On Convergence Rate of Concave-Convex Procedure

Ian En-Hsu Yen
Department of Computer Science
National Taiwan University
Taipei 106, Taiwan
r00922017@csie.ntu.edu.tw

Nanyun Peng
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
npeng1@jhu.edu

Po-Wei Wang
Department of Computer Science
National Taiwan University
Taipei 106, Taiwan
b97058@csie.ntu.edu.tw

Shou-De Lin
Department of Computer Science
National Taiwan University
Taipei 106, Taiwan
sdlin@csie.ntu.edu.tw

Abstract

Concave-Convex Procedure (CCCP) has been widely used to solve nonconvex d.c.(difference of convex function) programs occur in learning problems, such as sparse support vector machine (SVM), transductive SVM, sparse principal componenent analysis (PCA), etc. Although the global convergence behavior of CCCP has been well studied, the convergence rate of CCCP is still an open problem. Most of d.c. programs in machine learning involve constraints or nonsmooth objective function, which prohibits the convergence analysis via differentiable map. In this paper, we approach this problem in a different manner by connecting CCCP with more general block coordinate decent method. We show that the recent convergence result [1] of coordinate gradient descent on nonconvex, nonsmooth problem can also apply to exact alternating minimization. This implies the convergence rate of CCCP is at least linear, if in d.c. program the nonsmooth part is piecewise-linear and the smooth part is strictly convex quadratic. Many d.c. programs in SVM literature fall in this case.

1 Introduction

Concave-Convex Procedure is a popular method for optimizing d.c.(difference of convex function) program of the form:

$$\begin{aligned} \min_{x} \quad & u(x) - v(x) \\ s.t. \quad & f_i(x) \leq 0,\ i = 1..p \\ & g_j(x) = 0,\ j = 1..q \end{aligned} \tag{1}$$

where $u(x), v(x)$, and $f_i(x)$ being convex functions, $g_j(x)$ being affine function, defined on $\mathbb{R}^n$. Suppose $v(x)$ is (piecewise) differentiable, the Concave-Convex Procedure iteratively solves a sequence of convex program defined by linearizing the concave part:

$$\begin{aligned} x^{t+1} \in \ \underset{x}{argmin} \quad & u(x) - \nabla v(x^t)^T x \\ s.t. \quad & f_i(x) \leq 0,\ i = 1..p \\ & g_j(x) = 0,\ j = 1..q \end{aligned} \tag{2}$$

This procedure is originally proposed by Yuille etal.[2] to deal with unconstrained d.c. program with smooth objective function. Nevertheless, many d.c. programs in machine learning come with

constraints or nonsmooth functions. For example, [4] and [5] propose using a concave function to approximate l_o-loss for sparse PCA and SVM feature selection repectively, which results in d.c. programs with convex constraints. In [11], [12], R.Collobert etal. formulate ramp-loss and transductive SVM as d.c. programs with nonsmooth $u(x), v(x)$. Other examples such as [9] and [13] use (1) with piecewise-linear $u(x), v(x)$ to handle nonconvex tighter bound [9] and hidden variables [13] in structural-SVM.

Though CCCP is extensively used in machine learning, its convergence behavior has not been fully understood. Yuille etal. give an analysis of CCCP's global convergence in the original paper [2], which, however, is not complete [14]. The global convergence of CCCP is proved in [10] and [14] via different approaches. However, as [14] pointed out, the converence rate of CCCP is still an open problem. [6] and [7] have analyzed the local convergence behavior of Majorization Minimiation algorithm, where CCCP is a special case, by taking (2) as a differentiable map $x^{t+1} = M(x^t)$, but the analysis only applies to the unconstrained, differetiable version of d.c. program. Thus it cannot be used for examples mentioned above.

In this paper, we approach this problem in a different way by connecting CCCP with more general block coordinate decent method. We show that the recent convergence result [1] of coordinate gradient descent on nonconvex, nonsmooth problem can also apply to block coordinate descent. Since CCCP, and more general Majorization-Minimization algorithm are special cases of block coordinate descent, this connection provides a simple way to prove convergence rate of CCCP. In particular, we show that the sequence $\{x^t\}_{t=0}^{\infty}$ provided by (2) converges at least linearly to a stationary point of (1) if the nonsmooth part in $u(x)$ and $v(x)$ are convex pieceiwise-linear and the smooth part in $u(x)$ is strictly convex quadratic. Many d.c. programs in SVM literature fall in this case, such as that in ramp-loss SVM [11], transductive SVM [12], and structual-SVM with hidden variables [13] or nonconvex tighter bound [9].

2 Majorization Minimization as Block Coordinate Descent

In this section, we show CCCP is a special case of Majorization Minimization algoritm, and thus can be viewed as block cooridnate descent on surrogate function. Then we introduce an alternative formulation of algorithm (2) when $v(x)$ is piecewise-linear that fits better for convergence analysis.

The Majorization Minimization (MM) principle is as follows. Suppose our goal is to minimize $f(x)$ over $\Omega \subset \mathbb{R}^n$. We first construct a *majorization function* $g(x, y)$ over $\Omega \times \Omega$ such that

$$\begin{cases} f(x) \leq g(x, y), & x, y \in \Omega \\ f(x) = g(x, x), & x \in \Omega \end{cases} \tag{3}$$

The MM algorithm, therefore, minimizes an upper bound of $f(x)$ at each iteration

$$x^{t+1} \in \underset{x \in \Omega}{argmin} \quad g(x, x^t) \tag{4}$$

until fix point of (4). Since the minimum of $g(x^t, y)$ occurs at $y = x^t$, (4) can be seen as block coordinate descent on $g(x, y)$ with two blocks x, y.

CCCP is a special case of (4) with $f(x) = u(x) - v(x)$ and $g(x, y) = u(x) - v'(y)^T(x - y) - v(y)$. The condition (3) holds because, by convexity of $v(x)$, the first-order Taylor approximiation is less or equal to $v(x)$, with eqaulity when $y = x$. Thus we have

$$\begin{cases} f(x) = u(x) - v(x) \leq u(x) - v'(y)^T(x - y) - v(y) = g(x, y) & , x, y \in \Omega \\ f(x) = g(x, x) & , x \in \Omega \end{cases} \tag{5}$$

However, when $v(x)$ is piecewise-linear, the term $-v'(y)^T(x - y) - v(y)$ is discontinuous and nonconvex. For convergence analysis, we express $v(x)$, in another way, as $\max_{i=1}^m (a_i^T x + b_i)$ with $v'(x) = a_k$ piecewisely, where $k = \arg\max_i (a_i^T x + b_i)$. Thus we can express d.c. program (1) in another form

$$\begin{aligned} \underset{x \in \mathbb{R}^n, d \in \mathbb{R}^m}{min} \quad & u(x) - \sum_{i=1}^m d_i (a_i^T x + b_i) \\ s.t. \quad & \sum_{i=1}^m d_i = 1, \quad d_i \geq 0, \quad i = 1..m \\ & f_i(x) \leq 0, \; i = 1..p, \quad g_j(x) = 0, \; j = 1..q \end{aligned} \tag{6}$$

Alternaing minimizing (6) between x and d yields the same CCCP algorithm (2).[1]

This formulation replaces discontinous, nonconvex term $-v'(y)(x-y)-v(y)$ with quadratic term $-\sum_{i=1}^{m} d_i(a_i^T x+b_i)$, so (6) fits the form of nonsmooth problem studied in [1], that is, minimizing a function $F(x)=f(x)+cP(x)$, with $c>0$, $f(x)$ being smooth, and $P(x)$ being convex nonsmooth. In next section, we give convergence result of CCCP by showing that the alternating minimization of (6) yields the same sequence $\{x^t, y^t\}_{t=0}^{\infty}$ as that of a special case of Coordinate Gradient Descent proposed in [1].

3 Convergence Theorem

Lemma 3.1. *Consider the problem*

$$\min_{x,\ y} \quad F(x,y)=f(x,y)+cP(x,y) \tag{7}$$

where $f(x,y)$ is smooth and $P(x,y)$ is nonsmooth, convex, lower semi-continuous, and seperable for x and y. The sequence $\{x^t, y^t\}_{t=0}^{\infty}$ produced by alternating minimization

$$x^{t+1} = \underset{x}{argmin} \quad F(x, y^t) \tag{8}$$

$$y^{t+1} = \underset{y}{argmin} \quad F(x^{t+1}, y) \tag{9}$$

(a) converges to a stationary point of (7), if in (8) and (9), or their equivalent problems[2], the objective functions have smooth parts $f(x)$, $f(y)$, that are strictly convex quadratic.

(b) with linear convergence rate, if $f(x,y)$ is quadratic (maybe nonconvex), $P(x,y)$ is polyhedral, in addition to assumption in (a).

Proof. Let the smooth parts of objective function in (8) and (9), or their equivalent problems, be $f(x)$ and $f(y)$. We can define a special case of Coordinate Gradient Descent (CGD) proposed in [1] which yields the same sequence $\{x^t, y^t\}_{t=0}^{\infty}$ as that produced by the alternating minimization, where for each iteration k of the CGD algorithm, we use hessian matrix $H_{xx}^k = \nabla^2 f(x) \succ 0_n$ for block $J=x$ and $H_{yy}^k = \nabla^2 f(y) \succ 0_m$ for block $J=y$, with exact line search (which satisfies Armijo Rule)[3]. By Theorem 1 of [1], the sequence $\{x^t, y^t\}_{t=0}^{\infty}$ converges to a stationary point of (7). If we further assume that $f(x,y)$ is quadratic and $P(x,y)$ is polyhedral, by applying Theorem 2 and 4 in [1], the convergence rate of $\{x^t, y^t\}_{t=0}^{\infty}$ is at least linear. □

Lemma 3.1 provides a basis for the convergence of (6), and thus the CCCP (2). However, when minimizing over d, the objective in (6) is linear instead of strictly convex quadratic as required by lemma 3.1. The next lemma shows that the problem, nevertheless, has equivalent strictly convex quadratic problem that gives the same solution.

Lemma 3.2. *There exist $\epsilon_0 > 0$ and d^* such that the problem*

$$\begin{aligned} \underset{d\in\mathbb{R}^m}{argmin} \quad & -c^T d + \frac{\epsilon}{2}\|d\|^2 \\ s.t. \quad & \sum_{i=1}^{m} d_i = 1, \quad d_i \geq 0, \quad i = 1..m \end{aligned} \tag{10}$$

has the same optimal solution d^* for $\forall \epsilon < \epsilon_0$.

[1] Let set $S = \arg\min_i (-a_i^T x - b_i)$. When minimizing (6) over d, an optimal solution just set $d_{k\in S} = 1/|S|$ and $d_{j\notin S} = 0$. When minimizing over x, the problem becomes $x^{t+1} = \arg\min_x (u(x) - a_S^T x - b_S) = \arg\min_x (u(x) - v'(x^t)^T x)$ subject to constraints in (1), where a_S, b_S are averages over a_k, b_k for $k \in S$, and $v'(x^t) = a_S$ is a sub-gradient of $v(x)$ at x^t.

[2] By "equivalent", we means the two problems have the same optimal solution.

[3] Note, in [1], the Hassian matrix H^k affect CGD algorithm only through H_{JJ}^k, so we can always have a positive-definite H^k when H_{JJ}^k is positive definite by assigning, other than H_{JJ}^k, 1 to diagnoal elements, 0 to non-diagnoal elements.

Proof. Let set $S = \arg\max_i (c_i)$ and $c_{max} = \max_i (c_i)$. As $\epsilon = 0$, we can obtain a optimal solution d^* by setting $d^*_{k\in S} = \frac{1}{|S|}$ and $d^*_{j\notin S} = 0$. Let α, β, be Lagrange multipliers of affine and non-negative constraints in (10) respectively. By KKT condtion, the solution d^*, α^*, β^* must have $-c - \alpha^* e - \beta^* = 0$, with $\beta^*_{k\in S} = 0, \quad \alpha^* = -c_{max}$, and $\beta^*_{j\notin S} = c_{max} - c_j$, where $e = [1, .., 1]^T_m$.

When $\epsilon > 0$, the KKT condition for (10) only differs by the equation $\epsilon d^* - c - \tilde{\alpha} e - \tilde{\beta} = 0$, which can be satisfied by setting $\tilde{\beta}_{k\in S} = 0, \quad \tilde{\alpha} = \alpha^* + \frac{\epsilon}{|S|}$, and $\tilde{\beta}_{j\notin S} = \beta^*_{j\notin S} - \frac{\epsilon}{|S|}$. If $\frac{\epsilon}{|S|} < \beta^*_j = c_{max} - c_j$ for $\forall j \notin S$, then $d^*, \tilde{\alpha}, \tilde{\beta}$ still satisfy the KKT condition. In other words, d^* is still optimal if $\epsilon < \epsilon_0$, where $\epsilon_0 = \min_{j\notin S} (c_{max} - c_j)|S| > 0$. □

When minimizing (6) over d, the problem is of form (10) with $\epsilon = 0$ and $c^t_i = a^T_i x^t + b_i$, $i = 1..m$. Lemma (3.2) says that we can always find equivalent strictly convex quadratic problem by setting positive $\epsilon^t < \min_{j\notin S} (c^t_{max} - c^t_j)|S|$. Now we are ready to give the main theorem.

Theorem 3.3. *The Concave-Convex Procedure (2) converges to a stationary point of (1) in at least linear rate, if the nonsmooth part of $u(x)$ and $v(x)$ are convex piecewise-linear, the smooth part of $u(x)$ is strictly convex quadratic, and the domain formed by $f_i(x) \leq 0$, $g_i(x) = 0$ is polyhedral.*

Proof. Let $u(x) - v(x) = f_u(x) + P_u(x) - v(x)$, where $f_u(x)$ is strictly convex quadratic and $P_u(x)$, $v(x)$ are convex piecewise-linear. We can formulate CCCP as an alternating minimization on (6) between x and d. The constraints $f_i(x) \leq 0$ and $g_j(x) = 0$ form a polyhedral domain, so we transform them into a polyhedral, lower semi-continuous function $P_{dom}(x) = 0$ if $f_i(x) \leq 0$, $g_j(x) = 0$ and $P_{dom}(x) = \infty$ otherwise. By the same way, the domain constraints on d are also transformed into polyhedral, lower semi-continuous function $P_{dom}(d)$. Then (6) can be expressed as $\min_{x,d} F(x, d) = f(x, d) + P(x, d)$, where $f(x, d) = f_u(x) - \sum^m_{i=1} d_i(a^T_i x + b_i)$ is quadratic and $P(x, d) = P_u(x) + P_{dom}(x) + P_{dom}(d)$ is polyhedral, lower-semi-continuous, separable for x and d. When alternating minimizing $F(x, d)$, the smooth part of subproblem $min_x F(x, d^t)$ is strictly convex quadratic, and the subproblem $min_d F(x^{t+1}, d)$, by Lemma 3.2, has equivalent strictly convex quadratic problem. Hence, the alternating minimization of $F(x, d)$, that is, the CCCP algorithm (2), converges at least linearly to a stationary point of (1) by Lemma 3.1. □

To our knowledge, this is the first result on convergence rate of CCCP for nonsmooth problem. Specifically, we show that the CCCP algorithm used in [9], [11], [12] and [13] for structual-SVM with tighter bound, transductive SVM, ramp-loss SVM and structual-SVM with hidden variables has at least linear convergence rate.

Acknowledgments

This work was also supported by National Science Council, National Taiwan University and Intel Cooperation under Grants NSC 100-2911-I-002-001, and 101R7501.

References

[1] Paul Tseng and Sangwoon Yun. (2009) A coordinate gradient descent method for nonsmooth separable minimization. Mathematical Programming.

[2] A. L. Yuille and A. Rangarajan. The concave-convex procedure. Neural Computation, 15:915V936, 2003

[3] L. Wang, X. Shen, and W. Pan. On transductive support vector machines. In J. Verducci, X. Shen, and J. Lafferty, editors, Prediction and Discovery. American Mathematical Society, 2007.

[4] B. K. Sriperumbudur, D. A. Torres, and G. R. G. Lanckriet. Sparse eigen methods by d.c. programming. In Proc. of the 24th Annual International Conference on Machine Learning, 2007.

[5] J. Neumann, C. Schnorr, and G. Steidl. Combined SVM-based feature selection and classification. Machine Learning, 61:129V150, 2005.

[6] X.-L. Meng. Discussion on optimization transfer using surrogate objective functions. Journal of Computational and Graphical Statistics, 9(1):35V43, 2000.

[7] R. Salakhutdinov, S. Roweis, and Z. Ghahramani. On the convergence of bound optimization algorithms. In Proc. 19th Conference in Uncertainty in Arti?cial Intelligence, pages 509V516, 2003.

[8] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, Advances in Neural Information Processing Systems 13, pages 556V562. MIT Press, Cambridge, 2001.

[9] C. B. Do, Q. V. Le, C. H. Teo, O. Chapelle, and A. J. Smola. Tighter bounds for structured estimation. In Advances in Neural Information Processing Systems 21, 2009. To appear.

[10] T. Pham Dinh and L. T. Hoai An. Convex analysis approach to d.c. programming: Theory, algorithms and applications. Acta Mathematica Vietnamica, 22(1):289V355, 1997.

[11] R. Collobert, F. Sinz, J. Weston, and L. Bottou. Large scale transductive SVMs. Journal of Machine Learning Research, 7:1687V1712, 2006.

[12] Collobert, R., Weston, J., Bottou, L. (2006). Trading convexity for scalability. ICML06, 23rd International Conference on Machine Learning. Pittsburgh, USA.

[13] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In International Conference on Machine Learning (ICML), 2009

[14] B. Sriperumbudur and G. Lanckriet, On the convergence of the concave-convex procedure, in Neural Information Processing Systems, 2009.