A Proof of Convergence of the Concave-Convex Procedure Using Zangwill's Theory

Bharath K. Sriperumbudur

bharath@gatsby.ucl.ac.uk

Gatsby Unit, University College London, Alexandra House, 17 Queen Square, London WC1N 3AR, UK.

Gert R. G. Lanckriet

gert@ece.ucsd.edu

Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093-0407, USA.

Keywords: Concave-convex procedure (CCCP), D.c. programs, Zangwill's global convergence theory.

Abstract

The concave-convex procedure (CCCP) is an iterative algorithm that solves d.c. (difference of convex functions) programs as a sequence of convex programs. In machine learning, CCCP is extensively used in many learning algorithms like sparse support vector machines (SVMs), transductive SVMs, sparse principal component analysis, etc. Though widely used in many applications, the convergence behavior of CCCP has not gotten a lot of specific attention. Yuille and Rangarajan analyzed its convergence in their original paper, however we believe the analysis is not complete. On the other hand, the convergence of CCCP can be derived from the convergence of the d.c. algorithm (DCA) — proposed in the global optimization literature to solve general d.c. programs — whose proof relies on d.c. duality. In this paper, we follow a different reasoning and show how Zangwill's *global convergence* theory of iterative algorithms provides a natural framework to prove the convergence of CCCP. This underlines Zangwill's theory as a powerful and general framework to deal with the convergence issues of iterative algorithms, after also being used to prove the convergence of algorithms like Expectation-Maximization, generalized alternating minimization, etc. In this paper, we provide a rigorous analysis of the convergence of CCCP by addressing these questions: (i) When does CCCP find a local minimum or a stationary point of the d.c. program under consideration? (ii) When does the sequence generated by CCCP converge? We also present an open problem on the issue of *local convergence* of CCCP.

1 Introduction

The concave-convex procedure (CCCP) (Yuille and Rangarajan, 2003) is a popularly used algorithm to solve d.c. (difference of convex functions) programs of the form,

$$\min_{x} \left\{ f(x) := u(x) - v(x) : f_i(x) \leq 0,\, i \in [m] \right\}, \tag{1}$$

where $u,\, v$ and $\{f_i\}_{i=1}^m$ are real-valued convex functions, all defined on $\mathbb{R}^n$. Here, $[m] := \{1, \ldots, m\}$. Suppose v is differentiable. The CCCP algorithm is an iterative procedure that solves (1) as the following sequence of convex programs,

$$x^{(l+1)} \in \arg\min_{x} \left\{ u(x) - x^T \nabla v(x^{(l)}) : f_i(x) \leq 0,\, i \in [m] \right\}. \tag{2}$$

As can be seen from (2), the idea of CCCP is to linearize the concave part of f, which is $-v$, around a solution obtained in the current iteration so that $u(x) - x^T \nabla v(x^{(l)})$ is convex in x, and therefore the non-convex program in (1) is solved as a sequence of convex programs as shown in (2). CCCP has been extensively used in solving many non-convex programs (of the form in (1)) that appear in machine learning. For example, Bradley and Mangasarian (1998) proposed a successive linear approximation (SLA) algorithm for feature selection in support vector machines, which can be seen as a special case of CCCP. Other applications where CCCP has been used include sparse principal component analysis (Sriperumbudur et al., 2007), transductive SVMs (Fung and Mangasarian, 2001; Collobert et al., 2006; Wang et al., 2007), feature selection in

SVMs (Neumann et al., 2005), structured estimation (Do et al., 2009), missing data problems in Gaussian processes and SVMs (Smola et al., 2005), etc.

The algorithm in (2) starts at some random point $x^{(0)} \in \Omega := \{x : f_i(x) \leq 0, i \in [m]\}$, iteratively solves the program in (2) and therefore generates a sequence $\{x^{(l)}\}_{l=0}^{\infty}$. The goal of this paper is to study the convergence of $\{x^{(l)}\}_{l=0}^{\infty}$: (i) When does CCCP find a local minimum or a stationary point[1] of the program in (1)? (ii) Does $\{x^{(l)}\}_{l=0}^{\infty}$ converge? If so, to what and under what conditions? From a practical perspective, these questions are highly relevant, given that CCCP is widely applied in machine learning.

In their original CCCP paper, Yuille and Rangarajan (2003, Theorem 2) analyzed its convergence, but we believe the analysis is not complete. They showed that $\{x^{(l)}\}_{l=0}^{\infty}$ satisfies the monotone descent property, i.e., $f(x^{(l+1)}) \leq f(x^{(l)})$ and argued that this descent property ensures the convergence of $\{x^{(l)}\}_{l=0}^{\infty}$ to a minimum or saddle point of the program in (1). However, their analysis is not complete as the monotone descent property by itself is not sufficient to claim the convergence of $\{x^{(l)}\}_{l=0}^{\infty}$.

In the d.c. programming literature, Pham Dinh and Le Thi (1997) proposed a primal-dual subdifferential method called DCA (d.c. algorithm) for solving general d.c. programs of the form $\min\{u(x) - v(x) : x \in \mathbb{R}^n\}$, where it is assumed that u and v are proper lower semi-continuous convex functions, which form a larger class of convex functions than the class of differentiable convex functions (note that in the case of CCCP, v is assumed to be differentiable). Unlike in CCCP, DCA involves constructing two sets of convex programs (called the primal and dual programs) and solving them iteratively in succession such that the solution of the primal is the initialization to the dual and vice-versa. However, when v is differentiable, DCA and CCCP can be shown to be equivalent. Pham Dinh and Le Thi (1997, Theorem 3) provide a proof for the convergence of DCA for general d.c. programs, which therefore proves the convergence of CCCP. The proof exploits d.c. duality (and, as such, follows an approach that is tailored specifically to d.c. programs solved by DCA). We refer the reader to Section 2 for a brief review of d.c. duality and a summary of convergence results for DCA.

[1] x_* is said to be a stationary point of a constrained optimization problem if it satisfies the corresponding Karush-Kuhn-Tucker (KKT) conditions. Assuming constraint qualification, KKT conditions are necessary for the local optimality of x_*. See Bonnans et al. (2006, Section 11.3) for details.

In this paper, we follow a fundamentally different approach and show that the convergence of CCCP, specifically, can be analyzed by relying on Zangwill's *global convergence* theory of iterative algorithms. The tools employed in our proof are of completely different flavor than the ones used in the proof of DCA convergence: DCA convergence analysis exploits d.c. duality while we use the notion of point-to-set maps as introduced by Zangwill. Zangwill's theory is a powerful and general framework to deal with the convergence issues of iterative algorithms. It has also been used to prove the convergence of the Expectation-Maximization (EM) algorithm (Wu, 1983), generalized alternating minimization algorithms (Gunawardana and Byrne, 2005), multiplicative updates in non-negative quadratic programming (Sha et al., 2007), etc. and is therefore a natural framework to analyze the convergence of CCCP in a direct way.

The paper is organized as follows. Following Pham Dinh and Le Thi (1997, 1998), in Section 2, we review d.c. duality and summarize the convergence results obtained for DCA. In Section 3, we present Zangwill's theory of global convergence, which is a general framework to analyze the convergence behavior of iterative algorithms. This theory is used to address the *global convergence* of CCCP in Section 4.1. This involves analyzing the *fixed points* of the CCCP algorithm in (2) and then showing that the fixed points are the stationary points of the program in (1). The results in Section 4.1 are extended in Section 4.2 to analyze the convergence of the *constrained concave-convex procedure* that was proposed by Smola et al. (2005) to deal with d.c. programs involving d.c. constraints (note that in contrast, CCCP in (2) deals with convex constraints). We briefly discuss the *local convergence* issues of CCCP in Section 5 and conclude the section with an example and an open question.

2 Review of D.C. Duality, DCA and Convergence of DCA

Let $X = \mathbb{R}^n$, which means the dual space Y of X can be identified with X itself. Suppose $\Gamma_0(X)$ is the set of all proper lower semicontinuous convex functions on X. The conjugate function u^* of $u \in \Gamma_0(X)$ is a function belonging to $\Gamma_0(Y)$, defined as

$$u^*(y) = \sup\{x^T y - u(x) : x \in X\}.$$

Pham Dinh and Le Thi (1998) considered d.c. programs of the form

$$\alpha = \inf\{f(x) := u(x) - v(x) : x \in X\}, \tag{P}$$

where $u, v \in \Gamma_0(X)$. Note that (P) can handle the minimization of f over a closed convex subset, C of X. This is because the constraint set, $\{x : x \in C\}$ can be absorbed into the objective function through its indicator, $\chi_C(x) = 0$ if $x \in C$, $+\infty$ otherwise — the modified objective would be $(u + \chi_C) - v$. Using the definition of conjugate functions, we have

$$\begin{aligned} \alpha &= \inf\{u(x) - v(x) : x \in X\} = \inf\{u(x) - \sup\{x^T y - v^*(y) : y \in Y\} : x \in X\} \\ &= \inf\{\beta(y) : y \in Y\}, \end{aligned}$$

where $\beta(y) = \inf\{u(x) - (x^T y - v^*(y)) : x \in X\}$. It is clear that $\beta(y) = v^*(y) - u^*(y)$ if $y \in \text{dom}\, v^*$, $+\infty$ otherwise. Therefore, the dual problem can be written as

$$\alpha = \inf\{v^*(y) - u^*(y) : y \in \text{dom}\, v^*\},$$

which is equivalent to

$$\alpha = \inf\{v^*(y) - u^*(y) : y \in Y\}. \tag{D}$$

The perfect symmetry between the primal and dual programs (P) and (D) — the dual to (D) is exactly (P) — is referred to as the d.c. duality.

Based on the above d.c. duality, Pham Dinh and Le Thi (1997, 1998) proposed DCA, which involves constructing two sets of sequences $\{x^{(l)}\}$ and $\{y^{(l)}\}$, starting from a given point $x^{(0)} \in \text{dom}\, u$, by setting

$$y^{(l)} \in \partial v(x^{(l)}); \;\; x^{(l+1)} \in \partial u^*(y^{(l)}),$$

where $\partial v(x^{(0)})$ is the subdifferential of v at $x^{(0)}$, i.e., $\partial v(x^{(0)}) = \{y \in \mathbb{R}^n : v(x) \geq v(x^{(0)}) + (x - x^{(0)})^T y, \forall\, x \in \mathbb{R}^n\}$. Lemma 3.6 in Pham Dinh and Le Thi (1998) shows that the sequences $\{x^{(l)}\}$ and $\{y^{(l)}\}$ are well-defined if and only if $\text{dom}\, \partial u \subset \text{dom}\, \partial v$ and $\text{dom}\, \partial v^* \subset \text{dom}\, \partial u^*$. DCA can be interpreted as follows: at each iteration l, we have

$$y^{(l)} \in \partial v(x^{(l)}) = \arg\min\{v^*(y) - (u^*(y^{(l-1)}) + (y - y^{(l-1)})^T x^{(l)}) : y \in Y\}, \tag{D$_l$}$$

$$x^{(l+1)} \in \partial u^*(y^{(l)}) = \arg\min\{u(x) - (v(x^{(l)}) + (x - x^{(l)})^T y^{(l)}) : x \in X\}. \qquad (\mathrm{P}_l)$$

Note that (P$_l$) is a convex program obtained from (P) by replacing v with its affine minorization defined by $y^{(l)} \in \partial v(x^{(l)})$. Similarly, the convex problem (D$_l$) is obtained from (D) by using the affine minorization of u^* defined by $x^{(l+1)} \in \partial u^*(y^{(l)})$. Suppose v is differentiable. Then DCA reduces to CCCP as shown below:

$$\begin{aligned} x^{(l+1)} \in \partial u^*(\nabla v(x^{(l)})) &= \arg\min\{u(x) - (v(x^{(l)}) + (x - x^{(l)})^T \nabla v(x^{(l)})) : x \in X\} \\ &= \arg\min\{u(x) - x^T \nabla v(x^{(l)}) : x \in X\}. \end{aligned}$$

We now summarize the convergence of DCA for general d.c. programs. To do that, we need some definitions. A point x^* is said to be a *critical point* of $u - v$ if $\partial u(x^*) \cap \partial v(x^*) \neq \emptyset$. If u and v are differentiable, then x^* is a stationary point of $u - v$ as $\nabla u(x^*) = \nabla v(x^*)$. Let $\rho \geq 0$ and C be a convex subset of X. A function $\theta : C \to \mathbb{R} \cup \{+\infty\}$ is said to be ρ-*convex* if

$$\theta(\lambda x + (1-\lambda)x') \leq \lambda\theta(x) + (1-\lambda)\theta(x') - \frac{\lambda(1-\lambda)}{2}\rho\|x - x'\|_2^2, \;\forall\, \lambda \in (0,1),\; \forall\, x, x' \in C,$$

where $\|x\|_2^2 = x^T x$. This is equivalent to saying that $\theta - \frac{\rho}{2}\|\cdot\|_2^2$ is convex on C. The modulus of strong convexity of θ on C, denoted by $\rho(\theta, C)$ is given by $\rho(\theta, C) = \sup\{\rho \geq 0 : \theta - \frac{\rho}{2}\|\cdot\|_2^2 \text{ is convex on } C\}$. θ is said to be *strongly convex* on C if $\rho(\theta, C) > 0$.

Convergence of DCA: (a) Pham Dinh and Le Thi (1998, Theorem 3.7) showed that DCA is a descent method for both (P) and (D): $(u - v)(x^{(l+1)}) \leq (v^* - u^*)(y^{(l)}) \leq (u - v)(x^{(l)})$ with equality if $x^{(l)} \in \partial u^*(y^{(l)})$, $y^{(l)} \in \partial v(x^{(l)})$, u, v are strongly convex on X and u^*, v^* are strongly convex on Y. In addition, when equality holds, $x^{(l)}$ and $y^{(l)}$ are the critical points of (P) and (D) respectively.

(b) If α is finite, then the decreasing sequences $\{(u-v)(x^{(l)})\}$ and $\{(v^*-u^*)(y^{(l)})\}$ converge to the same limit $\beta \geq \alpha$, i.e., $\lim_{l\to\infty}(u - v)(x^{(l)}) = \lim_{l\to\infty}(v^* - u^*)(y^{(l)}) = \beta$. In addition, if the sequences $\{x^{(l)}\}$ and $\{y^{(l)}\}$ are bounded, then for every limit point x^* of $\{x^{(l)}\}$ (resp. y^* of $\{y^{(l)}\}$) there exists a limit point y^* of $\{y^{(l)}\}$ (resp. x^* of $\{x^{(l)}\}$) such that $(u - v)(x^*) = (v^* - u^*)(y^*) = \beta$. This means, every limit point x^* of $\{x^{(l)}\}$ is a critical point of $u - v$.

(c) The convergence of the whole sequence $\{x^{(l)}\}$ (resp. $\{y^{(l)}\}$) can be ensured if the following hold: (i) $\{x^{(l)}\}$ is bounded, (ii) the set of limit points of $\{x^{(l)}\}$ is finite and (iii) $\lim_{l\to\infty}\|x^{(l+1)} - x^{(l)}\| = 0$.

Having summarized the convergence properties of DCA (and therefore of CCCP), in Section 4, we state and prove the convergence results for CCCP using a completely different framework, i.e., Zangwill's global convergence theory, which is briefly discussed in the following section.

3 Global Convergence Theory of Iterative Algorithms

For an iterative procedure like CCCP to be useful, it must converge to a local optimum or a stationary point from all or at least a significant number of initialization states and not exhibit other nonlinear system behaviors, such as divergence or oscillation. This behavior can be analyzed by using the global convergence theory of iterative algorithms developed by Zangwill (1969). Note that the word "global convergence" is a misnomer. We will clarify it below and also introduce some notation and terminology.

To understand the convergence of an iterative procedure like CCCP, we need to understand the notion of a *set-valued mapping*, or *point-to-set mapping*, which is central to the theory of global convergence.[2] A point-to-set map Ψ from a set X into a set Y is defined as $\Psi : X \to \mathscr{P}(Y)$, which assigns a subset of Y to each point of X, where $\mathscr{P}(Y)$ denotes the power set of Y. We introduce some definitions related to the properties of point-to-set maps that will be used later. Suppose X and Y are two topological spaces. A point-to-set map Ψ is said to be *closed* at $x_0 \in X$ if $x_k \to x_0$ as $k \to \infty$, $x_k \in X$ and $y_k \to y_0$ as $k \to \infty$, $y_k \in \Psi(x_k)$, imply $y_0 \in \Psi(x_0)$. This concept of *closure* generalizes the concept of continuity for ordinary point-to-point mappings. A point-to-set map Ψ is said to be closed on $S \subset X$ if it is closed at every point of S. A *fixed point* of the map $\Psi : X \to \mathscr{P}(X)$ is a point x for which $\{x\} = \Psi(x)$, whereas a *generalized fixed point* of Ψ is a point for which $x \in \Psi(x)$. Ψ is said to

[2] Note that depending on the objective and constraints, the minimizer of (2) need not be unique. Therefore, the algorithm takes $x^{(l)}$ as its input and returns a set of minimizers from which an element, $x^{(l+1)}$ is chosen. Hence, the notion of point-to-set maps appears naturally in such iterative algorithms.

be *uniformly compact* on X if there exists a compact set H independent of x such that $\Psi(x) \subset H$ for all $x \in X$. Note that if X is compact, then Ψ is uniformly compact on X. Let $\phi : X \to \mathbb{R}$ be a continuous function. Ψ is said to be *monotone* with respect to ϕ whenever $y \in \Psi(x)$ implies that $\phi(y) \leq \phi(x)$. If, in addition, $y \in \Psi(x)$ and $\phi(y) = \phi(x)$ imply that $y = x$, then we say that Ψ is *strictly monotone*.

Many iterative algorithms in mathematical programming can be described using the notion of point-to-set maps. Let X be a set and $x_0 \in X$ a given point. Then, an *algorithm*, $\mathcal{A}$, with initial point x_0 is a point-to-set map $\mathcal{A} : X \to \mathscr{P}(X)$ which generates a sequence $\{x_k\}_{k=1}^{\infty}$ via the rule $x_{k+1} \in \mathcal{A}(x_k)$, $k = 0, 1, \ldots$. $\mathcal{A}$ is said to be *globally convergent* if *for any chosen initial point* x_0, the sequence $\{x_k\}_{k=0}^{\infty}$ generated by $x_{k+1} \in \mathcal{A}(x_k)$ (or a subsequence) converges to a point for which a necessary condition of optimality holds. The property of global convergence expresses, in a sense, the certainty that the algorithm works. It is very important to stress the fact that it does not imply (contrary to what the term might suggest) convergence to a global optimum for all initial points x_0.

With the above mentioned concepts in place, we now state Zangwill's global convergence theorem (Zangwill, 1969, Convergence theorem A, page 91).

Theorem 1 (Zangwill (1969))**.** *Let $\mathcal{A} : X \to \mathscr{P}(X)$ be a point-to-set map (an algorithm) that given a point $x_0 \in X$ generates a sequence $\{x_k\}_{k=0}^{\infty}$ through the iteration $x_{k+1} \in \mathcal{A}(x_k)$. Also let a solution set $\Gamma \subset X$ be given. Suppose*

(1) All points x_k are in a compact set $S \subset X$.

(2) There is a continuous function $\phi : X \to \mathbb{R}$ such that:

(a) $x \notin \Gamma \Rightarrow \phi(y) < \phi(x), \forall\, y \in \mathcal{A}(x)$,

(b) $x \in \Gamma \Rightarrow \phi(y) \leq \phi(x), \forall\, y \in \mathcal{A}(x)$.

(3) $\mathcal{A}$ is closed at x if $x \notin \Gamma$.

Then, the limit of any convergent subsequence of $\{x_k\}_{k=0}^{\infty}$ is in Γ. Furthermore,

$$\lim_{k \to \infty} \phi(x_k) = \phi(x_*)$$

for all limit points x_*.

The general idea when proving the global convergence of an algorithm, $\mathcal{A}$ is to invoke Theorem 1 by appropriately defining ϕ and Γ. For an algorithm $\mathcal{A}$ that solves the minimization problem, $\min\{f(x) : x \in \Omega\}$, the solution set, Γ is usually chosen to be the set of corresponding stationary points and ϕ can be chosen to be the objective function itself, i.e., f, if f is continuous. In Theorem 1, the convergence of $\phi(x_k)$ to $\phi(x_*)$ does not automatically imply the convergence of x_k to x_*. However, if $\mathcal{A}$ is strictly monotone with respect to ϕ, then Theorem 1 can be strengthened by using the following result due to Meyer (1976, Theorem 3.1, Corollary 3.2).

Theorem 2 (Meyer (1976))**.** *Let $\mathcal{A} : X \to \mathscr{P}(X)$ be a point-to-set map such that $\mathcal{A}$ is uniformly compact, closed and strictly monotone on X, where X is a closed subset of $\mathbb{R}^n$. If $\{x_k\}_{k=0}^{\infty}$ is any sequence generated by $\mathcal{A}$, then all limit points will be fixed points of $\mathcal{A}$, $\phi(x_k) \to \phi(x_*) =: \phi^*$ as $k \to \infty$, where x_* is a fixed point, $\|x_{k+1} - x_k\| \to 0$, and either $\{x_k\}_{k=0}^{\infty}$ converges or the set of limit points of $\{x_k\}_{k=0}^{\infty}$ is connected. Define $\mathscr{F}(a) := \{x \in \mathscr{F} : \phi(x) = a\}$ where $\mathscr{F}$ is the set of fixed points of $\mathcal{A}$. If $\mathscr{F}(\phi^*)$ is finite, then any sequence $\{x_k\}_{k=0}^{\infty}$ generated by $\mathcal{A}$ converges to some x_* in $\mathscr{F}(\phi^*)$.*

Using these results on the global convergence of algorithms, Wu (1983) has studied the convergence properties of the EM algorithm, while Gunawardana and Byrne (2005) analyzed the convergence of generalized alternating minimization procedures. In the following section, we use these results to analyze the convergence of CCCP.

4 Main Results

In Section 4.1, we analyze the global convergence of CCCP. In Section 4.2, we extend these results and present a global convergence theorem for the *constrained concave-convex procedure*, a generalization of CCCP proposed by Smola et al. (2005), to deal with d.c. programs involving d.c. constraints. Proofs for the results in Sections 4.1 and 4.2 are provided in Section 4.3.

4.1 Convergence Theorems for CCCP

To analyze the global convergence of the CCCP algorithm in (2), pertaining to the d.c. program in (1), we consider the point-to-set map $\mathcal{A}_{cccp}$, defined as follows:

$$\mathcal{A}_{cccp}(y) = \arg\min_x \left\{u(x) - x^T \nabla v(y) : x \in \Omega\right\}, \tag{3}$$

where $\Omega := \{x : f_i(x) \leq 0,\, i \in [m]\}$. We now present two global convergence theorems for CCCP.

Theorem 3 (Global convergence of CCCP$-$I)**.** *Let u, $\{f_i\}_{i=1}^m$ be real-valued continuous convex functions and v be a real-valued differentiable convex function, all defined on $\mathbb{R}^n$. Suppose ∇v is continuous. Let $\{x^{(l)}\}_{l=0}^\infty$ be any sequence generated by $\mathcal{A}_{cccp}$ defined by (3). Suppose $\mathcal{A}_{cccp}$ is uniformly compact*[3] *on $\Omega := \{x : f_i(x) \leq 0,\, i \in [m]\}$ and $\mathcal{A}_{cccp}(x)$ is nonempty for every $x \in \Omega$. Then, assuming suitable constraint qualification,*[4] *all the limit points of $\{x^{(l)}\}_{l=0}^\infty$ are generalized fixed points of $\mathcal{A}_{cccp}$, which are stationary points of (1). In addition $\lim_{l\to\infty}(u(x^{(l)}) - v(x^{(l)})) = u(x_*) - v(x_*)$, where x_* is some generalized fixed point of $\mathcal{A}_{cccp}$.*

Remark 4. *(i) Note that if Ω is compact, then $\mathcal{A}_{cccp}$ is uniformly compact on Ω. In addition, since u is continuous on Ω, by the Weierstrass theorem*[5] *(Minoux, 1986), it follows that $\mathcal{A}_{cccp}(x)$ is nonempty for every $x \in \Omega$ and therefore is also closed on Ω (by Lemma 9, see Appendix). Therefore, the assumptions of uniform compactness and nonemptiness of $\mathcal{A}_{cccp}$ are trivially satisfied if Ω is compact.*

(ii) The result obtained in Theorem 3 is similar to the convergence result for DCA but with slightly stronger assumptions. In Theorem 3, we require u to be continuous, v to be

[3] Instead of uniform compactness, one could also assume that for every $x \in \Omega$, the set $H(x) := \{y : u(y) - u(x) \leq v(y) - v(x),\, y \in \mathcal{A}_{cccp}(\Omega)\}$ is bounded, for the claims in Theorem 3 to hold.

[4] Examples include Slater's qualification, Mangasarian-Fromovitz qualification, etc. See Bonnans et al. (2006, p. 201) for details.

[5] The Weierstrass theorem states: If f is a real-valued continuous function on a compact set $K \subset \mathbb{R}^n$, then the problem $\min\{f(x) : x \in K\}$ has an optimal solution $x^* \in K$.

differentiable and ∇v to be continuous while DCA requires u and v to be lower semi-continuous convex functions on $\mathbb{R}^n$. However, the assumptions on u and v as mentioned in Theorem 3 are usually satisfied in machine learning applications, the examples of which include sparse principal component analysis (Sriperumbudur et al., 2007), feature selection in SVMs (Neumann et al., 2005), transductive SVMs (Collobert et al., 2006), etc.

In Theorem 3, we considered the generalized fixed points of $\mathcal{A}_{cccp}$. The disadvantage with this case is that it does not rule out "oscillatory" behavior (Meyer, 1976). To elaborate, let us consider $\{x_*\} \subset \mathcal{A}_{cccp}(x_*)$. For example, let $\Omega_0 = \{x_1, x_2\}$ and let $\mathcal{A}_{cccp}(x_1) = \mathcal{A}_{cccp}(x_2) = \Omega_0$ and $u(x_1) - v(x_1) = u(x_2) - v(x_2) = 0$. Then, the sequence $\{x_1, x_2, x_1, x_2, \ldots\}$ could be generated by $\mathcal{A}_{cccp}$, with the convergent subsequences converging to the generalized fixed points x_1 and x_2. Such an oscillatory behavior can be avoided if we ensure $\mathcal{A}_{cccp}$ to have fixed points instead of generalized fixed points. With appropriate assumptions on u, the following stronger result can be obtained on the convergence of CCCP.

Theorem 5 (Global convergence of CCCP$-$II)**.** *Let u be a real-valued strictly convex function, $\{f_i\}_{i=1}^m$ be real-valued continuous convex functions and v be a differentiable convex function with continuous ∇v, all defined on $\mathbb{R}^n$. Let $\{x^{(l)}\}_{l=0}^\infty$ be any sequence generated by $\mathcal{A}_{cccp}$ defined by (3). Suppose $\mathcal{A}_{cccp}$ is uniformly compact on $\Omega := \{x : f_i(x) \le 0,\, i \in [m]\}$ and $\mathcal{A}_{cccp}(x)$ is nonempty for every $x \in \Omega$. Then, assuming suitable constraint qualification, all the limit points of $\{x^{(l)}\}_{l=0}^\infty$ are fixed points of $\mathcal{A}_{cccp}$, which are stationary points of the d.c. program in (1), $u(x^{(l)}) - v(x^{(l)}) \to u(x_*) - v(x_*) =: f^*$ as $l \to \infty$, for some fixed point x_* (also a stationary point of (1)), $\|x^{(l+1)} - x^{(l)}\| \to 0$, and either $\{x^{(l)}\}_{l=0}^\infty$ converges or the set of limit points of $\{x^{(l)}\}_{l=0}^\infty$ is a connected and compact subset of $\mathscr{S}(f^*)$, where $\mathscr{S}(a) := \{x \in \mathscr{S} : u(x) - v(x) = a\}$ and $\mathscr{S}$ is the set of fixed points of $\mathcal{A}_{cccp}$. If $\mathscr{S}(f^*)$ is finite, then any sequence $\{x^{(l)}\}_{l=0}^\infty$ generated by $\mathcal{A}_{cccp}$ converges to some x_* in $\mathscr{S}(f^*)$.*

Note that the main difference between the assumptions in Theorems 3 and 5 is that u is assumed to be strictly convex in Theorem 5. This is not a strong assumption as it can be achieved as follows. Suppose u is convex but not strictly convex. Let t be a real-valued strictly convex function defined on $\mathbb{R}^n$. Then $\tilde{u} := u + t$ is strictly convex on $\mathbb{R}^n$ and

$$f := u - v = (u + t) - (v + t) =: \tilde{u} - \tilde{v}.$$

If t is continuously differentiable with ∇t continuous (for e.g., $t(x) = \lambda \|x\|_2^2$, $\lambda > 0$), then it is clear that $\tilde{u}$ and $\tilde{v}$ satisfy the conditions in Theorem 5, which means with the same assumptions of Theorem 3, we obtain a stronger result in Theorem 5. However, since Theorem 5 is applied to $\tilde{u}$ and $\tilde{v}$, it has to be noted that the sequence $\{x^{(l)}\}_{l=0}^{\infty}$ is generated by the following point-to-set map

$$\mathcal{A}_{cccp}(y) = \arg\min_x \left\{ u(x) + t(x) - x^T \left(\nabla v(y) + \nabla t(y) \right) : x \in \Omega \right\} \tag{4}$$

instead of (3), which is the point-to-set map corresponding to Theorem 3, which is applied to u and v.

Given the stronger guarantees about the convergence behavior of $\{x^{(l)}\}_{l=0}^{\infty}$ in (4), as provided by Theorem 5, it may be preferable to use (4) instead of (3) to solve (1) when u is convex (but not strictly convex). On the other hand, (3) may be computationally simpler and more efficient to solve than (4) — e.g., if u is linear and Ω is a polyhedral set. In case the latter is more desirable, then Theorem 3 can be used to provide convergence guarantees and therefore, Theorem 3 is not completely redundant.

From Theorem 5, it should be clear that convergence of $f(x^{(l)})$ to f^* does not automatically imply the convergence of $x^{(l)}$ to x_*. The convergence in the latter sense requires more stringent conditions like the finiteness of the set of stationary points of (1) that assume the value of f^*. Note that a similar condition of the set of limit points of $\{x^{(l)}\}$ being finite is also required for the convergence of the whole DCA sequence.

4.2 Extensions

So far, we have considered d.c. programs where the constraint set is convex and analyzed the global convergence behavior of CCCP — using Zangwill's theory — that is used to solve such programs. In the following, we consider general d.c. programs where the constraints need not be convex and present the global convergence analysis (using Zangwill's theory) of an iterative algorithm (which is an extension of CCCP) that solves such general d.c. programs. Note that DCA can be used to solve such general d.c. programs (see (5)),[6] whose convergence properties are summarized in Section 2.

Let us consider a general d.c. program (Horst and Thoai, 1999), given by

$$\min_x \{u_0(x) - v_0(x) : u_i(x) - v_i(x) \le 0,\, i \in [m]\}, \tag{5}$$

where $\{u_i\}_{i=0}^m$, $\{v_i\}_{i=0}^m$ are real-valued continuous convex functions defined on $\mathbb{R}^n$ with $\{v_i\}_{i=0}^m$ being continuously differentiable. While dealing with kernel methods for missing variables, Smola et al. (2005) encountered a problem of the form in (5) for which they proposed a *constrained concave-convex procedure* given by

$$x^{(l+1)} \in \arg\min_x \left\{u_0(x) - \widehat{v_0}(x; x^{(l)}) : u_i(x) - \widehat{v_i}(x; x^{(l)}) \le 0,\, i \in [m]\right\}, \tag{6}$$

where

$$\widehat{v_i}(x; x^{(l)}) := v_i(x^{(l)}) + (x - x^{(l)})^T \nabla v_i(x^{(l)}).$$

Note that, similar to CCCP, the algorithm in (6) is a sequence of convex programs. Although Smola et al. (2005, Theorem 1) provided some convergence analysis for the

[6] While (5) is not a d.c. program in the sense of (1) where the constraint set is convex, an exact penalty approach can be used to transform (5) into a d.c. program. Consider a modified form of (5),

$$\min_x \{u_0(x) - v_0(x) : u_i(x) - v_i(x) \le 0,\, i \in [m], x \in K\}, \tag{A}$$

where K is a non-empty closed convex set in $\mathbb{R}^n$. A penalty approach penalizes the constraints and introduces the following non-differentiable d.c. program:

$$\min_x \left\{ u_0(x) + t\sum_{i=1}^m \max\left(u_i(x), v_i(x)\right) - \left(v_0(x) + t\sum_{i=1}^m v_i(x)\right) \,:\, x \in K \right\}, \tag{B}$$

with $t > 0$, which can be solved using DCA. To solve general d.c. programs via DCA on (B) — note that to apply DCA to (B), the continuity and differentiability conditions on $\{u_i\}_{i=0}^m$ and $\{v_i\}_{i=0}^m$ mentioned in the paragraph following (5) are not needed — exact penalty must hold, i.e., the existence of $t_0 \ge 0$ such that (A) and (B) are equivalent for all $t > t_0$. As far as we know, the existence of such t_0 is guaranteed if K is a non-empty bounded polyhedral convex set in $\mathbb{R}^n$ and the feasible set of (A) is nonempty. For details, we refer the reader to Pham Dinh and Le Thi (1997, Section 8.1) and Le Thi et al. (1999).

algorithm in (6), their analysis is not complete due to the fact that the convergence of $\{x^{(l)}\}_{l=0}^{\infty}$ is assumed. In this section, we provide its convergence analysis, following an approach similar to what we did for CCCP, by considering a point-to-set map $\mathcal{B}_{c-ccp}$, associated with the iterative algorithm in (6), where $x^{(l+1)} \in \mathcal{B}_{c-ccp}(x^{(l)})$. Note that unlike in (2), the constraint set in (6) varies with l and $\{x : u_i(x) - \widehat{v_i}(x; x^{(l)}) \leq 0,\, i \in [m]\} \subset \{x : u_i(x) - v_i(x) \leq 0,\, i \in [m]\} =: \Omega$ for any $x^{(l)}$, which therefore implies $x^{(l+1)} \in \Omega$. In Theorem 6, we provide the global convergence result for the constrained concave-convex procedure, which is an equivalent version of Theorem 5 for CCCP. Theorem 6 provides a result similar to the convergence result for DCA but under slightly stronger assumptions of $\{u_i\}_{i=0}^{m}$, ∇v_0 being continuous and $\{v_i\}_{i=1}^{m}$ being differentiable on $\mathbb{R}^n$.

Theorem 6 (Global convergence of constrained CCP)**.** *Let u_0 be a real-valued continuous and strictly convex function, $\{u_i\}_{i=1}^{m}$ be real-valued continuous convex functions and $\{v_i\}_{i=0}^{m}$ be real-valued convex differentiable functions with continuous ∇v_0, all defined on $\mathbb{R}^n$. Let $\{x^{(l)}\}_{l=0}^{\infty}$ be any sequence generated by $\mathcal{B}_{c-ccp}$ defined in (6). Suppose $\mathcal{B}_{c-ccp}$ is uniformly compact on $\Omega := \{x : u_i(x) - v_i(x) \leq 0,\, i \in [m]\}$ and $\mathcal{B}_{c-ccp}(x)$ is nonempty for every $x \in \Omega$. Then, assuming suitable constraint qualification, all the limit points of $\{x^{(l)}\}_{l=0}^{\infty}$ are fixed points of $\mathcal{B}_{c-ccp}$, which are stationary points of the d.c. program in (5), $u_0(x^{(l)}) - v_0(x^{(l)}) \to u_0(x_*) - v_0(x_*) =: f^*$ as $l \to \infty$, for some fixed point, x_* of $\mathcal{B}_{c-ccp}$ (also a stationary point of (5)), $\|x^{(l+1)} - x^{(l)}\| \to 0$, and either $\{x^{(l)}\}_{l=0}^{\infty}$ converges or the set of limit points of $\{x^{(l)}\}_{l=0}^{\infty}$ is a connected and compact subset of $\mathscr{S}(f^*)$, where $\mathscr{S}(a) := \{x \in \mathscr{S} : u_0(x) - v_0(x) = a\}$ and $\mathscr{S}$ is the set of fixed points of $\mathcal{B}_{c-ccp}$. If $\mathscr{S}(f^*)$ is finite, then any sequence $\{x^{(l)}\}_{l=0}^{\infty}$ generated by $\mathcal{B}_{c-ccp}$ converges to some x_* in $\mathscr{S}(f^*)$.*

In the following section, we present the proofs of Theorems 3, 5 and 6.

4.3 Proofs

Proof of Theorem 3. The assumption of $\mathcal{A}_{cccp}$ being uniformly compact on Ω ensures that condition (1) in Theorem 1 is satisfied. Let Γ be the set of all generalized fixed points of $\mathcal{A}_{cccp}$ and let $\phi = f = u - v$. Because of the descent property, $f(x^{(l+1)}) \leq f(x^{(l)})$ as shown in Yuille and Rangarajan (2003), condition (2) in Theorem 1 is satisfied. By our assumption on u and v, we have $g(x,y) := u(x) - x^T\nabla v(y)$ is continuous in x and y. Therefore, by Lemma 9 (see Appendix), the assumption of nonemptiness of $\mathcal{A}_{cccp}(x)$ for every $x \in \Omega$ ensures that $\mathcal{A}_{cccp}$ is closed on Ω and so satisfies condition (3) in Theorem 1. Therefore, by Theorem 1, all the limit points of $\{x^{(l)}\}_{l=0}^{\infty}$ are the generalized fixed points of $\mathcal{A}_{cccp}$ and $\lim_{l\to\infty}(u(x^{(l)}) - v(x^{(l)})) = u(x_*) - v(x_*)$, where x_* is some generalized fixed point of $\mathcal{A}_{cccp}$. We now show that any generalized fixed point of $\mathcal{A}_{cccp}$ is a stationary point of (1), therefore proving the result.

Suppose x_* is a generalized fixed point of $\mathcal{A}_{cccp}$, i.e., $x_* \in \mathcal{A}_{cccp}(x_*)$. Since the constraints in (3) are qualified at x_*, there exists Lagrange multipliers $\{\eta_i^*\}_{i=1}^m \subset \mathbb{R}_+$ such that the following KKT conditions hold:

$$\begin{cases} 0 \in \partial u(x_*) - \nabla v(x_*) + \sum_{i=1}^m \eta_i^* \partial f_i(x_*), \\ f_i(x_*) \le 0,\ \eta_i^* \ge 0,\ f_i(x_*)\eta_i^* = 0,\ \forall\, i \in [m]. \end{cases} \tag{7}$$

(7) is exactly the set of KKT conditions of (1) which are satisfied by $(x_*, \{\eta_i^*\})$ and therefore, x_* is a stationary point of (1). □

Proof of Theorem 5. Since u is strictly convex, the strict descent property holds, i.e., $f(x^{(l+1)}) < f(x^{(l)})$ unless $x^{(l+1)} = x^{(l)}$ and therefore $\mathcal{A}_{cccp}$ is strictly monotone with respect to f. The assumption of nonemptiness of $\mathcal{A}_{cccp}(x)$ for every $x \in \Omega$ ensures that $\mathcal{A}_{cccp}$ is closed on Ω (which follows from Lemma 9 in Appendix). By assumption, since $\mathcal{A}_{cccp}$ is uniformly compact on Ω, invoking Theorem 2 provides that all the limit points of $\{x^{(l)}\}_{l=0}^{\infty}$ are fixed points of $\mathcal{A}_{cccp}$, which either converge or form a connected compact set. Since any fixed point of $\mathcal{A}_{cccp}$ is a generalized fixed point which is also a stationary point of (1) (see the proof of Theorem 3), the desired result follows. □

Proof of Theorem 6. The proof is very similar to that of Theorem 5. Note that $u_0(x^{(l+1)}) - v_0(x^{(l+1)}) \le u_0(x^{(l+1)}) - \widehat{v_0}(x^{(l+1)}; x^{(l)}) \le u_0(x^{(l)}) - \widehat{v_0}(x^{(l)}; x^{(l)}) = u_0(x^{(l)}) - v_0(x^{(l)})$. Since u_0 is strictly convex, we have $u_0(x^{(l+1)}) - \widehat{v_0}(x^{(l+1)}; x^{(l)}) < u_0(x^{(l)}) - \widehat{v_0}(x^{(l)}; x^{(l)})$ unless $x^{(l+1)} = x^{(l)}$, which means $u_0(x^{(l+1)}) - v_0(x^{(l+1)}) < u_0(x^{(l)}) - v_0(x^{(l)})$ unless $x^{(l+1)} = x^{(l)}$ and therefore $\mathcal{B}_{c-ccp}$ is strictly monotone. Since u_0 and ∇v_0 are continuous and $\mathcal{B}_{c-ccp}(x)$ is nonempty for every $x \in \Omega$, by invoking Lemma 9 (see Appendix), we obtain that $\mathcal{B}_{c-ccp}$ is closed on Ω. The result therefore follows from Theorem 2 which shows that all the limit points of $\{x^{(l)}\}_{l=0}^{\infty}$ are fixed points of $\mathcal{B}_{c-ccp}$, which either converge or form a connected compact set. We now show that any fixed point of $\mathcal{B}_{c-ccp}$

is a stationary point of (5).

Suppose x_* is a fixed point of $\mathcal{B}_{c-ccp}$ and assume that constraints in (6) are qualified at x_*. Then, there exist Lagrange multipliers $\{\eta_i^*\}_{i=1}^m \subset \mathbb{R}_+$ such that the following KKT conditions hold:

$$\begin{cases} 0 \in \partial u_0(x_*) - \nabla v_0(x_*) + \sum_{i=1}^m \eta_i^*(\partial u_i(x_*) - \nabla v_i(x_*)), \\ u_i(x_*) - v_i(x_*) \le 0,\ \eta_i^* \ge 0,\ i \in [m], (u_i(x_*) - v_i(x_*))\eta_i^* = 0,\ i \in [m], \end{cases} \tag{8}$$

which is exactly the KKT conditions for (5) satisfied by $(x_*, \{\eta_i^*\})$ and, therefore, x_* is a stationary point of (5). □

5 Local Convergence Analysis of CCCP

The study so far has been devoted to the global convergence analysis of CCCP and the constrained concave-convex procedure. As mentioned before, we say an algorithm is globally convergent if for *any* chosen starting point, x_0, the sequence $\{x_k\}_{k=0}^\infty$ generated by $x_{k+1} \in \mathcal{A}(x_k)$ converges to a point for which a necessary condition of optimality holds. In the results so far, we have shown that all the limit points of any sequence generated by CCCP (*resp.* its constrained version) are the stationary points (local extrema or saddle points) of the program in (1) (*resp.* (5)). Suppose that x_0 is chosen such that it lies in an ϵ-neighborhood around a local minimum, x_*. Then, will the CCCP sequence converge to x_*? If so, what is the rate of convergence? These are questions of *local convergence*.

Salakhutdinov et al. (2003) studied the local convergence of bound optimization algorithms (of which CCCP is an example) to compare the rate of convergence of such methods to that of gradient and second-order methods. In their work, they considered the unconstrained version of CCCP with $\mathcal{A}_{cccp}$ as a point-to-point map that is differentiable. They showed that, depending on the curvature of u and v, CCCP will exhibit either quasi-Newton behavior with fast, typically superlinear convergence or extremely slow, first-order convergence behavior. However, extending these results to the constrained setup in (2) is not obvious. The following result due to Ostrowski which can be found in Ortega and Rheinboldt (1970, Theorem 10.1.3) provides a way to study the local convergence of iterative algorithms.

Proposition 7 (Ostrowski)**.** *Suppose that $\Psi : U \subset \mathbb{R}^n \to \mathbb{R}^n$ has a fixed point $x_* \in$ int(U) and Ψ is Fréchet-differentiable at x_*. If the spectral radius, $\rho(\Psi'(x_*))$ of $\Psi'(x_*)$ satisfies $\rho(\Psi'(x_*)) < 1$, and if x_0 is sufficiently close to x_*, then the iterates $\{x_k\}$ defined by $x_{k+1} = \Psi(x_k)$ all lie in U and converge to x_*.*

We now discuss how Proposition 7 can be used to study the local convergence of CCCP. First note that Proposition 7 treats Ψ (in our case, $\mathcal{A}_{cccp}$) as a point-to-point map which can be obtained by choosing u to be strictly convex so that $x^{(l+1)}$ is the unique minimizer of (2). Suppose, we choose x_* in Proposition 7 to be a local minimum of (1). Then, the desired result of local convergence with at least linear rate of convergence is obtained if we show that $\rho(\mathcal{A}'_{cccp}(x_*)) < 1$. However, currently we are not aware of a way to compute the Fréchet differential of $\mathcal{A}_{cccp}$ and, moreover, to impose conditions on the functions in (2) so that $\mathcal{A}_{cccp}$ is a Fréchet-differentiable map. This is an open question coming out of this work. However, in the following, we present a simple example for which $\mathcal{A}_{cccp}$ is Fréchet differentiable and $\rho(\mathcal{A}'_{cccp}(x_*)) < 1$.

Example 8. *Consider the following non-convex program,*

$$\min_x \{x^T Ax + b^T x + c : Cx = d\}, \tag{9}$$

where $A \in \mathbb{S}^n$ (space of $n \times n$ symmetric matrices over $\mathbb{R}$), $b \in \mathbb{R}^n$, $c \in \mathbb{R}$, $C \in \mathbb{R}^{m\times n}$ and $d \in \mathbb{R}^m$. Assume rank$(C) = m$, where $m < n$. Although the objective in (9) need not be convex, it can be written as a difference of convex functions:

$$(\rho\|x\|_2^2 + b^T x + c) - (x^T(\rho I_n - A)x),$$

where $\rho > \max(0, \lambda_{max}(A))$, so that $\rho I_n - A$ is positive definite. Here I_n denotes the $n \times n$ identity matrix and $\lambda_{max}(A)$ is the largest eigenvalue of A. Define $u(x) := \rho\|x\|_2^2 + b^T x + c$ and $v(x) := x^T(\rho I_n - A)x$. Using CCCP, (9) can be solved as

$$x^{(l+1)} = \arg\min_x \left\{\rho\|x\|_2^2 + x^T(b - 2(\rho I_n - A)x^{(l)}) + c : Cx = d\right\}. \tag{10}$$

By solving the Lagrangian of (10), we get

$$x^{(l+1)} = \rho^{-1}(I_n - C^+C)(\rho I_n - A)x^{(l)} + C^+d + (2\rho)^{-1}(C^+C - I_n)b, \tag{11}$$

where $C^+ := C^T(CC^T)^{-1}$. Note that the point-to-point map $\mathcal{A}_{cccp}$, defined as $x^{(l+1)} = \mathcal{A}_{cccp}(x^{(l)})$ in (11), is linear and therefore is Fréchet differentiable at any $x \in \mathbb{R}^n$.

Suppose A, C and ρ are such that $\max_{i\in[n]}(|\lambda_i|) < 1$, where $\{\lambda_i\}$ are the eigenvalues of $\rho^{-1}(I_n - C^+C)(\rho I_n - A)$, then the conditions in Proposition 7 are satisfied. Therefore, if x_* is a local minimum of $\mathcal{A}_{cccp}$, then choosing any $x^{(0)}$ that is sufficiently close to x_* results in a sequence of iterates, $\{x^{(l)}\}$ converging to x_* with a rate of convergence that is at least linear.

6 Conclusion

The concave-convex procedure (CCCP) is widely used in machine learning. In this work, we provide a proof of its global convergence by using results from the global convergence theory of iterative algorithms. The proposed approach is fundamentally different from the approach used for the convergence of DCA. It illustrates the power and generality of Zangwill's global convergence theory as a framework for proving the convergence of iterative algorithms. We also briefly discuss the local convergence of CCCP and present an open question, the settlement of which would address the local convergence behavior of CCCP.

Acknowledgments

The authors thank anonymous reviewers for their constructive comments that greatly improved the manuscript. The work was carried out when B. K. S. was a Ph. D. student at the University of California, San Diego. The authors wish to acknowledge support from the National Science Foundation (grants DMS-MSPA 0625409 and IIS-1054960).

Appendix: Supplementary Result

The following result[7] from Gunawardana and Byrne (2005, Proposition 7) shows that the minimization of a continuous function forms a closed point-to-set map. A similar

[7] Lemma 9 is a slight modification to Proposition 7 of Gunawardana and Byrne (2005) — with exactly the same proof — wherein the latter deals with unconstrained minimization, i.e., the minimization in (12) is over Y, while Lemma 9 deals with constrained minimization carried out over a subset, Ω of Y.

sufficient condition is also provided in Wu (1983, Equation 10).

Lemma 9 (Gunawardana and Byrne (2005))**.** *Given a real-valued continuous function* h *on* $X \times Y$*, define the point-to-set map* $\Psi : X \to \mathscr{P}(\Omega)$ *by*

$$\Psi(x) = \arg \min_{y' \in \Omega \subset Y} h(x, y') = \{y : h(x, y) \leq h(x, y'), \forall\, y' \in \Omega \subset Y\}\,. \tag{12}$$

Then, Ψ *is closed at* x *if* $\Psi(x)$ *is nonempty.*

References

Bonnans, J. F., Gilbert, J. C., Lemaréchal, C., and Sagastizábal, C. A. (2006). *Numerical Optimization: Theoretical and Practical Aspects*. Springer-Verlag.

Bradley, P. S. and Mangasarian, O. L. (1998). Feature selection via concave minimization and support vector machines. In *Proc. 15th International Conf. on Machine Learning*, pages 82–90. Morgan Kaufmann, San Francisco, CA.

Collobert, R., Sinz, F., Weston, J., and Bottou, L. (2006). Large scale transductive SVMs. *Journal of Machine Learning Research*, 7:1687–1712.

Do, C. B., Le, Q. V., Teo, C. H., Chapelle, O., and Smola, A. J. (2009). Tighter bounds for structured estimation. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 281–288.

Fung, G. and Mangasarian, O. L. (2001). Semi-supervised support vector machines for unlabeled data classification. *Optimization Methods and Software*, 15:29–44.

Gunawardana, A. and Byrne, W. (2005). Convergence theorems for generalized alternating minimization procedures. *Journal of Machine Learning Research*, 6:2049–2073.

Horst, R. and Thoai, N. V. (1999). D.c. programming: Overview. *Journal of Optimization Theory and Applications*, 103:1–43.

Le Thi, H. A., Pham Dinh, T., and Le Dung, M. (1999). Exact penalty in d.c. programming. *Vietnam Journal of Mathematics*, 27(2):169–179.

Meyer, R. R. (1976). Sufficient conditions for the convergence of monotonic mathematical programming algorithms. *Journal of Computer and System Sciences*, 12:108–121.

Minoux, M. (1986). *Mathematical Programming: Theory and Algorithms*. John Wiley & Sons Ltd.

Neumann, J., Schnörr, C., and Steidl, G. (2005). Combined SVM-based feature selection and classification. *Machine Learning*, 61:129–150.

Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York.

Pham Dinh, T. and Le Thi, H. A. (1997). Convex analysis approach to d.c. programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355.

Pham Dinh, T. and Le Thi, H. A. (1998). D.c. optimization algorithms for solving the trust region subproblem. *SIAM Journal of Optimization*, 8:476–505.

Salakhutdinov, R., Roweis, S., and Ghahramani, Z. (2003). On the convergence of bound optimization algorithms. In *Proc. 19th Conference in Uncertainty in Artificial Intelligence*, pages 509–516.

Sha, F., Lin, Y., Saul, L. K., and Lee, D. D. (2007). Multiplicative updates for nonnegative quadratic programming. *Neural Computation*, 19:2004–2031.

Smola, A. J., Vishwanathan, S. V. N., and Hofmann, T. (2005). Kernel methods for missing variables. In *Proc. of the Tenth International Workshop on Artificial Intelligence and Statistics*.

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

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

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95–103.

Yuille, A. L. and Rangarajan, A. (2003). The concave-convex procedure. *Neural Computation*, 15:915–936.

Zangwill, W. I. (1969). *Nonlinear Programming: A Unified Approach*. Prentice-Hall, Englewood Cliffs, N.J.