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 component 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 descent 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:

\[
\min_x u(x) - v(x) \\
\text{s.t. } f_i(x) \leq 0, \ i = 1..p \\
g_j(x) = 0, \ j = 1..q
\]

(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:

\[
x^{t+1} \in \arg\min_x u(x) - \nabla v(x^t)^T x \\
\text{s.t. } f_i(x) \leq 0, \ i = 1..p \\
g_j(x) = 0, \ j = 1..q
\]

(2)

This procedure is originally proposed by Yuille et al. [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 $L_p$-loss for sparse PCA and SVM feature selection respectively, which results in d.c. programs with convex constraints. In [11], [12], R. Collobert et al. 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 et al. 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 convergence rate of CCCP is still an open problem. [6] and [7] have analyzed the local convergence behavior of Majorization Minimization 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, differentiable 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 piecewise-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 in ramp-loss SVM [11], transductive SVM [12], and structural-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 algorithm, and thus can be viewed as block coordinate decent 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

$$
\begin{align*}
    f(x) &\leq g(x, y), \quad x, y \in \Omega \\
    f(x) &\leq g(x, x), \quad x \in \Omega
\end{align*}
$$

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

$$
    x^{t+1} \in \text{argmin}_{x \in \Omega} g(x, x^t)
$$

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 approximation is less or equal to $v(x)$, with equality when $y = x$. Thus we have

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

(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 $k = \arg \max_x(a_i^T x + b_i)$. Thus we can express d.c. program (1) in another form

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

(6)
Consider the problem

Let the smooth parts of the objective function in (8) and (9), or their equivalent problems, be non-diagonal elements. There exist positive-definite \(v\) and \(H\) for (6) to yield the same sequence \(\{x', y'\}_{t=0}^{\infty}\) as that of a special case of Coordinate Gradient Descent proposed in [1].

3 Convergence Theorem

Lemma 3.1. Consider the problem

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

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

\[
x^{t+1} = \arg\min_x F(x, y^t) \tag{8}
\]

\[
y^{t+1} = \arg\min_y F(x^{t+1}, y) \tag{9}
\]

(a) converges to a stationary point of (7), if in (8) and (9), or their equivalent problems, 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 the 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 \(H_{xx} = \nabla^2 f(x) - 0_n\) for block \(J = x\) and \(H_{yy} = \nabla^2 f(y) - 0_n\) for block \(J = y\), with exact line search (which satisfies Armijo Rule). 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

\[
\min_{d \in \mathbb{R}^m} -c^T d + \frac{\epsilon}{2} \|d\|^2 \tag{10}
\]

subject to \(\sum_{i=1}^m d_i = 1\), \(d_i \geq 0\), \(i = 1..m\)

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

---

1Let set \(S = \arg\min_{x} (-a_i^T x - b_i)\). 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 \(x^{t+1} = \arg\min_x (a(x) - a_k^T x - b_k)\) subject to constraints in (1), where \(a, b\) are averages over \(a_k, b_k\) for \(k \in S\), and \(v(x^*) = a_S\) is a sub-gradient of \(v(x)\) at \(x^*\).

2By “equivalent”, we mean the two problems have the same optimal solution.

3Note, in [1], the Hessian matrix \(H\) affect CGD algorithm only through \(H_{xx}\), so we can always have a positive-definite \(H\) when \(H_{xx}\) is positive definite by assigning, other than \(H_{xx}\), 1 to diagonal elements, 0 to non-diagonal elements.
Proof. Let set $S = \arg \max_{e} (c_{e})$ and $c_{\max} = \max_{i} (c_{i})$. As $\epsilon = 0$, we can obtain a optimal solution $d^{*}$ by setting $d_{e}^{*} \in S$ and $d_{j}^{*} S = 0$. Let $\alpha$, $\beta$, be Lagrange multipliers of affine and non-negative constraints in (10) respectively. By KKT condition, the solution $d^{*}$, $\alpha^{*}$, $\beta^{*}$ must have $-c - \alpha^{*} e - \beta^{*} = 0$, with $\beta_{e} S = 0$, $\alpha^{*} = -c_{\max}$, and $\beta_{j} S = c_{\max} - c_{j}$, where $e = [1 \ldots 1]^{T}$.

When $\epsilon > 0$, the KKT condition for (10) only differs by the equation $cd^{*} - \alpha^{*} e - \beta^{*} = 0$, which can be satisfied by setting $\beta_{e} S = 0$, $\bar{\alpha} = \alpha^{*} + \frac{\epsilon}{|S|}$, and $\bar{\beta}_{j} S = \beta_{j} S - \frac{\epsilon}{|S|}$. If $\frac{\epsilon}{|S|} < \beta_{j} S = c_{\max} - c_{j}$ for $\forall j \not\in S$, then $d^{*}$, $\bar{\alpha}$, $\bar{\beta}$ still satisfy the KKT condition. In other words, $d^{*}$ is still optimal if $\epsilon < \epsilon_{0}$, where $\epsilon_{0} = \min_{j \not\in S} (c_{\max} - c_{j}) |S| > 0$.

When minimizing (6) over $d$, the problem is of form (10) with $\epsilon = 0$ and $c_{i}^{*} = a_{i}^{T} x^{*} + b_{i}, i = 1..m$.

Lemma (3.2) says that we can always find equivalent strictly convex quadratic problem by setting positive $c_{i}^{*} < \min_{j \not\in S} (c_{\max} - c_{j}) |S|$. 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 $f_{i}(x)$, $i \in 0$ 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 $f(x, d) = f_{u}(x) - \sum_{i=1}^{m} d_{i} (a_{i}^{T} x + b_{i})$ is quadratic and $P(x, d) = P_{u}(x) + P_{dom}(x) + P_{dom}(d)$ is polyhedral, lower-semicontinuous, separable for $x$ and $d$. When alternating minimizing $F(x, d)$, the smooth part of subproblem $\min_{x} F(x, d^{*})$ is strictly convex quadratic, and the subproblem $\min_{d} F(x^{*}, d)$, 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 structural-SVM with tighter bound, transductive SVM, ramp-loss SVM and structural-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 101I7501.

**References**


