On Convergence Rate of Concave-Convex Procedure

Ian E.H. Yen, Nanyun Peng, Po-Wei Wang, and Shou-De Lin

National Taiwan University

OPT 2012

Outline

- Difference of Convex Functions (d.c.) Program
  - Applications in SVM literature

- Concave-Convex Procedure (CCCP)
  - Majorization-Minimization (MM) algorithm
  - Block Coordinate Descent (BCD)

- Convergence Analysis
  - Alternative BCD Formulation
  - Convergence Theorem

D.C. Program

Let u(x), v(x), $f_i(x)$ be convex function defined on $\mathbf{R}^n$, $g_j(x)$ be affine function on $\mathbf{R}^n$.
A *Difference of Convex Function* (*D.C.) Program* is defined as:

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

D.C. Program

Let u(x), v(x), $f_i(x)$ be convex function defined on $\mathbf{R}^n$, $g_j(x)$ be affine function on $\mathbf{R}^n$.
A *Difference of Convex Function* (*D.C.) Program* is defined as:

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

Ex. Structural SVM with hidden variables: [C.N.J. Yu and T. Joachims, 2009]

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\left(\max_{(\hat{y},\hat{h})\in\mathcal{Y}\times\mathcal{H}} [\boldsymbol{w}\cdot\Phi(x_i,\hat{y},\hat{h}) + \Delta(y_i,\hat{y},\hat{h})]\right) - C\sum_{i=1}^{n}\left(\max_{h\in\mathcal{H}} \boldsymbol{w}\cdot\Phi(x_i,y_i,h)\right)$$

D.C. Program

Let u(x), v(x), $f_i(x)$ be convex function defined on $\mathbf{R^n}$, $g_j(x)$ be affine function on $\mathbf{R^n}$.
A *Difference of Convex Function* (*D.C.) Program* is defined as:

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

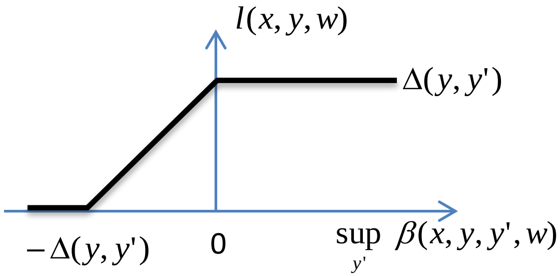
Ex. Structural SVM with hidden variables: [C.N.J. Yu and T. Joachims, 2009]

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\left(\max_{(\hat{y},\hat{h})\in\mathcal{Y}\times\mathcal{H}} [\boldsymbol{w}\cdot\Phi(x_i,\hat{y},\hat{h}) + \Delta(y_i,\hat{y},\hat{h})]\right) - C\sum_{i=1}^{n}\left(\max_{h\in\mathcal{H}} \boldsymbol{w}\cdot\Phi(x_i,y_i,h)\right)$$

Ex. Structural SVM with non-convex tighter bound: [C. B. Do et al., 2009]

$$\min_{w} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} l(x,y,w)$$

$$where\ \ l(x,y,w) = \sup_{y'}[\,\beta(x,y,y',w) + \Delta(y,y')\,] - \sup_{y'} \beta(x,y,y',w)$$



D.C. Program

Let u(x), v(x), $f_i(x)$ be convex function defined on $\mathbf{R^n}$, $g_j(x)$ be affine function on $\mathbf{R^n}$.
A *Difference of Convex Function* (*D.C.) Program* is defined as:

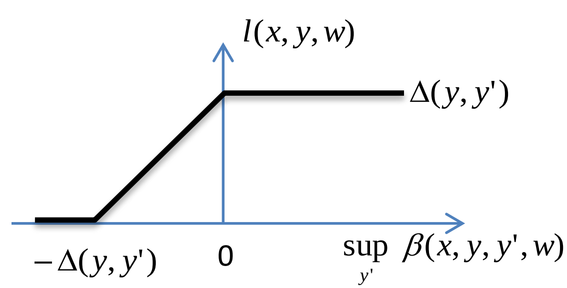
$$\begin{aligned} \min_{x} \quad & u(x) - v(x) \\ s.t. \quad & \mathrm{f}_{\mathrm{i}}(x) \le 0, \ i = 1 \ldots p \\ & g_j(x) = 0, \ j = 1 \ldots q \end{aligned}$$

Convergence rate is hard to analyze in *non-smooth* problem. In this work, we handle the special case when the *smooth part* of u(x) is *strictly convex quadratic*, and v(x) is *piecewise-linear*.

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\left(\max_{(\hat{y},\hat{h})\in\mathcal{Y}\times\mathcal{H}} [\boldsymbol{w}\cdot\Phi(x_i,\hat{y},\hat{h}) + \Delta(y_i,\hat{y},\hat{h})]\right) - C\sum_{i=1}^{n}\left(\max_{h\in\mathcal{H}} \boldsymbol{w}\cdot\Phi(x_i,y_i,h)\right)$$

$$\min_{\mathrm{w}} \ \frac{1}{2}\| w\|^2 + C\sum_{i=1}^{N} l(x,y,w)$$

$$where \ \ l(x,y,w) = \sup_{y'}[\ \beta(x,y,y',w) + \Delta(y,y')\ ] - \sup_{y'} \ \beta(x,y,y',w)$$



Concave-Convex Procedure

Suppose we can compute the *sub-gradient of v(x)*, the *Concave-Convex Procedure (CCCP)* solves a *D.C. Program* by a series of convex problem: [Yuille and Rangarajan, 2003]:

$$
\begin{aligned}
x^{(t+1)} = \; & \arg\min_{\mathrm{x}} \quad u(x) - \nabla v(x^{(t)})^T x \\
s.t. \quad & \mathrm{f_i}(x) \le 0, \; i = 1 \ldots p \\
& g_j(x) = 0, \; j = 1 \ldots q
\end{aligned}
\tag{1}
$$

[Yuille and Rangarajan, 2003] shows (1) guarantees descent of the D.C. Program.

[B. Sriperumbudur et al., 2009] provided *Global Convergence* of (1) via Zanwill's thoery. However, they pointed out the *Local Convergence Rate* of (1) is an open problem.

Goal:

Show that (1) has at least *Linear Convergence Rate* via the connection to more general Block Coordinate Descent (BCD) algorithm.

CCCP as Majorization Minimization (MM)

CCCP is a special case of *Majorization Minimization (MM)*, where we construct a majorization function g(x,y) of objective function f(x)=u(x)-v(x):

$$\begin{cases} f(x) \le g(x,y), \ \ x, y \in \Omega \\ f(x) = g(x,x), \ \ x \in \Omega \end{cases}$$

where Ω is the feasible domain. Then the MM algorithm solves:

$$x^{(t+1)} = \arg\min_{x \in \Omega} \ g(x, x^{(t)}) \qquad (2)$$

CCCP as Majorization Minimization (MM)

CCCP is a special case of *Majorization Minimization (MM)*, where we construct a majorization function g(x,y) of objective function f(x)=u(x)-v(x):

$$\begin{cases} f(x) \le g(x,y), \ \ x, y \in \Omega \\ f(x) = g(x,x), \ \ x \in \Omega \end{cases}$$

where Ω is the feasible domain. Then the MM algorithm solves:

$$x^{(t+1)} = \arg\min_{x \in \Omega} \ g(x, x^{(t)}) \qquad (2)$$

In *CCCP*, g(x,y) is constructed by *1st order Taylor Approximation* of v(x) at point y:

$$\begin{cases} f(x) = u(x) - v(x) \ \ \le \ \ u(x) - v(y) - \nabla v(y)^T (x - y) = g(x,y), \ \text{for } x, y \in \Omega \\ f(x) = g(x,x), \ \text{for } x \in \Omega \end{cases}$$

Therefore,

$$x^{(t+1)} = \arg\min_{\mathrm{x} \in \Omega} \mathrm{g}(\mathrm{x}, \mathrm{x}^{(\mathrm{t})}) = \arg\min_{\mathrm{x} \in \Omega} \ u(x) - \nabla v(x^{(t)})^T x$$

CCCP as Majorization Minimization (MM)

CCCP is a special case of *Majorization Minimization (MM)*, where we construct a majorization function g(x,y) of objective function f(x)=u(x)-v(x):

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

where Ω is the feasible domain. Then the MM algorithm solves:

$$x^{(t+1)} = \arg\min_{x \in \Omega} g(x, x^{(t)}) \qquad (2)$$

[R. Salakhutdinov, 2003] analyzed *local convergence rate* of general MM algorithm by taking (2) as a *differentiable map* $x^{(t+1)} = \psi(x^{(t)})$. However, $\psi(x)$ is not differentiable when there are *constraints* or *non-smooth* function.

Here we took another view of (2) to analyze convergence.

MM as Block Coordinate Descent

Since the minimum of $g(x^{(t)},y)$ occurs at $y=x^{(t)}$, we can view MM algorithm as Block Coordinate Descent over x and y:

$$x^{(t+1)} = \arg\min_{x\in\Omega} g(x, y^{(t)})$$

$$y^{(t+1)} = \arg\min_{y\in\Omega} g(x^{(t+1)}, y) = x^{(t+1)}$$

However, when v(x) is piecewise-linear, the master problem

$$\min_{x,y\in\Omega} \quad g(x,y) = u(x) - v(y) - \nabla v(y)^T (x - y)$$

is discontinuous and hard to analyze.

MM as Block Coordinate Descent

Since the minimum of $g(x^{(t)},y)$ occurs at $y=x^{(t)}$, we can view MM algorithm as Block Coordinate Descent over x and y:

$$x^{(t+1)} = \arg\min_{x\in\Omega} \; g(x, y^{(t)})$$
$$y^{(t+1)} = \arg\min_{y\in\Omega} \; g(x^{(t+1)}, y) = x^{(t+1)}$$

We can take an *alternative formulation* by observing:

$$v(x) = \max_{i} \; (\mathrm{a}_\mathrm{i}^\mathrm{T}\mathrm{x} + \mathrm{b}_\mathrm{i})$$
$$\nabla \mathrm{v(x)} = \mathrm{a}_{\mathrm{k(x)}}, \;\text{ where }\; k(x) = \arg\min_{i} \; (\mathrm{a}_\mathrm{i}^\mathrm{T}\mathrm{x} + \mathrm{b}_\mathrm{i})$$

Block Coordinate Descent over *x and d* on the *alternative formulation*:

$$\min_{\mathrm{x}\in\Omega, d\in R^m} \quad u(x) - \sum_{i=1}^{m} d_i (a_i^T x + b_i)$$
$$s.t. \qquad \sum_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{d_i} = 1 \;\text{ and }\; \mathrm{d_i} \geq 0, \;\; \mathrm{i} = 1...\mathrm{m}$$

yields the same CCCP algorithm.

Block Coordinate Descent for Non-convex, Non-smooth Problem

Lemma 1

Consider the problem:

$$\min_{x,y} F(x,y) = f(x,y) + cP(x,y) \quad (3)$$

where *f(x,y) is smooth* and *P(x,y) is nonsmooth, convex, lower semi-continuous, and separable for x and y*. The *Block Coordinate Descent*

$$x^{(t+1)} = \arg\min_{x} F(x, y^{(t)}) \quad (4)$$

$$y^{(t+1)} = \arg\min_{y} F(x^{(t+1)}, y) \quad (5)$$

Converges to a stationary point of (3) with at least linear rate if the *smooth part* of (4), (5) are *strictly convex quadratic*, f(x,y) is *quadratic*, and P(x,y) is *polyhedral*.

Proof. Since (4), (5) are *strictly convex quadratic*, the BCD correspond to *Coordinate Gradient Descent (CGD)* in [Paul Tseng, etal., 2009] with exact *Hessian matrix* and *line search*. The result holds by Theorem 1, 2, 4 of their paper.

Convergence Theorem of CCCP

Theorem

The *CCCP* converges to stationary point of D.C. Program with *at least linear rate*, if the *non-smooth* part of u(x) and v(x) are *piecewise-linear*, the *smooth part* of u(x) is *strictly convex quadratic*, and the domain Ω is *polyhedral*.

Proof.

The CCCP can be interpreted as *BCD over x and d* of

$$\min_{x\in\Omega, d\in R^m} \quad u(x) - \sum_{i=1}^{m} d_i (a_i^T x + b_i)$$

$$s.t. \qquad \sum_{i=1}^{m} \mathrm{d_i} = 1 \ \text{ and } \ \mathrm{d_i} \geq 0, \ \ \mathrm{i} = 1\ldots\mathrm{m}$$

Which can be also written as

$$\min_{x\in R^n, d\in R^m} \quad \left\{ \mathrm{f}_u(x) - \sum_{i=1}^{m} d_i (a_i^T x + b_i) \right\} + \left\{ \mathrm{P_u}(x) + \mathrm{P}_\Omega(x) + \mathrm{P}(d) \right\}$$

Where smooth part f(x,d) is *quadratic*, and P(x,d) is *polyhedral separable*.
Minimizing over *x*, the problem *strictly convex quadratic*.
Minimizing over *d*, there is equivalent *strictly convex quadratic* problem (Lemma 2 in paper).

Reference

[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 Artificial 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.