Complexity of Single Loop Algorithms for Nonlinear Programming with Stochastic Objective and Constraints

Ahmet Alacaoglu
University of Wisconsin–Madison

Stephen J. Wright
University of Wisconsin–Madison

Abstract

We analyze the sample complexity of single-loop quadratic penalty and augmented Lagrangian algorithms for solving nonconvex optimization problems with functional equality constraints. We consider three cases, in all of which the objective is stochastic, that is, an expectation over an unknown distribution that is accessed by sampling. The nature of the equality constraints differs among the three cases: deterministic and linear in the first case, deterministic and nonlinear in the second case, and stochastic and nonlinear in the third case. Variance reduction techniques are used to improve the complexity. To find a point that satisfies ε-approximate first-order conditions, we require $\widetilde{O}(\varepsilon^{-3})$ complexity in the first case, $\widetilde{O}(\varepsilon^{-4})$ in the second case, and $\widetilde{O}(\varepsilon^{-5})$ in the third case. For the first and third cases, they are the first algorithms of "single loop" type that also use $O(1)$ samples at each iteration and still achieve the best-known complexity guarantees.

1 INTRODUCTION

Augmented Lagrangian and quadratic penalty algorithms have been a mainstay for solving nonlinear optimization problems for several decades (Hestenes, 1969; Powell, 1969; Fiacco and McCormick, 1968; Bertsekas, 2014). We consider the nonlinear programming template:

$$\min_{x\in X} f(x) \text{ subject to } c(x) = 0, \tag{1.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ and $c : \mathbb{R}^d \to \mathbb{R}^m$. Historically, algorithms for (1.1) are analyzed for the case of nonlinear and nonconvex f and c. Typical results show *asymptotic* iterate convergence with *local linear* or *superlinear* rate guarantees. In the last two decades, with the influence of the emerging field of data science, there has been wider interest in global convergence rates, including *sublinear* rates. Nonasymptotic convergence rate analyses of augmented Lagrangian method (ALM) and quadratic penalty method (QPM) for nonlinear programming in both convex (Lan and Monteiro, 2013, 2016; Xu, 2017, 2021) and nonconvex cases (Hong, 2016; Xie and Wright, 2021; Li et al., 2021; Lin et al., 2022a; Lu, 2022; Huang and Lin, 2023; Kong et al., 2023; Sun and Sun, 2021; He et al., 2023) are surprisingly recent.

With large data sets fueling many recent advances in machine learning and data sciences, *stochastic* algorithms for solving (1.1) have become a necessity. These algorithms generally work with a single-sample or a mini-batch of the full dataset at each iteration. In many applications, even one pass over the data can be prohibitive. *Unconstrained* and *simple-constrained* versions of (1.1), where $c(x)$ is absent, have been solved with stochastic *projected* gradient algorithms for convex or nonconvex f (Lan, 2020; Davis and Drusvyatskiy, 2019; Cutkosky and Orabona, 2019). In this paper, we focus on three subclasses of (1.1) in which $c(x)$ is nontrivial, so that projection onto the feasible set of (1.1) is too expensive to be practical.

In all problems considered in this paper, the objective f has expectation form, so we restate (1.1) more narrowly as follows:

$$\min_{x\in X} \left\{ f(x) := \mathbb{E}_\xi[\tilde{f}(x,\xi)] \right\} \text{ subject to } c(x) = 0, \tag{1.2}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ and $c\colon \mathbb{R}^d \to \mathbb{R}^m$ are functions with Lipschitz continuous gradients that can be nonconvex, while the set $X \subseteq \mathbb{R}^d$ is closed and convex. The function $\tilde{f}$ maps $\mathbb{R}^d \times \Xi$ to $\mathbb{R}$, where Ξ is the sample space with $\xi \in \Xi$ and ξ is distributed according to P_Ξ. $\mathbb{E}_\xi$ is the expectation over the distribution of ξ. (We sometimes abbreviate $\mathbb{E}_\xi$ as $\mathbb{E}$ when the context is clear.)

Three cases. We consider three instances of the template (1.2), motivated by applications in machine

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

learning and data science.

I. Let $c(x) := Ax - b$ and $X = \mathbb{R}^d$, yielding

$$\begin{aligned} &\min_{x \in \mathbb{R}^d} \{f(x) := \mathbb{E}[\tilde{f}(x, \xi)]\} \\ &\text{subject to } Ax = b, \end{aligned} \tag{I}$$

where $A \in \mathbb{R}^{m \times d}$. This problem arises in the context of distributed optimization where the linear constraints enforce consensus (Hong et al., 2018; Hong, 2016). It also arises in resource allocation (Boyd et al., 2011, Section 7.3), reformulations of problems involving the composition of convex or nonconvex functions with linear operators, and other contexts (see also (Hong, 2016, Section 1.1)). Oracle accesses in this case require stochastic gradients of f and matrix multiplications by A and $A^\top$. We denote by $\delta > 0$ the smallest nonzero eigenvalue of $A^\top A$, so that

$$\|A^\top \lambda\| \geq \sqrt{\delta}\|\lambda\| \quad \text{for all } \lambda \in \text{Range}(A). \tag{1.3}$$

II. In the second case, $c : \mathbb{R}^d \to \mathbb{R}^m$ is a deterministic nonlinear function, yielding

$$\begin{aligned} &\min_{x \in X} \{f(x) := \mathbb{E}[\tilde{f}(x, \xi)]\} \\ &\text{subject to } c(x) = 0, \end{aligned} \tag{II}$$

with $X \subseteq \mathbb{R}^d$ a closed, convex set. Problems of this type arise in optimization problems with partial differential equation (PDE) constraints (Kupfer and Sachs, 1992; Rees et al., 2010; Curtis et al., 2021). The oracle access requires stochastic gradients of f, and evaluations of constraint function and gradient, c and ∇c.

III. The third problem has the constraint c as a nonlinear function defined as an expectation

$$\begin{aligned} &\min_{x \in X} \{f(x) := \mathbb{E}_\xi[\tilde{f}(x, \xi)]\}, \\ &\text{subject to } c(x) = \mathbb{E}_\zeta[\tilde{c}(x, \zeta)] = 0, \end{aligned} \tag{III}$$

where $\tilde{c} : \mathbb{R}^d \times Z \to \mathbb{R}^m$, with Z being the sample space and $X \subseteq \mathbb{R}^d$ is closed and convex. The oracle access requires stochastic gradients and stochastic function evaluations for both objective and constraint. This problem is motivated by recent applications of neural network training with output constraints, in such tasks as out-of-distribution detection or fair machine learning (Katz-Samuels et al., 2022; Liu et al., 2020; Dener et al., 2020; Zafar et al., 2019).

Problems (I), (II), (III) are studied in Sections 2, 4.2, and 3, respectively. Inequality constraints can be accommodated into (II) and (III) via the use of slack variables, which can be constrained to be nonnegative by membership in an appropriately defined set X.

First-order stationarity. We say that $\bar{x}$ is ε-*stationary* for (1.2) if there exists $\bar{\lambda} \in \mathbb{R}^m$ such that

$$\begin{aligned} \mathsf{d}(\nabla f(\bar{x}) + \nabla c(\bar{x})^\top \bar{\lambda}, -N_X(\bar{x})) &\leq \varepsilon, \\ \|c(\bar{x})\| &\leq \varepsilon, \end{aligned} \tag{1.4}$$

where $\mathsf{d}(x, C) = \min_{y \in C} \|x - y\|$ is the distance function and $N_X(\bar{x})$ is the normal cone to X at $\bar{x}$. This is the same first-order stationarity definition as in Sahin et al. (2019); Li et al. (2021); Lin et al. (2022a); Li et al. (2023) and generalizes Xie and Wright (2021) to the case when X is present in the problem formulation.

We say that $\bar{x}$ is an ε-*stationary point in expectation* if (1.4) holds in expectation and we consider oracle (defined in the sequel) and sample complexities.

Augmented Lagrangian and Quadratic Penalty. Augmented Lagrangian methods were proposed to overcome both the theoretical and practical drawbacks of quadratic penalty (QP) approaches (Hestenes, 1969). Instances of ALM are traditionally equipped with stronger guarantees than QPM; they incorporate dual updates that help with feasibility guarantees and subproblem conditioning (Bertsekas, 2014). However, for existing nonasymptotic analyses in the nonconvex cases, the literature does not reflect these advantages. In the deterministic case, the existing analyses for ALM with best-known guarantees require the penalty parameters to increase rapidly to infinity, so the dual step size effectively needs to decay (Sahin et al., 2019; Li et al., 2021). The only work to our knowledge with a constant dual step size and penalty parameter is Xie and Wright (2021), which has worse complexity than the methods with growing penalty parameters and decaying dual step sizes.

Since constant penalty parameter and constant dual step sizes are the main features of ALM, we focus on a version of ALM for problem (I) that uses constant penalty parameters and constant dual step sizes. For the more general problems, we focus on QP-based algorithms and their extensions to ALM-type algorithms with small dual step sizes.

Oracle model. We assume throughout to have access to an unbiased oracle for gradients, a standard setting used for example in (Arjevani et al., 2022; Cutkosky and Orabona, 2019). That is, there exists $\tilde{\nabla} f(x, \xi)$ such that

$$\begin{aligned} \mathbb{E}_\xi[\tilde{\nabla} f(x, \xi)] &= \nabla f(x), \\ \mathbb{E}_\xi\|\tilde{\nabla} f(x, \xi) - \tilde{\nabla} f(y, \xi)\|^2 &\leq L^2\|x - y\|^2. \end{aligned} \tag{1.5}$$

Algorithmic Approaches and Contributions. We handle the functional constraints via two classical approaches: quadratic penalty and augmented Lagrangian (Bertsekas, 2014; Nocedal and Wright, 2006).

These algorithmic frameworks normally require solution of subproblems at every iteration. Instead of solving these subproblems exactly, we perform one stochastic gradient descent step on each subproblem, yielding an overall approach that is single-loop in nature. This technique is also known as *linearization* in the context of ALM and QPM (Ouyang et al., 2015). To obtain improved sample complexity guarantees, we also use variance reduction techniques (see Cutkosky and Orabona (2019)). Our main aim in the paper is to provide *simple* and *easily implementable* single-loop algorithms for the described problem classes with optimal or best-known complexity results.[1]

In the case of linear constraints studied in Section 2, we use constant penalty parameter and constant dual step sizes for an ALM with variance reduction. In this case, we show the complexity $\widetilde{O}(\varepsilon^{-3})$ which is optimal (up to a log factor) even for unconstrained, smooth, stochastic optimization (Arjevani et al., 2022).

With functional constraints, consistent with the literature on deterministic instances of our template, we use increasing penalty parameters with quadratic penalty (and decreasing dual step sizes with ALM in Section 4.1). We show $\widetilde{O}(\varepsilon^{-4})$ complexity with deterministic constraints in Section 4.2 and $\widetilde{O}(\varepsilon^{-5})$ complexity with stochastic constraints in Section 3.

Besides being single-loop and requiring only $O(1)$ samples at each iteration, each iteration of our algorithms requires only simple projections and simple vector operations. We do not require complicated auxiliary subproblems to be solved. As a consequence, our sample complexity and computational complexity results are essentially the same.

1.1 Related Works

Algorithms for nonlinear programming have been studied for many decades, but recent years have seen more focus on the case of functions defined as expectations. There is focus also on algorithms that identify an approximate solution in some finite time, expressed in terms of a parameter $\varepsilon > 0$ that quantifies the inexactness in the solution of the problem. For ease of presentation we discuss the related works separately for each of our three special cases.

Problem (I): Nonconvex stochastic optimization with linear constraints. Complexity of ALM in the case of deterministic f is studied in many works; see for example Zhang and Luo (2020, 2022); Hong (2016); Hong et al. (2018). A complexity of $O(\varepsilon^{-2})$ is typical for identifying a point $\bar{x}$ that satisfies first-order conditions ε-approximately in the sense of (1.4), for some $\bar{\lambda}$. Among the works mentioned, Hong (2016); Hong et al. (2018) focus on the unconstrained case and Zhang and Luo (2020, 2022); Zhang et al. (2022) focus on the case in which additional polyhedral constraints (or more general nonlinear functional constraints with further assumptions) are present, requiring error bounds to estimate distances to the optimal set. One feature of the methods in these works is that both the penalty parameter and the dual step size in ALM are constant.

With nonconvex and stochastic objective, Huang et al. (2019) obtained complexity $O(\varepsilon^{-3})$ with additional assumptions such as A having a full rank, large batch sizes depending on accuracy ε, and a uniform upper bound on $\|\nabla f(x)\|^2$. (The latter often does not hold for problems in the form (I).) See Sec. 2.2 for the details and how we address these shortcomings[2].

The work of Zhang et al. (2021) focuses on a consensus-optimization instance of (I) with nonconvex stochastic objective, and uses gradient tracking and the variance-reduction approach from Cutkosky and Orabona (2019) to obtain $\widetilde{O}(\varepsilon^{-3})$ complexity. We achieve the same rate for a more general problem than consensus optimization; see for instance (Hong, 2016, Sec. 1), (Boyd et al., 2011, Section 7) for a "sharing problem" example or Boţ and Nguyen (2020), for standard splitting approaches to represent composite optimization problems as linearly constrained optimization. Our ALM type method is more general and different from the problem-specific method of Zhang et al. (2021).

Problem (II): Nonconvex stochastic optimization with nonlinear deterministic constraints. For this problem, Shi et al. (2022) analyzes an algorithm similar to ours except that the penalty parameter is fixed (ours is variable) and depends on the predefined number of iterations K. Their approach involves an initial stage of finding a feasible point of the nonconvex constraint and has complexity $O(\varepsilon^{-4})$. We show that the initial stage is unnecessary when we use variable parameters that depend on the current iterate k. The sequential quadratic programming (SQP) method of Curtis et al. (2021) has sample complexity $\widetilde{O}(\varepsilon^{-4})$, but this paper does not address iteration complexity or computational complexity directly. In addition, each iteration requires solution of a linear system, a more expensive operation than the vector operations required at each iteration of an ALM.

[1] Such a goal is nontrivial and not always achievable. See e.g., Ji et al. (2022) and Zhang et al. (2020) for different settings where one currently needs multiple loops for best complexity and research for single-loop methods is active.

[2] The same limitations are present in (Lin et al., 2022b, Thm. 5.6).

Problem (III): Nonconvex stochastic optimization with nonlinear stochastic constraints. The two works building on a *regularization* idea for solving problem (III), with inequality constraints instead of equalities, are Boob et al. (2022) and Ma et al. (2020). Both assume the existence of a strictly feasible solution, so their applicability to equality constraints is not clear. Both describe a complexity of $O(\varepsilon^{-6})$ on a slightly weaker assumption on Lipschitz continuity of the gradient. The algorithms of these papers have a double loop structure, compared to the single-loop algorithm that we analyze.

The recent independent work of Li et al. (2023) considered a similar idea of using STORM estimator for this problem to obtain complexity $O(\varepsilon^{-5})$. In contrast to us, they analyzed an inexact ALM. Apart from the complicated structure of a double loop method, an important drawback of this approach is that termination rule of the inner loop is generally not implementable. This is because the number of required iterations of the inner loop depends on the optimal value of the subproblems, variance upper bounds or other unknown values[3]. The other alternative for termination of the inner loop requires computing first order stationarity, which in turn requires the computation of full gradients, an operation that is not practical with stochastic algorithms. By contrast, single-loop any-time algorithms like ours have a straightforward implementation both conceptually and in practice.

Notation. To improve readability, we use standard asymptotic notations such as O, Ω, $\asymp$ in the main text by suppressing universal constants. The distance between a point $x \in \mathbb{R}^n$ and a set $C \subseteq \mathbb{R}^n$ is denoted as $\mathsf{d}(x, C) = \min_{y \in C} \|x - y\|$. *Any-time* refers to an algorithm that does not require setting ε in advance.

2 LINEAR CONSTRAINTS: PROBLEM (I)

2.1 Algorithm and the Main Result

In this section, we address (I), restated here as

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) := \mathbb{E}[\tilde{f}(x, \xi)] \right\} \text{ subject to } Ax = b, \qquad \text{(I)}$$

for which the augmented Lagrangian is

$$\mathcal{L}_\rho(x, \lambda) := f(x) + \langle \lambda, Ax - b \rangle + \frac{\rho}{2} \|Ax - b\|^2,$$

for parameter $\rho > 0$. We describe and analyze a linearized ALM given in Algorithm 1, in which a single step of stochastic gradient descent replaces the minimization of augmented Lagrangian with respect to the primal variable. This algorithm can be seen as a variance-reduced version of ALM with constant step sizes, studied in the deterministic setting by Hong (2016). Due to stochasticity in the objective, we use a variance-reduced estimator of $\nabla f(x_k)$ based on sampling of the oracle $\tilde{\nabla} f(x_i, \xi_i)$ for $i = 0, 1, \ldots, k$; see (1.5) for the oracle description. The output of the algorithm (denoted as $\bar{x}$ in Thm 2.1), after running for K iterations, is a randomly selected primal-dual pair, i.e., $(x_{\hat{k}}, \lambda_{\hat{k}})$ where $\hat{k}$ is selected uniformly at random from $\{1, 2, \ldots, K\}$.

In this section, we obtain optimal complexity results with constant penalty parameter / dual step size ρ in ALM. The latter feature of the algorithm is the main challenge in the analysis, and is the reason for our separate focus on the linearly constrained case.

We make the following assumptions in this case (see also (1.5)):

$$\begin{aligned} \mathbb{E}_\xi \|\tilde{\nabla} f(u, \xi) - \tilde{\nabla} f(v, \xi)\|^2 &\le L_f^2 \|u - v\|^2, \\ \mathbb{E}_\xi \|\tilde{\nabla} f(x, \xi) - \nabla f(x)\|^2 &\le V^2, \\ f(x) &\ge 0, \quad \forall x. \end{aligned} \qquad \text{(A1)}$$

The first assumption in (A1) is Lipschitz continuity of the gradients on average (also called mean-square smoothness, see (Arjevani et al., 2022, eq. (4))) while the second is a standard variance bound. By Jensen's inequality, the first inequality in (A1) also implies that $\|\nabla f(u) - \nabla f(v)\| \le L_f \|u - v\|$. The last assumption in (A1), also made in Hong (2016) is without loss of generality[4].

In the following subsection, we prove the following result, stated informally here for simplicity. The choices of parameters ρ, α_k, and η_k, and the full result appear as Theorem 2.4 and Corollary B.5.

Theorem 2.1 (Informal). *With the assumptions in (A1) and suitable choices of $\eta_k, \rho, \alpha_{k+1}$ as in (2.1), Algorithm 1 outputs $(\bar{x}, \bar{\lambda})$ such that*

$$\mathbb{E}\|\nabla f(\bar{x}) + A^\top \bar{\lambda}\| \le \varepsilon \quad \text{and} \quad \mathbb{E}\|A\bar{x} - b\| \le \varepsilon,$$

after $K = \widetilde{O}(\varepsilon^{-3})$ iterations, thus requiring $\widetilde{O}(\varepsilon^{-3})$ evaluations of stochastic gradients of f.

Remark 2.2. An important aspect of this result, which is critical for ALM, is that the penalty parameter and the dual step size ρ is a constant and independent of final accuracy ε or iteration counter k.

[3] (Li et al., 2023, Lemma 5) suggests that optimal value of subproblems can be replaced by other values such as the diameter of balls containing the iterates, upper bound of function values or the parameter of regularity-condition (δ in (A5) in our notation, v in (Li et al., 2023, Assumption 3)) to set the number of inner iterations. Unfortunately, these values are also normally unknown.

[4] As mentioned in (Hong, 2016, footnote 1, pg 5) this assumption is essentially equivalent to lower boundedness of f.

Algorithm 1 Stochastic Linearized Augmented Lagrangian Method with Variance Reduction for (I)

Input: Initialize λ_0, x_0, g_0 arbitrarily and α_k, ρ, η_k as in (2.2).
for $k = 0, 1, \ldots$ **do**
  $x_{k+1} = x_k - \eta_{k+1}(g_k + A^\top \lambda_k + \rho A^\top (Ax_k - b))$
  $\lambda_{k+1} = \lambda_k + \rho(Ax_{k+1} - b)$
  Sample $\xi_{k+1} \sim P_\Xi$ and set $g_{k+1} = \tilde{\nabla} f(x_{k+1}, \xi_{k+1}) + (1 - \alpha_{k+1})(g_k - \tilde{\nabla} f(x_k, \xi_{k+1}))$
end for

The choices α_k, η_k are independent of ε and they depend on k because of the stochastic setting we focus on. The independence from ε is critical to ensuring that a certain potential function can increase only by a controlled amount at every iteration, which in turn is important for lower boundedness of the expected potential. Further explanations appear in Section 2.2.

2.2 Analysis

We use the following parameters which are written with asymptotic notations for readability (recall (1.3) and (A1)). We suppress only universal constants; full specifications are provided in (B.19).

$$\begin{aligned} \rho &\asymp \frac{L_f}{\delta}, \quad k_0 = \Omega(\mathrm{poly}(\delta^{-1})), \\ \eta &= \frac{1}{11(L_f + \rho\|A\|^2)}, \\ \eta_k &= \frac{\eta}{(k+k_0)^{1/3}\log(k+k_0)}, \quad \alpha_k = 121 L_f^2 \eta_k^2. \end{aligned} \tag{2.1}$$

We start with a lemma that analyzes a single iteration of the algorithm. This lemma constructs a potential function Y_k that we show later to be non-increasing in expectation, up to a small error. Due to the combination of ALM with constant dual step sizes/penalty parameters and the use of variance-reduction techniques, each with complicated constants, the coefficients in this lemma are rather involved. We only provide the orders of some terms, which suffice to convey the central ideas in the lemma.

Lemma 2.3. *Let the assumptions in (A1) hold. Set ρ, η_k, α_k as (2.1). For the iterates of Alg. 1, we have*

$$\begin{aligned} \mathbb{E}Y_{k+1} \le \mathbb{E}Y_k &- \frac{\eta_{k+1}}{2}\mathbb{E}\|g_k - \nabla f(x_k)\|^2 \\ &+ (\beta_{1,k+1} + \beta_{2,k+1})\mathbb{E}\|x_{k+1} - x_k\|^2 \\ &+ w_k + v_k V^2, \end{aligned} \tag{2.2}$$

where

$$\begin{aligned} Y_{k+1} = L_\rho(x_{k+1}, \lambda_{k+1}) &+ \frac{\rho m}{2\eta_{k+2}}\|Ax_{k+1} - b\|^2 \\ &+ \frac{m}{2\eta_{k+1}}\|x_{k+1} - x_k\|^2_{Q_{k+1}} + \beta_{1,k+1}\|x_{k+1} - x_k\|^2 \\ &+ \frac{2}{c\eta_{k+1}}\|g_{k+1} - \nabla f(x_{k+1})\|^2 \\ &+ \left(\frac{6(1+c_1)}{\rho\delta} + \frac{2m}{L_f \eta_k}\right)\|g_k - \nabla f(x_k)\|^2 \end{aligned}$$

with $c_1 > 0$, $c = 121 L_f^2$, $m \asymp \frac{1}{L_f}$, and

$$\begin{aligned} \beta_{1,k} &\asymp L_f + \eta_{k+1}^{-1}, \quad \beta_{2,k} = O\left(\frac{L_f}{\delta} + \eta_{k+1}^{-1}\right) - \frac{1}{2\eta_k}, \\ w_k &= \frac{3(1+c_1)}{\delta\rho}\mathbb{E}\|x_{k+1} - 2x_k + x_{k-1}\|^2_{Q_{k+1}^\top Q_{k+1}} \\ &\quad - \frac{m}{2\eta_{k+1}}\mathbb{E}\|x_{k+1} - 2x_k + x_{k-1}\|^2_{Q_{k+1}}, \\ v_k &= O\left(\frac{\alpha_k^2}{\eta_k}\right), \quad Q_k = \eta_k^{-1} I - \rho A^\top A \succeq 0. \end{aligned}$$

At a high level, the lemma requires us to show that *(i)* $\beta_{1,k+1} + \beta_{2,k+1} < 0$, *(ii)* $w_k \le 0$ and *(iii)* $\sum_{k=1}^\infty v_k < \infty$ to obtain that the function Y_k is non-increasing in expectation up to a small error (see (2.3)). The parameter choices of (2.1) with constants chosen as in in (B.19) can be shown to achieve the required properties, by a tedious but straightforward analysis.

The main theorem of this section utilizes the single-iteration inequality described in Lemma 2.3 to show that both scaled iterate differences and variance term are small. Approximate stationarity follows in view of Theorem 2.1 via standard reductions described in Corollary B.5. We provide a proof sketch to illustrate the main ideas; details are deferred to Section B.3.

Theorem 2.4. *Let the assumptions in (A1) hold and suppose that η_k and the other algorithmic parameters are chosen as in (2.1) (see also (B.19)). Then, for the iterates of Algorithm 1, we have for any $K > 1$ that*

$$\begin{aligned} &\frac{1}{K}\sum_{k=1}^{K+1} \mathbb{E}\left[\|\eta_k^{-1}(x_k - x_{k-1})\|^2 + \|g_{k-1} - \nabla f(x_{k-1})\|^2\right] \\ &= \widetilde{O}(K^{-2/3}). \end{aligned}$$

Proof sketch. In the result of Lemma 2.3, we use the parameter choices given in (2.1) to obtain

$$\begin{aligned} \mathbb{E}Y_{k+1} \le \mathbb{E}Y_k &- \frac{1}{16\eta_{k+1}}\mathbb{E}\|x_{k+1} - x_k\|^2 \\ &- \frac{\eta_{k+1}}{2}\mathbb{E}\|g_k - \nabla f(x_k)\|^2 + v_k V^2. \end{aligned} \tag{2.3}$$

Note that by the choices of η_k, α_k and the definition of v_k, we have that $\sum_{k=1}^\infty v_k = O(1)$. By adjusting (2.3)

and summing for $k \geq 1$, we obtain

$$\frac{\eta_{K+1}}{32} \sum_{k=1}^{K+1} \left(\mathbb{E}\|x_k - x_{k-1}\|^2 + \mathbb{E}\|g_{k-1} - \nabla f(x_{k-1})\|^2 \right)$$
$$\leq \mathbb{E}Y_1 + \frac{1}{32\eta_1}\|x_1 - x_0\|^2 + \frac{\eta_1}{4}\|g_0 - \nabla f(x_0)\|^2$$
$$- \mathbb{E}Y_{K+1} + O(1).$$

To ensure that the the right-hand side is upper bounded by a constant, we need to show $\mathbb{E}Y_{k+1}$ is lower bounded. This is not immediate, because our use of a constant dual step size blocks the derivation of a uniform upper bound on the norm of dual variable λ_k. Lack of monotonicity of $\mathbb{E}Y_k$ also prevents us from using the estimates available in deterministic cases; see Hong (2016). In Lemma B.8, we show that the *almost* monotonicity of $\mathbb{E}Y_k$ given in (2.3) is sufficient to show lower boundedness. This fact leads to the result. □

With this result, we can use the standard reductions of Corollary B.5 to prove Theorem 2.1. It is worth noting that most of the estimations in the analysis would simplify if we were to fix all η_k, α_k at values that depend on the final iterate K (or equivalently ε). However, such choices would not suffice to show lower boundedness of $\mathbb{E}Y_k$ mentioned above, which is necessary to obtain the right constants. Our use of variable step sizes also allows us to derive an "any-time" algorithm with no need to set an accuracy ε in advance. For example, Huang et al. (2019) needed to assume uniform boundedness of $\|\nabla f(x_k)\|^2$ which trivially implies a lower bound for the potential; see (Huang et al., 2019, eq. (69)). However, this assumption does not hold normally and the limitations of bounded gradient assumption are well-known; see, for example, (Yang et al., 2023, Section 3), Faw et al. (2022). Our analysis does not need this restriction to obtain a lower bound for the potential.

3 STOCHASTIC CONSTRAINTS: PROBLEM (III)

3.1 Algorithm and the Main Result

In this section, we address (III), restated here as

$$\min_{x \in X} \{f(x) := \mathbb{E}[\tilde{f}(x, \xi)]\}$$
$$\text{subject to } c_i(x) := \mathbb{E}_\zeta[\tilde{c}_i(x, \zeta)] = 0, \quad \forall i \in \{1, \ldots, m\},$$

where X is convex and closed, f and c_i, $i = 1, 2, \ldots, m$ are smooth functions, $c(x) = (c_1(x), \ldots, c_m(x))^\top$, and ∇c is the Jacobian.[5] The notation P_X denotes projection onto X. In addition to the oracle model described in (1.5), in this section we also have access to $\tilde{\nabla} c_i$ such that $\mathbb{E}[\tilde{\nabla} c_i(x, \zeta)] = \nabla c_i(x)$. We assume that there are constants $\tilde{L}_{\nabla f}$, $\tilde{L}_{\nabla c}$, and $\tilde{L}_c$ such that for all x, y, we have

$$\begin{aligned} \mathbb{E}_\xi\|\tilde{\nabla} f(x, \xi) - \tilde{\nabla} f(y, \xi)\|^2 &\leq \tilde{L}^2_{\nabla f}\|x - y\|^2, \\ \mathbb{E}_\zeta\|\tilde{\nabla} c(x, \zeta) - \tilde{\nabla} c(y, \zeta)\|^2 &\leq \tilde{L}^2_{\nabla c}\|x - y\|^2, \\ \mathbb{E}_\zeta\|\tilde{c}(x, \zeta) - \tilde{c}(y, \zeta)\|^2 &\leq \tilde{L}^2_c\|x - y\|^2. \end{aligned} \tag{A2}$$

These conditions are stronger than mere smoothness of f, c but are necessary for variance reduction in general (Arjevani et al., 2022). Other recent works (for example, (Boob et al., 2022)) do not use this assumption but obtain a slightly worse complexity result (see Section 1.1).

Algorithm 2 is based on the quadratic penalty function

$$Q_\rho(x) = f(x) + \frac{\rho}{2} \sum_{i=1}^m (c_i(x))^2.$$

The variance reduced estimator g_{k+1} is based on STORM (Cutkosky and Orabona, 2019). The quadratic terms $(c_i(x))^2$ are not in the suitable form to apply SGD due to their *compositional* structure. However, it is a special form for which simply using independent samples can give an unbiased sample for the gradient. (This observation appeared in the recent independent work (Li et al., 2023).) Let us define the stochastic oracle

$$\begin{aligned} &\tilde{\nabla} Q_\rho(x, B) \\ &= \tilde{\nabla} f(x, \xi^0) + \rho \sum_{i=1}^m \tilde{\nabla} c_i(x, \zeta^1)\tilde{c}_i(x, \zeta^2) \end{aligned} \tag{3.1}$$

where $B = (\xi^0, \zeta^1, \zeta^2) \in \Xi \times Z^2$ with ζ^1 and ζ^2 i.i.d. We then have that $\mathbb{E}[\tilde{\nabla} f(x, \xi^0)] = \nabla f(x_k)$. Additionally, $\forall i = 1, 2, \ldots, m$, we have

$$\begin{aligned} &\mathbb{E}_{\zeta^1, \zeta^2}[\tilde{\nabla} c_i(x, \zeta^1)\tilde{c}_i(x, \zeta^2)] \\ &= \mathbb{E}_{\zeta^1}[\tilde{\nabla} c_i(x, \zeta^1)]\mathbb{E}_{\zeta^2}[\tilde{c}_i(x, \zeta^2)] \\ &= \nabla c_i(x) c_i(x), \end{aligned}$$

where the first step is by independence of ζ^1 and ζ^2. Hence, we have $\mathbb{E}\tilde{\nabla} Q_\rho(x) = \nabla Q_\rho(x)$. We assume that there are positive constants σ_f, $\sigma_{\nabla c}$, and σ_c such that

$$\begin{aligned} \mathbb{E}\|\tilde{\nabla} f(x, \xi) - \nabla f(x)\|^2 &\leq \sigma_f^2, \\ \mathbb{E}\|\tilde{\nabla} c(x, \xi) - \nabla c(x)\|^2 &\leq \sigma_{\nabla c}^2, \\ \mathbb{E}\|c(x, \xi) - c(x)\|^2 &\leq \sigma_c^2, \end{aligned} \tag{A3}$$

a set of assumptions also made in (Boob et al., 2022, eq. (2.9)). Other assumptions include the following:

$$\begin{aligned} &\|\nabla c(x)\| \leq C_{\nabla c}, \quad \|c(x)\| \leq C_c, \\ &\|\tilde{\nabla} c(x, \zeta)\| \leq \tilde{C}_{\nabla c}, \quad \|c(x, \zeta)\| \leq \tilde{C}_c, \\ &|f(x_k)| \leq B_f, \quad \|\nabla f(x)\| \leq C_{\nabla f}, \\ &Q_\rho(x) \geq \underline{Q} > -\infty \quad \forall \rho, x. \end{aligned} \tag{A4}$$

[5] One can consider $m = 1$ in the first reading for simplicity.

Algorithm 2 Stochastic Linearized Quadratic Penalty Method with Variance Reduction for (III)

Input: Initialize $x_1 \in X$ and $g_1 = \tilde{\nabla} Q_{\rho_1}(x_1, B_1)$ and α_k, ρ_k, η_k as in Theorem 3.1
for $k = 1, 2, \dots$ **do**
  $x_{k+1} = P_X(x_k - \eta_k g_k)$
  Sample ξ^0_{k+1}, ζ^1_{k+1}, ζ^2_{k+1} to get $B_{k+1} = (\xi^0_{k+1}, \zeta^1_{k+1}, \zeta^2_{k+1}) \in \Xi \times Z^2$ where ζ^1_{k+1} and ζ^2_{k+1} are i.i.d.
  $\tilde{\nabla} Q_\rho(x, B_{k+1}) = \tilde{\nabla} f(x, \xi^0_{k+1}) + \rho \sum_{i=1}^m \tilde{\nabla} c_i(x, \zeta^1_{k+1}) \tilde{c}_i(x, \zeta^2_{k+1})$
  $g_{k+1} = \tilde{\nabla} Q_{\rho_{k+1}}(x_{k+1}, B_{k+1}) + (1 - \alpha_{k+1})(g_k - \tilde{\nabla} Q_{\rho_k}(x_k, B_{k+1}))$
end for

These boundedness assumptions are widespread for nonconvex constrained problems, even with deterministic objective and constraints, see for example (Sahin et al., 2019; Lin et al., 2022a; Li et al., 2021).

Under these assumptions, we have that $x \mapsto Q_{\rho_k}(x)$ is L_{ρ_k}-smooth with $L_{\rho_k} = \rho_k(L_{\nabla f} + m(C_c L_{\nabla c} + C_{\nabla c} L_c))$ which, for example, we can see by direct calculation on the gradient. For variance reduction, we use

$$\begin{aligned} &\mathbb{E}_B \|\tilde{\nabla} Q_{\rho_k}(x, B) - \tilde{\nabla} Q_{\rho_k}(y, B)\|^2 \\ &\leq \tilde{L}^2_{\rho_k} \|x - y\|^2, \end{aligned} \tag{3.2}$$

where $\tilde{L}^2_\rho = \tilde{L}\rho^2$, $\tilde{L} := 4\tilde{L}^2_{\nabla f} + 4m^2(\tilde{C}^2_c \tilde{L}^2_{\nabla c} + \tilde{C}^2_{\nabla c} \tilde{L}^2_c)$, with $\rho_k \geq 1$. This can be shown the same way as the L_ρ, by using (A2) and (A4) (see also (C.3) and (C.2)).

We also use a generalization of the full rank assumption on the Jacobian. Recall that without the set inclusion constraint $x \in X$, this means $\|\nabla c(x_k)^\top c(x_k)\| \geq \delta |c(x_k)|$. With constraints, we assume

$$\mathsf{d}(\nabla c(x_k)^\top c(x_k), -N_X(x_k)) \geq \delta \|c(x_k)\|. \tag{A5}$$

This assumption is common in the existing literature of deterministic or stochastic algorithms with nonconvex functional constraints, see e.g., Sahin et al. (2019); Li et al. (2021); Lin et al. (2022a); Li et al. (2023). Let us note that this is implied by assuming LICQ in the whole space. Bolte et al. (2018) make a similar assumption that they term uniform regularity. Lin et al. (2022a) considered the relationship of this assumption with Kurdyka-Łojasiewicz and constraint qualifications.

We state now the main result for this section.

Theorem 3.1. *Let the assumptions in (A2), (A3), (A4), (A5) hold. Set the parameters of Algorithm 2 as*

$$\eta_k = \frac{1}{9\tilde{L}\rho(k+1)^{3/5}}, \rho_k = \rho k^{1/5}, \alpha_{k+1} = \frac{72}{81(k+1)^{4/5}},$$

for some $\rho > 1$. Then, there exists λ such that

$$\begin{aligned} \mathbb{E}\mathsf{d}(\nabla f(x_{\hat{k}+1}) + \nabla c(x_{\hat{k}+1})^\top \lambda, -N_X(x_{\hat{k}+1})) \leq \varepsilon, \\ \mathbb{E}\|c(x_{\hat{k}+1})\| \leq \varepsilon. \end{aligned}$$

with number of iterations K of Algorithm 2 bounded by $\widetilde{O}(\varepsilon^{-5})$ and $\hat{k}$ selected uniformly at random from $\{1, \dots, K\}$.

Remark 3.2. This is an iteration complexity result that directly translates to $\widetilde{O}(\varepsilon^{-5})$ sample complexity and computational complexity. This is because each iteration of Algorithm 2 requires one sample of each stochastic function f, 2 samples of $(c_i)_{i=1}^m$ and each iteration only involves simple projections to X and vector operations.

Remark 3.3. It is worth noting that it is straightforward to get rid of the logarithmic terms in the above bound by using parameters η_k, ρ_k that depend on the final iterate. However, this would require an initial preprocessing stage to get a near-feasible point for getting the best complexity. Shi et al. (2022) used this approach to study the deterministic constraints setting.

Remark 3.4. We state our results for the constrained case for simplicity. The extension to the *proximal case*, where we have the additional proper convex lower semicontinuous function instead of the constraint $x \in X$, is straightforward with our analysis template.

3.2 Analysis

As in the previous section, we start with the one iteration analysis of the algorithm for which we suppress some of the universal constants for readability. Statement of the lemma with details appears in Sec. C.2.

Lemma 3.5. *Under the assumptions in (A2), (A3), (A4), (A5) and the parameters (see also (3.2), (A4))*

$$\begin{aligned} \eta_k &= \frac{1}{9\tilde{L}\rho(k+1)^{3/5}}, \quad \rho_k = \rho k^{1/5}, \\ \alpha_{k+1} &= \frac{72}{81(k+1)^{4/5}}, \end{aligned}$$

for some constant $\rho > 1$, we have that

$$\begin{aligned} &\frac{\eta_k}{72} \mathbb{E}\mathsf{d}^2(\nabla f(x_{k+1}) + \rho_k \nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1})) \\ &\leq \mathbb{E}[Y_k - Y_{k+1} + |Q_{\rho_k}(x_{k+1}) - Q_{\rho_{k+1}}(x_{k+1})|] + \mathcal{E}_{k+1}, \end{aligned}$$

where

$$\begin{aligned} Y_{k+1} &= Q_{\rho_{k+1}}(x_{k+1}) \\ &\quad + \frac{1}{72\tilde{L}^2 \rho^2_{k+1} \eta_k} \|g_{k+1} - \nabla Q_{\rho_{k+1}}(x_{k+1})\|^2, \end{aligned}$$

$$\mathcal{E}_{k+1} = O\left(\frac{(\rho_k - \rho_{k+1})^2}{\rho_{k+1}^2 \eta_k}\right) + O\left(\frac{\alpha_{k+1}^2}{\eta_k}\right).$$

Remark 3.6. The first term of $\mathcal{E}_{k+1}$ has the order $O(k^{-7/5})$ and the second term of $\mathcal{E}_{k+1}$ has the order $O(k^{-1})$, therefore $\sum_{k=1}^K \mathcal{E}_{k+1} = O(\log(K+1))$.

Proof sketch of Theorem 3.1. In view of Remark C.2, it is easy to see that the only remaining piece we need on top of Lemma 3.5 is the control over the penalty parameter changes. For this, we show in Lemma C.3 that

$$\sum_{k=1}^{\infty} \mathbb{E}|Q_{\rho_k}(x_{k+1}) - Q_{\rho_{k+1}}(x_{k+1})| = O(1). \tag{3.3}$$

The main idea in this lemma is to use the estimate of Lemma 3.5 with the uniform upper bound on $\|c(x_k)\|^2$ from (A4) and take advantage of the decay of $|\rho_k - \rho_{k+1}|$ and (A5) to obtain (3.3). Using this estimate in Lemma 3.5 gives the result. □

4 EXTENSIONS

In this section, we consider two extensions (with details and proofs given in Sec. D.1 and D.2) and show how they follow by minor adjustments on our analysis.

4.1 Dual Variable Updates

In the context of nonconvex optimization with nonconvex functional constraints and ALM, the standard way of incorporating dual updates is to use small step sizes and large penalty parameters to ensure boundedness of the dual variable, see Li et al. (2021); Sahin et al. (2019); Shi et al. (2022). Rapidly increasing, unbounded, penalty parameter is then used to obtain feasibility guarantees. One exception is Xie and Wright (2021) which, unfortunately comes with worse complexity guarantees for first-order stationarity, compared to Li et al. (2021); Lin et al. (2022a). We considered the quadratic penalty method in the previous section for simplicity, but we show in this section that dual updates can be incorporated as done in Li et al. (2021); Sahin et al. (2019); Shi et al. (2022), with small step sizes. The modification compared to Algorithm 2 consists of changing the definition of g_{k+1} and incorporating a dual update step. In particular, we will change the step for g_{k+1} as

$$\begin{aligned} g_{k+1} = &\tilde{\nabla} Q_{\rho_{k+1}}(x_{k+1}, \lambda_{k+1}, B_{k+1}) \\ &+ (1-\alpha_{k+1})(g_k - \tilde{\nabla} Q_{\rho_k}(x_k, \lambda_k, B_{k+1})), \end{aligned}$$

where

$$\tilde{\nabla} Q_\rho(x, \lambda, B) = \tilde{\nabla} f(x, \xi^0) + \sum_{i=1}^m \tilde{\nabla} c_i(x, \zeta^1)\lambda_i + \rho \sum_{i=1}^m \tilde{\nabla} c_i(x, \zeta^1)\tilde{c}_i(x, \zeta^2).$$

We also add the dual update step as

$$\lambda_{k+2,i} = \lambda_{k+1,i} + \gamma_{k+1,i}\tilde{c}_i(x_{k+1}, \zeta_{k+1}^2), \tag{4.1}$$

for all $i \in \{1, \dots, m\}$. As alluded earlier, dual steps generally require a decaying step size as γ_{k+1} (or clipping the contribution of the previous dual parameter by a constant amount as in Lu (2022)) for getting the best-known guarantees.

Theorem 4.1. *For the algorithm described in Section 4.1, let*

$$\eta_k = \frac{1}{9\tilde{L}\rho(k+1)^{3/5}}, \quad \rho_k = \rho k^{1/5},$$

$$\gamma_{k,i} = \frac{\gamma}{k(\log(k+1))^2|\tilde{c}_i(x_k, \zeta_k^2)|}, \quad \alpha_{k+1} = \frac{72}{81(k+1)^{4/5}}$$

for some constant $\rho > 1$, $\gamma > 0$. Also let the assumptions in (A2), (A3), (A4), (A5) hold. We have that there exists λ such that

$$\mathbb{E}\mathsf{d}(\nabla f(x_{\hat{k}+1}) + \nabla c(x_{\hat{k}+1})^\top \lambda, -N_X(x_{\hat{k}+1})) \le \varepsilon,$$

with number of iterations bounded by $\widetilde{O}(\varepsilon^{-5})$. Moreover, we also have that $\mathbb{E}\|c(x_{\hat{k}+1})\| \le \varepsilon$.

4.2 Deterministic Functional Constraints

In this section, we consider the case when the constraints are deterministic. In this case, we set the parameters accordingly to get the complexity $\tilde{O}(\varepsilon^{-4})$.

Theorem 4.2. *For Algorithm 2, set*

$$\eta_k = \frac{1}{9\tilde{L}\rho(k+1)^{1/2}}, \quad \rho_k = \rho k^{1/4},$$

$$\alpha_{k+1} = \frac{72}{81(k+1)^{1/2}},$$

for some constant $\rho > 1$. Also let the assumptions in (A2), (A3), (A4), (A5) hold with a deterministic $c(x)$. We have that there exists λ such that

$$\mathbb{E}[\mathsf{d}(\nabla f(x_{\hat{k}+1}) + \nabla c(x_{\hat{k}+1})^\top \lambda, -N_X(x_{\hat{k}+1}))] \le \varepsilon,$$
$$\mathbb{E}\|c(x_{\hat{k}+1})\| \le \varepsilon,$$

with number of iterations bounded by $\widetilde{O}(\varepsilon^{-4})$.

We note that a similar result with a single-loop algorithm is obtained in Shi et al. (2022) with parameters depending on the last iteration (or equivalently, on the final accuracy). This results requires a pre-processing step to get an almost feasible point to get the complexity $O(\varepsilon^{-4})$, which deteriorates to $O(\varepsilon^{-5})$ otherwise. Hence, obtaining the more favorable complexity leads to a two-stage approach and also needing to set the final accuracy. Our approach leads to an algorithm that is both single stage and any-time.

5 CONCLUSIONS & OPEN QUESTIONS

This paper focused on improving the understanding of single-loop algorithms for stochastic optimization with functional constraints belonging to three different classes highlighted in Section 1. In contrast, *(i)* for nonlinearly constrained problems, previous work relied on double loop algorithms that have non-implementable stopping criteria for their inner loops (unless strong assumptions for access to structural constants made), *(ii)* for linearly constrained problems, previous work relied on increasing batch sizes with restrictive assumptions on the gradient norm. Our work helps overcome these drawbacks and opens up some directions that we sketch below.

1. An important point that is already emphasized many times in our manuscript is the following: the existing analyses for ALM even for deterministic and nonlinearly constrained problems require increasing penalty parameters (or equivalently, large penalty parameters depending inversely on the target accuracy) and small dual updates to obtain best-known complexity guarantees. This can be observed by the analysis frameworks in these works which analyze ALM like a quadratic penalty method with perturbation due to dual parameter updates. This results in ALM guarantees to be worse than guarantees for quadratic penalty methods.

   This is also the reason why we used increasing penalty parameters in our Section 3. On the other hand, one of the main points for ALM historically was its ability to work with non-increasing penalty parameters. Xie and Wright (2021) provided an analysis with constant penalty parameters, albeit with a worse complexity guarantee. An important open question is to analyze ALM with constant penalty parameters (independent of target accuracy ε) and large dual step sizes for deterministic, nonlinearly constrained nonconvex problems and obtain the best-known $O(\varepsilon^{-3})$ complexity. This would also lead the way to design more efficient stochastic methods, to improve our results.

2. A surprising difficulty for nonconvex problems with linear constraints is incorporating projectable constraints on top of linear constraints. A recent discovery on this context have been the work of Zhang et al. (2022) that uses error bound theory and can solve problems with linear constraints and projectable constraints with single-loop methods and optimal complexity. It is interesting to investigate how the results in this paper can be extended to stochastic case or for problem with nonlinear and non-projectable constraints.

Acknowledgments

This research was supported in part by the NSF grant 2023239, the NSF grant 2224213, the AFOSR award FA9550-21-1-0084.

We are thankful to Julian Katz-Samuels and Jiawei Zhang for helpful discussions.

References

Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pages 1–50, 2022.

D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

J. Bolte, S. Sabach, and M. Teboulle. Nonconvex lagrangian-based optimization: monitoring schemes and global convergence. *Mathematics of Operations Research*, 43(4):1210–1232, 2018.

D. Boob, Q. Deng, and G. Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, pages 1–65, 2022.

R. I. Boţ and D.-K. Nguyen. The proximal alternating direction method of multipliers in the nonconvex setting: convergence analysis and rates. *Mathematics of Operations Research*, 45(2):682–712, 2020.

S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

F. E. Curtis, M. J. O'Neill, and D. P. Robinson. Worst-case complexity of an sqp method for nonlinear equality constrained stochastic optimization. *arXiv preprint arXiv:2112.14799*, 2021.

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

D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

A. Dener, M. A. Miller, R. M. Churchill, T. Munson, and C.-S. Chang. Training neural networks under physical constraints using a stochastic augmented lagrangian approach. *arXiv preprint arXiv:2009.07330*, 2020.

M. Faw, I. Tziotis, C. Caramanis, A. Mokhtari, S. Shakkottai, and R. Ward. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, pages 313–355. PMLR, 2022.

A. V. Fiacco and G. P. McCormick. *Nonlinear programming: sequential unconstrained minimization techniques*. John Wiley, 1968.

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

C. He, Z. Lu, and T. K. Pong. A newton-cg based augmented lagrangian method for finding a second-order stationary point of nonconvex equality constrained optimization with complexity guarantees. *SIAM Journal on Optimization*, 33(3):1734–1766, 2023.

M. R. Hestenes. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4 (5):303–320, 1969.

M. Hong. Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: Algorithms, convergence, and applications. *arXiv preprint arXiv:1604.00543*, 2016.

M. Hong, M. Razaviyayn, and J. Lee. Gradient primal-dual algorithm converges to second-order stationary solution for nonconvex distributed optimization over networks. In *International Conference on Machine Learning*, pages 2009–2018. PMLR, 2018.

F. Huang, S. Chen, and H. Huang. Faster stochastic alternating direction method of multipliers for nonconvex optimization. In *International conference on machine learning*, pages 2839–2848. PMLR, 2019.

Y. Huang and Q. Lin. Single-loop switching subgradient methods for non-smooth weakly convex optimization with non-smooth convex constraints. *arXiv preprint arXiv:2301.13314*, 2023.

K. Ji, M. Liu, Y. Liang, and L. Ying. Will bilevel optimizers benefit from loops. *Advances in Neural Information Processing Systems*, 35:3011–3023, 2022.

J. Katz-Samuels, J. B. Nakhleh, R. Nowak, and Y. Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pages 10848–10865. PMLR, 2022.

W. Kong, J. G. Melo, and R. D. Monteiro. Iteration complexity of a proximal augmented lagrangian method for solving nonconvex composite optimization problems with nonlinear convex constraints. *Mathematics of Operations Research*, 48(2):1066–1094, 2023.

F. Kupfer and E. W. Sachs. Numerical solution of a nonlinear parabolic control problem by a reduced sqp method. *Computational Optimization and Applications*, 1(1):113–135, 1992.

G. Lan. *First-order and stochastic optimization methods for machine learning*. Springer, 2020.

G. Lan and R. D. Monteiro. Iteration-complexity of first-order penalty methods for convex programming. *Mathematical Programming*, 138(1-2):115–139, 2013.

G. Lan and R. D. Monteiro. Iteration-complexity of first-order augmented lagrangian methods for convex programming. *Mathematical Programming*, 155 (1-2):511–547, 2016.

Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu. Rate-improved inexact augmented lagrangian method for constrained nonconvex optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2170–2178. PMLR, 2021.

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*, pages 1–31, 2023.

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, 2022a.

Z. Lin, H. Li, and C. Fang. Stochastic admm. In *Alternating Direction Method of Multipliers for Machine Learning*, pages 143–205. Springer, 2022b.

X. Liu, Q. Liu, S. Song, and J. Peng. A chance-constrained generative framework for sequence optimization. In *International Conference on Machine Learning*, pages 6271–6281. PMLR, 2020.

S. Lu. A single-loop gradient descent and perturbed ascent algorithm for nonconvex functional constrained optimization. In *International Conference on Machine Learning*, pages 14315–14357. PMLR, 2022.

R. Ma, Q. Lin, and T. Yang. Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints. In *International Conference on Machine Learning*, pages 6554–6564. PMLR, 2020.

A. Mokhtari, H. Hassani, and A. Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *Journal of machine learning research*, 2020.

J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, 2nd edition, 2006.

Y. Ouyang, Y. Chen, G. Lan, and E. Pasiliao Jr. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, 8(1):644–681, 2015.

M. J. Powell. A method for nonlinear constraints in minimization problems. *Optimization*, pages 283–298, 1969.

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.

A. Ruszczyński. A linearization method for nonsmooth stochastic programming problems. *Mathematics of Operations Research*, 12(1):32–49, 1987.

M. F. Sahin, A. Eftekhari, A. Alacaoglu, F. Latorre, and V. Cevher. An inexact augmented lagrangian framework for nonconvex optimization with nonlinear constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

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

K. Sun and A. Sun. Dual descent alm and admm. *arXiv preprint arXiv:2109.13214*, 2021.

Y. Xie and S. J. Wright. Complexity of proximal augmented lagrangian for nonconvex optimization with nonlinear equality constraints. *Journal of Scientific Computing*, 86(3):1–30, 2021.

Y. Xu. Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming. *SIAM Journal on Optimization*, 27 (3):1459–1484, 2017.

Y. Xu. Iteration complexity of inexact augmented lagrangian methods for constrained convex programming. *Mathematical Programming*, 185:199–244, 2021.

J. Yang, X. Li, I. Fatkhullin, and N. He. Two sides of one coin: the limits of untuned sgd and the power of adaptive methods. *arXiv preprint arXiv:2305.12475*, 2023.

Y. Yang, G. Scutari, D. P. Palomar, and M. Pesavento. A parallel decomposition method for nonconvex stochastic multi-agent optimization problems. *IEEE Transactions on Signal Processing*, 64 (11):2949–2964, 2016.

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778, 2019.

J. Zhang and Z.-Q. Luo. A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM Journal on Optimization*, 30(3):2272–2302, 2020.

J. Zhang and Z.-Q. Luo. A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization. *SIAM Journal on Optimization*, 32(3):2319–2346, 2022.

J. Zhang, P. Xiao, R. Sun, and Z. Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in neural information processing systems*, 33:7377–7389, 2020.

J. Zhang, W. Pu, and Z.-Q. Luo. On the iteration complexity of smoothed proximal alm for nonconvex optimization problem with convex constraints. *arXiv preprint arXiv:2207.06304*, 2022.

X. Zhang, J. Liu, Z. Zhu, and E. S. Bentley. Gtstorm: Taming sample, communication, and memory complexities in decentralized non-convex learning. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 271–280, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A PRELIMINARIES

Remarks regarding the stochastic oracle model in (1.5). One possible way to obtain $\tilde{\nabla} f$ as characterized in (1.5) is to compute $\nabla_x \tilde{f}(x,\xi)$ where $\tilde{f}$ is defined in (1.2). For this quantity to satisfy the requirement in (1.5), its expectation over ξ must be the full gradient. However, this would require the gradient and expectation operations to be interchangeable, which is not always true, especially in our nonconvex setting. It does however hold for functions in which the support of ξ is finite (that is, finite-sum functions). It also holds under fairly general conditions for smooth f. We make similar assumptions and apply similar conventions for the constraint functions c and $\tilde{c}$ and use the oracle model described in this paragraph. It is worth noting that this oracle access is standard in stochastic optimization with nonconvexity, see e.g., Ghadimi and Lan (2013); Arjevani et al. (2022); Cutkosky and Orabona (2019); Lan (2020).

A preliminary lemma. The variance reduction technique introduced in Cutkosky and Orabona (2019) uses a vector g_k at each k as a proxy for a gradient of an expectation function. The variable g_k accumulates information from earlier iterations and from many values of the random variables, thus has lower variance than the gradient evaluated at x_k and a single value of the random variable.

We need a lemma to bound the difference between g_k and the corresponding gradient, as it evolves across iterations. Here we state this lemma in a general form that can be applied in all the problem formulations considered in this paper. The lemma includes possibly iteration dependent functions to accommodate situations in which g_k represents the gradient of the augmented Lagrangian or quadratic penalty function with an iteration-dependent penalty parameter. This lemma builds on (Cutkosky and Orabona, 2019, Lemma 5, Theorem 2); see also Ruszczyński (1987); Mokhtari et al. (2020); Yang et al. (2016) for conceptually similar derivations in other settings.

Lemma A.1. *Let $G_k\colon \mathbb{R}^n \to \mathbb{R}^n$ and $\tilde{G}_k(x,\xi)$ be such that $\mathbb{E}_\xi[\tilde{G}_k(x_k,\xi)] = G_k(x_k)$, $\mathbb{E}_\xi[\tilde{G}_{k+1}(x_{k+1},\xi)] = G_{k+1}(x_{k+1})$. Define $g_{k+1} = \tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) + (1-\alpha_{k+1})(g_k - \tilde{G}_k(x_k,\xi_{k+1}))$ for $k \geq 0$ and $g_0 \in \mathbb{R}^n$. We then have*

$$\begin{aligned}\mathbb{E}_k\|g_{k+1} - G_{k+1}(x_{k+1})\|^2 \leq{}& (1-\alpha_{k+1})^2\|g_k - G_k(x_k)\|^2\\ &+ 7\mathbb{E}_k\|\tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) - \tilde{G}_{k+1}(x_k,\xi_{k+1})\|^2\\ &+ 42\mathbb{E}_k\|\tilde{G}_{k+1}(x_k,\xi_{k+1}) - \tilde{G}_k(x_k,\xi_{k+1})\|^2\\ &+ 3\alpha_{k+1}^2\mathbb{E}_k\|G_k(x_k) - \tilde{G}_k(x_k,\xi_{k+1})\|^2,\end{aligned}$$

where $\mathbb{E}_k$ is the expectation conditioned on all the history up to and including x_{k+1},ξ_k.

Proof. We have, by subtracting $G_{k+1}(x_{k+1})$ from both sides of the definition of g_{k+1}, that

$$g_{k+1} - G_{k+1}(x_{k+1}) = \tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) + (1-\alpha_{k+1})(g_k - \tilde{G}_k(x_k,\xi_{k+1})) - G_{k+1}(x_{k+1}).$$

On this identity, we use the simple decomposition

$$(1-\alpha_{k+1})(g_k - \tilde{G}_k(x_k,\xi_{k+1})) = (1-\alpha_{k+1})(g_k - G_k(x_k)) + (1-\alpha_{k+1})(G_k(x_k) - \tilde{G}_k(x_k,\xi_{k+1})),$$

to obtain

$$\begin{aligned}g_{k+1} - G_{k+1}(x_{k+1}) = (1-\alpha_{k+1})(g_k - G_k(x_k)) &+ (\tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) - G_{k+1}(x_{k+1}))\\ &+ (1-\alpha_{k+1})(G_k(x_k) - \tilde{G}_k(x_k,\xi_{k+1})).\end{aligned}$$

We next take the squared norm of both sides and expand the right-hand side

$$\begin{aligned}\|g_{k+1} - G_{k+1}(x_{k+1})\|^2 = (1-\alpha_{k+1})^2\|g_k - G_k(x_k)\|^2&\\ + 2(1-\alpha_{k+1})\langle g_k - G_k(x_k), \tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) - G_{k+1}(x_{k+1})\rangle&\\ + 2(1-\alpha_{k+1})^2\langle g_k - G_k(x_k), G_k(x_k) - \tilde{G}_k(x_k,\xi_{k+1})\rangle&\\ + \|\tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) - G_{k+1}(x_{k+1}) + (1-\alpha_{k+1})(G_k(x_k) - \tilde{G}_k(x_k,\xi_{k+1}))\|^2.&\end{aligned} \tag{A.1}$$

We now take conditional expectation of this equality, where $\mathbb{E}_k$ is as defined in the lemma statement:

$$\begin{aligned}\mathbb{E}_k\|g_{k+1} - G_{k+1}(x_{k+1})\|^2 &= (1-\alpha_{k+1})^2\|g_k - G_k(x_k)\|^2 \\ &\quad + \mathbb{E}_k\|\tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) - G_{k+1}(x_{k+1}) + (1-\alpha_{k+1})(G_k(x_k) - \tilde{G}_k(x_k,\xi_{k+1}))\|^2.\end{aligned} \tag{A.2}$$

This equality is because the inner product terms on (A.1) disappear after taking condition expectation: we have $g_k - G_k(x_k)$ is deterministic when we condition on x_{k+1} and also $\mathbb{E}_k[\tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) - G_{k+1}(x_{k+1})] = 0$ and $\mathbb{E}_k[G_k(x_k) - G_k(x_k,\xi_{k+1}))] = 0$, by the requirements on $\tilde{G}_k$ and $\tilde{G}_{k+1}$ given in the lemma.

For the last term on the right-hand side of (A.2), we use Young's inequalities to obtain

$$\begin{aligned}&\mathbb{E}_k\|\tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) - G_{k+1}(x_{k+1}) + (1-\alpha_{k+1})(G_k(x_k) - \tilde{G}_k(x_k,\xi_{k+1}))\|^2 \\ &\le 3\mathbb{E}_k\|G_{k+1}(x_{k+1}) - G_k(x_k)\|^2 + 3\mathbb{E}_k\|\tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) - \tilde{G}_k(x_k,\xi_{k+1})\|^2 \\ &\quad + 3\alpha_{k+1}^2\mathbb{E}_k\|G_k(x_k) - \tilde{G}_k(x_k,\xi_{k+1})\|^2 \\ &\le 6\mathbb{E}_k\|\tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) - \tilde{G}_k(x_k,\xi_{k+1})\|^2 + 3\alpha_{k+1}^2\mathbb{E}_k\|G_k(x_k) - \tilde{G}_k(x_k,\xi_{k+1})\|^2,\end{aligned} \tag{A.3}$$

where the final bound joins the first two terms in the previous bound, by Jensen's inequality, since

$$\mathbb{E}_k[\tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) - \tilde{G}_k(x_k,\xi_{k+1})] = G_{k+1}(x_{k+1}) - G_k(x_k).$$

For the first term on the right-hand side of (A.3), we add and subtract $\tilde{G}_{k+1}(x_k,\xi_{k+1})$ and use Young's inequality to find that

$$\begin{aligned}&\|\tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) - \tilde{G}_k(x_k,\xi_{k+1})\|^2 \\ &\le \frac{7}{6}\|\tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) - \tilde{G}_{k+1}(x_k,\xi_{k+1})\|^2 + 7\|\tilde{G}_{k+1}(x_k,\xi_{k+1}) - \tilde{G}_k(x_k,\xi_{k+1})\|^2.\end{aligned}$$

Thus (A.3) becomes

$$\begin{aligned}&\mathbb{E}_k\|\tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) - G_{k+1}(x_{k+1}) + (1-\alpha_{k+1})(G_k(x_k) - \tilde{G}_k(x_k,\xi_{k+1}))\|^2 \\ &\quad \le 7\mathbb{E}_k\|\tilde{G}_{k+1}(x_{k+1},\xi_{k+1}) - \tilde{G}_{k+1}(x_k,\xi_{k+1})\|^2 \\ &\qquad + 42\mathbb{E}_k\|\tilde{G}_{k+1}(x_k,\xi_{k+1}) - \tilde{G}_k(x_k,\xi_{k+1})\|^2 \\ &\qquad + 3\alpha_{k+1}^2\mathbb{E}_k\|G_k(x_k) - \tilde{G}_k(x_k,\xi_{k+1})\|^2.\end{aligned}$$

By using this inequality to bound the last term on the right-hand side of (A.2), we obtain the result. □

B LINEAR CONSTRAINTS: PROBLEM (I)

B.1 One-step recursion on augmented Lagrangian

We recall the augmented Lagrangian function for (I) as

$$\mathcal{L}_\rho(x,\lambda) := f(x) + \langle\lambda, Ax - b\rangle + \frac{\rho}{2}\|Ax - b\|^2, \tag{B.1}$$

for $\rho > 0$. In this section, we prove the following result concerning the change in $\mathbb{E}\mathcal{L}_\rho(x_k,\lambda_k)$ over one iteration. Note that the expectation $\mathbb{E}$ is with respect to the randomness of ξ_k for $k = 1, 2, \ldots$. Our result makes use of the Lipschitz constant of $\nabla_x\mathcal{L}_\rho(\cdot,\lambda_k)$, which is

$$L_\rho := L_f + \rho\|A\|^2. \tag{B.2}$$

In the rest of this section, we make frequent use of the matrix Q_k defined by

$$Q_k := \eta_k^{-1}I - \rho A^\top A. \tag{B.3}$$

Lemma B.1. *Let the assumptions in (A1) hold. Then for any $k \geq 1$, we have that*

$$\begin{aligned}\mathbb{E}\mathcal{L}_\rho(x_{k+1},\lambda_{k+1}) \leq \mathbb{E}\mathcal{L}_\rho(x_k,\lambda_k) + \left(\frac{L_\rho}{2} - \frac{1}{2\eta_{k+1}}\right)\mathbb{E}\|x_{k+1}-x_k\|^2 + \frac{\eta_{k+1}}{2}\mathbb{E}\|g_k - \nabla f(x_k)\|^2 \\ + \frac{1}{\rho}\mathbb{E}\|\lambda_{k+1}-\lambda_k\|^2.\end{aligned} \tag{B.4}$$

With the definition of Q_{k+1} as in (B.3), and assuming that η_{k+1} is chosen to ensure that $Q_{k+1} \succ 0$, we also have

$$\begin{aligned}\mathbb{E}\|\lambda_{k+1}-\lambda_k\|^2 \leq \frac{1}{\delta}\Bigg(6L_f^2\mathbb{E}\|x_k - x_{k-1}\|^2 + 6\alpha_k^2V^2 + 6\alpha_k^2\mathbb{E}\|g_{k-1}-\nabla f(x_{k-1})\|^2 \\ + 3\mathbb{E}\left\|x_{k+1}-2x_k+x_{k-1}\right\|^2_{Q_{k+1}^\top Q_{k+1}} + 3\left(\eta_k^{-1}-\eta_{k+1}^{-1}\right)^2\mathbb{E}\|x_k-x_{k-1}\|^2\Bigg).\end{aligned}$$

Before proving this result, we state and prove an immediate corollary.

Corollary B.2. *Under the same assumptions as Lemma B.1, for any positive constants c_1 and c_4, and with η_k defined as in (B.19) (where this definition depends on c_1, c_5, and other constants), we have*

$$\begin{aligned}\mathbb{E}\mathcal{L}_\rho(x_{k+1},\lambda_{k+1}) \leq \mathbb{E}\mathcal{L}_\rho(x_k,\lambda_k) + \frac{\eta_{k+1}}{2}\mathbb{E}\|g_k-\nabla f(x_k)\|^2 + \frac{6(1+c_1)\alpha_k^2}{\rho\delta}\mathbb{E}\|g_{k-1}-\nabla f(x_{k-1})\|^2 \\ + \left(\frac{L_\rho}{2} - \frac{1}{2\eta_{k+1}}\right)\mathbb{E}\|x_{k+1}-x_k\|^2 + \frac{(6+c_4)(1+c_1)L_f^2}{\rho\delta}\mathbb{E}\|x_k-x_{k-1}\|^2 \\ + \frac{3(1+c_1)}{\rho\delta}\mathbb{E}\|x_{k+1}-2x_k+x_{k-1}\|^2_{Q_{k+1}^\top Q_{k+1}} + \frac{6(1+c_1)\alpha_k^2V^2}{\rho\delta} - \frac{c_1}{\rho}\mathbb{E}\|\lambda_{k+1}-\lambda_k\|^2.\end{aligned}$$

Proof of Corollary B.2. The proof is immediate after adding and subtracting $\frac{c_1}{\rho}\mathbb{E}\|\lambda_{k+1}-\lambda_k\|^2$, using the upper bound of $\mathbb{E}\|\lambda_{k+1}-\lambda_k\|^2$ from Lemma B.1 and Fact B.7 to bound $3(\eta_k^{-1}-\eta_{k+1}^{-1})^2 \leq c_4L_f^2$. □

We now return to the proof of Lemma B.1.

Proof of Lemma B.1. By Lipschitz continuity of $\nabla\mathcal{L}_\rho(\cdot,\lambda_k)$, we have

$$\mathcal{L}_\rho(x_{k+1},\lambda_k) \leq \mathcal{L}_\rho(x_k,\lambda_k) + \langle\nabla_x\mathcal{L}_\rho(x_k,\lambda_k), x_{k+1}-x_k\rangle + \frac{L_\rho}{2}\|x_{k+1}-x_k\|^2. \tag{B.5}$$

By the definitions of $\nabla_x\mathcal{L}_\rho$ and x_{k+1} in Algorithm 1, we have that

$$\begin{aligned}\langle\nabla_x\mathcal{L}_\rho(x_k,\lambda_k), x_{k+1}-x_k\rangle &= \langle g_k + A^\top\lambda_k + \rho A^\top(Ax_k-b), x_{k+1}-x_k\rangle + \langle\nabla f(x_k)-g_k, x_{k+1}-x_k\rangle \\ &\leq -\frac{1}{\eta_{k+1}}\|x_{k+1}-x_k\|^2 + \frac{\eta_{k+1}}{2}\|\nabla f(x_k)-g_k\|^2 + \frac{1}{2\eta_{k+1}}\|x_{k+1}-x_k\|^2 \\ &= -\frac{1}{2\eta_{k+1}}\|x_{k+1}-x_k\|^2 + \frac{\eta_{k+1}}{2}\|\nabla f(x_k)-g_k\|^2,\end{aligned}$$

where the last two terms on the second line are from Young's inequality. By substituting in (B.5) and collecting like terms, we obtain

$$\mathcal{L}_\rho(x_{k+1},\lambda_k) \leq \mathcal{L}_\rho(x_k,\lambda_k) + \left(\frac{L_\rho}{2} - \frac{1}{2\eta_{k+1}}\right)\|x_{k+1}-x_k\|^2 + \frac{\eta_{k+1}}{2}\|g_k-\nabla f(x_k)\|^2. \tag{B.6}$$

We also have by the definition of $\mathcal{L}_\rho$ in (B.1) and λ_{k+1} in Algorithm 1 that

$$\mathcal{L}_\rho(x_{k+1},\lambda_{k+1}) - \mathcal{L}_\rho(x_{k+1},\lambda_k) = \langle\lambda_{k+1}-\lambda_k, Ax_{k+1}-b\rangle = \frac{1}{\rho}\|\lambda_{k+1}-\lambda_k\|^2.$$

By using this identity in (B.6), we obtain the first result (B.4) after taking expectation.

We now bound $\mathbb{E}\|\lambda_{k+1}-\lambda_k\|^2$ to get the second result. By using the definitions of x_{k+1} and λ_{k+1} in Algorithm 1, we obtain

$$\begin{aligned} x_{k+1} &= x_k - \eta_{k+1}(g_k + A^\top \lambda_k + \rho A^\top (Ax_k - b)) \\ &= x_k - \eta_{k+1}(g_k + A^\top \lambda_{k+1} + \rho A^\top A(x_k - x_{k+1})) \qquad \text{(B.7)} \\ \iff \left(\eta_{k+1}^{-1} I - \rho A^\top A\right)(x_{k+1} - x_k) &= -g_k - A^\top \lambda_{k+1}. \qquad \text{(B.8)} \end{aligned}$$

Replacing k by $k-1$ in this identity, we have

$$\begin{aligned} \left(\eta_k^{-1} I - \rho A^\top A\right)(x_k - x_{k-1}) &= -g_{k-1} - A^\top \lambda_k \\ \iff \left(\eta_{k+1}^{-1} I - \rho A^\top A\right)(x_k - x_{k-1}) + \left(\eta_k^{-1} - \eta_{k+1}^{-1}\right)(x_k - x_{k-1}) &= -g_{k-1} - A^\top \lambda_k. \qquad \text{(B.9)} \end{aligned}$$

Subtracting (B.9) from (B.8) (to use a similar technique to one used in Hong (2016); Hong et al. (2018) with the change of identifying the error coming from iteration-dependent parameters) gives

$$\begin{aligned} &\left(\eta_{k+1}^{-1} I - \rho A^\top A\right)(x_{k+1} - x_k - (x_k - x_{k-1})) - \left(\eta_k^{-1} - \eta_{k+1}^{-1}\right)(x_k - x_{k-1}) \\ &= g_{k-1} - g_k + A^\top(\lambda_k - \lambda_{k+1}). \qquad \text{(B.10)} \end{aligned}$$

By rearranging, taking squared norm of each side, and using Young's inequality, we obtain

$$\begin{aligned} \|A^\top(\lambda_k - \lambda_{k+1})\|^2 \le 3\|g_k - g_{k-1}\|^2 &+ 3\left\|\left(\eta_{k+1}^{-1} I - \rho A^\top A\right)(x_{k+1} - x_k - (x_k - x_{k-1}))\right\|^2 \\ &+ 3\left(\eta_k^{-1} - \eta_{k+1}^{-1}\right)^2 \|x_k - x_{k-1}\|^2. \qquad \text{(B.11)} \end{aligned}$$

For the first term in this bound, by the definition of g_{k+1} from Algorithm 1, we have that

$$g_{k+1} - g_k = \tilde{\nabla} f(x_{k+1}, \xi_{k+1}) - \tilde{\nabla} f(x_k, \xi_{k+1}) + \alpha_{k+1}(\tilde{\nabla} f(x_k, \xi_{k+1}) - g_k).$$

By taking squared norms, using Young's inequality and Lipschitzness of $\tilde{\nabla} f(\cdot, \xi)$ from (A1), we obtain

$$\begin{aligned} \mathbb{E}\|g_{k+1} - g_k\|^2 &\le 2L_f^2 \mathbb{E}\|x_{k+1} - x_k\|^2 + 2\alpha_{k+1}^2 \mathbb{E}\|\tilde{\nabla} f(x_k, \xi_{k+1}) - g_k\|^2 \\ &= 2L_f^2 \mathbb{E}\|x_{k+1} - x_k\|^2 + 2\alpha_{k+1}^2 \mathbb{E}\|\tilde{\nabla} f(x_k, \xi_{k+1}) - \nabla f(x_k)\|^2 + 2\alpha_{k+1}^2 \mathbb{E}\|g_k - \nabla f(x_k)\|^2 \\ &\le 2L_f^2 \mathbb{E}\|x_{k+1} - x_k\|^2 + 2\alpha_{k+1}^2 V^2 + 2\alpha_{k+1}^2 \mathbb{E}\|g_k - \nabla f(x_k)\|^2, \qquad \text{(B.12)} \end{aligned}$$

where V is the bound on variance from (A1). To obtain the equality in the derivation above, we used the tower rule: since $\mathbb{E}_{\xi_{k+1}}[\tilde{\nabla} f(x_k, \xi_{k+1})] = \nabla f(x_k)$ and $g_k, \nabla f(x_k)$ are deterministic when we are taking the conditional expectation, thus $\mathbb{E}\langle \tilde{\nabla} f(x_k, \xi_{k+1}) - \nabla f(x_k), g_k - \nabla f(x_k)\rangle = 0$.

Using the inequality (B.12) in (B.11) (with index $k-1$ instead of k), we have after taking expectation that

$$\begin{aligned} \mathbb{E}\|A^\top(\lambda_k - \lambda_{k+1})\|^2 \le 6L_f^2 \mathbb{E}\|x_k - x_{k-1}\|^2 &+ 6\alpha_k^2 V^2 + 6\alpha_k^2 \mathbb{E}\|g_{k-1} - \nabla f(x_{k-1})\|^2 \\ &+ 3\mathbb{E}\left\|\left(\eta_{k+1}^{-1} I - \rho A^\top A\right)(x_{k+1} - x_k - (x_k - x_{k-1}))\right\|^2 \\ &+ 3\left(\eta_k^{-1} - \eta_{k+1}^{-1}\right)^2 \mathbb{E}\|x_k - x_{k-1}\|^2. \qquad \text{(B.13)} \end{aligned}$$

Since $\lambda_{k+1} - \lambda_k$ is in the range of A and δ is the smallest nonzero eigenvalue of $A^\top A$ (see also (1.3)), we have the result after using the definition of Q_{k+1} from (B.3). □

B.2 One-step recursion on feasibility and iterate difference

Lemma B.3. *Let the assumptions in (A1) hold. For all $k \ge 0$, assume that η_{k+1} is chosen to ensure that $Q_{k+1} \succ 0$ (where Q_{k+1} is defined in (B.3)). Then for any $k \ge 1$, we have that*

$$\begin{aligned} &\frac{\rho}{2\eta_{k+2}}\mathbb{E}\|Ax_{k+1} - b\|^2 + \frac{1}{2\eta_{k+1}}\mathbb{E}\|x_{k+1} - x_k\|_{Q_{k+1}}^2 \\ &\le \frac{\rho}{2\eta_{k+1}}\mathbb{E}\|Ax_k - b\|^2 + \frac{1}{2\eta_k}\mathbb{E}\|x_k - x_{k-1}\|_{Q_k}^2 \end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{2\eta_{k+1}}\mathbb{E}\|x_{k+1} - 2x_k + x_{k-1}\|^2_{Q_{k+1}} + \frac{\eta_{k+2}^{-1} - \eta_{k+1}^{-1}}{2\rho}\mathbb{E}\|\lambda_{k+1} - \lambda_k\|^2 \\
&+ \left(\frac{L_f}{2\eta_{k+1}} + \frac{\eta_{k+1}^{-1} - \eta_k^{-1}}{2\eta_{k+1}}\right)\mathbb{E}\|x_{k+1} - x_k\|^2 + \left(\frac{L_f}{\eta_{k+1}} + \frac{\eta_{k+1}^{-1} - \eta_k^{-1}}{2\eta_{k+1}} + \frac{\eta_{k+1}^{-2} - \eta_k^{-2}}{2}\right)\mathbb{E}\|x_k - x_{k-1}\|^2 \\
&+ \frac{\alpha_k^2}{L_f\eta_{k+1}}\mathbb{E}\|g_{k-1} - \nabla f(x_{k-1})\|^2 + \frac{\alpha_k^2 V^2}{L_f\eta_{k+1}}.
\end{aligned}
$$

Before proving this result, we state and prove an immediate corollary.

Corollary B.4. *Suppose the assumptions of Lemma B.3 hold and that c_1, c_2, c_3 are arbitrary positive constants, with η_k and $m > 0$ defined from (B.19) (where the definitions of η_k and m depend on c_1, c_2, c_3, and other constants). We have*

$$
\begin{aligned}
&\frac{\rho}{2\eta_{k+2}}\mathbb{E}\|Ax_{k+1} - b\|^2 + \frac{1}{2\eta_{k+1}}\mathbb{E}\|x_{k+1} - x_k\|^2_{Q_{k+1}} \\
&\leq \frac{\rho}{2\eta_{k+1}}\mathbb{E}\|Ax_k - b\|^2 + \frac{1}{2\eta_k}\mathbb{E}\|x_k - x_{k-1}\|^2_{Q_k} \\
&\quad - \frac{1}{2\eta_{k+1}}\mathbb{E}\|x_{k+1} - 2x_k + x_{k-1}\|^2_{Q_{k+1}} + \frac{c_1}{\rho m}\mathbb{E}\|\lambda_{k+1} - \lambda_k\|^2 \\
&\quad + \frac{(1+2c_3)L_f}{2\eta_{k+1}}\mathbb{E}\|x_{k+1} - x_k\|^2 + \frac{(1+c_2+c_3)L_f}{\eta_{k+1}}\mathbb{E}\|x_k - x_{k-1}\|^2 \\
&\quad + \frac{\alpha_k^2}{L_f\eta_{k+1}}\mathbb{E}\|g_{k-1} - \nabla f(x_{k-1})\|^2 + \frac{\alpha_k^2 V^2}{L_f\eta_{k+1}}.
\end{aligned}
$$

Proof of Corollary B.4. The result is immediate after using (B.37a), (B.37b) and (B.37c) in Fact B.7 on the result of Lemma B.3 and rearranging terms. □

We now return to Lemma B.3

Proof of Lemma B.3. For this recursion, we use a similar technique to Hong (2016); Hong et al. (2018), extended to use variable step sizes. The additional error terms will appear due to using stochastic gradients and variance reduction in our case, which were not considered in Hong (2016); Hong et al. (2018). We start from (B.10) and use the definition of λ_{k+1} in Algorithm 1 (that is, $\lambda_{k+1} - \lambda_k = \rho(Ax_{k+1} - b)$) and the definition (B.3) of Q_{k+1} to obtain

$$
g_k - g_{k-1} + \rho A^\top(Ax_{k+1} - b) + Q_{k+1}(x_{k+1} - 2x_k + x_{k-1}) + (\eta_{k+1}^{-1} - \eta_k^{-1})(x_k - x_{k-1}) = 0.
$$

We next take the inner product of this expression with $x_{k+1} - x_k$, divide by η_{k+1} and rearrange to get

$$
\begin{aligned}
\frac{\rho}{\eta_{k+1}}\langle Ax_{k+1} - b, A(x_{k+1} - x_k)\rangle &= \frac{1}{\eta_{k+1}}\langle g_{k-1} - g_k - Q_{k+1}(x_{k+1} - 2x_k + x_{k-1}), x_{k+1} - x_k\rangle \\
&\quad + \frac{\eta_{k+1}^{-1} - \eta_k^{-1}}{\eta_{k+1}}\langle x_{k-1} - x_k, x_{k+1} - x_k\rangle. \qquad \text{(B.14)}
\end{aligned}
$$

We now obtain bounds on four parts of this expression in turn.

1. By using the identity $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$ for any vectors a and b, we have that

$$
\begin{aligned}
&\frac{\rho}{\eta_{k+1}}\langle Ax_{k+1} - b, A(x_{k+1} - x_k)\rangle \\
&= \frac{\rho}{2\eta_{k+1}}\left(\|Ax_{k+1} - b\|^2 + \|A(x_{k+1} - x_k)\|^2 - \|Ax_k - b\|^2\right) \\
&\geq \frac{\rho}{2\eta_{k+1}}\left(\|Ax_{k+1} - b\|^2 - \|Ax_k - b\|^2\right)
\end{aligned}
$$

$$= \frac{\rho}{2\eta_{k+2}}\|Ax_{k+1}-b\|^2 - \frac{\rho}{2\eta_{k+1}}\|Ax_k-b\|^2 + \frac{\eta_{k+1}^{-1}-\eta_{k+2}^{-1}}{2\rho}\|\lambda_{k+1}-\lambda_k\|^2, \tag{B.15}$$

where the last step used the definition of λ_{k+1} from Algorithm 1 again, together with addition and subtraction of a term involving η_{k+2}^{-1}.

2. By using the same identity and the definition (B.3) of $Q_{k+1} \succ 0$, we also have

$$\begin{aligned}
&-\frac{1}{\eta_{k+1}}\langle Q_{k+1}(x_{k+1}-2x_k+x_{k-1}), x_{k+1}-x_k\rangle \\
&= -\frac{1}{\eta_{k+1}}\langle x_{k+1}-2x_k+x_{k-1}, x_{k+1}-x_k\rangle_{Q_{k+1}} \\
&= -\frac{1}{2\eta_{k+1}}\left(\|x_{k+1}-2x_k+x_{k-1}\|_{Q_{k+1}}^2 + \|x_{k+1}-x_k\|_{Q_{k+1}}^2 - \|x_k-x_{k-1}\|_{Q_{k+1}}^2\right) \\
&\le -\frac{1}{2\eta_{k+1}}\|x_{k+1}-2x_k+x_{k-1}\|_{Q_{k+1}}^2 - \frac{1}{2\eta_{k+1}}\|x_{k+1}-x_k\|_{Q_{k+1}}^2 + \frac{1}{2\eta_k}\|x_k-x_{k-1}\|_{Q_k}^2 \\
&\quad + \frac{1}{2}\left(\eta_{k+1}^{-2}-\eta_k^{-2}\right)\|x_k-x_{k-1}\|^2,
\end{aligned} \tag{B.16}$$

where the last step uses (B.3) together with $0 < \eta_{k+1} \le \eta_k$, i.e.,

$$\begin{aligned}
&\frac{1}{2\eta_{k+1}}\|x_k-x_{k-1}\|_{Q_{k+1}}^2 - \frac{1}{2\eta_k}\|x_k-x_{k-1}\|_{Q_k}^2 \\
&= \frac{1}{2}\left(\eta_{k+1}^{-2}-\eta_k^{-2}\right)\|x_k-x_{k-1}\|^2 + \langle x_k-x_{k-1}, \left(-\eta_{k+1}^{-1}+\eta_k^{-1}\right)\rho A^\top A(x_k-x_{k-1})\rangle \\
&\le \frac{1}{2}\left(\eta_{k+1}^{-2}-\eta_k^{-2}\right)\|x_k-x_{k-1}\|^2.
\end{aligned}$$

3. By Young's inequality, we have

$$\frac{\eta_{k+1}^{-1}-\eta_k^{-1}}{\eta_{k+1}}\langle x_{k-1}-x_k, x_{k+1}-x_k\rangle \le \frac{\eta_{k+1}^{-1}-\eta_k^{-1}}{2\eta_{k+1}}\left(\|x_k-x_{k-1}\|^2 + \|x_{k+1}-x_k\|^2\right). \tag{B.17}$$

4. By Young's inequality and (B.12) (replacing k by $k-1$), we get

$$\begin{aligned}
&\frac{1}{\eta_{k+1}}\mathbb{E}\langle g_{k-1}-g_k, x_{k+1}-x_k\rangle \\
&\le \frac{1}{2L_f\eta_{k+1}}\mathbb{E}\|g_k-g_{k-1}\|^2 + \frac{L_f}{2\eta_{k+1}}\mathbb{E}\|x_{k+1}-x_k\|^2 \\
&\le \frac{1}{2L_f\eta_{k+1}}\left(2L_f^2\mathbb{E}\|x_k-x_{k-1}\|^2 + 2\alpha_k^2V^2 + 2\alpha_k^2\mathbb{E}\|g_{k-1}-\nabla f(x_{k-1})\|^2\right) \\
&\quad + \frac{L_f}{2\eta_{k+1}}\mathbb{E}\|x_{k+1}-x_k\|^2.
\end{aligned} \tag{B.18}$$

We get the result by taking expectation of (B.14) and substituting from (B.15), (B.16), (B.17), and (B.18). □

B.3 Main result

In this section, for arbitrary positive values of c_1, c_2, c_3, c_4 we define the parameters η_k, m, ρ, and α_k in the following way (recall the definitions of L_f and δ from (A1) and (1.3)):

$$\begin{aligned}
c &= 121L_f^2,\\
m &= \min\left(\frac{1}{448L_f}, \frac{1}{32(1+c_2+c_3)L_f}, \frac{1}{8(1+2c_3)L_f}\right)\\
\rho &= \max\left(\frac{7(1+c_1)}{m\delta}, \frac{4(6+c_4)(1+c_1)L_f}{\delta}, \frac{168(1+c_1)L_f}{\delta}\right),\\
\eta &= \frac{1}{11(L_f+\rho\|A\|^2)},\\
k_0 &= \max\left(\left(\frac{10m}{3c_1\eta}\right)^2, \left(\frac{20}{3\eta c_2 L_f}\right)^2, \left(\frac{10}{3c_3L_f}\right)^2, \frac{400}{3\eta^2c_4L_f^2}, \left(\frac{20}{c\eta^2}\right)^4, \left(\frac{50}{3c\eta^2}\right)^6, 2\right)\\
\eta_k &= \frac{\eta}{(k+k_0)^{1/3}\log(k+k_0)}\\
\alpha_k &= c\eta_k^2.
\end{aligned} \tag{B.19}$$

We should remark that the constants of these parameters and also constants in other bounds in the paper are not optimized, since we focus on the dependence on ε in this work. The constants certainly could be improved at the expense of even great complexity in the analysis.

Lemma 2.3. *Let the assumptions in (A1) hold and η_k be set as (B.19). For the iterates of Algorithm 1, we have for $k \geq 1$ that*

$$\mathbb{E}Y_{k+1} \leq \mathbb{E}Y_k - \frac{\eta_{k+1}}{2}\mathbb{E}\|g_k - \nabla f(x_k)\|^2 + (\beta_{1,k+1}+\beta_{2,k+1})\mathbb{E}\|x_{k+1}-x_k\|^2 + w_k + v_kV^2, \tag{B.20}$$

where

$$\begin{aligned}
Y_{k+1} = L_\rho(x_{k+1},\lambda_{k+1}) &+ \frac{\rho m}{2\eta_{k+2}}\|Ax_{k+1}-b\|^2 + \frac{m}{2\eta_{k+1}}\|x_{k+1}-x_k\|^2_{Q_{k+1}} + \beta_{1,k+1}\|x_{k+1}-x_k\|^2\\
&+ \frac{2}{c\eta_{k+1}}\|g_{k+1}-\nabla f(x_{k+1})\|^2 + \left(\frac{6(1+c_1)}{\rho\delta} + \frac{4m}{L_f\eta_k}\right)\|g_k-\nabla f(x_k)\|^2
\end{aligned} \tag{B.21}$$

and

$$\begin{aligned}
\beta_{1,k} &= \frac{(1+c_2+c_3)L_fm}{\eta_{k+1}} + \frac{(6+c_4)(1+c_1)L_f^2}{\rho\delta} + \frac{42(1+c_1)L_f^2}{\rho\delta} + \frac{28mL_f}{\eta_k},\\
\beta_{2,k} &= \frac{L_\rho}{2} + \frac{(1+2c_3)mL_f}{2\eta_k} + \frac{14L_f^2}{c\eta_k} - \frac{1}{2\eta_k},\\
w_k &= \frac{3(1+c_1)}{\delta\rho}\mathbb{E}\|x_{k+1}-2x_k+x_{k-1}\|^2_{Q_{k+1}^\top Q_{k+1}} - \frac{m}{2\eta_{k+1}}\mathbb{E}\|x_{k+1}-2x_k+x_{k-1}\|^2_{Q_{k+1}},\\
v_k &= \frac{6(1+c_1)\alpha_k^2}{\delta\rho} + \frac{\alpha_k^2m}{L_f\eta_{k+1}} + \frac{6\alpha_{k+1}^2}{c\eta_{k+1}} + \frac{18(1+c_1)\alpha_k^2}{\rho\delta} + \frac{12m\alpha_k^2}{L_f\eta_k}.
\end{aligned}$$

Proof. We combine the bounds of Corollaries B.2 and B.4 after multiplying the latter by the scalar m of (2.1) (by noting cancellation of the terms with $\|\lambda_{k+1}-\lambda_k\|^2$ and after adding and subtracting $\frac{\eta_{k+1}}{2}\mathbb{E}\|g_k-\nabla f(x_k)\|^2$):

$$\begin{aligned}
&\mathbb{E}L_\rho(x_{k+1},\lambda_{k+1}) + \frac{\rho m}{2\eta_{k+2}}\mathbb{E}\|Ax_{k+1}-b\|^2 + \frac{m}{2\eta_{k+1}}\mathbb{E}\|x_{k+1}-x_k\|^2_{Q_{k+1}}\\
&\leq \mathbb{E}L_\rho(x_k,\lambda_k) + \frac{\rho m}{2\eta_{k+1}}\mathbb{E}\|Ax_k-b\|^2 + \frac{m}{2\eta_k}\mathbb{E}\|x_k-x_{k-1}\|^2_{Q_k} - \frac{\eta_{k+1}}{2}\mathbb{E}\|g_k-\nabla f(x_k)\|^2\\
&\quad + \left(\frac{L_\rho}{2} + \frac{(1+2c_3)mL_f}{2\eta_{k+1}} - \frac{1}{2\eta_{k+1}}\right)\mathbb{E}\|x_{k+1}-x_k\|^2
\end{aligned}$$

$$+\left(\frac{(1+c_2+c_3)L_f m}{\eta_{k+1}}+\frac{(6+c_4)(1+c_1)L_f^2}{\rho\delta}\right)\mathbb{E}\|x_k-x_{k-1}\|^2$$
$$+\eta_{k+1}\mathbb{E}\|g_k-\nabla f(x_k)\|^2+\left(\frac{m\alpha_k^2}{L_f\eta_{k+1}}+\frac{6(1+c_1)\alpha_k^2}{\rho\delta}\right)\mathbb{E}\|g_{k-1}-\nabla f(x_{k-1})\|^2 \tag{B.22}$$
$$+\frac{3(1+c_1)}{\rho\delta}\mathbb{E}\|x_{k+1}-2x_k+x_{k-1}\|^2_{Q_{k+1}^\top Q_{k+1}}-\frac{m}{2\eta_{k+1}}\mathbb{E}\|x_{k+1}-2x_k+x_{k-1}\|^2_{Q_{k+1}}$$
$$+\left(\frac{6(1+c_1)\alpha_k^2}{\rho\delta}+\frac{\alpha_k^2 m}{L_f\eta_{k+1}}\right)V^2. \tag{B.23}$$

Since this expression looks rather daunting, we take a closer look at the different terms. On a high level, we see that the terms in the first and second lines telescope. For the terms in the third and fourth lines, and the terms in the sixth line, we show that our parameter choices make them nonpositive or telescoping. The size of the terms in the last line can be controlled by the choice of α_k. The details will be spelled out in Theorem 2.4. What remains is to handle the fifth line by variance reduction recursion from Lemma B.6.

We now use the definitions of $\beta_{1,k}$, $\beta_{2,k}$, and v_k in the statement of the lemma. These terms include not only the coefficients of $\mathbb{E}\|x_k-x_{k-1}\|^2$, $\mathbb{E}\|x_{k+1}-x_k\|^2$, and V^2 but also the contribution for the corresponding terms that will come when we use Lemma B.6 to bound the terms on (B.22), *cf.* (B.31). In particular, with Lemma B.6 to bound the terms on (B.22), and the definitions of $\beta_{1,k},\beta_{2,k+1},v_k,w_k$ in the statement of the lemma, we have

$$\begin{aligned}
&\mathbb{E}L_\rho(x_{k+1},\lambda_{k+1})+\frac{\rho m}{2\eta_{k+2}}\mathbb{E}\|Ax_{k+1}-b\|^2+\frac{m}{2\eta_{k+1}}\mathbb{E}\|x_{k+1}-x_k\|^2_{Q_{k+1}}\\
&\quad+\frac{2}{c\eta_{k+1}}\mathbb{E}\|g_{k+1}-\nabla f(x_{k+1})\|^2+\left(\frac{6(1+c_1)}{\rho\delta}+\frac{4m}{L_f\eta_k}\right)\mathbb{E}\|g_k-\nabla f(x_k)\|^2\\
&\le \mathbb{E}L_\rho(x_k,\lambda_k)+\frac{\rho m}{2\eta_{k+1}}\mathbb{E}\|Ax_k-b\|^2+\frac{m}{2}\mathbb{E}\|x_k-x_{k-1}\|^2_{Q_k}-\frac{\eta_{k+1}}{2}\mathbb{E}\|g_k-\nabla f(x_k)\|^2\\
&\quad+\frac{2}{c\eta_k}\mathbb{E}\|g_k-\nabla f(x_k)\|^2+\left(\frac{6(1+c_1)}{\rho\delta}+\frac{4m}{L_f\eta_{k-1}}\right)\mathbb{E}\|g_{k-1}-\nabla f(x_{k-1})\|^2\\
&\quad+\beta_{2,k+1}\mathbb{E}\|x_{k+1}-x_k\|^2+\beta_{1,k}\mathbb{E}\|x_k-x_{k-1}\|^2+w_k+v_kV^2.
\end{aligned}$$

By adding $\beta_{1,k+1}\mathbb{E}\|x_{k+1}-x_k\|^2$ to both sides and using the definition of Y_k, we have the result. □

We continue with the restatement and proof of Theorem 2.4.

Theorem 2.4. *Let the assumptions in (A1) hold and supppose that η_k and the other algorithmic parameters are chosen as in (B.19). Then, for the iterates of Algorithm 1, we have for any $K\ge 1$ that*

$$\frac{1}{K}\sum_{k=1}^{K+1}\mathbb{E}\left[\|\eta_k^{-1}(x_k-x_{k-1})\|^2+\|g_{k-1}-\nabla f(x_{k-1})\|^2\right]=O\left(\frac{1}{\eta_{K+1}K}\right)=\tilde{O}(K^{-2/3}).$$

Proof. We start with the result of Lemma 2.3. To show that the terms involving w_k, $\beta_{1,k+1}$, and $\beta_{2,k+1}$ in the right-hand side of (B.20) sum up to a nonpositive constant, we will show that

$$Q_{k+1}=\eta_{k+1}^{-1}I-\rho A^\top A\succ 0, \tag{B.24a}$$
$$\frac{m}{2\eta_{k+1}}Q_{k+1}-\frac{3(1+c_1)}{\rho\delta}Q_{k+1}^\top Q_{k+1}\succeq 0, \tag{B.24b}$$
$$\begin{aligned}
\beta_{1,k+1}+\beta_{2,k+1}&=\frac{L_\rho}{2}+\frac{(1+2c_3)mL_f}{2\eta_{k+1}}+\frac{14L_f^2}{c\eta_{k+1}}-\frac{1}{2\eta_{k+1}}\\
&\quad+\frac{(1+c_2+c_3)L_f m}{\eta_{k+2}}+\frac{(6+c_4)(1+c_1)L_f^2}{\rho\delta}+\frac{42(1+c_1)L_f^2}{\rho\delta}+\frac{28mL_f}{\eta_{k+1}}\le-\frac{1}{16\eta_{k+1}},
\end{aligned} \tag{B.24c}$$

where the definitions of $\beta_{1,k},\beta_{2,k}$ are given in Lemma 2.3.

First, we know that (B.24a) is satisfied since $\eta_{k+1} < \eta \leq 1/(\rho\|A\|^2)$. The positive semidefiniteness condition (B.24b) is satisfied when $m \geq \frac{7(1+c_1)}{\rho\delta}$ (which is ensured by the definition of ρ) since $\frac{m}{2} \geq \frac{3(1+c_1)}{\rho\delta}$ and since

$$Q_{k+1} = \frac{1}{\eta_{k+1}} I - \rho A^\top A \succ 0 \implies \frac{1}{\eta_{k+1}} Q_{k+1} - Q_{k+1}^\top Q_{k+1} \succeq 0,$$

as can be verified by direct substitution of Q_{k+1}. We thus have from the definition of w_k in Lemma 2.3 that $w_k \leq 0$.

Finally, we verify (B.24c). With m, ρ, c defined in (2.1), and using $\eta_{k+2}^{-1} \leq 2\eta_{k+1}^{-1}$, see for example Fact B.7, we have that

$$-\frac{1}{2\eta_{k+1}} + \frac{(1+2c_3)mL_f}{2\eta_{k+1}} + \frac{14L_f^2}{c\eta_{k+1}} + \frac{(1+c_2+c_3)L_f m}{\eta_{k+2}} + \frac{28mL_f}{\eta_{k+1}} \leq -\frac{3}{16\eta_{k+1}},$$
$$\frac{L_\rho}{2} + \frac{(6+c_4)(1+c_1)L_f^2}{\rho\delta} + \frac{42(1+c_1)L_f^2}{\rho\delta} \leq \frac{L_\rho}{2} + \frac{L_f}{2} \leq L_\rho,$$

where the final inequality follows from (B.2). Since the left-hand sides of these two inequalitites sum to $\beta_{1,k+1} + \beta_{2,k+1}$, (B.24c) holds if we can show that

$$L_\rho \leq \frac{1}{8\eta_{k+1}} \iff \eta_{k+1} \leq \frac{1}{8L_\rho},$$

which is implied by $\eta \leq \frac{1}{11L_\rho}$ in (2.1).

As shown above, we have $w_k \leq 0$ with the selected parameters. By substituting this bound together with (B.24c) into (B.20), we obtain

$$\mathbb{E}Y_{k+1} \leq \mathbb{E}Y_k - \frac{1}{16\eta_{k+1}}\mathbb{E}\|x_{k+1} - x_k\|^2 - \frac{\eta_{k+1}}{2}\mathbb{E}\|g_k - \nabla f(x_k)\|^2 + v_k V^2, \tag{B.25}$$

By adding to both sides $\frac{1}{32\eta_k}\mathbb{E}\|x_k - x_{k-1}\|^2 + \frac{\eta_k}{4}\mathbb{E}\|g_{k-1} - \nabla f(x_{k-1})\|^2$ and rearranging, we get

$$\begin{aligned}
&\frac{1}{32\eta_{k+1}}\mathbb{E}\|x_{k+1} - x_k\|^2 + \frac{1}{32\eta_k}\mathbb{E}\|x_k - x_{k-1}\|^2 + \frac{\eta_{k+1}}{4}\mathbb{E}\|g_k - \nabla f(x_k)\|^2 + \frac{\eta_k}{4}\mathbb{E}\|g_{k-1} - \nabla f(x_{k-1})\|^2 \\
&\leq \left(\mathbb{E}Y_k + \frac{1}{32\eta_k}\|x_k - x_{k-1}\|^2 + \frac{\eta_k}{4}\|g_{k-1} - \nabla f(x_{k-1})\|^2\right) \\
&\quad - \left(\mathbb{E}Y_{k+1} + \frac{1}{32\eta_{k+1}}\|x_{k+1} - x_k\|^2 + \frac{\eta_{k+1}}{4}\|g_k - \nabla f(x_k)\|^2\right) + v_k V^2,
\end{aligned} \tag{B.26}$$

From the definition $\alpha_k = c\eta_k^2$ with $\eta_k = \frac{\eta}{(k+k_0)^{1/3}\log(k+k_0)}$ in (B.19), we have $\sum_{k=1}^\infty \alpha_{k+1}^2/\eta_{k+1} = O(1)$ and $\sum_{k=1}^\infty \alpha_k^2 = O(1)$. It follows from the definition of v_k in Lemma 2.3 that $\sum_{k=1}^\infty v_k = O(1)$. Thus by summing the inequality (B.26) over $k = 1, 2, \ldots, K$ and telescoping the right-hand side, we obtain

$$\begin{aligned}
&\sum_{k=1}^K \frac{1}{32\eta_{k+1}}\mathbb{E}\|x_{k+1} - x_k\|^2 + \frac{1}{32\eta_k}\mathbb{E}\|x_k - x_{k-1}\|^2 + \frac{\eta_{k+1}}{4}\mathbb{E}\|g_k - \nabla f(x_k)\|^2 + \frac{\eta_k}{4}\mathbb{E}\|g_{k-1} - \nabla f(x_{k-1})\|^2 \\
&\leq \left(\mathbb{E}Y_1 + \frac{1}{32\eta_1}\|x_1 - x_0\|^2 + \frac{\eta_1}{4}\|g_0 - \nabla f(x_0)\|^2\right) \\
&\quad - \left(\mathbb{E}Y_{K+1} + \frac{1}{32\eta_{K+1}}\|x_{K+1} - x_K\|^2 + \frac{\eta_K}{4}\|g_K - \nabla f(x_K)\|^2\right) + O(1) \\
&\leq \left(\mathbb{E}Y_1 + \frac{1}{32\eta_1}\|x_1 - x_0\|^2 + \frac{\eta_1}{4}\|g_0 - \nabla f(x_0)\|^2\right) - \mathbb{E}Y_{K+1} + O(1).
\end{aligned} \tag{B.27}$$

For the terms on the left-hand side, we have for $k = 1, 2, \ldots, K$, using $\eta_k \geq \eta_{k+1} \geq \eta_{K+1}$, that

$$\frac{1}{32\eta_{k+1}}\mathbb{E}\|x_{k+1} - x_k\|^2 + \frac{1}{32\eta_k}\mathbb{E}\|x_k - x_{k-1}\|^2$$

$$= \frac{\eta_{k+1}}{32}\mathbb{E}\|\eta_{k+1}^{-1}(x_{k+1}-x_k)\|^2 + \frac{\eta_k}{32}\mathbb{E}\|\eta_k^{-1}(x_k-x_{k-1})\|^2$$
$$\geq \frac{\eta_{K+1}}{32}\left[\mathbb{E}\|\eta_{k+1}^{-1}(x_{k+1}-x_k)\|^2 + \mathbb{E}\|\eta_k^{-1}(x_k-x_{k-1})\|^2\right],$$

and similarly

$$\frac{\eta_{k+1}}{4}\mathbb{E}\|g_k-\nabla f(x_k)\|^2 + \frac{\eta_k}{4}\mathbb{E}\|g_{k-1}-\nabla f(x_{k-1})\|^2 \geq \frac{\eta_{K+1}}{4}\left[\mathbb{E}\|g_k-\nabla f(x_k)\|^2 + \mathbb{E}\|g_{k-1}-\nabla f(x_{k-1})\|^2\right].$$

Thus, by adding successive terms on the left-hand side of (B.27), and using these bounds, we have

$$\sum_{k=1}^{K}\frac{1}{32\eta_{k+1}}\mathbb{E}\|x_{k+1}-x_k\|^2 + \frac{1}{32\eta_k}\mathbb{E}\|x_k-x_{k-1}\|^2 + \frac{\eta_{k+1}}{4}\mathbb{E}\|g_k-\nabla f(x_k)\|^2 + \frac{\eta_k}{4}\mathbb{E}\|g_{k-1}-\nabla f(x_{k-1})\|^2$$
$$\geq \eta_{K+1}\sum_{k=1}^{K+1}\left(\frac{1}{32}\mathbb{E}\|\eta_k^{-1}(x_k-x_{k-1})\|^2 + \frac{1}{4}\mathbb{E}\|g_{k-1}-\nabla f(x_{k-1})\|^2\right).$$

By substituting this lower bound into (B.27), joining the constant coefficients in the left-hand side to the $O(1)$ term on the right-hand side and using $-Y_{K+1} = O(1)$ (by lower boundedness of the potential function to be shown below in Lemma B.8), we have the result after dividing both sides by $\eta_{K+1}K$. □

Corollary B.5. *Let the assumptions in (A1) hold. With the parameter choices in (B.19), Algorithm 1 outputs $(\bar{x},\bar{\lambda})$ such that*

$$\mathbb{E}\|\nabla f(\bar{x}) + A^\top\bar{\lambda}\| \leq \varepsilon, \text{ and } \mathbb{E}\|A\bar{x}-b\| \leq \varepsilon,$$

within $K = \tilde{O}(\varepsilon^{-3})$ iterations.

Proof. A useful preliminary result is as follows. For $\hat{k}$ selected uniformly at random from $\{1,2,\ldots,K\}$, we have

$$\mathbb{E}\left[\|\eta_{\hat{k}+1}^{-1}(x_{\hat{k}+1}-x_{\hat{k}})\|^2 + \|\eta_{\hat{k}}^{-1}(x_{\hat{k}}-x_{\hat{k}-1})\|^2 + \|g_{\hat{k}}-\nabla f(x_{\hat{k}})\|^2 + \|g_{\hat{k}-1}-\nabla f(x_{\hat{k}-1})\|^2\right]$$
$$= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\eta_{k+1}^{-1}(x_{k+1}-x_k)\|^2 + \|\eta_k^{-1}(x_k-x_{k-1})\|^2 + \|g_k-\nabla f(x_k)\|^2 + \|g_{k-1}-\nabla f(x_{k-1})\|^2\right]$$
$$\leq \frac{2}{K}\sum_{k=1}^{K+1}\mathbb{E}\left[\|\eta_k^{-1}(x_k-x_{k-1})\|^2 + \|g_{k-1}-\nabla f(x_{k-1})\|^2\right]$$
$$= \tilde{O}(K^{-2/3}), \tag{B.28}$$

where the last step used the result of Theorem 2.4. As a result, the right-hand side of (B.28) is guaranteed to be less that ε^2 within $K = \tilde{O}(\varepsilon^{-3})$ iterations.

To show the stationarity condition, we have by the definition of x_{k+1}, λ_{k+1} in Algorithm 1 and the triangle inequality that (see also (B.7))

$$x_{k+1} = x_k - \eta_{k+1}(g_k + A^\top\lambda_{k+1} + \rho A^\top A(x_k - x_{k+1}))$$
$$\iff \quad x_{k+1} + \eta_{k+1}(g_k - \nabla f(x_k)) = x_k - \eta_{k+1}(\nabla f(x_k) + A^\top\lambda_{k+1} + \rho A^\top A(x_k - x_{k+1}))$$
$$\iff \quad \nabla f(x_{k+1}) + A^\top\lambda_{k+1} = \eta_{k+1}^{-1}(x_k - x_{k+1}) + \rho A^\top A(x_{k+1} - x_k)$$
$$- (g_k - \nabla f(x_k)) + \nabla f(x_{k+1}) - \nabla f(x_k)$$
$$\implies \quad \|\nabla f(x_{k+1}) + A^\top\lambda_{k+1}\| \leq \|\eta_{k+1}^{-1}(x_k - x_{k+1})\| + \|g_k - \nabla f(x_k)\| + (\rho\|A\|^2 + L_f)\|x_k - x_{k+1}\|,$$

where, after taking expectation and using $k = \hat{k}$, all the terms on the right-hand side are smaller than ε in expectation after $K = \tilde{O}(\varepsilon^{-3})$ iterations, due to (B.28) and Jensen's inequality. Setting $\bar{x} = x_{\hat{k}+1}$ and $\bar{\lambda} = \lambda_{\hat{k}+1}$, we thus have $\mathbb{E}(\|\nabla f(\bar{x}) + A^\top\bar{\lambda}\|) \leq \varepsilon$.

It remains to show that $\|Ax_{\hat{k}+1} - b\| \leq \varepsilon$ for this same value of $\hat{k}$. For any k, since $\lambda_{k+1} - \lambda_k = \rho(Ax_{k+1} - b)$, $\lambda_{k+1} - \lambda_k$ is in the range of A. Since δ is the smallest nonzero eigenvalue of $A^\top A$, we have (see also (1.3))

$$\|A^\top(\lambda_{k+1} - \lambda_k)\|^2 \geq \delta\|\lambda_{k+1} - \lambda_k\|^2 = \rho^2\delta\|Ax_{k+1} - b\|^2. \tag{B.29}$$

By (B.11), together with $\eta_{k+1}^{-1} I \succ \eta_{k+1}^{-1} I - \rho A^\top A \succ 0$, $0 < \eta_{k+1} \leq \eta_k$, and Young's inequality, we have that

$$\begin{aligned}\|A^\top(\lambda_k - \lambda_{k+1})\|^2 &\leq 3\|g_k - g_{k-1}\|^2 + \frac{6}{\eta_{k+1}^2}\left(\|x_{k+1} - x_k\|^2 + \|x_k - x_{k-1}\|^2\right) + \frac{3}{\eta_{k+1}^2}\|x_k - x_{k-1}\|^2 \\ &\leq 9\|g_k - \nabla f(x_k)\|^2 + 9\|g_{k-1} - \nabla f(x_{k-1})\|^2 + 9\|\nabla f(x_k) - \nabla f(x_{k-1})\|^2 \\ &\quad + \frac{6}{\eta_{k+1}^2}\left(\|x_{k+1} - x_k\|^2 + \|x_k - x_{k-1}\|^2\right) + \frac{3}{\eta_{k+1}^2}\|x_k - x_{k-1}\|^2.\end{aligned}$$

As shown in Fact B.7 we have $\eta_k \leq 2\eta_{k+1}$ and consequently $\frac{1}{\eta_{k+1}} \leq \frac{2}{\eta_k}$. Combining these with Lipschitzness of ∇f and $L_f^2 \leq \eta_k^{-2}$ in the last inequality, we obtain

$$\begin{aligned}\|A^\top(\lambda_{k+1} - \lambda_k)\|^2 = O\Big(&\|g_k - \nabla f(x_k)\|^2 + \|g_{k-1} - \nabla f(x_{k-1})\|^2 \\ &+ \|\eta_{k+1}^{-1}(x_{k+1} - x_k)\|^2 + \|\eta_k^{-1}(x_k - x_{k-1})\|^2\Big).\end{aligned} \tag{B.30}$$

Particularly, by (B.28), the last estimate implies

$$\mathbb{E}\|A^\top(\lambda_{\hat{k}+1} - \lambda_{\hat{k}})\|^2 = \tilde{O}(K^{-2/3}).$$

In view of (B.29), this gives

$$\mathbb{E}\|Ax_{\hat{k}+1} - b\|^2 \leq \frac{1}{\rho^2\delta}\mathbb{E}\|A^\top(\lambda_{\hat{k}+1} - \lambda_{\hat{k}})\| = \tilde{O}(K^{-2/3}).$$

As a result, for $K = \tilde{O}(\varepsilon^{-3})$, by Jensen's inequality we have that $\mathbb{E}\|A\bar{x} - b\| \leq \varepsilon$ for $\bar{x} = x_{\hat{k}+1}$, as required. □

B.4 Auxiliary results used in the analysis

The following lemma is about control of the variance of the estimator, specializing the preliminary Lemma A.1 for the current purpose.

Lemma B.6. *Let the assumptions in (A1) hold and g_k be defined as Algorithm 1. For c such that $\alpha_k = c\eta_k^2 \leq 1$, η_k as in (B.19) and any m, c_1, ρ, δ, we have*

$$\begin{aligned}&\eta_{k+1}\mathbb{E}\|g_k - \nabla f(x_k)\|^2 + \left(\frac{m\alpha_k^2}{L_f\eta_{k+1}} + \frac{6(1+c_1)\alpha_k^2}{\delta\rho}\right)\mathbb{E}\|g_{k-1} - \nabla f(x_{k-1})\|^2 \\ &\leq \frac{2}{c\eta_k}\mathbb{E}\|g_k - \nabla f(x_k)\|^2 + \left(\frac{6(1+c_1)}{\rho\delta} + \frac{4m}{L_f\eta_{k-1}}\right)\|g_{k-1} - \nabla f(x_{k-1})\|^2 \\ &\quad - \frac{2}{c\eta_{k+1}}\mathbb{E}\|g_{k+1} - \nabla f(x_{k+1})\|^2 - \left(\frac{6(1+c_1)}{\rho\delta} + \frac{4m}{L_f\eta_k}\right)\mathbb{E}\|g_k - \nabla f(x_k)\|^2 \\ &\quad + \frac{14L_f^2}{c\eta_{k+1}}\mathbb{E}\|x_{k+1} - x_k\|^2 + \left(\frac{42(1+c_1)L_f^2}{\rho\delta} + \frac{28mL_f}{\eta_k}\right)\mathbb{E}\|x_k - x_{k-1}\|^2 \\ &\quad + \left(\frac{6\alpha_{k+1}^2}{c\eta_{k+1}} + \frac{18(1+c_1)\alpha_k^2}{\rho\delta} + \frac{12m\alpha_k^2}{L_f\eta_k}\right)V^2.\end{aligned} \tag{B.31}$$

Proof. We first recall the result from Lemma A.1. By substituting $(\tilde{G}_k(x,\xi), G_k(x)) = (\tilde{\nabla} f(x,\xi), \nabla f(x))$ for all k, using Lipschitzness of $\tilde{\nabla} f(x,\xi)$ and the variance bound in (A1), and taking total expectation, we have

$$\mathbb{E}\|g_{k+1} - \nabla f(x_{k+1})\|^2 \leq (1-\alpha_{k+1})^2\mathbb{E}\|g_k - \nabla f(x_k)\|^2 + 7L_f^2\mathbb{E}\|x_{k+1} - x_k\|^2 + 3\alpha_{k+1}^2V^2. \tag{B.32}$$

Since $\alpha_{k+1} \leq 1$ by the assumption of the lemma, we have that $(1-\alpha_{k+1})^2 \leq 1 - \alpha_{k+1}$ and therefore

$$\alpha_{k+1}\mathbb{E}\|g_k - \nabla f(x_k)\|^2 \leq \mathbb{E}\|g_k - \nabla f(x_k)\|^2 - \mathbb{E}\|g_{k+1} - \nabla f(x_{k+1})\|^2 + 7L_f^2\mathbb{E}\|x_{k+1} - x_k\|^2 + 3\alpha_{k+1}^2V^2. \tag{B.33}$$

After multiplying the last inequality (replacing k by $k-1$) with $\frac{6(1+c_1)}{\rho\delta}$ and using $\alpha_k^2 \leq \alpha_k$, we obtain

$$\frac{6(1+c_1)\alpha_k^2}{\rho\delta}\mathbb{E}\|g_{k-1} - \nabla f(x_{k-1})\|^2 \leq \frac{6(1+c_1)}{\rho\delta}\left(\mathbb{E}\|g_{k-1} - \nabla f(x_{k-1})\|^2 - \mathbb{E}\|g_k - \nabla f(x_k)\|^2\right)$$

$$+\frac{42(1+c_1)L_f^2}{\rho\delta}\mathbb{E}\|x_k-x_{k-1}\|^2+\frac{18(1+c_1)\alpha_k^2V^2}{\rho\delta}. \tag{B.34}$$

By using $\alpha_k=c\eta_k^2$, we have from (B.33) that

$$\begin{aligned}\eta_{k+1}\mathbb{E}\|g_k-\nabla f(x_k)\|^2\le\frac{1}{c\eta_{k+1}}\Big(&\mathbb{E}\|g_k-\nabla f(x_k)\|^2-\mathbb{E}\|g_{k+1}-\nabla f(x_{k+1})\|^2\\&+7L_f^2\mathbb{E}\|x_{k+1}-x_k\|^2+3\alpha_{k+1}^2V^2\Big).\end{aligned} \tag{B.35}$$

Formula (B.37e) from Fact B.7 sat that $\frac{1}{\eta_{k+1}}-\frac{1}{\eta_k}\le\frac{\eta_{k+1}c}{2}$, so we have

$$\begin{aligned}\frac{1}{c\eta_{k+1}}\|g_k-\nabla f(x_k\|^2&=\frac{1}{c\eta_k}\|g_k-\nabla f(x_k\|^2+\left(\frac{1}{c\eta_{k+1}}-\frac{1}{c\eta_k}\right)\|g_k-\nabla f(x_k\|^2\\&\le\frac{1}{c\eta_k}\|g_k-\nabla f(x_k\|^2+\frac{\eta_{k+1}}{2}\|g_k-\nabla f(x_k\|^2,\end{aligned}$$

which by substituting in (B.35) yields

$$\begin{aligned}\eta_{k+1}\mathbb{E}\|g_k-\nabla f(x_k)\|^2\le&\frac{1}{c\eta_k}\mathbb{E}\|g_k-\nabla f(x_k)\|^2+\frac{\eta_{k+1}}{2}\|g_k-\nabla f(x_k)\|^2-\frac{1}{c\eta_{k+1}}\mathbb{E}\|g_{k+1}-\nabla f(x_{k+1})\|^2\\&+\frac{7L_f^2}{c\eta_{k+1}}L_f^2\mathbb{E}\|x_{k+1}-x_k\|^2+\frac{3\alpha_{k+1}^2V^2}{c\eta_{k+1}}.\end{aligned}$$

By moving the second term on the right-hand side to the left, then multiplying both sides by 2, we obtain

$$\begin{aligned}&\eta_{k+1}\mathbb{E}\|g_k-\nabla f(x_k)\|^2\\&\le\frac{2}{c}\left(\frac{1}{\eta_k}\mathbb{E}\|g_k-\nabla f(x_k)\|^2-\frac{1}{\eta_{k+1}}\mathbb{E}\|g_{k+1}-\nabla f(x_{k+1})\|^2\right)+\frac{14L_f^2}{c\eta_{k+1}}\mathbb{E}\|x_{k+1}-x_k\|^2+\frac{6\alpha_{k+1}^2V^2}{c\eta_{k+1}}.\end{aligned} \tag{B.36}$$

By replacing k by $k-1$ and multiplying both sides by $\frac{2mc}{L_f}$, we get

$$\begin{aligned}&\frac{2mc\eta_k}{L_f}\mathbb{E}\|g_{k-1}-\nabla f(x_{k-1})\|^2\\&\le\frac{4m}{L_f}\left(\frac{1}{\eta_{k-1}}\mathbb{E}\|g_{k-1}-\nabla f(x_{k-1})\|^2-\frac{1}{\eta_k}\mathbb{E}\|g_k-\nabla f(x_k)\|^2\right)+\frac{28mL_f}{\eta_k}\mathbb{E}\|x_k-x_{k-1}\|^2+\frac{12m\alpha_k^2V^2}{L_f\eta_k}.\end{aligned}$$

By using $\alpha_k\le1$ (thus $\alpha_k^2\le\alpha_k$), $\alpha_k=c\eta_k^2$, and $\eta_k\le2\eta_{k+1}$ from (B.37f) in Fact B.7, we have

$$\frac{m\alpha_k^2}{L_f\eta_{k+1}}\le\frac{m\alpha_k}{L_f\eta_{k+1}}=\frac{mc\eta_k^2}{L_f\eta_{k+1}}\le\frac{2mc\eta_k}{L_f},$$

and so by replacing the left-hand side in the previous inequality, we obtain

$$\begin{aligned}&\frac{m\alpha_k^2}{L_f\eta_{k+1}}\mathbb{E}\|g_{k-1}-\nabla f(x_{k-1})\|^2\\&\le\frac{4m}{L_f}\left(\frac{1}{\eta_{k-1}}\mathbb{E}\|g_{k-1}-\nabla f(x_{k-1})\|^2-\frac{1}{\eta_k}\mathbb{E}\|g_k-\nabla f(x_k)\|^2\right)+\frac{28mL_f}{\eta_k}\mathbb{E}\|x_k-x_{k-1}\|^2+\frac{12m\alpha_k^2V^2}{L_f\eta_k}.\end{aligned}$$

The result follows by combining this bound with (B.34) and (B.36). □

We use the following fact to simplify the coefficients appearing in the analysis (before and after this point). We do not try to optimize the constants in this result or other parts of the paper since our focus is on the ε-dependence in our bounds. (The constants in this lemma are certainly improvable, since, for simplicity, we make use of loose inequalities to compare terms of the order $\log(x+1)$ and x^α for $\alpha>1/3$.)

Fact B.7. *Given* $\eta_k = \frac{\eta}{(k+k_0)^{1/3}\log(k+k_0)}$, *it holds that*

$$\frac{1}{2\rho}\left(\frac{1}{\eta_{k+2}} - \frac{1}{\eta_{k+1}}\right) \leq \frac{c_1}{\rho m}, \tag{B.37a}$$

$$\frac{\eta_{k+1}^{-2} - \eta_k^{-2}}{2} \leq \frac{c_2 L_f}{\eta_{k+1}}, \tag{B.37b}$$

$$\frac{\eta_{k+1}^{-1} - \eta_k^{-1}}{2\eta_{k+1}} \leq \frac{c_3 L_f}{\eta_{k+1}}, \tag{B.37c}$$

$$3(\eta_k^{-1} - \eta_{k+1}^{-1})^2 \leq c_4 L_f^2 \tag{B.37d}$$

$$\frac{1}{\eta_{k+1}} - \frac{1}{\eta_k} \leq \frac{\eta_{k+1} c}{2} \tag{B.37e}$$

$$\eta_{k+1} \leq 2\eta_{k+2}, \tag{B.37f}$$

where $k_0 = \max\left(\left(\frac{10m}{3c_1\eta}\right)^2, \left(\frac{20}{3\eta c_2 L_f}\right)^2, \left(\frac{10}{3c_3 L_f}\right)^2, \frac{400}{3\eta^2 c_4 L_f^2}, \left(\frac{20}{c\eta^2}\right)^4, \left(\frac{50}{3c\eta^2}\right)^6, 2\right)$, *for any absolute constants* c_1, c_2, c_3, c_4 *and any positive values of* m, η, c.

Proof. For (B.37a), by straightforward computation, we have

$$\begin{aligned}\frac{1}{\eta_{k+2}} - \frac{1}{\eta_{k+1}} &= \frac{(k+k_0+2)^{1/3}\log(k+k_0+2) - (k+k_0+1)^{1/3}\log(k+k_0+1)}{\eta} \\ &= \frac{\log(k+k_0+2)((k+k_0+2)^{1/3} - (k+k_0+1)^{1/3})}{\eta} \\ &\quad + \frac{(k+k_0+1)^{1/3}(\log(k+k_0+2) - \log(k+k_0+1)}{\eta}.\end{aligned} \tag{B.38}$$

Here, we have

$$\log(k+k_0+2) - \log(k+k_0+1) = \log\left(1 + \frac{1}{k+k_0+1}\right) \leq \frac{1}{k+k_0+1}$$

and

$$(k+k_0+2)^{1/3} - (k+k_0+1)^{1/3} = \frac{1}{(k+k_0+2)^{2/3} + (k+k_0+2)(k+k_0+1) + (k+k_0+1)^{2/3}} \leq \frac{1}{3(k+k_0+1)^{2/3}}.$$

With these, (B.38) yields

$$\frac{1}{\eta_{k+2}} - \frac{1}{\eta_{k+1}} \leq \frac{\log(k+k_0+2)}{3(k+k_0+1)^{2/3}\eta} + \frac{1}{\eta(k+k_0+1)^{2/3}} \leq \frac{4\log(k+k_0+2)}{3(k+k_0+1)^{2/3}\eta}$$

and hence the desired inequality is implied by $\frac{4\log(k+k_0+2)}{3(k+k_0+1)^{2/3}\eta} \leq \frac{2c_1}{m}$. By using $\log(k+k_0+2) \leq 5(k+k_0+1)^{1/6}$ for $k \geq 1$, the inequality is implied by $\frac{10m}{3c_1\eta} \leq (k+k_0+1)^{1/2}$ which is implied by $k_0 \geq \left(\frac{10m}{3c_1\eta}\right)^2$.

For (B.37b), using the first bound in the previous paragraph and $\frac{1}{\eta_k} \leq \frac{1}{\eta_{k+1}}$, we have

$$\frac{1}{\eta_{k+1}^2} - \frac{1}{\eta_k^2} \leq \frac{2}{\eta_{k+1}}\left(\frac{1}{\eta_{k+1}} - \frac{1}{\eta_k}\right) \leq \frac{4\log(k+k_0+1)}{3\eta(k+k_0)^{2/3}}\frac{2}{\eta_{k+1}},$$

and the inequality we want to prove is implied by $\frac{4\log(k+k_0+1)}{3\eta(k+k_0)^{2/3}} \leq c_2 L_f$. By using $\log(k+k_0+1) \leq 5(k+k_0)^{1/6}$, the assertion is implied by $\frac{20}{3\eta c_2 L_f} \leq (k+k_0)^{1/2}$ which is implied by $k_0 \geq \left(\frac{20}{3\eta c_2 L_f}\right)^2$.

It is straightforward to show (B.37c), (B.37d) by using the same estimates, which we omit for brevity.

For (B.37e), as in the beginning of the proof, we have

$$\frac{1}{\eta_{k+1}} - \frac{1}{\eta_k} \leq \frac{1}{\eta(k+k_0)^{2/3}}\left(1 + \frac{\log(k+k_0+1)}{3}\right),$$

and the desired inequality is implied by

$$\frac{1}{\eta(k+k_0)^{2/3}}\left(1+\frac{\log(k+k_0+1)}{3}\right)\leq\frac{c\eta}{2(k+k_0+1)^{1/3}\log(k+k_0+1)},$$

which in turn is implied by

$$2\log(k+k_0+1)+\frac{2(\log(k+k_0+1))^2}{3}\leq c\eta^2(k+k_0)^{1/3}.$$

By using $\log(k+k_0+1)\leq 5(k+k_0)^{1/12}$, this inequality is implied by

$$10(k+k_0)^{1/12}+\frac{25(k+k_0)^{1/6}}{3}\leq c\eta^2(k+k_0)^{1/3},$$

which in turn is implied by

$$10(k+k_0)^{1/12}\leq\frac{c\eta^2(k+k_0)^{1/3}}{2}\quad\text{and}\quad\frac{25(k+k_0)^{1/6}}{3}\leq\frac{c\eta^2(k+k_0)^{1/3}}{2}.$$

These bounds hold when $k_0\geq\max\left(\left(\frac{20}{c\eta^2}\right)^4,\left(\frac{50}{3c\eta^2}\right)^6\right)=\left(\frac{50}{3c\eta^2}\right)^6$.

The last assertion $\eta_{k+1}\leq 2\eta_{k+2}$ (B.37f) is implied by $(k+k_0+2)^{1/3}\log(k+k_0+2)\leq 2(k+k_0+1)^{1/3}\log(k+k_0+1)$ which is implied by $(k+k_0+2)^{1/3}\leq\frac{4}{3}(k+k_0+1)^{1/3}$ which holds for any k_0 and by $\log(k+k_0+2)\leq\frac{3}{2}\log(k+k_0+1)$ which holds, for example, when $k_0\geq 2$. □

The following lemma is showing the lower boundedness of the potential function Y_k defined in Lemma 2.3, which is important for ensuring that the complexity has the desired dependence on ε.

Lemma B.8. *Under the assumptions and the parameter choices of Theorem 2.4 and the definition of Y_k in (B.21), there exists $\underline{y}>-\infty$ such that $\mathbb{E}Y_k\geq\underline{y}$ for any k. In particular $\underline{y}=-2V^2\sum_{k=1}^{\infty}v_k>-\infty$ since*

$$v_k=\frac{6(1+c_1)\alpha_k^2}{\delta\rho}+\frac{\alpha_k^2 m}{L_f\eta_{k+1}}+\frac{6\alpha_{k+1}^2}{c\eta_{k+1}}+\frac{18(1+c_1)\alpha_k^2}{\rho\delta}+\frac{12m\alpha_k^2}{L_f\eta_k}=O\left(\frac{1}{(k+k_0)(\log(k+k_0))^3}\right),$$

due to

$$\eta_k=\frac{\eta}{(k+k_0)^{1/3}\log(k+k_0)},\quad\alpha_k=c\eta_k^2,$$

for the constants η and c from (B.19).

Proof. Note that all the terms are nonnegative in the definition of Y_k in Lemma 2.3) with the exception of $L_\rho(x_k,\lambda_k)$. We therefore focus on the latter term. Our argument extends that of (Hong, 2016, Lemma 3.5), the important difference being that, due to the stochasticity of our setting, we do not have non-increasing potential, which is critical in the argument. We show that our time-varying choices of step sizes still allow us to establish lower boundedness.

First, by using $f(x)\geq\underline{f}\geq 0$ and the definition of λ_{k+1}, which implies

$$\langle\lambda_{k+1},Ax_{k+1}-b\rangle=\rho^{-1}\langle\lambda_{k+1},\lambda_{k+1}-\lambda_k\rangle=\frac{1}{2\rho}\left(\|\lambda_{k+1}\|^2-\|\lambda_k\|^2+\|\lambda_{k+1}-\lambda_k\|^2\right),$$

we can show that the following holds for any K:

$$\begin{aligned}\sum_{k=0}^{K}\mathbb{E}L_\rho(x_{k+1},\lambda_{k+1})&=\sum_{k=0}^{K}\mathbb{E}\left[f(x_{k+1})+\langle\lambda_{k+1},Ax_{k+1}-b\rangle+\frac{\rho}{2}\|Ax_{k+1}-b\|^2\right]\\&\geq\frac{1}{\rho}\left(\mathbb{E}\|\lambda_{K+1}\|^2-\|\lambda_0\|^2\right)\geq-\frac{1}{\rho}\|\lambda_0\|^2.\end{aligned}$$

It follows that

$$\sum_{k=1}^{\infty} \mathbb{E} L_\rho(x_k, \lambda_k) > -\infty \quad \Longrightarrow \quad \sum_{k=1}^{\infty} \mathbb{E} Y_k \geq \underline{y}_1 > -\infty, \tag{B.39}$$

for some $\underline{y}_1$.

Next, we use the bound (B.25) which states that

$$\mathbb{E} Y_{k+1} \leq \mathbb{E} Y_k + v_k V^2, \tag{B.40}$$

where v_k from Lemma 2.3 is redefined in the statement of this lemma. For the $O()$ estimate $v_k = O\left(\frac{1}{(k+k_0)(\log(k+k_0))^3}\right)$, we have used $\eta_k = \frac{\eta}{(k+k_0)^{1/3}\log(k+k_0)}$, $\alpha_k = c\eta_k^2$ for constants η, c given in (B.19), together with with $k_0 \geq 2$. We define $C := V^2 \sum_{k=1}^{\infty} v_k$ which is finite due to the definition of v_k with $k_0 \geq 2$.

We consider three cases.

1. When $\mathbb{E} Y_k \geq 0$ for all k, the assertion follows immediately.
2. When $\mathbb{E} Y_{k_2} < 0$ for some k_2 and $\mathbb{E} Y_k \geq -2C$ for all $k \geq k_2$, the assertion also follows.
3. When there exists an index k_1 such that $\mathbb{E} Y_{k_1} < -2C$, assume without loss of generality that k_1 is the smallest index that satisfies this property. By the definition of this case, $\mathbb{E} Y_k \geq -2C$ for $k < k_1$. Since $V^2 \sum_{k=1}^{\infty} v_k = C$, we have from (B.40) that $\mathbb{E} Y_k < -C$ for all $k \geq k_1$. However, this would cause a contradiction with (B.39).

This concludes the proof. □

C STOCHASTIC CONSTRAINTS: PROBLEM (III)

C.1 Variance control

We start with a result for the variance of the estimator g_k. This result is essentially a corollary of Lemma A.1. We characterize the precise constants for the bound of this lemma which are important for getting the order of complexity. For ease of reference, let us recall here the relevant quantities from (A2), (A3), (A4):

$$\begin{aligned}
\mathbb{E}\|\tilde{\nabla} f(x,\xi) - \tilde{\nabla} f(y,\xi)\|^2 &\leq \tilde{L}_{\nabla f}^2 \|x-y\|^2, \\
\mathbb{E}\|\tilde{\nabla} c_i(x,\zeta) - \tilde{\nabla} c_i(y,\zeta)\|^2 &\leq \tilde{L}_{\nabla c}^2 \|x-y\|^2, \\
\mathbb{E}|\tilde{c}_i(x,\zeta) - \tilde{c}_i(y,\zeta)\|^2 &\leq \tilde{L}_c^2 \|x-y\|^2, \\
\mathbb{E}\|\tilde{\nabla} f(x,\xi) - \nabla f(x)\|^2 &\leq \sigma_f^2, \\
\mathbb{E}\|\tilde{\nabla} c_i(x,\zeta) - \nabla c_i(x)\|^2 &\leq \sigma_{\nabla c}^2, \\
\mathbb{E}\|\tilde{c}_i(x,\zeta) - c_i(x)\|^2 &\leq \sigma_c^2, \\
\|\nabla c_i(x)\| &\leq C_{\nabla c} \\
|c_i(x)| &\leq C_c, \\
\|\tilde{\nabla} c_i(x,\zeta)\| &\leq \tilde{C}_{\nabla c}, \\
|\tilde{c}_i(x,\zeta)| &\leq \tilde{C}_c,
\end{aligned}$$

for any $i \in \{1, \ldots, m\}$.

Lemma C.1. *Let the assumptions in (A2), (A3), (A4) hold and let the parameters of Algorithm 2 be given as*

$$\eta_k = \frac{1}{9\tilde{L}\rho(k+1)^{3/5}}, \quad \rho_k = \rho k^{1/5}, \quad \text{and } \alpha_{k+1} = 72\tilde{L}^2\rho_{k+1}^2\eta_k^2 = \frac{72}{81}(k+1)^{-4/5},$$

for some constant $\rho > 1$, $\tilde{L}^2 = 4\tilde{L}_{\nabla f}^2 + 4m^2(\tilde{C}_c^2\tilde{L}_{\nabla c}^2 + \tilde{C}_{\nabla c}^2\tilde{L}_c^2)$, we have

$$\eta_k \mathbb{E}\|g_k - \nabla Q_{\rho_k}(x_k)\|^2 \leq \frac{1}{72\tilde{L}^2\rho_k^2\eta_{k-1}}\mathbb{E}\|g_k - \nabla Q_{\rho_k}(x_k)\|^2 - \frac{1}{72\tilde{L}^2\rho_{k+1}^2\eta_k}\mathbb{E}\|g_{k+1} - \nabla Q_{\rho_{k+1}}(x_{k+1})\|^2$$

$$+\frac{7}{18\eta_k}\mathbb{E}\|x_{k+1}-x_k\|^2+\frac{7m^2\tilde{C}_{\nabla c}^2\tilde{C}_c^2}{12\tilde{L}^2\rho_{k+1}^2\eta_k}|\rho_{k+1}-\rho_k|^2$$
$$+\frac{\alpha_{k+1}^2}{12\tilde{L}^2\rho_{k+1}^2\eta_k}\left(\sigma_f^2+2m^2\rho_k^2\left(C_c^2\sigma_{\nabla c}^2+\tilde{C}_{\nabla c}^2\sigma_c^2\right)\right).$$

Proof. We apply Lemma A.1 with

$$G_k(x_k)=\nabla Q_{\rho_k}(x_k)=\nabla f(x_k)+\rho_k\sum_{i=1}^m\nabla c_i(x_k)c_i(x_k), \tag{C.1a}$$
$$G_{k+1}(x_{k+1})=\nabla Q_{\rho_{k+1}}(x_{k+1})=\nabla f(x_{k+1})+\rho_{k+1}\sum_{i=1}^m\nabla c_i(x_{k+1})c_i(x_{k+1}), \tag{C.1b}$$
$$\tilde{G}_k(x_k,B_{k+1})=\tilde{\nabla}Q_{\rho_k}(x_k,B_{k+1})=\tilde{\nabla}f(x_k,\xi_{k+1}^0)+\rho_k\sum_{i=1}^m\tilde{\nabla}c_i(x_k,\zeta_{k+1}^1)\tilde{c}_i(x_k,\zeta_{k+1}^2), \tag{C.1c}$$
$$\tilde{G}_{k+1}(x_{k+1},B_{k+1})=\tilde{\nabla}Q_{\rho_{k+1}}(x_{k+1},B_{k+1})=\tilde{\nabla}f(x_{k+1},\xi_{k+1}^0)+\rho_{k+1}\sum_{i=1}^m\tilde{\nabla}c_i(x_{k+1},\zeta_{k+1}^1)\tilde{c}_i(x_{k+1},\zeta_{k+1}^2). \tag{C.1d}$$

We note that $G_k(x_k)=\mathbb{E}_{B_{k+1}}\tilde{G}_k(x_k,B_{k+1})$ and $G_{k+1}(x_{k+1})=\mathbb{E}_{B_{k+1}}\tilde{G}_{k+1}(x_{k+1},B_{k+1})$, as required by Lemma A.1. This estimation is due to B_{k+1} being sampled after the computation of x_{k+1}, by the independence of ζ^1 and ζ^2 and by ρ_k being a deterministic sequence. We now estimate the error terms on the right-hand side of Lemma A.1. Recall that $\mathbb{E}_k$ the expectation conditioning on all the history up to and including x_{k+1}.

For lighter notation, we drop the subscripts from the random variables (ξ^0,ζ^1,ζ^2) that define $B_{k+1}=(\xi_{k+1}^0,\zeta_{k+1}^1,\zeta_{k+1}^2)$ in this lemma.

First, we estimate the second term on the right-hand side of the inequality in Lemma A.1. By Young's inequality, we have

$$\begin{aligned}&\mathbb{E}_k\|\tilde{G}_{k+1}(x_{k+1},B_{k+1})-\tilde{G}_{k+1}(x_k,B_{k+1})\|^2\\&\le 2\mathbb{E}_k\|\tilde{\nabla}f(x_{k+1},\xi^0)-\tilde{\nabla}f(x_k,\xi^0)\|^2\\&\quad+2\rho_{k+1}^2\mathbb{E}_k\left\|\sum_{i=1}^m\left(\tilde{\nabla}c_i(x_{k+1},\zeta^1)\tilde{c}_i(x_{k+1},\zeta^2)-\tilde{\nabla}c_i(x_k,\zeta^1)\tilde{c}_i(x_k,\zeta^2)\right)\right\|^2.\end{aligned} \tag{C.2}$$

By using (A2) and (A4), we have that

$$\begin{aligned}&\mathbb{E}_k\left\|\sum_{i=1}^m\left(\tilde{\nabla}c_i(x_{k+1},\zeta^1)\tilde{c}_i(x_{k+1},\zeta^2)-\tilde{\nabla}c_i(x_k,\zeta^1)\tilde{c}_i(x_k,\zeta^2)\right)\right\|^2\\&\le m\sum_{i=1}^m\mathbb{E}_k\left\|\tilde{\nabla}c_i(x_{k+1},\zeta^1)\tilde{c}_i(x_{k+1},\zeta^2)-\tilde{\nabla}c_i(x_k,\zeta^1)\tilde{c}_i(x_k,\zeta^2)\right\|^2\\&\le 2m\sum_{i=1}^m\mathbb{E}_k\left\|\tilde{\nabla}c_i(x_{k+1},\zeta^1)\left(\tilde{c}_i(x_{k+1},\zeta^2)-\tilde{c}_i(x_k,\zeta^2)\right)\right\|^2\\&\quad+2m\sum_{i=1}^m\mathbb{E}_k\left\|\left(\tilde{\nabla}c_i(x_{k+1},\zeta^1)-\tilde{\nabla}c_i(x_k,\zeta^1)\right)\tilde{c}_i(x_k,\zeta^2)\right\|^2\\&\le 2m^2\left(\tilde{C}_c^2\tilde{L}_{\nabla c}^2+\tilde{C}_{\nabla c}^2\tilde{L}_c^2\right)\|x_{k+1}-x_k\|^2,\end{aligned} \tag{C.3}$$

where the first inequality follows from the fact that for vectors Y_i, we have $\|\sum_{i=1}^m Y_i\|^2\le(\sum_{i=1}^m\|Y_i\|)^2\le m\sum_{i=1}^m\|Y_i\|^2$ and the second by Young's inequality. Using this in (C.2) along with the first line in (A2), we have

$$\mathbb{E}_k\|\tilde{G}_{k+1}(x_{k+1},B_{k+1})-\tilde{G}_{k+1}(x_k,B_{k+1})\|^2=\left(2\tilde{L}_{\nabla f}^2+4\rho_{k+1}^2m^2\left(\tilde{C}_c^2\tilde{L}_{\nabla c}^2+\tilde{C}_{\nabla c}^2\tilde{L}_c^2\right)\right)\|x_{k+1}-x_k\|^2$$

$$\leq 4\tilde{L}^2\rho_{k+1}^2\|x_{k+1}-x_k\|^2 \tag{C.4}$$

where the last line is by the definition of $\tilde{L}$, see also (3.2).

Second, we estimate the third term on the right-hand side in Lemma A.1:

$$\begin{aligned}\mathbb{E}_k\|\tilde{G}_{k+1}(x_k,B_{k+1})-\tilde{G}_k(x_k,B_{k+1})\|^2 &= \mathbb{E}_k\left\|(\rho_{k+1}-\rho_k)\sum_{i=1}^m\tilde{\nabla}c_i(x_k,\zeta^1)\tilde{c}_i(x_k,\zeta^2)\right\|^2\\ &\leq m^2\tilde{C}_{\nabla c}^2\tilde{C}_c^2|\rho_{k+1}-\rho_k|^2.\end{aligned} \tag{C.5}$$

Third, we estimate the fourth term on the right-hand side in Lemma A.1. From Young's inequality, we have

$$\begin{aligned}&\mathbb{E}\|G_k(x_k)-\tilde{G}_k(x_k,B_{k+1})\|^2\\ &\leq 2\mathbb{E}\left\|\nabla f(x_k)-\tilde{\nabla}f(x_k,\xi^0)\right\|^2+2\rho_k^2\mathbb{E}\left\|\sum_{i=1}^m\nabla c_i(x_k)c_i(x_k)-\sum_{i=1}^m\tilde{\nabla}c_i(x_k,\zeta^1)\tilde{c}_i(x_k,\zeta^2)\right\|^2\\ &\leq 2\sigma_f^2+4m^2\rho_k^2\left(C_c^2\sigma_{\nabla c}^2+\tilde{C}_{\nabla c}^2\sigma_c^2\right),\end{aligned} \tag{C.6}$$

where the estimations for the last inequality are similar to (C.3).

By substituting the bounds (C.4), (C.5), (C.6) into (A.1) and also using (C.1), we obtain

$$\begin{aligned}\mathbb{E}_k\|g_{k+1}-\nabla Q_{\rho_{k+1}}(x_{k+1})\|^2 \leq{}& (1-\alpha_{k+1})^2\|g_k-\nabla Q_{\rho_k}(x_k)\|^2\\ &+28\tilde{L}^2\rho_{k+1}^2\|x_{k+1}-x_k\|^2+42m^2\tilde{C}_{\nabla c}^2\tilde{C}_c^2|\rho_{k+1}-\rho_k|^2\\ &+6\alpha_{k+1}^2\left(\sigma_f^2+2m^2\rho_k^2\left(C_c^2\sigma_{\nabla c}^2+\tilde{C}_{\nabla c}^2\sigma_c^2\right)\right).\end{aligned} \tag{C.7}$$

In (C.7), dividing all terms by $72\tilde{L}^2\rho_{k+1}^2\eta_k$ gives

$$\begin{aligned}\frac{1}{72\tilde{L}^2\rho_{k+1}^2\eta_k}\mathbb{E}_k\|g_{k+1}-\nabla Q_{\rho_{k+1}}(x_{k+1})\|^2 \leq{}& \frac{(1-\alpha_{k+1})^2}{72\tilde{L}^2\rho_{k+1}^2\eta_k}\|g_k-\nabla Q_{\rho_k}(x_k)\|^2\\ &+\frac{7}{18\eta_k}\|x_{k+1}-x_k\|^2+\frac{7m^2\tilde{C}_{\nabla c}^2\tilde{C}_c^2}{12\tilde{L}^2\rho_{k+1}^2\eta_k}|\rho_{k+1}-\rho_k|^2\\ &+\frac{\alpha_{k+1}^2}{12\tilde{L}^2\rho_{k+1}^2\eta_k}\left(\sigma_f^2+2m^2\rho_k^2\left(C_c^2\sigma_{\nabla c}^2+\tilde{C}_{\nabla c}^2\sigma_c^2\right)\right).\end{aligned} \tag{C.8}$$

We focus on the first term on the right-hand side and will show next

$$\frac{(1-\alpha_{k+1})^2}{72\tilde{L}^2\rho_{k+1}^2\eta_k}\|g_k-\nabla Q_{\rho_k}(x_k)\|^2 \leq \left(\frac{1}{72\tilde{L}^2\rho_k^2\eta_{k-1}}-\eta_k\right)\|g_k-\nabla Q_{\rho_k}(x_k)\|^2. \tag{C.9}$$

By the definitions of $\alpha_{k+1},\eta_k,\rho_{k+1}$, we have that $\frac{-\alpha_{k+1}}{72\tilde{L}^2\rho_{k+1}^2\eta_k}=-\eta_k$ therefore (C.9) follows after showing that

$$\frac{1-\alpha_{k+1}+\alpha_{k+1}^2}{72\tilde{L}^2\rho_{k+1}^2\eta_k}\leq\frac{1}{72\tilde{L}^2\rho_k^2\eta_{k-1}} \iff \frac{1}{\rho_{k+1}^2\eta_k}-\frac{1}{\rho_k^2\eta_{k-1}}\leq\frac{\alpha_{k+1}(1-\alpha_{k+1})}{\rho_{k+1}^2\eta_k}. \tag{C.10}$$

Note that by definitions of η_k and ρ_k, we have that

$$\rho_{k+1}^2\eta_k=\rho^2(k+1)^{2/5}\frac{1}{9\tilde{L}\rho(k+1)^{3/5}}=\frac{\rho}{9\tilde{L}(k+1)^{1/5}}.$$

By substituting the values of $\rho_{k+1}^2\eta_k$ and $\alpha_{k+1}=72\tilde{L}^2\rho_{k+1}^2\eta_k^2=\frac{72}{81(k+1)^{4/5}}$, we find that (C.10) is equivalent to

$$\frac{9\tilde{L}}{\rho}\left((k+1)^{1/5}-k^{1/5}\right)\leq\frac{9\tilde{L}\alpha_{k+1}(1-\alpha_{k+1})(k+1)^{1/5}}{\rho} \iff (k+1)^{1/5}-k^{1/5}\leq\alpha_{k+1}(1-\alpha_{k+1})(k+1)^{1/5}.$$

First, we note that $(k+1)^{1/5}-k^{1/5} \leq \frac{1}{5k^{4/5}}$ and also $(1-\alpha_{k+1}) = 1-\frac{72}{81(k+1)^{4/5}} \geq \frac{4}{10}$ for $k \geq 1$. Therefore, (C.10) will be implied by

$$\frac{1}{5k^{4/5}} \leq \frac{288}{810(k+1)^{3/5}},$$

which holds for $k \geq 1$. Thus, (C.10) and consequently (C.9) hold for $k \geq 1$. Using (C.9) to bound the first term on the right-hand side of (C.8) and taking total expectation gives the result. □

C.2 One iteration inequality

Lemma 3.5. *Let the assumptions in (A2), (A3), (A4), (A5) hold and let the parameters of Algorithm 2 be given as*

$$\eta_k = \frac{1}{9\tilde{L}\rho(k+1)^{3/5}}, \quad \rho_k = \rho k^{1/5}, \quad \text{and } \alpha_{k+1} = 72\tilde{L}^2\rho_{k+1}^2\eta_k^2 = \frac{72}{81}(k+1)^{-4/5},$$

for some constant $\rho > 1$ and $\tilde{L}^2 = 4\tilde{L}_{\nabla f}^2 + 4m^2(\tilde{C}_c^2\tilde{L}_{\nabla c}^2 + \tilde{C}_{\nabla c}^2\tilde{L}_c^2)$. Then, we have that

$$\frac{\eta_k}{72}\mathbb{E}\left[\mathsf{d}^2(\nabla f(x_{k+1}) + \rho_k \nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1}))\right] \leq \mathbb{E}[Y_k - Y_{k+1} + |Q_{\rho_k}(x_{k+1}) - Q_{\rho_{k+1}}(x_{k+1})|] + \mathcal{E}_{k+1},$$

where

$$Y_{k+1} = Q_{\rho_{k+1}}(x_{k+1}) + \frac{1}{72\tilde{L}^2\rho_{k+1}^2\eta_k}\|g_{k+1} - \nabla Q_{\rho_{k+1}}(x_{k+1})\|^2$$

and

$$\mathcal{E}_{k+1} = \frac{7m^2\tilde{C}_{\nabla c}^2\tilde{C}_c^2}{12\tilde{L}^2\rho_{k+1}^2\eta_k}|\rho_{k+1} - \rho_k|^2 + \frac{\alpha_{k+1}^2}{12\tilde{L}^2\rho_{k+1}^2\eta_k}\left(\sigma_f^2 + 2m^2\rho_k^2\left(C_c^2\sigma_{\nabla c}^2 + \tilde{C}_{\nabla c}^2\sigma_c^2\right)\right).$$

Remark C.2. By the definitions of η_k, ρ_k, we have that the first term of $\mathcal{E}_{k+1}$ is $O(k^{-7/5})$, the second term of $\mathcal{E}_{k+1}$ is $O(k^{-1})$, therefore $\sum_{k=1}^K \mathcal{E}_{k+1} = O(\log(K+1))$.

Proof. By descent lemma applied on $x \mapsto Q_{\rho_k}(x)$, we have (by denoting the Lipschitz constant of $\nabla Q_{\rho_k}(x)$ as $L_{\rho_k} = \rho_k(L_{\nabla f} + m(C_c L_{\nabla c} + C_{\nabla c}L_c))$),

$$\begin{aligned}
Q_{\rho_k}(x_{k+1}) &\leq Q_{\rho_k}(x_k) + \langle \nabla Q_{\rho_k}(x_k), x_{k+1} - x_k\rangle + \frac{L_{\rho_k}}{2}\|x_{k+1} - x_k\|^2 \\
&= Q_{\rho_k}(x_k) + \langle g_k, x_{k+1} - x_k\rangle + \langle \nabla Q_{\rho_k}(x_k) - g_k, x_{k+1} - x_k\rangle + \frac{L_{\rho_k}}{2}\|x_{k+1} - x_k\|^2 \\
&\leq Q_{\rho_k}(x_k) + \langle g_k, x_{k+1} - x_k\rangle + \left(\frac{L_{\rho_k}}{2} + \frac{1}{2\eta_k}\right)\|x_{k+1} - x_k\|^2 + \frac{\eta_k}{2}\|\nabla Q_{\rho_k}(x_k) - g_k\|^2, \qquad \text{(C.11)}
\end{aligned}$$

where we added and subtracted $\langle g_k, x_{k+1} - x_k\rangle$ for the equality and then used Young's inequality.

By the definition of x_{k+1} in Algorithm 2 and $x_k \in X$, we have

$$\langle x_{k+1} - x_k + \eta_k g_k, x_k - x_{k+1}\rangle \geq 0 \iff \langle g_k, x_{k+1} - x_k\rangle \leq -\frac{1}{\eta_k}\|x_k - x_{k+1}\|^2.$$

By using this estimate in (C.11), and then splitting the last term, we have

$$\begin{aligned}
Q_{\rho_k}(x_{k+1}) &\leq Q_{\rho_k}(x_k) + \left(-\frac{1}{\eta_k} + \frac{L_{\rho_k}}{2} + \frac{1}{2\eta_k}\right)\|x_{k+1} - x_k\|^2 + \frac{\eta_k}{2}\|\nabla Q_{\rho_k}(x_k) - g_k\|^2 \\
&= Q_{\rho_k}(x_k) + \left(-\frac{1}{2\eta_k} + \frac{L_{\rho_k}}{2}\right)\|x_{k+1} - x_k\|^2 + \eta_k\|\nabla Q_{\rho_k}(x_k) - g_k\|^2 - \frac{\eta_k}{2}\|\nabla Q_{\rho_k}(x_k) - g_k\|^2.
\end{aligned}$$

We take expectation on this inequality and then use Lemma C.1 to bound the expectation of the third term on the right-hand side to get

$$\mathbb{E}Q_{\rho_k}(x_{k+1}) + \frac{1}{72\tilde{L}^2\rho_{k+1}^2\eta_k}\mathbb{E}\|g_{k+1} - \nabla Q_{\rho_{k+1}}(x_{k+1})\|^2$$

$$\leq \mathbb{E}Q_{\rho_k}(x_k) + \frac{1}{72\tilde{L}^2\rho_k^2\eta_{k-1}}\mathbb{E}\|g_k - \nabla Q_{\rho_k}(x_k)\|^2 - \frac{\eta_k}{2}\mathbb{E}\|g_k - \nabla Q_{\rho_k}(x_k)\|^2$$
$$+ \left(\frac{L_{\rho_k}}{2} + \frac{7}{18\eta_k} - \frac{1}{2\eta_k}\right)\mathbb{E}\|x_{k+1} - x_k\|^2 + \frac{7m^2\tilde{C}_{\nabla c}^2\tilde{C}_c^2}{12\tilde{L}^2\rho_{k+1}^2\eta_k}|\rho_{k+1} - \rho_k|^2$$
$$+ \frac{\alpha_{k+1}^2}{12\tilde{L}^2\rho_{k+1}^2\eta_k}\left(\sigma_f^2 + m^2\rho_k^2\left(C_c^2\sigma_{\nabla c}^2 + \tilde{C}_{\nabla c}^2\sigma_c^2\right)\right)$$

First, by the definition of η_k and by $L_{\rho_k} \leq \tilde{L}_{\rho_k} \leq \tilde{L}_{\rho_{k+1}} = \tilde{L}\rho(k+1)^{1/5} \leq \tilde{L}\rho(k+1)^{3/5}$ which is due to (3.2), Jensen's inequality and $\mathbb{E}[\tilde{\nabla}Q_{\rho_k}(x, B_{k+1})] = \nabla Q_{\rho_k}(x)$, we have that $\frac{L_{\rho_k}}{2} \leq \frac{\tilde{L}\rho(k+1)^{3/5}}{2} = \frac{1}{18\eta_k}$ since $\tilde{L}\rho(k+1)^{3/5} = \frac{1}{9\eta_k}$. We consequently have $\frac{L_{\rho_k}}{2} + \frac{7}{18\eta_k} - \frac{1}{2\eta_k} \leq \frac{4}{9\eta_k} - \frac{1}{2\eta_k} \leq -\frac{1}{18\eta_k}$. We then add to both sides $Q_{\rho_{k+1}}(x_{k+1})$, use the definitions of Y_k and $\mathcal{E}_k$ along with $\frac{\eta_k}{2} \geq \frac{\eta_k}{18}$ to get

$$\frac{\eta_k}{18}\mathbb{E}\left[\eta_k^{-2}\|x_{k+1} - x_k\|^2 + \|g_k - \nabla Q_{\rho_k}(x_k)\|^2\right] \leq \mathbb{E}\left[Y_k - Y_{k+1} + |Q_{\rho_k}(x_{k+1}) - Q_{\rho_{k+1}}(x_{k+1})|\right] + \mathcal{E}_{k+1}. \tag{C.12}$$

We will now show that

$$\mathsf{d}^2(\nabla f(x_{k+1}) + \rho_k\nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1})) \leq 4(\eta_k^{-2}\|x_{k+1} - x_k\|^2 + \|g_k - \nabla Q_{\rho_k}(x_k)\|^2). \tag{C.13}$$

By the definition of x_{k+1}, we have

$$\begin{aligned}
&0 \in x_{k+1} - x_k + \eta_k g_k + \partial i_X(x_{k+1}) \\
&\iff \eta_k^{-1}(x_k - x_{k+1}) + (\nabla Q_{\rho_k}(x_k) - g_k) + (\nabla Q_{\rho_k}(x_{k+1}) - \nabla Q_{\rho_k}(x_k)) \in \partial i_X(x_{k+1}) + \nabla Q_{\rho_k}(x_{k+1}) \\
&\iff \eta_k^{-1}(x_k - x_{k+1}) + (\nabla Q_{\rho_k}(x_k) - g_k) + (\nabla Q_{\rho_k}(x_{k+1}) - \nabla Q_{\rho_k}(x_k)) \\
&\qquad\qquad \in \partial i_X(x_{k+1}) + \nabla f(x_{k+1}) + \rho_k\nabla c(x_{k+1})^\top c(x_{k+1}),
\end{aligned}$$

where the last step also used the definition of $\nabla Q_{\rho_k}(x_{k+1})$. This gives

$$\begin{aligned}
&\mathsf{d}^2(\nabla f(x_{k+1}) + \rho_k\nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1})) \\
&\leq 3\eta_k^{-2}\|x_{k+1} - x_k\|^2 + \|g_k - \nabla Q_{\rho_k}(x_k)\|^2 + \|\nabla Q_{\rho_k}(x_k) - \nabla Q_{\rho_k}(x_{k+1})\|^2 \\
&\leq 4\eta_k^{-2}\|x_{k+1} - x_k\|^2 + \|g_k - \nabla Q_{\rho_k}(x_k)\|^2,
\end{aligned}$$

where the last step is due to ∇Q_{ρ_k} being L_{ρ_k}-Lipschitz, $L_{\rho_k} \leq \tilde{L}_{\rho_k} \leq \tilde{L}_{\rho_{k+1}}$ and hence $L_{\rho_k} \leq \tilde{L}\rho(k+1)^{1/5} < 9\tilde{L}\rho(k+1)^{3/5} = \eta_k^{-1}$ by the definition of η_k. Using (C.13) on (C.12) and taking total expectation gives the result. □

C.3 Controlling the change of penalty parameters

Using variable penalty parameters allows us to remove assumptions on initialization that was done in Shi et al. (2022) for solving a special case of our problem. To handle the effect of the change on penalty parameters we have the next lemma that uses (A5) and Lemma 3.5.

Lemma C.3. *Let the assumptions in (A2), (A3), (A4), (A5) hold and let the parameters of Algorithm 2 be given as*

$$\eta_k = \frac{1}{9\tilde{L}\rho(k+1)^{3/5}}, \quad \rho_k = \rho k^{1/5}, \quad \text{and } \alpha_{k+1} = 72\tilde{L}^2\rho_{k+1}^2\eta_k^2 = \frac{72}{81}(k+1)^{-4/5},$$

for some constant $\rho > 1$, we have that

$$\begin{aligned}
&\sum_{k=1}^K \mathbb{E}|Q_{\rho_k}(x_{k+1}) - Q_{\rho_{k+1}}(x_{k+1})| \\
&\leq bQ_{\rho_1}(x_1) - \frac{bQ}{(K+1)^{3/5}} + \frac{9b}{72\tilde{L}\rho}\|g_1 - \nabla Q_{\rho_1}(x_1)\|^2 \\
&\quad + \sum_{k=1}^K \frac{b(B_f + \rho C_c^2)}{k(k+1)^{2/5}} + \sum_{k=1}^K \frac{b\mathcal{E}_{k+1}}{(k+1)^{3/5}} + \sum_{k=1}^K \frac{b\rho C_c^2}{k^{4/5}(k+1)^{3/5}} + \sum_{k=1}^K \frac{2C_{\nabla f}^2}{5\rho\delta^2 k^{6/5}},
\end{aligned}$$

where $b_k = \frac{648\cdot 8\tilde{L}}{5\delta^2(k+1)^{3/5}}$, $b = \frac{648\cdot 8\tilde{L}}{5\delta^2}$ and $\mathcal{E}_{k+1}$ is as given in Lemma 3.5.

Remark C.4. In view of Remark C.2 and since $b_k = O(k^{-3/5})$, the right-hand side in this lemma is finite.

Proof. We start with the error term

$$|Q_{\rho_k}(x_{k+1}) - Q_{\rho_{k+1}}(x_{k+1})| = (\rho_{k+1} - \rho_k)\|c(x_{k+1})\|^2. \tag{C.14}$$

We have by the assumption in (A5) and triangle inequality that

$$\begin{aligned} &\mathsf{d}(\rho_k \nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1})) \geq \rho_k \delta \|c(x_{k+1})\| \\ &\iff \|c(x_{k+1})\| \leq \frac{1}{\rho_k \delta}\left(\|\nabla f(x_{k+1})\| + \mathsf{d}(\nabla f(x_{k+1}) + \rho_k \nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1}))\right). \end{aligned} \tag{C.15}$$

Using this in (C.14) gives

$$\begin{aligned} &|Q_{\rho_k}(x_{k+1}) - Q_{\rho_{k+1}}(x_{k+1})| \\ &\leq \frac{2|\rho_k - \rho_{k+1}|}{\rho_k^2 \delta^2}\left(\|\nabla f(x_{k+1})\|^2 + \mathsf{d}^2(\nabla f(x_{k+1}) + \rho_k \nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1}))\right). \end{aligned} \tag{C.16}$$

As a result, we wish to bound

$$\begin{aligned} &\sum_{k=1}^K |Q_{\rho_k}(x_{k+1}) - Q_{\rho_{k+1}}(x_{k+1})| \\ &\leq \sum_{k=1}^K \frac{2|\rho_k - \rho_{k+1}|}{\rho_k^2 \delta^2}\left(\|\nabla f(x_{k+1})\|^2 + \mathsf{d}^2(\nabla f(x_{k+1}) + \rho_k \nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1}))\right). \end{aligned} \tag{C.17}$$

For bounding the right-hand side of this inequality, we use a crude bound that can be obtained by Lemma 3.5. By using (C.14) with the uniform upper bound $\|c(x_{k+1})\| \leq C_c$ to bound the third term on the right-hand side of the result in Lemma 3.5, we get

$$\frac{\eta_k}{72}\mathbb{E}\left[\mathsf{d}^2(\nabla f(x_{k+1}) + \rho_k \nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1}))\right] \leq \mathbb{E}\left[Y_k - Y_{k+1}\right] + \mathcal{E}_{k+1} + |\rho_k - \rho_{k+1}|C_c^2, \tag{C.18}$$

Let $b_k = \frac{648 \cdot 8 \tilde{L}}{5\delta^2(k+1)^{3/5}}$ and $b = \frac{648 \cdot 8}{5\delta^2}$. First, note that for $k \geq 1$

$$b_k \cdot \frac{\eta_k}{72} = \frac{8}{5\rho\delta^2(k+1)^{6/5}} \geq \frac{2}{5\rho\delta^2(k)^{6/5}} \geq \frac{2|\rho_k - \rho_{k+1}|}{\rho^2\delta^2 k^{2/5}} = \frac{2|\rho_k - \rho_{k+1}|}{\rho_k^2\delta^2},$$

since $|\rho_{k+1} - \rho_k| \leq \frac{\rho}{5k^{4/5}}$ and $4(k)^{6/5} \geq (k+1)^{6/5}$ for $k \geq 1$. Hence, after multiplying (C.18) by b_k, we get

$$\begin{aligned} &\frac{2|\rho_k - \rho_{k+1}|}{\rho_k^2\delta^2}\mathbb{E}\left[\mathsf{d}^2(\nabla f(x_{k+1}) + \rho_k \nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1}))\right] \\ &\leq b_k(\mathbb{E}Y_k - \mathbb{E}Y_{k+1}) + b_k\mathcal{E}_{k+1} + b_k|\rho_k - \rho_{k+1}|C_c^2. \end{aligned} \tag{C.19}$$

In view of (C.17), this gives

$$\begin{aligned} &\sum_{k=1}^K \mathbb{E}|Q_{\rho_k}(x_{k+1}) - Q_{\rho_{k+1}}(x_{k+1})| \\ &\leq \sum_{k=1}^K \left(b_k(\mathbb{E}Y_k - \mathbb{E}Y_{k+1}) + b_k\mathcal{E}_{k+1} + b_k|\rho_k - \rho_{k+1}|C_c^2 + \frac{2|\rho_k - \rho_{k+1}|C_{\nabla f}^2}{\rho_k^2\delta^2}\right), \end{aligned} \tag{C.20}$$

after also using $\|\nabla f(x_{k+1})\|^2 \leq C_{\nabla f}^2$. By the definitions in Lemma 3.5, we have $b_k\mathcal{E}_{k+1} = O\left(\frac{1}{k^{8/5}}\right)$, $b_k|\rho_k - \rho_{k+1}| = O\left(\frac{1}{k^{7/5}}\right)$, and $\frac{|\rho_k - \rho_{k+1}|}{\rho_k^2} = O\left(\frac{1}{k^{6/5}}\right)$. Hence, we now bound the term $b_k(\mathbb{E}Y_k - \mathbb{E}Y_{k+1})$. By the definition of Y_{k+1} in Lemma 3.5, we have

$$\sum_{k=1}^K b_k(\mathbb{E}Y_k - \mathbb{E}Y_{k+1})$$

$$= \sum_{k=1}^{K} b_k \mathbb{E}\left[Q_{\rho_k}(x_k) - Q_{\rho_{k+1}}(x_{k+1})\right]$$
$$+ \sum_{k=1}^{K} b_k \mathbb{E}\left[\frac{1}{72\tilde{L}^2\rho_k^2\eta_{k-1}}\|g_k - \nabla Q_{\rho_k}(x_k)\|^2 - \frac{1}{72\tilde{L}^2\rho_{k+1}^2\eta_k}\|g_{k+1} - \nabla Q_{\rho_{k+1}}(x_{k+1})\|^2\right] \tag{C.21}$$

First, we have

$$\sum_{k=1}^{K} \frac{1}{(k+1)^{3/5}}\left(Q_{\rho_k}(x_k) - Q_{\rho_{k+1}}(x_{k+1})\right)$$
$$= \sum_{k=1}^{K}\left(\frac{1}{k^{3/5}}Q_{\rho_k}(x_k) - \frac{1}{(k+1)^{3/5}}Q_{\rho_{k+1}}(x_{k+1})\right) + \sum_{k=1}^{K}\left(\frac{1}{(k+1)^{3/5}} - \frac{1}{k^{3/5}}\right)Q_{\rho_k}(x_k)$$
$$\leq Q_{\rho_1}(x_1) - \frac{\underline{Q}}{(K+1)^{3/5}} + (B_f + \rho C_c^2)\sum_{k=1}^{K}\frac{1}{k(k+1)^{2/5}},$$

where the last step is by

$$|Q_{\rho_{k+1}}(x_{k+1})| = |f(x_{k+1}) + \rho_{k+1}\|c(x_{k+1}\|^2| \leq B_f + \rho(k+1)^{1/5}C_c^2$$

and

$$\left|\frac{1}{(k+1)^{3/5}} - \frac{1}{k^{3/5}}\right| \leq \frac{1}{(k+1)^{3/5}k},$$

since

$$\frac{1}{k^{3/5}} - \frac{1}{(k+1)^{3/5}} = \frac{(k+1)^{3/5} - k^{3/5}}{(k+1)^{3/5}k^{3/5}} = \frac{(k+1)^{3/5}k^{2/5} - k}{(k+1)^{3/5}k} \leq \frac{1}{(k+1)^{3/5}k}$$

and $Q_\rho(x) \geq \underline{Q} > -\infty$ by (A4). After multiplying by b and using $b_k = \frac{b}{(k+1)^{3/5}}$ on the last estimate gives

$$\sum_{k=1}^{K} b_k\left(Q_{\rho_k}(x_k) - Q_{\rho_{k+1}}(x_{k+1})\right) \leq bQ_{\rho_1}(x_1) - \frac{b\underline{Q}}{(K+1)^{3/5}} + b(B_f + \rho C_c^2)\sum_{k=1}^{K}\frac{1}{k(k+1)^{2/5}}. \tag{C.22}$$

Next, for the second term in (C.21), we have

$$\sum_{k=1}^{K}\frac{1}{(k+1)^{3/5}}\left(\frac{1}{72\tilde{L}^2\rho_k^2\eta_{k-1}}\|g_k - \nabla Q_{\rho_k}(x_k)\|^2 - \frac{1}{72\tilde{L}^2\rho_{k+1}^2\eta_k}\|g_{k+1} - \nabla Q_{\rho_{k+1}}(x_{k+1})\|^2\right)$$
$$\leq \sum_{k=1}^{K}\left(\frac{1}{k^{3/5}}\frac{1}{72\tilde{L}^2\rho_k^2\eta_{k-1}}\|g_k - \nabla Q_{\rho_k}(x_k)\|^2 - \frac{1}{(k+1)^{3/5}}\frac{1}{72\tilde{L}^2\rho_{k+1}^2\eta_k}\|g_{k+1} - \nabla Q_{\rho_{k+1}}(x_{k+1})\|^2\right)$$
$$\leq \frac{9}{72\tilde{L}\rho}\|g_1 - \nabla Q_{\rho_1}(x_1)\|^2.$$

Hence, after multiplying this estimate by b and using $b_k = \frac{b}{(k+1)^{3/5}}$, we obtain

$$\sum_{k=1}^{K}\frac{1}{(k+1)^{3/5}}\left(\frac{1}{72\tilde{L}^2\rho_k^2\eta_{k-1}}\|g_k - \nabla Q_{\rho_k}(x_k)\|^2 - \frac{1}{72\tilde{L}^2\rho_{k+1}^2\eta_k}\|g_{k+1} - \nabla Q_{\rho_{k+1}}(x_{k+1})\|^2\right)$$
$$\leq \frac{9b}{72\tilde{L}\rho}\|g_1 - \nabla Q_{\rho_1}(x_1)\|^2. \tag{C.23}$$

We take expectations and then combine (C.22) and (C.23) in (C.21) to have

$$\sum_{k=1}^{K} b_k(\mathbb{E}Y_k - \mathbb{E}Y_{k+1}) \leq bQ_{\rho_1}(x_1) - \frac{b\underline{Q}}{(K+1)^{3/5}} + b(B_f + \rho C_c^2)\sum_{k=1}^{K}\frac{1}{k(k+1)^{2/5}}$$
$$+ \frac{9b}{72\tilde{L}\rho}\|g_1 - \nabla Q_{\rho_1}(x_1)\|^2, \tag{C.24}$$

Using this estimate in (C.20) gives the result after also substituting the values of ρ_k, b_k.

□

C.4 Main theorem

Theorem 3.1. *Let the assumptions in (A2), (A3), (A4), (A5) hold. Let*

$$\eta_k = \frac{1}{9\tilde{L}\rho(k+1)^{3/5}}, \quad \rho_k = \rho k^{1/5}, \quad \text{and } \alpha_{k+1} = \frac{72}{81}(k+1)^{-4/5},$$

for some constant $\rho > 1$ and $\tilde{L}^2 = 4\tilde{L}_{\nabla f}^2 + 4m^2(\tilde{C}_c^2\tilde{L}_{\nabla c}^2 + \tilde{C}_{\nabla c}^2\tilde{L}_c^2)$, we have that there exists λ such that

$$\mathbb{E}\left[\mathsf{d}(\nabla f(x_{\hat{k}+1}) + \nabla c(x_{\hat{k}+1})^\top \lambda, -N_X(x_{\hat{k}+1}))\right] \leq \varepsilon,$$
$$\mathbb{E}\|c(x_{\hat{k}+1})\| \leq \varepsilon.$$

with number of iterations bounded by $\tilde{O}(\varepsilon^{-5})$.

Proof. We start by summing the one iteration inequality in Lemma 3.5

$$\sum_{k=1}^{K} \frac{\eta_k}{72}\mathbb{E}\left[\mathsf{d}^2(\nabla f(x_{k+1}) + \rho_k \nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1}))\right]$$
$$\leq \mathbb{E}Y_1 + \sum_{k=1}^{K} |Q_{\rho_k}(x_{k+1}) - Q_{\rho_{k+1}}(x_{k+1})| + \sum_{k=1}^{K} \mathcal{E}_{k+1}.$$

We use that $\eta_k \geq \eta_K$ and divide both sides of the inequality by K to derive

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\mathsf{d}^2(\nabla f(x_{k+1}) + \rho_k \nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1}))\right]$$
$$\leq \frac{72}{K\eta_K}\Big(\mathbb{E}Y_1 + \sum_{k=1}^{K} |Q_{\rho_k}(x_{k+1}) - Q_{\rho_{k+1}}(x_{k+1})| + \sum_{k=1}^{K} \mathcal{E}_{k+1}\Big).$$

In view of Lemma C.3, Remark C.2 and Remark C.4, we have that the sums in the right-hand side are either finite or increase logarithmically in K and since $\eta_K = O(K^{-3/5})$, we have

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\mathsf{d}^2(\nabla f(x_{k+1}) + \rho_k \nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1}))\right] = O\left(\frac{\log(K+1)}{K^{2/5}}\right), \tag{C.25}$$

along with $\|\nabla f(x)\|^2 \leq C_{\nabla f}^2$ as per (A4).

This estimate along with $\rho_k = \rho k^{1/5}$ in (C.15) gives

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\|c(x_{k+1})\|^2 \leq \frac{1}{K}\sum_{k=1}^{K} \frac{2}{\rho^2 k^{2/5}\delta^2}\mathbb{E}\left[\|\nabla f(x_{k+1})\|^2 + \mathsf{d}^2(\nabla f(x_{k+1}) + \rho_k \nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1}))\right]$$
$$= O\left(\frac{\log(K+1)}{K^{2/5}}\right),$$

This inequality with (C.25) gives

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\mathsf{d}^2(\nabla f(x_{k+1}) + \rho_k \nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1})) + \|c(x_{k+1})\|^2\right] = O\left(\frac{\log(K+1)}{K^{2/5}}\right).$$

Hence the claims follow by using $\lambda = \rho_{\hat{k}} c(x_{\hat{k}+1})$ and Jensen's inequality. □

D EXTENSIONS

Since the arguments in these parts are mostly the same as the previous section, the analyses in these two sections do not spell out all the details but mentions the changes compared to Section C. In this section, we will consider two extensions and show how they follow by minor adjustments on the analysis of the previous section.

D.1 Dual variable updates

The algorithm in this case, written explicitly, is

$$x_{k+1} = P_X(x_k - \eta_k g_k) \tag{D.1a}$$
$$\text{Sample } B_{k+1} = (\xi^0_{k+1}, \zeta^1_{k+1}, \zeta^2_{k+1}) \in \Xi \times Z^2 \text{ and set } \nabla Q_\rho(x, \lambda, B) \text{ as (D.2)} \tag{D.1b}$$
$$g_{k+1} = \tilde{\nabla} Q_{\rho_{k+1}}(x_{k+1}, \lambda_{k+1}, B_{k+1}) + (1-\alpha_{k+1})(g_k - \tilde{\nabla} Q_{\rho_k}(x_k, \lambda_k, B_{k+1})), \tag{D.1c}$$
$$\lambda_{k+2,i} = \lambda_{k+1,i} + \gamma_{k+1}\tilde{c}_i(x_{k+1}, \zeta^2_{k+1}) \quad \forall i = \{1, \dots, m\} \tag{D.1d}$$

where we have

$$\tilde{\nabla} Q_\rho(x, \lambda, B) = \tilde{\nabla} f(x, \xi^0) + \sum_{i=1}^m \lambda_i \tilde{\nabla} c_i(x, \zeta^1) + \rho \sum_{i=1}^m \tilde{\nabla} c_i(x, \zeta^1)\tilde{c}_i(x, \zeta^2). \tag{D.2}$$

Theorem 4.1. *For the algorithm described in (D.1a)-(D.1d) (as sketched in Section 4.1), let*

$$\eta_k = \frac{1}{9\tilde{L}\rho(k+1)^{3/5}}, \quad \rho_k = \rho k^{1/5}, \quad \gamma_k = \frac{\gamma}{k(\log(k+1))^2|\tilde{c}_i(x_k, \zeta^2_k)|}, \quad \text{and } \alpha_{k+1} = 72\tilde{L}^2\rho^2_{k+1}\eta^2_k = \frac{72}{81}(k+1)^{-4/5},$$

for some constant $\rho > 1$ and $\tilde{L}^2 = \frac{3}{4}\tilde{L}^2_{\nabla f} + \frac{3}{4}m^2(\|\lambda_1\|+4)^2\tilde{L}^2_{\nabla c} + \frac{3}{2}m^2(\tilde{C}^2\tilde{L}^2_{\nabla c} + \tilde{C}^2_{\nabla c}\tilde{L}^2_c)$. Let also the assumptions in (A2), (A3), (A4), (A5) hold. We have that there exists λ such that

$$\mathbb{E}\left[\mathsf{d}(\nabla f(x_{\hat{k}+1}) + \nabla c(x_{\hat{k}+1})^\top \lambda, -N_X(x_{\hat{k}+1}))\right] \le \varepsilon,$$
$$\mathbb{E}\|c(x_{\hat{k}+1})\|^2 \le \varepsilon$$

with number of iterations bounded by $\tilde{O}(\varepsilon^{-5})$.

Proof. For convenience, let us denote $\tilde{\gamma}_k = \gamma_{k,i}|\tilde{c}_i(x_k, \zeta^2_k)| = \frac{\gamma}{k(\log(k+1))^2}$.

• Modification of Lemma C.1.

We apply Lemma A.1 with (*cf.* Lemma C.1)

$$G_k(x_k) = \nabla Q_{\rho_k}(x_k, \lambda_k) = \nabla f(x_k) + \sum_{i=1}^m \nabla c_i(x_k)\lambda_{k,i} + \rho_k \sum_{i=1}^m \nabla c_i(x_k)c_i(x_k),$$
$$G_{k+1}(x_{k+1}) = \nabla Q_{\rho_{k+1}}(x_{k+1}, \lambda_{k+1}) = \nabla f(x_{k+1}) + \sum_{i=1}^m \nabla c_i(x_{k+1})\lambda_{k+1,i} + \rho_{k+1} \sum_{i=1}^m \nabla c_i(x_{k+1})c_i(x_{k+1}),$$
$$\tilde{G}_k(x_k, B_{k+1}) = \tilde{\nabla} Q_{\rho_k}(x_k, \lambda_k, B_{k+1}) = \tilde{\nabla} f(x_k, \xi^0_{k+1}) + \sum_{i=1}^m \tilde{c}_i(x_k, \zeta^1)\lambda_{k,i} + \rho_k \sum_{i=1}^m \tilde{\nabla} c_i(x_k, \zeta^1_{k+1})\tilde{c}_i(x_k, \zeta^2_{k+1}),$$
$$\begin{aligned}\tilde{G}_{k+1}(x_{k+1}, B_{k+1}) &= \tilde{\nabla} Q_{\rho_{k+1}}(x_{k+1}, \lambda_{k+1}, B_{k+1}) \\ &= \tilde{\nabla} f(x_{k+1}, \xi^0_{k+1}) + \sum_{i=1}^m \tilde{\nabla} c_i(x_{k+1}, \zeta^1)\lambda_{k+1,i} + \rho_{k+1} \sum_{i=1}^m \tilde{\nabla} c_i(x_{k+1}, \zeta^1_{k+1})\tilde{c}_i(x_{k+1}, \zeta^2_{k+1}).\end{aligned}$$

where $\mathbb{E}_{B_{k+1}}\tilde{G}_k(x_k, B_{k+1}) = G_k(x_k)$ and $\mathbb{E}_{B_{k+1}}\tilde{G}_{k+1}(x_{k+1}, B_{k+1}) = G_{k+1}(x_{k+1})$ as before. We also define

$$\begin{aligned}\tilde{G}_{k+1}(x_k, B_{k+1}) &= \tilde{\nabla} Q_{\rho_{k+1}}(x_k, \lambda_{k+1}, B_{k+1}) \\ &= \tilde{\nabla} f(x_k, \xi^0_{k+1}) + \sum_{i=1}^m \tilde{\nabla} c_i(x_k, \zeta^1)\lambda_{k+1,i} + \rho_{k+1} \sum_{i=1}^m \tilde{\nabla} c_i(x_k, \zeta^1_{k+1})\tilde{c}_i(x_k, \zeta^2_{k+1}).\end{aligned}$$

As a result, the norm of dual vector λ_k will affect the bounds in Lemma C.1. First note that by the definition of γ_k, we have $\lambda_{k+1,i} = \lambda_{k,i} + \frac{\gamma}{k(\log(k+1))^2|\tilde{c}_i(x_k,\zeta^2_k)|}\tilde{c}_i(x_k, \zeta^2_k)$ and hence $\lambda_{k+1,i} = \lambda_{1,i} + \sum_{j=1}^k \frac{\gamma}{j(\log(j+1))^2|\tilde{c}_i(x_j,\zeta^2_j)|}\tilde{c}_i(x_j, \zeta^2_j)$ and hence

$$\|\lambda_{k+1}\|^2 = \sum_{i=1}^m (\lambda_{k+1,i})^2$$

$$\leq \sum_{i=1}^{m} \left(2(\lambda_{1,i})^2 + 2 \left| \sum_{j=1}^{k} \frac{\gamma}{j(\log(j+1))^2 |\tilde{c}_i(x_j, \zeta_j^2)|} \tilde{c}_i(x_j, \zeta_j^2) \right|^2 \right)$$
$$\leq \sum_{i=1}^{m} (2(\lambda_{1,i})^2 + 32\gamma^2)$$
$$\leq 2\|\lambda_1\|^2 + 32m\gamma^2, \tag{D.4}$$

with $|\lambda_{k+1,i}| \leq |\lambda_{1,i}| + 4\gamma$, where the second step in the inequality chain is by Young's inequality and the third by

$$\left| \sum_{j=1}^{k} \frac{\gamma}{j(\log(j+1))^2 |\tilde{c}_i(x_j, \zeta_j^2)|} \tilde{c}_i(x_j, \zeta_j^2) \right| \leq \sum_{j=1}^{k} \left| \frac{\gamma}{j(\log(j+1))^2 |\tilde{c}_i(x_j, \zeta_j^2)|} \tilde{c}_i(x_j, \zeta_j^2) \right| \leq 4\gamma.$$

Note that instead of (C.2), we now have

$$\begin{aligned} \mathbb{E}_k \|\tilde{G}_{k+1}(x_{k+1}, B_{k+1}) - \tilde{G}_{k+1}(x_k, B_{k+1})\|^2 &\leq 3\mathbb{E}_k \|\tilde{\nabla} f(x_{k+1}, \xi^0) - \tilde{\nabla} f(x_k, \xi^0)\|^2 \\ &+ 3\mathbb{E}_k \left\| \sum_{i=1}^{m} \left(\tilde{\nabla} c_i(x_{k+1}, \zeta_{k+1}^1) - \tilde{\nabla} c_i(x_k, \zeta_{k+1}^1) \right) \lambda_{k+1,i} \right\|^2 \\ &+ 3\rho_{k+1}^2 \mathbb{E}_k \left\| \sum_{i=1}^{m} \left(\tilde{\nabla} c_i(x_{k+1}, \zeta^1) \tilde{c}_i(x_{k+1}, \zeta^2) - \tilde{\nabla} c_i(x_k, \zeta^1) \tilde{c}_i(x_k, \zeta^2) \right) \right\|^2. \end{aligned} \tag{D.5}$$

Note also that, for the second term on the right-hand side of this inequality, we have

$$\begin{aligned} \mathbb{E}_k \left\| \sum_{i=1}^{m} \left(\tilde{\nabla} c_i(x_{k+1}, \zeta_{k+1}^1) - \tilde{\nabla} c_i(x_k, \zeta_{k+1}^1) \right) \lambda_{k+1,i} \right\| &\leq \sum_{i=1}^{m} \mathbb{E}_k \left\| \left(\tilde{\nabla} c_i(x_{k+1}, \zeta_{k+1}^1) - \tilde{\nabla} c_i(x_k, \zeta_{k+1}^1) \right) \lambda_{k+1,i} \right\| \\ &\leq (\|\lambda_1\| + 4) \sum_{i=1}^{m} \mathbb{E}_k \|\tilde{\nabla} c_i(x_{k+1}, \zeta_{k+1}^1) - \tilde{\nabla} c_i(x_k, \zeta_{k+1}^1)\| \\ &\leq (\|\lambda_1\| + 4) m \tilde{L}_{\nabla c} \|x_{k+1} - x_k\|^2, \end{aligned} \tag{D.6}$$

where the second inequality used $\max_i |\lambda_{k+1,i}| \leq (\|\lambda_1\| + 4)$ and the last inequality used (A3) with Jensen's inequality.

We reuse the estimation in Lemma C.1 for the third term on the right-hand side of (D.5), see (C.3). For the second term on the right-hand side of (D.5), we use (D.6) and obtain (*cf.* (C.4))

$$\begin{aligned} &\mathbb{E}_k \|\tilde{G}_{k+1}(x_{k+1}, B_{k+1}) - \tilde{G}_{k+1}(x_k, B_{k+1})\|^2 \\ &\leq \left(3\tilde{L}_{\nabla f}^2 + 3m^2 \tilde{L}_{\nabla c}^2 (\|\lambda_1\| + 4)^2 + 6\rho_{k+1}^2 m^2 \left(\tilde{C}_c^2 \tilde{L}_{\nabla c}^2 + \tilde{C}_{\nabla c}^2 \tilde{L}_c^2 \right) \right) \|x_{k+1} - x_k\|^2 \\ &\leq 4\tilde{L}^2 \rho_{k+1}^2 \|x_{k+1} - x_k\|^2, \end{aligned} \tag{D.7}$$

where

$$\tilde{L}^2 = \frac{3}{4} \tilde{L}_{\nabla f}^2 + \frac{3}{4} m^2 (\|\lambda_1\| + 4)^2 \tilde{L}_{\nabla c}^2 + \frac{3}{2} m^2 (\tilde{C}^2 \tilde{L}_{\nabla c}^2 + \tilde{C}_{\nabla c}^2 \tilde{L}_c^2).$$

Instead of (C.5) we have

$$\begin{aligned} &\mathbb{E}_k \|\tilde{G}_{k+1}(x_k, B_{k+1}) - \tilde{G}_k(x_k, B_{k+1})\|^2 \\ &\leq 2\mathbb{E}_k \left\| (\rho_{k+1} - \rho_k) \sum_{i=1}^{m} \tilde{\nabla} c_i(x_k, \zeta^1) \tilde{c}_i(x_k, \zeta^2) \right\|^2 + 2\mathbb{E}_k \left\| \sum_{i=1}^{m} \nabla \tilde{c}_i(x_k)(\lambda_{k+1,i} - \lambda_{k,i}) \right\|^2 \\ &\leq 2m^2 \tilde{C}_{\nabla c}^2 \tilde{C}_c^2 |\rho_{k+1} - \rho_k|^2 + 2\tilde{C}_{\nabla c}^2 m^2 \tilde{\gamma}_k^2, \end{aligned} \tag{D.8}$$

where the last estimate is by (A4) and $|\lambda_{k+1,i} - \lambda_{k,i}| = \tilde{\gamma}_k$ which is due to the definition of λ_{k+1}.

Moreover, instead of (C.6), we have

$$\begin{aligned}
&\mathbb{E}\|G_k(x_k) - \tilde{G}_k(x_k, B_{k+1})\|^2 \\
&\leq 3\mathbb{E}\left\|\nabla f(x_k) - \tilde{\nabla} f(x_k, \xi^0)\right\|^2 + 3\left\|\sum_{i=1}^m (\tilde{\nabla} c_i(x_k, \zeta_{k+1}^2) - \nabla c_i(x_k))\lambda_{k,i}\right\|^2 \\
&\quad + 3\rho_k^2 \mathbb{E}\left\|\sum_{i=1}^m \nabla c_i(x_k) c_i(x_k) - \sum_{i=1}^m \tilde{\nabla} c_i(x_k, \zeta^1)\tilde{c}_i(x_k, \zeta^2)\right\|^2 \\
&\leq 3\sigma_f^2 + 3\sigma_{\nabla c}^2 m^2 (2\|\lambda_1\|^2 + 32m\gamma^2) + 6m^2\rho_k^2\left(C_c^2\sigma_{\nabla c}^2 + \tilde{C}_{\nabla c}^2\sigma_c^2\right),
\end{aligned} \tag{D.9}$$

where we used (D.4) and (A3) for the second term on the right-hand side and the estimation in (C.6) for the third term on the right-hand side.

By tracing the same calculations as Lemma C.1 (i.e., substituting (D.7), (D.8), (D.9) into (A.1), dividing by $72\tilde{L}^2\rho_{k+1}^2\eta_k$ and arguing the same way as (C.9) and the following estimations), we have

$$\begin{aligned}
\eta_k \mathbb{E}\|g_k - \nabla Q_{\rho_k}(x_k, \lambda_k)\|^2 \leq{}& \frac{1}{72\tilde{L}^2\rho_k^2\eta_{k-1}}\mathbb{E}\|g_k - \nabla Q_{\rho_k}(x_k, \lambda_k)\|^2 \\
&- \frac{1}{72\tilde{L}^2\rho_{k+1}^2\eta_k}\mathbb{E}\|g_{k+1} - \nabla Q_{\rho_{k+1}}(x_{k+1}, \lambda_{k+1})\|^2 \\
&+ \frac{7}{18\eta_k}\mathbb{E}\|x_{k+1} - x_k\|^2 + \frac{7m^2\tilde{C}_{\nabla c}^2\tilde{C}_c^2}{6\tilde{L}^2\rho_{k+1}^2\eta_k}|\rho_{k+1} - \rho_k|^2 + \frac{7\tilde{\gamma}_k^2 m^2\tilde{C}_{\nabla c}^2}{6\tilde{L}^2\rho_{k+1}^2\eta_k} \\
&+ \frac{\alpha_{k+1}^2}{8\tilde{L}^2\rho_{k+1}^2\eta_k}\left(\sigma_f^2 + \sigma_{\nabla c}^2 m^2(2\|\lambda_1\|^2 + 32m\gamma^2) + 2m^2\rho_k^2\left(C_c^2\sigma_{\nabla c}^2 + \tilde{C}_{\nabla c}^2\sigma_c^2\right)\right).
\end{aligned}$$

Note that the additions compared to Lemma C.1 are the constants in the last term and the fifth term (and as described above $\tilde{L}$ is different in this case). The fifth term is summable thanks to the definition of $\tilde{\gamma}_k$ hence the order of the bound is the same.

• Modification of Lemma 3.5.

In Lemma 3.5, the only change is that in addition to changing the penalty parameter, we also have to take into account the change in dual variable and also the effect of dual variable size on the Lipschitz constant. The latter is already reflected in the definition of $\tilde{L}$ earlier in this section. For changing the dual variable, note that

$$\begin{aligned}
|Q_{\rho_k}(x_{k+1}, \lambda_{k+1}) - Q_{\rho_k}(x_{k+1}, \lambda_k)| &= \left|\sum_{i=1}^m (\lambda_{k+1,i} - \lambda_{k,i})c_i(x_{k+1})\right| \leq C_c\sum_{i=1}^m |\lambda_{k+1,i} - \lambda_k| \\
&\leq mC_c\tilde{\gamma}_k,
\end{aligned}$$

where we used triangle inequality, (A4), and the definition of $\lambda_{k+1,i}$ that gives $|\lambda_{k+1,i} - \lambda_{k,i}| = \gamma_k$.

Consequently, the result of Lemma 3.5 becomes

$$\begin{aligned}
&\frac{\eta_k}{72}\mathbb{E}\mathsf{d}^2(\nabla f(x_{k+1}) + \nabla c(x_{k+1})^\top\lambda_k + \rho_k\nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1})) \\
&\leq \mathbb{E}[Y_k - Y_{k+1} + |Q_{\rho_k}(x_{k+1}, \lambda_{k+1}) - Q_{\rho_{k+1}}(x_{k+1}, \lambda_{k+1})|] + mC_c\tilde{\gamma}_k + \mathcal{E}_{k+1},
\end{aligned} \tag{D.10}$$

where

$$Y_{k+1} = Q_{\rho_{k+1}}(x_{k+1}, \lambda_{k+1}) + \frac{1}{72\tilde{L}^2\rho_{k+1}^2\eta_k}\|g_{k+1} - \nabla Q_{\rho_{k+1}}(x_{k+1}, \lambda_{k+1})\|^2$$

and

$$\mathcal{E}_{k+1} = \frac{7m^2\tilde{C}_{\nabla c}^2\tilde{C}_c^2}{6\tilde{L}^2\rho_{k+1}^2\eta_k}|\rho_{k+1} - \rho_k|^2 + \frac{7\gamma_k^2 m^2\tilde{C}_{\nabla c}^2}{6\tilde{L}^2\rho_{k+1}^2\eta_k}$$

$$+\frac{\alpha_{k+1}^2}{8\tilde{L}^2\rho_{k+1}^2\eta_k}\left(\sigma_f^2+\sigma_{\nabla c}^2 m^2(2\|\lambda_1\|^2+32m\gamma^2)+2m^2\rho_k^2\left(C_c^2\sigma_{\nabla c}^2+\tilde{C}_{\nabla c}^2\sigma_c^2\right)\right).$$

An important remark here is that the order of the dominant term in $\mathcal{E}_k$ is still $O\left(\frac{1}{k}\right)$ as before.

• Modification of Lemma C.3.

In this case, the dual variable update will change two estimations. First is (C.16) where we will now have

$$\begin{aligned}&|Q_{\rho_k}(x_{k+1},\lambda_{k+1})-Q_{\rho_{k+1}}(x_{k+1},\lambda_{k+1})|\\&\leq\frac{3|\rho_k-\rho_{k+1}|}{\rho_k^2\delta}\left(\|\nabla f(x_{k+1})\|^2+C_c^2\|\lambda_k\|^2\right.\\&\quad\left.+\mathsf{d}^2(\nabla f(x_{k+1})+\nabla c(x_{k+1})^\top\lambda_k+\rho_k\nabla c(x_{k+1})^\top c(x_{k+1}),-N_X(x_{k+1}))\right),\end{aligned}\tag{D.11}$$

where we note that $\|\lambda_k\|^2$ is finite as per (D.4) and hence do not change the order of the dominant terms in this bound.

The second is (C.18) where now we will have, in view of (D.10), that

$$\begin{aligned}&\frac{\eta_k}{72}\mathbb{E}\mathsf{d}^2(\nabla f(x_{k+1})+\nabla c(x_{k+1})^\top\lambda_k+\rho_k\nabla c(x_{k+1})^\top c(x_{k+1}),-N_X(x_{k+1}))\\&\leq\mathbb{E}\left[Y_k-Y_{k+1}\right]+\mathcal{E}_{k+1}+|\rho_k-\rho_{k+1}|C_c^2+mC_c\tilde{\gamma}_k,\end{aligned}\tag{D.12}$$

where the additional error term $\tilde{\gamma}_k$ is summable by definition and the rest of the proof of Lemma C.3 would be the same to get $\sum_{k=1}^K|Q_{\rho_k}(x_{k+1},\lambda_{k+1})-Q_{\rho_{k+1}}(x_{k+1},\lambda_{k+1})|=O(1)$.

By combining these modified results as in Theorem 3.1 we deduce the result. □

D.2 Deterministic functional constraints

In this section, we consider the case when the constraints are not given in the expectation form. In this case, we will set the parameters accordingly to get the complexity $\tilde{O}(\varepsilon^{-4})$.

Theorem 4.2. *For Algorithm 2, set*

$$\eta_k=\frac{1}{9\tilde{L}\rho(k+1)^{1/2}},\quad \rho_k=\rho k^{1/4},$$
$$\alpha_{k+1}=\frac{72}{81(k+1)^{1/2}},$$

for some $\rho>1$ and $\tilde{L}^2=4\tilde{L}_{\nabla f}^2+4m^2(\tilde{C}_c^2\tilde{L}_{\nabla c}^2+\tilde{C}_{\nabla c}^2\tilde{L}_c^2)$. Let the assumptions in (A2), (A3), (A4), (A5) hold with a deterministic $c(x)$. We have that there exists λ such that

$$\mathbb{E}\left[\mathsf{d}(\nabla f(x_{\hat{k}+1})+\nabla c(x_{\hat{k}+1})^\top\lambda,-N_X(x_{\hat{k}+1}))\right]\leq\varepsilon,$$
$$\mathbb{E}\|c(x_{\hat{k}+1})\|\leq\varepsilon,$$

with number of iterations bounded by $\widetilde{O}(\varepsilon^{-4})$.

Proof. Since the orders of the parameter choices differ in this case, we will show the changes in the analysis of Section 3.

First, in Lemma C.1, the main change will be that we use full gradients for the constraints. In particular, we have

$$G_k(x)=\nabla Q_{\rho_k}(x)=\nabla f(x)+\rho_k\sum_{i=1}^m c_i(x)\nabla c_i(x),$$
$$\tilde{G}_k(x,\xi)=\tilde{\nabla}Q_{\rho_k}(x,\xi)=\tilde{\nabla}f(x,\xi)+\rho_k\sum_{i=1}^m c_i(x)\nabla c_i(x).$$

As a result, the main change will be in the variance term, i.e.:

$$\mathbb{E}\|G_k(x_k) - \tilde{G}_k(x_k, \xi_{k+1})\|^2 = \mathbb{E}\|\nabla f(x) - \tilde{\nabla} f(x, \xi)\|^2 \leq \sigma_f^2.$$

Since in this case, we have $72\tilde{L}^2\rho_{k+1}^2\eta_k = \frac{72\tilde{L}\rho}{9}$, it follows that (C.9) holds by $\alpha_{k+1} \leq 1$ due to $1-\alpha_{k+1}+\alpha_{k+1}^2 \leq 1$ (note that this is sufficient as per (C.10) due to the term $72\tilde{L}^2\rho_{k+1}^2\eta_k$ being independent of k in this case). This gives, instead of the result of Lemma C.1, that

$$\begin{aligned}\eta_k\mathbb{E}\|g_k - \nabla Q_{\rho_k}(x_k)\|^2 \leq \frac{1}{72\tilde{L}^2\rho^2}\mathbb{E}\|g_k - \nabla Q_{\rho_k}(x_k)\|^2 - \frac{1}{72\tilde{L}^2\rho^2}\mathbb{E}\|g_{k+1} - \nabla Q_{\rho_{k+1}}(x_{k+1})\|^2 \\ + \frac{7}{18\eta_k}\|x_{k+1} - x_k\|^2 + \frac{7m^2\tilde{C}_{\nabla c}^2\tilde{C}_c^2}{12\tilde{L}^2\rho^2}|\rho_{k+1} - \rho_k|^2 + \frac{\alpha_{k+1}^2\sigma_f^2}{12\tilde{L}^2\rho^2}. \qquad \text{(D.13)}\end{aligned}$$

The numerical estimations in Lemma 3.5 are true with the new parameters since we still have $L_{\rho_k} \leq \tilde{L}_{\rho_k} \leq \tilde{L}_{\rho_{k+1}} \leq \tilde{L}\rho_{k+1}(k+1)^{1/4} = \frac{1}{9\eta_k}$ and hence $\frac{L_{\rho_k}}{2} + \frac{7}{18\eta_k} - \frac{1}{2\eta_k} \leq -\frac{1}{18\eta_k}$. Moreover, for the estimations at the end of the proof of Lemma 3.5, we still have that $L_{\rho_k} \leq \eta_k^{-1}$ and as a result we have (*cf.* the result of Lemma 3.5)

$$\frac{\eta_k}{72}\mathbb{E}\mathsf{d}^2(\nabla f(x_{k+1}) + \rho_k\nabla c(x_{k+1})^\top c(x_{k+1}), -N_X(x_{k+1})) \leq \mathbb{E}[Y_k - Y_{k+1} + |Q_{\rho_k}(x_{k+1}) - Q_{\rho_{k+1}}(x_{k+1})|] + \mathcal{E}_{k+1},$$

where

$$Y_{k+1} = Q_{\rho_{k+1}}(x_{k+1}) + \frac{1}{72\tilde{L}^2\rho^2}\|g_{k+1} - \nabla Q_{\rho_{k+1}}(x_{k+1})\|^2$$

and

$$\mathcal{E}_{k+1} = \frac{7m^2\tilde{C}_{\nabla c}^2\tilde{C}_c^2}{12\tilde{L}^2\rho^2}|\rho_{k+1} - \rho_k|^2 + \frac{\alpha_{k+1}^2\sigma_f^2}{12\tilde{L}^2\rho^2}.$$

Note that as Remark C.2, we have that $\sum_{k=1}^K \mathcal{E}_{k+1} = O(\log(K+1))$ by the definitions of α_{k+1} and ρ_k since $|\rho_k - \rho_{k+1}| \leq \frac{\rho}{4k^{3/4}}$ and $\alpha_k = \frac{72}{81(k+1)^{1/2}}$.

For Lemma C.3, we use $b_k = \frac{72\cdot 18\tilde{L}}{\delta(k+1)^{3/4}}$ to have

$$b_k \cdot \frac{\eta_k}{72} = \frac{2}{\rho\delta(k+1)^{5/4}} \geq \frac{1}{2\rho\delta(k)^{5/4}} \geq \frac{2|\rho_k - \rho_{k+1}|}{\rho_k^2\delta},$$

by also using $|\rho_k - \rho_{k+1}| \leq \frac{\rho}{4k^{3/4}}$ and $\rho_k = \rho k^{1/4}$. We note also that, in the same way as Lemma C.3, we have $\left|\frac{1}{(k+1)^{3/4}} - \frac{1}{k^{3/4}}\right| \leq \frac{1}{(k+1)^{3/4}k}$. With these estimations and by repeating the same arguments as Lemma C.3, we get

$$\begin{aligned}&\sum_{k=1}^K \mathbb{E}|Q_{\rho_k}(x_{k+1}) - Q_{\rho_{k+1}}(x_{k+1})| \\ &\leq bQ_{\rho_1}(x_1) - \frac{bQ}{(K+1)^{3/4}} + \frac{b}{72\tilde{L}^2\rho^2}\|g_1 - \nabla Q_{\rho_1}(x_1)\|^2 \\ &\quad + \sum_{k=1}^K \frac{b(B_f + \rho C_c^2)}{k(k+1)^{1/2}} + \sum_{k=1}^K \frac{b\mathcal{E}_{k+1}}{(k+1)^{3/4}} + \sum_{k=1}^K \frac{b\rho C_c^2}{k^{3/2}} + \sum_{k=1}^K \frac{2C_{\nabla f}^2}{4\rho\delta k^{5/4}},\end{aligned}$$

where $b_k = \frac{72\cdot 18\tilde{L}}{\delta(k+1)^{3/4}}$, $b = \frac{72\cdot 18\tilde{L}}{\delta}$. As $\mathcal{E}_k = O\left(\frac{1}{k}\right)$, the right-hand side on this inequality is finite. We can then combine these inequalities the same way as Theorem 3.1 and use the definitions of $\eta_k = \frac{1}{9\tilde{L}\rho(k+1)^{1/2}}$ and $\rho_k = \rho k^{1/4}$ to get the result. □