\
\leq&\begin{cases}1+\frac{1}{1-2r_\eta+r_\rho}(K^{1-2r_\eta+r_\rho}-1) & ,\text{if } -2r_\eta+r_\rho\in(-\infty,-1)\cup(-1,0]\\ 1+\ln K & ,\text{if } -2r_\eta+r_\rho=-1\end{cases}\\
\leq&\begin{cases}\frac{1}{1-2r_\eta+r_\rho}K^{1-2r_\eta+r_\rho} & ,\text{if } -2r_\eta+r_\rho\in(-1,0]\\ 1-\frac{1}{1-2r_\eta+r_\rho} & ,\text{if } -2r_\eta+r_\rho\in(-\infty,-1)\\ 1+\ln K & ,\text{if } -2r_\eta+r_\rho=-1\end{cases}\\
\overset{K\gg 1}{=}&\tilde{\mathcal{O}}(1+K^{1-2r_\eta+r_\rho}).
\end{aligned}\tag{68}$$

Combining (66) – (68), one has

$$\sum_{k=1}^{K}\eta_k^2\rho_k\leq\tilde{\mathcal{O}}(1+K^{1-2r_\eta+r_\rho})\text{ when }K\gg 1.\tag{69}$$

Next, we consider the last term in (62). Note that Lemma 2 defines two real scalars $k_{\tilde{\delta}}$ and $M_{\tilde{\delta}}^K$ in (25), where $k_{\tilde{\delta}}$ only dependes on the values of (r_η,r_ρ,r_ω), hence is independent of K. From (25), $M_{\tilde{\delta}}^K$ dependes on upper bound of $\tilde{\eta}_{1,K}$ which relies on upper bound of $M_{\rho c}^K$, hence $M_{\tilde{\delta}}^K$ relies on upper bound of $M_{\rho c}^K$. Since (50) assumes an uniformly upper bound of $M_{\rho c}^K$, one has that $M_{\tilde{\delta}}^K$ is also indepndent of K. To indicate the independence, one use $M_{\tilde{\delta}}$ to replace $M_{\tilde{\delta}}^K$ here. As a result, one has scalars $(k_{\tilde{\delta}},M_{\tilde{\delta}})\in\mathbb{N}\times\mathbb{R}_{>0}$ such that

$$\tilde{\delta}_k\leq M_{\tilde{\delta}}k^{-r_\omega+2r_{\tilde{\eta}}}\text{ for all }k\geq k_{\tilde{\delta}},\text{ where }r_{\tilde{\eta}}=\max\{r_\omega,-r_\eta\}.$$

Since $k_{\tilde{\delta}}<\infty$, one has the existence of a scalar $M'_{\tilde{\delta}}$ such that

$$\tilde{\delta}_k\leq M'_{\tilde{\delta}}k^{-r_\omega+2r_{\tilde{\eta}}}\text{ for all }k\geq 1.$$

As a result, one has

$$\begin{aligned}\sum_{k=1}^{K}\eta_k\rho_k^2\tilde{\delta}_{C_k} &\leq M'_{\tilde{\delta}}\eta_1\rho_1^2\sum_{k=1}^{K}k^{-r_\eta+2r_\rho-r_\omega+\max\{2r_\omega,-2r_\eta\}}\\ &\overset{K\gg 1}{=}\mathcal{O}(\sum_{k=1}^{K}k^{\max\{r_\omega-r_\eta+2r_\rho,-r_\omega-3r_\eta+2r_\rho\}})\\ &\overset{K\gg 1}{\leq}\tilde{\mathcal{O}}(1+K^{1+\max\{r_\omega-r_\eta+2r_\rho,-r_\omega-3r_\eta+2r_\rho\}}).\end{aligned}\tag{70}$$

Combining (59) – (70), from the convexity of the squared norm, Jensen's inequality, one has

$$\begin{aligned}(\mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_{\hat{K}}}(x_{\hat{K}})\|])^2 &\leq \mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_{\hat{K}}}(x_{\hat{K}})\|^2]\\ &\overset{K\gg 1}{\leq}\tilde{\mathcal{O}}(K^{r_\eta-1}+K^{-r_\eta+r_\rho}+K^{\max\{r_\omega+2r_\rho,-r_\omega-2r_\eta+2r_\rho\}})\\ &\leq\tilde{\mathcal{O}}((K^{\frac{r_\eta-1}{2}}+K^{\frac{-r_\eta+r_\rho}{2}}+K^{\frac{\max\{r_\omega+2r_\rho,-r_\omega-2r_\eta+2r_\rho\}}{2}})^2).\end{aligned}$$

Taking square roots on both sides gives

$$\mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_{\hat{K}}}(x_{\hat{K}})\|]\leq\tilde{\mathcal{O}}(K^{\frac{r_\eta-1}{2}}+K^{\frac{-r_\eta+r_\rho}{2}}+K^{\frac{\max\{r_\omega+2r_\rho,-r_\omega-2r_\eta+2r_\rho\}}{2}}).\tag{71}$$

On the other hand, we give a bound for $\mathbb{E}_{[\mathcal{S}_K]}[\|\mathcal{P}(c(x_{\hat{K}}))\|^2]$. From the definition of $x_{\hat{K}}$, $\{\rho_k\}_k$ in (3) and (50), one has

$$\mathbb{E}_{[\mathcal{S}_K]}[\|\mathcal{P}(c(x_{\hat{K}}))\|^2]=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2]\leq\frac{M'_{\rho c}}{K}\sum_{k=1}^{K}\rho_k^{-2}\leq\frac{M'_{\rho c}}{\rho_1^2K}\int_{k=1}^{K}k^{-2r_\rho}dk.\tag{72}$$

And further,

$$\begin{aligned}\frac{1}{K}\int_{k=1}^{K}k^{-2r_\rho}dk &\leq\begin{cases}\frac{1}{K}\frac{1}{1-2r_\rho}(K^{1-2r_\rho}-1) & ,\text{if } r_\rho\in[0,\frac{1}{2})\cup(\frac{1}{2},\infty)\\ \frac{1}{K}\ln K & ,\text{if } r_\rho=\frac{1}{2}.\end{cases}\\ &\overset{K\gg 1}{\leq}\tilde{\mathcal{O}}(K^{-2r_\rho}).\end{aligned}\tag{73}$$

Combing (72), (73), and follow similar arguments that lead to (71), one has

$$\mathbb{E}_{[\mathcal{S}_K]}[\|\mathcal{P}(c(x_{\hat{K}}))\|]\overset{K\gg 1}{\leq}\tilde{\mathcal{O}}(K^{-r_\rho}).\tag{74}$$

Letting $\lambda_{\hat{K}}=\rho_{\hat{K}}\mathcal{P}(c(x_{\hat{K}}))$, and via similar arguments that lead to (73), one has the

approximate slackness measure

$$\begin{aligned}\mathbb{E}_{[\mathcal{S}_k]}[\|\lambda_{\hat{K}}^T\mathcal{P}(c(x_{\hat{K}}))\|] &= \mathbb{E}_{[\mathcal{S}_k]}[\rho_{\hat{K}}\|\mathcal{P}(c(x_{\hat{K}}))\|^2] \\ &= \tfrac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{[\mathcal{S}_k]}[\rho_k\|\mathcal{P}(c(x_k))\|^2] \\ &\overset{K\gg 1}{\leq} \tilde{\mathcal{O}}(K^{-r_\rho}).\end{aligned} \tag{75}$$

Combining (71), (74), (75) and the linearity of expectation, one has

$$\begin{aligned}&\mathbb{E}_{[\mathcal{S}_k]}[\|\nabla f(x_{\hat{K}}) + \nabla c(x_{\hat{K}})^T\lambda_{\hat{K}}\| + \|\mathcal{P}(c(x_{\hat{K}}))\| + \|\lambda_{\hat{K}}^T\mathcal{P}(c(x_{\hat{K}}))\|] \\ \overset{K\gg 1}{\leq}\ &\tilde{\mathcal{O}}(K^{\frac{r_\eta-1}{2}} + K^{\frac{-r_\eta+r_\rho}{2}} + K^{\frac{\max\{r_\omega+2r_\rho,-r_\omega-2r_\eta+2r_\rho\}}{2}} + K^{-r_\rho}).\end{aligned} \tag{76}$$

Combining with the requirements in (58), one can find an optimal choice for the orders $(r_\eta, r_\rho, r_\omega)$ by solving the following constraint problem

$$\begin{aligned}\min_{(r_\eta,r_\rho,r_\omega)}\ &\max\{r_\eta - 1, -r_\eta + r_\rho, r_\omega + 2r_\rho, -r_\omega - 2r_\eta + 2r_\rho, -2r_\rho\} \\ \text{s.t. } &r_\eta > 0,\ r_\rho > 0,\ \textit{and (58) is satisfied.}\end{aligned} \tag{77}$$

In the Appendix A, one is able to transform (77) into a linear programming problem. By calling a MATLAB solver, one obtains an optimal solution

$$(r_\eta, r_\rho, r_\omega) = (\tfrac{2}{3}, \tfrac{1}{6}, -\tfrac{2}{3})$$

with an objective value of (77) being $-\frac{1}{3}$, meaning that when $K \gg 1$, one has

$$\mathbb{E}_{[\mathcal{S}_k]}[\|\nabla f(x_{\hat{K}}) + \nabla c(x_{\hat{K}})^T\lambda_{\hat{K}}\| + \|\mathcal{P}(c(x_{\hat{K}}))\| + \|\lambda_{\hat{K}}^T\mathcal{P}(c(x_{\hat{K}}))\|] \leq \tilde{\mathcal{O}}(K^{-\frac{1}{6}}). \tag{78}$$

Equavalently, for any $\epsilon > 0$, when the right-hand side of (78) to be the size $\tilde{\mathcal{O}}(\epsilon)$, one has

$$\tilde{\mathcal{O}}(K^{-\frac{1}{6}}) = \epsilon \Longleftrightarrow K = \tilde{\mathcal{O}}(\epsilon^{-6}),$$

indicating that one can find an ϵ-stochastic stationary point in at most $\tilde{\mathcal{O}}(\epsilon^{-6})$ iterations. Also note that in Algorithm 1 line 4, in each iteration, only three data points are drawn. Hence, Algorithm 1 has the same sample complexity as the iteration complexity, i.e., at most $\tilde{\mathcal{O}}(\epsilon^{-6})$. □

Remark 3. Let us consider when $\delta_{\nabla c} = 0$ and $\delta_{\nabla f} = 0$. One has that (23) to be

$$\tilde{\delta}_{C_{k+1}} \leq (1 - \omega_{k+1})\,\tilde{\delta}_{C_k} + 3\omega_{k+1}^2\delta_c^2 + 12L_{\tilde{c}}^2\eta_k^2\mathbb{E}_{\mathcal{S}_k}[\|\nabla m_{\rho_k}(x_k)\|^2]. \tag{79}$$

Recall the merit descent result (40)

$$\begin{aligned}&\mathbb{E}_{[\mathcal{S}_k]}\left[m_{\rho_{k+1}}(x_{k+1})\right]\\ \leq&\mathbb{E}_{[\mathcal{S}_k]}[m_{\rho_k}(x_k)] - \tfrac{1}{2}\eta_k\mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\|^2] + 3L_{\nabla m_1}\delta_{\nabla f}^2\eta_k^2\rho_k\\ &+ 3L_{\bar{c}}^2\eta_k\rho_k^2\tilde{\delta}_{C_k} + (2^\beta\beta\tfrac{\rho_k}{k} + 3L_{\nabla m_1}\delta_{\nabla c}^2\eta_k^2\rho_k^3)\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2]).\end{aligned} \tag{80}$$

The potential function approach is as follows: assuming the existence of a sequence $\{u_k\}_k$ such that

$$\tfrac{1-\omega_{k+1}}{u_{k+1}} \leq \tfrac{1}{u_k} - \eta_k\rho_k^2.$$

Then, multiplying both sides of (79) by $\frac{3L_{\bar{c}}^2}{u_{k+1}}$ and use the above inequality

$$\tfrac{3L_{\bar{c}}^2}{u_{k+1}}\tilde{\delta}_{C_{k+1}} \leq \tfrac{3L_{\bar{c}}^2}{u_k}\tilde{\delta}_{C_k} - 3L_{\bar{c}}^2\eta_k\rho_k^2\tilde{\delta}_{C_k} + \tfrac{9L_{\bar{c}}^2\omega_{k+1}^2}{u_{k+1}}\delta_c^2 + \tfrac{36L_{\bar{c}}^4\eta_k^2}{u_{k+1}}\mathbb{E}_{\mathcal{S}_k}[\|\nabla m_{\rho_k}(x_k)\|^2]. \tag{81}$$

Sum (81) to (80) on both sides, and define

$$M_k = \mathbb{E}_{[\mathcal{S}_k]}[m_{\rho_k}(x_k)] + \tfrac{3L_{\bar{c}}^2}{u_k}\tilde{\delta}_{C_k} \text{ as a potential function at } x_k,$$

one has

$$\begin{aligned}M_{k+1} \leq M_k - (\tfrac{1}{2} - \tfrac{36L_{\bar{c}}^4\eta_k}{u_{k+1}})\eta_k\mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\|^2] + 3L_{\nabla m_1}\delta_{\nabla f}^2\eta_k^2\rho_k\\ + (2^\beta\beta\tfrac{\rho_k}{k} + 3L_{\nabla m_1}\delta_{\nabla c}^2\eta_k^2\rho_k^3)\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2]) + \tfrac{9L_{\bar{c}}^2\omega_{k+1}^2}{u_{k+1}}\delta_c^2.\end{aligned} \tag{82}$$

Then, performing summation and telescoping cancellation for M_k, one can obtain optimal choices for parameters in this case.

2.4 Proof for unifomrly boundness of $\{M_{\rho c}^K\}_K$

Lemma 5. Under Assumptions 1 – 5 and suppose inequalities (5) and (22) hold. In addition, assume that for all $k \in \mathbb{N}$, it holds

$$\eta_k\rho_k \leq \min\{\tfrac{\sigma_{\nabla c}^2}{4(9L_{\|\mathcal{P}\circ c\|^2}\delta_{\nabla c}^2 + 6L_{\bar{c}}^2L_{\|\mathcal{P}\circ c\|^2})}, \tfrac{L_{\bar{c}}^4}{12L_{\mathcal{P}\circ c}\kappa_{\nabla\bar{c}}^2\sigma_{\nabla c}^2}, \tfrac{\kappa_{\nabla f}^2L_{\bar{c}}^2}{4\sigma_{\nabla c}^2(2\kappa_{\nabla f}^2L_{\|\mathcal{P}\circ c\|^2} + 3L_{\|\mathcal{P}\circ c\|^2}\delta_{\nabla f}^2)}\}. \tag{83}$$

Then, one has

$$\begin{aligned}\mathbb{E}_{[\mathcal{S}_k]}[\tfrac{1}{2}\|\mathcal{P}(c(x_{k+1}))\|^2] \leq (1 - \tfrac{\sigma_{\nabla c}^2}{2}\eta_k\rho_k)\mathbb{E}_{[\mathcal{S}_k]}[\tfrac{1}{2}\|\mathcal{P}(c(x_k))\|^2] + \tfrac{L_{\bar{c}}^4}{\sigma_{\nabla c}^2}\eta_k\rho_k\tilde{\delta}_{C_k}\\ + \tfrac{\kappa_{\nabla f}^2L_{\bar{c}}^2}{\sigma_{\nabla c}^2}\eta_k\rho_k^{-1}.\end{aligned} \tag{84}$$

Proof. First, one can observe that (83) is compatible with (5) and (22) by first choosing any $\rho_1 \geq 1$, and then choosing a small enough η_1. Via Lemma 1 and Taylor's expansion on the function $\frac{1}{2}\|\mathcal{P} \circ c\|^2$ at x_k, one has

$$\begin{aligned}
&\mathbb{E}_{[\mathcal{S}_k]}[\tfrac{1}{2}\|\mathcal{P}(c(x_{k+1}))\|^2] \\
=&\mathbb{E}_{[\mathcal{S}_k]}[\tfrac{1}{2}\|\mathcal{P}(c(x_k - \eta_k \bar{d}_k))\|^2] \\
\leq&\mathbb{E}_{[\mathcal{S}_k]}[\tfrac{1}{2}\|\mathcal{P}(c(x_k))\|^2 - \eta_k \mathcal{P}(c(x_k))^T \nabla c(x_k)\bar{d}_k + \tfrac{1}{2}L_{\|\mathcal{P}\circ c\|^2}\eta_k^2 \left\|\bar{d}_k\right\|^2] \\
=&\mathbb{E}_{[\mathcal{S}_k]}[\tfrac{1}{2}\|\mathcal{P}(c(x_k))\|^2] - \eta_k \mathbb{E}_{[\mathcal{S}_k]}[\mathcal{P}(c(x_k))^T \nabla c(x_k)\bar{d}_k] + \tfrac{1}{2}L_{\|\mathcal{P}\circ c\|^2}\eta_k^2 \mathbb{E}_{[\mathcal{S}_k]}[\left\|\bar{d}_k\right\|^2].
\end{aligned} \tag{85}$$

Recall that $\bar{d}_k = \nabla \overline{f}(x_k, \xi_k^f) + \rho_k [\nabla \bar{c}(x_k, \xi_k^J)]^T \mathcal{P}(C_k)$. By adding and subtracting a term, and note that x_k depends only on x_1 and $[\mathcal{S}_{k-1}]$, one has

$$\begin{aligned}
&\mathbb{E}_{[\mathcal{S}_k]}[\mathcal{P}(c(x_k))^T \nabla c(x_k)\bar{d}_k] \\
=&\mathbb{E}_{[\mathcal{S}_k]}[\mathcal{P}(c(x_k))^T \nabla c(x_k) \nabla \overline{f}(x_k, \xi_k^f)] + \rho_k \mathbb{E}_{[\mathcal{S}_k]}[[\nabla \bar{c}(x_k, \xi_k^J)]^T \mathcal{P}(C_k)] \\
=&\mathbb{E}_{[\mathcal{S}_k]}[\mathcal{P}(c(x_k))^T \nabla c(x_k) \nabla \overline{f}(x_k, \xi_k^f)] + \rho_k \mathbb{E}_{[\mathcal{S}_k]}[\mathcal{P}(c(x_k))^T \nabla c(x_k)[\nabla \bar{c}(x_k, \xi_k^J)]^T \mathcal{P}(c(x_k))] \\
&\quad + \rho_k \mathbb{E}_{[\mathcal{S}_k]}[\mathcal{P}(c(x_k))^T \nabla c(x_k)[\nabla \bar{c}(x_k, \xi_k^J)]^T (\mathcal{P}(C_k) - \mathcal{P}(c(x_k)))] \\
=&\mathbb{E}_{[\mathcal{S}_k]}[\mathcal{P}(c(x_k))^T \nabla c(x_k) \nabla f(x_k)] + \rho_k \mathbb{E}_{[\mathcal{S}_k]}[\mathcal{P}(c(x_k))^T \nabla c(x_k)[\nabla c(x_k)]^T \mathcal{P}(c(x_k))] \\
&\quad + \rho_k \mathbb{E}_{[\mathcal{S}_k]}[\mathcal{P}(c(x_k))^T \nabla c(x_k)[\nabla \bar{c}(x_k, \xi_k^J)]^T (\mathcal{P}(C_k) - \mathcal{P}(c(x_k)))].
\end{aligned} \tag{86}$$

Note that C_k, $\nabla \bar{c}(x_k, \xi_k^J)$ are independent when conditioned on x_k, one has

$$\begin{aligned}
&\mathbb{E}_{[\mathcal{S}_k]}[\mathcal{P}(c(x_k))^T \nabla c(x_k)[\nabla \bar{c}(x_k, \xi_k^J)]^T (\mathcal{P}(C_k) - \mathcal{P}(c(x_k)))] \\
=&\mathbb{E}_{[\mathcal{S}_{k-1}]}[\mathbb{E}_{\mathcal{S}_k}[\mathcal{P}(c(x_k))^T \nabla c(x_k)[\nabla \bar{c}(x_k, \xi_k^J)]^T (\mathcal{P}(C_k) - \mathcal{P}(c(x_k)))\big|[\mathcal{S}_{k-1}]] \\
=&\mathbb{E}_{[\mathcal{S}_k]}[\mathcal{P}(c(x_k))^T \nabla c(x_k)[\nabla c(x_k)]^T (\mathcal{P}(C_k) - \mathcal{P}(c(x_k)))].
\end{aligned} \tag{87}$$

Combining (86) and (87). From submultiplicity, Assumption 2, Assumption 5 and Lemma 1, one has

$$\begin{aligned}
&- \eta_k \mathbb{E}_{[\mathcal{S}_k]}[\mathcal{P}(c(x_k))^T \nabla c(x_k)\bar{d}_k] \\
\leq&\eta_k \mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|\|\nabla c(x_k)\|\|\nabla f(x_k)\|] \\
&\quad - \eta_k \rho_k \mathbb{E}_{[\mathcal{S}_k]}[\mathcal{P}(c(x_k))^T \nabla c(x_k)[\nabla c(x_k)]^T \mathcal{P}(c(x_k))] \\
&\quad + \eta_k \rho_k \mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|\|\nabla c(x_k)\|\|\nabla c(x_k)\|\|\mathcal{P}(C_k) - \mathcal{P}(c(x_k))\|] \\
\leq&\eta_k L_{\bar{c}} \kappa_{\nabla f} \mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|] - \eta_k \rho_k \sigma_{\nabla c}^2 \mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2] \\
&\quad + L_{\bar{c}}^2 \eta_k \rho_k \mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|\|C_k - c(x_k)\|].
\end{aligned} \tag{88}$$

Moreover, from Young's inequality, one has

$$\begin{aligned}\eta_k L_{\bar{c}}\kappa_{\nabla f}\|\mathcal{P}(c(x_k))\| &\leq \tfrac{1}{3}\eta_k\rho_k\sigma_{\nabla c}^2\|\mathcal{P}(c(x_k))\|^2 + \frac{3\left(\eta_k\kappa_{\nabla f}L_{\bar{c}}\right)^2}{4\eta_k\rho_k\sigma_{\nabla c}^2}, \text{ and}\\ L_{\bar{c}}^2\eta_k\rho_k\|\mathcal{P}(c(x_k))\|\|C_k - c(x_k)\| &\leq \tfrac{1}{3}\eta_k\rho_k\sigma_{\nabla c}^2\|\mathcal{P}(c(x_k))\|^2 \\ &\quad + \frac{3\left(\eta_k\rho_k L_{\bar{c}}^2\right)^2\|C_k - c(x_k)\|^2}{4\eta_k\rho_k\sigma_{\nabla c}^2}.\end{aligned} \tag{89}$$

And recall (32) that

$$\|\nabla m_{\rho_k}(x_k)\|^2 \leq 2\kappa_{\nabla f}^2 + 2L_{\bar{c}}^2\rho_k^2\|\mathcal{P}(c(x_k))\|^2.$$

Taking the total expectation on both sides of (32) and (89), combining (85) – (88), Lemma 1 and (83), one has

$$\begin{aligned}&\mathbb{E}_{[\mathcal{S}_k]}[\tfrac{1}{2}\|\mathcal{P}(c(x_{k+1}))\|^2]\\ &\leq \mathbb{E}_{[\mathcal{S}_k]}[\tfrac{1}{2}\|\mathcal{P}(c(x_k))\|^2] - \tfrac{1}{3}\eta_k\rho_k\sigma_{\nabla c}^2\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2] + \tfrac{3\kappa_{\nabla f}^2L_{\bar{c}}^2}{4\sigma_{\nabla c}^2}\eta_k\rho_k^{-1}\\ &\quad + \tfrac{3L_{\bar{c}}^4}{4\sigma_{\nabla c}^2}\eta_k\rho_k\mathbb{E}_{[\mathcal{S}_k]}[\|C_k - c(x_k)\|^2] + \tfrac{1}{2}L_{\|\mathcal{P}\circ c\|^2}\eta_k^2\mathbb{E}_{[\mathcal{S}_k]}[\left\|\bar{d}_k\right\|^2]\\ &\leq \mathbb{E}_{[\mathcal{S}_k]}[\tfrac{1}{2}\|\mathcal{P}(c(x_k))\|^2] - \tfrac{1}{3}\eta_k\rho_k\sigma_{\nabla c}^2\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2] + \tfrac{3\kappa_{\nabla f}^2L_{\bar{c}}^2}{4\sigma_{\nabla c}^2}\eta_k\rho_k^{-1}\\ &\quad + \tfrac{3L_{\bar{c}}^4}{4\sigma_{\nabla c}^2}\eta_k\rho_k\tilde{\delta}_{C_k} + \tfrac{1}{2}L_{\|\mathcal{P}\circ c\|^2}\eta_k^2(2\mathbb{E}_{[\mathcal{S}_k]}[\|\nabla m_{\rho_k}(x_k)\|^2] + 6\delta_{\nabla f}^2\\ &\qquad + 6\kappa_{\nabla\bar{c}}^2\rho_k^2\tilde{\delta}_{C_k} + 6\delta_{\nabla c}^2\rho_k^2\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2])\\ &\leq \mathbb{E}_{[\mathcal{S}_k]}[\tfrac{1}{2}\|\mathcal{P}(c(x_k))\|^2] - \tfrac{1}{3}\eta_k\rho_k(\sigma_{\nabla c}^2 - 9L_{\|\mathcal{P}\circ c\|^2}\delta_{\nabla c}^2\eta_k\rho_k)\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2]\\ &\quad + \tfrac{3\kappa_{\nabla f}^2L_{\bar{c}}^2}{4\sigma_{\nabla c}^2}\eta_k\rho_k^{-1} + \eta_k\rho_k(\tfrac{3L_{\bar{c}}^4}{4\sigma_{\nabla c}^2} + 3L_{\|\mathcal{P}\circ c\|^2}\kappa_{\nabla\bar{c}}^2\eta_k\rho_k)\tilde{\delta}_{C_k}\\ &\quad + L_{\|\mathcal{P}\circ c\|^2}\eta_k^2(2\kappa_{\nabla f}^2 + 2L_{\bar{c}}^2\rho_k^2\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2]) + 3L_{\|\mathcal{P}\circ c\|^2}\delta_{\nabla f}^2\eta_k^2\\ &= \mathbb{E}_{[\mathcal{S}_k]}[\tfrac{1}{2}\|\mathcal{P}(c(x_k))\|^2]\\ &\quad - \tfrac{1}{3}\eta_k\rho_k(\sigma_{\nabla c}^2 - 9L_{\|\mathcal{P}\circ c\|^2}\delta_{\nabla c}^2\eta_k\rho_k - 6L_{\bar{c}}^2L_{\|\mathcal{P}\circ c\|^2}\eta_k\rho_k)\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_k))\|^2]\\ &\quad + \eta_k\rho_k(\tfrac{3L_{\bar{c}}^4}{4\sigma_{\nabla c}^2} + 3L_{\|\mathcal{P}\circ c\|^2}\kappa_{\nabla\bar{c}}^2\eta_k\rho_k)\tilde{\delta}_{C_k}\\ &\quad + (2\kappa_{\nabla f}^2L_{\|\mathcal{P}\circ c\|^2}\eta_k\rho_k + 3L_{\|\mathcal{P}\circ c\|^2}\delta_{\nabla f}^2\eta_k\rho_k + \tfrac{3\kappa_{\nabla f}^2L_{\bar{c}}^2}{4\sigma_{\nabla c}^2})\eta_k\rho_k^{-1}.\end{aligned} \tag{90}$$

Combining with (83) which gives

$$\begin{aligned}\sigma_{\nabla c}^2 - 9L_{\|\mathcal{P}\circ c\|^2}\delta_{\nabla c}^2\eta_k\rho_k - 6L_{\bar{c}}^2L_{\|\mathcal{P}\circ c\|^2}\eta_k\rho_k &\geq \tfrac{3\sigma_{\nabla c}^2}{4},\\ \tfrac{3L_{\bar{c}}^4}{4\sigma_{\nabla c}^2} + 3L_{\|\mathcal{P}\circ c\|^2}\kappa_{\nabla\bar{c}}^2\eta_k\rho_k &\leq \tfrac{L_{\bar{c}}^4}{\sigma_{\nabla c}^2},\\ \text{and } 2\kappa_{\nabla f}^2L_{\|\mathcal{P}\circ c\|^2}\eta_k\rho_k + 3L_{\|\mathcal{P}\circ c\|^2}\delta_{\nabla f}^2\eta_k\rho_k &\leq \tfrac{\kappa_{\nabla f}^2L_{\bar{c}}^2}{\sigma_{\nabla c}^2},\end{aligned}$$

one has (84) as desired.

□

Theorem 1. Under Assumption 1 – 5. When the parameters $(\eta_1, \rho_1, \omega_1)$ and $(r_\eta, r_\rho, r_\omega)$ satisfies

$$\begin{cases} (r_\eta, r_\rho, r_\omega) = (\frac{2}{3}, \frac{1}{6}, -\frac{2}{3}), \\ \rho_1 \geq 1,\ \eta_1 \leq 1,\ \omega_1 = 72 L_{\tilde{c}}^2 \kappa_{\nabla \tilde{c}}^2 \eta_1^2 \rho_1^2 \text{ and} \\ \eta_1 \rho_1 \leq \min\{ \frac{1}{4 L_{\nabla m_1}}, \frac{L_{\tilde{c}}^2}{3 L_{\nabla m_1} \kappa_{\nabla \tilde{c}}^2}, \frac{\sigma_{\nabla c}^2}{4(9 L_{\|\mathcal{P}\circ c\|^2} \delta_{\nabla c}^2 + 6 L_{\tilde{c}}^2 L_{\|\mathcal{P}\circ c\|^2})}, \\ \qquad \frac{L_{\tilde{c}}^4}{12 L_{\|\mathcal{P}\circ c\|^2} \kappa_{\nabla \tilde{c}}^2 \sigma_{\nabla c}^2}, \frac{\kappa_{\nabla f}^2 L_{\tilde{c}}^2}{4 \sigma_{\nabla c}^2 (2 \kappa_{\nabla f}^2 L_{\|\mathcal{P}\circ c\|^2} + 3 L_{\|\mathcal{P}\circ c\|^2} \delta_{\nabla f}^2)}, \frac{1}{12 L_{\tilde{c}} \kappa_{\nabla \tilde{c}}} \}. \end{cases} \tag{91}$$

One has the existence of a finite scalar $M'_{\rho c} < \infty$ such that

$$\mathbb{E}_{[\mathcal{S}_k]} \left[\rho_k^2 \|\mathcal{P}(c(x_k))\|^2 \right] \leq M'_{\rho c} \text{ for all } k \in \mathbb{N}. \tag{92}$$

Moreover, Algorithm 1 produces an ϵ-stochastic stationary point $x_{\hat{K}}$ of (1) with the lower bound worse-case sample complexity, when $\{\mathbb{E}_{[\mathcal{S}_k]} [m_{\rho_k}(x_k)]\}_{k \in \mathbb{N}}$ is bounded below. Specifically, in at most $\mathcal{O}(\epsilon^{-6})$ number of iterations and with access of at most $\mathcal{O}(\epsilon^{-6})$ single datapoints, $x_{\hat{K}}$ will satisfy

$$\mathbb{E}_{[\mathcal{S}_K]} [\|\nabla f(x_{\hat{K}}) + \nabla c(x_{\hat{K}})^T \lambda_{\hat{K}}\| + \|\mathcal{P}(c(x_{\hat{K}}))\| + \|\lambda_{\hat{K}}^T \mathcal{P}(c(x_{\hat{K}}))\|] \leq \epsilon,$$

where $\lambda_{\hat{K}} = \rho_{\hat{K}} \mathcal{P}(c(x_{\hat{K}})) \in \mathbb{R}^m$.

Proof. We will utilize Lemma 2 and Lemma 5 here. Observe that Assumptions 1 – 5 are assumed here and (91) implies (83), since $\eta_k \rho_k = \eta_1 \rho_1 k^{-r_\eta + r_\rho} = \eta_1 \rho_1 k^{-1/2} \leq \eta_1 \rho_1$. Also observe that (91) implies (49), which implies (5) and (22), as argued in Lemma 4. Hence, Lemma 5 can be applied here. Lemma 2 can be applied here as its assumptions are satisfied here, and the inequality (22) on which it relies holds here.

Before showing the existence of $M'_{\rho c}$ in (92), one requires a few more parameters at hand. Define scalars

$$\begin{cases} \kappa & = \max\{\tilde{\delta}_{k_{\tilde{\delta}}} k_{\tilde{\delta}}^{2/3}, \frac{2\tilde{\eta}_1^2}{\omega_1}\}, \\ k_{\rho c} & = \max\{\lceil \sqrt[1/2]{\frac{4}{3\sigma_{\nabla c}^2 \eta_1 \rho_1}} \rceil, k_{\tilde{\delta}}, (\frac{16 L_{\tilde{c}}^4 \kappa \rho_1^2}{\sigma_{\nabla c}^4})^3\}, \\ M_{\rho c} & = \max_{k \in [k_{\rho c}]} \{\mathbb{E}_{[\mathcal{S}_k]} [\rho_k^2 \|\mathcal{P}(c(x_k))\|^2], 1, \frac{16 \kappa_{\nabla f}^2 L_{\tilde{c}}^2}{\sigma_{\nabla c}^4}\}, \\ M_{\tilde{\delta}} & = \kappa M_{\rho c}, \end{cases} \tag{93}$$

where $\tilde{\eta}_1$ is defined in (25) and independent of the iterates k. The $k_{\tilde{\delta}} = \lceil \sqrt[r_\omega + 1]{\frac{2(r_\omega - 2 r_{\tilde{\eta}})}{\omega_1}} \rceil \} = \left\lceil \sqrt[1/3]{\frac{4}{3\omega_1}} \right\rceil$ is independent of k. Hence, κ and $k_{\rho c}$ are also independent of k. Apply Lemma 2 with $K = k_{\rho c} \geq k_{\tilde{\delta}}$. Then, one has the following for all $k \in [k_{\tilde{\delta}}, k_{\rho c}]$

$$\mathbb{E}_{[\mathcal{S}_k]} [\rho_k^2 \|\mathcal{P}(c(x_k))\|^2] \leq M_{\rho c}, \text{ and } \tilde{\delta}_k \leq M_{\tilde{\delta}} k^{-2/3}. \tag{94}$$

Next, we focus on proving the following statement by using induction, that is

$$\mathbb{E}_{[\mathcal{S}_k]}\left[\rho_k^2\|\mathcal{P}(c(x_k))\|^2\right] \leq M_{\rho c} \text{ for all } k \geq k_{\rho c}. \tag{95}$$

The base case we consider is when the iterate $k = k_{\rho c}$, which holds by the construction of $M_{\rho c}$. By induction, suppose (95) holds for any $k \in (k_{\rho c}, k']$ where $\mathbb{N} \ni k' > k_{\rho c}$. From Lemma 2, one has (94) holds for all $k \in (k_{\rho c}, k']$, with the same constants $(M_{\rho c}, M_{\tilde{\delta}})$ as the base case. Dividing ρ_k^2 on both sides of (95) and plugging in the definition of ρ_k, one has

$$\mathbb{E}_{[\mathcal{S}_k]}\left[\|\mathcal{P}(c(x_k))\|^2\right] \leq M_{\rho c}\rho_k^{-2} = \tfrac{M_{\rho c}}{\rho_1^2}k^{-\frac{1}{3}}. \tag{96}$$

From Lemma 5, (94) and (96), one has

$$\begin{aligned}
&\mathbb{E}_{[\mathcal{S}_k]}[\tfrac{1}{2}\|\mathcal{P}(c(x_{k+1}))\|^2] \\
&\leq (1 - \tfrac{\sigma_{\nabla c}^2}{2}\eta_k\rho_k)\mathbb{E}_{[\mathcal{S}_k]}[\tfrac{1}{2}\|\mathcal{P}(c(x_k))\|^2] + \tfrac{L_{\tilde{c}}^4}{\sigma_{\nabla c}^2}\eta_k\rho_k\tilde{\delta}_{C_k} + \tfrac{\kappa_{\nabla f}^2 L_{\tilde{c}}^2}{\sigma_{\nabla c}^2}\eta_k\rho_k^{-1} \\
&\leq \tfrac{M_{\rho c}}{2\rho_1^2}(1 - \tfrac{\sigma_{\nabla c}^2\eta_1\rho_1}{2}k^{-\frac{1}{2}})k^{-\frac{1}{3}} + \tfrac{L_{\tilde{c}}^4 M_{\tilde{\delta}}\eta_1\rho_1}{\sigma_{\nabla c}^2}k^{-\frac{7}{6}} + \tfrac{\kappa_{\nabla f}^2 L_{\tilde{c}}^2\eta_1}{\sigma_{\nabla c}^2\rho_1}k^{-\frac{5}{6}} \\
&= \tfrac{M_{\rho c}}{2\rho_1^2}k^{-\frac{1}{3}} - (\tfrac{M_{\rho c}\sigma_{\nabla c}^2\eta_1}{4\rho_1} - (\tfrac{L_{\tilde{c}}^4\kappa M_{\rho c}\eta_1\rho_1}{\sigma_{\nabla c}^2}k^{-\frac{1}{3}} + \tfrac{\kappa_{\nabla f}^2 L_{\tilde{c}}^2\eta_1}{\sigma_{\nabla c}^2\rho_1}))k^{-\frac{5}{6}}
\end{aligned} \tag{97}$$

From (93), one has

$$\begin{cases}
k \geq (\tfrac{16L_{\tilde{c}}^4\kappa\rho_1^2}{\sigma_{\nabla c}^4})^3 & \Longrightarrow \tfrac{L_{\tilde{c}}^4\kappa M_{\rho c}\eta_1\rho_1}{\sigma_{\nabla c}^2}k^{-\frac{1}{3}} \leq \tfrac{M_{\rho c}\sigma_{\nabla c}^2\eta_1}{16\rho_1} \\
\tfrac{16\kappa_{\nabla f}^2 L_{\tilde{c}}^2}{\sigma_{\nabla c}^4} \leq M_{\rho c} & \Longrightarrow \tfrac{\kappa_{\nabla f}^2 L_{\tilde{c}}^2\eta_1}{\sigma_{\nabla c}^2\rho_1} \leq \tfrac{M_{\rho c}\sigma_{\nabla c}^2\eta_1}{16\rho_1},
\end{cases}$$

which combining with (97) and multiplying two on both sides, gives

$$\mathbb{E}_{[\mathcal{S}_k]}[\|\mathcal{P}(c(x_{k+1}))\|^2] \leq \tfrac{M_{\rho c}}{\rho_1^2}k^{-\frac{1}{3}} - \tfrac{M_{\rho c}\sigma_{\nabla c}^2\eta_1}{4\rho_1}k^{-\frac{5}{6}}. \tag{98}$$

On the other hand, applying (31) since $-\frac{1}{3} \leq 0$ and $\frac{1}{k} \in (0, 2)$, for all $k \geq 1$. It gives

$$\begin{aligned}
M_{\rho c}\rho_{k+1}^{-2} = \tfrac{M_{\rho c}}{\rho_1^2}(k+1)^{-\frac{1}{3}} &= \tfrac{M_{\rho c}}{\rho_1^2}k^{-\frac{1}{3}}\left(1 + \tfrac{1}{k}\right)^{-\frac{1}{3}} \\
&\geq \tfrac{M_{\rho c}}{\rho_1^2}k^{-\frac{1}{3}}\left(1 - \tfrac{1}{3k}\right) = \tfrac{M_{\rho c}}{\rho_1^2}k^{-\frac{1}{3}} - \tfrac{M_{\rho c}}{3\rho_1^2}k^{-\frac{4}{3}}.
\end{aligned} \tag{99}$$

Since $k \geq k_{\rho c}$, one has

$$k \geq k_{\rho c} \geq \left\lceil \sqrt[1/2]{\tfrac{4}{3\sigma_{\nabla c}^2\eta_1\rho_1}} \right\rceil \Longrightarrow \tfrac{3\sigma_{\nabla c}^2\eta_1\rho_1}{4} \geq k^{-\frac{1}{2}} \Longleftrightarrow \tfrac{M_{\rho c}\sigma_{\nabla c}^2\eta_1}{4\rho_1}k^{-\frac{5}{6}} \geq \tfrac{M_{\rho c}}{3\rho_1^2}k^{-\frac{4}{3}}. \tag{100}$$

Combining (98) – (100), one has

$$\mathbb{E}_{[\mathcal{S}_{k+1}]}[\rho_{k+1}^2\|\mathcal{P}(c(x_{k+1}))\|^2] = \mathbb{E}_{[\mathcal{S}_k]}[\rho_{k+1}^2\|\mathcal{P}(c(x_{k+1}))\|^2] \leq M_{\rho c}.$$

The induction is complete.

Note that (95) implies (92). Indeed, $k_{\rho c}$ is independent of k and there are only finite number of iterates for $k \leq k_{\rho c}$, hence one can choose

$$M'_{\rho c} = \max_{k \in [k_{\rho c}]} \{M_{\rho c}, \mathbb{E}_{[\mathcal{S}_k]}[\rho_k^2 \|\mathcal{P}(c(x_k))\|^2]\}$$

such that (92) holds. As a result, the inequality (50) in Lemma 4 holds here. Moreover, one can verify that the assumptions assumed in Lemma 4 are also assumed here, and the requirements (91) imply requirements (49) in Lemma 4. One is able to apply the result in Lemma 4 here, which gives the desired result.

□

A Transforming (77) into LPs.

Introducing a dummy variable $v \in \mathbb{R}$ and replacing strict inequalities by nonstrict inequalities, one is able to obtain an optimal choice for the orders $(r_\eta, r_\rho, r_\omega)$ by solving the following linear constraint problem, which is

$$\begin{aligned} &\min_{r_\eta, r_\rho, r_\omega, v} \ v \\ &\text{s.t.} \quad \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 2 & 1 & -1 \\ -2 & 2 & -1 & -1 \\ 0 & -2 & 0 & -1 \\ -2 & 2 & -1 & 0 \\ -1 & 1 & 0 & 0 \\ -2 & 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} r_\eta \\ r_\rho \\ r_\omega \\ v \end{bmatrix} \leq \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \\ &\qquad r_\eta \geq 0, \ r_\rho \geq 0, \ -1 \leq r_\omega \leq 0. \end{aligned}$$

By calling a MATLAB solver, one obtains an optimal value of -1/3 with a solution $r_\eta = 2/3, r_\rho = 1/6, r_\omega = -2/3, v = -1/3$.

References

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

[2] Yawen Cui et al. "An exact penalty method for stochastic equality-constrained optimization". In: *Optimization Online* (2025).

[3] Gang Li et al. "A Retention-Centric Framework for Continual Learning with Guaranteed Model Developmental Safety". In: *arXiv preprint arXiv:2410.03955* (2024).

[4] Wei Liu and Yangyang Xu. "A SPIDER-Type Stochastic Subgradient Method for Expectation-Constrained Nonconvex Nonsmooth Optimization". In: *SIAM Journal on Optimization* 36.2 (2026), pp. 1125–1153.