A stochastic Lagrangian-based method for nonconvex empirical risk minimization with nonlinear constraints

Dimitri Papadimitriou*[1] and Bằng Công Vũ[2]

[1]Université libre de Bruxelles, Belgium, [2] BeRC, Leuven, Belgium

Abstract: To tackle the computational and statistical challenges of constrained training at large scale, we propose a stochastic Augmented Lagrangian Method (ALM) method that performs on a mini-batch of randomly selected points to minimize the empirical risk defined as a smooth possibly nonconvex function subject to both nonlinear equality and inequality constraints. The convergence properties (in expectation) of the proposed algorithm and its complexity are then thoroughly demonstrated under very general assumptions and compared against inexact state-of-the-art ALM methods.

1 Introduction

Minimization problems involving both equality and inequality constraints have received significant attention in statistical/neural learning community due to the need to incorporate various constraints during the training process. These constraints range from hardware-specifics (e.g., energy/power consumption, memory space/access time) to learning tasks themselves (e.g., fairness and robustness). With current training models, these requirements are often enforced using penalties, i.e., by augmenting the empirical risk minimization (ERM) objective with constraint violation costs. This augmentation leads in turn to weight decay in the update steps of the minibatch stochastic gradient (SG) algorithm. It is relatively straightforward to find penalty terms leading to optimal solutions when the objective of the ERM problem and the constraints are convex (implying that equality constraint functions must be affine and inequality constraint functions convex). However, the training of neural networks leads to nonconvex objective functions subject to nonlinear constraints. Enforcing their behavior to tackle fairness, robustness, and safety requires thus to solve instead nonlinearly constrained nonconvex problems. Consequently, the update step of the minibatch SG algorithm is not limited anymore to a weight decay.

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,

$$\text{minimize } h(x) \quad \text{subject to} \quad c_1(x) = b_1,\ c_2(x) \leq b_2,\ x \in C, \tag{1}$$

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 $\mathbb{R}^K$. 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. Introduced by Powell and Hestenes in 1969 [29] [19], ALM is one of the most common approaches for solving linear and nonlinear constrained problems. However, for non-convex objectives, handling both equality and inequality constraints that are nonlinear remains challenging.

In this respect, the main objective is to extend the ALM [13] [23] for the nonlinear setting described by the above model. For this purpose, we propose a stochastic ALM (sALM) with backtracking line search that performs on a mini-batch of randomly selected points for solving such nonlinearly constrained nonconvex problems. The considered class of problems includes both nonlinear equality and inequality constraints; thus, the minimization of the (possibly) nonconvex objective function h can be subject to nonlinear equality and inequality constraints without imposing convexity of its functions. Moreover, our method relies on line search that performs on a subset of randomly selected points only; hence, the sALM 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. The main motivation for the design of a Lagrangian-based algorithm that relies on the mini-batch SG can be stated as follows. The SG method was first introduced in [34]. This method, as well as its extension, the stochastic proximal gradient, have been widely adopted nowadays as optimization method in machine learning (statistical learning, deep learning, etc.), linear inverse problem, and game theory; see [1] [7] [8] [24] [18] for examples. A main property of the SG 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 SG; see [8] for a detailed development. The major advantage of the mini-batch SG is the reduction of variance when the mini-batch size increases [8] [22] [12].

*dimitrios.papadimitriou@ulb.be

2 Algorithm

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 $\mathrm{prox}_f\colon \mathbb{R}^K \to \mathbb{R}^K\colon x \mapsto \underset{y\in\mathbb{R}^K}{\operatorname{argmin}}\, f(y) + \frac{1}{2}\|x-y\|^2$. The conjugate function of f is denoted by $f^\star$. When f is the indicator function of some closed convex $S \subset \mathbb{R}^K$, which is denoted by $\iota_S\colon x \mapsto 0$ if $x \in S, +\infty$ if $x \notin S$, 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 designated by $J_c(u)$ and its conjugate 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 $\langle\cdot \mid\mid \cdot\rangle$ defined for $y := (y_q)_{1\leq q\leq M}$ and $v := (v_q)_{1\leq q\leq M}$ as*

$$\langle\cdot \mid\mid \cdot\rangle : (y,v) \mapsto \sum_{q=1}^M \omega_q \langle v_q \mid y_q\rangle \tag{2}$$

$$\text{with vector norm } |||\cdot|||\colon v \mapsto \sqrt{\langle v \mid\mid v\rangle}, \tag{3}$$

Definition 2 *[9] 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*

$$\Delta_0\varphi(u;d) = f(u+d) + g\big(c(u) - b + J_c(u)d\big) - \varphi(u) < 0, \tag{4}$$

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.

The following lemma generalizes the definition of descent direction d_k to nonconvex functions denoted 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 1 *Assume $\overline{c}\colon \mathbb{R}^K \to \mathbb{R}^m \times]-\infty,+\infty]\colon u \mapsto \overline{c}(u) = (c(u), c_0(u))$ 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 Ψ_k is convex. The composition $(\Psi_k \circ \overline{c})$ verifies the identity*

$$(\Psi_k \circ \overline{c})(u) = \psi_k \circ c(u) + \mathrm{Id}_R \circ c_0(u) \equiv \varphi_k(u), \tag{5}$$

where $\varphi_k\colon u \mapsto h(u) + \psi_k \circ c(u)$. Moreover, by defining, for every $u \in \mathrm{dom}(\varphi)$ and $d \in \mathbb{R}^K$,

$$\Delta\varphi_k(u;d) = \psi_k(c(u) + J_c(u)d) + \langle \nabla h(u) \mid d\rangle - \psi_k(c(u)), \tag{6}$$

the following identity is verified

$$\Delta_0(\Psi_k \circ \overline{c})(u;d) \equiv \Delta\varphi_k(u;d). \tag{7}$$

Using Lemma 1 (proof, see [27, Lemma 4]), 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.

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*

$$g_\rho\colon (u,\lambda) \mapsto \sup_{y\in\mathbb{R}^m} \Big(\langle c(u) - b \mid\mid y\rangle - g^\star(y) - \frac{1}{2\rho}|||y-\lambda|||^2\Big), \tag{8}$$

where ρ is referred to as the smoothing parameter and $g^\star$ denotes the Fenchel conjugate of the function g that is defined by $g^\star\colon u \mapsto \sup_{x\in\mathbb{R}^m}\big(\langle u \mid\mid x\rangle - g(x)\big)$.
***Note**: the function g_ρ provides a smooth approximation of g, which is known as the smoothing technique.*

Problem: 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

$$\text{minimize } h(u) = \sum_{q=1}^M \omega_q h_q(u) \tag{9}$$

$$\text{subject to } (\forall q \in \{1,\ldots,M\})\ c_q(u) - b_q \in S_q, u \in C. \tag{10}$$

The main features of the proposed algorithm are the following. Firstly, it is structured as a *single-loop* algorithm; more precisely, it does not require calling a first-order method (such as the proximal gradient) to compute inner iterations. 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 stepsizes. The design principles of the proposed single-loop of Algorithm 2 for solving Problem 9 are i) by *smoothing the nonlinear constraints* $c_q(u) - b_q \in S_q$, formulate a generalization of the augmented Lagrangian function $\mathcal{L}_\rho(u, \lambda)$ that 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 multiplier λ, 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; and iii) *backtracking* to find the primal t_k and the dual stepsizes s_k and then update the dual variable λ as $\lambda^+ = \lambda + s_k \nabla_2 \mathcal{L}_\rho(u^+, \lambda)$. 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. In Section 3, 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 assume the following:

Assumption 1 Let C be a closed convex subset of $\mathbb{R}^K$ and μ_c be a positive constant. The Jacobian J_c of the constraints c verifies

$$\mu_0 = \sup_{u\in C} \|J_c(u)^\intercal\| < +\infty \quad \text{and} \quad (\forall (u, \tilde{u}) \in C \times C)\ \|J_c(u) - J_c(\tilde{u})\| \leq \mu_c \|u - \tilde{u}\|, \tag{11}$$

Assumption 2 We further assume that the variance of d_{k,i_p} in (18) denoted by $\mathrm{Var}(d_{k,i_p})$ is bounded. More precisely, 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 $k \in \mathbb{N}$,

$$\mathrm{Var}(d_{k,i_p}) = \mathbf{E}_{i_p}[\|d_{k,i_p} - \nabla \mathcal{L}_{\rho_k}(u_k, \lambda_k)\|^2 | \mathcal{E}_k] \leq \sigma_k^2 < +\infty, \tag{12}$$

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,

$$\mathrm{Var}(d_k) \leq \frac{1}{\mathtt{m}_k^2} \sum_{p=1}^{\mathtt{m}_k} \mathrm{Var}(d_{k,i_p}), \tag{13}$$

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

$$f_{\lambda_k,\xi_k}(\cdot) = \frac{1}{\mathtt{m}_k} \sum_{p=1}^{\mathtt{m}_k} \Big(h_{i_p}(\cdot) + \frac{\rho_k}{2} d^2_{S_{i_p}}(c_{i_p}(\cdot) - b_{i_p} + \rho_k^{-1} \lambda_{k,i_p}) - \frac{1}{2\rho_k} \|\lambda_{k,i_p}\|^2 \Big). \tag{14}$$

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 $S = \prod_{q=1}^M S_q$.

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

Lemma 2 *[27, Lemma 7] The line search Algorithm 1 terminates after a finite number of steps, i.e., there exists $t_k > 0$ such that*

$$f_{\lambda_k,\xi_k}(\bar{u}_{k+1}) < f_{\lambda_k,\xi_k}(u_k) + \nu t_k \Delta f_{\lambda_k,\xi_k}(u_k; -d_k) + \mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big). \tag{15}$$

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

$$\nu\beta_k - \frac{t_k}{2}\Big((1 + t_k\ell_{\xi_k})^2 - (1 + t_k\ell_k)^2 - 2(1+\varepsilon)\Big) - \varsigma_{1,k} \geq \frac{\varepsilon}{2}. \tag{16}$$

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, 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

$$f_{\lambda_k,\xi_k}(\bar{u}_{k+1}) < f_{\lambda_k,\xi_k}(u_k) + \nu t_\theta \Delta f_{\lambda_k,\xi_k}(u_k; -d_k) + \mathcal{O}\left(\frac{1}{(k+1)^{1+\varepsilon}}\right) \tag{17}$$

  then break
  end if
end for

▷ **Step 2: Final step size**

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

Algorithm 2 : Stochastic ALM algorithm

▷ **Initialization**

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

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

Compute μ_0 from (11)

▷ **Main Loop**

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

▷ **Step 1**: Select $\rho_k \in]0, \infty[$ such that

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

▷ **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}

$$\begin{aligned} v_{k,i_p} &= \lambda_{k,i_p} + \rho_k\big(c_{i_p}(u_k) - b_{i_p} - P_{S_{i_p}}(c_{i_p}(u_k) - b_{i_p} + \rho_k^{-1}\lambda_{k,i_p})\big) \\ d_{k,i_p} &= \nabla h_{i_p}(u_k) + J_{c_{i_p}}(u_k)^{\intercal} v_{k,i_p} \end{aligned} \tag{18}$$

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

▷ **Step 3**

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

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

▷ **Step 4**

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

Update $\quad \lambda_{k+1} = \lambda_k + s_k\Big(c(u_{k+1}) - b - P_S\big(c(u_{k+1}) - b + \rho_k^{-1}\lambda_k\big)\Big)$

end for

3 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 3 are well defined. They are presented in [27, Lemma 10] and [27, Lemma 7]. In particular, we obtain the descent property of the stochastic function f_{ξ_k,λ_k} as in (17).

(ii) We further estimate $\Delta f_{\lambda_k,\xi_k}(u_k; -d_k) \leq -\beta_k\|d_k\|^2$ as proved in [27, Lemma 9]. Combining this result to (17), we obtain the descent of f_{ξ_k,λ_k} with respect to d_k as

$$f_{\lambda_k,\xi_k}(u_{k+1}) < \mathcal{L}_{\rho_k,\xi_k}(u_k, \lambda_{k,\xi_k}) - t_k\beta_k\nu\|d_k\|^2 + \mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big). \tag{19}$$

(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 [27, Lemma 8] where we show that the Lagrange function satisfies a sufficient decrease condition and [27, Lemma 11] to derive the descent property of $\big(\mathcal{L}_{\rho_k}(u_k, \lambda_k)\big)_{k\in\mathbb{N}}$ from (19) as in (20)

$$\begin{aligned} &\mathbf{E}_{\xi_k}\Big[\mathcal{L}_{\rho_{k+1}}(u_{k+1}, \lambda_{k+1}) + 2(1+\varepsilon)t_k^2\|d_k\|^2\Big] \\ &\quad\leq \mathcal{L}_{\rho_k}(u_k, \lambda_k) + 2(1+\varepsilon)t_{k-1}^2\|d_{k-1}\|^2 \\ &\qquad - \mathbf{E}_{\xi_k}\Big[t_k\big(\nu\beta_k - \frac{1}{2}t_k(1+t_k\ell_{\xi_k})^2 - t_k(1+t_k\ell_k)^2 - 2(1+\varepsilon)t_k - \varsigma_{1,k}\big)\|d_k\|^2\Big] + \mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big) \\ &\quad\leq \mathcal{L}_{\rho_k}(u_k, \lambda_k) + 2(1+\varepsilon)t_{k-1}^2\|d_{k-1}\|^2 - \frac{\varepsilon}{2}\mathbf{E}_{\xi_k}\left[t_k\|d_k\|^2\right] + \mathcal{O}\Big(\frac{1}{(k+1)^{1+\varepsilon}}\Big). \end{aligned} \tag{20}$$

(iv) From (20), it is easy to find the convergence property of the proposed method as in Theorem 3.

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*

$$m_k \geq \mathcal{O}\left(\sigma_k^2 t_{k,max}(k+1)^{1+\varepsilon}\right) \tag{21}$$

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.*

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

(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.*

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

Proof. see [27, Theorem 3]□

Next, we demonstrate the (local) convergence of the sequence $(u_k,\lambda_k)_{k\in\mathbb{N}}$ to a stationary point of the augmented Lagrangian function $\mathcal{L}_\rho$. For this purpose, in addition to the lower boundedness of $\mathcal{L}_{\rho_k}(u_k,\lambda_k)$ for all $k\in\mathbb{N}$, the following conditions are assumed to be verified. Let $(u_k)_{k\in\mathbb{N}} \subset Z \subseteq C$.

C1 Constraints c verifies the MFCQ conditions on Z, i.e., $\exists\,\zeta\in\,]0,+\infty[$ such that $\forall\, v\in Y = c(Z)-b$ of $\mathbb{R}^m$, the following inequality is verified $\forall\, u\in Z\ :\ \zeta\|v\| \leq \|J_c(u)^\intercal v\|$.

C2 The sequence $(\rho_k)_{k\in\mathbb{N}}$ is bounded from above.

C3 The primal sequence $(u_k)_{k\in\mathbb{N}}$ generated by Algorithm 2 is bounded.

The demonstration proceeds as follows: first, show (cf. [27, Proposition 1]) that the limit points of the subsequences produced by Algorithm 2 verify the first-order KKT conditions. The latter can be stated as follows: If $N_C(u^\dagger)$ defines the normal cone of C at a local minimum $u^\dagger$ (that satisfies the condition **C1**); then, $\exists\,\lambda\in\mathbb{R}^m$, where $m=\sum_{q=1}^M m_q$, such that i) $-\Big(\nabla h(u^\dagger)+\sum_{i=1}^m \lambda_i J_{c_i}(u^\dagger)\Big) \in N_C(u^\dagger)$, ii) $c(u^\dagger)-b\in S$, iii) $\lambda\in S^\ominus$ (polar cone of S), iv) $\langle\lambda \mid c(u^\dagger)-b\rangle = 0$. The next step requires proving that the set of limits points is non-empty (cf. [27, Proposition 2]). Knowing this property, the last step consists of proving that, under the above conditions, the sequences produced by the algorithm converge to such limit point (cf. [27, Corollary 5]).

4 Iteration Complexity

In this section, we characterize the iteration complexity of the proposed sALM algorithm in terms of the difference $\Delta\mathcal{L}_k(\cdot,\lambda_k)$ and the feasibility. Iteration complexity refers to the number of iterations required to obtain an approximate ε-KKT point of the Problem 9 by means of the sALM algorithm proposed in Section 2.

Theorem 4 *Suppose that $C=\mathcal{H}\times\mathcal{G}$. Let $\mu_{c_{i_p}}$ and $\mu_{h_{i_p}}$ be the Lipschitz constant of $J_{c_{i_p}}$ and ∇h_{i_p}. Set*

$$\mu_{c,\xi_k} = \frac{1}{2m_k}\sum_{p=1}^{m_k}\mu_{c_{i_p}} \text{ and } \mu_{h,\xi_k} = \frac{1}{2m_k}\sum_{p=1}^{m_k}\mu_{h_{i_p}} \tag{22}$$

Assume that the conditions stated in Theorem 3 are satisfied. Then, t_k and u_k verify the following

$$t_k \geq \frac{2\theta(1-\nu)\beta_{\xi_k}}{\beta_{\xi_k}\mu_{c,\xi_k}\rho_k+\mu_{h,\xi_k}} \text{ and } \min_{0\leq i\leq k}\frac{2\theta(1-\nu)\beta_{\xi_i}}{\beta_{\xi_i}\mu_{c,\xi_i}\rho_i+\mu_{h,\xi_i}}\|\frac{u_{i+1}-u_i}{t_i}\|^2 = \mathcal{O}(1/(k+1)), \tag{23}$$

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

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

Suppose that there exists a positive constant β such that $\beta_{\xi_k}t_k\leq\beta$ and $2\theta(1-\nu)\beta_{\xi_k}-t_k\mu_{h,\xi_k}\geq\epsilon_1$ 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,

$$\|\|c(u_k)-b-P_S\big(c(u_k)-b+\rho_k^{-1}\lambda_k\big)\|\| \leq \mathcal{O}(1/\sqrt{k}). \tag{24}$$

Proof. see [27, Theorem 4]□

5 Comparison and Related Work

Recently, several ALM-based methods have been proposed to deal with the minimization of nonconvex objective functions subject to nonconvex equality constraints [30] and even fewer with inequality constraints [32]. In [32], the proposed method, referred to as Rate-Improved (RI)-iALM, aims 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 $d(x) \leq 0$, where c, d are vector functions from $\mathbb{R}^n \to \mathbb{R}^l$. As part of the objective, the function f is assumed continuously differentiable but possibly nonconvex, g closed convex but possibly nonsmooth. 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 the applicability of the method. Constraints are then handled by introducing slack variables $s \geq 0$, leading to 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.

None of the methods proposed in these two references meet the properties of a single-loop algorithm. The former reference [30] applies the accelerated proximal gradient method as proposed by [15] to find an approximate primal solution to the ALM subproblems. The latter [32] 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 sALM algorithm. In [32], 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 [30] 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 1 compares the proposed stochastic ALM algorithm with inexact ALM (iALM) [30] and Rate-Improved ALM (RI-ALM) [32]. The complexity in the number of iterations (last column) is demonstrated in Section 4.

Table 1: Comparison of ALM methods for nonconvex nonlineary constrained problems

Method	**Type**	**Objective**	**Constraints**	**Type**	**Regularity Condition**	**Complexity**
iALM [21]	Inexact	Convex	Convex	Inequality		$\tilde{O}(\varepsilon^{-1})$
iALM [30]	Inexact	Nonconvex	Nonconvex	Equality	[30, Equation 18]	$\tilde{O}(\varepsilon^{-4})$
RI-iALM [32]	Inexact	Nonconvex	Convex Nonconvex	Equality Inequality†	[32, Assumption 3]	$\tilde{O}(\varepsilon^{-3})$
This paper	Inexact (Line Search)	Nonconvex	Convex Nonconvex	Equality Inequality	Assumption 1	$\tilde{O}(1/\sqrt{k})$ $\sim \tilde{O}(\varepsilon^{-2})$

References

[1] 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.

[2] H. H. Bauschke and P. L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer, New York, 2nd ed., 2017.

[3] 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.

[4] A. Beck and M. Teboulle, Smoothing and first order methods: A unified framework, *SIAM J. Optim.*, Vol. 22, pp.557–580, 2012

[5] D. P. Bertsekas, Constrained Optimization and Lagrange Multiplier Methods, Academic Press, London, 2014.

[6] 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.

[7] L. Bottou, Stochastic gradient learning in neural networks, *Proceedings of Neuro-Nîmes 91*, EC2, Nimes, France, 1991.

[8] L. Bottou, F. E. Curtis and J. Nocedal, Optimization Methods for Large-Scale Machine Learning, *SIAM Review*, Vol. 60, pp. 223-311, 2018.

[9] J. V. Burke and A. Engle, Line search and trust-region methods for convex-composite optimization, https://arxiv.org/abs/1806.05218, 2018.

[10] M. Carey, Optimal Time Varying Flows On Congested Networks, *Oper. Res.*, Vol. 35, pp.58-69, 1987.

[11] P. L. Combettes, Đ. Dũng and B. C. Vu, Proximity for sums of composite functions, *J. Math. Anal.*, Vol., 380, pp. 680-688, 2011.

[12] S. Cui and U. V. Shanbhag, Variance-reduced splitting schemes for monotone stochastic generalized equations, *IEEE Trans. Autom. Control*, 2023.

[13] 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.

[14] 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.

[15] S. Ghadimi and G. Lan, Accelerated gradient methods for nonconvex nonlinear and stochastic programming, *Math. Program.*, Vol. 156, pp. 59-99, 2016.

[16] F. Glover, Improved Linear Integer Programming Formulations of Nonlinear Integer Problems, *Manag. Sci.*, Vol. 22, pp. 455-460, 1975.

[17] 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.

[18] 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.

[19] M. R. Hestenes, Multiplier and gradient methods, *J. Optim. Theory Appl.*, Vol 4, pp. 303-320, 1969.

[20] A. S. Lewis and S. J. Wright, A proximal method for composite minimization, *Math. Program.*, Vol. 158, pp. 501–546, 2016.

[21] 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

[22] 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

[23] David G. Luenberger, Yinyu Ye, et al., Linear and nonlinear programming, Volume 2, Springer, 2007, Third edition.

[24] 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.

[25] Y. Nesterov, Smooth minimization of non-smooth functions, *Math. Program.,* Vol. 103, pp. 127-152, 2005.

[26] D. Papadimitriou and B. Vu, An augmented Lagrangian method for nonconvex composite optimization problems with nonlinear constraints, *Optimization and Engineering*, Springer, Nov. 2023.

[27] D. Papadimitriou and B. Vu, A stochastic Lagrangian-based method for nonconvex optimization with nonlinear constraints, optimization-online.org, available at https://optimization-online.org/?p=29110

[28] 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

[29] 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.

[30] 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.

[31] T. Valkonen, A primal-dual hybrid gradient method for nonlinear operators with applications to MRI, *Inverse Probl.*, Vol.30, 055012, 2014.

[32] 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.

[33] 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.

[34] H. Robbins and S. Monro, A Stochastic Approximation Method, *Ann. Math. Statist.*, Vol. 22, pp. 400-407, 1951.

[35] 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.