On the Convergence of the Concave-Convex Procedure

Bharath K. Sriperumbudur and Gert R. G. Lanckriet

UC San Diego

OPT 2009
Outline

- Difference of convex functions (d.c.) program
  - Applications in machine learning
- The concave-convex procedure (CCCP)
  - Majorization-minimization (MM) algorithm
- Convergence analysis of CCCP
  - Point-to-set maps
  - Zangwill’s global convergence theorem
- Open question: Local convergence of CCCP.
**D.C. Program**

- **D.c. function**

  Let $\Omega$ be a convex set in $\mathbb{R}^n$. A real valued function $f : \Omega \rightarrow \mathbb{R}$ is called a **d.c. function** on $\Omega$, if there exist two convex functions $u, v : \Omega \rightarrow \mathbb{R}$ such that $f$ can be expressed in the form

  $$f(x) = u(x) - v(x), \; x \in \Omega.$$ 

- **D.c. program**

  $$\min_{x \in \Omega} f_0(x)$$

  $$\text{s.t.} \quad f_i(x) \leq 0, \; i = 1, \ldots, m,$$

  where $f_i = g_i - h_i, \; i = 0, \ldots, m$, are d.c. functions.

- **Computationally hard to solve!!**

- **Applications in machine learning**
  - Sparse PCA, transductive SVMs, feature selection in SVMs, etc.
Sparse Support Vector Machines

Consider

\[
\min_{w \in \mathbb{R}^n} \|\xi\|_1 + \lambda \text{card}(w)
\]

s.t. \(y_i(w^T x_i + b) \geq 1 - \xi_i, \ i = 1, \ldots, n,\)
\(\xi \succeq 0,\)

where \(\lambda > 0.\) Using the approximation
\(\|w\|_\varepsilon := \sum_{i=1}^n \frac{\log(1 + |w_i|\varepsilon^{-1})}{\log(1 + \varepsilon^{-1})}\)
for sufficiently small \(\varepsilon > 0\) as

\[
\text{card}(w) = \lim_{\varepsilon \to 0} \sum_{i=1}^n \frac{\log(1 + |w_i|\varepsilon^{-1})}{\log(1 + \varepsilon^{-1})},
\]

we have

\[
\min_{w \in \mathbb{R}^n} \|\xi\|_1 + \lambda \sum_{i=1}^n \log(|w_i| + \varepsilon)
\]

s.t. \(y_i(w^T x_i + b) \geq 1 - \xi_i, \ i = 1, \ldots, n,\)
\(\xi \succeq 0,\)

which is a \textit{d.c. program}.
The Concave-Convex Procedure

- \( v \): differentiable

- Assume \( \{f_i\}_{i=1}^m \) are convex functions. Define \( \Omega := \{x : f_i(x) \leq 0, \ i = 1, \ldots, m\} \).

Algorithm [Yuille and Rangarajan, 2003]

- Choose \( x^{(0)} \in \Omega \).

- \[ x^{(l+1)} \in \arg \min_{x \in \Omega} u(x) - x^T \nabla v(x^{(l)}), \tag{2} \]

- until convergence.

Goal: analyze the convergence of CCCP.

- When does CCCP find a local minimum or a stationary point of (1)?
- Does \( \{x^{(l)}\}_{l=0}^\infty \) converge? If so, when?
**Majorization-Minimization Algorithm**

Suppose we want to minimize $f$ over $\Omega \in \mathbb{R}^n$. Construct a majorization function $g$ such that

\[
\begin{cases}
  f(x) \leq g(x, y), \ \forall x, y \in \Omega \\
  f(x) = g(x, x), \ \forall x \in \Omega
\end{cases}
\]

$g$ as a function of $x$ is an upper bound on $f$ and coincides with $f$ at $y$.

**Algorithm** [Hunter and Lange, 2004]

- Choose $x^{(0)} \in \Omega$.

- $x^{(l+1)} \in \arg \min_{x \in \Omega} g(x, x^{(l)})$,

- until $x^{(l)} \in \arg \min_{x \in \Omega} g(x, x^{(l)})$.

\[
f(x^{(l+1)}) \leq g(x^{(l+1)}, x^{(l)}) \leq g(x^{(l)}, x^{(l)}) = f(x^{(l)}).
\]
Linear Majorization

- $f = u - v$
- $u$ and $v$ real-valued convex functions on $\mathbb{R}^n$. 
- $v$ is differentiable.
- $f(x) \leq u(x) - v(y) - (x - y)^T \nabla v(y) =: g(x, y)$. 
- What we get is CCCP.
Convergence Analysis of CCCP

- Since $f(x^{(l+1)}) \leq f(x^{(l)})$, [Yuille and Rangarajan, 2003] claimed that $\{x^{(l)}\}_{l=0}^{\infty}$ converges to a local minimum or a saddle point of (1).

- Expectation-Maximization (EM) is a special case of MM and satisfies the descent property.

- [Arslan et al., 1993] showed that EM algorithm may converge to a local minimum.

- Cycling behavior.

Goal: analyze the convergence of CCCP.

- When does CCCP find a local minimum or a stationary point of (1)?
- Does $\{x^{(l)}\}_{l=0}^{\infty}$ converge? If so, when?
Global Convergence of Iterative Algorithms

- **Point-to-set map** from $X$ into $Y$ is defined as $\Psi : X \rightarrow \mathcal{P}(Y)$, which assigns a subset of $Y$ to each point of $X$, where $\mathcal{P}(Y)$ denotes the power set of $Y$.

- **Algorithm**, $\mathcal{A}$ is a point-to-set map, $\mathcal{A} : X \rightarrow \mathcal{P}(X)$, via the rule:

  \[ x_{k+1} \in \mathcal{A}(x_k). \]  

- $\mathcal{A}$ is globally convergent: for any chosen initial point $x_0$, $\{x_k\}_{k=0}^\infty$ generated by (*) converges to a point for which the necessary condition of optimality holds.

- **Global convergence does not imply** convergence to a global optimum for all $x_0$. 
Point-to-set Map

- $X$ and $Y$ are topological spaces.

- $\Psi$ is said to be closed at $x_0 \in X$ if

\[ x_k \xrightarrow{k \to \infty} x_0, \quad x_k \in X \quad \text{and} \quad y_k \xrightarrow{k \to \infty} y_0, \quad y_k \in \Psi(x_k) \implies y_0 \in \Psi(x_0). \]

- $\Psi$ is closed on $S \subset X$ if it is closed at every point of $S$.

- Fixed point of $\Psi : X \to \mathcal{P}(X)$ is a point $x$ for which $\{x\} = \Psi(x)$.

- Generalized fixed point of $\Psi$ is a point for which $x \in \Psi(x)$.

- $\Psi$ is said to be uniformly compact on $X$ if there exists a compact set $H$ independent of $x$ such that $\Psi(x) \subset H$ for all $x \in X$. 
Zangwill’s Global Convergence Theorem

**Theorem ([Zangwill, 1969])**

Let \( A : X \to \mathcal{P}(X) \) 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

\[
    x_{k+1} \in A(x_k).
\]

Also let a solution set \( \Gamma \subset X \) be given. Suppose

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

2. **(2)** There is a continuous function \( \phi : X \to \mathbb{R} \) such that:
   
   - (a) \( x \notin \Gamma \Rightarrow \phi(y) < \phi(x), \forall y \in A(x) \),
   - (b) \( x \in \Gamma \Rightarrow \phi(y) \leq \phi(x), \forall y \in A(x) \).

3. **(3)** \( A \) is closed at \( x \) if \( x \notin \Gamma \).

Then the limit of any convergent subsequence of \( \{x_k\}_{k=0}^\infty \) is in \( \Gamma \). Furthermore, \( \lim_{k \to \infty} \phi(x_k) = \phi(x_*) \) for all limit points \( x_* \).
Global Convergence Theorem for CCCP-I

\[ \mathcal{A}_{\text{cccp}}(y) = \arg \min \{ u(x) - x^T \nabla v(y) : x \in \Omega \}. \]  

(3)

Theorem

- \( u, v \): real-valued differentiable convex functions defined on \( \mathbb{R}^n \).
- \( \nabla v \): continuous
- \( \{ f_i \} \): differentiable convex functions defined on \( \mathbb{R}^n \).
- \( \{ x^{(l)} \}_{l=0}^{\infty} \): any sequence generated by \( \mathcal{A}_{\text{cccp}} \).
- \( \mathcal{A}_{\text{cccp}} \) is uniformly compact on \( \Omega \).
- \( \mathcal{A}_{\text{cccp}}(x) \) is non-empty for any \( x \in \Omega \).

Assuming suitable constraint qualification, all the limit points of \( \{ x^{(l)} \}_{l=0}^{\infty} \) are stationary points of the d.c. program in (1). In addition

\[ \lim_{l \to \infty} (u(x^{(l)}) - v(x^{(l)})) = u(x_*) - v(x_*), \]

where \( x_* \) is some stationary point of \( \mathcal{A}_{\text{cccp}} \).
Proof Idea

- Show that any generalized fixed point of $A_{cccp}$ is a stationary point of (1).

- Analyze the generalized fixed points of $A_{cccp}$.
  - Choose $\Gamma$ to the set of all generalized fixed points of $A_{cccp}$.
  - Let $\phi = u - v$.
  - Invoke Zangwill’s global convergence theorem.

Issues: oscillatory behavior.

- Let $\Omega_0 = \{x_1, x_2\}$ and let $A_{cccp}(x_1) = A_{cccp}(x_2) = \Omega_0$ 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 $A_{cccp}$, with the convergent subsequences converging to the generalized fixed points $x_1$ and $x_2$. 
Global Convergence Theorem for CCCP-II

**Theorem**

- \( u, v : \) real-valued differentiable strictly convex functions defined on \( \mathbb{R}^n \).
- other conditions in Global Convergence Theorem for CCCP-I hold.

Assuming suitable constraint qualification, the following hold:

- All the limit points of \( \{x^{(l)}\}_{l=0}^{\infty} \) 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 stationary point \( x^* \).

- \( \|x^{(l+1)} - x^{(l)}\| \to 0 \), 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 \( \mathcal{S}(f^*) \), where \( \mathcal{S}(a) := \{x \in \mathcal{S} : u(x) - v(x) = a\} \) and \( \mathcal{S} \) is the set of stationary points of (1).

- If \( \mathcal{S}(f^*) \) is finite, then any sequence \( \{x^{(l)}\}_{l=0}^{\infty} \) generated by \( \mathcal{A}_{cccp} \) converges to some \( x^* \) in \( \mathcal{S}(f^*) \).
Extensions

\[
\min_x \quad u_0(x) - v_0(x) \\
\text{s.t.} \quad u_i(x) - v_i(x) \leq 0, \quad i \in 1, \ldots, m,
\]

where \(\{u_i\}, \{v_i\}\) are real-valued convex and differentiable functions defined on \(\mathbb{R}^n\).

**Algorithm (constrained concave-convex procedure)** [Smola et al., 2005]

\[
x^{(l+1)} \in \arg \min_x \quad u_0(x) - \hat{v}_0(x; x^{(l)}) \\
\text{s.t.} \quad u_i(x) - \hat{v}_i(x; x^{(l)}) \leq 0, \quad i \in 1, \ldots, m,
\]

where \(\hat{v}_i(x; x^{(l)}) := v_i(x^{(l)}) + (x - x^{(l)})^T \nabla v_i(x^{(l)})\).
Global Convergence Theorem for Constrained CCP

Theorem

- $\{u_i\}, \{v_i\}$: real-valued differentiable convex functions defined on $\mathbb{R}^n$.
- $\nabla v_0$: continuous
- $\{x^{(l)}\}_{l=0}^{\infty}$: any sequence generated by $\mathcal{B}_{ccp}$ defined in (5).
- $\mathcal{B}_{ccp}$ is uniformly compact on
  $\Omega := \{x : u_i(x) - v_i(x) \leq 0, i = 1, \ldots, m\}$.
- $\mathcal{B}_{ccp}(x)$ is non-empty for any $x \in \Omega$.

Assuming suitable constraint qualification, all the limit points of $\{x^{(l)}\}_{l=0}^{\infty}$ are stationary points of the d.c. program in (4). In addition

$$\lim_{l \to \infty} (u_0(x^{(l)}) - v_0(x^{(l)})) = u_0(x^*_*) - v_0(x^*_*),$$

where $x^*_*$ is some stationary point of $\mathcal{B}_{ccp}$. 
**Local Convergence of CCCP**

*Open question*: Suppose, if $x_0$ is chosen such that it lies in an $\epsilon$-neighborhood around a local minima, $x_\star$, then will the CCCP sequence converge to $x_\star$? If so, what is the rate of convergence?

**Proposition (Ostrowski)**

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

**Remarks:**

- $\Psi$ is a point-to-point map: choose $u$ and $v$ in (1) to be strictly convex.

- **Issue**: differentiability of $A_{cccp}$ and $B_{ccp}$. 
Summary

- Convergence of CCCP is analyzed using the global convergence theory of iterative algorithms.
- Applicable to many iterative algorithms in machine learning.
  - alternating minimization, non-negative matrix factorization, etc.
- Local convergence analysis: open problem.


