Augmented Lagrangian Optimization under Fixed-Point Arithmetic

Yan Zhang, Michael M. Zavlanos

Mechanical Engineering and Material Science, Duke University, Durham, NC 27708, USA

Abstract

In this paper, we propose an inexact Augmented Lagrangian Method (ALM) for the optimization of convex and nonsmooth objective functions subject to linear equality constraints and box constraints where errors are due to fixed-point data. To prevent data overflow we also introduce a projection operation in the multiplier update. We analyze theoretically the proposed algorithm and provide convergence rate results and bounds on the accuracy of the optimal solution. Since iterative methods are often needed to solve the primal subproblem in ALM, we also propose an early stopping criterion that is simple to implement on embedded platforms, can be used for problems that are not strongly convex, and guarantees the precision of the primal update. To the best of our knowledge, this is the first fixed-point ALM that can handle non-smooth problems, data overflow, and can efficiently and systematically utilize iterative solvers in the primal update. Numerical simulation studies on a logistic regression problem are presented that illustrate the proposed method.

Key words: Convex optimization, Augmented Lagrangian Method, embedded systems, fixed-point arithmetic.

1 Introduction

Embedded computers, such as FPGAs (Field-Programmable Gate Arrays), are typically low-cost, low-power, perform fast computations, and for these reasons they have long been used for the control of systems with fast dynamics and power limitations, e.g., automotive, aerospace, medical, and robotics. While embedded computers have been primarily used for low-level control, they have recently also been suggested to obtain real-time solutions to more complex optimization problems Jerez et al. (2014); Richter et al. (2017). The main challenges in implementing advanced optimization algorithms on resource-limited embedded devices are providing complexity certifications and designing the data precision to control the solution accuracy and the dynamic range to avoid overflow. While recent methods, such as Jerez et al. (2014); Richter et al. (2017), address these challenges and provide theoretical guarantees, they do so for quadratic problems. In this paper, we provide such guarantees for general convex problems. Specifically, we propose an Augmented Lagrangian Method (ALM) to solve problems of the form

\[
\min f(x) \; \text{s.t.} \; Ax = b, \; x \in \mathcal{X}
\]

on embedded platforms under fixed-point arithmetic, where \(x \in \mathbb{R}^n\), \(A \in \mathbb{R}^{p \times n}\), \(b \in \mathbb{R}^p\) and \(f(x)\) is a scalar-valued function. The set \(\mathcal{X}\) is convex. ALM falls in the class of first order methods, which have been demonstrated to be efficient in solving problem (1) on embedded computers due to their simple operations and less memory requirements, Richter et al. (2017).

Recent work on error analysis of inexact first-order methods is presented in Devolder, Glineur & Nesterov (2014); Necoara & Patrascu (2016); Patrinos, Guiggiani & Bemporad (2015). Specifically, Devolder et al. (2014) proposed an inexact oracle to characterize the iteration complexity and suboptimality of the primal gradient and fast gradient method. Necoara & Patrascu (2016); Patrinos et al. (2015) extend these results to the dual domain. However, these analyses assume strong convexity of the objective functions. Necoara, Patrascu & Glineur (2017); Nedelcu, Necoara & Tran-Dinh (2014) analyze the convergence of ALM using an inexact oracle for general convex problems. The inexactness comes from the approximate solution of the subproblems in ALM. Similar analysis has been conducted in Eckstein & Silva (2013); Lan & Monteiro (2016); Rockafellar (1976b). However, none of above works on ALM consider the error in the multiplier update under fixed-point arithmetic. Moreover, since no projection is used in the mul-

* This work is supported in part by AFOSR under award FA9550-19-1-0169 and by NSF under award CNS-1932011.

Email addresses: yz227@duke.edu (Yan Zhang), michael.zavlanos@duke.edu (Michael M. Zavlanos).
multiplier update, the above works cannot provide an upper bound on the multiplier iterates and therefore cannot avoid data overflow. Perhaps the most relevant work to the method proposed here is Jerez, Goulart, Richter, Constantinides, Kerrigan & Morari (2014). Specifically, Jerez et al. (2014) analyzes the behavior of the Alternating Direction Method of Multipliers (ADMM), a variant of the ALM, on fixed-point platforms. However, the analysis in Jerez et al. (2014) can only be applied to quadratic objective functions. Moreover, to prevent data overflow, the proposed method needs to monitor the iteration history of the algorithm to estimate a bound on the multiplier iterates.

Compared to existing literature on inexact ALM, we propose a new inexact ALM that incorporates errors in both the primal and dual updates and contains a projection operation in the dual update. Assuming a uniform upper bound on the norm of the optimal multiplier is known, this projection step can guarantee no data overflow during the whole iteration history. To the best of our knowledge, this is the first work to provide such guarantees for general convex and non-smooth problems. Furthermore, we show that our proposed algorithm has $O(1/K)$ convergence rate and provide bounds on the achievable primal and dual suboptimality and infeasibility. In general, iterative solvers are needed to solve the subproblems in ALM but the theoretical complexity of such solvers is usually conservative. Therefore, we present a stopping criterion that allows us to terminate early the primal iteration in the ALM while guaranteeing the precision of the primal update. This stopping condition is simple to check on embedded platforms and can be used for problems that are not necessarily strongly convex. We note that in this paper we do not provide a theoretical uniform upper bound on the optimal multiplier for general convex problems. Instead, our contribution is to develop a new projected ALM method that relies on such bounds to control data overflow on fixed-point platforms.

The rest of this paper is organized as follows. In Section 2, we formulate the problem and introduce necessary notations and lemmas needed to prove the main results. In Section 3, we characterize the convergence rate of the algorithm and present bounds on the primal suboptimality and infeasibility of the solution. In Section 4, we present the stopping criterion for the solution of the primal subproblem under fixed-point arithmetic. In Section 5, we present simulations that verify the theoretical analysis in the previous sections. In Section 6, we conclude the paper.

2 Preliminaries

We make the following assumptions on problem (1).

**Assumption 2.1** The function $f(x)$ is convex and is not necessarily differentiable. The problem (1) is feasible.

**Algorithm 1 Augmented Lagrangian Method**

**Require:** Initialize $\lambda_0 \in \mathbb{R}^p$, $k=0$

1: while $Ax_k \neq b$ do
2: $x_k = \arg \min_{x \in X} L_\rho(x; \lambda_k)$
3: $\lambda_{k+1} = \lambda_k + \rho (Ax_k - b)$
4: $k \leftarrow k + 1$
5: end while

The Lagrangian function of problem (1) is defined as $L(x; \lambda) = f(x) + \langle Ax - b, \lambda \rangle$ and the dual function is defined as $\Phi(\lambda) = \min_{x \in X} L(x; \lambda)$, where $\lambda \in \mathbb{R}^p$ is the Lagrangian multiplier (Ruszczynski, 2006) and $\langle \cdot, \cdot \rangle$ is the inner product between two vectors. Then the dual problem associated with problem (1) can be defined as $\max_{\lambda \in \mathbb{R}^p} \Phi(\lambda)$. Suppose $x^\star$ is an optimal solution of the primal problem (1) and $\lambda^\star$ is an optimal solution of the dual problem. Then, we make the following assumption:

**Assumption 2.2** Strong duality holds for the problem (1). That is, $f(x^\star) = \Phi(\lambda^\star)$.

Assumption 2.2 implies that $(x^\star, \lambda^\star)$ is a saddle point of the Lagrangian function $L(x; \lambda)$ (Ruszczynski, 2006). That is, for all $x \in X$, $\lambda \in \mathbb{R}^p$,

$$L(x^\star; \lambda) \leq L(x^\star; \lambda^\star) \leq L(x; \lambda^\star).$$

(2)

The Augmented Lagrangian function of the primal problem (1) is defined by $L_\rho(x; \lambda) = L(x; \lambda) + \frac{\rho}{2} \|Ax - b\|^2$, where $\rho$ is the penalty parameter and $\| \cdot \|$ is the Euclidean norm of a vector. Moreover, we define the augmented dual function $\Phi_\rho(\lambda) = \min_{x \in X} L_\rho(x; \lambda)$. Note that $\Phi_\rho(\lambda)$ is always differentiable with respect to $\lambda$ and its gradient is $\nabla \Phi_\rho(\lambda) = Ax^\star - b$, where $x^\star = \arg \min_{x \in X} L_\rho(x; \lambda)$. Moreover, $\nabla \Phi_\rho(\lambda)$ is Lipschitz continuous with constant $L_\Phi = \frac{1}{\rho}$ (Rockafellar, 1976a). The ALM can be viewed as a gradient ascent method on the multiplier $\lambda$ with step size $\frac{1}{L_\Phi} = \frac{1}{\rho}$. We present the ALM in Algorithm 1. Discussion on the convergence of Algorithm 1 can be found in Rockafellar (1976a); Ruszczyński (2006) and the references therein. ALM converges faster than dual method when the problem is not strongly convex due to its smoothing effect on the dual objective function $\Phi(\lambda)$, (Rockafellar, 1976a).

In practice, Algorithm 1 cannot be implemented exactly on a fixed-point platform. We present the modified ALM in Algorithm 2 to include fixed-point arithmetic errors. In Algorithm 2, $x_k^\star = \arg \min_{x \in X} L_\rho(x; \lambda_k)$. The effect of the fixed-point arithmetic is incorporated in the error terms $e_k^\epsilon$ and $e_k^\zeta$. Moreover, $K_{\text{out}}$ is the number of iterations in the ALM and $L$ is the step size used in the dual update and is defined in Lemma 3.2. Finally, $D$ is a convex and compact set containing at least one optimal multiplier $\lambda^\star$ and $\Pi_D$ denotes the projection onto the set $D$. This projection step is the major difference that makes the analysis in this paper different from other works on inexact ALM, e.g., Necoara et al. (2017); Nedelcu et al. (2014). The set $D$ is predetermined before running the
Algorithm 2 Augmented Lagrangian Method under inexactness

Require: Initialize $\lambda_0 \in \mathbb{R}^p$, $k=0$
1: while $k \leq K_{\text{out}}$ do
2: $\bar{x}_k \approx \arg \min_{x \in \mathcal{X}} L_p(x; \lambda_k)$ so that:
3: $L_p(\bar{x}_k, \lambda_k) - L_p(x_k^*, \lambda_k) \leq \epsilon^{k}_n$
4: $\lambda_{k+1} = \Pi_D[\lambda_k + \frac{1}{L}(A\bar{x}_k - b + \epsilon^{k}_{\text{out}})]$
5: $k \leftarrow k + 1$
end while

algorithm. We make the following assumption on $D$:

Assumption 2.3 The set $D$ is a box that contains 0, $2\lambda^*$ and $\lambda^* + 1$, where 1 is a vector of appropriate dimension and its entries are all 1.

The above choice of $D$ is discussed in more details in Section 3. Note that this set $D$ depends on a uniform bound on $\|\lambda^*\|$ over a set of problem data. The methods proposed in Devolder, Glineur & Nesterov (2012); Mangasarian (1985); Nedić & Ozdaglar (2009); Ruszczynski (2006) establish such bounds on $\|\lambda^*\|$ for fixed problem parameters. On the other hand, Patrinos & Bemporad (2014); Richter, Morari & Jones (2011) provide uniform bounds on $\|\lambda^*\|$ assuming $b \in B$ in the constraints in problem (1). However, the methods in Patrinos & Bemporad (2014); Richter et al. (2011) can only be applied to quadratic problems. For general problems considered in this paper, the interval analysis method in Hansen & Walster (1993) can be used to estimate a uniform bound on $\|\lambda^*\|$. However, the interval analysis method requires restrictive assumptions on the set of problem data and gives impractical bounds (Jerez, Constantinides & Kerrigan, 2015). In practice, we can approximate these bounds using sampling and then apply an appropriate scaling factor. Our contribution in this paper is to develop a new projected AML method that employs such bounds to control data overflow on fixed-point platforms.

3 Convergence Analysis

In this section, we show convergence of Algorithm 2 to a neighborhood of the optimal dual and primal objective value. First, we make some necessary assumptions on the boundedness of errors appearing in the algorithm.

Assumption 3.1 At every iteration of Algorithm 2, the errors are uniformly bounded, i.e.,

$$0 \leq \epsilon^{k}_n \leq B_{\text{in}}, \quad \|\epsilon^{k}_{\text{out}}\| \leq B_{\text{out}}, \text{ for all } k.$$

This assumption is satisfied by selecting appropriate data precision and subproblem solver parameters. Specifically, $\epsilon^{k}_n \geq 0$ if $\bar{x}_k$ is always feasible, that is, $\bar{x}_k \in \mathcal{X}$, which is possible if, e.g., the projected gradient method is used to solve the subproblem in line 2 in Algorithm 2. Due to the projection operation in line 3 of Algorithm 2, we have that for $\forall \lambda_1, \lambda_2 \in D_\delta$, $\|\lambda_1 - \lambda_2\| \leq B_{\lambda}$, where $D_\delta = D \cup \{\lambda \in \mathbb{R}^m : \lambda_k = \lambda + \epsilon^{k}_{\text{out}} \text{ for all } \lambda \in D\}$. Since $D$ is compact and $\|\epsilon^{k}_{\text{out}}\|$ is bounded according to Assumption 3.1, $D_\delta$ is also compact. $D_\delta$ is only introduced for the analysis. In practice, we only need to compute the size of $D$ to implement Algorithm 2.

3.1 Inexact Oracle

Consider the concave function $\Phi_p(\lambda)$ with $L_\Phi$-Lipschitz continuous gradient. For any $\lambda_1$ and $\lambda_2 \in \mathbb{R}^p$, we have

$$0 \geq \Phi_p(\lambda_2) - \Phi_p(\lambda_1) + (\nabla \Phi_p(\lambda_1), \lambda_2 - \lambda_1) \geq - \frac{L_\Phi}{2} \|\lambda_1 - \lambda_2\|^2.$$

Recall the expression of $\nabla \Phi_p(\lambda)$. Since step 2 of Algorithm 2 can only be solved approximately, $\nabla \Phi_p(\lambda)$ can only be evaluated inexactly. Therefore, the above inequalities can not be satisfied exactly. We extend the results in Devolder et al. (2014); Necoara et al. (2017) and propose the following inexact oracle to include the effect of $\epsilon^{k}_{\text{out}}$.

Lemma 3.2 (Inexact Oracle) Let assumptions 2.1, 2.2 and 3.1 hold. Moreover, consider the approximations $\Phi_{s,L}(\lambda_k) = L_p(\bar{x}_k; \lambda_k) + B_{\text{out}}B_\lambda$ to $\Phi_p(\lambda_k)$ and $s_{s,L}(\lambda_k) = A\bar{x}_k - b + \epsilon^{k}_{\text{out}}$ to $\nabla \Phi_p(\lambda_k)$. Then these approximations constitute a $(\delta, L)$ inexact oracle to the concave function $\Phi_p(\lambda)$ in the sense that, for all $\lambda \in D_\delta$,

$$0 \geq \Phi_p(\lambda) - (\Phi_{s,L}(\lambda_k) + (s_{s,L}(\lambda_k), \lambda - \lambda_k)) \geq - \frac{L}{2} \|\lambda - \lambda_k\|^2 - \delta,$$

where $L = 2L_\Phi = \frac{4}{\rho}$ and $\delta = 2B_{\text{in}} + 2B_{\text{out}}B_\lambda$.

PROOF. The proof is similar to Necoara et al. (2017) and therefore is omitted.

3.2 Dual Suboptimality

Showing the convergence of the dual variable is similar to showing the convergence of the projected gradient method with the inexact oracle used in Patrinos et al. (2015). Therefore, we omit the proof and directly summarize the dual suboptimality results of our algorithm.

Specifically, we have the following inequality:

$$\Phi_p(\lambda^*) - \Phi_p(\lambda_{k+1}) \leq \frac{L}{2} \|\lambda_k - \lambda^*\|^2 - \|\lambda_k - \lambda_{k+1} - \lambda^*\|^2 + \delta.$$

Furthermore, similar to the Theorem 5 in Patrinos et al. (2015), we have the following convergence result for the dual variable:

Theorem 3.3 Let assumptions 2.1, 2.2 and 3.1 hold. Define $\lambda_K = \frac{1}{2\rho} \sum_{k=1}^{K} \lambda_k$. Then, we have

$$\Phi_p(\lambda^*) - \Phi_p(\lambda_K) \leq \frac{L}{2\rho} \|\lambda_0 - \lambda^*\|^2 + \delta.$$
3.3 Primal Infeasibility and Suboptimality

Define the Lyapunov/Merit function $\phi^k(\lambda) = \frac{k}{2} \|\lambda_k - \lambda\|^2 + \frac{1}{2} \|\lambda_k - \lambda\|^2 - \lambda^*\lambda$. Also define the residual function $r(x) = Ax - b$. We have the following intermediate result:

**Lemma 3.4** Let assumptions 2.1, 2.2 and 3.1 hold. For all $k \geq 0$ and for all $\lambda \in D$, we have that

$$f(\tilde{x}_k) - f(x^*) + \langle \lambda, r(\tilde{x}_k) \rangle \leq \phi^k(\lambda) - \phi^{k+1}(\lambda) + E,$$

where $E = (1 + \frac{1}{2}) B_\lambda B_{\text{out}} + (1 + \frac{1}{2}) B_{\text{in}} + (\frac{1}{2} + \frac{1}{2\gamma}) B_{\text{out}}^2$. 

**PROOF.** We recall that $\tilde{x}_k$ is a suboptimal solution as defined in line 2 of Algorithm 2 that satisfies $L_p(\tilde{x}_k; \lambda_k) - L_p(x^*_k; \lambda_k) \leq \epsilon^*_k$. Moreover, due to the optimality of $x^*_k$ we also have that $L_p(x^*_k; \lambda_k) \leq L_p(x^*_k; \lambda_k) = f(x^*)$, where $x^*$ is the optimal solution to problem (1). The equality is because $r(x^*) = Ax^* - b = 0$. Combining these two inequalities, we have that $L_p(\tilde{x}_k; \lambda_k) - f(x^*) \leq \epsilon^*_k$. Expanding $L_p(\tilde{x}_k; \lambda_k)$ and rearranging terms, we get

$$f(\tilde{x}_k) - f(x^*) \leq -\langle \lambda, r(\tilde{x}_k) \rangle - \frac{\rho}{2} \|r(\tilde{x}_k)\|^2 + \epsilon^*_k.$$

Adding $\langle \lambda, r(\tilde{x}_k) \rangle$ to both sides of the above inequality, we get

$$f(\tilde{x}_k) - f(x^*) + \langle \lambda, r(\tilde{x}_k) \rangle \leq -\frac{\rho}{2} \|r(\tilde{x}_k)\|^2 + \epsilon^*_k. \quad (6)$$

In what follows, we show that the right hand side of (6) is upper bounded by $\phi^k(\lambda) - \phi^{k+1}(\lambda) + E$. First, we focus on the term $\langle \lambda, r(\tilde{x}_k) \rangle$. For all $\lambda \in D$, we have that $\|\lambda_k - \lambda\|^2 = \Pi_{D}(\lambda_k + \frac{1}{\rho} r(\tilde{x}_k) + \epsilon^*_k) - \|\lambda_k\|^2 \leq \|\lambda_k - \lambda\|^2 + \frac{1}{\rho} \|r(\tilde{x}_k)\|^2 + \epsilon^*_k$. Applying Assumption 2.1 and the Cauchy-Schwarz inequality, we obtain

$$\langle \lambda, r(\tilde{x}_k) \rangle \leq \frac{L}{2} (\|\lambda_k - \lambda\|^2 - \|\lambda_k - \lambda\|^2) \quad (7)$$

$$+ B_{\text{out}} B_\lambda + \frac{1}{2L} \|r(\tilde{x}_k)\|^2 + \frac{1}{2L} \|r(\tilde{x}_k)\|^2.$$ 

To upper bound the term $\langle \lambda, r(\tilde{x}_k) \rangle$ in (7), first we add and subtract $\epsilon^*_k$ from $r(\tilde{x}_k)$, and rearrange terms to get $\langle \lambda_k, \epsilon^*_k \rangle - \|\epsilon^*_k\|^2$. Defining $\lambda_{\text{in}} = \lambda_k + \epsilon^*_k$, recalling the definition of $s(\lambda, \lambda_k, \lambda)$ in Lemma 3.2, and ignoring the term $-\|\epsilon^*_k\|^2$, we obtain $\langle \lambda_{\text{in}}, r(\tilde{x}_k) \rangle \leq \langle s(\lambda, \lambda_k, \lambda) \rangle, \lambda_{\text{in}} - \lambda_k \rangle$. We apply the second inequality in Lemma 3.2 to upper bound $s(\lambda, \lambda_k, \lambda_{\text{in}} - \lambda_k)$. In order to apply Lemma 3.2, both $\lambda_{\text{in}}$ and $\lambda_{\text{in}} - \lambda_k$ need to belong to $D_k$. Due to the projection in line 3 in Algorithm 2, $\lambda_{\text{in}}$ always belongs to $D$. Recalling the definition of $D_k$, it is straightforward to verify that $\lambda_k, \lambda_{\text{in}} \in D_k$. Thus applying Lemma 3.2 we get

$$\langle \lambda_{\text{in}}, r(\tilde{x}_k) \rangle \leq \langle s(\lambda_{\text{in}}, \lambda_k, \lambda) \rangle, \lambda_{\text{in}} - \lambda_k \rangle \leq \Phi(\lambda_{\text{in}}) - \Phi(\lambda_k) + \frac{L}{2} B_{\text{out}}^2.$$ 

Since $\lambda^*$ is the global maximizer of the function $\Phi(\lambda)$, we get $\Phi(\lambda^*) \geq \Phi(\lambda_k)$. We can also show that $\Phi(\lambda_{\text{in}}) \geq \Phi(\lambda_k)$ always holds because $\Phi(\lambda_{\text{in}}) = L_p(\tilde{x}_k; \lambda_{\text{in}}) - B_{\text{out}} B_\lambda \geq L_p(\tilde{x}_k; \lambda_k)$. Combining these two inequalities we obtain that $\Phi(\lambda^*) - \Phi(\lambda_{\text{in}}) \geq \Phi(\lambda_k) - \Phi(\lambda_{\text{in}})$. Substituting this inequality into (8), we have that

$$\langle \lambda, r(\tilde{x}_k) \rangle \leq -\frac{L}{2} (\|\lambda_k - \lambda\|^2 - \|\lambda_k - \lambda\|^2) + \frac{1}{2} \|r(\tilde{x}_k)\|^2 \quad (9)$$

$$+ B_{\text{out}} B_\lambda + \frac{1}{2L} B_{\text{out}}^2 + \frac{2}{L} \delta.$$ 

Combining (9) with (6), we have that $f(\tilde{x}_k) - f(x^*) + \langle \lambda, r(\tilde{x}_k) \rangle \leq \frac{1}{2} (\|\lambda_k - \lambda\|^2 - \|\lambda_k - \lambda\|^2) + \frac{1}{2} \|r(\tilde{x}_k)\|^2$.

Theorem 3.5 (Primal Suboptimality and Infeasibility)

**Let assumptions 2.1, 2.2 and 3.1 hold. Define $\tilde{x}_K = \frac{1}{K} \sum_{k=1}^{K} \tilde{x}_k$. Then, we have that (a) primal optimality:**

$$-\frac{1}{K} \phi^1(2\lambda^*) + E \leq f(\tilde{x}_K) - f(x^*) \leq \frac{1}{K} \phi^1(0) + E,$$

(b) primal feasibility:

$$\|r(\tilde{x}_K)\| \leq \frac{1}{K} \phi^1(\lambda^* + r(\tilde{x}_K) \|r(\tilde{x}_K)\|) + E.$$ 

**PROOF.** Summing inequality (5) in Lemma 3.4 for $k = 1, 2, \ldots, K$, we have that $\sum_{k=1}^{K} f(\tilde{x}_k) - K f(x^*) + \langle \lambda, \sum_{k=1}^{K} r(\tilde{x}_k) \rangle \leq \phi^1(\lambda) - \phi^{K+1}(\lambda) + KE \leq \phi^1(\lambda) + KE$, where the second inequality follows from $\phi^k(\lambda) \geq 0$. Dividing both sides of the above inequality by $K$ and using the fact that $\frac{1}{K} \sum_{k=1}^{K} \lambda_k = r(\tilde{x}_K)$, we get

$$\|r(\tilde{x}_K)\| \leq \frac{1}{K} \sum_{k=1}^{K} f(\tilde{x}_k) - f(x^*) + \langle \lambda, r(\tilde{x}_K) \rangle.$$ 

From the convexity of the function $f(x)$, we have that

$$f(\tilde{x}_K) \leq \frac{1}{K} \sum_{k=1}^{K} f(\tilde{x}_k).$$

Combining the above two inequalities, we obtain

$$f(\tilde{x}_K) - f(x^*) + \langle \lambda, r(\tilde{x}_K) \rangle \leq \frac{1}{K} \phi^1(\lambda) + E.$$ 

(12)
To prove the right inequality in (10), let $\lambda = 0$ in (12). Then, we have that $f(\bar{x}_e) - f(x^*) \leq \frac{1}{\rho} \phi^1(0) + E$. To prove the left inequality in (10), according to the inequality (2) and Assumption 2.2, we have that

$$f(x^*) - f(\bar{x}_e) \leq \langle \lambda^*, r(\bar{x}_e) \rangle. \tag{13}$$

Adding $(\lambda^*, r(\bar{x}_e))$ to both sides of (13) and rearranging terms, we have that

$$(\lambda^*, r(\bar{x}_e)) \leq f(\bar{x}_e) - f(x^*) + (2\lambda^*, r(\bar{x}_e)). \tag{14}$$

Combining inequalities (13) and (14), we obtain $f(x^*) - f(\bar{x}_e) \leq f(\bar{x}_e) - f(x^*) + (2\lambda^*, r(\bar{x}_e))$. Combining this inequality and the inequality in (12) with $\lambda = 2\lambda^*$, we get $f(x^*) - f(\bar{x}_e) \leq \frac{1}{\rho} \phi^1(2\lambda^*) + E$.

To prove (11), let $\lambda = \lambda^* + \frac{r(\bar{x}_e)}{\|r(\bar{x}_e)\|}$ in (12), which also belongs in $D$ by Assumption 2.3. Rearranging terms, we obtain $f(\bar{x}_e) - f(x^*) + (\lambda^*, r(\bar{x}_e)) + \|r(\bar{x}_e)\| \leq \frac{1}{\rho} \phi^1(\lambda^* + \frac{r(\bar{x}_e)}{\|r(\bar{x}_e)\|}) + E$. Letting $x = \bar{x}_e$ in the second inequality in (2), we have that $f(\bar{x}_e) + (\lambda^*, r(\bar{x}_e)) - f(x^*) \geq 0$. Combining this inequality with the above inequality, we get $\|r(\bar{x}_e)\| \leq \frac{1}{\rho} \phi^1(\lambda^* + \frac{r(\bar{x}_e)}{\|r(\bar{x}_e)\|}) + E$, which completes the proof. \(\square\)

4 Fixed-point Implementation

To apply the Algorithm 2 to solve the problem (1) on fixed-point platforms, we need another assumption.

**Assumption 4.1** The set $\mathcal{X}$ is compact and simple.

The compactness is needed to bound the primal iterates $\bar{x}_k$ in Algorithm 2. $\lambda^*$ is simple in the sense that computing the projection onto $\mathcal{X}$ is of low complexity on fixed-point platforms, for example, $\lambda^*$ is a box. We note that our convergence analysis in Section 3 does not rely on the Assumption 4.1. This assumption is needed only for fixed-point implementation. Due to the projection operation in the multiplier update in Algorithm 2, $\|\lambda^*_k\|$ is always bounded. Since now both $x_k$ and $\lambda^*_k$ lie in compact sets, the bounds on the norms of these variables can be used to design the word length of the fixed-point data to avoid overflow, similar to Section 5.3 in Patrinos et al. (2015). Meanwhile, according to the results in (10) and (11), we can choose the number $K_{out}$ of the outer iterations, the accuracy of the multiplier update $B_{out}$ and the accuracy of the subproblem solution $B_{in}$ in Algorithm 2 to achieve an $\varepsilon$-solution to the problem (1).

Specifically, to achieve the accuracy $B_{out}$, we can determine the iteration complexity of the subproblem solver based on the studies in Devolder et al. (2014); Patrinos et al. (2015); Schmidt et al. (2011). However, these results are usually conservative in practice. Therefore, in this section, we propose a stopping criterion for early termination of the subproblem solver with guarantees on the solution accuracy. This stopping criterion can be used for box-constrained convex optimization problems that are not necessarily strongly convex, and is simple to check on embedded devices. A similar stopping criterion has been proposed in Nedelcu et al. (2014), which however requires the objective function of the subproblem to be strongly convex. In what follows, we relax the strong convexity assumption so that this stopping criterion can be applied.

**Lemma 4.2** Consider a convex and nonsmooth function $f(x)$, with support $\mathcal{X} \subset \mathbb{R}^n$ and set of minimizers $X^*$. Assume $f(x)$ satisfies the quadratic growth condition,

$$f(x) \geq f^* + \frac{\sigma}{2} \text{dist}^2(x, X^*), \quad \forall x \in \mathcal{X}, \tag{15}$$

where $\text{dist}(x, X^*)$ is the distance from $x$ to the set $X^*$, and $\sigma > 0$ is a scalar. Then, we have that for any $x \in \mathcal{X}$ and any $s \in N_\mathcal{X}(x),$

$$f(x) - f^* \geq \frac{\sigma}{2} \|\partial f(x) + s\|^2, \tag{16}$$

where $N_\mathcal{X}(x)$ is the normal cone at $x$ with respect to $\mathcal{X}$.

**PROOF.** The proof is the similar to the proof in Nedelcu et al. (2014) and therefore is omitted. \(\square\)

Necoara, Nesterov & Glineur (2018) have shown that the function $h(Ax)$ satisfies the inequality (15) when the set $\mathcal{X}$ is polyhedral and $h(y)$ is strongly convex. However, since the subproblem objective $L_p(x, \lambda_k)$ is in the form $\sum_i h_i(A_i x)$ rather than $h(Ax)$, in what follows we present a generalization of Theorem 8 in Necoara et al. (2018) so that the function $L_p(x, \lambda_k)$ satisfies the inequality (15) and the condition (16) can be applied.

**Lemma 4.3** Consider a convex function $H(x) = \sum_i h_i(A_i x - b_i)$ with the polyhedral support $\mathcal{X} = \{x : C x \leq d\}$, where $h_i(y)$ is $\sigma_i$-strongly convex with respect to $y$ for any $i$. Then we have

$$H(x) \geq H^* + \frac{\sigma}{2} \text{dist}^2(x, X^*), \tag{17}$$

where $\sigma = \min_{i \in [n]} \sigma_i$, the matrix $A$ is in the form $[\cdots | A_i^T | \cdots]^T$ and $\theta(A, C)$ is a constant only related to the matrices $\{A_i\}$ and $C$.

**PROOF.** Given $x^* \in X^*$, for any $x \in \mathcal{X}$, from the strong convexity of the function $h_i$, we have that $h_i(A_i x - b_i) \geq h_i(A_i x^* - b_i) + \langle \partial h_i(y)|_{y=A_i x^* - b_i}, x - x^* \rangle + \frac{\sigma_i}{2} \|A_i x - A_i x^*\|^2 = h_i(A_i x^* - b_i) + \langle x - x^*, A_i^T \partial h_i(y)|_{y=A_i x^* - b_i} \rangle + \frac{\sigma_i}{2} \langle x - x^* \rangle^T A_i^T A_i (x - x^*)$. Adding up the above inequalities for all $i$, we have $H(x) \geq H(x^*) + \sum_i \sigma_i \|x - x^*\|^2 A_i^T A_i (x - x^*) + \sum \frac{\sigma_i}{2} \langle x - x^* \rangle^T A_i^T A_i (x - x^*)$. Recall that $\partial H(x^*) = \sum_i \partial h_i(y)|_{y=A_i x^* - b_i}$. From the optimality of $x^*$, we have that $\sum \frac{\sigma_i}{2} \langle x - x^* \rangle^T A_i^T A_i (x - x^*) \geq 0$ for any $x \in \mathcal{X}$. Therefore, we obtain

$$H(x) \geq H(x^*) + \sum_i \sigma_i \langle x - x^* \rangle^T A_i^T A_i (x - x^*). \tag{18}$$

Next, we show that the optimal solution set $X^* = \{\tilde{x}^*: A \tilde{x}^* = t^*; C \tilde{x}^* \leq d\}$, where $t^* = A x^*$. To do so, we
decompose the space $\mathbb{R}^n$ into two mutually orthogonal subspaces, $\mathcal{V}$ and $\mathcal{U}$, where $\mathcal{V}$ is the intersection of the kernel spaces of all matrices $A_i$ and $\mathcal{U}$ is the union of the row space of all matrices $A_i$. Showing that $X^* = \{x^* : A^* x^* = t^*; C^* x^* \leq d\}$ is equivalent to showing that $X^* = \hat{X}^*$, where $X^* = \{x^* : \hat{x}^* = x^* + v, \text{ where } v \in \mathcal{V}; C^* x^* \leq d\}$. First, we show that $\hat{X}^* \subset X^*$. If $\hat{x}^* \in X^*$, we have that $\hat{x}^*$ is feasible and $H(x^*) = H(x^*)$, because $A_i \hat{x}^* = A_i(x^* + v) = A_i x^*$, for any $i$. Therefore, $X^* \subset \hat{X}^*$. Second, we prove that $\hat{X}^* \subset X^*$ by showing that if $\hat{x} \notin X^*$, then $\hat{x}$ is not optimal. If $C \hat{x} > d$, then $\hat{x}$ is not feasible. On the other hand, if $C \hat{x} \leq d$ but $\hat{x} = x^* + \alpha v + u$, where $\alpha$ is any real number and $u \in \mathcal{U}$, then $H(\hat{x}) > H(x^*)$. This is due to the inequality (18) and $\sum_i 2\|x^* - x^i\|A_i^T A_i(x^* - x^i) > 0$. Therefore, $X^* \subset \hat{X}^*$ and we have shown that $X^* = \hat{X}^*$. Since $X^* = \{x^* : A^* x^* = t^*; C^* x^* \leq d\}$, according to Lemma 15 in Wang & Lin (2014), for any $x \in X$, we have that dist$(x, X^*) \leq \theta(\hat{A}, C)|\hat{A}(x - x^*)|$. Furthermore, we have that dist$(x, X^*) \leq \theta(\hat{A}, C)|\hat{A}(x - x^*)|^2 \leq \theta^2(\hat{A}, C)\sum_i \sigma_i(x - x^i)^T A_i^T A_i(x - x^i)$. Combining this inequality with (18), we get the desired result (17). □

The constant $\theta(\hat{A}, C)$ can be obtained as in Necocorea, Nesterov & Glinsker (2018). If the objective function $f(x)$ in problem (1) is of the form $h(Ax - b)$ where $h(y)$ is strongly convex and the set $\mathcal{X} = \{x : l \leq x \leq u\}$, according to Lemma 4.3, the subproblem objective $L_p(x; \lambda_k)$ satisfies the inequality (15). Therefore, according to Lemma 4.2, we have that the iterate $x_k$ returned from the subproblem solver satisfies $L_p(x_k; \lambda_k) \leq B_k \|\nabla L_p(x_k; \lambda_k) + s^*_k\| \leq \sqrt{\frac{2}{\theta(\hat{A}, C)}}$, where $s^*_k = \arg \min_{s \in \mathcal{X}} \|\nabla L_p(x_k; \lambda_k) + s\|$. As discussed in Nedic, et al. (2014), $\|\nabla L_p(x_k; \lambda_k) + s^*_k\|$ can be efficiently evaluated on the evaluated platform.

5 Numerical Simulations

In this section, we present simulation results for a fairness constrained logistic regression problem to verify the theoretical results in Section 3 and the fixed-point implementation in Section 4. The simulations are conducted using the Fixed-Point Designer in Matlab R2020a on a Macbook Pro with 2.3GHz Quad-Core Intel Core i5. Specifically, we consider the dataset $\{d_i, z_i, y_i\}_{i=1:N}$, where $d_i \in \mathbb{R}^d$ is a vector of features at sample $i$, $z_i$ is a sensitive attribute associated with sample $i$, e.g., gender or age, and $y_i = \{+1, -1\}$ is the corresponding label. As shown in Zafar et al. (2019), to find a classifier with parameter $x$ that can both predict the label for a sample with feature $d_i$ and mitigate the correlation between the sample’s attribute and label, we can formulate the following optimization problem

$$\min_{x, \hat{x}} \frac{1}{N} \sum_{i=1}^{N} \log \left( \Pr(y_i | d_i, x) \right)$$  \hspace{1cm} (19)

s.t. $\frac{1}{N} \sum_{i=1}^{N} (z_i - \hat{y}_i) d_i^T x = c$, $x_i \leq x \leq x_u$, $c_i \leq c \leq c_u$,

where $\Pr(y_i | d_i, x) = 1 / (1 + \exp(-y_i d_i^T x))$, $\hat{y}$ is the mean of the attributes in the dataset, and $c$ represents the covariance between the labels and the attributes. In addition, $x_i, x_u, c_i$ and $c_u$ denote the bounds on these decision variables. \footnote{The logarithmic objectives as in (19) also appear in networked control problems studied, e.g., in Chang et al. (2015); Chatzipa, giotis et al. (2015); Chatzispanagiotis & Zavlanos (2016, 2017); Lee et al. (2018). In these networked control problems, fixed-point methods can improve the efficiency of solving the local subproblems as well as quantize the information shared among the agents to reduce communication overhead while maintaining the solution accuracy. Extension of the proposed fixed-point optimization method to this setting is part of our future research agenda.}

Table 1. Primal optimality and feasibility achieved by Algorithm 2 for Problem (19)

<table>
<thead>
<tr>
<th>$\epsilon$</th>
<th>$f/\mu$</th>
<th>low. (10),</th>
<th>opt.</th>
<th>up. (10)</th>
<th>feas.</th>
<th>(11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15-19</td>
<td>-0.7008</td>
<td>0.0218</td>
<td>0.6608</td>
<td>0.0198</td>
<td>0.9633</td>
</tr>
<tr>
<td>0.1</td>
<td>19-23</td>
<td>-0.0715</td>
<td>0.0071</td>
<td>0.0672</td>
<td>0.0019</td>
<td>0.9996</td>
</tr>
<tr>
<td>0.01</td>
<td>22-26</td>
<td>-0.0072</td>
<td>0.0045</td>
<td>0.0067</td>
<td>0.0004</td>
<td>0.1000</td>
</tr>
</tbody>
</table>

For our numerical simulations, we generate 10 datasets $\{d_i, z_i, y_i\}_i$. Each dataset contains 100 samples. The feature vectors $d_i \in \mathbb{R}^d$ and the true decision parameter $x^* \in \mathbb{R}^d$ in each dataset are sampled from a uniform distribution $[-1, 1]^d$. For a sample with the feature vector $d_i$, the corresponding label is $y_i = +1$ if $d_i^T x^* \geq 0$. Each sample is associated with a binary attribute $z_i = \{+1, -1\}$ with probability (0.5, 0.5). The bounds on $\|x^*\|$ for each problem instance were obtained by solving this problem using the Matlab function fmincon. Table 1 shows the achieved optimality and feasibility for the worst-case scenario and for three desired solution accuracies, $\epsilon = 1, 0.1$ and 0.01. The fractional and word lengths in Table 1 are chosen using the bounds (10) and (11) in Theorem 3.5. In Table 1, we observe that the optimality achieved by Algorithm 2 is bounded by the lower and upper bounds in (10), and the achieved feasibility is upper bounded by the bound in (11). Finally, notice
that the bounds in (10) and (11) are at most 50 times larger than the actual algorithm performance. This suggests that the bounds can be used in practice to design the proposed fixed-point implementation.

6 Conclusion
In this paper we proposed an Augmented Lagrangian Method to solve convex and non-smooth optimization problems using fixed-point arithmetic. To avoid data overflow, we introduced a projection operation in the multiplier update. Moreover, we present a stopping criterion to terminate early the primal subproblem iteration, while ensuring a desired accuracy of the solution. We presented convergence rate results as well as bounds on the optimality and feasibility gaps. To the best of our knowledge, this is the first fixed-point ALM that can handle non-smooth problems, data overflow, and can efficiently and systematically utilize iterative solvers in the primal update.

References