Research Article

An Accelerated Proximal Gradient Algorithm for Singly Linearly Constrained Quadratic Programs with Box Constraints

Congying Han,¹ Mingqiang Li,² Tong Zhao,¹ and Tiande Guo¹

¹ School of Mathematical Sciences, University of Chinese Academy of Sciences, Shijingshan District, Beijing 100049, China

² College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China

Correspondence should be addressed to Tong Zhao; zhaotong@ucas.ac.cn

Received 4 August 2013; Accepted 23 September 2013

Academic Editors: I. Ahmad and P.-y. Nie

Copyright © 2013 Congying Han et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Recently, the existed proximal gradient algorithms had been used to solve non-smooth convex optimization problems. As a special nonsmooth convex problem, the singly linearly constrained quadratic programs with box constraints appear in a wide range of applications. Hence, we propose an accelerated proximal gradient algorithm for singly linearly constrained quadratic programs with box constraints. At each iteration, the subproblem whose Hessian matrix is diagonal and positive definite is an easy model which can be solved efficiently via searching a root of a piecewise linear function. It is proved that the new algorithm can terminate at an ε -optimal solution within $O(1/\sqrt{\varepsilon})$ iterations. Moreover, no line search is needed in this algorithm, and the global convergence can be proved under mild conditions. Numerical results are reported for solving quadratic programs arising from the training of support vector machines, which show that the new algorithm is efficient.

1. Introduction

In this paper, we mainly consider the following quadratic programming problem:

$$\min_{x \in \Omega} f(x) = \frac{1}{2} x^T G x - b^T x, \quad (1)$$

where $G \in R^{n \times n}$ is symmetric and positive semidefinite, $x, b \in R^n$, and the feasible region Ω is defined by

$$\Omega = \{x \in R^n, l \leq x \leq u, c^T x = d\}, \quad (2)$$

where $l, u, c \in R^n$ and d is a scalar.

The singly linearly constrained quadratic programs with box constraints appear in a wide range of applications such as image processing, biological information and machine learning. Specifically, support vector machine (SVM) is one of the most classical models of (1). It is a promising technique for solving a variety of machine learning and function estimation problems. The SVM learning methodology has been shown to give good performance in a wide variety of problems such as face detection, text categorization, and handwritten character

recognition. The number of variables in SVM is so huge that traditional optimization methods cannot be directly applied. Some decomposition method [1–4] with its subproblem being a special case of (1) is the main approach for large-scale SVM problems. The solution of the subproblem by generalized variable projection method (GVPM) and projected gradient method (PGM) is introduced in [5, 6], respectively. Moreover, Zanni, and so forth, also proposes the parallel decomposition algorithms based on these two methods in [7, 8]. For more general large-scale model, some of the parallelization pieces of literature in [9, 10] recently are proposed, but these methods cannot be applied specifically to SVM so far.

In this work, we will give an accelerated proximal gradient algorithm for (1). First, we consider the following nonsmooth convex optimization problem:

$$\min_{x \in R^n} F(x) = h(x) + g(x), \quad (3)$$

where $g : R^n \rightarrow (-\infty, \infty]$ is a proper, lower semicontinuous (lsc), convex function and h is convex smooth (i.e., continuously differentiable) on an open subset of R^n containing

$\text{dom } g = \{x \mid g(x) < \infty\}$. ∇h is Lipschitz continuous on $\text{dom } g$. That is,

$$\|\nabla h(x) - \nabla h(y)\| \leq L \|x - y\| \quad \forall x, y \in \text{dom } g \quad (4)$$

for some $L > 0$, where, and in what follows, $\|\cdot\|$ denotes the spectral norm. Obviously, problem (1) is a special case of (3) with $h(x) = f(x)$ and $g(x)$ being the indicator function for the feasible region Ω defined by

$$g(x) = \begin{cases} 0 & \text{if } x \in \Omega, \\ +\infty & \text{else.} \end{cases} \quad (5)$$

Recently, great attention has been paid to the solution of (3). Nesterov and Nemirovski [11, 12] study the accelerated proximal gradient method for (1) with an attractive iteration complexity of $O(1/\sqrt{\varepsilon})$ for achieving ε -optimality. Almost at the same time, Beck and Teboulle [13] give a fast iterative shrinkage-thresholding algorithm (FISTA) which achieves the same convergence rate. After that, Tseng [14] summarizes these algorithms and presents a unified treatment of these methods. All these algorithms have a good performance on large-scale problems, such as linear inverse problems, matrix game problems, and matrix completion. Motivated by the successful use of the accelerated proximal gradient method for (3), we extend Beck and Teboulle's algorithm to solve (1). In particular, the subproblem is solved by searching a root of a piecewise linear continuous function. Numerical results show that the new algorithm is efficient.

The paper is organized as follows. In Section 2, the proximal gradient algorithm and its accelerated version are presented for (1). The convergence of these algorithms is also discussed. In Section 3, the method to efficiently solve the subproblem is introduced. Numerical results on SVM classification problems generated by the random and real world data sets are shown in Section 4. Section 5 is devoted to some conclusions and further study.

2. Proximal Gradient Algorithms

In this section, we introduce the proximal gradient algorithm and its accelerated version which can be applied to solve (1). For any $z \in R^n$, consider the following quadratic approximation of $f(x)$ at z :

$$Q_L(x, z) = f(z) + \langle x - z, \nabla f(z) \rangle + \frac{L}{2} \|x - z\|^2 \quad (6)$$

$$= \frac{L}{2} x^T x + (Gz - Lz - b)^T x + \text{const}, \quad (7)$$

where $L > 0$ is a given parameter, $\text{const} = (L/2)z^T z - \nabla f(z)^T z + f(z)$. Since (6) is a strongly convex function of x , $Q_L(x, z)$ has a unique minimizer in Ω which we denote by

$$P_L(z) = \text{argmin} \{Q_L(x, z) : x \in \Omega\}. \quad (8)$$

First, we present a simple result which characterizes the optimality of $P_L(z)$.

Lemma 1. For any $z \in R^n$, $w = P_L(z)$ if and only if there exists $\gamma(z) \in N_\Omega(w)$, such that

$$\nabla f(z) + L(w - z) + \gamma(z) = 0, \quad (9)$$

where $N_\Omega(w)$ is the normal cone of Ω at the point w denoted by

$$N_\Omega(w) = \{w^* \in R^n \mid \langle w^*, v - w \rangle \leq 0 \quad \forall v \in \Omega\}. \quad (10)$$

Proof. This result can be obtained directly from the optimality conditions of (8), which is a strongly convex problem. $\square$

Then, similarly to Lemma 2.3 in [13], we have the following key result.

Lemma 2. Let $z \in R^n$ and $L > 0$ such that

$$f(P_L(z)) \leq Q(P_L(z), z). \quad (11)$$

Then for any $x \in \Omega$,

$$f(x) - f(P_L(z)) \geq \frac{L}{2} \|P_L(z) - z\|^2 + L \langle z - x, P_L(z) - z \rangle. \quad (12)$$

Proof. Since $\gamma(z) \in N_\Omega(w)$ and f is convex, it follows that

$$f(x) \geq f(z) + \langle x - z, \nabla f(z) \rangle + \langle \gamma(z), x - P_L(z) \rangle. \quad (13)$$

Hence, from (9), (11), (13), and the definition of $P_L(z)$ we have

$$\begin{aligned} f(x) - f(P_L(z)) &\geq f(x) - Q(P_L(z), z) \\ &\geq -\frac{L}{2} \|P_L(z) - z\|^2 + \langle x - P_L(z), \nabla f(z) + \gamma(z) \rangle \\ &= -\frac{L}{2} \|P_L(z) - z\|^2 + L \langle x - P_L(z), z - P_L(z) \rangle \\ &= \frac{L}{2} \|P_L(z) - z\|^2 + L \langle z - x, P_L(z) - z \rangle. \end{aligned} \quad (14)$$

This completes the proof. $\square$

Note that $f(x)$ is a convex quadratic function and (11) can always be satisfied for $P_L(z)$ if $L \geq \lambda_{\max}(G)$, where $\lambda_{\max}$ denotes the maximum eigenvalue of G . The proximal gradient algorithm for (1) is as follows.

Algorithm 3. We have the following steps.

Step 1. Choose $x_0 \in R^n$.

Step 2. While (x_{k-1}) does not satisfy the terminal conditions)

$$\begin{aligned} x_k &= P_L(x_{k-1}) \\ k &= k + 1. \end{aligned} \quad (15)$$

The following theorem is about the iteration complexity of Algorithm 3. In what follows, X^* and f^* denote the set of optimal solutions and optimal value of (1), respectively.

Theorem 4. Let $\{x_k\}$ and $\{f(x_k)\}$ be the sequences generated by Algorithm 3 with $L \geq \lambda_{\max}(G)$; then, for $k \geq 1$, the following results hold:

(a) $\{f(x_k)\}$ is nonincreasing and

$$f(x_k) - f(x_{k+1}) \geq \frac{L}{2} \|x_k - x_{k+1}\|^2; \quad (16)$$

(b) $f(x_k) - f(x^*) \leq L\|x_0 - x^*\|^2/2k$, for all $x^* \in X^*$;

(c) if $\{x_{k_j}\}$ is a convergent subsequence of $\{x_k\}$ and $\lim_{j \rightarrow \infty} x_{k_j} = x^{**}$ then $x^{**} \in X^*$;

(d) if G is positive definite, then

$$\|x_k - x^*\| \leq \sqrt{\frac{L}{2\lambda_{\min}(G)k}} \|x_0 - x^*\|. \quad (17)$$

Proof. (a) Let $z = x_k$, $x_{k+1} = P_L(z)$; then from (11) and the optimality of $P_L(z)$ we have

$$f(x_{k+1}) \leq Q(x_{k+1}, x_k) \leq Q(x_k, x_k) = f(x_k). \quad (18)$$

Furthermore, take $x = x_k$ in (12); we get

$$f(x_k) - f(x_{k+1}) \geq \frac{L}{2} \|x_k - x_{k+1}\|^2. \quad (19)$$

(b) We have

$$\begin{aligned} & k(f(x_k) - f(x^*)) \\ &= kf(x_k) - \sum_{n=0}^{k-1} f(x_{n+1}) + \sum_{n=0}^{k-1} f(x_{n+1}) - kf(x^*) \\ &\leq \sum_{n=0}^{k-1} (f(x_{n+1}) - f(x^*)) \\ &\leq -\sum_{n=0}^{k-1} \left(\frac{L}{2} \|x_{n+1} - x_n\|^2 - L \langle x_n - x^*, x_{n+1} - x_n \rangle \right) \quad (20) \\ &= -\frac{L}{2} \left(\sum_{n=0}^{k-1} (\|x^* - x_{n+1}\|^2 - \|x^* - x_n\|^2) \right) \\ &= -\frac{L}{2} (\|x^* - x_k\|^2 - \|x_0 - x^*\|^2) \\ &\leq \frac{L}{2} \|x_0 - x^*\|^2, \end{aligned}$$

where the first inequality uses the fact that $\{f(x_k)\}$ is non-increasing and the second inequality is established by taking $x = x^*$, $z = x_n$ in (12). Dividing both sides by k in the last inequality, we get

$$f(x_k) - f(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2k}. \quad (21)$$

(c) It follows from (b) that, for all $x^* \in X^*$,

$$f(x_{k_j}) - f(x^*) \leq \frac{L\|x_0 - x^*\|^2}{2k^j}. \quad (22)$$

Let $j \rightarrow \infty$ in the above inequality; we have $f(x^{**}) \leq f(x^*)$. On the other hand, by the definition of x^* , we have $f(x^{**}) \geq f(x^*)$. Hence, $f(x^{**}) = f(x^*)$, which implies that $x^{**} \in X^*$.

(d) It follows from the definition of $f(x)$ that, for all $x^* \in X^*$,

$$\begin{aligned} & f(x_k) - f(x^*) \\ &= \nabla f(x^*)^T (x_k - x^*) + (x_k - x^*)^T G (x_k - x^*) \\ &\geq (x_k - x^*)^T G (x_k - x^*) \\ &\geq \lambda_{\min}(G) \|x_k - x^*\|^2, \end{aligned} \quad (23)$$

where the first inequality is established by the definition of x^* and the second inequality, with $\lambda_{\min}(G)$ denoting the minimum eigenvalue of G , is obtained from the positive definite property of G .

From (23) and (b), we get

$$\|x_k - x^*\| \leq \sqrt{\frac{L}{2\lambda_{\min}(G)k}} \|x_0 - x^*\|. \quad (24) \quad \square$$

Theorem 4 shows that Algorithm 3 has a result complexity of $O(1/k)$. Now we introduce a new algorithm with an improved complexity result of $O(1/k^2)$.

Algorithm 5. We have the following steps.

Step 1. Choose $z_1 = x_0 \in R^n$, set $k = 1$, $t_1 = 1$.

Step 2. While (x_{k-1}) does not satisfy the terminal conditions)

$$x_k = P_L(z_k), \quad (25)$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \quad (26)$$

$$\begin{aligned} z_{k+1} &= x_k + \left(\frac{t_k - 1}{t_{k+1}} \right) (x_k - x_{k-1}), \\ k &= k + 1. \end{aligned} \quad (27)$$

Remark 6. In Algorithm 5, the operator $P_L(\cdot)$ is employed on the point z_k which is a specific linear combination of the previous two points $\{x_{k-1}, x_{k-2}\}$. However, in Algorithm 3, the operator only uses the previous point x_{k-1} . On the other hand, the computational effort of these two algorithms is almost the same except for the computation of (26) in Algorithm 5 which is negligible.

Now we give the promising improved complexity result for Algorithm 5.

Theorem 7. Let $\{x_k\}$ and $\{f(x_k)\}$ be the sequences generated by Algorithm 5 with $L \geq \lambda_{\max}(G)$. Then, for $k \geq 1$, the following results hold:

(a) $f(x_k) - f(x^*) \leq 2L\|x_0 - x^*\|^2/(k+1)^2$, for all $x^* \in X^*$;

(b) if $\{x_{k_j}\}$ is a convergent subsequence of $\{x_k\}$ and $\lim_{j \rightarrow \infty} x_{k_j} = x^{**}$, then $x^{**} \in X^*$;

(c) if G is positive definite, then

$$\|x_k - x^*\| \leq \sqrt{\frac{L}{2\lambda_{\min}(G)(k+1)^2}} \|x_0 - x^*\|. \quad (28)$$

Proof. (a) Using Lemma 2, the proof is similar to the proof of Theorem 4.4 in [13]. (b) and (c) can be proved in the same way as that of (c) and (d) in Theorem 4, respectively. $\square$

3. How to Solve $P_L(\cdot)$

In this section, we extend the method studied by Dai and Fletcher [6] to solve $P_L(\cdot)$. In order to obtain $P_L(\cdot)$ in (8), we need to solve the following problem:

$$\begin{aligned} \min \quad & \frac{L}{2} x^T x - q^T x \\ \text{s.t.} \quad & c^T x = d, \quad l \leq x \leq u, \end{aligned} \quad (29)$$

where $q = Lz + b - Gz$. For a given value of λ , we consider the following box constrained QP problem:

$$\begin{aligned} \min \quad & \frac{L}{2} x^T x - q^T x - \lambda (c^T x - d) \\ \text{s.t.} \quad & l \leq x \leq u \end{aligned} \quad (30)$$

and denote the minimizer of (30) as $x(\lambda)$. Then,

$$x(\lambda) = \text{mid}(l, \alpha, u), \quad (31)$$

where $\alpha = \alpha(\lambda)$ has components $\alpha_i = (q_i + \lambda c_i)/L$ ($i = 1, 2, \dots, n$) and $\text{mid}(l, \alpha, u)$ is the componentwise operation that supplies the median of its three arguments. That is,

$$\begin{aligned} [x(\lambda)]_i = & l_i + \alpha_i - u_i - \max(l_i, \alpha_i, u_i) \\ & - \min(l_i, \alpha_i, u_i) \quad (i = 1, 2, \dots, n). \end{aligned} \quad (32)$$

Define

$$r(\lambda) = c^T x(\lambda) - d; \quad (33)$$

then from [6], we know that $r(\lambda)$ is a piecewise linear continuous and monotonically increasing function of λ . Furthermore, $x(\lambda^*)$ is the optimal of (29) if λ^* is located such that $r(\lambda^*) = 0$. Hence, the main task of solving (29) is to find a λ^* such that $r(\lambda^*) = 0$. For this purpose, we adopt the algorithm introduced in [6] which consists of a *bracketing phase* and a *secant phase*. Numerical experiments in the next section show that this algorithm achieves high efficiency.

4. Numerical Experiments

As we know, one of the important applications of SVM is classification. Given a training set

$$D = \{(z_i, y_i), i = 1, \dots, n, z_i \in R^m, y_i \in \{-1, 1\}\}, \quad (34)$$

we need to find a hyperplane to separate the two classes of points. This can be done by solving the following convex quadratic programming problem:

$$\begin{aligned} \min \quad & \frac{1}{2} x^T Q x - e^T x \\ \text{s.t.} \quad & y^T x = 0, \quad 0 \leq x \leq C e, \end{aligned} \quad (35)$$

where $y = [y_1, y_2, \dots, y_n]^T$, $e \in R^n$ is the vector of all ones, and C is a positive scalar. Q is a $n \times n$ symmetric and positive semidefinite matrix with entries $Q_{ij} = y_i y_j K(x_i, x_j)$, $i, j = 1, 2, \dots, n$, where $K : R^m \times R^m \rightarrow R$ is some kernel function.

In this section, we illustrate the performance of Algorithms 3 and 5 with Matlab 7.10 on a Windows XP professional computer (Pentium R, 2.79 GHZ). We conduct a test on SVM classification problems with the random data sets and the real world data sets.

The generation of the random data sets is based on four parameters n, m, β , and η , where n is the number of samples and m is the dimension of z_i ($i = 1, \dots, n$). Each element of z_i is randomly generated in $[\beta, \eta]$ ($i = 1, \dots, n$), and -1 or 1 randomly emerges in the i th entry of y ($i = 1, \dots, n$).

We have generated three random data sets with $\beta = -1$, $\eta = 1$, $m = 5$, and $n = 200, 600$, and 1000 , respectively. The two real world data sets are the UCI adult data set and the heart data set. For the UCI adult, the versions with $n = 1605, 2265$, and 3185 , are considered. The Gauss radial basis function

$$K(z_i, z_j) = \exp\left(-\frac{\|z_i - z_j\|^2}{\sigma^2}\right) \quad (36)$$

is used in our tests. The parameters C in (35) and σ in (36) are set to $(10, \sqrt{40})$, $(1, \sqrt{13})$, and $(1, 1000)$ for random data sets, heart, and UCI adult data sets respectively.

For all test problems, the initial point is chosen as $x_0 = (0, \dots, 0)^T$ and the const $L = \sum_{i=1}^n Q_{ii}$. The terminal rules for solving $P_L(\cdot)$ are $|r(\lambda)| \leq 10^{-8}$.

For the test problem with random data set ($n = 200$) and heart data set, the running steps are 8000 and 1000, respectively. In order to make a visual comparison between Algorithms 3 and 5, we plot the values of $f(x_k)$ with iterations in Figure 1. Furthermore, we compute

$$CR(j) = \sum_{k=1}^M cr(M \times (j-1) + k) \quad (j = 1, 2, \dots, J), \quad (37)$$

where $cr(M \times (j-1) + k)$ denotes the number of calculating $r(\lambda)$ in step $M \times (j-1) + k$. In Figure 2, we plot the values of $CR(j)$ ($j = 1, 2, \dots, J$) with $M = 100$, $J = 80$ and $M = 20$, $J = 50$, respectively.

For the test problems with the rest data sets, the terminal condition is based on the fulfilment of the KKT conditions within 0.001 (see [15]). The numerical results are shown in

FIGURE 1: Function value curve with iterations.

FIGURE 2: Number of calculating $r(\lambda)$ for every finite step.

Tables 1 and 2, where the meaning of the indexes are as follows:

- (i) n : the scale of the data sets,
- (ii) sec: the computing time (in seconds),
- (iii) k : the number of iterations,
- (iv) raver: the average number of calculating $r(\lambda)$ in each step.

From Figure 2, we can see that although the total number of calculating $r(\lambda)$ every 100 (27) steps in Algorithm 3 is less than that the number in Algorithm 5. However, Algorithm 5 uses much less iterations and time to achieve the approximate solution, compared with Algorithm 3 from Figure 1 and Tables 1 and 2. For the subproblem, less than 4 iterations are used to achieve a very accurate solution on average, which shows that $P_L(\cdot)$ can be solved efficiently. Moreover, from the column raver in Tables 1 and 2, we can also see that the average

TABLE 1: Numerical results for Algorithm 3.

Problems	n	Sec	k	Raver
Random	600	19.97	19524	2.98
	1000	78.23	32153	2.99
	1605	6.81	1064	1.34
Adult	2265	20.45	1515	1.50
	3185	49.50	2103	1.50

TABLE 2: Numerical results for Algorithm 5.

Problems	n	Sec	k	Raver
Random	600	5.33	5021	3.68
	1000	19.05	7095	3.30
	1605	1.41	88	2.82
Adult	2265	2.84	106	2.97
	3185	6.30	125	2.98

number of calculating $r(\lambda)$ with real world data sets is slightly less than that with random data sets.

5. Conclusion and Future Work

We have extended the accelerated proximal point method to the solution of singly linearly constrained quadratic programming with box constraints. The new algorithm is proved to be globally convergent. Numerical results also show that the new algorithm performs well on medium-scale quadratic programs. On the other hand, Solving the subproblem by searching a root of a piecewise linear continuous function is a very cheap. Considering the good performance of the new Algorithm 5, we can apply it to the solution of the subproblems in decomposition methods for large-scale SVM problems. Moreover, a parallel version of this algorithm combined with the theory in [9, 10] is also a direction of our future research.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (11101420, 71271204) and the Excellent Young Scientist Foundation of Shandong Province (2010BSE06047).

References

- [1] L. Zanni, "An improved gradient projection-based decomposition technique for support vector machines," *Computational Management Science*, vol. 3, no. 2, pp. 131–145, 2006.
- [2] S. Lucidi, L. Palagi, A. Risi, and M. Sciandrone, "A convergent decomposition algorithm for support vector machines," *Computational Optimization and Applications*, vol. 38, no. 2, pp. 217–234, 2007.
- [3] T. Serafini and L. Zanni, "On the working set selection in gradient projection-based decomposition techniques for support vector machines," *Optimization Methods and Software*, vol. 20, no. 4–5, pp. 583–596, 2005.
- [4] P. Tseng and S. Yun, "A coordinate gradient descent method for nonsmooth separable minimization," *Mathematical Programming B*, vol. 117, no. 1–2, pp. 387–423, 2009.
- [5] T. Serafini, G. Zanghirati, and L. Zanni, "Gradient projection methods for large quadratic programs and applications in training support vector machines," *Optimization Methods and Software*, vol. 20, pp. 353–378, 2003.
- [6] Y. Dai and R. Fletcher, "New algorithms for singly linearly constrained quadratic programs subject to lower and upper bounds," *Mathematical Programming*, vol. 106, no. 3, pp. 403–421, 2006.
- [7] G. Zanghirati and L. Zanni, "A parallel solver for large quadratic programs in training support vector machines," *Parallel Computing*, vol. 29, no. 4, pp. 535–551, 2003.
- [8] L. Zanni, T. Serafini, and G. Zanghirati, "Parallel software for training large scale support vector machines on multiprocessor systems," *Journal of Machine Learning Research*, vol. 7, pp. 1467–1492, 2006.
- [9] H. Congying, W. Yongli, and H. Guoping, "On the convergence of asynchronous parallel algorithm for large-scale linearly constrained minimization problem," *Applied Mathematics and Computation*, vol. 2, pp. 434–441, 2009.
- [10] Z. Fangying, H. Congying, and W. Yongli, "Parallel SSLE algorithm for large scale constrained optimization," *Applied Mathematics and Computation*, vol. 217, no. 12, pp. 5377–5384, 2011.
- [11] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, no. 1, pp. 127–152, 2005.
- [12] A. Nemirovski, "Prox-method with rate of convergence $O(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems," *SIAM Journal on Optimization*, vol. 15, no. 1, pp. 229–251, 2005.
- [13] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, pp. 183–202, 2009.
- [14] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization," *SIAM Journal on Optimization*. In press.
- [15] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods: Support Vector Learning*, B. Scholkopf, C. J. C. Burges, and A. J. Smola, Eds., pp. 169–184, MIT Press, Cambridge, Mass, USA, 1998.