Pure and Applied Functional Analysis
Volume 3, Number 3, 2018, 417–428

Yokohama Publishers
ISSN 2189-3764 ONLINE JOURNAL

CONVERGENCE ANALYSIS OF AUGMENTED LAGRANGIAN - FAST PROJECTED GRADIENT METHOD FOR CONVEX QUADRATIC PROBLEMS

IGOR GRIVA

Abstract. The necessity of solving large scale nonlinear optimization problems stimulated recent development and application of first-order methods for nonlinear optimization. Numerical experiments demonstrated efficient convergence of the Augmented Lagrangian - Fast Projected Gradient Method (AL-FPGM) for such important applications as training support vector machines that represent large scale convex quadratic optimization problems with linear constraints and simple bounds. This paper provides theoretical analysis of the convergence of the AL-FPGM for solving this class of problems.

1. Introduction

Over the past thirty years, methods that require solving linear systems of equations such as interior-point methods (IPM) or exterior-point methods (EPM) have been used to solve quadratic programming problems (see e.g. [6,15]). Those methods are efficient for solving medium size problems up to a few thousands of variables.

Some modern applications, however, require solving nonlinear optimization problems with hundreds of thousands or even millions of variables. Such applications include positron emission tomography (see [1]) and machine learning methods based on training support vector machines that are used for classification and regression (see [16]).

In the past ten years a considerable amount of work has been done to analyze a family of fast gradient methods for unconstrained minimization (see e.g. [2]) that are related to optimal gradient methods introduced by Nesterov (see [8]). Fast gradient methods have laid a foundation in development of new algorithms for solving large optimization problems. For example, Polyak (see [9, 10]) analyzed convergence of a fast projected gradient method in which a trajectory of the fast gradient method is projected onto a box set.

Recently, Bloom [4] have suggested using a combination of the fast projected gradient method together with augmented Lagrangian framework (AL-FPGM) to

2010 *Mathematics Subject Classification.* 49, 90.

Key words and phrases. Quadratic problems, convex problems, machine learning, optimization, support vector machines, fast gradient.

train Support Vector Machines with tens of thousands data points. While the AL-FPGM demonstrated promising numerical results, theoretical convergence analysis of the AL-FPGM has been missing.

The main contribution of this paper is to provide a mathematical convergence justification for the AL-FPGM applied to convex quadratic optimization problems with linear constraints and simple bounds.

The remainder of this paper is organized as follows: Section 2 describes the convex quadratic optimization problem with linear constraints and simple bounds, Section 3 describes the augmented Lagrangian method, Section 4 describes the fast projected gradient method, Section 5 presents the convergence analysis for the AL-FPGM, and Section 6 presents the concluding remarks.

2. Quadratic convex optimization problem with linear constraints and simple bounds

Let us introduce the following functions

$$f(x) = \frac{1}{2}x^T Q x + q^T x,$$

$$g(x) = Ax - b,$$

where Q is an $n \times n$ positive semidefinite matrix, A is an $m \times n$ matrix, $q \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $m < n$. We assume that the system $Ax - b = 0$ is consistent.

Note that $g : \mathbb{R}^n \to \mathbb{R}^m$ is an m dimensional vector function. We emphasize the fact that Q does not have to be nonsingular, but it needs to be positive semidefinite so the resulting problem would be convex. We also introduce the bounded set

$$B = \{x \in \mathbb{R}^n : l_i \leq x_i \leq u_i,\ i = 1, \dots n\}.$$

Then the optimization problem that needs to be solved is as follows:

$$\begin{aligned} &\underset{x \in B}{\text{minimize}}\ f(x) \\ &\text{subject to}\ g(x) = 0. \end{aligned} \tag{2.1}$$

We can define the Lagrangian

$$L(x, \lambda) = f(x) - \lambda^T g(x),$$

and the augmented Lagrangian as follows

$$\mathcal{L}_k(x, \lambda) = f(x) - \lambda^T g(x) + \frac{k}{2} g(x)^T g(x),$$

where $\lambda \in \mathbb{R}^m$ is a vector of Lagrange multipliers that corresponds to the equality constraints and $k > 0$ is the scaling parameter.

The necessary and sufficient first-order optimality conditions x^* to be the solution for problem (2.1) is the existence of the dual vector $\lambda^* \in \mathbb{R}^m$ such that the pair (x^*, λ^*) such as $g(x^*) = 0$ satisfies the following variational inequality

$$\langle \nabla_x L(x^*, \lambda^*), x - x^* \rangle \geq 0, \quad \forall x \in B, \tag{2.2}$$

or, equivalently the following system for each $i = 1, \ldots, n$

$$\begin{array}{ll} \nabla_x L(x^*, \lambda^*)_i \geq 0 & \text{if} \quad x_i^* = l_i \\ \nabla_x L(x^*, \lambda^*)_i = 0 & \text{if} \quad l_i < x_i^* < b_i \\ \nabla_x L(x^*, \lambda^*)_i \leq 0 & \text{if} \quad x_i^* = u_i. \end{array} \tag{2.3}$$

3. Augmented Lagrangian method

The augmented Lagrangian method consists of a sequence of inexact minimizations of $\mathcal{L}_k(x, \lambda)$ in x on the B set

$$\hat{x} \approx \hat{x}(\lambda) = \underset{x \in B}{\operatorname{argmin}}\ \mathcal{L}_k(x, \lambda). \tag{3.1}$$

followed by updating the Lagrange multiplier:

$$\hat{\lambda} = \lambda - kg(\hat{x}).$$

We add the quadratic proximal term $\frac{1}{2k}(x-a)^T(x-a)$ to $\mathcal{L}_k$:

$$\mathcal{L}_k(x, a, \lambda) = f(x) - \lambda^T g(x) + \frac{k}{2} g(x)^T g(x) + \frac{1}{2k}(x-a)^T(x-a).$$

Adding the proximal term helps in two ways. First, it makes the augmented Lagrangian strongly convex with the modulus of at least $1/k$. As a result, the minimization on B (3.1) can be performed more efficiently. Second, for an iterative method used for solving (3.1), we can define an explicit stopping criteria for the unconstrained minimization based on the gradient of $\mathcal{L}_k$. On the other hand, if Q is nonsingular then adding the proximal term $\frac{1}{2k}(x-a)^T(x-a)$ does not spoil convergence and gives a way to estimate the strong convexity modulus, which will be larger than $1/k$.

For the stopping criteria, we use the following function that measures the violation of the first order optimality conditions for problem (3.1):

$$\mu(x, a, \lambda) = \max_{1 \leq i \leq m} \mu_i(x, a, \lambda), \tag{3.2}$$

where

$$\mu_i(x, a, \lambda) = \begin{cases} |(\nabla_x \mathcal{L}_k(x, a, \lambda))_i|, & \text{if } l_i < x_i < u_i, \\ \max\{0, -(\nabla_x \mathcal{L}_k(x, a, \lambda))_i\}, & \text{if } x_i = l_i, \\ \max\{0, (\nabla_x \mathcal{L}_k(x, a, \lambda))_i\}, & \text{if } x_i = u_i, \end{cases} \tag{3.3}$$

Note that the function $accur(x, a, \lambda) =: \max\{\mu(x, x, \lambda), \|g(x)\|\}$ measures the violation of the optimality conditions for the problem (2.1), and $accur(x, x, \lambda) = 0$ is equivalent to satisfying the first order optimality conditions.

Figure 1 describes the augmented Lagrangian method for solving problem (2.1).

The inexact minimization on the B set performed in Step 2 of Figure 1 is the most computationally expensive part of the augmented Lagrangian algorithm. Therefore, the efficiency of the AL methods depends on that of Step 2. We believe that an efficient minimization can be performed using a variant of the Fast Projected Gradient algorithm, which we describe in the next section.

1. Set $x \in B$, $\lambda = 0$, $rec = accur(x, x, \lambda)$.
   Select $k > 0$, $\epsilon > 0$, $0 < \theta < 1$, $\delta \geq 1$.
2. Find $\hat{x} \approx \underset{v \in B}{\operatorname{argmin}}\, \mathcal{L}_k(v, x, \lambda)$ with FPGM such that $\mu(\hat{x}, x, \lambda) \leq \epsilon/k$
3. Set $rec := accur(\hat{x}, x, \lambda)$
4. Find $\hat{\lambda} = \lambda - kg(\hat{x})$.
5. Set $x := \hat{x}$, $\lambda := \hat{\lambda}$, $\epsilon := \theta\epsilon$, $k := \delta k$.
6. If $rec > RequiredAccuracy$ then Goto 2.
7. Stop.

FIGURE 1. Boxed Augmented Lagrangian FPG Method

4. Fast projected gradient method

The fast projected gradient method (FPGM) requires estimation of the Lipschitz constant $L > 0$ of the gradient of $\mathcal{L}_k$ so that the inequality

$$\|\nabla_x \mathcal{L}_k(x_1, a, \lambda) - \nabla_x \mathcal{L}_k(x_2, a, \lambda)\| \leq L\|x_1 - x_2\|. \tag{4.1}$$

holds for any $x_1, x_2 \in \mathbb{R}^m$.

The gradient and the Hessian of $\mathcal{L}_k(x, a, \lambda)$ are as follows

$$\nabla_x \mathcal{L}_k(x, a, \lambda) = Qx - q - A^T(\lambda - k(Ax - b)) + \frac{1}{k}(x - a), \tag{4.2}$$

$$\nabla^2_{xx} \mathcal{L}_k(x, a, \lambda) = Q + kA^TA + \frac{1}{k}I_n, \tag{4.3}$$

where I_n is $n \times n$ identity matrix.

Since $\mathcal{L}_k$ is a quadratic form with respect to x, $L = \|\nabla^2_{xx} \mathcal{L}_k(x, a, \lambda)\| = \|Q + kA^TA + \frac{1}{k}I_n\|$ where the matrix spectral norm is the largest singular value of a matrix, i.e. the constant that depends only on Q, A and the parameter k.

Note that the matrix-vector products Qx and $A^T\lambda$ are the most computationally expensive parts for the $\nabla_x \mathcal{L}_k(x, a, \lambda)$ calculation, which takes $\mathcal{O}(n^2 + nm)$ basic arithmetic operations in case of dense matrices (Steps 3 and 7 in Figure 3). The projection operator $P_B : \mathbb{R}^m \to B$ (Step 3) is computationally inexpensive (see Figure 2) and requires only $\mathcal{O}(n)$ basic arithmetic operations. The other steps combined have less than a dozen arithmetic operations. Keeping in mind that $m < n$, one iteration of FPGM requires $\mathcal{O}(n^2)$ operations. Figure 3 describes the fast projected gradient method (FPGM) used in Step 2 of the augmented Lagrangian algorithm.

1. Loop over all $i = 1, \ldots, n$.
2. If $x_i < l_i$ then Set $x_i = l_i$
3. If $x_i > u_i$ then Set $x_i = u_i$
4. Return x.

FIGURE 2. Operator P_B : Projection of $x \in \mathbb{R}^m$ onto the set B

1. Input (x, λ), $v := x$.
2. Set $\bar{v} = v$, $t = 1$. Select $L > 0$.
3. Set $\hat{v} = P_B(v - \frac{1}{L}\nabla_v \mathcal{L}_k(v, x, \lambda))$
4. Set $\bar{t} = 0.5(1 + \sqrt{1 + 4t^2})$
5. Set $v = \hat{v} + (\hat{v} - \bar{v})(t - 1)/\bar{t}$
6. Set $\bar{v} = \hat{v}$, $t = \bar{t}$
7. If $\mu(\hat{v}, x, \lambda) > RequiredAccuracy$, Goto 3.
8. Output $\hat{v}$.

FIGURE 3. Fast Projected Gradient Method

5. Convergence analysis of AL-FPGM

Convergence analysis of AL-FPGM relies on convergence properties of the fast projected gradient method and the augmented Lagrangian method. The AL method falls into the category of the proximal-point algorithms that have been extensively investigated (see [12]). The FPGM is analyzed in more recent time (see [9, 10]). This paper shows how the parts of the theory on proximal-point and fast projected gradient methods work together.

Convergence of the general proximal-point algorithm is analyzed in [12]. We will show how the AL-FPGM described in Figures 1-3 can be viewed as a general proximal-point method.

Consider an elementary barrier function

$$I_B(x) = \begin{cases} 0, & \text{if} \quad x \in B, \\ +\infty, & \text{if} \quad x \notin B. \end{cases}$$

Then the objective function $\hat{f}$ in the extended form is

$$\hat{f}(x) = f(x) + I_B(x) = \begin{cases} f(x), & \text{if} \quad x \in B, \\ +\infty, & \text{if} \quad x \notin B. \end{cases}$$

The original problem (2.1) is equivalent to the following problem:

$$\begin{aligned} &\underset{x \in \mathbb{R}^n}{\text{minimize}} \ \hat{f}(x) \\ &\text{subject to } g(x) = 0. \end{aligned} \tag{5.1}$$

We define the corresponding Lagragian for problem (5.1) as

$$\hat{L}(x, \lambda) = \hat{f}(x) - \lambda^T g(x) = L(x, \lambda) + I_B(x),$$

and the corresponding augmented Lagrangian as

$$\hat{\mathcal{L}}_k(x, a, \lambda) = \hat{f}(x) - \lambda^T g(x) + 0.5k g(x)^T g(x) + \frac{1}{2k}(x - a)^T(x - a) =$$

$$\mathcal{L}_k(x, a, \lambda) + I_B(x),$$

where $\lambda \in \mathbb{R}^m$ is a vector of Lagrange multipliers that corresponds to the equality constraints and $k > 0$ is the scaling parameter.

Note that $\hat{f}(x)$, $\hat{L}(x,\lambda)$ and $\hat{\mathcal{L}}(x,a,\lambda)$ are convex lower semicontinous proper functions of x. As functions of x they are differentiable only in the interior of the B set, however subdifferentials exist for any $x \in \mathbb{R}^n$. For example,

$$\partial_x \hat{L}(x,\lambda) = \partial_x \hat{L}(x,\lambda)_1 \times \partial_x \hat{L}(x,\lambda)_2 \times \cdots \times \partial_x \hat{L}(x,\lambda)_n,$$

where

$$\partial_x \hat{L}(x,\lambda)_i = \begin{cases} \nabla_x L(x,\lambda)_i, & \text{if} \quad l_i < x_i < u_i \\ (-\infty, \nabla_x L(x,\lambda)_i], & \text{if} \quad x_i \le l_i, \\ [\nabla_x L(x,\lambda)_i, +\infty), & \text{if} \quad x_i \ge u_i. \end{cases}$$

For the primal-dual pair $z^* = (x^*, \lambda^*)$ to be the solution to problem (5.1), it is necessary and sufficient that (x^*, λ^*) satisfies the following conditions

$$0 \in \partial_x \hat{L}(x^*, \lambda^*), \quad g(x^*) = 0. \tag{5.2}$$

Consider a primal-dual pair $z = (x, \lambda)$, and the maximal monotone operator $T(z) = (\partial_x \hat{L}(x,\lambda), -\partial_\lambda \hat{L}(x,\lambda)) = (\partial_x \hat{L}(x,\lambda), g(x))$. Therefore problems (2.1) and (5.1) are equivalent to finding z^* such that

$$0 \in T(z^*).$$

We will show that the primal-dual sequence $z_s = (x_s, \lambda_s)$ generated by the AL algorithm in Figure 1 converges to z^*.

Note that the condition $0 \in \partial_x L(x^*, \lambda^*)$ is equivalent to (2.3), while the condition $0 \in \partial_\lambda L(x^*, \lambda^*)$ is equivalent to $g(x^*) = 0$ (see [14] for the discussion of the subdifferential and subgradients)

The augmented Lagrangian method for problem (5.1) consists of a sequence of inexact minimizations of $\hat{\mathcal{L}}_k(v,x,\lambda)$ in v on $\mathbb{R}^n$

$$\hat{x} \approx \hat{x}(x,\lambda) = \underset{v \in \mathbb{R}^n}{\operatorname{argmin}} \hat{\mathcal{L}}_k(v,x,\lambda). \tag{5.3}$$

followed by updating the Lagrange multiplier

$$\hat{\lambda} = \lambda - kg(\hat{x}).$$

To represent the stopping criteria for the unconstrained minimization $\mu(v,x,\lambda) \le \epsilon/k$ in terms of $\hat{\mathcal{L}}_k$ we need to consider the subdifferential of $\hat{\mathcal{L}}_k$ in detail:

$$\partial_v \hat{\mathcal{L}}_k(v,x,\lambda) = \partial_v \hat{\mathcal{L}}_k(v,x,\lambda)_1 \times \partial_v \hat{\mathcal{L}}_k(v,x,\lambda)_2 \times \cdots \times \partial_v \hat{\mathcal{L}}_k(v,x,\lambda)_n,$$

where

$$\partial_v \hat{\mathcal{L}}_k(v,x,\lambda)_i = \begin{cases} \nabla_v \mathcal{L}_k(v,x,\lambda)_i, & \text{if} \quad l_i < v_i < u_i \\ (-\infty, \nabla_v \mathcal{L}_k(v,x,\lambda)_i], & \text{if} \quad v_i \le l_i, \\ [\nabla_v \mathcal{L}_k(v,x,\lambda)_i, +\infty), & \text{if} \quad v_i \ge u_i. \end{cases}$$

Therefore

$$\mu_i(v,x,\lambda) = \operatorname{dist}(0, \partial_v \hat{\mathcal{L}}_k(v,x,\lambda)_i), \forall i = 1, ..., n,$$

and

$$\mu(v,x,\lambda) = \operatorname{dist}(0, \partial_v \hat{\mathcal{L}}_k(v,x,\lambda)), \tag{5.4}$$

where the distance between 0 and a set Y is defined as

$$\operatorname{dist}(0,Y) = \min_{y \in Y} \max_{1 \le i \le m} |y_i|$$

based on L_∞ norm.

Therefore the AL algorithms described in Figures 1- 2 can be reformulated in the equivalent form described in Figure 4. The main difference of the AL algorithm described here is the unconstrained minimization of $\hat{\mathcal{L}}_k(v,x,\lambda)$ in v in Step 2 performed instead of the minimization of $\mathcal{L}_k(v,x,\lambda)$ in v on the set B in Figure 1.

1. Set $x \in B$, $\lambda = 0$, $rec = accur(x,x,\lambda)$.
   Select $k > 0$, $\epsilon > 0$, $0 < \theta < 1$, $\delta \geq 1$.
2. Find $\hat{x} \approx \operatorname{argmin}_{v \in \mathbb{R}^n} \hat{\mathcal{L}}_k(v,x,\lambda)$ such that $\operatorname{dist}(0, \partial_v \hat{\mathcal{L}}_k(\hat{x},x,\lambda)) \leq \epsilon/k$ with FPGM
3. Set $rec := accur(\hat{x},x,\lambda)$
4. Find $\hat{\lambda} = \lambda - kg(\hat{x})$.
5. Set $x := \hat{x}$, $\lambda := \hat{\lambda}$, $\epsilon := \theta\epsilon$, $k := \delta k$.
6. If $rec > RequiredAccuracy$ then Goto 2.
7. Stop.

FIGURE 4. UnBoxed Augmented Lagrangian FPG Method

To establish convergence of the AL-FPGM algorithm described in Figure 4 we need the following lemmas.

Lemma 5.1. *The augmented Lagrangian method in Figure 4 is equivalent to the following proximal point method*

$$\text{Find } z_{p+1} : \operatorname{dist}(0, S_p(z_{p+1})) \leq \frac{\epsilon_p}{k}\|z_{p+1} - z_p\| \tag{5.5}$$

where $z_p = (x_p, \lambda_p)$, $\sum \epsilon_p < \infty$,

$$\begin{aligned} S_p(z) &= T(z) + k^{-1}(z - z_p), \\ T(z) &= \partial_z \hat{L}(z) = (\partial_x \hat{L}(x,\lambda), -\partial_\lambda \hat{L}(x,\lambda)), \\ \hat{L}(x,\lambda) &= \hat{f}(x) - \lambda^T g(x). \end{aligned}$$

Clearly, $\hat{L}(x,\lambda)$ *is the Lagrangian of problem (5.1) and* $T(z)$ *is its pseudo subdifferential.*

Proof. Let x^*_{p+1} denote an exact minimizer of $\hat{\mathcal{L}}_k(v,x_p,\lambda_p) = \hat{\mathcal{L}}_k(v,z_p)$ in v:

$$x^*_{p+1} = \operatorname*{argmin}_{v \in \mathbb{R}^n} \hat{\mathcal{L}}_k(v,z_p).$$

Then we have $\mu(x^*_{p+1}, x_p, \lambda_p) = 0$ and

$$0 \in \partial_x \hat{\mathcal{L}}_k(x^*_{p+1}, z_p) =$$

$$\partial_x \hat{f}(x^*_{p+1}) - \lambda_p^T \nabla g(x^*_{p+1}) + kg(x^*_{p+1})^T \nabla g(x^*_{p+1}) + \frac{1}{k}(x^*_{p+1} - x_p).$$

Using $\lambda^*_{p+1} = \lambda_p - kg(x^*_{p+1})$, the above expression can be rewritten as

$$\begin{aligned} 0 \in \partial_x \hat{\mathcal{L}}_k(x^*_{p+1}, z_p) &= \partial_x \hat{f}(x^*_{p+1}) - {\lambda^*_{p+1}}^T \nabla g(x^*_{p+1}) + \frac{1}{k}(x^*_{p+1} - x_p) \\ &= \partial_x \hat{L}(x^*_{p+1}, \lambda^*_{p+1}) + \frac{1}{k}(x^*_{p+1} - x_p) \end{aligned}$$

Therefore x^*_{p+1} is also the minimizer of $\hat{L}(x,\lambda^*_{p+1})+\frac{1}{2k}(x-x_p)^T(x-x_p)+C$ in x, where C is some constant.

Now consider the maximization of a smooth and strongly concave function

$$\max_{\lambda\in\mathbb{R}^m} d(\lambda),$$

where

$$d(\lambda)=\hat{L}(x^*_{p+1},\lambda)+0.5k(x^*_{p+1}-x_p)^T(x^*_{p+1}-x_p)-\frac{1}{2k}(\lambda-\lambda_p)^T(\lambda-\lambda_p)$$

is smooth and strongly concave in λ. Therefore the maximizer $\lambda_{\max}$ satisfies the following equation

$$\nabla d(\lambda_{\max})=-g(x_{p+1})^*-\frac{1}{k}(\lambda_{\max}-\lambda_p)=0.$$

Rewriting the above equation yields

$$\lambda_{\max}=\lambda_p-kg(x^*_{p+1}).$$

Therefore $\lambda^*_{p+1}=\lambda_{\max}$ and thus λ^*_{p+1} maximizes $d(\lambda)$ in λ.

To summarize, what we obtained so far is that for the function

$$Y_p(x,\lambda)=\hat{L}(x,\lambda)+\frac{1}{2k}(x-x_p)^T(x-x_p)-\frac{1}{2k}(\lambda-\lambda_p)^T(\lambda-\lambda_p),$$

which is strongly convex in x and strongly concave in λ we have

$$Y_p(x^*_{p+1},\lambda^*_{p+1})=\min_{x\in\mathbb{R}^n} Y_p(x,\lambda^*_{p+1})=\max_{\lambda\in\mathbb{R}^m} Y_p(x^*_{p+1},\lambda)=\min_{x\in\mathbb{R}^n}\max_{\lambda\in\mathbb{R}^m} Y_p(x,\lambda).$$

In other words, the pair $(x^*_{p+1},\lambda^*_{p+1})$ is the solution of the minmax problem with the convex-concave objective function $Y_p(x,\lambda)$ and therefore

$$0\in\partial_x\hat{L}(x^*_{p+1},\lambda^*_{p+1})+\frac{1}{k}(x^*_{p+1}-x_p),$$

and

$$0\in-\partial_\lambda\hat{L}(x^*_{p+1},\lambda^*_{p+1})+\frac{1}{k}(\lambda_{p+1}-\lambda_p),$$

or, to combine the last two inclusions

$$\begin{aligned}0\in(\partial_x\hat{L}(z^*_{p+1}),-\partial_\lambda\hat{L}(z^*_{p+1}))+\frac{1}{k}(z^*_{p+1}-z_p)&=T(z^*_{p+1})+\frac{1}{k}(z^*_{p+1}-z_p)\\&=S_p(z^*_{p+1}).\end{aligned}$$

The above condition corresponds to $\mu(x^*_{p+1},x_p,\lambda_p)=0$.

Now we will show that the stopping condition for an inexact minimizer

$$\mu(x_{p+1},x_p,\lambda_p)\le\epsilon_p/k$$

equivalent to

$$\operatorname{dist}(0,S_p(z_{p+1}))\le\epsilon_p/k,$$

where $\epsilon_p=\epsilon\theta^p$.

We have already established (5.4). Since the smooth and strongly concave maximization of $Y_p(x_{p+1},\lambda)$ in λ is solved precisely by

$$\lambda_{p+1}=\lambda_p-kg(x_{p+1}),$$

then $\operatorname{dist}(0, S_p(z_{p+1})) = \operatorname{dist}(0, \partial_v \mathcal{L}_k(x_{p+1}, x_p, \lambda_p)$. Therefore an iteration of the augmented Lagrangian algorithm is equivalent to (5.5). Lemma 5.1 is proven. □

The following lemma establishes convergence of the FPGM to a minimizer on the B set and is proven in [9].

Lemma 5.2. *For the sequence $\{x_s\}$ generated by the FPGM in Figure 3 for a convex quadratic problem the following bound takes place*

$$\mathcal{L}_k(x_s, x, \lambda) - \mathcal{L}_k(x_{\min}(x, \lambda), x, \lambda) \leq \frac{2L\|x_0 - x_{\min}(x, \lambda)\|^2}{(s+1)^2} \tag{5.6}$$

where $L > 0$ is the Lipschitz constant mentioned earlier in the text and

$$x_{\min}(x, \lambda) = \operatorname*{argmin}_{v \in B} \mathcal{L}_k(v, x, \lambda).$$

Remark 5.3. Lemma 5.2 does not require for $\mathcal{L}(v, x, \lambda)$ to be strongly convex in v. Since the proposed algorithm has the proximal term added, the strong convexity of $\mathcal{L}(v, x, \lambda)$ in v allows us to establish a stronger result described in the following lemma.

Lemma 5.4. *For the sequence $\{x_s\}$ generated by the FPGM in Figure 3 the following bound takes place*

$$\|x_{s+1} - x_{\min}(x, \lambda)\| \leq C\|x_s - x_{\min}(x, \lambda)\| \leq C^s\|x_0 - x_{\min}(x, \lambda)\| \tag{5.7}$$

where $C = \sqrt{\frac{kL-1}{kL+1}} < 1$.

Proof. Let l be the strong convexity modulus constant. According to (4.3),

$$l \geq 1/k.$$

Keeping in mind that $L \geq l$, the steps size $t = 1/L$ satisfies the following inequalities

$$t = \frac{1}{L} = \frac{2}{2L} \leq \frac{2}{l+L},$$

$$kL \geq \frac{L}{l} \geq 1,$$

and

$$0 \leq 1 - t\frac{2lL}{l+L} \leq \frac{kL-1}{kL+1} < 1.$$

Therefore the bound (5.7) follows from Theorem 3 (see [10]). Lemma 5.4 is proven. □

Remark 5.5. Since the function $\mathcal{L}_k(v, x, \lambda)$ is strongly convex in v then for any $v \in B$ the following inequality holds:

$$\frac{l}{2}\|v - x_{\min}(x, \lambda)\|^2 \leq \mathcal{L}_k(v, x, \lambda) - \mathcal{L}_k(x_{\min}(x, \lambda), x, \lambda),$$

where l is the convexity modulus. Therefore keeping in mind (5.6) and the above inequality we have

$$\|x_{s+1} - x_{\min}(x, \lambda)\| \leq 2\sqrt{\frac{L}{l}}\frac{\|x_0 - x_{\min}(x, \lambda)\|}{s+1} \leq 2\sqrt{kL}\frac{\|x_0 - x_{\min}(x, \lambda)\|}{s+1} \tag{5.8}$$

Remark 5.6. Both (5.7) and (5.8) guarantee the strong convergence of the sequence generated by the FPGM to the minimizer on the B set:

$$\lim_{s\to\infty} x_s = x_{\min}(x,\lambda).$$

Therefore we have

$$\lim_{s\to\infty} \mu(x_s,x,\lambda) = 0,$$

and there exists a finite index $\bar{s}$ such that

$$\mu(x_{\bar{s}},x,\lambda) \le \epsilon/k$$

for any $\epsilon > 0$. Therefore Line 2 in the augmented Lagrangian method described in Figures 1 and 4 is well defined.

Equivalently, an iteration of the proximal-point method (5.5) described in Lemma 5.1 is also well defined and we are ready to formulate the main convergence result.

Theorem 5.7. *The sequence $\{(x_p,\lambda_p)\}$ generated by the AL-FPGM in Figure 1 has a unique cluster primal-dual pair (x^*,λ^*) that satisfies the first order optimality conditions*

$$\langle \nabla_x L(x^*,\lambda^*), x - x^* \rangle \ge 0 \quad \forall x \in B$$

and

$$g(x^*) = 0.$$

i.e. $\{(x_p,\lambda_p)\}$ converges to the optimal solution of problem (2.1) in the weak sense.

Proof. Consider the following claims.

1) The sequence $\{z_p\} = \{(x_p,\lambda_p)\}$ generated by the AL-FPGM is well defined in Lines 2 and 4. The primal sequence $\{x_p\}$ is well defined by Remark 5.6. Once x_p is available then λ_p is calculated by the explicit formulas in Line 4, so the dual sequence $\{\lambda_p\}$ is also well defined. Therefore by Lemma 5.1 the sequence $\{z_p\}$ generated in (5.5) is also well defined.

2) Since $\delta \ge 1$, the sequence of the scaling parameters $k_p = k_0\delta^p$ is strictly positive and increasing if $k_0 > 0$.

3) Since $0 < \theta < 1$, the sequence $\{\epsilon_p = \epsilon_0\theta^p\}$ satisfies $\sum_{p=0}^{\infty} \epsilon_p < \infty$.

4) The solution x^* to (2.1) exists since f is a continuous convex quadratic function. The interior of the B set is not empty, so the Slater's condition is satisfied. Therefore for the solution x^* there exists a vector of dual variables λ^* such that the first-order optimality conditions hold.

5) The above optimality condition is equivalent to the following optimality conditions for the problem (2.1)

$$0 \in (\partial_x \hat{L}(x^*,\lambda^*), -\partial_\lambda \hat{L}(x^*,\lambda^*))$$

or

$$0 \in T(z^*)$$

in the notations of Lemma 5.1.

6) The existence of the solution to the equivalent problem

$$\text{Find } z^* : 0 \in T(z^*)$$

guarantees the boundedness of the sequence $\{z_p\}$ generated in (5.5) (see [12]).

7) All the conditions of Theorem 1 in [12] are satisfied. Therefore the sequence $\{z_p\}$ generated in (5.5) converges to some z^* in a weak sense, i.e. there exists a unique cluster point z^* the satisfies

$$0 \in T(z^*).$$

The above inclusion also means that the primal-dual sequence generated by the AL-FPGM in Figures 1-3 converges to the prial-dual solution (x^*, λ^*) in a weak sense. Theorem 5.7 is proven. □

6. Concluding remarks

This paper demonstrates theoretical convergence of the Augmented Lagrangian - Fast Projected Gradient Method for solving convex quadratic optimization problems with linear constraints and simple bounds.

Such an algorithm is well suited for solving large scale optimization problems as was demonstrated with numerical experiments (see [4]). According to the numerical results, for large problems (tens of thousands of variables) AL-FPGM has the potential to outperform methods that solve linear systems of equations such as interior- or exterior-point methods as the latter require a dense system of a linear equations to be solved at each step.

As the size of optimization problems grows further, we expect that the first-order methods such as AL-FPGM will demonstrate an increasing practical importance. In the future, we are planning to generalize the algorithm for parallel computations in order to solve problems with hundreds of thousands of variables.

Acknowledgment

The author would like to thank the anonymous referee and Michael Libman for their helpful comments and suggestions that led to improvement of the paper.

References

[1] D. L. Bailey, D. W. Townsend, P. E. Valk and M. N. Maisey, *Positron Emission Tomography: Basic Sciences*, Secaucus, NJ: Springer-Verlag, 2005.
[2] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci. **2** (2009), 183–202.
[3] V. Bloom, I. Griva, B. Kwon and A.-R. Wolff, *Exterior-point method for support vector machines*, IEEE Trans. Neural Netw. **25** (2014), 1390–1393.
[4] V. Bloom, I. Griva and F. Quijada, *Fast projected gradient method for support vector machines*, Optimization and Engineering **17** (2016), 651–662.
[5] A. Frank and A. Asuncion, *UCI machine learning repository*, 2010. [Online]. Available: http://archive.ics.uci.edu/ml
[6] I. Griva and R. Polyak, *1.5-Q-superlinear convergence of an exterior-point method for constrained optimization*, J. Global Optim. **40** (2008), 679–695.
[7] M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appl. **4** (1969), 303–320.
[8] Y. Nesterov, *Introductory Lectures on Convex Optimization*, Kluwer, Dordrecht, 2004.
[9] R. A. Polyak, J. Costa and S. Neyshabouri, *Dual fast projected gradient method for quadratic programming*, Optimization Letters **7** (2013), 631–645.
[10] R. A. Polyak, *Projected gradient method for non-negative least square*, Contemporary Mathematics **636** (2015), 167–179.

[11] M. J. D. Powell, *A method for nonlinear optimization in minimization problems*, in: Optimization, R. Fletcher (ed.), Academic Press, New York, 1969, pp. 283–298.
[12] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim. **14** (1976), 887–898.
[13] R. T. Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res. **1** (1976), 97–116.
[14] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[15] R. J. Vanderbei and D. F. Shanno, *An interior-point algorithm for nonconvex nonlinear programming*, Comp.l Optim. Appl. **13** (1999), 231–252.
[16] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed, Springer, 2000.

Manuscript received December 31 2017
revised February 9 2018

Igor Griva
Department of Mathematical Sciences, George Mason University, VA, USA
E-mail address: igriva@gmu.edu