manuscript No.
(will be inserted by the editor)

A stochastic Lagrangian-based method for nonconvex optimization with nonlinear constraints

Dimitri Papadimitriou · Bằng Công Vũ

Received: date / Accepted: date

Abstract The Augmented Lagrangian Method (ALM) is one of the most common approaches for solving linear and nonlinear constrained problems. However, for non-convex objectives, handling nonlinear inequality constraints remains challenging. In this paper, we propose a stochastic ALM with Backtracking Line Search that performs on a subset (mini-batch) of randomly selected points for the solving of nonconvex problems. The considered class of problems include both nonlinear equality and inequality constraints. Together with the formal proof of the convergence properties (in expectation) of the proposed algorithm and its computational complexity, the performance of the proposed algorithm are then compared against inexact state-of-the-art ALM methods.

Keywords Nonlinear optimization · Constrained optimization · Augmented Lagrangian · Nonconvex

Mathematics Subject Classification (2020) 65K05 · 68Q25 · 90C46 · 90C30 · 90C25

1 Introduction

Minimization problems involving both equality and inequality nonlinear constraints are of significant interest as shown by an abundant literature, e.g. [43] [30] [12] [40] to cite a few. The generic problem in nonlinear optimization is to minimize a smooth (possibly nonconvex) function $h\colon \mathbb{R}^K \to \mathbb{R}$ subject to nonlinear equality constraints and nonlinear inequality constraints. More formally,

$$\begin{aligned} &\text{minimize } h(x) \\ &\text{subject to } c_1(x) = b_1, c_2(x) \leq b_2, x \in C, \end{aligned} \tag{1.1}$$

where c_1 and c_2 are smooth vector functions from $\mathbb{R}^K$ to $\mathbb{R}^m$, $(b_1, b_2) \in \mathbb{R}^m \times \mathbb{R}^m$ and C is a closed convex subset of $\mathbb{R}^K$. This typical problem finds applicability in mathematical optimization, in semidefinite programming, and nonlinear split feasibility problems. This problem covers a wide class of applications in the domain of signal processing, including image recovery problems [5] [40], in machine learning through various constrained problems in statistical learning and in operational research with, for example, network design problems. In this context, the augmented Lagrangian-based methods (ALM) can be considered as a major breakthrough in constrained optimization, providing the basis for fundamental algorithms that have been extensively studied for various classes of problems.

Dimitri Papadimitriou (Corresponding author)
3nLab & Universite Libre de Bruxelles (ULB)
3001 Leuven, Belgium
dpapadimitriou@3nlab.org

Bằng Công Vũ
Belgium Research Center (BeRC), Huawei
3001 Leuven, Belgium
bangcvvn@gmail.com

The main objective of this paper in this respect is to design the additional mechanisms and tools required to achieve a wider applicability of augmented Lagrangian-based methods (ALM) [19] [33] in the nonlinear setting described by the above model. Introduced by Powell and Hestenes in 1969 [38] [29], ALM alternates updates of the primal variable by minimizing the Augmented Lagrangian function and the Lagrangian multiplier by dual gradient ascent. Although the latter leads to the loss of the decomposability property, the resulting method shows improved convergence properties. Since then, this method has been subject to a vast amount of studies for the solving of both convex and nonconvex problems involving linear and nonlinear constraints. Indeed, in many of the applications described above, the optimization model turns out to include nonlinearities that the nonlinear composite problem (1) essentially captures. However, constraints are often assumed to be convex meaning that the feasible set is convex; in turn, this assumption implies that equality constraint functions must be affine and inequality constraint functions must be convex. With the proposed method, the minimization of the (possibly) nonconvex objective function h can be subject to nonlinear equality and inequality constraints without imposing convexity of its functions (or operators). Moreover, our method relies on line search that performs on a subset of randomly selected points only; hence, the stochastic ALM algorithm does not require the evaluation of all gradients (of objective function and constraints) at each iteration. This property enables, as long as the selected mini-batch verifies a well-defined minimum size criterion, the solving of larger scale nonconvex problems without compromising on convergence properties and computational complexity compared to its deterministic variant.

Following these formal developments, various numerical solving frameworks and methods based on ALM have been developed since the early 90's (and even before). As part of them, the ALGENCAN algorithmic scheme [1] [2] aims to provide a general method to solve smooth (non)convex optimization problems subject to non-linear equality and inequality constraints. That is, in ALGENCAN, the Augmented Lagrangian is defined not only with respect to equality constraints, but also with respect to inequalities (without slack variables). Recall from this perspective that no ALM algorithm can solve such problem without assuming either the solving of nonconvex subproblems to their global minima or updating penalty sequence to remain bounded on the problem at hand. Hence, it aims at preserving the property of external penalty methods that global minimizers of the original problem can be obtained if each outer iteration computes a global minimizer of the subproblem. The general algorithm belongs to the Powell-Hestenes-Rockafellar (PHR) Augmented Lagrangian type. PHR-based Augmented Lagrangian methods are based on the iterative (approximate) minimization of the Lagrangian followed by the updating of the penalty parameter and the KKT multipliers approximations. It is a safeguarded Augmented Lagrangian method in the sense that approximations of the Lagrange multipliers are estimated at every iteration. The primal subproblems are solved using GENCAN [11]. GENCAN (that is included in ALGENCAN) is a Fortran code for minimizing a smooth function with a potentially large number of variables and box constraints. The framework does not use matrix manipulations at all and, to enable solving large problems with moderate computer time.

More recently, several efforts have been dedicated to tackle composite nonconvex problems of the form $h(x) = f(x) + g(x)$, where f is continuously differentiable but possibly nonconvex and g is closed convex but possibly nonsmooth, subject to (possibly nonlinear) equality constraints vector function with continuously differentiable components $c(x) = 0$ [39] and (possibly nonlinear) inequality constraints $d(x) \leq 0$ [43]. For the latter, the authors propose their equivalent reformulation as equality constraints $d(x) + s = 0$ by enforcing the non-negativity of the slack variable s. Moreover, ALM generally uses a sequence of penalty parameters $\{\rho_k\}$, which is nondecreasing and possibly unbounded. However, when the penalty parameter ρ_k becomes too large, the ALM subproblem can become ill-conditioned. Therefore, instead using bounded ρ_k sequences is desirable, although for general nonconvex (and nonsmooth) problems, this condition might not be sufficient for the convergence of ALM [9, Section 2.1]. Further comparison against ALM methods is detailed in Section 6.

Alternatively, one could think of extending the applicability of the alternating direction method of multipliers (ADMM) [20] so that it can also solve Problem (1.1). This extension could be realized by adding non-negative slack variables s to the set of optimization variables. Now, it is fundamental to observe here that Problem (1.1) includes both nonlinear equality and inequality constraints. The usual trick of adding nonnegative slack variables s does not transform the nature of the constraints

and the complexity of the problem but only if the nonlinear constraints $c_2(x) \leq b_2$ are affine, that is $c_2(x) = Lx$. Hence, ADMM can straightforwardly deal with linear inequality constraints by adding nonnegative slack variables. For nonlinear inequality constraints, the situation is completely different. Adding such equality constraints would transform the nature of the problem and the solving of its subproblems. Furthermore, transforming the constraints into indicator functions and adding them to the objective function implies in turn to compute (in every iteration) a projection onto the more complicated feasible set $\{x \mid c_2(x) \leq b_2\}$. Few papers in the literature deal with this specific issue and mostly in the convex setting [22]; therefore, we defer this study to a dedicated paper.

Contribution: The main contribution of this paper is twofold.

- First, we propose a stochastic Augmented Lagrangian Method (ALM) method relying on Backtracking Line Search that performs on a subset (mini-batch) of randomly selected points to solve optimization problems involving the minimization of a smooth (possibly nonconvex) objective function subject to both nonlinear equality and inequality constraints.
- The convergence properties (in expectation) of the proposed algorithm are then thoroughly demonstrated under very general assumptions. The main features of the proposed algorithm compared to [43] [39] are the following. Firstly, it is structured as a *single-loop* algorithm; more precisely, and does not require calling a first-order method (such as proximal gradient descent) to compute inner iterations. For example, [43] further involves the use of an intermediate interior Proximal Point (iPP) method to solve approximately the primal subproblems of the ALM. Secondly, since performing on a mini-batch whose size is $\ll M$, the proposed algorithm does not require the evaluation of all gradients (of the objective function and constraints) at each iteration. Thirdly, it uses the *backtracking line search* technique to find both primal and dual stepsize.

Structure: The remainder of this paper is structured as follows. After introducing in Section 2 the preliminary notations and definitions used throughout this paper, the proposed algorithm, namely, the stochastic ALM with Backtracking Line Search is specified in Section 3.2. Its convergence properties are thoroughly detailed in Sections 4.2 and 4.3, which determine the conditions for local convergence (in expectation) of the sequences produced by the proposed Algorithm to a critical point of the augmented Lagrangian function. Section 5 characterizes the iteration complexity of the proposed Algorithm together with its formal proof. The comparison against inexact ALM methods [43] [39] is documented in Section 6.

2 Preliminaries

The generic formulation of the problem dealt with in this paper can be stated as follows.

Problem 1 Let M and K be strictly positive integers, let $(m_q)_{q=1}^M$ be a finite sequence of strictly positive integers with $\sum_{q=1}^M m_q = m < \infty$. Let $(\omega_q)_{1\leq q\leq M}$ be a sequence in $[0,1]^M$ with $\sum_{q=1}^M \omega_q = 1$. For every $q \in \{1,\ldots,M\}$, let $h_q\colon \mathbb{R}^K \to]-\infty,+\infty]$ and $c_q\colon \mathbb{R}^K \to \mathbb{R}^{m_q}$ be smooth functions with Lipschitz continuous gradients. Let $b = (b_q)_{1\leq q\leq M} \in \oplus_{q=1}^M \mathbb{R}^{m_q}$, and S_q be a closed convex cone of $\mathbb{R}^{m_q}$. Let C be a closed convex subset of $\mathbb{R}^K$. The problem is to

$$\text{minimize } h(u) = \sum_{q=1}^{M} \omega_q h_q(u) \tag{2.1}$$

$$\text{subject to } (\forall q \in \{1,\ldots,M\})\ c_q(u) - b_q \in S_q, u \in C. \tag{2.2}$$

Notations. Denote by $\Gamma_0(\mathbb{R}^K)$ the class of all proper lower semicontinuous convex functions from $\mathbb{R}^K$ to $]-\infty,+\infty]$. The proximity operator of $f \in \Gamma_0(\mathbb{R}^K)$ is

$$\operatorname{prox}_f\colon \mathbb{R}^K \to \mathbb{R}^K\colon x \mapsto \operatorname*{argmin}_{y\in\mathbb{R}^K} f(y) + \frac{1}{2}\|x-y\|^2.$$

The conjugate function of f is denoted by f^*. When f is the indicator function of some closed convex $S \subset \mathbb{R}^K$, which is denoted by ι_S,

$$\iota_S \colon x \mapsto \begin{cases} 0 & \text{if } x \in S \\ +\infty & \text{if } x \notin S, \end{cases}$$

the proximity operator of f reduces to the projection operator denoted by P_S. The distance from $x \in \mathbb{R}^K$ to S is $d_S(x) = \|x - P_S x\|$. Note that the conjugate function of ι_S is the support function of S and is denoted by σ_S. The normal cone operator of some closed convex set C is N_C. When S is a closed convex cone, the polar cone $S^{\ominus}$ of S is defined as $S^{\ominus} = \big\{u \mid \sup \langle S \mid u\rangle \leq 0\big\}$.

Let $g \colon \mathbb{R}^K \times \mathbb{R}^m \to]-\infty, +\infty]$ be a differentiable function. We denote by $\nabla_1 g$ the gradient of g with respect to the first variable when the second variable is fixed. The notation $\nabla_2 g$ is defined similarly. Let $c \colon \mathbb{R}^K \to \mathbb{R}^m$ be a differentiable (smooth) mapping, the Jacobian of c at $u \in \mathbb{R}^K$ is denoted by $J_c(u)$ and its conjugate is denoted by $J_c(u)^\intercal$. Let $\nu > 0$, the class of all smooth mappings $c \colon \mathbb{R}^K \to \mathbb{R}^m$ with ν-Lipschitzian Jacobian is denoted by $\mathcal{C}^1_\nu(\mathbb{R}^K, \mathbb{R}^m)$.
The development of this paper relies on the following definitions.

Definition 1 Let M be a strictly positive integer. Let $(\omega_q)_{1\leq q\leq M}$ be a sequence in $[0,1]^M$ with $\sum_{q=1}^M \omega_q = 1$. The weighted inner product on the Hilbert space V, maps each pairs of vectors $(y, v) =\in V \times V$ to the scalar $\langle \cdot \,||\, \cdot \rangle$ defined as

$$\langle \cdot \,||\, \cdot \rangle : (y, v) \mapsto \sum_{q=1}^{M} \omega_q \langle v_q \mid y_q \rangle \tag{2.3}$$

$$\text{with vector norm } |||\cdot||| \colon v \mapsto \sqrt{\langle v \,||\, v\rangle}, \tag{2.4}$$

where $y = (y_q)_{1\leq q\leq M}$ and $v = (v_q)_{1\leq q\leq M}$.

Definition 2 [15] Let $f \in \Gamma_0(\mathbb{R}^K)$, $g \in \Gamma_0(\mathbb{R}^m)$, $c \in \mathcal{C}^1_\nu(\mathbb{R}^K, \mathbb{R}^m)$, and $b \in \mathbb{R}^m$. A vector $d \in \mathbb{R}^K$ defines a descent direction of $\varphi \mapsto f(u) + g(c(u) - b)$ at u, if the difference $\Delta_0\varphi(u; d)$ verifies the strict inequality

$$\Delta_0\varphi(u; d) = f(u + d) + g\big(c(u) - b + J_c(u)d\big) - \varphi(x) < 0, \tag{2.5}$$

where $J_c(u)$ denotes the Jacobian of the function c at u. A method for which, at each iteration k, the descent direction d_k, at current point u_k, verifies the strict inequality $\Delta_0\varphi(u_k; d_k) < 0$ is referred to as a descent method.

Definition 3 Let $g \in \Gamma_0(\mathbb{R}^m)$, let $b \in \mathbb{R}^m$ and $\mathcal{C}^1_\nu(\mathbb{R}^K, \mathbb{R}^m) \ni c \colon u \mapsto c(u) - b$. For every $\rho \in]0, +\infty[$, and $(u, \lambda) \in \mathbb{R}^K \times \mathbb{R}^m$, the smooth approximation of $g(c(\cdot) - b)$ is defined by

$$g_\rho \colon (u, \lambda) \mapsto \sup_{y\in\mathbb{R}^m} \Big(\langle c(u) - b \,||\, y\rangle - g^\star(y) - \frac{1}{2\rho}|||y - \lambda|||^2 \Big), \tag{2.6}$$

where ρ is referred to as the smoothing parameter and $g^\star$ denotes the Fenchel conjugate of the function g that is defined by $g^\star \colon u \mapsto \sup_{x\in\mathbb{R}^m} \big(\langle u \,||\, x\rangle - g(x)\big)$.

The function g_ρ provides a smooth approximation of g, which is known as the smoothing technique. Various numerical methods have been developed by means of this technique; see, for instance, [35,37, 6]. Several examples where g_β admits a closed-form expression can be found in [35,4].

We recall the following result concerning the differentiability of g_ρ.

Lemma 1 *For every $\rho > 0$, let the function g_ρ be defined by (2.6). Then, g_ρ is a differentiable function with respect to the variable u, and, for every $(u, \lambda) \in \mathbb{R}^K \times \mathbb{R}^m$,*

$$\nabla_1 g_\rho(u, \lambda) = (J_c(u))^\intercal \operatorname{prox}_{\rho^{-1}g^\star}(\rho^{-1}(c(u) - b) + \lambda), \tag{2.7}$$

where $(J_c(u))^\intercal$ is the (conjugate) transpose of the linear operator $J_c(u)$.

We extend this result to the case where the function g admits a separable structure. More precisely, we have the following result.

Lemma 2 *Let $g_q = \iota_{S_q}$, where ι_{S_q} denotes the indicator function of the closed convex subset S_q of $\mathbb{R}^{m_q}$, and define the function $g\colon (v_q)_{1\leq q\leq M} \mapsto \sum_{q=1}^{M} \omega_q g_q(v_q)$, where $(\omega_q)_{1\leq q\leq M}$ denotes a sequence in $[0,1]^M$ with $\sum_{q=1}^{M} \omega_q = 1$. Then, for every $\rho > 0$ and for every $(u,\lambda) \in \mathbb{R}^K \times \mathbb{R}^m$,*

$$g_\rho(u,\lambda) = \sum_{q=1}^{M} \omega_q g_{\rho,q}(u,\lambda_q), \tag{2.8}$$

$$\text{where } g_{q,\rho}(u,\lambda_q) = \sup_{y_q\in\mathbb{R}^{m_q}} \big(\langle c_q(u) - b_q \mid \lambda_q\rangle - g_q^\star(y_q) - \frac{1}{2\rho}\|y_q - \lambda_q\|^2\big), \tag{2.9}$$

is a differentiable function whose gradient with respect to the first variable u is given by

$$\nabla_1 g_\rho(u,\lambda) = \rho \sum_{q=1}^{M} \omega_q (J_{c_q}(u))^\intercal \Big(c_q(u) - b_q + \rho^{-1}\lambda_q - P_{S_q}\big(c_q(u) - b_q + \rho^{-1}\lambda_q\big)\Big). \tag{2.10}$$

Proof. Following Definition 3, the conjugate $g^\star$ of the function g can be expressed as

$$\begin{aligned} g^\star\colon v \mapsto \sup_{y\in\mathbb{R}^m} (\langle v \mid\mid y\rangle - g(y)) &= \sup_{(y_q)_{1\leq q\leq M}\in\mathbb{R}^m} \sum_{q=1}^{M} (\omega_q \langle v_q \mid y_q\rangle - \omega_q g_q(y_q)) \\ &= \sum_{q=1}^{M} \omega_q g_q^\star(v_q). \end{aligned} \tag{2.11}$$

Therefore, the smooth approximation of g with parameter ρ, $g_\rho(u,\lambda)$, is defined by

$$g_\rho(u,\lambda) = \sup_{y\in\mathbb{R}^m} \big(\langle c(u) - b \mid\mid y\rangle - g^\star(y) - \frac{1}{2\rho}|||y-\lambda|||^2\big) \tag{2.12}$$

$$\begin{aligned} &= \sum_{q=1}^{M} \omega_q \sup_{y_q\in\mathbb{R}^{m_q}} \big(\langle c_q(u) - b_q \mid y_q\rangle - {g_q}^\star(y_q) - \frac{1}{2\rho}\|y_q - \lambda_q\|^2\big) \\ &= \sum_{q=1}^{M} \omega_q g_{q,\rho}(u,\lambda_q), \end{aligned} \tag{2.13}$$

which proves (2.8). Next, it follows from (2.8) and Lemma 1 that

$$\begin{aligned} \nabla_1 g_\rho(u,\lambda) &= \sum_{q=1}^{M} \omega_q \nabla_1 g_{q,\rho}(u,\lambda_q) \\ &= \sum_{q=1}^{M} \omega_q (J_{c_q}(u))^\intercal \operatorname{prox}_{\rho g_q^\star}(\rho(c_q(u) - b_q) + \lambda_q) \\ &= \rho \sum_{q=1}^{M} \omega_q (J_{c_q}(u))^\intercal \Big(c_q(u) - b_q + \rho^{-1}\lambda_q - P_{S_q}\big(c_q(u) - b_q + \rho^{-1}\lambda_q\big)\Big), \end{aligned} \tag{2.14}$$

where the last equality follows from the Moreau's identity $(\operatorname{prox}_f(x) + \operatorname{prox}_{f^\star}(x) = x)$ and the property $\operatorname{prox}_{\iota_{S_q}} = P_{S_q}$. □

Lemma 3 *Let $\lambda_q \in \mathbb{R}^{m_q}$ and $\rho \in]0, +\infty[$. Let $(g_q)_{1\leq q\leq M}$ be defined as Lemma 2. Let*

$$e_{q,\rho} \colon u \mapsto P_{S_q}(c_q(u) - b_q + \rho^{-1}\lambda_q). \tag{2.15}$$

Then, for every $(u, \lambda) \in \mathbb{R}^K \times \mathbb{R}^m$,

$$g_{q,\rho}(u, \lambda_q) = \langle c_q(u) - b_q - e_{q,\rho}(u) \mid \lambda_q\rangle + \frac{\rho}{2}\|c_q(u) - b_q - e_{q,\rho}(u)\|^2 \tag{2.16}$$

$$= \frac{\rho}{2} d^2_{S_q}(c_q(u) - b_q + \rho^{-1}\lambda_q) - \frac{1}{2\rho}\|\lambda_q\|^2, \tag{2.17}$$

where $d_{S_q} \colon v_q \mapsto \|v_q - P_{S_q} v_q\|$ defines the distance function d_{S_q}.

Proof. Let us define $\lambda_q^\dagger := \lambda_q + \rho(c_q(u) - b_q - e_{q,\rho}(u))$. Then, the Moreau's identity gives

$$\lambda_q^\dagger = \operatorname{prox}_{\rho\sigma_{S_q}}\left(\lambda_q + \rho(c_q(u) - b_q)\right) \text{ and } \sigma_{S_q}(\lambda_q^\dagger) = \left\langle \lambda_q^\dagger \mid e_{q,\rho}(u)\right\rangle. \tag{2.18}$$

Therefore, it follows from the definition of $g_{q,\rho}$ (see Definition 3) that

$$\begin{aligned} g_{q,\rho}(u, \lambda_q) &= \sup_{v_q\in\mathbb{R}^{m_q}} \left(\langle c_q(u) - b_q \mid v_q\rangle - g_q^\star(v_q) - \frac{1}{2\rho}\|v_q - \lambda_q\|^2\right) \\ &= \left\langle c_q(u) - b_q \mid \lambda_q^\dagger\right\rangle - g_q^\star(\lambda_q^\dagger) - \frac{1}{2\rho}\|\lambda_q^\dagger - \lambda_q\|^2 \\ &= \left\langle c_q(u) - b_q - e_{q,\rho}(u) \mid \lambda_q^\dagger\right\rangle - \frac{1}{2\rho}\|\lambda_q^\dagger - \lambda_q\|^2 \\ &= \langle c_q(u) - b_q - e_{q,\rho}(u) \mid \lambda_q\rangle + \frac{\rho}{2}\|c_q(u) - b_q - e_{q,\rho}(u)\|^2, \end{aligned} \tag{2.19}$$

which proves (2.16). Next, we have

$$\rho\left\langle c_q(u) - b_q - e_{q,\rho}(u) \mid \rho^{-1}\lambda_q\right\rangle = \frac{\rho}{2}\|c_q(u)-b_q-e_{q,\rho}(u)+\rho^{-1}\lambda_q\|^2-\frac{\rho}{2}\|c_q(u)-b_q-e_{q,\rho}(u)\|^2-\frac{\rho}{2}\|\rho^{-1}\lambda_q\|^2,$$

which implies that

$$g_{q,\rho}(u, \lambda_q) = \frac{\rho}{2}\|c_q(u) - b_q - e_{q,\rho}(u) + \rho^{-1}\lambda_q\|^2 - \frac{1}{2\rho}\|\lambda_q\|^2 \tag{2.20}$$

$$= \frac{\rho}{2} d^2_{S_q}(c_q(u) - b_q + \rho^{-1}\lambda_q) - \frac{1}{2\rho}\|\lambda_q\|^2, \tag{2.21}$$

where the last equality follows from the definition of d_{S_q}. Hence, (2.17) is verified. □
Let $(u, \lambda) \in \mathbb{R}^K \times \mathbb{R}^m$ and $\rho > 0$. By using (2.13) and (2.21)

$$g_\rho(u, \lambda) = \sum_{q=1}^M \omega_q g_{q,\rho}(u, \lambda_q),$$

$$\text{where } g_{q,\rho}(u, \lambda_q) := \iota_{S_q,\rho}(u, \lambda_q) = \frac{\rho}{2} d^2_{S_q}(c_q(u) - b_q + \rho^{-1}\lambda_q) - \frac{1}{2\rho}\|\lambda_q\|^2,$$

we can define the smooth approximation of the augmented objective function $\mathcal{L}_\rho$ by involving the indicator functions $\iota_{S_q,\rho}(u, \lambda_q)$ as follows.

$$\mathcal{L}_\rho \colon (u, \lambda) \mapsto \sum_{q=1}^M \left(\omega_q h_q(u) + \frac{\rho\omega_q}{2} d^2_{S_q}(c_q(u) - b_q + \rho^{-1}\lambda_q)\right) - \frac{1}{2\rho}|||\lambda|||^2. \tag{2.22}$$

Moreover, assuming that the smoothing parameter ρ_k and multiplier λ_k are given at iteration k, one can define the function φ_k by

$$\varphi_k \colon u \mapsto \mathcal{L}_{\rho_k}(u, \lambda_k) = h(u) + \psi_k \circ c(u), \tag{2.23}$$

where $\circ$ denotes the function composition, and

$$\psi_k\colon (w_q)_{1\leq q\leq M} \mapsto \sum_{q=1}^{M} \frac{\rho\omega_q}{2} d^2_{S_q}(w_q - b_q + \rho^{-1}\lambda_{k,q}) - \frac{1}{2\rho}|||\lambda_k|||^2. \tag{2.24}$$

The following Lemma generalizes the definition of the descent direction d_k to nonconvex functions φ_k. This result is obtained by defining the function φ_k as the composition of a convex and a nonconvex function set as the argument of the former (convex) function.

Lemma 4 *Assume* $\overline{c}\colon \mathbb{R}^K \to \mathbb{R}^m \times]-\infty,+\infty] : u \mapsto \overline{c}(u) = (c(u), c_0(u))$ *together with* $c\colon \mathbb{R}^K \to \mathbb{R}^m\colon u \mapsto c(u)$ *and* $c_0 = h$. *Define the function* $\Psi_k\colon \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}\colon (u,\xi) \mapsto \psi_k(u) + \mathrm{Id}_R(\xi)$, *where* $\mathrm{Id}_R : \mathbb{R} \ni \xi \mapsto \xi$.
If the function $\psi_k\colon \mathbb{R}^m \to]-\infty,+\infty]$ *is convex; then, the function* Ψ_k *is convex. The composition* $(\Psi_k \circ \overline{c})$ *verifies the identity*

$$(\Psi_k \circ \overline{c})(u) = \psi_k \circ c(u) + \mathrm{Id}_R \circ c_0(u) \equiv \varphi_k(u), \tag{2.25}$$

where φ_k *is defined by* (2.23). *Moreover, by defining, for every* $u \in \mathrm{dom}(\varphi)$ *and* $d \in \mathbb{R}^K$,

$$\Delta\varphi_k(u;d) = \psi_k(c(u) + J_c(u)d) + \langle \nabla h(u) \mid d\rangle - \psi_k(c(u)), \tag{2.26}$$

the following identity is verified

$$\Delta_0(\Psi_k \circ \overline{c})(u;d) \equiv \Delta\varphi_k(u;d). \tag{2.27}$$

Proof. The proof follows the same reasoning as the one used when $h \equiv f + g$, see [36]. For the sake of completeness, we reproduce it here with this setting. The function $\Psi_k(u,\xi)$ defined by $\psi_k(u) + \mathrm{Id}_R(\xi)$ is convex since the identity function on $\mathbb{R}$ is convex, by assumption, the function $\psi_k(u)$ is convex, and the sum of two convex functions is again convex. The expression (2.25) follows from the definition of composition functions. Let us now prove (2.27). By definition of Δ_0 in (2.5), we get

$$\Delta_0(\Psi_k \circ \overline{c})(u;d) = \Psi_k(\overline{c}(u) + J_{\overline{c}}(u)d) - \Psi_k(\overline{c}(u))$$

By expanding the last equality using the definition of $(\Psi_k \circ \overline{c})(u;d)$ given by (2.25), we obtain

$$\begin{aligned}\Delta_0(\Psi_k \circ \overline{c})(u;d) &= \psi_k(c(u) + J_c(u)d) + \mathrm{Id}_R(c_0(u) + \langle \nabla c_0(u) \mid d\rangle) - \mathrm{Id}_R \circ c_0(u) - \psi_k \circ c(u)\\ &= \psi_k(c(u) + J_c(u)d) + \mathrm{Id}_R \circ \langle \nabla c_0(u) \mid d\rangle - \psi_k \circ c(u)\end{aligned}$$

Then, since $c_0\colon u \mapsto h(u)$ and the scalar product $\langle \nabla c_0(u) \mid d\rangle \in \mathbb{R}$, we deduce the expression

$$\Delta_0(\Psi_k \circ \overline{c})(u;d) = \psi_k(c(u) + J_c(u)d) + \langle \nabla h(u) \mid d\rangle - \psi_k(c(u)) \equiv \Delta\varphi_k(u;d), \tag{2.28}$$

which completes of the proof. □

Using Lemma 3, one can then prove that at each iteration k the descent direction computer at u_k verifies the strict inequality $\Delta\varphi(u_k;d_k) < 0$; hence, it can be referred to as defining a descent method.

We recall the basic properties of the projection operator onto the nonempty closed convex subset S_q denoted P_{S_q}, that will be used in Section 3.2.

Lemma 5 [4, Proposition 29.3, Theorem 3.16] *Let* $q \in \{1,\ldots,M\}$, *let* S_q *be a non-empty closed convex subset in* $\mathbb{R}^{m_q}$ *and* $S = \prod_{q=1}^{M} S_q$. *Then, the following hold.*

(i) *For any* $v = (v_q)_{q=1}^M \in \oplus_{q=1}^M \mathbb{R}^{m_q}$, $P_S v = (P_{S_q} v_q)_{1\leq q\leq M}$.

(ii) *For any* $v = (v_q)_{q=1}^M \in \oplus_{q=1}^M \mathbb{R}^{m_q}$,

$$p = P_S v \iff (\forall w \in S)\ \langle v - p \mid\mid w - p\rangle \leq 0. \tag{2.29}$$

Let $(\Omega, \mathcal{F}, \text{Prob})$ be a probability space and $\mathcal{H} = \mathbb{R}^K$. A $\mathcal{H}$-valued random variable is a measurable function $X : \Omega \to \mathcal{H}$, where $\mathcal{H}$ is endowed with the Borel σ-algebra. We denote by $\sigma(X)$ the σ-field generated by X. The expectation of a random variable X is denoted by $\mathbf{E}[X]$. The conditional expectation of X given a σ-field $\mathcal{A} \subset \mathcal{F}$ is denoted by $\mathbf{E}[X|\mathcal{A}]$. A $\mathcal{H}$-valued random process is a sequence $(x_k)_{k\in\mathbb{N}}$ of $\mathcal{H}$-valued random variables. The abbreviation a.s. stands for 'almost surely'.

Lemma 6 ([44, Theorem 1]) *Let $(\mathcal{F}_k)_{k\in\mathbb{N}}$ be an increasing sequence of sub-σ-algebras of $\mathcal{F}$, let $(z_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}$, $(\zeta_k)_{k\in\mathbb{N}}$ and $(t_k)_{k\in\mathbb{N}}$ be sequences of $[0, +\infty[$-valued random variables such that, for every $k \in \mathbb{N}$, z_k, θ_k, ζ_k and t_k are $\mathcal{F}_k$-measurable. Moreover, assume that $\sum_{k\in\mathbb{N}} t_k < +\infty$, $\sum_{k\in\mathbb{N}} \zeta_k < +\infty$ a.s. and*

$$(\forall k \in \mathbb{N})\ \mathbf{E}[z_{k+1}|\mathcal{F}_k] \leq (1 + t_k)z_k + \zeta_k - \theta_k \ \text{a.s.}.$$

Then $(z_k)_{k\in\mathbb{N}}$ converges a.s. to a $[0, +\infty[$-valued random variable and $(\theta_k)_{k\in\mathbb{N}}$ is summable a.s..

Corollary 1 ([46, Corollary 2.6]) *Let $(\mathcal{F}_k)_{k\in\mathbb{N}}$ be an increasing sequence of sub-σ-algebras of $\mathcal{F}$, let $(x_k)_{k\in\mathbb{N}}$ be a $[0, +\infty[$-valued random sequence such that, for every $k \in \mathbb{N}$, x_{k-1} is $\mathcal{F}_k$-measurable and*

$$\sum_{k\in\mathbb{N}} \mathbf{E}[x_k|\mathcal{F}_k] < +\infty \quad \text{a.s.}. \tag{2.30}$$

Then, $\sum\limits_{k\in\mathbb{N}} x_k < +\infty$ a.s..

3 Algorithm

In this section, we detail the specification of Algorithm 2 for solving Problem 1. The main design principles of this single-loop algorithm can be summarized as follows:

(i) Formulate a generalization of the augmented Lagrangian function by *smoothing the nonlinear constraints* $c_q(u) - b_q \in S_q$. This function is the sum of smoothed functions with respect to the primal variable u and the dual variable λ.

(ii) Then, given a point u and the Lagrangian multiplier λ, we apply the *projected mini-batch stochastic gradient* to update the primal variable as $u^+ = P_C(u - t_k d_k)$, where t_k is the primal stepsize and d_k is the mini-batch stochastic gradient; provided the size of the mini-batch satisfies a well-defined minimum size criteria.

(iii) We use the *backtracking technique* to find the primal stepsizes t_k and dual stepsizes σ_k. Then, the update of the dual variable λ is performed as $\lambda^+ = \lambda + \sigma_k \nabla_2 \mathcal{L}_\rho(u^+, \lambda)$.

Thus, this algorithm does not involve any subsolver or auxiliary solver to compute the values of primal or dual variables; hence, it is referred to as a single-loop algorithm.

The main motivation for the design of a Lagrangian-based algorithm that rely on the mini-batch stochastic gradient can be stated as follows. The stochastic gradient method was first introduced in [45]. This method, as well as its extension, the stochastic proximal gradient method, have been widely adopted nowadays as optimization method in machine learning (statistical learning, deep learning, etc.), linear inverse problem, and game theory; see [3,13,14,34,28] for examples. A main feature of the stochastic gradient is that it uses only one sample point per iteration compared to the full gradient whose computational cost becomes prohibitive when the number of points of points is large. Nevertheless, the stochastic gradient does not guarantee convergence of the iterations without either ensuring the sequence of stepsizes decreases (leading to a decreasing stepsize method) or involving a variance reduction technique. Relaxation consists of using a mini-batch approach, where only a subset of samples is used per iteration. This idea leads to the mini-batch stochastic gradient; see [14] for a detailed development. The major advantage of the mini-batch stochastic gradient is the reduction of variance when the mini-batch size increases [14,32,18]. Further comparison against inexact augmented Lagrangian methods such as [43] and [39] is provided in Section 6.

In Section 4, we characterize the convergence properties of the sequences $(u_k, \lambda_k)_{k\in\mathbb{N}}$ generated by the proposed algorithm. For this purpose, we suppose that the Jacobian J_c of the constraints c verifies the following assumption.

Assumption 1 *Let C a closed convex subset of $\mathbb{R}^K$. Assume*

$$\mu_0 = \sup_{u\in C} \|J_c(u)^\intercal\| < +\infty \quad \textit{and} \quad (\forall (u,\tilde{u}) \in C\times C)\ \|J_c(u) - J_c(\tilde{u})\| \leq \mu_c \|u-\tilde{u}\|, \tag{3.1}$$

where μ_c is a positive constant.

We further assume that the variance of d_{k,i_p} in (3.16) denoted by $\mathrm{Var}(d_{k,i_p})$ is bounded. More precisely, we need the following.

Assumption 2 *Let i_p be a random variable. The probability $\mathrm{Prob}(i_p = q)$ that the random variable i_p takes the value q verifies the property $\mathrm{Prob}(i_p = q) = \omega_q$ with $0 \leq \omega_q \leq 1$. Let d_{k,i_p} be defined by Step 2 of Algorithm 2. Assume that for all $k \in \mathbb{N}$,*

$$\mathrm{Var}(d_{k,i_p}) = \mathbf{E}_{i_p}[\|d_{k,i_p} - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2 | \mathcal{E}_k] \leq \overline{\sigma}_k^2 < +\infty, \tag{3.2}$$

where $\mathcal{E}_k$ is the $\sigma-$algebra generated by $u_0, u_1, \ldots, u_k$.

Consequently the (sample) variance of the estimator of the descent direction $d_k \in \mathbb{R}^K$ is also bounded. More precisely,

$$\mathrm{Var}(d_k) \leq \frac{1}{\mathtt{m}_k^2} \sum_{p=1}^{\mathtt{m}_k} \mathrm{Var}(d_{k,i_p}), \tag{3.3}$$

where $\mathtt{m}_k$ denotes the size of the sample. Given $\lambda_k \in S^{\ominus}$ and $\xi_k = (i_p)_{1\leq p\leq \mathtt{m}_k}$, define

$$f_{\lambda_k,\xi_k}(\cdot) = \frac{1}{\mathtt{m}_k} \sum_{p=1}^{\mathtt{m}_k} \Big(h_{i_p}(\cdot) + \frac{\rho_k}{2} d^2_{S_{i_p}}(c_{i_p}(\cdot) - b_{i_p} + \rho_k^{-1}\lambda_{k,i_p}) - \frac{1}{2\rho_k}\|\lambda_{k,i_p}\|^2 \Big). \tag{3.4}$$

In the remainder of this paper, ℓ_{ξ_k} refers to the Lipschitz constant of $\nabla f_{\lambda_k,\xi_k}$. The Lipschitz constant of $\nabla\mathcal{L}_{\rho_k}(\cdot,\lambda_k)$ is denoted by ℓ_k. Recall also that the set $S = \prod_{q=1}^M S_q$.

3.1 Line Search Procedure

Let $(\theta,\nu) \in]0,1[^2$ and $\varepsilon > 0$. We denote by $t_k = \mathrm{LS}(f_{\lambda_k,\xi_k}, u_k, \lambda_k, d_k; \theta, \nu, \varepsilon)$ the line search procedure detailed here below.

Lemma 7 *The line search Algorithm 3.1 terminates after a finite number of steps, i.e., there exists $t_k > 0$ such that*

$$f_{\lambda_k,\xi_k}(\bar{u}_{k+1}) < f_{\lambda_k,\xi_k}(u_k) + \nu t_k \Delta f_{\lambda_k,\xi_k}(u_k; -d_k) + \mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big). \tag{3.6}$$

Moreover,

$$\nu\beta_k - \frac{1}{2}t_k(1+t_k\ell_{\xi_k})^2 - t_k\big(1+t_k\ell_k\big)^2 - 2(1+\varepsilon)t_k - \varsigma_{1,k} \geq \frac{\varepsilon}{2}, \tag{3.7}$$

where $\varsigma_{1,k} \coloneqq (1+\varepsilon)4\sigma_k\Big[4\mu_0^2 + \mu_c^2 t_k^2\|d_k\|^2\Big]t_k$.

Algorithm 1 : Step Size Selection $t_k = \mathrm{LS}(f_{\lambda_k,\xi_k}, u_k, \lambda_k, d_k; \theta, \nu, \varepsilon)$

Require: Current iterate u_k, descent direction d_k, objective function f_{λ_k,ξ_k}, $\Delta f_{\lambda_k,\xi_k}(u_k; -d_k)$, projection operator P_C

Require: Parameters $\theta \in]0,1[$, $\nu \in]0,1[$, $\varepsilon > 0$, $\beta_k > 0$, $\ell_k \geq 0$, $\ell_{\xi_k} \geq 0$

▷ **Step 1: Backtracking line search**

for $(j = 0\ ;\ j > -1\ ;\ j$++$)$ **do**

  $t_\theta \leftarrow \theta^j$

  $\bar{u}_{k+1} \leftarrow P_C(u_k - t_\theta d_k)$

  if

$$f_{\lambda_k,\xi_k}(\bar{u}_{k+1}) < f_{\lambda_k,\xi_k}(u_k) + \nu t_\theta \Delta f_{\lambda_k,\xi_k}(u_k; -d_k) + \mathcal{O}\left(\frac{1}{(k+1)^{1+\varepsilon}}\right) \tag{3.5}$$

  then

    break

  end if

end for

▷ **Step 2: Final step size**

  $s_k \leftarrow \min\{1/\ell_k, 1/\ell_{\xi_k}, (\nu\beta_k - \varepsilon)/(12 + 4(1+\varepsilon))\}$

  $t_k \leftarrow \min\{t_\theta, s_k\}$

  return t_k

Proof. In view of [15, Lemma 5.1], for a fixed $\nu \in]0,1[$, there exists a finite upper limit $\overline{t_k} > 0$ of the primal stepsize interval such that for all primal stepsizes included in the open interval $]0, \overline{t_k}[$, the function f_{λ_k,ξ_k} verifies the following inequality

$$(\forall t \in \left]0, \overline{t_k}\right[)\ f_{\lambda_k,\xi_k}(u_k - td_k) \leq f_{\lambda_k,\xi_k}(u_k) + t\nu\Delta f_{\lambda_k,\xi_k}(u_k; -d_k), \tag{3.8}$$

Since $\lim_{t\downarrow 0} P_C(u_k - td_k) = u_k$ and f_{λ_k,ξ_k} is continuous, we obtain

$$\lim_{t\downarrow 0} |f_{\lambda_k,\xi_k}(P_C(u_k - td_k)) - f_{\lambda_k,\xi_k}(u_k - td_k)| = 0. \tag{3.9}$$

Therefore, there exists

$$t_k \in \left]0, \bar{t}_k\right[\tag{3.10}$$

such that

$$f_{\lambda_k,\xi_k}(P_C(u_k - td_k)) \leq f_{\lambda_k,\xi_k}(u_k - td_k) + \mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big), \tag{3.11}$$

which implies that the condition (3.5) is well-defined. We next prove (3.7). In view of (3.19), we have

$$\frac{1}{2}\nu\beta_k - \varsigma_{1,k} \geq 0. \tag{3.12}$$

Using the conditions $1 + t_k\ell_k \leq 2$ and $1 + t_k\ell_{\xi_k} \leq 2$, we obtain

$$\frac{1}{2}\nu\beta_k - \frac{1}{2}(1 + t_k\ell_{\xi_k})^2 - t_k(1 + t_k\ell_k)^2 - 2(1+\varepsilon)t_k \geq \frac{\varepsilon}{2}. \tag{3.13}$$

Summing the last inequalities, we obtain (3.7). □

3.2 Main Algorithm

Algorithm 2 ALM algorithm with backtracking

▷ **Initialization**

Set $u_0 \in C$, $u_{-1} \neq u_0$, $\lambda_0 \in S^{\ominus}$

Set $\sigma_{-1} \gg 1$, $\rho_{-1} \in]0, \infty[$, $1 \gg \varepsilon > 0$, $\theta \in]0, 1[$, $\nu \in]0, 1[$, $n \in \mathbb{N}$

Compute μ_0 from (3.1)

▷ **Main Loop**

for $k \leftarrow 0 : n$ **do**

▷ **Step 1**

Select $\rho_k \in]0, \infty[$ such that

$$\begin{cases} \beta_k := 1 - \dfrac{\rho_k}{2}\|J_c(u_k)\|^2 > \varepsilon \\ \sqrt{\rho_k}\|\|c(u_k) - b - P_S\big(c(u_k) - b + \rho_k^{-1}\lambda_k\big)\|\| \leq \min_{1\leq i\leq k}\|u_i - u_{i-1}\| \\ \rho_k < \rho_{k-1} + \varepsilon\sigma_{k-1} \end{cases} \tag{3.14}$$

▷ **Step 2**

Select mini-batch size $\mathtt{m}_k \in \mathbb{N}$.

Generate $\mathtt{m}_k$ random variables $\xi_k = (i_p)_{1\leq p\leq \mathtt{m}_k}$ with $\mathrm{Prob}(i_p = q) = \omega_q$

Compute v_{k,i_p} and d_{k,i_p}

$$v_{k,i_p} = \lambda_{k,i_p} + \rho_k\big(c_{i_p}(u_k) - b_{i_p} - P_{S_{i_p}}(c_{i_p}(u_k) - b_{i_p} + \rho_k^{-1}\lambda_{k,i_p})\big) \tag{3.15}$$

$$d_{k,i_p} = \nabla h_{i_p}(u_k) + J_{c_{i_p}}(u_k)^{\intercal} v_{k,i_p} \tag{3.16}$$

Compute d_k

$$d_k = \frac{1}{\mathtt{m}_k}\sum_{p=1}^{\mathtt{m}_k} d_{k,i_p} \tag{3.17}$$

▷ **Step 3**

Find $t_k = \mathrm{LS}(f_{\lambda_k,\xi_k}, u_k, \lambda_k, d_k, \theta, \nu, \varepsilon)$

Update

$$u_{k+1} = P_C(u_k - t_k d_k) \tag{3.18}$$

▷ **Step 4**

Compute

$$\sigma_k = \min\left\{\rho_k, \frac{1}{8t_k}\nu\beta_k(1+\varepsilon)^{-1}\Big[4\mu_0^2 + \mu_c^2 t_k^2\|d_k\|^2\Big]^{-1}\right\} \tag{3.19}$$

Update

$$\lambda_{k+1} = \lambda_k + \sigma_k\Big(c(u_{k+1}) - b - P_S\big(c(u_{k+1}) - b + \rho_k^{-1}\lambda_k\big)\Big) \tag{3.20}$$

end for

4 Convergence Properties

Before presenting our main convergence results, we summarize the general strategy followed. The main principle is to derive the descent property of the Lagrange function values $\big(\mathcal{L}_{\rho_k}(u_k, \lambda_k)\big)_{k\in\mathbb{N}}$ with respect to $(t_k\|d_k\|^2)_{k\in\mathbb{N}}$. To reach this goal, we consider the following steps:

(i) We first need to show that Step 1 and Step 3 are well defined. They are presented in Lemma 10 as well as in Lemma 7. In particular, we obtain the descent property of the stochastic function f_{ξ_k,λ_k} as in (3.5).

(ii) We further estimate $\Delta f_{\lambda_k,\xi_k}(u_k;-d_k) \leq -\beta_k\|d_k\|^2$ as proved in Lemma 9. Combining this result to (3.5), we obtain the descent of f_{ξ_k,λ_k} with respect to d_k as written in (4.47).

(iii) Based on $\mathbf{E}_{\xi_k}[\mathcal{L}_{\rho_k,\xi_k}(u_k,\lambda_{k,\xi_k})] = \mathcal{L}_{\rho_k}(u_k,\lambda_k)$, we use the results obtained in Lemma 8 where we show that the Lagrange function satisfies a sufficient decrease condition and Lemma 11 to derive the descent property of $\big(\mathcal{L}_{\rho_k}(u_k,\lambda_k)\big)_{k\in\mathbb{N}}$ from (4.47) as in (4.63).

(iv) From (4.63), it is easy to find the convergence property of the proposed method as in Theorem 3.

4.1 Auxiliary Results

We first prove several auxiliary results, which will be used in the proof of the main Theorem part of this section.

Lemma 8 *Let $k \in \mathbb{N}$. Then,*

$$\mathcal{L}_{\rho_{k+1}}(u_{k+1},\lambda_{k+1}) \leq \mathcal{L}_{\rho_k}(u_{k+1},\lambda_k) + \frac{(\sigma_k + 0.5(\rho_{k+1}-\rho_k))}{\sigma_k}|||\lambda_{k+1}-\lambda_k|||^2. \tag{4.1}$$

Proof. In view of Lemma 3, for $e_{k+1} = P_S(c(u_{k+1}) - b + \rho_{k+1}^{-1}\lambda_{k+1})$, we have

$$\mathcal{L}_{\rho_{k+1}}(u_{k+1},\lambda_{k+1}) = h(u_{k+1}) + \langle c(u_{k+1}) - b - e_{k+1} \mid\mid \lambda_{k+1}\rangle + \frac{\rho_{k+1}}{2}|||c(u_{k+1}) - b - e_{k+1}|||^2. \tag{4.2}$$

By defining $p_{k+1} := P_S(c(u_{k+1}) - b + \rho_k^{-1}\lambda_k)$, we can express the third term in the right hand side of (4.2) as

$$\begin{aligned}\frac{\rho_{k+1}}{2}|||c(u_{k+1}) - b - e_{k+1}|||^2 = \frac{\rho_{k+1}}{2}|||c(u_{k+1}) - b - p_{k+1}|||^2 + \frac{\rho_{k+1}}{2}|||e_{k+1} - p_{k+1}|||^2 \\ + \rho_{k+1}\langle c(u_{k+1}) - b - p_{k+1} \mid\mid p_{k+1} - e_{k+1}\rangle.\end{aligned} \tag{4.3}$$

The second term in the right hand side of (4.2) can be written as

$$\begin{aligned}\langle c(u_{k+1}) - b - e_{k+1} \mid\mid \lambda_{k+1}\rangle = \langle c(u_{k+1}) - b - p_{k+1} \mid\mid \lambda_k\rangle + \langle c(u_{k+1}) - b - e_{k+1} \mid\mid \lambda_{k+1}\rangle \\ - \langle c(u_{k+1}) - b - p_{k+1} \mid\mid \lambda_k\rangle.\end{aligned} \tag{4.4}$$

Using the definition of p_{k+1}, the update rule of the dual variables can be written as

$$\lambda_{k+1} = \lambda_k + \sigma_k(c(u_{k+1} - b - p_{k+1})). \tag{4.5}$$

Thus, it follows that

$$\begin{aligned}\langle c(u_{k+1}) - b - e_{k+1} \mid\mid \lambda_{k+1}\rangle &= \langle c(u_{k+1}) - b - p_{k+1} \mid\mid \lambda_k\rangle + \langle c(u_{k+1}) - b - e_{k+1} \mid\mid \lambda_{k+1}\rangle \\ &\quad - \langle c(u_{k+1}) - b - p_{k+1} \mid\mid \lambda_{k+1}\rangle + \sigma_k|||c(u_{k+1}) - b - p_{k+1}|||^2 \\ &= \langle c(u_{k+1}) - b - p_{k+1} \mid\mid \lambda_k\rangle + \langle p_{k+1} - e_{k+1} \mid\mid \lambda_{k+1}\rangle \\ &\quad + \sigma_k|||c(u_{k+1}) - b - p_{k+1}|||^2.\end{aligned} \tag{4.6}$$

Therefore, (4.2) becomes

$$\mathcal{L}_{\rho_{k+1}}(u_{k+1},\lambda_{k+1}) = \mathcal{L}_{\rho_k}(u_{k+1},\lambda_k) + (\sigma_k + \frac{\rho_{k+1}-\rho_k}{2})|||c(u_{k+1}) - b - p_{k+1}|||^2 + o_k, \tag{4.7}$$

where we set

$$\begin{aligned} o_k &\coloneqq \langle p_{k+1} - e_{k+1} \mid\mid \lambda_{k+1} \rangle + \rho_{k+1} \langle c(u_{k+1}) - b - p_{k+1} \mid\mid p_{k+1} - e_{k+1} \rangle + \frac{\rho_{k+1}}{2} |||e_{k+1} - p_{k+1}|||^2 \\ &= \langle p_{k+1} - e_{k+1} \mid\mid \lambda_{k+1} \rangle + \rho_{k+1} \langle c(u_{k+1}) - b - e_{k+1} \mid\mid p_{k+1} - e_{k+1} \rangle - \frac{\rho_{k+1}}{2} |||p_{k+1} - e_{k+1}|||^2 \\ &\leq \rho_{k+1} \Big(\langle p_{k+1} - e_{k+1} \mid\mid \rho_{k+1}^{-1} \lambda_{k+1} \rangle + \langle c(u_{k+1}) - b - e_{k+1} \mid\mid p_{k+1} - e_{k+1} \rangle \Big) \\ &= \rho_{k+1} \Big(\langle p_{k+1} - e_{k+1} \mid\mid \rho_{k+1}^{-1} \lambda_{k+1} + c(u_{k+1}) - b - e_{k+1} \rangle \Big) \\ &\leq 0, \end{aligned} \tag{4.8}$$

where the last inequality follows from Lemma 5. Therefore, using the expression (4.5), the conclusion follows from (4.7). □

Lemma 9 *Let $k \in \mathbb{N}$ and let β_k, d_k and f_{λ_k,ξ_k} be defined, respectively, by Step 1, Step 2 and Step 3 of Algorithm 2. Then,*

$$\Delta f_{\lambda_k,\xi_k}(u_k; -d_k) \leq -\beta_k \|d_k\|^2. \tag{4.9}$$

Proof. At each iteration $k \in \mathbb{N}$, define

$$\begin{cases} c_{\xi_k} = (c_{i_p})_{1 \leq p \leq \mathtt{m}_k} \\ v_{\xi_k} = (v_{k,i_p})_{1 \leq p \leq \mathtt{m}_k} \\ b_{\xi_k} = (b_{i_p})_{1 \leq p \leq \mathtt{m}_k} \\ \lambda_{k,\xi_k} = (\lambda_{k,i_p})_{1 \leq p \leq \mathtt{m}_k} \\ S_{\xi_k} = (S_{i_p})_{1 \leq p \leq \mathtt{m}_k} \\ h_{\xi_k}(\cdot) = \dfrac{1}{\mathtt{m}_k} \displaystyle\sum_{p=1}^{\mathtt{m}_k} h_{i_p}(\cdot) \\ \psi_{\xi_k}(\cdot) = \dfrac{1}{\mathtt{m}_k} \displaystyle\sum_{p=1}^{\mathtt{m}_k} \left(\frac{\rho_k}{2} d^2_{S_{i_p}}((\cdot) - b_{i_p} + \rho_k^{-1} \lambda_{k,i_p}) - \frac{1}{2\rho_k} \|\lambda_{k,i_p}\|^2 \right) \\ \mathcal{L}_{\rho_k,\xi_k}(\cdot, \lambda_k) = h_{\xi_k}(\cdot) + (\psi_{\xi_k} \circ c_{\xi_k})(\cdot). \end{cases}$$

Then, we obtain

$$f_{\lambda_k,\xi_k}(\cdot) = \mathcal{L}_{\rho_k,\xi_k}(\cdot, \lambda_k). \tag{4.10}$$

The direction $d_k \in \mathbb{R}^K$ defined by (3.17) satisfies following (2.26),

$$\Delta f_{\lambda_k,\xi_k}(u_k; d_k) = \psi_{\xi_k}(c_{\xi_k}(u_k) + J_{c_{\xi_k}}(u_k) d_k) + \langle \nabla h_{\xi_k}(u_k) \mid d_k \rangle - \psi_{\xi_k}(c_{\xi_k}(u_k)). \tag{4.11}$$

For the sake of clarity and conciseness, let us define the following

$$\begin{cases} e_{\xi_k} \coloneqq P_{S_{\xi_k}}(c_{\xi_k}(u_k) - b_{\xi_k} + \rho_k^{-1} \lambda_{\xi_k}), \\ z_{\xi_k} \coloneqq P_{S_{\xi_k}}(c_{\xi_k}(u_k) + J_{c_{\xi_k}}(u_k) d_k - b_{\xi_k} + \rho_k^{-1} \lambda_k), \\ s_{\xi_k} \coloneqq J_{c_{\xi_k}}(u_k) d_k - z_{\xi_k} + e_{\xi_k}. \end{cases} \tag{4.12}$$

We also use the following scalar product

$$\langle\langle \cdot \mid \cdot \rangle\rangle \colon (w_{\xi_k}, v_{\xi_k}) \mapsto \sum_{p=1}^{\mathtt{m}_k} \langle w_{i_p} \mid v_{i_p} \rangle$$

$$\text{with vector norm } |||\cdot||| \colon v_{\xi_k} \mapsto \sqrt{\langle\langle v_{\xi_k} \mid v_{\xi_k} \rangle\rangle}.$$

Using these notations, by Lemma 3, we have

$$\begin{aligned}
&\mathtt{m}_k\psi_{\xi_k}(c_{\xi_k}(u_k)+J_{c_{\xi_k}}(u_k)d_k)\\
&\quad=\langle\langle c_{\xi_k}(u_k)+J_{c_{\xi_k}}(u_k)d_k-b_{\xi_k}-z_{\xi_k}\mid\lambda_{\xi_k}\rangle\rangle+\frac{\rho_k}{2}|||c_{\xi_k}(u_k)+J_{c_{\xi_k}}(u_k)d_k-b_{\xi_k}-z_{\xi_k}|||^2\\
&\quad=\langle\langle c_{\xi_k}(u_k)-b_{\xi_k}-e_{\xi_k}+s_{\xi_k}\mid\lambda_{\xi_k}\rangle\rangle+\frac{\rho_k}{2}|||c_{\xi_k}(u_k)-b_{\xi_k}-e_{\xi_k}+s_{\xi_k}|||^2\\
&\quad=\langle\langle c_{\xi_k}(u_k)-b_{\xi_k}-e_{\xi_k}\mid\lambda_{\xi_k}\rangle\rangle+\langle\langle s_{\xi_k}\mid\lambda_{\xi_k}\rangle\rangle+\frac{\rho_k}{2}|||c_{\xi_k}(u_k)-b_{\xi_k}-e_{\xi_k}|||^2\\
&\qquad+\rho_k\langle\langle c_{\xi_k}(u_k)-b_{\xi_k}-e_{\xi_{\xi_k}}\mid s_{\xi_k}\rangle\rangle+\frac{\rho_k}{2}|||s_{\xi_k}|||^2,
\end{aligned}\tag{4.13}$$

which implies, using the definition of e_{ξ_k}, that

$$\begin{aligned}
&\mathtt{m}_k\big(\psi_{\xi_k}(c_{\xi_k}(u_k)+J_{c_{\xi_k}}(u_k)d_k)-\psi_{\xi_k}(c_{\xi_k}(u_k))\big)\\
&\qquad=\langle\langle s_{\xi_k}\mid\lambda_{\xi_k}\rangle\rangle+\rho_k\langle\langle c_{\xi_k}(u_k)-b_{\xi_k}-e_{\xi_k}\mid s_{\xi_k}\rangle\rangle+\frac{\rho_k}{2}|||s_{\xi_k}|||^2\\
&\qquad=\langle\langle\lambda_{\xi_k}+\rho_k\big(c_{\xi_k}(u_k)-b_{\xi_k}-e_{\xi_k}\big)\mid s_{\xi_k}\rangle\rangle+\frac{\rho_k}{2}|||s_{\xi_k}|||^2\\
&\qquad=\langle\langle v_{\xi_k}\mid s_{\xi_k}\rangle\rangle+\frac{\rho_k}{2}|||s_{\xi_k}|||^2.
\end{aligned}\tag{4.14}$$

Note that

$$|||s_{\xi_k}|||^2=|||J_{c_{\xi_k}}(u_k)d_k|||^2+2\langle\langle J_{c_{\xi_k}}(u_k)d_k\mid e_{\xi_k}-z_{\xi_k}\rangle\rangle+|||e_{\xi_k}-z_{\xi_k}|||^2.\tag{4.15}$$

Therefore,

$$\begin{aligned}
&\mathtt{m}_k\big(\psi_{\xi_k}(c_{\xi_k}(u_k)+J_{c_{\xi_k}}(u_k)d_k)-\psi_{\xi_k}(c_{\xi_k}(u_k))\big)\\
&\le\langle\langle v_{\xi_k}\mid s_{\xi_k}\rangle\rangle+\frac{\rho_k}{2}|||J_{c_{\xi_k}}(u_k)d_k|||^2+\rho_k\langle\langle J_{c_{\xi_k}}(u_k)d_k\mid e_{\xi_k}-z_{\xi_k}\rangle\rangle+\frac{\rho_k}{2}|||e_{\xi_k}-z_{\xi_k}|||^2\\
&=\langle\langle J_{c_{\xi_k}}(u_k)^{\intercal}v_{\xi_k}\mid d_k\rangle\rangle+\frac{\rho_k}{2}|||J_{\xi_k}(u_k)d_k|||^2+\langle\langle v_{\xi_k}+\rho_k J_{\xi_k}(u_k)d_k\mid e_{\xi_k}-z_{\xi_k}\rangle\rangle+\frac{\rho_k}{2}|||e_{\xi_k}-z_{\xi_k}|||^2.
\end{aligned}\tag{4.16}$$

The weighted inner product $\langle\langle v_{\xi_k}+\rho_k J_{c_{\xi_k}}(u_k)d_k\mid e_{\xi_k}-z_{\xi_k}\rangle\rangle$ satisfies

$$\begin{aligned}
&\langle\langle v_{\xi_k}+\rho_k J_{c_{\xi_k}}(u_k)d_k\mid e_{\xi_k}-z_{\xi_k}\rangle\rangle\\
&\qquad=\rho_k\langle\langle c_{\xi_k}(u_k)-b_{\xi_k}+J_{c_{\xi_k}}(u_k)d_k+\rho_k^{-1}\lambda_{\xi_k}-e_{\xi_k}\mid e_{\xi_k}-z_{\xi_k}\rangle\rangle\\
&\qquad=\rho_k\langle\langle c_{\xi_k}(u_k)-b_{\xi_k}+J_{c_{\xi_k}}(u_k)d_k+\rho_k^{-1}\lambda_{\xi_k}-z_{\xi_k}\mid e_{\xi_k}-z_{\xi_k}\rangle\rangle-\rho_k|||z_{\xi_k}-e_{\xi_k}|||^2\\
&\qquad\le-\rho_k|||z_{\xi_k}-e_{\xi_k}|||^2,
\end{aligned}\tag{4.17}$$

where the last inequality follows from Lemma 5. In turn,

$$\psi_{\xi_k}(c_{\xi_k}(u_k)+J_{c_{\xi_k}}(u_k)d_k)-\psi_{\xi_k}(c_{\xi_k}(u_k))\le\frac{1}{\mathtt{m}_k}\Big(\langle\langle J_{c_{\xi_k}}(u_k)^{\intercal}v_{\xi_k}\mid d_k\rangle\rangle+\frac{\rho_k}{2}|||J_{c_{\xi_k}}(u_k)d_k|||^2\Big).\tag{4.18}$$

Adding $\langle\langle\nabla h_{\xi_k}(u_k)\mid d_k\rangle\rangle$ to both sides of (4.18) and using the definition of the descent direction d_k, we obtain

$$\Delta f_{\lambda_k,\xi_k}(u_k;d_k)\le\langle\langle\nabla h_{\xi_k}(u_k)+\frac{1}{\mathtt{m}_k}J_{c_{\xi_k}}(u_k)^{\intercal}v_{\xi_k}\mid d_k\rangle\rangle+\frac{\rho_k}{2\mathtt{m}_k}|||J_{c_{\xi_k}}(u_k)d_k|||^2\tag{4.19}$$

Observe that the stochastic direction d_k is the gradient of f_{λ_k,ξ_k} at the current point u_k, i.e., $d_k = \nabla f_{\lambda_k,\xi_k}(u_k)$. Hence,

$$\begin{aligned}\Delta f_{\lambda_k,\xi_k}(u_k;-d_k) &\leq -\|d_k\|^2 + \frac{\rho_k}{2\mathfrak{m}_k}\|\|J_{c_{\xi_k}}(u_k)d_k\|\|^2 \\ &\leq -\left(1 - \frac{\rho_k}{2\mathfrak{m}_k}\|J_{c_{\xi_k}}(u_k)\|^2\right)\|d_k\|^2 \\ &\leq -\beta_k\|d_k\|^2 \\ &< 0, \end{aligned} \tag{4.20}$$

where the second inequality follows from (3.14). □

Lemma 10 *The sequence $(\lambda_k)_{k\in\mathbb{N}}$ belongs to the polar cone $S^{\ominus}$ when $\lambda_0 \in S^{\ominus}$ and Step 1 of Algorithm 2 is well defined.*

Proof. Suppose that $\lambda_k \in S^{\ominus}$. Let $u \in C$ and $\rho > 0$ and set $a = c(u) - b$. Then, it follows from [4, Theorem 6.30(i)] and [4, Proposition 29(ii)] that

$$\begin{aligned}\rho\Big(a - P_S\big(a + \rho^{-1}\lambda_k\big)\Big) &= \rho\Big(P_{S^{\ominus}}\big(a + \rho^{-1}\lambda_k\big) - \rho^{-1}\lambda_k\Big) \\ &= \rho P_{S^{\ominus}}\big((\rho a + \lambda_k)/\rho\big) - \lambda_k \\ &= \rho P_{S^{\ominus}/\rho}\big((\rho a + \lambda_k)/\rho\big) - \lambda_k \\ &= P_{S^{\ominus}}\big(\rho a + \lambda_k\big) - \lambda_k. \end{aligned} \tag{4.21}$$

The latter equality implies that, for $u = u_k$, the following identity is verified

$$\lambda_k = P_{S^{\ominus}}\big(\rho(c(u_k) - b) + \lambda_k\big) - \rho\big(c(u_k) - b - P_S(c(u_k) - b + \rho^{-1}\lambda_k)\big). \tag{4.22}$$

Therefore,

$$\rho\|\|c(u_k) - b - P_S(c(u_k) - b + \rho^{-1}\lambda_k)\|\| = \|\|\lambda_k - P_{S^{\ominus}}\big(\rho(c(u_k) - b) + \lambda_k\big)\|\| \leq \rho\|\|c(u_k) - b\|\|. \tag{4.23}$$

Hence, by choosing

$$\rho_k = \rho \leq \min\left\{\Big(\min_{1\leq i\leq k}\|u_i - u_{i-1}\|\Big)^2\|\|c(u_k) - b\|\|^{-2}, \rho_{k-1} + \varepsilon\sigma_{k-1}, 2(1-\varepsilon)\|J_c(u_k)\|^{-2}\right\}, \tag{4.24}$$

we get

$$\begin{cases} 1 - \dfrac{\rho_k}{2}\|J_c(u_k)\|^2 > \varepsilon \\ \sqrt{\rho_k}\|\|c(u_k) - b - P_S\big(c(u_k) - b + \rho_k^{-1}\lambda_k\big)\|\| \leq \min\limits_{1\leq i\leq k}\|u_i - u_{i-1}\| \\ \rho_k < \rho_{k-1} + \varepsilon\sigma_{k-1}. \end{cases} \tag{4.25}$$

Consequently, Step 1 is well defined when $\lambda_k \in S^{\ominus}$. We next prove $\lambda_{k+1} \in S^{\ominus}$. Indeed, we have

$$\lambda_k = P_{S^{\ominus}}\big(\rho_k(c(u_{k+1}) - b) + \lambda_k\big) - \rho_k\big(c(u_{k+1}) - b - P_S(c(u_{k+1}) - b + \rho_k^{-1}\lambda_k)\big). \tag{4.26}$$

Thus

$$\lambda_{k+1} = (1 - \sigma_k/\rho_k)\lambda_k + (\sigma_k/\rho_k)P_{S^{\ominus}}(c(u_{k+1}) - b + \rho_k^{-1}\lambda_k) \in S^{\ominus}, \tag{4.27}$$

where the last inclusion follows from $\lambda_k \in S^{\ominus}$ and $\sigma_k \leq \rho_k$. Therefore, the lemma is proved by induction. □

Lemma 11 *Let d_k be defined by (3.17). Set*

$$\varsigma_{1,k} := (1+\varepsilon)4\sigma_k\Big[4\mu_0^2 + \mu_c^2 t_k^2\|d_k\|^2\Big]t_k. \tag{4.28}$$

Then

$$\mathbf{E}_{\xi_k}\Big[\frac{1+\varepsilon}{\sigma_k}|||\lambda_{k+1}-\lambda_k|||^2|\mathcal{E}_k\Big] \leq \mathbf{E}_{\xi_k}\left[\varsigma_{1,k}t_k\|d_k\|^2|\mathcal{E}_k\right] + 2(1+\varepsilon)t_{k-1}^2\|d_{k-1}\|^2. \tag{4.29}$$

Proof. Define

$$(\forall k\in\mathbb{N})\ e_k := P_S(c(u_k)-b+\rho_k^{-1}\lambda_k) \text{ and } q_k := c(u_k)-b-e_k. \tag{4.30}$$

and

$$(\forall k\in\mathbb{N})\ p_{k+1} := P_S(c(u_{k+1})-b+\rho_k^{-1}\lambda_k) \text{ and } \bar{q}_{k+1} := c(u_{k+1})-b-p_{k+1}. \tag{4.31}$$

Then, by the update rules

$$\begin{cases} v_k = \lambda_k + \rho_k q_k \\ \lambda_{k+1} = \lambda_k + \sigma_k\bar{q}_{k+1}, \end{cases} \tag{4.32}$$

we obtain the following inequality

$$\begin{aligned} |||\lambda_{k+1}-\lambda_k|||^2 &= \sigma_k^2\|\bar{q}_{k+1}\|^2 \\ &\leq 2\sigma_k^2\left(|||q_k|||^2 + |||q_k\bar{q}_{k+1}|||^2\right) \\ &= 2\sigma_k^2\left(|||q_k|||^2 + |||c(u_{k+1})-c(u_k)-p_{k+1}+e_k|||^2\right) \\ &\leq 2\sigma_k^2|||q_k|||^2 + 4\sigma_k^2|||c(u_{k+1})-c(u_k)|||^2 + 4\sigma_k^2|||p_{k+1}-e_k|||^2. \end{aligned} \tag{4.33}$$

Using (4.31), the third term in the RHS of (4.33) becomes

$$\begin{aligned} |||p_{k+1}-e_k|||^2 &= |||P_S(c(u_{k+1})-b+\rho_k^{-1}\lambda_k) - P_S(c(u_k)-b+\rho_k^{-1}\lambda_k)|||^2 \\ &\leq |||c(u_{k+1})-c(u_k)|||^2. \end{aligned} \tag{4.34}$$

Therefore, inequality (4.33) can be written as

$$|||\lambda_{k+1}-\lambda_k|||^2 \leq 2\sigma_k^2|||q_k|||^2 + 8\sigma_k^2|||c(u_{k+1})-c(u_k)|||^2. \tag{4.35}$$

By the Step 1 of Algorithm 2 and $\sigma_k \leq \rho_k$, the first term in the RHS of (4.35) verifies the inequality

$$2\sigma_k^2|||q_k|||^2 \leq 2\sigma_k|||u_k-u_{k-1}|||^2. \tag{4.36}$$

Since the Jacobian $J_c(u_k)$ of c is μ_c-Lipschitz continuous on the subset C of $\mathbb{R}^m$, the second term in the RHS of (4.35) satisfies the inequality

$$\begin{aligned} |||c(u_{k+1})-c(u_k)|||^2 &\leq \Big(|||J_c(u_k)(u_{k+1}-u_k)||| + (\mu_c/2)\|u_{k+1}-u_k\|^2\Big)^2 \\ &\leq 2|||J_c(u_k)(u_{k+1}-u_k)|||^2 + (\mu_c^2/2)\|u_{k+1}-u_k\|^4 \\ &\leq 2\|J_c(u_k)\|^2\|(u_{k+1}-u_k)\|^2 + (\mu_c^2/2)\|u_{k+1}-u_k\|^4. \end{aligned} \tag{4.37}$$

By our assumption, $\sup_{k\in\mathbb{N}}\|J_c(u_k)\| \leq \mu_0$ is finite. It follows that

$$8\sigma_k^2|||c(u_{k+1})-c(u_k)|||^2 \leq 16\mu_0^2\sigma_k^2\|(u_{k+1}-u_k)\|^2 + 4\mu_c^2\sigma_k^2\|u_{k+1}-u_k\|^4. \tag{4.38}$$

Summing the RHS of (4.36) and (4.38), we deduce from (4.35) that

$$|||\lambda_{k+1}-\lambda_k|||^2 \leq 16\mu_0^2\sigma_k^2\|u_{k+1}-u_k\|^2 + 4\mu_c^2\sigma_k^2\|u_{k+1}-u_k\|^4 + 2\sigma_k\|u_k-u_{k-1}\|^2. \tag{4.39}$$

Since $\|u_{k+1}-u_k\| \leq t_k\|d_k\|$, we further bound (4.39) as

$$\frac{1+\varepsilon}{\sigma_k}|||\lambda_{k+1}-\lambda_k|||^2 \leq (1+\varepsilon)\Big[16\mu_0^2\sigma_k t_k^2\|d_k\|^2 + 4\mu_c^2\sigma_k t_k^4\|d_k\|^4 + 2t_{k-1}^2\|d_{k-1}\|^2\Big]. \tag{4.40}$$

$$\begin{aligned} &= (1+\varepsilon)\Big[16\mu_0^2\sigma_k t_k + 4\mu_c^2\sigma_k t_k^3\|d_k\|^2\Big]t_k\|d_k\|^2 + 2(1+\varepsilon)t_{k-1}^2\|d_{k-1}\|^2 \\ &= \varsigma_{1,k}t_k\|d_k\|^2 + 2(1+\varepsilon)t_{k-1}^2\|d_{k-1}\|^2, \end{aligned} \tag{4.41}$$

which proves (4.29) by taking the expectation with respect to ξ_k on both sides of (4.41) and using $\mathbf{E}_{\xi_k}[t_{k-1}^2\|d_{k-1}\|^2] = t_{k-1}^2\|d_{k-1}\|^2$. □

4.2 Main Theorem

Theorem 3 *Let $((u_k,\lambda_k))_{k\in\mathbb{N}}$ be the primal-dual sequence generated by Algorithm 2. Suppose that Assumptions 1 & 2 are satisfied and $(\mathcal{L}_{\rho_k}(u_k,\lambda_k))_{k\in\mathbb{N}}$ is bounded below. Further assume that, the size m_k of the mini-batch selected at each iteration k, verifies*

$$m_k \geq \mathcal{O}\left(\overline{\sigma}_k^2 t_{k,max}(k+1)^{1+\varepsilon}\right) \tag{4.42}$$

together with $(1+t_k^2\ell_{\xi_k}) \leq t_{k,max} < +\infty$ a.s., and $(1+t_k^2\ell_k) \leq t_{k,max} < +\infty$ a.s., where $t_{k,max}$ is independent of ξ_k.

Then, the following hold.

(i) *The sequence $(\mathbf{E}_{\xi_k}\left[\|\frac{u_{k+1}-u_k}{\sqrt{t_k}}\|^2|\mathcal{E}_k\right])_{k\in\mathbb{N}}$ is summmable, i.e.,*

$$\sum_{k\in\mathbb{N}} \mathbf{E}_{\xi_k}\left[\|\frac{u_{k+1}-u_k}{\sqrt{t_k}}\|^2|\mathcal{E}_k\right] < +\infty. \tag{4.43}$$

(ii) *The sequence $(\mathbf{E}_{\xi_k}\left[\sigma_k\||c(u_{k+1})-b-P_S(c(u_{k+1}-b+\rho_k^{-1}\lambda_k)\|\|^2|\mathcal{E}_k\right])_{k\in\mathbb{N}}$ is summable, i.e.,*

$$\sum_{k\in\mathbb{N}} \mathbf{E}_{\xi_k}\left[\sigma_k\||c(u_{k+1})-b-P_S(c(u_{k+1}-b+\rho_k^{-1}\lambda_k)\|\|^2|\mathcal{E}_k\right] < +\infty. \tag{4.44}$$

(iii) *Define $u_{k+1}^e = P_C(u_k - t_k\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k))$. Then, the sequence $(\mathbf{E}\left[\|\frac{u_{k+1}^e-u_k}{\sqrt{t_k}}\|^2|\mathcal{E}_k\right])_{k\in\mathbb{N}}$ is summmable, i.e.,*

$$\sum_{k\in\mathbb{N}} \mathbf{E}_{\xi_k}\left[\|\frac{u_{k+1}^e-u_k}{\sqrt{t_k}}\|^2|\mathcal{E}_k\right] < +\infty, \text{ a.s.} \tag{4.45}$$

(iv) *Choosing σ_k such that $\sup_{k\in\mathbb{N}}\sigma_k \leq \sigma_\infty < +\infty$ where σ_∞ is independent of ξ_k. Then, the sequence $(\mathbf{E}_{\xi_k}\left[\sigma_k\||c(u_{k+1}^e)-b-P_S(c(u_{k+1}^e)-b+\rho_k^{-1}\lambda_k)\|\|^2|\mathcal{E}_k\right])_{k\in\mathbb{N}}$ is summable, i.e.,*

$$\sum_{k\in\mathbb{N}} \mathbf{E}_{\xi_k}\left[\sigma_k\||c(u_{k+1}^e)-b-P_S(c(u_{k+1}^e)-b+\rho_k^{-1}\lambda_k)\|\|^2|\mathcal{E}_k\right] < +\infty, \text{ a.s.} \tag{4.46}$$

Proof. In this proof, we denote by $\mathbf{E}_{\xi_k}[X] = \mathbf{E}_{\xi_k}[X|\mathcal{E}_k]$ the conditional expectation of X with respect to $\mathcal{E}_k$. Using Lemma 9 and Lemma 7, we obtain the following

$$f_{\lambda_k,\xi_k}(u_{k+1}) < \mathcal{L}_{\rho_k,\xi_k}(u_k,\lambda_{k,\xi_k}) - t_k\beta_k\nu\|d_k\|^2 + \mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big). \tag{4.47}$$

Note that $u_{k+1}^e = P_C(u_k - t_k\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k))$. Then, it follows from the nonexpansiveness of the projection operator P_C that

$$\|u_{k+1}^e - u_{k+1}\|^2 = \|P_C(u_k - t_k\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)) - P_C(u_k - t_k d_k)\|^2 \leq t_k^2\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2. \tag{4.48}$$

Let ℓ_{ξ_k} be the Lipschitz constant of $\nabla f_{\lambda_k,\xi_k}$. Then, it follows from the Descent Lemma [4, Lemma 2.64] that

$$\begin{aligned} f_{\lambda_k,\xi_k}(u_{k+1}^e) - f_{\lambda_k,\xi_k}(u_{k+1}) &\leq \left\langle \nabla f_{\lambda_k,\xi_k}(u_{k+1}) \mid u_{k+1}^e - u_{k+1}\right\rangle + \frac{\ell_{\xi_k}}{2}\|u_{k+1}^e - u_{k+1}\|^2 \\ &\leq \|\nabla f_{\lambda_k,\xi_k}(u_{k+1})\|\|u_{k+1}^e - u_{k+1}\| + \frac{\ell_{\xi_k}}{2}\|u_{k+1}^e - u_{k+1}\|^2 \\ &\leq \Big(\|\nabla f_{\lambda_k,\xi_k}(u_k)\| + \|\nabla f_{\lambda_k,\xi_k}(u_{k+1}) - \nabla f_{\lambda_k,\xi_k}(u_k)\|\Big)\|u_{k+1}^e - u_{k+1}\| \\ &\quad + \frac{\ell_{\xi_k}}{2}\|u_{k+1}^e - u_{k+1}\|^2. \end{aligned} \tag{4.49}$$

Since $\|\nabla f_{\lambda_k,\xi_k}(u_{k+1}) - \nabla f_{\lambda_k,\xi_k}(u_k)\| \leq \ell_{\xi_k}\|u_{k+1} - u_k\|$, the RHS of (4.49) verifies

$$\begin{aligned}
&\Big(\|\nabla f_{\lambda_k,\xi_k}(u_k)\| + \|\nabla f_{\lambda_k,\xi_k}(u_{k+1}) - \nabla f_{\lambda_k,\xi_k}(u_k)\|\Big)\|u^e_{k+1} - u_{k+1}\| + \frac{\ell_{\xi_k}}{2}\|u^e_{k+1} - u_{k+1}\|^2 \\
&\qquad \leq (1 + t_k\ell_{\xi_k})\|d_k\|\|u^e_{k+1} - u_{k+1}\| + \frac{\ell_{\xi_k}}{2}\|u^e_{k+1} - u_{k+1}\|^2 \\
&\qquad \leq t_k(1 + t_k\ell_{\xi_k})\|d_k\|\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\| + \frac{\ell_{\xi_k}}{2}t_k^2\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2 \\
&\qquad \leq \frac{1}{2}t_k^2(1 + t_k\ell_{\xi_k})^2\|d_k\|^2 + \frac{1}{2}(1 + t_k^2\ell_{\xi_k})\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2. \qquad (4.50)
\end{aligned}$$

Combining (4.50) with (4.47), we deduce

$$\begin{aligned}
f_{\lambda_k,\xi_k}(u^e_{k+1}) &< \mathcal{L}_{\rho_k,\xi_k}(u_k,\lambda_{k,\xi_k}) - t_k\big(\nu\beta_k - \frac{1}{2}t_k(1 + t_k\ell_{\xi_k})^2\big)\|d_k\|^2 \\
&\quad + \frac{1}{2}(1 + t_k^2\ell_{\xi_k})\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2 + \mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big). \qquad (4.51)
\end{aligned}$$

Let ℓ_k be the Lipschitz constant of $\nabla\mathcal{L}_{\rho_k}(\cdot,\lambda_k)$. Then, it follows from the Descent Lemma [4, Lemma 2.64] that

$$\mathcal{L}_{\rho_k}(u_{k+1},\lambda_k) - \mathcal{L}_{\rho_k}(u^e_{k+1},\lambda_k) \leq \big\langle \nabla\mathcal{L}_{\rho_k}(u^e_{k+1},\lambda_k) \mid u_{k+1} - u^e_{k+1}\big\rangle + \frac{\ell_k}{2}\|u^e_{k+1} - u_{k+1}\|^2 \qquad (4.52)$$

The RHS of (4.52) verifies

$$\begin{aligned}
&\big\langle \nabla\mathcal{L}_{\rho_k}(u^e_{k+1},\lambda_k) \mid u_{k+1} - u^e_{k+1}\big\rangle + \frac{\ell_k}{2}\|u^e_{k+1} - u_{k+1}\|^2 \\
&\quad \leq \|\nabla\mathcal{L}_{\rho_k}(u^e_{k+1},\lambda_k)\|\|u^e_{k+1} - u_{k+1}\| + \frac{\ell_k}{2}\|u^e_{k+1} - u_{k+1}\|^2 \\
&\quad \leq \Big(\|\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\| + \|\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k) - \nabla\mathcal{L}_{\rho_k}(u^e_{k+1},\lambda_k)\|\Big)\|u^e_{k+1} - u_{k+1}\| + \frac{\ell_k}{2}\|u^e_{k+1} - u_{k+1}\|^2 \\
&\quad \leq \Big(\|\mathbf{E}_{\xi_k}[d_k]\| + \ell_k\|u_k - u^e_{k+1}\|\Big)\|u^e_{k+1} - u_{k+1}\| + \frac{\ell_k}{2}\|u^e_{k+1} - u_{k+1}\|^2
\end{aligned}$$

Using (4.48), we obtain

$$\begin{aligned}
&\Big(\|\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\| + \ell_k\|u_k - u^e_{k+1}\|\Big)\|u^e_{k+1} - u_{k+1}\| + \frac{\ell_k}{2}\|u^e_{k+1} - u_{k+1}\|^2 \\
&\quad \leq t_k\Big(\|\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\| + t_k\ell_k\|\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|\Big)\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\| + \frac{\ell_k}{2}t_k^2\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2 \\
&\quad \leq t_k\|\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|\big(1 + t_k\ell_k\big)\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\| + \frac{\ell_k}{2}t_k^2\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2 \\
&\quad \leq \frac{1}{2}t_k^2\big(1 + t_k\ell_k\big)^2\|\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2 + \frac{1}{2}(1 + t_k^2\ell_k)\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2 \\
&\quad \leq t_k^2\big(1 + t_k\ell_k\big)^2\|d_k\|^2 + \frac{3}{2}(1 + t_k^2\ell_k)\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2 \qquad (4.53)
\end{aligned}$$

Combining (4.51) and (4.53), we deduce

$$\begin{aligned}
&\mathcal{L}_{\rho_k}(u_{k+1},\lambda_k) - \mathcal{L}_{\rho_k}(u^e_{k+1},\lambda_k) + f_{\lambda_k,\xi_k}(u^e_{k+1}) \\
&\quad \leq \mathcal{L}_{\rho_k,\xi_k}(u_k,\lambda_{k,\xi_k}) - t_k\big[\big(\nu\beta_k - \frac{1}{2}t_k(1 + t_k\ell_{\xi_k})^2\big)\|d_k\|^2 - t_k\big(1 + t_k\ell_k\big)^2\|d_k\|^2\big] \\
&\qquad + \frac{1}{2}(1 + t_k^2\ell_{\xi_k})\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2 + \frac{3}{2}(1 + t_k^2\ell_k)\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2 + \mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big).
\end{aligned} \qquad (4.54)$$

Taking the conditional expectation with respect to ξ_k, using Lemma 8, we derive from (4.54) that

$$\begin{aligned}\mathbf{E}_{\xi_k}[\mathcal{L}_{\rho_{k+1}}(u_{k+1},\lambda_{k+1})] \leq \mathcal{L}_{\rho_k}(u_k,\lambda_k) - \mathbf{E}_{\xi_k}\Big[t_k\Big(\nu\beta_k - \frac{1}{2}t_k(1+t_k\ell_{\xi_k})^2 - t_k\big(1+t_k\ell_k\big)^2\Big)\|d_k\|^2\Big] \\ + \mathbf{E}_{\xi_k}\Big[\frac{1}{2}(1+t_k^2\ell_{\xi_k})\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2 + \frac{3}{2}(1+t_k^2\ell_k)\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2 \\ + \frac{1+\varepsilon}{\sigma_k}\|\|\lambda_{k+1}-\lambda_k\|\|^2\Big] + \mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big). \end{aligned} \tag{4.55}$$

Then, by Lemma 11,

$$\begin{aligned}\mathbf{E}_{\xi_k}[\mathcal{L}_{\rho_{k+1}}(u_{k+1},\lambda_{k+1})] \leq \mathcal{L}_{\rho_k}(u_k,\lambda_k) - \mathbf{E}_{\xi_k}\Big[t_k\Big(\nu\beta_k - \frac{1}{2}t_k(1+t_k\ell_{\xi_k})^2 - t_k\big(1+t_k\ell_k\big)^2\Big)\|d_k\|^2\Big] \\ + \mathbf{E}_{\xi_k}\Big[\frac{1}{2}(1+t_k^2\ell_{\xi_k})\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2\Big] + \mathbf{E}_{\xi_k}\Big[\frac{3}{2}(1+t_k^2\ell_k)\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2\Big] \\ + \mathbf{E}_{\xi_k}\big[\varsigma_{1,k}t_k\|d_k\|^2\big] + 2(1+\varepsilon)t_{k-1}^2\|d_{k-1}\|^2 + \mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big). \end{aligned} \tag{4.56}$$

Next, using Assumption 1, $(1+t_k^2\ell_k) \leq t_{k,\max}$ and Assumption 2, we get for the third term of the RHS of (4.56)

$$\begin{aligned}\mathbf{E}_{\xi_k}\Big[\frac{1}{2}(1+t_k^2\ell_{\xi_k})\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2\Big] &\leq \frac{1}{2}t_{k,\max}\mathbf{E}_{\xi_k}\big[\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2\big] \\ &\leq \frac{1}{2}\frac{t_{k,\max}\overline{\sigma}_k^2}{\mathtt{m}_k} \end{aligned} \tag{4.57}$$

and in the same manner, for the fourth term of the RHS of (4.56)

$$\begin{aligned}\mathbf{E}_{\xi_k}\Big[\frac{3}{2}(1+t_k^2\ell_k)\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2\Big] &\leq \frac{3}{2}t_{k,\max}\mathbf{E}_{\xi_k}\big[\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2\big] \\ &\leq \frac{3}{2}\frac{t_{k,\max}\overline{\sigma}_k^2}{\mathtt{m}_k}. \end{aligned} \tag{4.58}$$

Adding (4.57) and (4.58), we obtain the following

$$\mathbf{E}_{\xi_k}\Big[\frac{1}{2}(1+t_k^2\ell_{\xi_k})\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2\Big] + \mathbf{E}_{\xi_k}\Big[\frac{3}{2}(1+t_k^2\ell_k)\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2\Big] \leq 2\frac{t_{k,\max}\overline{\sigma}_k^2}{\mathtt{m}_k}. \tag{4.59}$$

Consequently, in order to satisfy (4.42), the following inequality must be verified

$$2\frac{t_{k,\max}\overline{\sigma}_k^2}{\mathtt{m}_k} \leq \mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big). \tag{4.60}$$

Then, using the LHS of (4.59), we can simplify the RHS of (4.56) as

$$\begin{aligned}\mathbf{E}_{\xi_k}\big[\mathcal{L}_{\rho_{k+1}}(u_{k+1},\lambda_{k+1})\big] \leq \mathcal{L}_{\rho_k}(u_k,\lambda_k) - \mathbf{E}_{\xi_k}\Big[t_k\big(\nu\beta_k - \frac{1}{2}t_k(1+t_k\ell_{\xi_k})^2 - t_k\big(1+t_k\ell_k\big)^2\big)\|d_k\|^2\Big] \\ + \mathbf{E}_{\xi_k}\big[\varsigma_{1,k}t_k\|d_k\|^2\big] + 2(1+\varepsilon)t_{k-1}^2\|d_{k-1}\|^2 + \mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big). \end{aligned} \tag{4.61}$$

Adding $\mathbf{E}_{\xi_k}\big[2(1+\varepsilon)t_k^2\|d_k\|^2\big]$ to both sides of (4.61), we get

$$\begin{aligned}&\mathbf{E}_{\xi_k}\Big[\mathcal{L}_{\rho_{k+1}}(u_{k+1},\lambda_{k+1}) + 2(1+\varepsilon)t_k^2\|d_k\|^2\Big] \\ &\leq \mathcal{L}_{\rho_k}(u_k,\lambda_k) - \mathbf{E}_{\xi_k}\Big[t_k\big(\nu\beta_k - \frac{1}{2}t_k(1+t_k\ell_{\xi_k})^2 - t_k\big(1+t_k\ell_k\big)^2\big)\|d_k\|^2\Big] \\ &\quad + \mathbf{E}_{\xi_k}\Big[\varsigma_{1,k}t_k\|d_k\|^2 + 2(1+\varepsilon)t_k^2\|d_k\|^2\Big] + 2(1+\varepsilon)t_{k-1}^2\|d_{k-1}\|^2 + \mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big). \end{aligned} \tag{4.62}$$

We can further obtain from (4.62) that

$$\begin{aligned}
&\mathbf{E}_{\xi_k}\Big[\mathcal{L}_{\rho_{k+1}}(u_{k+1},\lambda_{k+1})+2(1+\varepsilon)t_k^2\|d_k\|^2\Big]\\
&\quad\leq \mathcal{L}_{\rho_k}(u_k,\lambda_k)+2(1+\varepsilon)t_{k-1}^2\|d_{k-1}\|^2\\
&\qquad-\mathbf{E}_{\xi_k}\Big[t_k\big(\nu\beta_k-\frac{1}{2}t_k(1+t_k\ell_{\xi_k})^2-t_k(1+t_k\ell_k)^2-2(1+\varepsilon)t_k-\varsigma_{1,k}\big)\|d_k\|^2\Big]+\mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big)\\
&\quad\leq \mathcal{L}_{\rho_k}(u_k,\lambda_k)+2(1+\varepsilon)t_{k-1}^2\|d_{k-1}\|^2-\frac{\varepsilon}{2}\mathbf{E}_{\xi_k}\big[t_k\|d_k\|^2\big]+\mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big).
\end{aligned}\tag{4.63}$$

Since the sequence $(\mathcal{L}_{\rho_k}(u_k,\lambda_k))_{k\in\mathbb{N}}$ is bounded below, by adding $-\inf_{k\in\mathbb{N}}\mathcal{L}_{\rho_k}(u_k,\lambda_k)$ to both sides of (4.63), we derive from the Lemma 6 that

$$\sum_{k\in\mathbb{N}}\mathbf{E}_{\xi_k}[t_k\|d_k\|^2]<+\infty.\tag{4.64}$$

As a consequence, by Corollary 1, we obtain

$$\sum_{k\in\mathbb{N}}t_k\|d_k\|^2<+\infty\ \mathbf{a.s.}\tag{4.65}$$

(i): This conclusion follows directly from (4.65) and $\|\frac{u_{k+1}-u_k}{\sqrt{t_k}}\|\leq\sqrt{t_k}\|d_k\|$.

(ii): Note that under the conditions (3.7), $\varsigma_{1,k}<1$ and $t_{k-1}\leq 1$ a.s. Hence, we can derive from Lemma 11 that

$$\begin{aligned}
\sum_{k\in\mathbb{N}}\mathbf{E}_{\xi_k}\Big[\frac{1+\varepsilon}{\sigma_k}\|\lambda_{k+1}-\lambda_k\|^2\Big]&\leq\sum_{k\in\mathbb{N}}\mathbf{E}_{\xi_k}[\varsigma_{1,k}t_k\|d_k\|^2]+\sum_{k\in\mathbb{N}}2(1+\varepsilon)t_{k-1}^2\|d_{k-1}\|^2\\
&\leq\sum_{k\in\mathbb{N}}\mathbf{E}_{\xi_k}\big[t_k\|d_k\|^2+2(1+\varepsilon)t_{k-1}^2\|d_{k-1}\|^2\big]\\
&<+\infty,
\end{aligned}\tag{4.66}$$

where the last inequality follows from (i) and (4.63). Therefore, the conclusion follows from (4.66) and the update rule of λ_{k+1}.

(iii): We have

$$\begin{aligned}
\mathbf{E}_{\xi_k}\Big[\|\frac{u_{k+1}^e-u_k}{\sqrt{t_k}}\|^2\Big]&\leq 2\mathbf{E}_{\xi_k}\Big[\|\frac{u_{k+1}^e-u_{k+1}}{\sqrt{t_k}}\|^2\Big]+2\mathbf{E}_{\xi_k}\Big[\|\frac{u_{k+1}-u_k}{\sqrt{t_k}}\|^2\Big]\\
&\leq 2\mathbf{E}_{\xi_k}[t_k\|d_k-\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2]+2\mathbf{E}_{\xi_k}\Big[\|\frac{u_{k+1}-u_k}{\sqrt{t_k}}\|^2\Big]
\end{aligned}\tag{4.67}$$

Since $t_k\in]0,1[$, the first term of the RHS of (4.67) verifies $\mathbf{E}_{\xi_k}[t_k\|d_k-\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2]\leq\mathbf{E}_{\xi_k}[\|d_k-\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2]$; hence,

$$\begin{aligned}
&2\mathbf{E}_{\xi_k}[t_k\|d_k-\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2]+2\mathbf{E}_{\xi_k}\Big[\|\frac{u_{k+1}-u_k}{\sqrt{t_k}}\|^2\Big]\\
&\qquad\leq 2\mathbf{E}_{\xi_k}\Big[\|d_k-\nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2\Big]+2\mathbf{E}_{\xi_k}\Big[\|\frac{u_{k+1}-u_k}{\sqrt{t_k}}\|^2\Big]\\
&\qquad\leq 2\frac{\overline{\sigma}_k^2}{\mathtt{m}_k}+2\mathbf{E}_{\xi_k}\Big[\|\frac{u_{k+1}-u_k}{\sqrt{t_k}}\|^2\Big]\\
&\qquad=2\mathbf{E}_{\xi_k}\Big[\|\frac{u_{k+1}-u_k}{\sqrt{t_k}}\|^2\Big]+\mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big),
\end{aligned}\tag{4.68}$$

which implies that the sequence $\Big(\mathbf{E}\Big[\|\frac{u^e_{k+1}-u_k}{\sqrt{t_k}}\|^2\Big]\Big)_{k\in\mathbb{N}}$ is summable.

(iv): Let us set

$$\begin{cases} r^e_{k+1} = c(u^e_{k+1}) - b - P_S(c(u^e_{k+1}) - b + \rho_k^{-1}\lambda_k) \\ r_{k+1} = c(u_{k+1}) - b - P_S(c(u_{k+1}) - b + \rho_k^{-1}\lambda_k). \end{cases}$$

Since the projection operator P_S is nonexpansive,

$$\begin{aligned} \sigma_k|||r_{k+1} - r^e_{k+1}|||^2 &\leq 2\sigma_k|||c(u^e_{k+1}) - c(u_{k+1})|||^2 \\ &\leq 4\sigma_k|||J_c(u^e_{k+1})(u_{k+1} - u^e_{k+1})|||^2 + \sigma_k\mu_c^2\|u_{k+1} - u^e_{k+1}\|^4. \end{aligned} \tag{4.69}$$

Then, using Assumption 1 & 2, and $t_k \leq 1$, the RHS of (4.69) verifies

$$\begin{aligned} &\mathbf{E}_{\xi_k}\Big[4\sigma_k|||J_c(u^e_{k+1})(u_{k+1} - u^e_{k+1})|||^2 + \sigma_k\mu_c^2\|u_{k+1} - u^e_{k+1}\|^4\Big] \qquad (4.70) \\ &\qquad \leq \mathbf{E}_{\xi_k}\Big[4\sigma_k\mu_0^2\|u_{k+1} - u^e_{k+1}\|^2 + \sigma_k\mu_c^2\|u_{k+1} - u^e_{k+1}\|^4\Big] \\ &\qquad \leq \mathbf{E}_{\xi_k}\Big[4\sigma_k\mu_0^2t_k^2\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2 + \sigma_k\mu_c^2t_k^4\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^4\Big] \\ &\qquad \leq \mathbf{E}_{\xi_k}\Big[4\mu_0^2\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^2 + \mu_c^2\|d_k - \nabla\mathcal{L}_{\rho_k}(u_k,\lambda_k)\|^4\Big]\sup_{k\in\mathbb{N}}(\sigma_k) \\ &\qquad \leq \Big(4\mu_0^2\frac{\overline{\sigma}_k^2}{\mathtt{m}_k} + \mu_c^2(\frac{\overline{\sigma}_k^2}{\mathtt{m}_k})^2\Big)\sup_{k\in\mathbb{N}}(\sigma_k) \\ &\qquad = \mathcal{O}\Big(\frac{1}{(1+k)^{1+\varepsilon}}\Big). \qquad (4.71) \end{aligned}$$

Therefore, the sequence $(\mathbf{E}_{\xi_k}\big[4\sigma_k|||J_c(u^e_{k+1})(u_{k+1}-u^e_{k+1})|||^2 + \sigma_k\mu_c^2\|u_{k+1}-u^e_{k+1}\|^4)_{k\in\mathbb{N}}$ is summable. By (4.69), the sequence $(\sigma_k|||r_{k+1} - r^e_{k+1}|||^2)_{k\in\mathbb{N}}$ is also summable. Hence, in view of (ii) and

$$|||r^e_{k+1}|||^2 \leq 2|||r_{k+1} - r^e_{k+1}|||^2 + |||r_{k+1}|||^2,$$

It follows that the sequence $(|||r^e_{k+1}|||^2)_{k\in\mathbb{N}}$ is summable, and thus the result is proved. □

Corollary 2 *Under the same conditions stated in Theorem 3. The followings hold almost surely,*

(i) *The sequence* $(\|\frac{u_{k+1}-u_k}{\sqrt{t_k}}\|)_{k\in\mathbb{N}}$ *is square summable.*

(ii) *The sequence* $\Big(\sqrt{\sigma_k}|||c(u_{k+1}) - b - P_S(c_2(u_{k+1}) - b + \rho_k^{-1}\lambda_k)|||\Big)_{k\in\mathbb{N}}$ *is square summable.*

(iii) *The sequence* $(\|\frac{u^e_{k+1}-u_k}{\sqrt{t_k}}\|)_{k\in\mathbb{N}}$ *is square summable.*

(iv) *The sequence* $\Big(\sqrt{\sigma_k}|||c(u^e_{k+1}) - b - P_S(c_2(u^e_{k+1}) - b + \rho_k^{-1}\lambda_k)|||\Big)_{k\in\mathbb{N}}$ *is square summable.*

Proof. The results follow from Theorem 3 as well as Corollary 1. □

4.3 Convergence to KKT points

In this section, we determine the conditions for the local convergence of the sequences $(u_k,\lambda_k)_{k\in\mathbb{N}}$ produced by the Algorithm 2 to a critical point of the augmented Lagrangian function $\mathcal{L}_\rho$ defined by (2.22).

Let us first recall the first-order KKT conditions for the constrained optimization problem at hand. If $N_C(u^\dagger)$ defines the normal cone of the set C at the point $u^\dagger$, a local minimum of Problem 1, that satisfies the regularity conditions stated here below; then, there exists a vector $\lambda \in \mathbb{R}^m$, where $m = \sum_{q=1}^M m_q$, such that the following conditions hold:

$$\begin{cases} \text{Stationarity: } -\left(\nabla h(u^\dagger) + \sum_{i=1}^{m} \lambda_i J_{c_i}(u^\dagger)\right) \in N_C(u^\dagger), \\ \text{Primal feasibility: } c(u^\dagger) - b \in S, \\ \text{Dual feasibility: } \lambda \in S^{\ominus}{}^{1}, \\ \text{Complementary slackness: } \langle \lambda \mid c(u^\dagger) - b \rangle = 0. \end{cases} \tag{4.72}$$

Throughout this paper, the set $\mathcal{K}$ of KKT points is non-empty. The first-order KKT conditions hold if some regularity conditions, called constraint qualification (CQ) conditions, are satisfied by feasible points. The constraint qualification of c on $Z \subseteq C$, a non-empty closed convex subset of $\mathcal{H}$, can be stated as follows: there exists a strictly positive constant $\zeta \in]0, +\infty[$ such that for all $v \in Y = c(Z) - b$ of $\mathbb{R}^m$, the following inequality is verified for all $u \in Z$

$$\zeta \|v\| \leq \|J_c(u)^\intercal v\|. \tag{4.73}$$

In nonlinear programming, see, e.g., [12] [39], the uniform regularity condition (4.73) is equivalent to the well-known Mangasarian-Fromovitz Constraint Qualification (MFCQ) of c on Z. Let $u^\dagger$ be a local minimizer for Problem 1. The MFCQ conditions holding at $u^\dagger$ guarantee the existence[2] and the boundedness -but not necessarily the uniqueness- of KKT multipliers (λ) at $u^\dagger$.

Throughout this paper, in addition to the lower boundedness of $\mathcal{L}_{\rho_k}(u_k, \lambda_k)$ for all $k \in \mathbb{N}$, the following conditions and properties are assumed to be verified. Let $(u_k)_{k\in\mathbb{N}} \subset Z \subseteq C$.

P1 Constraint c verifies the MFCQ conditions on Z with constant $\zeta \in]0, +\infty[$;
P2 The sequence $(\rho_k)_{k\in\mathbb{N}}$ is bounded from above.
P3 The primal sequence $(u_k)_{k\in\mathbb{N}}$ generated by Algorithm 2 is bounded.

The reasoning developed is to first demonstrate by means of Proposition 1 that the limit points of the subsequences produced by Algorithm 2 verify the first-order KKT conditions (4.72). The next step consists of proving that the set of limits points is non-empty (cf. Proposition 2). Knowing this property, the last step then requires to prove that, under certain conditions, the sequences produced by the algorithm converge to such limit point (cf. Corollary 3).

Proposition 1 *Assume that the conditions stated for Theorem 3 hold. Suppose, according to property **P2**, that $(\rho_k)_{k\in\mathbb{N}}$ is bounded from above. Let $((u_{n_k}, \lambda_{n_k}))_{k\in\mathbb{N}}$ be a subsequence of $((u_k, \lambda_k))_{k\in\mathbb{N}}$ such that*

$$\begin{cases} (u_{n_k}, \lambda_{n_k}) \to (u^\dagger, \lambda^\dagger), \\ (u_{n_k+1} - u_{n_k})/t_{n_k} \to 0, \\ c(u_{n_k+1}) - b - P_S(c(u_{n_k+1}) - b + \rho_{n_k}^{-1}\lambda_{n_k}) \to 0. \end{cases} \tag{4.74}$$

Then, the limit point $(u^\dagger, \lambda^\dagger)$ verifies the KKT conditions (4.72).

Proof. (i). Primal feasibility. Since $t_k\ (= \theta^j, j \in \mathbb{N}) \leq 1$,

$$\|u_{n_k+1} - u_{n_k}\| \leq \|(u_{n_k+1} - u_{n_k})/t_{n_k}\|. \tag{4.75}$$

Hence, $u_{k+1} - u_k \to 0$ implies that

$$u_{n_k+1} \to u^\dagger. \tag{4.76}$$

Since P_S and c are continuous, it follows that

$$\lim_{k\to\infty} P_S(c(u_{n_k+1}) - b) = P_S(c(u^\dagger) - b). \tag{4.77}$$

By assumption

$$d_S(c(u_{n_k+1}) - b) \leq \|\|c(u_{n_k+1}) - b - P_S(c(u_{n_k+1}) - b + \rho_{n_k}^{-1}\lambda_{n_k})\|\| \to 0;$$

[1] where $S^\ominus$ refers to the polar cone of S, see infra for its definition.
[2] The set of KKT multipliers (λ) at $u^\dagger$ is nonempty.

hence, it follows that

$$c(u^\dagger) - b = P_S(c(u^\dagger) - b) \in S. \tag{4.78}$$

ii) Dual feasibility and Complementarity slackness: consider the negative of the dual cone of S[3], i.e., the polar cone $S^\ominus$ of S defined as $S^\ominus = \{u \mid \sup \langle S \mid u\rangle \leq 0\}$. Then, by [4, Theorem 6.30], we have

$$(\forall a \in \mathbb{R}^m) \quad a = P_S a + P_{S^\ominus} a, \tag{4.79}$$

In turn, by using [4, Proposition 29(ii)], we obtain for the constraints $c(u) - b \in S$, and $a_{k+1} = c(u_{k+1}) - b$ that

$$\begin{aligned}
\rho_k\Big(a_{k+1} - P_S\big(a_{k+1} + \rho_k^{-1}\lambda_k\big)\Big) &= \rho_k\Big(P_{S^\ominus}\big(a_{k+1} + \rho_k^{-1}\lambda_k\big) - \rho_k^{-1}\lambda_k\Big) \\
&= \rho_k P_{S^\ominus}\big((\rho_k a_{k+1} + \lambda_k)/\rho_k\big) - \lambda_k \\
&= \rho_k P_{S^\ominus/\rho_k}\big((\rho_k a_{k+1} + \lambda_k)/\rho_k\big) - \lambda_k \\
&= P_{S^\ominus}\big(\rho_k a_{k+1} + \lambda_k\big) - \lambda_k.
\end{aligned} \tag{4.80}$$

The latter equality implies that for all k, the following identity is verified

$$\lambda_k = P_{S^\ominus}\big(\rho_k(c(u_{k+1}) - b) + \lambda_k\big) - \rho_k\big(c(u_{k+1}) - b - P_S(c(u_{k+1}) - b + \rho_k^{-1}\lambda_k)\big). \tag{4.81}$$

Following property **P2**, the sequence $(\rho_k)_{k\in\mathbb{N}}$ is bounded; hence, we derive from (4.81) that

$$\lambda^\dagger = P_{S^\ominus}\big(\rho^\dagger(c(u^\dagger) - b) + \lambda^\dagger\big) \in S^\ominus, \tag{4.82}$$

where $\rho^\dagger$ is a cluster point of $(\rho_{n_k})_{k\in\mathbb{N}}$. Moreover, since $\lambda^\dagger/2 \in S^\ominus$, by the Projection theorem [4, Theorem 3.16], we get

$$\big\langle c(u^\dagger) - b \,||\, \lambda^\dagger\big\rangle = 0. \tag{4.83}$$

iii) Stationarity: We also have

$$\begin{aligned}
v_{n_k} &:= \lambda_{n_k} + \rho_{n_k}\big(c(u_{n_k}) - b - P_S(c(u_{n_k}) - b + \rho_{n_k}^{-1}\lambda_{n_k})\big) \to \lambda^\dagger \\
\nabla\mathcal{L}_{\rho_k}(u_k, \lambda_k) &= \big(\nabla h(u_{n_k}) + J_c(u_{n_k})^\intercal v_{n_k}\big) \to d^\dagger = \big(\nabla h(u^\dagger) + J_c(u^\dagger)^\intercal\lambda^\dagger\big).
\end{aligned} \tag{4.84}$$

Next, we deduce from the definition of u^e_{k+1} that

$$(u_{n_k} - u^e_{n_k+1})/t_{n_k} + \nabla\mathcal{L}_{\rho_k}(u_k, \lambda_k) \in N_C(u^e_{n_k+1}). \tag{4.85}$$

By Corollary 2(iii), $(u_{n_k} - u^e_{n_k+1})/t_{n_k} \to 0$ and $u^e_{n_k+1} \to u^\dagger$. Hence, it follows from (4.84) that

$$d^\dagger \in N_C(u^\dagger). \tag{4.86}$$

The expressions (4.78), (4.82), (4.83) and (4.86) are exactly the first-order KKT conditions (4.72). Consequently, the limit point $(u^\dagger, \lambda^\dagger) \in \mathcal{K}$. □

Note that when the set C is bounded, the primal sequence is bounded. In the general case, we have the following result.

Proposition 2 *Assume that the conditions stated for Theorem 3 hold, the sequence $(u_k)_{k\in\mathbb{N}}$ is bounded (property **P3**), and the sequence $(\rho_k)_{k\in\mathbb{N}}$ is bounded from above (property **P2**). We further suppose that the subsequences $(d_{n_k})_{k\in\mathbb{N}}$ is bounded. Then, the subsequence $(u_{n_k}, \lambda_{n_k})_{k\in\mathbb{N}}$ is bounded. Consequently, the set of cluster points of $(u_k, \lambda_k)_{k\in\mathbb{N}}$ is non-empty.*

[3] The dual cone $S^\star$ is always convex irrespective of the original set

Proof. It follows from our assumption that there exists a constant $\mathcal{O}(1)$ such that $(\forall k \in \mathbb{N})\ \|d_{n_k}\| \leq \mathcal{O}(1)$. Hence,

$$(\forall k \in \mathbb{N})\ \|\nabla \mathcal{L}_{\rho_{n_k}}(u_{n_k}, \lambda_{n_k})\| = \|\mathbf{E}_{\xi_{n_k}}[d_{n_k}]\| \leq \mathbf{E}_{\xi_{n_k}}[\|d_{n_k}\|] \leq \mathcal{O}(1). \tag{4.87}$$

Note that

$$\begin{aligned}\nabla \mathcal{L}_{\rho_k}(u_k, \lambda_k) &= \nabla h(u_{n_k}) + J_c(u_{n_k})^\intercal v_{n_k} \\ &\text{with } v_{n_k} = \lambda_{n_k} + \rho_{n_k}\big(c(u_{n_k}) - b - P_S(c(u_{n_k}) - b + \rho_{n_k}^{-1}\lambda_{n_k})\big),\end{aligned} \tag{4.88}$$

which implies that

$$\|J_c(u_{n_k})^\intercal v_{n_k}\| \leq \|\nabla \mathcal{L}_{\rho_{n_k}}(u_{n_k}, \lambda_{n_k})\| + \|\nabla h(u_{n_k})\|. \tag{4.89}$$

Since the primal sequence $(u_k)_{k\in\mathbb{N}}$ is bounded and since the function h is differentiable with μ_h-Lipschitz continuous gradient ∇h, it follows from (4.87) and (4.89) that

$$\|J_c(u_{n_k})^\intercal v_{n_k}\| \leq \mathcal{O}(1). \tag{4.90}$$

Now, using the Mangasarian-Fromowitz (MF) condition of c on Z (property **P1**), there exists a strictly positive constant ζ_c such that

$$\zeta_c \||(v_{n_k})|\| \leq \|J_c(u_{n_k})^\intercal v_{n_k}\|. \tag{4.91}$$

Hence, $(v_{n_k})_{k\in\mathbb{N}}$ defines a sequence bounded by

$$\zeta_c \||(v_{n_k})|\| \leq \|J_c(u_{n_k})^\intercal v_{n_k}\| \leq \mathcal{O}(1).$$

Since $(\rho_k)_{k\in\mathbb{N}}$ is bounded and $(u_k)_{k\in\mathbb{N}}$ is bounded, By Step 1 of Algorithm 2, the sequence $\Big(\rho_{n_k}\big(c(u_{n_k}) - b - P_S(c(u_{n_k}) - b + \rho_{n_k}^{-1}\lambda_{n_k})\big)\Big)_{k\in\mathbb{N}}$ is bounded. In turn, by the definition of v_{n_k} defined by (4.88), the sequence $(\lambda_{n_k})_{k\in\mathbb{N}}$ is also bounded. Consequently, the set of cluster points of $(u_k, \lambda_k)_{k\in\mathbb{N}}$ is non-empty. □

Corollary 3 *Assume the conditions stated for Theorem 3 hold. Suppose, the properties **P1**, **P2**, and **P3** are satisfied. If $\sum_{k\in\mathbb{N}} t_k = \infty$, and $\|d_k\| = \mathcal{O}(1)$, and $k\rho_k \to \infty$. Then, there exists a subsequence $(u_{n_k}, \lambda_{n_k})_{k\in\mathbb{N}}$ of $(u_k, \lambda_k)_{k\in\mathbb{N}}$ that converges to a limit point $(u^\dagger, \lambda^\dagger) \in \mathcal{K}$.*

Proof. Since $\sum_{k\in\mathbb{N}} t_k = \infty$, and by Theorem 3, $\sum_{k\in\mathbb{N}} t_k \|d_k\|^2 < +\infty$, we get $\inf_{k\in\mathbb{N}} \|d_k\| = 0$. Then there exists a subsequence $(d_{p_k})_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} d_{p_k} = 0$. By Proposition 2, there exists a subsequence $(u_{n_k}, \lambda_{n_k})_{k\in\mathbb{N}}$ of $(u_{p_k}, \lambda_{p_k})_{k\in\mathbb{N}}$ of such that $(u_{n_k}, \lambda_{n_k}) \to (u^\dagger, \lambda^\dagger)$. We first have

$$\|\frac{u_{n_k+1} - u_{n_k}}{t_{n_k}}\| \leq \|d_{n_k}\| \to 0. \tag{4.92}$$

Moreover, it follows Step 1 of Algorithm 2 that

$$\begin{aligned}\rho_k \|| c(u_k) - b - P_S\big(c(u_k) - b + \rho_k^{-1}\lambda_k\big)|\|^2 &\leq \min_{1\leq i\leq k} \|u_i - u_{i-1}\|^2 \\ &\leq \frac{1}{k}\sum_{i=1}^{k} \|u_i - u_{i-1}\|^2 \\ &\leq \frac{1}{k}\sum_{i=1}^{k} t_i^2 \|d_i\|^2 \\ &\leq \mathcal{O}(1/k),\end{aligned} \tag{4.93}$$

where the last estimation follows from Corollary 2. Since $k\rho_k \to \infty$, we get

$$\|| c(u_k) - b - P_S\big(c(u_k) - b + \rho_k^{-1}\lambda_k\big)|\| \to 0;$$

and thus,

$$\|| c(u_{k+1}) - b - P_S\big(c(u_{k+1}) - b + \rho_k^{-1}\lambda_k\big)|\| \to 0. \tag{4.94}$$

Hence, in view of Proposition 1, the limit point $(u^\dagger, \lambda^\dagger) \in \mathcal{K}$. □

5 Iteration Complexity

In this Section, we characterize the iteration complexity of the proposed Algorithm in terms of the difference $\Delta\mathcal{L}_k(\cdot,\lambda_k)$ and the feasibility. By iteration complexity, we refer here to the number of iterations required to obtain an approximate ε-KKT point of Problem 1 by means of the proposed algorithm (cf. Section 3.2).

Theorem 4 *Let $\mu_{c_{i_p}}$ and $\mu_{h_{i_p}}$ be, respectively, the Lipschitz constant of $J_{c_{i_p}}$ and ∇h_{i_p}. Set*

$$\mu_{c,\xi_k} = \frac{1}{2m_k}\sum_{p=1}^{m_k}\mu_{c_{i_p}} \text{ and } \mu_{h,\xi_k} = \frac{1}{2m_k}\sum_{p=1}^{m_k}\mu_{h_{i_p}} \tag{5.1}$$

Suppose that $C = \mathcal{H}\times\mathcal{G}$. Assume that the conditions stated in Theorem 3 are satisfied. Then, t_k and u_k verify the following

$$t_k \geq \frac{2\theta(1-\nu)\beta_{\xi_k}}{\beta_{\xi_k}\mu_{c,\xi_k}\rho_k + \mu_{h,\xi_k}} \text{ and } \min_{0\leq i\leq k}\frac{2\theta(1-\nu)\beta_{\xi_i}}{\beta_{\xi_i}\mu_{c,\xi_i}\rho_i + \mu_{h,\xi_i}}\|\frac{u_{i+1}-u_i}{t_i}\|^2 = \mathcal{O}(1/(k+1)), \tag{5.2}$$

where the constant of $\mathcal{O}$ is a random variable which is independent of k, and

$$\beta_{\xi_k} \coloneqq \max_{0\leq t\leq 1, 1\leq p\leq m_k}\Big(d_{S_{i_p}}(c_{i_p}(u_k - td_k) - b_{i_p} + \rho_k^{-1}\lambda_{k,i_p}) + d_{S_{i_p}}(c_{i_p}(u_k) - J_{c_{i_p}}(u_k)(td_k) - b_{i_p} + \rho_k^{-1}\lambda_{k,i_p})\Big). \tag{5.3}$$

Suppose that there exists a positive constant β such that

$$\beta_{\xi_k}t_k \leq \beta \text{ and } 2\theta(1-\nu)\beta_{\xi_k} - t_k\mu_{h,\xi_k} \geq \epsilon_1 \tag{5.4}$$

Then, ρ_k is bounded below by $\rho_{\min} := \epsilon_1/(\beta\mu_c^e)$ with $\mu_c^e := \mathbf{E}_{\xi_k}[\mu_{c,\xi_k}]$. Moreover,

$$\|\|c(u_k) - b - P_S\big(c(u_k) - b + \rho_k^{-1}\lambda_k\big)\|\| \leq \mathcal{O}(1/\sqrt{k}). \tag{5.5}$$

Proof. Suppose that the line search step (3.5) does not yet terminate at a certain $t = \theta^j,\ \theta\in\,]0,1]$. Then, we have

$$\nu t\Delta f_{\lambda_k,\xi_k}(u_k;-d_k) \leq f_{\lambda_k,\xi_k}(u_k - td_k) - f_{\lambda_k,\xi_k}(u_k). \tag{5.6}$$

Following the definition of the function f_{λ_k,ξ_k}, the terms of the right-hand side in (5.6) can be written respectively as

$$f_{\lambda_k,\xi_k}(u_k - td_k) = \psi_{\xi_k}(c_{\xi_k}(u_k - td_k)) + h_{\xi_k}(u_k - td_k), \tag{5.7}$$
$$f_{\lambda_k,\xi_k}(u_k) = \psi_{\xi_k}(c_{\xi_k}(u_k)) + h_{\xi_k}(u_k). \tag{5.8}$$

Thus, the right-hand side of (5.6) becomes

$$f_{\lambda_k,\xi_k}(u_k - td_k) - f_{\lambda_k,\xi_k}(u_k) = [\psi_{\xi_k}(c_{\xi_k}(u_k - td_k)) - \psi_{\xi_k}\big(c_{\xi_k}(u_k)\big)] + [h_{\xi_k}(u_k - td_k) - h_{\xi_k}(u_k)]. \tag{5.9}$$

Using the definition of $\Delta f_{\lambda_k,\xi_k}(u_k;-td_k)$, the second term

$$-\psi_{\xi_k}(c_{\xi_k}(u_k)) = \Delta f_{\lambda_k,\xi_k}(u_k;-td_k) - \psi_{\xi_k}(c_{\xi_k}(u_k) - J_{c_{\xi_k}}(u_k)td_k) + \langle\nabla h_{\xi_k}(u_k)\mid td_k\rangle;$$

thus, the right-hand side of (5.6) can be expressed as

$$\begin{aligned} f_{\lambda_k,\xi_k}(u_k - td_k) - f_{\lambda_k,\xi_k}(u_k) = \Delta f_{\lambda_k,\xi_k}(u_k;-td_k) + \psi_{\xi_k}(c_{\xi_k}(u_k - td_k)) - \psi_{\xi_k}(c_{\xi_k}(u_k) - J_{c_{\xi_k}}(u_k)td_k) \\ + h_{\xi_k}(u_k - td_k) - h_{\xi_k}(u_k) + \langle\nabla h_{\xi_k}(u_k)\mid td_k\rangle. \end{aligned} \tag{5.10}$$

By Lemma 4, the function $\Psi_{\xi_k}\colon \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}\colon (u,\xi) \mapsto \psi_{\xi_k}(u) + \mathrm{Id}_R(\xi)$ is convex. Hence, by defining $\overline{c}_{\xi_k} : \mathbb{R}^K \to \mathbb{R}^m \times]-\infty,+\infty] : u \mapsto \overline{c}_{\xi_k}(u) = (c_{\xi_k}(u), c_0(u))$, it follows from [15, Lemma 3.1] that

$$\begin{aligned}\Delta f_{\lambda_k,\xi_k}(u_k; td_k) &= \Delta_0(\Psi_{\xi_k} \circ \overline{c}_{\xi_k})(u_k; td_k)\\ &\leq t\Delta_0(\Psi_{\xi_k} \circ \overline{c}_{\xi_k})(u_k; d_k) = t\Delta f_{\lambda_k,\xi_k}(u_k; d_k).\end{aligned} \tag{5.11}$$

In turn, simple calculations show that

$$\begin{aligned}&\Delta f_{\lambda_k,\xi_k}(u_k, -td_k) + \psi_{\xi_k}(c_{\xi_k}(u_k - td_k)) - \psi_{\xi_k}(c_{\xi_k}(u_k) - J_{c_{\xi_k}}(u_k)td_k)\\ &\qquad \leq t\Delta f_{\lambda_k,\xi_k}(u_k; -d_k) + \psi_{\xi_k}(c_{\xi_k}(u_k - td_k)) - \psi_{\xi_k}(c_{\xi_k}(u_k) - J_{c_{\xi_k}}(u_k)td_k).\end{aligned} \tag{5.12}$$

To determine the upper bound of the third term in the RHS of (5.12), we make use of the Assumption 1 and the $\mu_{c_{i_p}}$- Lipschitz continuity property of $J_{c_{i_p}}$ to obtain

$$\begin{aligned}&\psi_{\xi_k}(c_{\xi_k}(u_k - td_k)) - \psi_{\xi_k}(c_{\xi_k}(u_k) - J_{c_{\xi_k}}(u_k)td_k)\\ &= \frac{\rho_k}{2\mathtt{m}_k}\sum_{p=1}^{\mathtt{m}_k}\left(d^2_{S_{i_p}}(c_{i_p}(u_k - td_k) - b_{i_p} + \rho_k^{-1}\lambda_{k,i_p}) - d^2_{S_{i_p}}(c_{i_p}(u_k) - J_{c_{i_p}}(u_k)td_k - b_{i_p} + \rho_k^{-1}\lambda_{k,i_p})\right)\\ &\leq \frac{\beta_{\xi_k}\rho_k}{2\mathtt{m}_k}\sum_{p=1}^{\mathtt{m}_k}\left(d_{S_{i_p}}\big(c_{i_p}(u_k - td_k) - b_{i_p} + \rho_k^{-1}\lambda_{i_p}\big) - d_{S_{i_p}}\big(c_{i_p}(u_k) - J_{c_{i_p}}(u_k)td_k - b_{i_p} + \rho_k^{-1}\lambda_{i_p}\big)\right)\\ &\leq \frac{\beta_{\xi_k}\rho_k}{2\mathtt{m}_k}\sum_{p=1}^{\mathtt{m}_k}\|c_{i_p}(u_k - td_k) - c_{i_p}(u_k) - J_{c_{i_p}}(u_k)td_k\|\\ &\leq \frac{\beta_{\xi_k}\mu_{c,\xi_k}\rho_k}{2}\|td_k\|^2.\end{aligned} \tag{5.13}$$

Moreover, since the gradient of h_{ξ_k} is μ_{h,ξ_k}-Lipschitz continuous, we also have

$$h_{\xi_k}(u_k - td_k) - h_{\xi_k}(u_k) + \langle \nabla h_{\xi_k}(u_k) \mid td_k\rangle \leq \frac{\mu_{h,\xi_k}}{2}t^2\|d_k\|^2. \tag{5.14}$$

Therefore, we derive from (5.6), (5.12), (5.13) and (5.14) that

$$\nu t\Delta f_{\lambda_k,\xi_k}(u_k; -d_k) \leq t\Delta f_{\lambda_k,\xi_k}(u_k; -d_k) + \frac{1}{2}\big(\beta_{\xi_k}\mu_{c,\xi_k}\rho_k + \mu_{h,\xi_k}\big)t^2\|d_k\|^2, \tag{5.15}$$

which implies that

$$t \geq \frac{2(1-\nu)}{\big(\beta_{\xi_k}\mu_{c,\xi_k}\rho_k + \mu_{h,\xi_k}\big)\|d_k\|^2}|\Delta f_{\lambda_k,\xi_k}(u_k; -d_k)|. \tag{5.16}$$

In turn, the line search step (3.5) terminates at $t_k > 0$ (since $\nu \in]0,1[$) with

$$t_k \geq \frac{2\theta(1-\nu)}{\big(\beta_{\xi_k}\mu_{c,\xi_k}\rho_k + \mu_{h,\xi_k}\big)\|d_k\|^2}|\Delta f_{\lambda_k,\xi_k}(u_k; -d_k)|. \tag{5.17}$$

In view of Lemma 9,

$$|\Delta f_{\lambda_k,\xi_k}(u_k; -d_k)| \geq \beta_k\|d_k\|^2. \tag{5.18}$$

It follows by combining (5.18) with (5.17) that

$$t_k \geq \frac{2\theta(1-\nu)\beta_k}{\beta_{\xi_k}\mu_{c,\xi_k}\rho_k + \mu_{h,\xi_k}}, \tag{5.19}$$

which is the first assertion in (5.2). Hence, by involving Corollary 2, we deduce

$$\sum_{k\in\mathbb{N}}\frac{2\theta(1-\nu)\beta_k\|d_k\|^2}{\beta_{\xi_k}\mu_{c,\xi_k}\rho_k + \mu_{h,\xi_k}} < +\infty, \tag{5.20}$$

which implies the second assertion in (5.2). Moreover, from (5.3) and (5.19), we also obtain

$$\beta\mu_{c,\xi_k}\rho_k \geq \beta_{\xi_k}\mu_{c,\xi_k}\rho_k t_k \geq 2\theta(1-\nu)\beta_{\xi_k} - t_k\mu_{h,\xi_k} \geq \epsilon_1. \tag{5.21}$$

This inequality implies that the sequence $(\rho_k)_{k\in\mathbb{N}}$ is bounded below by $\rho_{\min}$. Note also that

$$\sum_{k\in\mathbb{N}} \|u_{k+1} - u_k\|^2 \leq \sum_{k\in\mathbb{N}} t_k\|d_k\|^2 < +\infty. \tag{5.22}$$

Hence, using the Step 1 of Algorithm 2, it follows that

$$\sqrt{\rho_{\min}}^2 \|\|c(u_k) - b - P_S\big(c(u_k) - b + \rho_k^{-1}\lambda_k\big)\|\|^2 \leq \min_{1\leq i\leq k} \|u_k - u_{k-1}\|^2 = \mathcal{O}(1/k), \tag{5.23}$$

which proves the last conclusion. □

6 Comparison and Related Work

The augmented Lagrangian method (ALM) is one of the most common approaches for solving nonlinear constrained problems. However, as stated in the Introduction section, constraints are often assumed to be convex implying that equality constraint functions must be affine and inequality constraint functions must be convex. With the proposed method, the minimization of the (possibly) nonconvex objective function h can be subject to nonlinear equality and inequality constraints.

The handling of such problems has been the subject of significant efforts, including LANCELOT, GENCAN, and ALGENCAN due to the ability of ALM to solve large-scale problems. The latter (and most recent) algorithmic scheme iterates by approximately minimizing the so-called PHR-Augmented Lagrangian function subject to bound constraints as well as updating both the penalty parameter and the Lagrange multipliers. ALGENCAN includes a decision that takes into account improvements in both the feasibility and complementary conditions. If both feasibility and complementary conditions were improved, it is considered that the penalty parameter is sufficiently large; thus, it is not further increased. Otherwise, it is multiplied by factor large than 1. ALGENCAN imposes that the KKT multiplier estimates must be bounded by explicitly projecting the estimates on a compact box after each update. The main reason invoked is to preserve the property of external penalty methods such that global minimizers of the original problem are obtained if each outer iteration computes a global minimizer of the subproblem. The boundedness of penalty parameters imposes in turn to assume that the KKT multipliers are within the bounds imposed by the algorithm. For these purposes, ALGENCAN uses safeguarded KKT multipliers such that limit points converge to KKT points under the Constant Positive Linear Dependence (CPLD) constraint qualification –which is weaker than MFCQ– and exhibit good properties in terms of penalty parameter boundedness. Although insufficient to ensure convergence for general nonconvex problems, as shown, for instance, in [9, Section 2.1], the properties of this algorithmic scheme have been shown to be competitive against alternatives such as interior point methods.

Recently, several ALM-based methods have been proposed to deal with the minimization of nonconvex objective functions subject to nonconvex equality constraints [39] and even fewer with inequality constraints [43]. In [43], authors aim to minimize over $x \in \mathbb{R}^n$, the composite function $f(x) + g(x)$ subject to equality constraints $c(x) = 0$ and inequality constraints $d(x) \leq 0$ with f continuously differentiable but possibly nonconvex, g closed convex but possibly nonsmooth, and c, d being vector functions from $\mathbb{R}^n \to \mathbb{R}^l$. Further, in addition to uniform regularity conditions (to ensure near feasibility of a near-stationary point to the augmented Lagrangian function), their proposed method assumes weak convexity of both the function f and each component of the vector function c; these assumptions significantly restrict its applicability. Constraints are then handled by introducing slack variables $s \geq 0$, leading the reformulation of the inequality constraints as $d(x) + s = 0$. Using the boundedness of the multipliers $\{y_k\}$, authors then show that their algorithm enables to reach an $\epsilon-$KKT point $(\bar{x}; \bar{s})$ with a corresponding multiplier $(\bar{y}, \bar{z})$. It turns out that $\bar{x}$ is an $O(\epsilon)-$KKT point of the original problem in terms of primal feasibility, dual feasibility, and the complementarity condition.

The former [39] applies the accelerated proximal gradient method as proposed by [21] to find an approximate primal solution to the ALM subproblems. The latter [43], referred to as Rate-Improved (RI)-iALM uses an inexact proximal point (iPP) method to approximately solve each ALM subproblem. The iPP procedure itself relies on the accelerated proximal gradient (APG) algorithm to solve each iPP subproblem. This combination yields a triple loop algorithm: each iteration k of the main ALM routine calls the iPP procedure to compute a x_{k+1} iterate that is itself the output obtained after running t iterations of the APG algorithm. This triple loop structure contrasts with the single-loop characterizing the proposed stochastic ALM algorithm. In [43], authors report that this change of subroutine for the solving of nonconvex subproblems enables to obtain order-reduced complexity by geometrically increasing the penalty parameter in ALM compared to [39] as well as more stable and efficient numerical results under the same assumptions. The complexity result of iPP has the best dependence on the smoothness and weak convexity constant (per iteration); however, for most problems, their explicit formula remains unknown and the corresponding parameters tuned. Table 6 compares the proposed stochastic ALM algorithm with inexact ALM (iALM) [39] and Rate-Improved ALM (RI-ALM) [43]. The complexity in the number of iterations (last column) is demonstrated in Section 5.

Table 1 Comparison of ALM methods for nonconvex nonlineary constrained problems

Method	**Type**	**Objective**	**Constraints**	**Type**	**Regularity Condition**	**Complexity**
iALM [31]	Inexact	Convex	Convex	Inequality		$\tilde{O}(\varepsilon^{-1})$
iALM [39]	Inexact	Nonconvex	Nonconvex	Equality	[39, Equation 18]	$\tilde{O}(\varepsilon^{-4})$
RI-iALM [43]	Inexact	Nonconvex	Convex Nonconvex	Equality Inequality†	[43, Assumption 3]	$\tilde{O}(\varepsilon^{-3})$
This paper	Inexact (Line Search)	Nonconvex	Convex Nonconvex	Equality Inequality	Assumption 1	$\tilde{O}(1/\sqrt{k})$ $\sim \tilde{O}(\varepsilon^{-2})$

Ethical Approval

Not applicable.

Disclosure statement

The authors report there are no competing interests to declare, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Funding

Not applicable.

Authors' contributions

Authors have contributed equally to this manuscript. All authors reviewed the manuscript.

Acknowledgments

Not applicable.

Availability of supporting data

Not applicable.

References

1. R. Andreani, E. Birgin, J. Martınez and M. L. Schuverdt, Augmented Lagrangian methods under the Constant Positive Linear Dependence constraint qualification, *Math. Program.*, Vol. 111, pp. 5-32, 2008.
2. R. Andreani, E. G. Birgin, J. M. Martınez and M. L. Schuverdt, On Augmented Lagrangian Methods with General Lower-Level Constraints, *SIAM J. Optim.*, Vol. 18, pp. 1286-1309, 2008.
3. F. Bach and E. Moulines, Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning, *Advances in Neural Information Processing Systems*, pp. 451-459, 2011.
4. H. H. Bauschke and P. L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer, New York, 2nd ed., 2017.
5. S. Becker, J. Bobin and E. J. Candès, NESTA: A fast and accurate first-order method for sparse recovery, *SIAM J. Imag. Sci.*, Vol. 4, pp. 1-39, 2011.
6. A. Beck and M. Teboulle, Smoothing and first order methods: A unified framework, *SIAM J. Optim.*, Vol. 22, pp.557–580, 2012
7. W. Ben-Ameur and A. Ouorou, Mathematical models of the delay constrained routing problem, *Algorithmic Oper. Res.*, Vol.1, pp. 94-103, 2006.
8. A. Ben-Tal, A. Goryashko, E. Guslitzer and A. Nemirovski, Adjustable robust solutions of uncertain linear programs, *Math. Program.*, Vol. 99, pp. 351–376, 2004.
9. D. P. Bertsekas, Constrained Optimization and Lagrange Multiplier Methods, Academic Press, London, 2014.
10. J. R. Birge and J. K. Ho, Optimal Flows in Stochastic Dynamic Networks with Congestion, *Oper. Res.*, Vol. 41, pp. 203-216, 1993.
11. E. G. Birgin and J. M. Martınez, Large-scale active-set box-constrained optimization method with spectral projected gradients, *Comput. Optim. Appl.*, Vol. 23, pp.101-125, 2002.
12. J. Bolte, S. Sabach and M. Teboulle, Nonconvex lagrangian-based optimization: monitoring schemes and global convergence, *Math. Oper. Res.*, Vol. 43, pp. 1210-1232, 2018.
13. L. Bottou, Stochastic gradient learning in neural networks, *Proceedings of Neuro-Nîmes 91*, EC2, Nimes, France, 1991.
14. L. Bottou, F. E. Curtis and J. Nocedal, Optimization Methods for Large-Scale Machine Learning, *Siam Review*, Vol. 60, pp. 223-311, 2018.
15. J. V. Burke and A. Engle, Line search and trust-region methods for convex-composite optimization, https://arxiv.org/abs/1806.05218, 2018.
16. M. Carey, Optimal Time Varying Flows On Congested Networks, *Oper. Res.*, Vol. 35, pp.58-69, 1987.
17. P. L. Combettes, Đ. Dũng and B. C. Vu, Proximity for sums of composite functions, *J. Math. Anal.*, Vol., 380, pp. 680-688, 2011.
18. S. Cui and U. V. Shanbhag, Variance-reduced splitting schemes for monotone stochastic generalized equations, *IEEE Trans. Autom. Control*, 2023.
19. D. Gabay, Applications of the method of multipliers to variational inequalities, in: M. Fortinand R. Glowinski (Eds.), Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems, North-Holland, Amsterdam, 1983.
20. D. Gabay and B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, *Comput. Math. with Appl.*, Vol. 2, pp. 17-40, 1976.
21. S. Ghadimi and G. Lan, Accelerated gradient methods for nonconvex nonlinear and stochastic programming, *Math. Program.*, Vol. 156, pp. 59-99, 2016.
22. J. Giesen and S. Laue, Distributed Convex Optimization with Many Non-Linear Constraints, https://arxiv.org/pdf/1610.02967.pdf, 2018.
23. F. Glover, Improved Linear Integer Programming Formulations of Nonlinear Integer Problems, *Manag. Sci.*, Vol. 22, pp. 455-460, 1975.
24. M. R. Garey, D. S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W.H. Freeman, ISBN 0-7167-1045-5, 1979.
25. M. X. Goemans and D. P. Williamson, Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming, *J. ACM*, Vol 42, pp. 1115–1145, 1995.
26. A. Gupta, J. Kleinberg, A. Kumar, R. Rastogi and B. Yener, Provisioning a virtual private network: A network design problem for multicommodity flow, Proc 33rd annual ACM Symposium on Theory of Computing (STOC 2001), Heraklion, Crete, Greece, 2001, pp.389–398.
27. H. L. Hijazi, P. Bonami and A. Ouorou, Robust delay-constrained routing in telecommunications, *Ann. Oper. Res.*, Vol. 206, pp. 163-181, 2013.
28. E. Hazan and S. Kale, Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization, *J. Mach. Learn. Res.*, Vol. 15, pp. 2489-2512, 2014.
29. M. R. Hestenes, Multiplier and gradient methods, *J. Optim. Theory Appl.*, Vol 4, pp. 303-320, 1969.
30. A. S. Lewis and S. J. Wright, A proximal method for composite minimization, *Math. Program.*, Vol. 158, pp. 501–546, 2016.

31. Z. Li and Y. Xu, Augmented Lagrangian based first-order methods for convex and nonconvex programs: nonergodic convergence and iteration complexity, 2021. Available at https://arxiv.org/abs/2003.08880
32. H. Xhang, S. Ghadimi and G. Lan, Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization, *Math. Program., Ser. A,* Vol. 155, pp. 267–305, 2016
33. David G. Luenberger, Yinyu Ye, et al., Linear and nonlinear programming, Volume 2, Springer, 2007, Third edition.
34. A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro, Robust stochastic approximation approach to stochastic programming, *SIAM J. Optim.*, Vol. 19, pp. 1574–1609, 2008.
35. Y. Nesterov, Smooth minimization of non-smooth functions, *Math. Program.,* Vol. 103, pp. 127-152, 2005.
36. D. Papadimitriou and B. Vu, An augmented Lagrangian method for nonconvex composite optimization problems with nonlinear constraints, *Optimization and Engineering*, Springer, Nov. 2023.
37. Q. Tran-Dinh, O. Fercoq and V. Cevher, A smooth primal-dual optimization framework for nonsmooth composite convex minimization, *SIAM J. Optim.*, Vol. 28, pp. 96-134,2018
38. M. J. D. Powell, A method for non-linear constraints in minimization problems, Optimization, R. Fletcher Ed., Academic Press, New York, NY, pp. 283–298, 1969.
39. M. F. Sahin, A. Alacaoglu, F. Latorre and V. Cevher, An inexact augmented Lagrangian framework for nonconvex optimization with nonlinear constraints, *In Advances in Neural Information Processing Systems*, pp. 13943–13955, 2019.
40. T. Valkonen, A primal-dual hybrid gradient method for nonlinear operators with applications to MRI, *Inverse Probl.*, Vol.30, 055012, 2014.
41. http://proximity-operator.net/
42. https://sites.google.com/site/fomsolv
43. Z. Li, P. Y. Chen, S. Liu, S. Lu and Y. Xu, Rate-improved Inexact Augmented Lagrangian Method for Constrained Nonconvex Optimization, *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, San Diego, California, USA. PMLR, Vol. 130, 2021.
44. H. Robbins and D. Siegmund, A convergence theorem for non negative almost supermartingales and some applications, In: Rustagi JS, editor. *Optimizing methods in statistic*, New York (NY): Academic Press, pp. 233-257, 1971.
45. H. Robbins and S. Monro, A Stochastic Approximation Method, *Ann. Math. Statist.*, Vol. 22, pp. 400-407, 1951.
46. V. D. Nguyen and B. C. Vũ, Convergence analysis of the stochastic reflected forward–backward splitting algorithm, *Optim. Lett.*, Vol. 16, pp. 2649-2679, 2022.