Contemporary Mathematics
Volume **636**, 2015
http://dx.doi.org/10.1090/conm/636/12735

Projected Gradient Method for Non-Negative Least Square

Roman A. Polyak

Abstract. The projected gradient (PG) method was introduced for convex optimization in the sixties. It has mainly theoretical value because even in case of linear constraints it requires at each step solving a quadratic programming (QP) problem. On the other hand, in case of simple constraints the PG method can be very efficient.

In this paper, we apply the PG method to non-negative least squares (NNLS). The NNLS is critical in a number of real world applications because often the underlying parameters represent quantities that cannot be negative. The NNLS problem plays a key role in statistical learning theory in general and in Support Vector Machines (SVM) in particular.

In contrast to active set and interior point methods, which for a long time were the main tools for solving NNLS, the PG does not require solving at each step a linear system of equations. It rather requires matrix by vector multiplication as the main operation per step. Therefore, the critical issue is the convergence rate of the PG methods. The purpose of this paper is to establish convergence rates and to estimate the complexity bounds for PG methods under various assumptions on the input data.

1. Introduction

Let $A : \mathbb{R}^n \to \mathbb{R}^m$ $(m >> n)$ be the LS matrix and $b \in \mathbb{R}^m$ be the right-hand side. The non-negative least square (NNLS) problem consists in finding

$$f^* = f(x^*) = \min \left\{ f(x) = \frac{1}{2}\|Ax - b\|^2 | x \in \mathbb{R}^n_+ \right\},$$

where $\|a\| = (a, a)^{\frac{1}{2}}$.

The NNLS is one of the main linear algebra problems, which has been studied for a long time. The research on NNLS was summarized in the classical monograph by C. Lawson and R. Hanson [**10**]. Since the 70s their active set method and its modifications (see for example [**3**]-[**4**]) were the main tools for solving NNLS.

The active set approach requires at each step solving a standard LS sub-problem, which is equivalent to solving a linear system of equations. Moreover, the combinatorial nature of the active set methods does not allow establishing meaningful bounds for the number of steps.

On the other hand, NNLS is a quadratic programming (QP) problem and can be solved by interior point methods (IPMs) (see, for example, [**2**],[**12**]) in polynomial time. In fact, it takes $O(n^3 \ln \varepsilon^{-1})$ operations to find an $\varepsilon > 0$ approximation for f^*

2010 *Mathematics Subject Classification.* Primary 65B99, 90C20, 90C25.

(see, for example, [**15**]). The IPMs also require solving a linear system of equations at each step, which for very large scale NNLS can be difficult or even impossible.

In this note, we apply the PG methods [**7**], [**11**] (see also [**5**], [**6**]) for NNLS. Instead of solving a linear system of equations , the PG at each step require matrix by vector multiplication. What is even more important is the fact that the PG methods have no combinatorial features, because the generated sequence is projected on the entire feasible set. It allows establishing both convergence rate and complexity bounds under various assumptions on the input data.

Particular emphasis will be given to the fast projected gradient (FPG), which is based on Yu. Nesterov's gradient mapping theory [**14**] and closely related to the Dual Fast Projected Gradient (DFPG) method for QP [**16**] (see also A. Beck and M. Teboulle's FISTA algorithm [**1**]).

The FPG requires $O(\lambda^{\frac{1}{2}}\|x_0 - x^*\|n^2\varepsilon^{-\frac{1}{2}})$ operations for finding $f(x_k) : \Delta_k = f(x_k) - f^* \leq \varepsilon$ where $\lambda = \text{maxeigeval}\, A^TA$, $\varepsilon > 0$ is the required accuracy and x_0 is the starting point. So, for large n, FPG has the potential to be an efficient alternative for IPMs. Moreover, matrix by vector multiplication is much cheaper than solving the same size system of linear equations and it admits fast parallel computations, which can substantially speed up the process and improve the complexity bound (see, for example, [**9**]).

The paper is organized as follows. In the next section, we recall some basic results. In Sections 3 and 4 we consider the PG and FPG methods. In Section 5, we consider the PG method for full rank NNLS. In Section 6, we discuss an opportunity of using FPG for solving SVM. We conclude this note by pointing out a few topics for further research.

2. Problem formulation and some preliminary results

To cover a wider class of application, we consider the LS problem under box constraints, i.e.,

$$f^* = f(x^*) = \min\left\{f(x) = \frac{1}{2}\|Ax - b\|^2 | x \in \Omega\|\right\}, \tag{2.1}$$

where $c \in \mathbb{R}^n_{++}$ and

$$\Omega = \{x \in \mathbb{R}^n : 0 \leq x_i \leq c_i, i = 1, .., n\}.$$

The gradient

$$\nabla f(x) = A^T(Ax - b) = Qx - q,$$

where $Q = A^TA : \mathbb{R}^n \to \mathbb{R}$ and $q = A^Tb \in \mathbb{R}^n$, satisfies the Lipschitz condition

$$\|\nabla f(x) - \nabla f(y)\| \leq \|Q\|\|x - y\|.$$

Therefore for any $L \geq \text{maxeigval}Q = \lambda$, we obtain

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \tag{2.2}$$

for any x and y from $\mathbb{R}^n$.

The Lipschitz constant $L > 0$ plays a key role in the PG theory; therefore, finding a close to $\lambda > 0$ upper bound is an important part of the PG methods. One can find the upper bound for $\lambda > 0$ by using the following power method.

For any $1 \leq i \leq n$, we have

$$\lambda = \lim_{s\to\infty} \frac{y_i^{s+1}}{y_i^s},$$

where $y^0 \in \mathbb{R}^n$ and $y^s = A^s y_0$. In other words, one can find the upper bound L for λ by using a few matrices by vector multiplication.

The optimality criteria, for any $x^* \in X^* = \operatorname{argmin}\{f(x)|x \in \Omega\}$, is given by the following inequality:

$$(\nabla f(x^*), X - x^*) \geq 0, \forall X \in \Omega. \tag{2.3}$$

For a given $x \in \mathbb{R}^n$, let us consider the following quadratic approximation of f

$$\psi_L(x, X) = f(x) + (X - x, \nabla f(x)) + \frac{L}{2}\|X - x\|^2.$$

There exists a unique minimizer

$$x_\Omega^L \equiv x_\Omega^L(x) = \operatorname{argmin}\{\psi_L(x, X)|X \in \Omega\}. \tag{2.4}$$

The optimality criteria for x_Ω^L is given by the following inequality:

$$(\nabla_X \psi_L(x, x_\Omega^L), X - x_\Omega^L) \geq 0, \forall X \in \Omega. \tag{2.5}$$

One obtains the solution x_Ω^L in (2.4) by solving n one dimensional problems

$$x_{i,\Omega}^L = \operatorname{argmin}\left\{\frac{\partial f(x)}{\partial x_i}(X_i - x_i) + \frac{L}{2}(X_i - x_i)^2 | 0 \leq X_i \leq c_i\right\}, i = 1, \ldots, n. \tag{2.6}$$

Using the optimality criteria (2.5) for $x_{i,\Omega}^L$ in (2.6), we obtain the following solution:

$$x_{i,\Omega}^L = \begin{cases} 0, & \text{if } x_i - \frac{1}{L}\frac{\partial f(x)}{\partial x_i} \leq 0 \\ x_i - \frac{1}{L}\frac{\partial f(x)}{\partial x_i}, & \text{if } 0 < x_i - \frac{1}{L}\frac{\partial f(x)}{\partial x_i} < c_i \\ c_i, & \text{if } x_i - \frac{1}{L}\frac{\partial f(x)}{\partial x_i} \geq c_i. \end{cases}$$

Therefore, the problem (2.4) admits the closed form solution

$$x_\Omega^L = P_\Omega(x - \frac{1}{L}\nabla f(x)), \tag{2.7}$$

where the projection of $u \in \mathbb{R}^n$ on Ω is defined as follows

$$P_\Omega u = \operatorname{argmin}\{\|u - v\||v \in \Omega\}.$$

3. Projected gradient method

Starting with $x_0 \in \mathbb{R}^n$ and reiterating (2.7), we obtain the projected gradient (PG) method

$$x_{s+1} = P_\Omega(x_s - L^{-1}\nabla f(x_s)) \tag{3.1}$$

for solving NNLS (2.1).

Due to (2.4), the PG method (2.7) reminds us of the linearization method introduced by B. Pschenichny [**17**] in the 70s. On the other hand, it has a flavor of Quadratic Prox (see [**8**]), which will play an important role in our considerations.

Due to the Lipschitz condition (2.2), for any pair $(X; x) \in \mathbb{R}^n \times \mathbb{R}^n$, we have

$$f(X) - f(x) - (X - x, \nabla f(x)) \leq \frac{L}{2}\|X - x\|^2.$$

Therefore,

$$f(X) \leq \psi_L(x, X) = f(x) + (X - x, \nabla f(x)) + \frac{L}{2}\|X - x\|^2.$$

The following Lemma is similar to Lemma 1 in [**16**] (see also Lemma 2.3 in [**1**]).

LEMMA 1. *For any given $x \in \mathbb{R}^n$ and $L > 0$ such that*

$$f(x_\Omega^L) \le \psi_L(x, x_\Omega^L), \tag{3.2}$$

the following inequality holds for any $X \in \Omega$

$$f(X) - f(x_\Omega^L) \ge \frac{L}{2}\|x_\Omega^L - x\|^2 + L(x - X, x_\Omega^L - x). \tag{3.3}$$

PROOF. From (3.2) and convexity f(x), we have

$$\begin{aligned} f(X) - f(x_\Omega^L) &\ge f(X) - \psi(x, x_\Omega^L) \\ &= f(X) - f(x) - (x_\Omega^L - x, \nabla f(x)) - \frac{L}{2}\|x_\Omega^L - x\|^2 \\ &\ge f(x) + (\nabla f(x), X - x) - f(x) - (x_\Omega^L - x, \nabla f(x)) - \frac{L}{2}\|x_\Omega^L - x\|^2 \\ &= \frac{L}{2}\|x_\Omega^L - x\|^2 + (\nabla f(x), X - x_\Omega^L) - L\|x_\Omega^L - x\|^2. \end{aligned} \tag{3.4}$$

From the optimality criteria (2.5) applied to (2.4), we obtain

$$(\nabla f(x) + L(x_\Omega^L - x), X - x_\Omega^L) \ge 0, \forall X \in \Omega,$$

or

$$(\nabla f(x), X - x_\Omega^L) \ge -L(x_\Omega^L - x, X - x_\Omega^L), \forall X \in \Omega. \tag{3.5}$$

Therefore combining (3.4) and (3.5), we obtain

$$\begin{aligned} d(X) - d(x_\Omega^L) &\ge \frac{L}{2}\|x_\Omega^L - x\|^2 - L(x_\Omega^L - x, X - x_\Omega^L) - L(x_\Omega^L - x, x_\Omega^L - x) \\ &= \frac{L}{2}\|x_\Omega^L - x\|^2 + L(x - X, x_\Omega^L - x), \forall X \in \Omega. \end{aligned}$$

The most costly part of PG method (3.1) is computing the gradient $\nabla f(x_s) = Qx_s - q$, which requires matrix by vector multiplication. It takes at most $O(n^2)$ operations.

The following theorem establishes the convergence of the PG method (3.1) and estimate the convergence rate.

THEOREM 1. *The PG method (3.1) converges in value and*

$$\Delta_k = f(x_k) - f^* \le \frac{L}{2k}\|x_0 - x^*\|^2.$$

PROOF. Let us consider (3.3) with $X = x^*, x = x_s$, and $x_\Omega^L = x_{s+1}$. Then we have

$$\begin{aligned} \frac{2}{L}(f(x^*) - f(x_{s+1})) &\ge \|x_{s+1} - x_s\|^2 + 2(x_s - x^*, x_{s+1} - x_s) \\ &= (x_{s+1}, x_{s+1}) - 2(x_{s+1}, x_s) + (x_s, x_s) + 2(x_s, x_{s+1}) \\ &\quad - 2(x^*, x_{s+1}) - 2(x_s, x_s) + 2(x^*, x_s) + (x^*, x^*) - (x^*, x^*) \\ &= \|x_{s+1} - x^*\|^2 - \|x_s - x^*\|^2. \end{aligned}$$

Summing up the last inequality from $s = 0$ to $s = k - 1$, we obtain

$$kf(x^*) - \sum_{s=0}^{k-1} f(x_{s+1}) \ge \frac{L}{2}[\|x^* - x_k\|^2 - \|x^* - x_0\|^2]. \tag{3.6}$$

Using (3.3) with $X = x = x_s$ and $x_\Omega^L = x_{s+1}$, we obtain

$$\frac{2}{L}(f(x_s) - f(x_{s+1})) \geq \|x_{s+1} - x_s\|^2,$$

or

$$s(f(x_s) - f(x_{s+1})) \geq \frac{L}{2} s\|x_{s+1} - x_s\|^2,$$

i.e.,

$$sf(x_s) - (s+1)f(x_{s+1}) + f(x_{s+1}) \geq \frac{L}{2} s\|x_{s+1} - x_s\|^2.$$

Summing up the last inequality from s=0 to $s = k-1$, we obtain

$$-kf(x_k) + \sum_{s=0}^{k-1} f(x_{s+1}) \geq \frac{L}{2} \sum_{s=0}^{k-1} s\|x_{s+1} - x_s\|^2. \tag{3.7}$$

From (3.6) and (3.7) follows

$$k(f(x^*) - f(x_k)) \geq \frac{L}{2}\left[\sum_{s=0}^{k-1} s\|x_{s+1} - x_s\|^2 + \|x^* - x_k\|^2 - \|x^* - x_0\|^2\right],$$

or

$$\Delta_k = f(x_k) - f(x^*) \leq \frac{L}{2k}\|x_0 - x^*\|^2. \tag{3.8}$$

It follows from (3.8) that for a given $\varepsilon > 0$, it takes $k = O(L\|x_0 - x^*\|^2\varepsilon^{-1})$ steps to get $\Delta_k \leq \varepsilon$. Matrix by vector multiplication requires at most $O(n^2)$ operations; therefore for the PG complexity bound, we obtain

$$\mathrm{Comp}(PG) = O(L\|x_0 - x^*\|^2 n^2 \varepsilon^{-1}). \tag{3.9}$$

It turns out that the PG complexity can be drastically improved practically without increasing numerical effort per step.

In the following section, we consider the fast projected gradient (FPG) method for NNLS. The FPG is based on Yu. Nesterov's gradient mapping theory [**15**] and closely related to DFPG [**16**] for QP and FISTA algorithm by A. Beck and M. Teboulle [**1**].

4. Fast Projected Gradient

At each step, FPG generates a predictor vector x_k and a corrector vector X_k. The predictor x_k is computed as an extrapolation of two successive correctors. One obtains the corrector X_k as a result of one PG step with x_k as a starting point.

FPG method

(1) Input: $L > 0$ the upper bound for the Lipschitz constant of the gradient ∇f.

$$0 < x_0 = x_1 < c$$

$$t_1 = 1$$

(2) Step k

a) using the predictor x_k we find the corrector

$$X_k = \mathrm{argmin}\{\psi(x_k, X) = f(x_k) + (X - x_k, \nabla f(x_k)) + \frac{L}{2}\|X - x_k\|^2 | X \in \Omega\};$$

b) update step length $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
c) find new predictor

$$x_{k+1} = X_k + \frac{t_k - 1}{t_{k+1}}(X_k - X_{k-1}).$$

The corrector

$$X_k = P_\Omega(x_k - \frac{1}{L}\nabla f(x_k))$$

is the new approximation for x^*.

In other words, the corrector X_k one obtains as a result of one step of the PG method for NNLS (2.1) with a starting point x_k and step-length L^{-1}. Although FPG does not require much extra work as compared with PG (3.1), the FPG has much better convergence rate.

Moreover, it is impossible to improve the FPG convergence rate in the class of gradient methods (see [**13**], [**14**]). In other words, the FPG is optimal in the class of gradient methods.

Let $\Delta_k = f(x_k) - f^*; y_k = t_k X_k + (t_k - 1)X_{k-1} - x^*$.

The following inequality (see Lemma 2.3 in [**1**] and (17) in [**16**]) is critical for the proof of FPG convergence rate.

LEMMA 2. *The following inequality holds:*

$$t_k^2\Delta_k - t_{k+1}^2\Delta_{k+1} \geq \frac{L}{2}[\|y_{y+1}\| - \|y_k\|^2]. \tag{4.1}$$

For completeness we will sketch the proof in Appendix 1.

THEOREM 2. *For the sequence $\{X_k\}_{k=1}$ generated by FPG (a)-(c), the following bound holds:*

$$\Delta_k \leq \frac{2L\|x_0 - x^*\|^2}{(k+2)^2}. \tag{4.2}$$

PROOF. First of all, from (b) follows $t_k \geq \frac{1}{2}(k+1), \forall k \geq 1$. It is obvious for $k = 1$. Assuming $t_k \geq \frac{1}{2}(k+1)$ from (b), we obtain

$$t_{k+1} = \frac{1}{2}(1 + \sqrt{1+4t_k^2}) \geq \frac{1}{2}(1 + \sqrt{1+(k+1)^2}) = \frac{1}{2}(k+2).$$

From (4.1), we have

$$\begin{aligned} t_{k+1}^2\Delta_{k+1} + \frac{L}{2}\|y_{k+1}\|^2 &\leq t_k^2\Delta_k + \frac{L}{2}\|y_k\|^2 \\ &\leq t_{k-1}^2\Delta_{k-1} + \frac{L}{2}\|y_{k-1}\|^2 \\ &\cdot \\ &\cdot \\ &\leq t_1^2\Delta_1 + \frac{L}{2}\|y_1\|^2. \end{aligned}$$

Keeping in mind $t_1 = 1$ and $y_1 = X_1 - x^*$, we obtain

$$t_{k+1}^2\Delta_{k+1} \leq t_1^2\Delta_1 + \frac{L}{2}\|y_1\|^2 \leq \Delta_1 + \frac{L}{2}\|X_1 - x^*\|^2. \tag{4.3}$$

Using again (3.3) with $X = x^*$, $x_\Omega^L = X_1$ and $x = x_0$, we obtain

$$f(x^*) - f(X_1) \geq \frac{L}{2}\|X_1 - x_0\|^2 + L(x_0 - x^*, X_1 - x_0) = \frac{L}{2}[\|X_1 - x^*\|^2 - \|x_0 - x^*\|^2].$$

Therefore

$$\Delta_1 = f(X_1) - f(x^*) \leq \frac{L}{2}\|x_0 - x^*\|^2 - \frac{L}{2}\|X_1 - x^*\|^2.$$

Adding the last inequality with (4.3), we obtain

$$t_{k+1}^2 \Delta_{k+1} \leq \frac{L}{2}\|x_0 - x^*\|^2.$$

Keeping in mind $t_{k+1} \geq k+2$ we obtain (4.2). It follows from (4.2) that for a given $\varepsilon > 0$, it takes $k = O(\sqrt{L}\|x_0 - x^*\|\varepsilon^{-\frac{1}{2}})$ steps to get $\Delta_k \leq \varepsilon$. Again each FPG step requires at most $O(n^2)$ operations; therefore for the FPG complexity, we obtain

$$\text{Comp}(FPG) = O(\sqrt{L}\|x_0 - x^*\|n^2\varepsilon^{-\frac{1}{2}}). \tag{4.4}$$

It follows from (4.4) that for large n in a number of instances the FPG provides an alternative for IPMs, for which the complexity bound is $O(n^3 \ln \varepsilon^{-1})$. Moreover, for large n solving at each step, a system of linear equations can drastically reduce IPMs efficiency. On the other hand, the FPG complexity can be improved by using fast and parallel computations for matrix by vector multiplication [**9**] .

In the following section, we show that if $\operatorname{rank} A = n$, then the bound (4.4) can be substantially improved.

5. Projected Gradient for full rank NNLS

If A is a full rank matrix, i.e., $\operatorname{rank} A = n$, then $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex and the gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is a strongly monotone operator, i.e., there exists $l > 0$:

$$(\nabla f(x) - \nabla f(y), x - y) \geq l\|x - y\|^2, \forall x, y \in \mathbb{R}^n; \tag{5.1}$$

and for $Q : \mathbb{R}^n \to \mathbb{R}^n$, we have

$$(Qx, x) \geq l\|x\|^2, \forall x \in \mathbb{R}^n.$$

We recall that the gradient ∇f satisfies Lipschitz condition (2.2).

The following inequality (see, for example, [**15**]) will be used later to prove the Q-linear convergence rate of the PG method for full rank NNLS.

LEMMA 3. *For a strongly convex function with modulus convexity $l > 0$ and Lipschitz continuous gradient ∇f with a constant $L > l$, the following inequality holds:*

$$(\nabla f(x) - \nabla f(y), x - y) \geq \frac{lL}{l+L}\|x - y\|^2 + \frac{1}{l+L}\|\nabla f(x) - \nabla f(y)\|^2. \tag{5.2}$$

We will sketch the proof in Appendix 2.

Now we need two basic properties of the projection on a convex set Ω.

First, iff

$$x^* = \operatorname{argmin}\{f(x) | x \in \Omega\},$$

then for any $t > 0$ we have

$$P_\Omega(x^* - t\nabla f(x^*)) = x^*. \tag{5.3}$$

Second, the operator $P_\Omega(x)$ is a continuous and nonexpansive, i.e., for any pair x and y from $\mathbb{R}^n$, we have

$$\|P_\Omega x - P_\Omega y\| \le \|x - y\|. \tag{5.4}$$

Obviously, f satisfies a Lipschitz condition on Ω, i.e., there is $L_0 > 0$ such that the following inequality

$$|f(x) - f(y)| \le L_0\|x - y\| \tag{5.5}$$

holds for any x and y from Ω. The projected gradient method is defined by the formula

$$x_{s+1} = P_\Omega(x_s - t\nabla f(x_s)). \tag{5.6}$$

The convergence rate and PG complexity establishes the following Theorem.

THEOREM 3. *If* $\operatorname{rank} A = n$, *then*

(1) *for* $0 < t < 2/(l + L)$ *the following bound holds:*

$$\|x_{s+1} - x^*\|^2 \le \left(1 - t\frac{2lL}{l + L}\right)\|x_s - x^*\|^2; \tag{5.7}$$

(2) *for* $t = 2/(l + L)$ *we have*

$$\|x_{s+1} - x^*\| \le \left(\frac{1 - \varkappa}{1 + \varkappa}\right)\|x_s - x^*\|, \tag{5.8}$$

where $0 < \varkappa = l/L < 1$ *is the condition number of the matrix* $Q = A^TA$;

(3) *for* $t = 2/(l + L)$ *the following bound holds:*

$$f(x_k) - f(x^*) \le L_0\left(\frac{1 - \varkappa}{1 + \varkappa}\right)^k\|x_0 - x^*\|; \tag{5.9}$$

(4) *let* $\varepsilon > 0$ *be the given accuracy, then the complexity of the PG method* (5.6) *is given by the following formula*

$$\operatorname{Comp}(PG) = O(n^2\varkappa^{-1}\ln\varepsilon^{-1}). \tag{5.10}$$

PROOF. First of all, we recall that

$$x^* = P_\Omega(x^* - t\nabla f(x^*)), \forall t \ge 0.$$

Therefore, in view of (5.4) for the PG method (5.6), we obtain

$$\begin{aligned}\|x_{s+1} - x^*\|^2 &= \|P_\Omega(x_s - t\nabla f(x_s)) - P_\Omega(x^* - t\nabla f(x^*))\|^2\\ &\le \|x_s - t\nabla f(x_s) - x^* + t\nabla f(x^*)\|^2\\ &= \|x_s - x^*\|^2 - 2t(\nabla f(x_s) - \nabla f(x^*), x_s - x^*) + t^2\|\nabla f(x_s) - \nabla f(x^*)\|^2.\end{aligned} \tag{5.11}$$

Keeping in mind the Lipschitz condition (2.2) and strong monotonicity (5.1) from (5.2) with $x = x_s$ and $y = x^*$, we have

$$(\nabla f(x_s) - \nabla f(x^*), x_s - x^*) \ge \frac{lL}{l + L}\|x_s - x^*\|^2 + \frac{1}{l + L}\|\nabla f(x_s) - \nabla f(x^*)\|^2. \tag{5.12}$$

Therefore from (5.11) and (5.12) follows

$$\|x_{s+1} - x^*\| \le \|x_s - x^*\|^2 - 2t\frac{lL}{l + L}\|x_s - x^*\|^2 + t\left(t - \frac{2}{l + L}\right)\|\nabla f(x_s) - \nabla f(x^*)\|^2;$$

hence for $0 < t < 2/(l + L)$ the bound (5.7) holds.

For $t = 2/(l + L)$ from (5.7) follows

$$\|x_{s+1} - x^*\|^2 \leq \left(1 - \frac{4lL}{(l+L)^2}\right) \|x_s - x^*\|^2,$$

or

$$\|x_{s+1} - x^*\|^2 \leq (\frac{L-l}{L+l})^2 \|x_s - x^*\|^2.$$

Therefore for $\varkappa = l/L$ the bound (5.8) holds.

Keeping in mind (5.5) from (5.8) follows (5.9). Therefore x_k is an $\varepsilon > 0$ approximation of f^* if

$$\Delta_k = f(x_k) - f^*(x) \leq L_0 \left(\frac{1-\varkappa}{1+\varkappa}\right)^k \|x_0 - x^*\| \leq \varepsilon.$$

Then

$$k \ln \frac{1-\varkappa}{1+\varkappa} \leq \ln \frac{\varepsilon}{L_0 \|x_0 - x^*\|},$$

or

$$k \geq \frac{\ln \frac{L_0\|x_0 - x^*\|}{\varepsilon}}{\ln \frac{1+\varkappa}{1-\varkappa}} = \frac{\ln \frac{L_0\|x_0 - x^*\|}{\varepsilon}}{\ln(1 + \frac{2\varkappa}{1-\varkappa})}.$$

Keeping in mind $\ln(1+x) \leq x$ for the number of steps $k > 0$ which guarantee $\varepsilon > 0$ approximation for $f(x^*)$, we obtain

$$k \geq \frac{1-\varkappa}{2\varkappa}(\ln[L_0\|x_0 - x^*\|] + \ln \varepsilon^{-1}).$$

Therefore for the PG complexity, we obtain the bound (5.10). In contrast to (4.4), the bound (5.10) is not the worst case bound. It is rather a bound which is defined by the condition number of Q. It shows that for the full rank NNLS in a number of instances the FPG complexity can be substantially improved.

6. Projected Gradient Method for SVM

Constructing soft margin separating hyperplanes leads to NNLS type problem (2.1) with one extra equality constraint. In other words, one has to solve the following NNLS (see [**18**, p. 137])

$$f(x^*) = \min\{f(x) = \frac{1}{2}\|Ax - b\|^2 | 0 \leq x \leq c, (b, x) = 0\} \tag{6.1}$$

where $b = (b_1, ..., b_n)$ and $b_i \in \{-1, 1\}$ $i = 1, .., n$.

Let us remove the box constraints from the set of constraints in (6.1) and consider the Lagrangian for the only equality constraint $(b, x) = 0$ left. We have

$$\mathcal{L}(x, \lambda) = f(x) - \lambda(b, x)$$

The problem (6.1) can be solved by the FPG method applied to

$$\mathcal{L}(x^*, \lambda^*) = \min\{\mathcal{L}(x, \lambda^*) | x \in \Omega\} \tag{6.2}$$

if the optimal multiplier λ^*, which corresponds to equality $(b, x) = 0$, is given.

Obviously it is not the case; however, the fact that (6.1) has only one extra constraint on top of box constraints is very helpful.

To estimate λ^*, we will use the dual function

$$d(\lambda) = \min\{\mathcal{L}(x, \lambda) | x \in \Omega\}. \tag{6.3}$$

The dual function $d(\lambda)$ is concave and continuous in R. By computing the dual function value in two nearby points, one can find accent direction for $d(\lambda)$ at any given point $\lambda \in \mathbb{R}^n$. So the interval, which contains λ^*, can be shrunk by half in two dual function evaluations. Therefore, for a given $\varepsilon > 0$, localization of λ^* in an $\varepsilon > 0$ interval will take $O(\ln \varepsilon^{-1})$ function evaluation. Therefore, one obtains the overall complexity of FPG for QP (6.1) as a product of the bound (4.4) and $O(\ln \varepsilon^{-1})$. Hence, for large n, the FPG can be considered as an alternative to IPMs for SVM calculations. The key advantage of FPG, however, is the necessity to perform at each step matrix by vector multiplication instead of solving the same size linear system.

7. Concluding Remarks

The PG approach is fundamentally different from both active set methods and IPMs. The active set methods deal with active constraint sets locally at the current approximation. The combinatorial flavor of these methods is evident; it makes very difficult the establishment of a meaningful upper bound for the number of steps. At the same time, the active set methods require solving a LS sub-problem or a linear system of equations at each step.

The IPMs eliminate the combinatorial nature of the NNLS by treating the non-negative constraints with the log-banier function. The IPMs guarantee the well-known complexity bound $O(n^3 \ln \varepsilon^{-1})$, but they also require solving a linear system of equation at each step, which for large scale NNLS can be very difficult.

The PG method eliminates both the combinational nature of the box constraints and the necessity of solving a linear system of equation at each step.

A few important issues are left for further research.

First of all, it would be important to incorporate the only equality constraint $(x, b) = 0$ in the FPG method for solving (6.1) – in other words, to avoid the necessity of solving (6.3) several times to locate λ^*.

Second, the main operation in all PG methods has to be done using parallel computations.

Third, extensive numerical experiments with NNLS in general and with SVM problems in particular are necessary to understand the real efficiency of the PG methods for NNLS.

8. Appendix 1

From (3.3) with $X = x^*$, $x = x_{k+1}$ and $x^L_\Omega = X_{k+1}$ follows

$$-\frac{2}{L}\Delta_{k+1} \geq \|X_{k+1} - x_{k+1}\|^2 + 2(x_{k+1} - x^*, X_{k+1} - x_{k+1}). \tag{8.1}$$

On the other hand, for $X = X_k, x = x_{k+1}$ and $x^L_\Omega = X_{k+1}$ from (3.3), we have

$$\frac{2}{L}(\Delta_k - \Delta_{k+1}) \geq \|X_{k+1} - x_{k+1}\|^2 + 2(x_{k+1} - X_k, X_{k+1} - x_{k+1}). \tag{8.2}$$

After multiplying both sides of (8.2) by $(t_{k+1} - 1) > 0$ and adding to (8.1), we obtain

$$\frac{2}{L}[(t_{k+1} - 1)\Delta_k - t_{k+1}\Delta_{k+1}] \geq \tag{8.3}$$

$$t_{k+1}\|X_{k+1} - x_{k+1}\|^2 + 2(X_{k+1} - x_{k+1}, t_{k+1}x_{k+1} - (t_{k+1} - 1)X_k - x^*).$$

From the step length update (b), we have

$$t_k^2 = t_{k+1}(t_{k+1} - 1). \tag{8.4}$$

Therefore after multiplying both sides of (8.3) by t_{k+1} and keeping in mind (8.4) from (8.3) follows

$$\frac{2}{L}[t_k^2\Delta_k - t_{k+1}^2\Delta_{k+1}] \geq \tag{8.5}$$
$$\|t_{k+1}(X_{k+1} - x_k)\|^2 + 2t_{k+1}(X_{k+1} - x_{k+1}, t_{k+1}x_{k+1} - (t_{k+1} - 1)X_k - x^*).$$

Let $t_{k+1}x_{k+1} = a$, $t_{k+1}X_{k+1} = b$ and $(t_{k+1}-1)X_k+x^* = c$, then using the standard three vector identity

$$\|b - a\|^2 + 2(b - a, a - c) = \|b - c\|^2 - \|a - c\|^2$$

from (8.4), we obtain

$$\frac{2}{L}[t_k^2\Delta_k - t_{k+1}^2\Delta_{k+1}] \geq \tag{8.6}$$
$$\|t_{k+1}X_{k+1} - (t_{k+1} - 1)X_k - x^*\|^2 - \|t_{k+1}x_{k+1} - (t_{k+1} - 1)X_k - x^*\|^2.$$

From (c) follows

$$t_{k+1}x_{k+1} = t_{k+1}X_k + (t_k - 1)(X_k - X_{k+1}). \tag{8.7}$$

Keeping in mind

$$y_k = t_kX_k + (t_k - 1)X_{k-1} - x^*$$

from (8.6) and (8.7), we obtain (4.1).

9. Appendix 2

To prove the inequality (5.2), let us first consider the so-called co-coercitivity property

$$(\nabla f(x) - \nabla f(y), x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2, \tag{9.1}$$

which is true for any convex function f with Lipschitz continuous gradient ∇f.

First of all, we recall that from (2.2) and convexity f follows

$$o \leq f(y) - f(x) - (\nabla f(x), y - x) \leq \frac{L}{2}\|x - y\|^2. \tag{9.2}$$

For a given $x \in \mathbb{R}^n$, we consider $\varphi : \mathbb{R}^n \to \mathbb{R}$ defined by the formula

$$\varphi(y) = f(y) - (\nabla f(x), y).$$

The φ is convex and due to (2.2) the gradient

$$\nabla\varphi(y) = \nabla f(y) - \nabla f(x)$$

satisfies the Lipschitz condition. Also $\nabla\varphi(x) = 0$, therefore

$$\varphi(x) \leq \varphi(y - L^{-1}\nabla\varphi(y)). \tag{9.3}$$

By applying (9.2) to $\varphi(y - L^{-1}\nabla\varphi(y))$ and keeping in mind (9.3), we obtain

$$\varphi(x) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|^2$$

or

$$f(y) \geq f(x) + (\nabla f(x), y - x) + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2. \tag{9.4}$$

By interchanging x and y in (9.4), we obtain

$$f(x) \geq f(y) + (\nabla f(y), x - y) + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2. \tag{9.5}$$

We obtain the co-coercitivity property (9.1) by adding (9.4) and (9.5). From strong convexity f follows convexity $\psi(x) = f(x) - \frac{l}{2}\|x\|^2$. Also

$$\|\nabla\psi(x) - \nabla\psi(y)\| \leq (L - l)\|x - y\|.$$

Application of the co-coercitivity property (9.1) to $\psi(x)$ leads to the following inequality:

$$(\nabla f(x) - \nabla f(y) - l(x - y), x - y) \geq \frac{1}{L - l}\|\nabla f(x) - \nabla f(y) - l(x - y)\|^2,$$

i.e.,

$$\begin{aligned}(\nabla f(x) - \nabla f(y), x - y) &\geq l\|x - y\|^2 \\ &+ \frac{1}{L - l}[\|\nabla f(x) - \nabla f(y)\|^2 - 2l(\nabla f(x) - \nabla f(y), x - y) + l^2\|x - y\|^2] \\ &= \frac{Ll}{L - l}\|x - y\|^2 - \frac{2l}{L - l}(\nabla f(x) - \nabla f(y), x - y) + \frac{1}{L - l}\|\nabla f(x) - \nabla f(y)\|^2\end{aligned}$$

or

$$\frac{L + l}{L - l}(\nabla f(x) - \nabla f(y), x - y) \geq \frac{Ll}{L - l}\|x - y\|^2 + \frac{1}{L - l}\|\nabla f(x) - \nabla f(y)\|. \tag{9.6}$$

Dividing both sides of (9.6) by $\frac{L+l}{L-l} > 0$, we obtain (5.2).

Acknowledgement

I am grateful to Dr. V. Vapnik for stimulating discussions.

References

[1] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci. **2** (2009), no. 1, 183–202, DOI 10.1137/080716542. MR2486527 (2010d:35390)

[2] S. Bellavia, M. Macconi, and B. Morini, *An interior point Newton-like method for nonnegative least-squares problems with degenerate solution*, Numer. Linear Algebra Appl. **13** (2006), no. 10, 825–846, DOI 10.1002/nla.502. MR2278195 (2007k:90166)

[3] M. Benthem and M. Keenan, *Fast algorithm for the solution of large-scale non- negativity constrained least squares problems*, J. Chemometrics **18** (2004), 441-450.

[4] R. Bro and S. Jong, *A fast non-negativity-constrained least squares algorithm*, J. Chemometrics **11** (1997), no. 5, 393-401.

[5] Y. Censor, A. Gibali, and S. Reich, *The subgradient extragradient method for solving variational inequalities in Hilbert space*, J. Optim. Theory Appl. **148** (2011), no. 2, 318–335, DOI 10.1007/s10957-010-9757-3. MR2780566 (2011k:49014)

[6] Y. Censor, A. Gibali, and S. Reich, *Strong convergence of subgradient extragradient methods for the variational inequality problem in Hilbert space*, Optim. Methods Softw. **26** (2011), no. 4-5, 827–845, DOI 10.1080/10556788.2010.551536. MR2837800 (2012g:49012)

[7] A. A. Goldstein, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc. **70** (1964), 709–710. MR0165982 (29 #3262)

[8] O. Güler, *New proximal point algorithms for convex minimization*, SIAM J. Optim. **2** (1992), no. 4, 649–664, DOI 10.1137/0802032. MR1186167 (93j:90076)

[9] M. Gusev and D. Evans, *The fastest matrix vector multiplication*, Parallel Algorithms Appl. **1** (1993), no. 1, 57-67.

[10] C. L. Lawson and R. J. Hanson, *Solving least squares problems*, Classics in Applied Mathematics, vol. 15, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1995. Revised reprint of the 1974 original. MR1349828 (96d:65067)

[11] E. S. Levitin and B. T. Poljak, *Minimization methods in the presence of constraints* (Russian), Ž. Vyčisl. Mat. i Mat. Fiz. **6** (1966), 787–823. MR0211590 (35 #2468)
[12] M. Merritt and Y. Zhang, *An interior-point gradient method for large-scale totally nonnegative least squares problems*, Technical Report TR04-08, Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005, U.S.A., 2004.
[13] A. S. Nemirovsky and D. B. Yudin, *Problem complexity and method efficiency in optimization*, A Wiley-Interscience Publication, John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson; Wiley-Interscience Series in Discrete Mathematics. MR702836 (84g:90079)
[14] Yu. E. Nesterov, *A method for solving the convex programming problem with convergence rate* $O(1/k^2)$ (Russian), Dokl. Akad. Nauk SSSR **269** (1983), no. 3, 543–547. MR701288 (84i:90119)
[15] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, Applied Optimization, vol. 87, Kluwer Academic Publishers, Boston, MA, 2004. MR2142598 (2005k:90001)
[16] R. A. Polyak, J. Costa, and S. Neyshabouri, *Dual fast projected gradient method for quadratic programming*, Optim. Lett. **7** (2013), no. 4, 631–645, DOI 10.1007/s11590-012-0476-6. MR3035519
[17] B. Pschenichny, *Algorithms for general problems of mathematical programming*, Kibernetica **6** (1970), 120-125.
[18] V. N. Vapnik, *The nature of statistical learning theory*, 2nd ed., Statistics for Engineering and Information Science, Springer-Verlag, New York, 2000. MR1719582 (2001c:68110)

Department of Mathematics, Technion – Israeli Institute of Technology, Haifa, Israel

E-mail address: rpolyak@techunix.technion.ac.il