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

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

- *D.c. program*

$$\begin{aligned} \min_{x \in \Omega} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0,\ i = 1, \ldots, m, \end{aligned} \tag{1}$$

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

$$\begin{aligned} \min_{w \in \mathbb{R}^n} \quad & \|\xi\|_1 + \lambda \, \mathrm{card}(w) \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i, \; i = 1, \ldots, n, \\ & \xi \succeq 0, \end{aligned}$$

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

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

we have

$$\begin{aligned} \min_{w \in \mathbb{R}^n} \quad & \|\xi\|_1 + \lambda \sum_{i=1}^n \log(|w_i| + \varepsilon) \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i, \; i = 1, \ldots, n, \\ & \xi \succeq 0, \end{aligned}$$

which is a *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 $\Psi : X \longrightarrow \mathscr{P}(Y)$, which assigns a subset of Y to each point of X, where $\mathscr{P}(Y)$ denotes the power set of Y.
- *Algorithm, $\mathcal{A}$* is a point-to-set map, $\mathcal{A} : X \longrightarrow \mathscr{P}(X)$, via the rule:

$$x_{k+1} \in \mathcal{A}(x_k). \qquad (\star)$$

- *$\mathcal{A}$ is globally convergent* : *for any chosen initial point* x_0, $\{x_k\}_{k=0}^{\infty}$ generated by $(\star)$ 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.
- Ψ is said to be closed at $x_0 \in X$ if

$$x_k \xrightarrow{k\to\infty} x_0,\ x_k \in X \textit{ and } y_k \xrightarrow{k\to\infty} y_0,\ y_k \in \Psi(x_k) \Longrightarrow y_0 \in \Psi(x_0).$$

- *Ψ is closed on $S \subset X$* if it is closed at every point of S.
- *Fixed point* of $\Psi : X \to \mathscr{P}(X)$ is a point x for which $\{x\} = \Psi(x)$.
- *Generalized fixed point* of Ψ is a point for which $x \in \Psi(x)$.
- Ψ 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 $\mathcal{A} : X \to \mathscr{P}(X)$ 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_*.

Global Convergence Theorem for CCCP-I

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

Theorem

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

Proof Idea

- Show that *any generalized fixed point of $\mathcal{A}_{cccp}$ is a stationary point of (1).*
- Analyze the generalized fixed points of $\mathcal{A}_{cccp}$.
  - Choose Γ *to the set of all generalized fixed points of $\mathcal{A}_{cccp}$.*
  - Let $\phi = u - v$.
  - Invoke Zangwill's global convergence theorem.

Issues: oscillatory behavior.

- 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

$$\{x_1, x_2, x_1, x_2, \dots\}$$

could be generated by $\mathcal{A}_{cccp}$, 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 $\mathscr{S}(f^*)$, where $\mathscr{S}(a) := \{x \in \mathscr{S} : u(x) - v(x) = a\}$ and $\mathscr{S}$ is the set of stationary points of (1).
- If $\mathscr{S}(f^*)$ is finite, then any sequence $\{x^{(l)}\}_{l=0}^{\infty}$ generated by $\mathcal{A}_{cccp}$ converges to some x_* in $\mathscr{S}(f^*)$.

Extensions

$$\begin{aligned} \min_x \quad & u_0(x) - v_0(x) \\ \text{s.t.} \quad & u_i(x) - v_i(x) \leq 0,\ i \in 1, \ldots, m, \end{aligned} \tag{4}$$

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]

$$\begin{aligned} x^{(l+1)} \in \arg\min_x \quad & u_0(x) - \widehat{v_0}(x; x^{(l)}) \\ \text{s.t.} \quad & u_i(x) - \widehat{v_i}(x; x^{(l)}) \leq 0,\ i \in 1, \ldots, m, \end{aligned} \tag{5}$$

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

Global Convergence Theorem for Constrained CCP

Theorem

- $\{u_i\}$, $\{v_i\}$: real-valued differentiable convex functions defined on $\mathbb{R}^n$.
- ∇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 ϵ-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 $x_* \in int(U)$ and Ψ is Fréchet-differentiable at x_*. If the spectral radius 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_*.

Remarks:

- Ψ is a point-to-point map : choose u and v in (1) to be strictly convex.
- *Issue* : differentiability of $\mathcal{A}_{cccp}$ and $\mathcal{B}_{ccp}$.

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.

References

- Arslan, O., Constable, P. D. L., and Kent, J. T. (1993).
  Domains of convergence for the EM algorithm: a cautionary tale in a location estimation problem.
  Statist. Comput., 3:103–108.
- Hunter, D. R. and Lange, K. (2004).
  A tutorial on MM algorithms.
  The American Statistician, 58:30–37.
- 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*.
- 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.