Difference Of Convex (DC) Functions and DC Programming

Songcan Chen
Outline

1. A Brief History
2. DC Functions and their Property
3. Some examples
4. DC Programming
5. Case Study
6. Our next work
1. A Brief History

- 1964, Hoang Tuy, (incidentally in his convex optimization paper),
- 1979, J. F. Toland, Duality formulation
- 1985, Pham Dinh Tao, DC Algorithm
- 1990 --, Pham Dinh Tao, et al
- …

Applicable Fields

- For smooth/non-smooth and convex/non-convex optimization problems, especially,
- For large-scale DC problems → Robust and efficient in solving!

Hence
- Machine learning (Clustering, Kernel optimization, Feature selection,...)
- Engineering (Quality control,...)
- ...

2.1 DC Functions

- **Definition 2.1.** Let $C$ be a convex subset of $\mathbb{R}^n$. A real-valued function $f : C \rightarrow \mathbb{R}$ is called DC on $C$, if there exist two convex functions $g, h : C \rightarrow \mathbb{R}$ such that $f$ can be expressed in the form

$$f(x) = g(x) - h(x) \quad (1)$$

$h(x)$ convex $\Rightarrow$ -$h(x)$ concave.

If $C = \mathbb{R}^n$, then $f$ is simply called a DC function.

Notice: DC representation for $f$ is NOT unique, in fact, can have infinite decompositions!
2.2 Their Properties

Let $f$ and $f_i$, $i = 1, \ldots, m$, be DC functions. Then, the following functions are also DC:

1) $\sum_{i=1}^{m} \lambda_i f_i$, $\lambda_i \in \mathbb{R}, i = 1, 2, \ldots, m$. 

2) $\max_{i=1,2,\ldots,m} \{f_i\}$ and $\min_{i=1,2,\ldots,m} \{f_i\}$

3) $|f(x)|$

4) $\prod_{i=1}^{m} f_i$
2.2 Their Properties (Cont’d)

1) Every function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ whose second partial derivatives are continuous everywhere is DC.

2) Let $C$ be a compact convex subset of $\mathbb{R}^n$. Then for any continuous function $c: C \rightarrow \mathbb{R}$ and for any $\varepsilon > 0$, there exists a DC function $f: C \rightarrow \mathbb{R}$ such that

$$|c(x) - f(x)| < \varepsilon, \text{ for any } x \text{ in } C.$$

3) Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be DC, and let $g: \mathbb{R} \rightarrow \mathbb{R}$ be convex. Then, the composite function $(g \circ f)(x) = g(f(x))$ is DC.
3. Some simple examples

1) $x^t Q x$, $Q = A - B$, $A$ and $B$ are positive semi-definite.
2) $x^t y$,
3) Let $d_M$ be a distance function, then $d_M(x) = \inf \{ \|x - y\| : y \in M \}$. 
Proof of 3)

Proof: We have

\[ d_M^2(x) = \inf \{ \|x - y\|^2 : y \in M \} \]
\[ = \|x\|^2 + \inf \{ -\|x\|^2 + \|x - y\|^2 : y \in M \} \]
\[ = \|x\|^2 - \sup \{ \|x\|^2 - \|x - y\|^2 : y \in M \} \]
\[ = \|x\|^2 - \sup \{ 2x^T y - \|y\|^2 : y \in M \}. \]

The norm \( p(x) = \|x\|^2 \) is convex, and the function \( q(x) := \sup \{ 2x^T y - \|y\|^2 : y \in M \} \) is the pointwise supremum of a family of affine functions, and hence convex.
4. DC Programming

• 4.1 Primal Problem
• 4.2 Dual Problem
• 4.3 DC Algorithm (DCA)
4.1 Primal Problem

- A general form

\[(P_{dc}) \quad \alpha = \inf \{ f_0(x) : x \in X \subseteq \mathbb{R}^n, f_i(x) \leq 0, i = 1, 2, \ldots, m \}\]

Where \(f_i = g_i - h_i\), \(i = 1, 2, \ldots, m\) are DC functions and \(X\) is a closed convex subset of \(\mathbb{R}^n\).

Constrained (closed) Set \(X\) can be represented by a convex indicator function which is added to the \(g_0(x)\) \((f_0 = g_0 - h_0)\): \(I_X(x) = 0\) if \(x\) in \(X\), \(+\infty\) otherwise.
4.1 Primal Problem (Cont’d)

When $X$ is constrained by a set of linear inequality equations and the objective function is linear, the optimization problem is called polyhedral DC, solving it amounts to solving a linear programming. For example, selecting features based on SVM and $l_0$ norm [5].
Notations

1) The conjugate function $g^*$ of $g$ is defined by

$$g^*(y) = \sup\{\langle x, y \rangle - g(x) : x \in X\}$$

2) Support Domain of $g(x)$

$$\text{dom } g = \{x \in X : g(x) < +\infty\}$$

3) $\epsilon$-subdifferential of $g(x)$ at $x^0$, when $\epsilon = 0$, simply called subdifferential.

$$\partial_\epsilon g(x^0) = \{y \in Y : g(x) \geq g(x^0) + \langle x - x^0, y \rangle - \epsilon \quad \forall x \in X\}$$
Notations (Cont’d)

Support Domain of the subdifferential \( \partial g \)

\[
\text{dom } \partial g = \{x \in X : \partial g(x) \neq \emptyset\}
\]

Range Domain of \( \partial g \)

\[
\text{range } \partial g = \bigcup \{\partial g(x) : \tilde{x} \in \text{dom } \partial g\}
\]
4.2 Dual Problem

Using the definition of conjugate functions, we have

\[
\alpha = \inf \{ g(x) - h(x) : x \in X \} \\
= \inf \{ g(x) - \sup \{ \langle x, y \rangle - h^*(y) : y \in Y \} : x \in X \} \\
= \inf \{ \beta(y) : y \in Y \}
\]

with

\[
(P_y) \quad \beta(y) = \inf \{ g(x) - (\langle x, y \rangle - h^*(y)) : x \in X \}
\]

\[
\beta(y) = h^*(y) - g^*(y) \text{ if } y \in \text{dom } h^*, \quad +\infty \text{ otherwise.}
\]
4.2 Dual Problem (Cont’d)

Dual Formulation:

\[
(D) \quad \alpha = \inf \{ h^*(y) - g^*(y) : y \in Y \}
\]

Where \( Y = \text{dom} \ \partial h^* \).

A perfect symmetry exists between the primal and its dual programs (P) and (D):

the dual program to (D) is exactly (P).
4.2 Dual Problem (Cont’d)

• The necessary local optimality condition for $P_{dc}$ is
  \[ \partial h(x^*) \text{ in } \partial g(x^*) \]

• A point that $x^*$ that verifies the generalized Kuhn-Tucker condition
  \[ \partial h(x^*) \cap \partial g(x^*) \neq \emptyset \]
  is called a critical point of $g-h$. 
4.3 DCA

**DCA Scheme**

**INPUT**
- Let \( x^0 \in \mathbb{R}^p \) be a best guest, \( 0 \leftarrow k \).

**REPEAT**
- Calculate \( y^k \in \partial h(x^k) \).
- Calculate

\[
x^{k+1} \in \arg\min \left\{ g(x) - h(x^k) - \langle x - x^k, y^k \rangle \mid s.t. x \in \mathbb{R}^p \right\}. \quad (P_k)
\]
- \( k + 1 \leftarrow k \).

**UNTIL**{convergence of \( x^k \).}

Affine majorization of the concave part \(-h(x)\)!
4.3 DCA (Cont’d)

• Different decompositions → thus make trade-off between Complexity of each step,
  • number of iterations.
  • Local convergence, empirically: “good” optima.
4.3 DCA (Cont’d)

Convergence properties

– DCA is a descent method (i.e., the sequences \( \{g(x^k) - h(x^k)\} \) and \( \{h^*(y^k) - g^*(y^k)\} \) are both decreasing) without linesearch;

– If the optimal value \( \alpha \) of problem \((P_{dc})\) is finite and the infinite sequences \( \{x^k\} \) and \( \{y^k\} \) are bounded, then every limit point \( x^* \) (resp. \( y^* \)) of \( \{x^k\} \) (resp. \( \{y^k\} \)) is a critical point of \( g - h \) (resp. \( h^* - g^* \)), i.e., \( \partial h(x^*) \cap \partial g(x^*) = \emptyset \) (resp. \( \partial h^*(y^*) \cap \partial g^*(y^*) = \emptyset \)).

– DCA has a linear convergence for general DC programs.
5. Case Study

• 5.1 Fuzzy c-means Clustering
• 5.2 Feature Selection and Classification
5.1 Fuzzy c-means Clustering

$$\min J_m(U, V) := \sum_{k=1}^{n} \sum_{i=1}^{c} u_{i,k}^m \| x_k - v_i \|^2$$

$$\text{s.t. } u_{i,k} \in [0, 1] \text{ for } i = 1, \ldots, c \quad k = 1, \ldots, n$$

$$\sum_{i=1}^{c} u_{i,k} = 1, \quad k = 1, \ldots, n$$
FCM (Cont’d)

• How to be changed to DC
  1) g and h?
  2) X – a convex set of variables (U, V)?
Characterization of Convex Set

• From the centers’ solution $V=\{v_i, \ i=1,2,\ldots,c\}$,

$$
v_i \sum_{k=1}^{n} u_{i,k}^m = \sum_{k=1}^{n} u_{i,k}^m x_k$$

$$
\|v_i\|^2 \leq \frac{\left( \sum_{k=1}^{n} u_{i,k}^m \|x_k\| \right)^2}{\left( \sum_{k=1}^{n} u_{i,k}^m \right)^2} \leq \sum_{k=1}^{n} \|x_k\|^2 := r^2
$$

Leading to the Euclidean ball $R_i$ with radius $r$. It is convex!

In fact, $\|v_i\| \leq \max\{\|x_k\|, \ k=1, 2, \ldots, n\}$ for all $i$. 
Characterization of Convex Set

• For $U$, let

$$u_{i,k} = t_{i,k}^2$$

• Constraints

$$\sum_{i=1}^{c} u_{i,k} = 1$$

$$\sum_{i=1}^{c} t_{i,k}^2 = 1 \text{ or } \|t_k\|^2 = 1 \text{ with } t_k \in \mathbb{R}^c$$

Leading to the Euclidean sphere $S_k$ with radius 1. It is NOT convex.
Equivalent Formulation to FCM

\[
\begin{align*}
\min J_{2m}(T, V) := & \sum_{k=1}^{n} \sum_{i=1}^{c} t_{i,k}^{2m} \| x_k - v_i \|^2 \\
\text{s.t.} & \quad T \in S := \Pi_{k=1}^{n} S_k, \quad V \in C := \Pi_{i=1}^{c} R_i
\end{align*}
\]

A DC decomposition of the above objective function

\[
J_{2m}(T, V) = \frac{\rho}{2} (\| T \|^2 + \| V \|^2) - \left[ \frac{\rho}{2} \|(T, V)\|^2 - J_{2m}(T, V) \right]
\]
DC Formulation

For all \((T, V) \in \mathcal{S} \times \mathcal{C}\)

\[
J_{2m}(T, V) = \frac{\rho}{2} n + \frac{\rho}{2} \|V\|^2 - H(T, V)
\]

with

\[
H(T, V) := \frac{\rho}{2} \|(T, V)\|^2 - J_{2m}(T, V)
\]

A Question: is \(H(T, V)\) unconditionally convex? \textbf{No!}
Condition ensuring $H(T, V)$

Proposition 1. Let $B := \prod_{k=1}^{n} B_k$, where $B_k$ is the ball of centre 0 and radius 1 in $\mathbb{R}^c$. The function $H(T, V)$ is convex on $B \times C$ for all values of $\rho$ such that

$$\rho \geq \frac{m}{n} (2m - 1) \alpha^2 + 1 + \sqrt{\left[\frac{m}{n} (2m - 1) \alpha^2 + 1\right]^2 + \frac{16}{n} m^2 \alpha^2},$$

(8)

where

$$\alpha = r + \max_{1 \leq k \leq n} \|x_k\|.$$

(9)

Notice here B denotes a Ball and thus is convex!

In fact, $\alpha$ can be $2 \max\{ \|x_k\|, k=1, 2, \ldots, n\}$!
Proof (1)

Proof: from

\[ H(T, V) = \sum_{k=1}^{n} \sum_{i=1}^{c} \left[ \frac{\rho}{2} t_{i,k}^2 + \frac{\rho}{2n} \| v_i \|^2 - t_{i,k}^{2m} \| x_k - v_i \|^2 \right] \]

Just prove the function are convex for all \( i \) and \( k \)

\[ h_{i,k}(t_{i,k}, v_i) := \frac{\rho}{2} t_{i,k}^2 + \frac{\rho}{2n} \| v_i \|^2 - t_{i,k}^{2m} \| x_k - v_i \|^2 \]

Define

\[ f : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \]
\[ f(x, y) = \frac{\rho}{2} x^2 + \frac{\rho}{2n} y^2 - x^{2m} y^2 \]

Its Hessian

\[ J_f(x, y) = \begin{pmatrix} \rho - 2m(2m - 1)y^2 x^{2m-2} - 4mx^{2m-1}y \\ -4mx^{2m-1}y \\ \frac{n}{n} - 2x^{2m} \end{pmatrix} \]
Proof (2)

For all \((x, y): 0 \leq x \leq 1; \|y\| \leq \alpha\)

\[
| J_f(x, y) | = (\rho - 2m(2m - 1)y^2x^{2m-2}) \left(\frac{\rho}{n} - 2x^{2m}\right) - 16m^2x^{4m-2}y^2 \\
\geq \frac{1}{n} \rho^2 - \left[ \frac{2m}{n} (2m - 1)y^2x^{2m-2} + 2x^{2m} \right] \rho - 16m^2x^{4m-2}y^2 \\
\geq \frac{1}{n} \rho^2 - 2 \left( \frac{m}{n} (2m - 1)\alpha^2 + 1 \right) \rho - 16m^2\alpha^2.
\]

So \(f(x, y)\) is convex on \([0, 1] \times [-\alpha, \alpha]\)
Proof (3)

implying

\[ \theta_{i,k}(t_{i,k}, v_i) := \frac{\rho}{2} t_{i,k}^2 + \frac{\rho}{2n} \|x_k - v_i\|^2 - t_{i,k}^2 \|x_k - v_i\|^2 \]

is convex on \( \{0 \leq t_{i,k} \leq 1, \|v_i\| \leq r\} \)

Further \( h_{i,k} \) is convex

\[ h_{i,k}(t_{i,k}, v_i) = \theta_{i,k}(t_{i,k}, v_i) + \frac{\rho}{n} \langle x_k, v_i \rangle - \frac{\rho}{2n} \|x_k\|^2 \]

Finally, the function \( H(T, V) \) is convex on \( B \times C \).
Proof (4)

For all $T \in B$ (closed ball) and a given matrix $V \in C$, the function $J_{2m}(T,V)$ is concave in variable $T$ (since $H(T,V)$ is convex). Hence $S$ (sphere, i.e., boundary) contains minimizers (reaching at boundary) of $J_{2m}(T,V)$ on $B$, i.e.,

$$
\min \left\{ \frac{\rho}{2} \| V \|^2 - H(T,V) : (T,V) \in \mathcal{B} \times \mathcal{C} \right\}
$$

$$
= \min \left\{ \frac{\rho}{2} \| V \|^2 - H(T,V) : (T,V) \in \mathcal{S} \times \mathcal{C} \right\}
$$
DC Formulation

\[
\min \left\{ \frac{\rho}{2} \|V\|^2 - H(T, V) : (T, V) \in \mathcal{B} \times \mathcal{C} \right\}
\]

\[
\min \left\{ \chi_{\mathcal{B} \times \mathcal{C}}(T, V) + \frac{\rho}{2} \|V\|^2 - H(T, V) \right\} \\
\text{s.t. } (T, V) \in \mathbb{R}^{c \times n} \times \mathbb{R}^{c \times p}.
\]

\[
\chi_{\mathcal{B} \times \mathcal{C}}(T, V) + \frac{\rho}{2} \|V\|^2 - H(T, V) := G(T, V) - H(T, V)
\]

where \[
G(T, V) := \chi_{\mathcal{B} \times \mathcal{C}}(T, V) + \frac{\rho}{2} \|V\|^2
\]
Solving FCM by DCA (1)

A key: construct two sequences \((Y^l, Z^l) \in \partial H (T^l, V^l)\) and

\[(T^{l+1}, V^{l+1}) \in \text{arg min} \left\{ \frac{\rho}{2} \|V\|^2 - \langle (T, V), (Y^l, Z^l) \rangle \right\} \ 	ext{s.t.} \ (T, V) \in \mathcal{B} \times \mathcal{C}.\]

\(H\) is differentiable and its gradient at the point \((T^l, V^l)\):

\[
\nabla H(T^l, V^l) = \rho(T^l, V^l) - (2m t_{i,k}^{2m-1} \|x_k - v_i\|^2, 2 \sum_{k=1}^{n} (v_i - x_k) t_{i,k}^{2m})
\]

\[ (14) \]
Algorithm 1. DCA applied to FCM

INPUT
- $T^0 \in \mathbb{R}^{c \times n}$ and $V^0 \in \mathbb{R}^{c \times p}$.
- $l = 0$. Let $\epsilon > 0$ be sufficiently small number.

REPEAT
- Calculate $(Y^l, Z^l) = \nabla H(T^l, V^l)$ via (14);
- Calculate $(T^{l+1}, V^{l+1})$ via (15) and (16);
- $l + 1 \leftarrow l$.

UNTIL $\| (T^{l+1}, V^{l+1}) - (T^l, V^l) \| \leq \epsilon (\| (T^{l+1}, V^{l+1}) \|)$
Solving FCM by DCA (2)

\[ T^{l+1} = \text{Proj}_B(Y^l), \quad V^{l+1} = \text{Proj}_C\left(\frac{1}{\rho} Z^l\right) \]

More precisely:

\[ V_{i, .}^{l+1} = \begin{cases} \frac{(Z^l)_{i, .}}{\rho} \text{ if } \| (Z^l)_{i, .} \| \leq \rho r \\ \frac{1}{\| (Z^l)_{i, .} \|} \text{ otherwise} \end{cases}, \quad i = 1, \ldots, c, \tag{15} \]

\[ T_{. , k}^{l+1} = \begin{cases} \frac{Y_{. , k}^l}{(Y^l)_{. , k}} \text{ if } \| Y_{. , k}^l \| \leq 1 \\ \frac{(Y^l)_{. , k}}{\| (Y^l)_{. , k} \|} \text{ otherwise} \end{cases}, \quad k = 1, \ldots, n. \tag{16} \]
Algorithm 2. Combined FCM-DCA algorithm

INPUT

- Let $U^0$ and $V^0$ be the membership and the cluster centers randomly generated.
- Set $l = 0$. Let $\epsilon > 0$ be sufficiently small number.

REPEAT
i. One iteration of FCM:
Accelerating DCA -- FCM-DCM (2)

- Compute the cluster centers $V^l$ via

$$v_i = \sum_{k=1}^{n} u_{ik}^m x_k / \sum_{k=1}^{n} u_{ik}^m \quad \forall i = 1,..,c. \quad (17)$$

- Compute the membership $U^l$ via

$$u_{ik} = \left[ \sum_{j=1}^{c} \frac{\|x_k - v_i\|^{2/(m-1)}}{\|x_k - v_j\|^{2/(m-1)}} \right]^{-1} \cdot \left[ \sum_{j=1}^{c} \frac{\|x_k - v_i\|^{2/(m-1)}}{\|x_k - v_j\|^{2/(m-1)}} \right]^{-1} \cdot$$

- Set $t_{ik} = \sqrt{u_{ik}}, \forall i = 1,..,c$ and $\forall k = 1,..,n.$
ii. One iteration of DCA:

- Calculate \((Y^l, Z^l) = \nabla H(T^l, V^l)\) via (14);
- Calculate \((T^{l+1}, V^{l+1})\) via (15) and (16);
- \(l + 1 \leftarrow l\)

\[\text{UNTIL}\left\{\| (T^{l+1}, V^{l+1}) - (T^l, V^l) \| \leq \epsilon(\| (T^{l+1}, V^{l+1}) \|) \right\}\]
Two phase algorithm 3

**INPUT**

- Let $U^0$ and $V^0$ be the membership and the cluster centers randomly generated.
- Set $l = 0$. Let $\epsilon > 0$ be sufficiently small number.

**PHASE 1:**

- Perform $q$ iterations of Algorithm 2 for obtaining $(T^{q+1}, V^{q+1})$.
- Update $(T^0, V^0) \leftarrow (T^{q+1}, V^{q+1})$

**PHASE 2:**

- Apply Algorithm 1 from the initial point $(T^0, V^0)$ until the convergence.
Partial results

Table 1. Computation time of FCM Algorithm and Algorithm 2, 3

<table>
<thead>
<tr>
<th>Data</th>
<th>FCM</th>
<th>Algorithm 2</th>
<th>Algorithm 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N^\circ$ Size $c$ $N^\circ F$ Time</td>
<td>$N^\circ I$ Time</td>
<td>$q$ $N^\circ D$ Time</td>
<td></td>
</tr>
<tr>
<td>1 $128^2$ 2 24 1.453</td>
<td>16 1.312 12 10 1.219</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 $128^2$ 2 17 1.003</td>
<td>12 0.985 10 2 0.765</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 $256^2$ 3 36 15.340</td>
<td>24 13.297 20 2 10.176</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 $256^2$ 3 75 31.281</td>
<td>57 30.843 30 12 26.915</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 $256^2$ 3 39 15.750</td>
<td>27 14.687 20 14 13.125</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 $256^2$ 5 91 84.969</td>
<td>75 86.969 40 78 61.500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 $256^2$ 3 73 31.094</td>
<td>62 34.286 15 21 24.188</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 $256^2$ 3 78 34.512</td>
<td>52 32.162 20 13 29.182</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 $512^2$ 3 49 92.076</td>
<td>41 102.589 30 46 74.586</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 $512^2$ 5 246 915.095</td>
<td>196 897.043 120 86 691.854</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fig. 1. The original noisy image and the results of segmentation ($c=3$)

(a) (resp. (b)) corresponds to the original image without (resp. with) noise; (c) (resp. (d)) represents the resulting image given by FCM Algorithm without (resp. with) spatial information. (e) represents the resulting image given by Algorithm 2 without spatial information (f) (resp. (g)) represents the resulting image given by Algorithm 3 without (resp. with) spatial information.
Fig. 2. The medical noisy image and the results of segmentation ($c=3$)
Fig. 3. The medical noisy image and the results of segmentation (c=3)
Fig. 4. The Blume noisy image and the results of segmentation ($c=5$)
5.2 Feature Selection and Classification

Formulation of problem

• Given two finite point sets $A$ and $B$ in $\mathbb{R}^n$ represented by the matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{k \times n}$, respectively. Discriminate these sets by a separating plane $(w \in \mathbb{R}^n, \gamma \in \mathbb{R})$

$$P = \{x \mid x \in \mathbb{R}^n, x^Tw = \gamma\} \quad (1)$$

which uses as few features as possible.
The optimization problem

\[
\begin{align*}
\min_{w, \gamma, y, z} & \quad (1 - \lambda)(\frac{1}{m} e^T y + \frac{1}{k} e^T z) + \lambda \|w\|_0 \\
\text{s.t.} & \quad -Aw + e\gamma + e \leq y \\
& \quad Bw - e\gamma + e \leq z \\
& \quad y \geq 0, \quad z \geq 0.
\end{align*}
\]

(2)

Where \( y_i, i=1,2,\ldots,m \) and \( z_j, j=1,2,\ldots,k \) are non-negative slack variables, \( e \) is a vector with all entries of 1.

The zero-norm:
\[
\|w\|_0 := \text{card} \{w_i : w_i \neq 0\}
\]

Optimization Difficulty of Zero-Norm

- Discontinuity at the origin
- NP-Hard

Solution: Approximation to Zero-norm!
for example,

\[ \|v\|_0 \simeq e^T (e - e^{-\alpha v}) \]
Approximate Zero-norm

\[ \|w\|_0 \approx \sum_{i=1}^{n} \eta(\alpha, w_i). \]

where

\[ \eta(x, \alpha) = \begin{cases} 
1 - \varepsilon^{-\alpha x} & \text{if } x \geq 0 \\
1 - \varepsilon^{\alpha x} & \text{if } x < 0 
\end{cases}, \alpha > 0. \]
Reformulation of the optimization

\[
\min \left\{ \begin{array}{l}
F(y, z, w, \gamma) := (1 - \lambda)\left( \frac{e^T y}{m} + \frac{e^T z}{k} \right) \\
+ \lambda \sum_{i=1}^{n} \eta(w_i) : (y, z, w, \gamma) \in K
\end{array} \right\}
\]

where $K$ is the polyhedral convex set defined by:

\[
K := \left\{ (y, z, w, \gamma) \in \mathbb{R}^{m+k+n+1} : \\
-Aw + e\gamma + e \leq y, \\
Bw - e\gamma + e \leq z
\right\}.
\]
A DC decomposition of the approximation

$$\eta(x) = \overline{g(x)} - \underline{h(x)}.$$ 

where

$$g(x) = \begin{cases} 
\alpha x & \text{if } x \geq 0 \\
-\alpha x & \text{if } x < 0 
\end{cases}$$

$$h(x) = g(x) - \eta(x) = \begin{cases} 
\alpha x - 1 + e^{-\alpha x} & \text{if } x \geq 0 \\
-\alpha x - 1 + e^{\alpha x} & \text{if } x < 0 
\end{cases}$$

They are both convex!
Illustration
DC Decomposition of the Objective

\[ F(y, z, w) := G(y, z, w) - H(y, z, w) \]

where:

\[
G(y, z, w) := (1 - \lambda) \left( \frac{e^T y}{m} + \frac{e^T z}{k} \right) + \lambda \sum_{j=1}^{n} g(w_j)
\]

\[
H(y, z, w) := \lambda \sum_{j=1}^{n} h(w_j)
\]
A Final Formulation: FSC-DC

$$\min \{ G(y,z,w) - H(y,z,w) : (y,z,w,\gamma) \in K \}$$

$$\min \left\{ \chi_K(y,z,w,\gamma) + G(y,z,w) - H(y,z,w) : (y,z,w,\gamma) \in \mathbb{R}^m \times \mathbb{R}^k \times \mathbb{R}^n \times \mathbb{R}, \right\}$$

(13)
DCA Revisited

Generic DCA scheme:
Initialization: Let $x^0 \in \mathbb{R}^p$ be a best guest,
0 $\leftarrow$ $k$.
iteration $k = 0, 1, ...$
Calculate $y^k \in \partial H(x^k)$
Calculate

$$x^{k+1} \in \arg \min \left\{ G(x) - H(x^k) - \langle x - x^k, y^k \rangle : x \in \mathbb{R}^p \right\}$$

$k + 1 \leftarrow k$
Until convergence of $x^k$. 
DCA for FSC

Initialization Let $\tau$ be a tolerance sufficiently small, set $k = 0$.

Choose $(y^0, z^0, w^0, \gamma^0) \in \mathbb{R}^m \times \mathbb{R}^k \times \mathbb{R}^n \times \mathbb{R}$.

Repeat

- Compute $v^k \in \partial H(w^k)$ via (15).
- Solve the linear program (16) to obtain $(y^{k+1}, z^{k+1}, w^{k+1}, \gamma^{k+1})$
- $k + 1 \leftarrow k$

Until

$$\| y^k - y^{k-1} \| + \| z^k - z^{k-1} \| + \| w^k - w^{k-1} \| + \| \gamma^k - \gamma^{k-1} \| \leq \tau (1 + \| y^k \| + \| z^k \| + \| w^k \| + \| \gamma^k \|)$$
(15) And (16)

\[ v_j = \begin{cases} 
\alpha(1 - e^{-\alpha w_j}) & \text{if } w_j \geq 0 \\
-\alpha(1 - e^{\alpha w_j}) & \text{if } w_j < 0 
\end{cases} \]

\[
\min\{G(y, z, w) - \langle v^k, w \rangle : (y, z, w, \gamma) \in K\}
\]

\[
= \min \left\{ \left(1 - \lambda\right)\left(\frac{e^T y}{m} + \frac{e^T z}{k}\right) + \lambda \sum_{j=1}^{n} \max\{\alpha w_j, -\alpha w_j\} - \langle v^k, w \rangle \right\}
\]

s.t. \quad (y, z, w) \in K

\[
\Leftrightarrow \min \left\{ \left(1 - \lambda\right)\left(\frac{e^T y}{m} + \frac{e^T z}{k}\right) + \lambda \sum_{j=1}^{n} t_j 
- \langle v^k, w \rangle : (y, z, w, \gamma, t) \in \Omega \right\}
\]

(16)
Feasible Domain

\[
\Omega := \left\{ \left( y, z, w, \gamma, t \right) \in \mathbb{R}^{m+k+n+1+n} : \\
\left( y, z, w, \gamma \right) \in K, \\
-\alpha w_j \leq t_j, \alpha w_j \leq t_j, j = 1..n \right\}
\]
An Important Theorem

Theorem 1 (Convergence properties of Algorithm DCA)

(i) DCA generates a sequence \( \{(y^k, z^k, w^k, \gamma^k)\} \) such that the sequence \( \{F(y^k, z^k, w^k)\} \) is monotonously decreasing.

(ii) The sequence \( \{(y^k, z^k, w^k, \gamma^k)\} \) converges to \( (y^*, z^*, w^*, \gamma^*) \) after a finite number of iterations.

(iii) The point \( (y^*, z^*, w^*) \) is a critical point of the objective function \( F \) in Problem (13).
## Experimental Result

<table>
<thead>
<tr>
<th>Data set</th>
<th>FSV</th>
<th></th>
<th>DCA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>selected</td>
<td>correctness (%)</td>
<td>selected</td>
<td>correctness (%)</td>
</tr>
<tr>
<td></td>
<td>feature (%)</td>
<td>train</td>
<td>test</td>
<td>feature (%)</td>
</tr>
<tr>
<td>Pima Indian</td>
<td>66</td>
<td>75.22</td>
<td>74.60</td>
<td>50</td>
</tr>
<tr>
<td>BUPA Liver</td>
<td>75</td>
<td>68.18</td>
<td>65.20</td>
<td>50</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>31</td>
<td>90.47</td>
<td>84.07</td>
<td>9</td>
</tr>
<tr>
<td>WPBC (24 mo)</td>
<td>12</td>
<td>73.97</td>
<td>66.42</td>
<td>9</td>
</tr>
<tr>
<td>WPBC (60 mo)</td>
<td>8</td>
<td>70.70</td>
<td>67.05</td>
<td>6</td>
</tr>
<tr>
<td>Average</td>
<td>38.4</td>
<td>75.70</td>
<td>71.47</td>
<td>24.8</td>
</tr>
</tbody>
</table>
Selection of the $\lambda : CV$

Step 1. Set aside 10\% of the training data as a "tuning" set.

Step 2. Obtain a classifier for the given value of $\lambda$.

Step 3. Determine correctness on the "tuning" set.

Step 4. Repeat steps 1-3 10 times, each time setting aside a different 10\% portion of the training data. The "score" for this value of $\lambda$ is the average of the 10 correctness values determined in Step 3.
6. Our next work: Applying DCP

- (Structured) AUC-FSC-DC (based SVM)
- Asymmetric (SVM) FSC-DC
- FSC based on DR and LR
- KFCM-DC and Combination
- NMF-DC and Combination
- SPP-DC (mainly in Sparse solving)
- ...
Reference


• Thanks a lot!

• Q & A