On the convergence rate of Difference-of-Convex Algorithm (DCA)

Hadi Abbaszadehpeivasti (joint work with Etienne de Klerk and Moslem Zamani)

December 9, 2021

Tilburg University

Contents

DC Optimization

Preliminaries

Let $L \in (0, \infty]$ and $\mu \in [0, \infty)$ and let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a closed proper convex function.

- The function f is called *L-smooth* if for any $x_1, x_2 \in \mathbb{R}^n$,

$$\|g_1 - g_2\| \leq L\|x_1 - x_2\| \ \ \forall g_1 \in \partial f(x_1), \ g_2 \in \partial f(x_2).$$

- The function f is called *μ-strongly* convex function if the function $x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex.

Preliminaries

Let $L \in (0, \infty]$ and $\mu \in [0, \infty)$ and let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a closed proper convex function.

- The function f is called *L-smooth* if for any $x_1, x_2 \in \mathbb{R}^n$,

$$\|g_1 - g_2\| \leq L\|x_1 - x_2\| \quad \forall g_1 \in \partial f(x_1),\ g_2 \in \partial f(x_2).$$

- The function f is called *μ-strongly* convex function if the function $x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex.

We denote the set of L-smooth and μ-strongly convex function by $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$.

DC Optimization

$$\min f(x) \quad \text{s.t. } x \in K \tag{DCO}$$

where $f : \mathbb{R}^n \to [-\infty, \infty]$ is a *difference of convex (DC)* function,

$$f = f_1 - f_2,$$

and f_1, f_2 are convex functions.

DC Optimization

$$\min f(x) \qquad \text{(DCO)}$$
$$\text{s.t. } x \in K$$

where $f : \mathbb{R}^n \to [-\infty, \infty]$ is a *difference of convex (DC)* function,

$$f = f_1 - f_2,$$

and f_1, f_2 are convex functions.

- $K \subseteq \mathbb{R}^n$ is a closed convex set.
- The function f is closed.
- The functions $f_1 \in \mathcal{F}_{\mu_1, L_1}(\mathbb{R}^n)$, $f_2 \in \mathcal{F}_{\mu_2, L_2}(\mathbb{R}^n)$ for some $\mu_1, \mu_2 \in [0, \infty)$ and $L_1, L_2 \in (0, \infty]$.
- $f^\star > -\infty$ is a lower bound of (DCO).

Some classes of DC functions

The class of DC functions is quite large:

- Continuous *piece-wise linear* functions.
- *Twice continuously differentiable* functions on any convex subset of $\mathbb{R}^n$.

Some classes of DC functions

The class of DC functions is quite large:

- Continuous *piece-wise linear* functions.
- *Twice continuously differentiable* functions on any convex subset of $\mathbb{R}^n$.

Moreover,

- Every continuous function on a convex compact set can be *approximated by a DC function* with a given accuracy.

Difference of Convex Algorithm (DCA)

Algorithm 1 Unconstrained DCA

Pick $x^1 \in \mathbb{R}^n$, $N \in \mathbb{N}$, and $\epsilon > 0$.

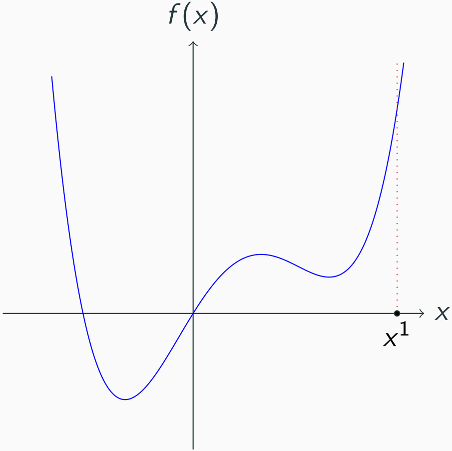
For $k = 1, 2, \ldots, N$ perform the following steps:

1. Choose $g_1^k \in \partial f_1(x^k)$ and $g_2^k \in \partial f_2(x^k)$. If $\|g_1^k - g_2^k\| \leq \epsilon$, then stop.
2. Choose

$$x^{k+1} \in \mathrm{argmin}_{x \in \mathbb{R}^n} f_1(x) - \left(f_2(x^k) + \langle g_2^k, x - x^k \rangle \right).$$

One iteration of DCA

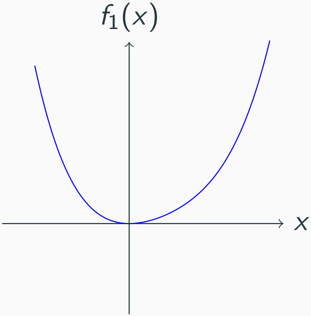
$$\min_{x \in \mathbb{R}} f(x) := \frac{1}{4}x^4 - \frac{2}{3}x^3 - \frac{1}{2}x^2 + 2x, \quad x^1 = 3.$$

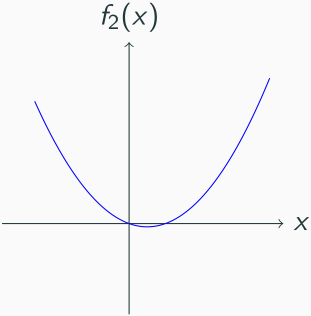


One iteration of DCA

$$\min_{x\in\mathbb{R}} f(x) := \frac{1}{4}x^4 - \frac{2}{3}x^3 - \frac{1}{2}x^2 + 2x, \quad x^1 = 3.$$

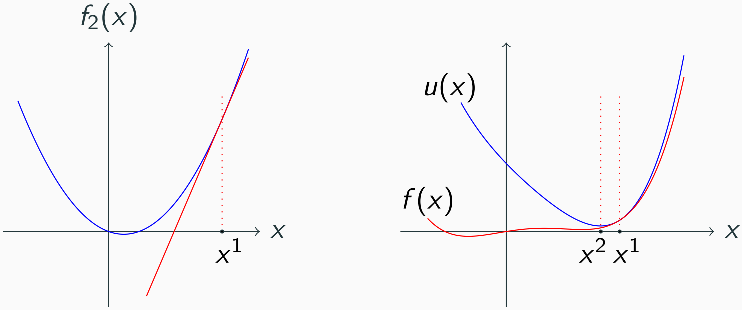
$$f(x) = \underbrace{\left(\frac{1}{4}x^4 - \frac{2}{3}x^3 + 2x^2\right)}_{f_1} - \underbrace{\left(\frac{1}{2}x^2 - 2x + 2x^2\right)}_{f_2}$$





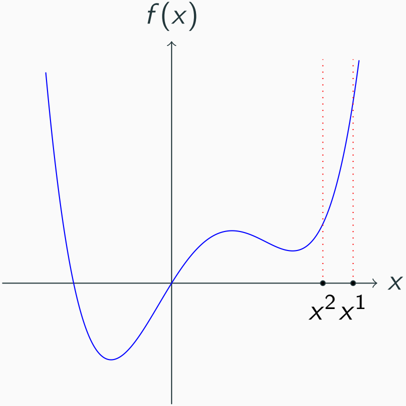
One iteration of DCA

$$\min_{x\in\mathbb{R}} u(x) := \underbrace{\left(\tfrac{1}{4}x^4 - \tfrac{2}{3}x^3 + 2x^2\right)}_{f_1} - \underbrace{(16.5 + 13(x-3))}_{f_2(x^1)+f_2'(x^1)(x-x^1)}$$



One iteration of DCA

$$\min_{x\in\mathbb{R}} f(x) := \frac{1}{4}x^4 - \frac{2}{3}x^3 - \frac{1}{2}x^2 + 2x$$



- DCA generates $x^2 = 2.5$.

Convergence rate result in the literature

Theorem (Thi, Dinh)

Assume that the following conditions hold:

i) *Either f_1 or f_2 is differentiable with locally Lipschitz derivative on all stationary points* (DCO).

ii) $\mu_1 + \mu_2 > 0$.

iii) $\{x^k\}$ *is bounded.*

iv) *The Łojasiewicz gradient inequality for all stationary points.*

Then we have the linear convergence rate for a suitable Łojasiewicz exponent.

H. A. Le Thi, T. P. Dinh. Convergence analysis of difference-of-convex algorithm with subanalytic data. *Journal of Optimization Theory and Applications* 179, 103–126 (2018)

Performance Estimation (PEP)

Performance Estimation Problem

$$
\begin{aligned}
\max \ & \left(\min_{1 \le k \le N+1} \left\| g_1^k - g_2^k \right\|^2 \right) \\
& f_1 \in \mathcal{F}_{\mu_1, L_1}(\mathbb{R}^n), f_2 \in \mathcal{F}_{\mu_2, L_2}(\mathbb{R}^n) \\
& f_1(x) - f_2(x) \ge f^\star \quad \forall x \in \mathbb{R}^n \\
& f_1(x^1) - f_2(x^1) - f^\star \le \Delta \\
& g_1^{N+1}, g_2^{N+1}, x^{N+1}, \ldots, x^2 \text{ are generated by DCA w.r.t. } f_1, f_2, x^1 \\
& x^1 \in \mathbb{R}^n,
\end{aligned}
$$

Performance Estimation Problem

$$
\begin{aligned}
\max \ & \left(\min_{1 \le k \le N+1} \left\| g_1^k - g_2^k \right\|^2 \right) \\
& f_1 \in \mathcal{F}_{\mu_1, L_1}(\mathbb{R}^n), f_2 \in \mathcal{F}_{\mu_2, L_2}(\mathbb{R}^n) \\
& f_1(x) - f_2(x) \ge f^\star \quad \forall x \in \mathbb{R}^n \\
& f_1(x^1) - f_2(x^1) - f^\star \le \Delta \\
& g_1^{N+1}, g_2^{N+1}, x^{N+1}, \ldots, x^2 \text{ are generated by DCA w.r.t. } f_1, f_2, x^1 \\
& x^1 \in \mathbb{R}^n,
\end{aligned}
$$

- *Decision variables:* f_1, f_2 and x^k, g_1^k, g_2^k $(k \in \{1, ..., N+1\})$.

Performance Estimation Problem

$$
\begin{aligned}
\max \ & \left(\min_{1 \leq k \leq N+1} \left\| g_1^k - g_2^k \right\|^2 \right) \\
& f_1 \in \mathcal{F}_{\mu_1, L_1}(\mathbb{R}^n), f_2 \in \mathcal{F}_{\mu_2, L_2}(\mathbb{R}^n) \\
& f_1(x) - f_2(x) \geq f^\star \quad \forall x \in \mathbb{R}^n \\
& f_1(x^1) - f_2(x^1) - f^\star \leq \Delta \\
& g_1^{N+1}, g_2^{N+1}, x^{N+1}, \ldots, x^2 \text{ are generated by DCA w.r.t. } f_1, f_2, x^1 \\
& x^1 \in \mathbb{R}^n,
\end{aligned}
$$

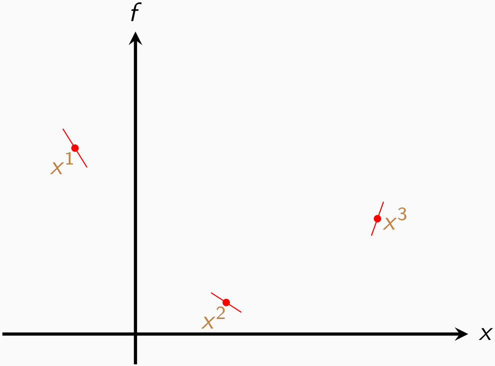
- *Decision variables:* f_1, f_2 and x^k, g_1^k, g_2^k $(k \in \{1, ..., N+1\})$.
- *Fixed parameters:* $\Delta, \mu_1, L_1, \mu_2, L_2, N$

L-smooth and μ-strongly Convex Interpolation Problem

Consider a finite index set I, and given triple $\left\{(\mathbf{x}^k, \mathbf{g}^k, f^k)\right\}_{k\in I}$ where $\mathbf{x}^k \in \mathbb{R}^n$, $\mathbf{g}^k \in \mathbb{R}^n$ and $f^k \in \mathbb{R}$.

L-smooth and μ-strongly Convex Interpolation Problem

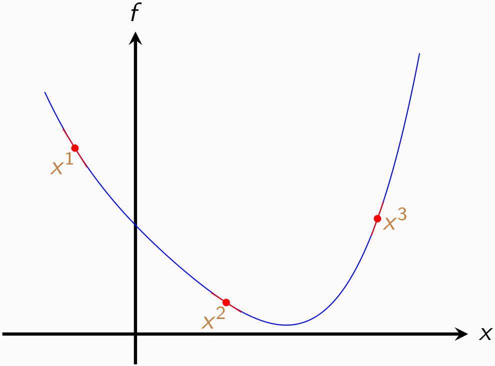
Consider a finite index set I, and given triple $\left\{(\mathbf{x}^k, \mathbf{g}^k, f^k)\right\}_{k\in I}$ where $\mathbf{x}^k \in \mathbb{R}^n$, $\mathbf{g}^k \in \mathbb{R}^n$ and $f^k \in \mathbb{R}$.



$$?\exists\ f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n):\ f(\mathbf{x}^k) = f^k, \quad \text{and} \quad \mathbf{g}^k \in \partial f(\mathbf{x}^k), \qquad \forall k \in I.$$

L-smooth and μ-strongly Convex Interpolation Problem

Consider a finite index set I, and given triple $\left\{(\mathbf{x}^k, \mathbf{g}^k, f^k)\right\}_{k \in I}$ where $\mathbf{x}^k \in \mathbb{R}^n$, $\mathbf{g}^k \in \mathbb{R}^n$ and $f^k \in \mathbb{R}$.



$$?\exists \;\; f \in \mathcal{F}_{\mu,L}(\mathbb{R}^n):\; f(\mathbf{x}^k) = f^k, \quad \text{and} \quad \mathbf{g}^k \in \partial f(\mathbf{x}^k), \qquad \forall k \in I.$$

If yes, we say $\left\{(\mathbf{x}^k, \mathbf{g}^k, f^k)\right\}_{k \in I}$ is $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$-interpolable.

L-smooth and μ-strongly Interpolation

Theorem (Taylor, Hendrickx, and Glineur (2017))

The following statements are equivalent:

1. $\left\{(\mathbf{x}^i, \mathbf{g}^i, f^i)\right\}_{i\in I}$ *is* $\mathcal{F}_{\mu,L}(\mathbb{R}^n)$*-interpolable;*
2. $\forall i, j \in I$:

$$\frac{1}{2(1-\frac{\mu}{L})}\left(\frac{1}{L}\left\|g^i - g^j\right\|^2 + \mu\left\|x^i - x^j\right\|^2 - \frac{2\mu}{L}\left\langle g^j - g^i, x^j - x^i\right\rangle\right) \leq f^i - f^j - \left\langle g^j, x^i - x^j\right\rangle.$$

A.B. Taylor, J.M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming* 161.1-2, 307–345 (2017)

Reformulation of PEP

$$
\begin{aligned}
\max \ & \left(\min_{1 \le k \le N+1} \left\| g_1^k - g_2^k \right\|^2 \right) \\
\text{s.t. } & \frac{1}{2(1-\frac{\mu_1}{L_1})} \left(\frac{1}{L_1} \left\| g_1^i - g_1^j \right\|^2 + \mu_1 \left\| x^i - x^j \right\|^2 - \frac{2\mu_1}{L_1} \left\langle g_1^j - g_1^i, x^j - x^i \right\rangle \right) \\
& \qquad \le f_1^i - f_1^j - \left\langle g_1^j, x^i - x^j \right\rangle \quad i, j \in \{1, \dots, N+1\} \\
& \frac{1}{2(1-\frac{\mu_2}{L_2})} \left(\frac{1}{L_2} \left\| g_2^i - g_2^j \right\|^2 + \mu_2 \left\| x^i - x^j \right\|^2 - \frac{2\mu_2}{L_2} \left\langle g_2^j - g_2^i, x^j - x^i \right\rangle \right) \\
& \qquad \le f_2^i - f_2^j - \left\langle g_2^j, x^i - x^j \right\rangle \quad i, j \in \{1, \dots, N+1\} \\
& g_1^{k+1} = g_2^k \quad k \in \{1, \dots, N\} \\
& f_1^k - f_2^k \ge f^\star \quad k \in \{1, \dots, N+1\} \\
& f_1^1 - f_2^1 - f^\star \le \Delta.
\end{aligned}
$$

Covergence rate of DCA

Performance estimation technique

- The last problem may be rewritten as a *semidefinite programming problem (SDP)* by replacing all inner products by the entries of an unknown Gram matrix.

Performance estimation technique

- The last problem may be rewritten as a *semidefinite programming problem (SDP)* by replacing all inner products by the entries of an unknown Gram matrix.
- We employ *weak duality* to bound the optimal value of the last problem by constructing a dual feasible solution of SDP.

- The last problem may be rewritten as a *semidefinite programming problem (SDP)* by replacing all inner products by the entries of an unknown Gram matrix.
- We employ *weak duality* to bound the optimal value of the last problem by constructing a dual feasible solution of SDP.
- The dual feasible solution is constructed empirically by doing *numerical experiments* with fixed values of the parameters $\Delta, N, \mu_1, L_1, \mu_2, L_2$.

Performance estimation technique

- The last problem may be rewritten as a *semidefinite programming problem (SDP)* by replacing all inner products by the entries of an unknown Gram matrix.
- We employ *weak duality* to bound the optimal value of the last problem by constructing a dual feasible solution of SDP.
- The dual feasible solution is constructed empirically by doing *numerical experiments* with fixed values of the parameters $\Delta, N, \mu_1, L_1, \mu_2, L_2$.
- The *analytical expressions* of the dual multipliers and optimal value *are guessed* and the guess is verified analytically.

Convergence rate of of unconstrained DCA

Theorem

Let $f_1 \in \mathcal{F}_{\mu_1,L_1}(\mathbb{R}^n)$ and $f_2 \in \mathcal{F}_{\mu_2,L_2}(\mathbb{R}^n)$. If L_1 or L_2 is finite, then after N iterations of DCA, one has:

i) *If $L_1 = \infty$, $L_2 < \infty$, then*

$$\min_{1\le k\le N+1} \left\| g_1^k - g_2^k \right\| \le \sqrt{\left(\frac{2L_2^2}{L_2+\mu_1}\right) \frac{f(x^1)-f^\star}{N}}.$$

Convergence rate of of unconstrained DCA

Theorem

Let $f_1 \in \mathcal{F}_{\mu_1,L_1}(\mathbb{R}^n)$ *and* $f_2 \in \mathcal{F}_{\mu_2,L_2}(\mathbb{R}^n)$*. If* L_1 *or* L_2 *is finite, then after* N *iterations of DCA, one has:*

i) *If* $L_1 = \infty$, $L_2 < \infty$, *then*

$$\min_{1\leq k\leq N+1} \left\| g_1^k - g_2^k \right\| \leq \sqrt{\left(\frac{2L_2^2}{L_2+\mu_1}\right)\frac{f(x^1)-f^\star}{N}}.$$

ii) *If* $L_2 = \infty$, $L_1 < \infty$, *then*

$$\min_{1\leq k\leq N+1} \left\| g_1^k - g_2^k \right\| \leq \sqrt{\left(\frac{2L_1^2}{L_1+\mu_2}\right)\frac{f(x^1)-f^\star}{N}}.$$

Convergence rate

Theorem

Suppose that $f_1 \in \mathcal{F}_{\mu_1,L_1}(\mathbb{R}^n)$ *and* $f_2 \in \mathcal{F}_{\mu_2,L_2}(\mathbb{R}^n)$*. If* L_1 *or* L_2 *is finite, then after* N *iterations of DCA, one has:*

iii) *If* $L_1 = L_2 = L$*, then*

$$\min_{1\le k\le N+1} \left\| g_1^k - g_2^k \right\| \le \sqrt{L\left(\frac{f(x^1)-f^\star}{N}\right)}.$$

Convergence rate

Theorem

Suppose that $f_1 \in \mathcal{F}_{\mu_1,L_1}(\mathbb{R}^n)$ *and* $f_2 \in \mathcal{F}_{\mu_2,L_2}(\mathbb{R}^n)$. *If* L_1 *or* L_2 *is finite, then after* N *iterations of DCA, one has:*

iii) *If* $L_1 = L_2 = L$, *then*

$$\min_{1\le k\le N+1} \left\| g_1^k - g_2^k \right\| \le \sqrt{L\left(\frac{f(x^1) - f^\star}{N}\right)}.$$

iv) *If* $L_1, L_2 < \infty$, *and* $\mu_1 = \mu_2 = 0$ *then*

$$\min_{1\le k\le N+1} \left\| g_1^k - g_2^k \right\| \le \sqrt{\left(\frac{2L_1L_2}{L_1+L_2}\right)\left(\frac{f(x^1) - f^\star}{N}\right)}.$$

Lower bound

- As the theorem shows, the worst case convergence rate of DCA is of $O(\frac{1}{\sqrt{N}})$.

Lower bound

- As the theorem shows, the worst case convergence rate of DCA is of $O(\frac{1}{\sqrt{N}})$.
- *There exists a DC function f and initial point x^1 that DCA performs at least N iterations for obtaining the accuracy of $\frac{1}{\sqrt{N}}$.*

Constrained DC Optimization

$$\begin{aligned} \min\ & f(x) = f_1(x) - f_2(x) \\ \text{s.t. } & x \in K. \end{aligned}$$

Constrained DC Optimization

$$\begin{aligned} &\min\ f(x) = f_1(x) - f_2(x) \\ &\text{s.t. } x \in K. \end{aligned}$$

$$T(x^{k+1}) := f_1(x^k) - f_1(x^{k+1}) - \left\langle g_2^k, x^k - x^{k+1} \right\rangle .$$

- $T(x^{k+1}) \geq 0$.
- $T(x^{k+1}) = 0$ implies that x^k is a *critical point* of (DCO).

Constrained DCA (CDCA)

Algorithm 2 CDCA

Pick $x^1 \in K$, $N \in \mathbb{N}$, and $\epsilon > 0$.

For $k = 1, 2, \ldots, N$ perform the following steps:

1. Choose $g_2^k \in \partial f_2(x^k)$ and
$$x^{k+1} \in \operatorname{argmin}_{x \in K} f_1(x) - f_2(x^k) - \left\langle g_2^k, x - x^k \right\rangle.$$
2. If $f_1(x^k) - f_1(x^{k+1}) - \left\langle g_2^k, x^k - x^{k+1} \right\rangle \leq \epsilon$, then stop.

Convergence rate of Constrained DCA

Using *performance estimation* as before, we can prove the following.

Theorem

Let $f_1 \in \mathcal{F}_{\mu_1,L_1}(\mathbb{R}^n)$ and $f_2 \in \mathcal{F}_{\mu_2,L_2}(\mathbb{R}^n)$ and let K be a closed convex set. Then, after $N \geq 2$ iterations of CDCA, one has:

$$\min_{1\leq k\leq N} f_1(x^k) - f_1(x^{k+1}) - \left\langle g_2^k, x^k - x^{k+1} \right\rangle \leq \frac{L_2}{(L_2+\mu_1)\,N - \mu_1} \left(f(x^1) - f^\star \right).$$

Conclusion

Future work

- Convergence of the DCA on more restricted classes of DC problems, e.g. f is a *polynomial function*, ((extended) *trust region problems* in constraint case).
- *Undominated DC decompositions* to obtain the sharpest possible results.
- Understanding the class of DC functions defined by $\mathcal{F}_{\mu_1,L_1} - \mathcal{F}_{\mu_2,L_2}$ since some of our results only hold for this class with at least one of L_1 or L_2 finite.

Future work

- Convergence of the DCA on more restricted classes of DC problems, e.g. f is a *polynomial function*, ((extended) *trust region problems* in constraint case).
- *Undominated DC decompositions* to obtain the sharpest possible results.
- Understanding the class of DC functions defined by $\mathcal{F}_{\mu_1,L_1} - \mathcal{F}_{\mu_2,L_2}$ since some of our results only hold for this class with at least one of L_1 or L_2 finite.

H. Abbaszadehpeivasti, E. de Klerk, and M. Zamani. On the rate of convergence of the Difference-of-Convex Algorithm (DCA). *arXiv preprint arXiv:2109.13566* (2021)

The End