First-Order Methods through Partial Linearization

Alp Yurtsever
Umeå University, Sweden

based on joint works with Suvrit Sra, Hoomaan Maskan, and others

May 2026
ELLIIT Focus Period
Lund, Sweden

DC Template

$$\min_{\boldsymbol{x}\in\mathbb{R}^m} \quad \phi(\boldsymbol{x}) := f(\boldsymbol{x}) - h(\boldsymbol{x}), \tag{DC}$$

- f and h are lower semi-continuous convex functions,
- we assume that the optimal value $\phi_\star$ is finite.

DC Algorithm

The key idea of DCA (Pham Dinh & Souad, 1986) is to linearize the concave component $-h(\boldsymbol{x})$ to obtain a global convex upper bound:

$$\hat{\phi}(\boldsymbol{x}; \boldsymbol{y}) := f(\boldsymbol{x}) - h(\boldsymbol{y}) - \langle \nabla h(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle.$$

DCA then updates its estimation by solving convex subproblems

$$\boldsymbol{x}_{k+1} \in \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^m} \hat{\phi}(\boldsymbol{x}; \boldsymbol{x}_k) \tag{1}$$

A solution to (1) always exists as $\phi_\star$ is finite by assumption and

$$\min_{\boldsymbol{x} \in \mathbb{R}^m} \hat{\phi}(\boldsymbol{x}; \boldsymbol{x}_k) \geq \min_{\boldsymbol{x} \in \mathbb{R}^m} \phi(\boldsymbol{x}) = \phi_\star.$$

DC Algorithm

1. DCA is a specific Majorize-Minorize algorithm.
2. If the functions are differentiable, DCA amounts to the implicit update rule:
$$\nabla f(\boldsymbol{x}^{k+1}) = \nabla h(\boldsymbol{x}^k).$$
3. DCA generates a monotonically decreasing sequence $\phi(\boldsymbol{x}^k)$.
4. Numerous algorithms can be viewed as special cases of DCA:
   - Sinkhorn's method for matrix scaling
   - EM algorithm for natural exponential family distributions
   - Proximal gradient method, Mirror descent
   - Proximal point method, Mirror prox

DC Critical Points

Definition

We measure convergence in terms of the following gap function:

$$\mathrm{gap}_{\mathrm{DC}}(\boldsymbol{y}) = \max_{\boldsymbol{x}\in\mathbb{R}^d} \min_{\boldsymbol{u}\in\partial h(\boldsymbol{y})} \left\{ f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \boldsymbol{u}, \boldsymbol{y} - \boldsymbol{x} \rangle \right\}.$$

Lemma

- $\mathrm{gap}_{\mathrm{DC}}(\boldsymbol{y}) \geq 0$ *for all* $\boldsymbol{y} \in \mathbb{R}^d$.
- $\mathrm{gap}_{\mathrm{DC}}(\boldsymbol{y}) = 0$ *if and only if* $\boldsymbol{y}$ *is a critical point satisfying*

$$\partial f(\boldsymbol{y}) \cap \partial h(\boldsymbol{y}) \neq \emptyset.$$

If h is differentiable, then any such $\boldsymbol{y}$ is a first-order stationary point.

Theoretical guarantees for DCA

Theorem

Let the sequence $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^K$ be generated by an inexact DCA, where each subproblem is solved approximately to satisfy

$$\hat{\phi}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) - \min_{x \in \mathbb{R}^d} \hat{\phi}(\boldsymbol{x}, \boldsymbol{x}^k) \leq \frac{\epsilon}{2}$$

for some target accuracy $\epsilon > 0$. Then,

$$\frac{1}{K} \sum_{k=1}^{K} \text{gap}_{\text{DC}}(\boldsymbol{x}^k) \;\leq\; \frac{\phi(\boldsymbol{x}^1) - \phi^\star}{K} + \frac{\epsilon}{2}.$$

Implications:

- Iteration complexity of DCA is $\mathcal{O}(1/\epsilon)$.
- Complexity is independent of d (assuming $\phi(\boldsymbol{x}^1) - \phi^\star$ is constant).

Proof: Theoretical guarantees for DCA

Denoting $\boldsymbol{u}^k \in \partial h(\boldsymbol{x}^k)$, by the convexity of h,

$$\phi(\boldsymbol{x}^{k+1}) = f(\boldsymbol{x}^{k+1}) - h(\boldsymbol{x}^{k+1}) \leq f(\boldsymbol{x}^{k+1}) - h(\boldsymbol{x}^k) - \langle \boldsymbol{u}^k, \boldsymbol{x}^{k+1} - \boldsymbol{x}^k \rangle.$$

Then, by definition of $\boldsymbol{x}^{k+1}$:

$$\phi(\boldsymbol{x}^{k+1}) \leq f(\boldsymbol{x}) - h(\boldsymbol{x}^k) - \langle \nabla h(\boldsymbol{x}^k), \boldsymbol{x} - \boldsymbol{x}^k \rangle + \epsilon/2, \quad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Add and subtract $f(\boldsymbol{x}^k)$ on the right side:

$$\begin{aligned} \phi(\boldsymbol{x}^{k+1}) &\leq f(\boldsymbol{x}^k) - f(\boldsymbol{x}^k) + f(\boldsymbol{x}) - h(\boldsymbol{x}^k) - \langle \nabla h(\boldsymbol{x}^k), \boldsymbol{x} - \boldsymbol{x}^k \rangle + \epsilon/2 \\ &\leq \phi(\boldsymbol{x}^k) - \mathrm{gap}_{\mathrm{DC}}(\boldsymbol{x}^k) + \epsilon/2. \end{aligned}$$

We complete the proof by averaging over $k = 1, \ldots, K$.

⚠ Convexity of f is not used!

PPM via DCA

$$\min_{\boldsymbol{x}\in\mathbb{R}^m} \quad F(\boldsymbol{x}).$$

Add and subtract a strongly convex functions:

$$\min_{\boldsymbol{x}\in\mathbb{R}^m} \quad \underbrace{F(\boldsymbol{x}) + \tfrac{1}{2\lambda}\|\boldsymbol{x}\|^2}_{f(\boldsymbol{x})} - \underbrace{\tfrac{1}{2\lambda}\|\boldsymbol{x}\|^2}_{h(\boldsymbol{x})}.$$

DCA update is

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^m} \left\{F(\boldsymbol{x}) + \tfrac{1}{2\lambda}\|\boldsymbol{x}\|^2 - \tfrac{1}{\lambda}\langle \boldsymbol{x}_k, \boldsymbol{x}-\boldsymbol{x}_k\rangle\right\} \\ &= \operatorname{prox}_{\lambda F}(\boldsymbol{x}_k)\end{aligned}$$

DCA convergence theorem implies

$$\frac{1}{K}\sum_{k=1}^{K} \max_{\boldsymbol{x}\in\mathbb{R}^m} \left\{F(\boldsymbol{x}_k) - F(\boldsymbol{x}) - \tfrac{1}{2\lambda}\|\boldsymbol{x}-\boldsymbol{x}_k\|^2\right\} \le \frac{\phi(\boldsymbol{x}_1)-\phi_\star}{K}$$

$$\implies \frac{1}{K}\sum_{k=1}^{K} \frac{1}{2(L+\lambda^{-1})}\|\boldsymbol{x}_k - \operatorname{prox}_{\lambda F}(\boldsymbol{x}_k)\|^2 \le \frac{\phi(\boldsymbol{x}_1)-\phi_\star}{K}$$

PGM via DCA

$$\min_{\boldsymbol{x}\in\mathbb{R}^m} \quad F(\boldsymbol{x}) + G(\boldsymbol{x})$$

Add and subtract a strongly convex functions:

$$\min_{\boldsymbol{x}\in\mathbb{R}^m} \quad \underbrace{G(\boldsymbol{x}) + \tfrac{L}{2}\|\boldsymbol{x}\|^2}_{f(\boldsymbol{x})} - \underbrace{\left(\tfrac{L}{2}\|\boldsymbol{x}\|^2 - F(\boldsymbol{x})\right)}_{h(\boldsymbol{x})}$$

DCA update is

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^m} \left\{G(\boldsymbol{x}) + \tfrac{L}{2}\|\boldsymbol{x}\|^2 - \langle L\boldsymbol{x}_k - \nabla F(\boldsymbol{x}_k),\, \boldsymbol{x} - \boldsymbol{x}_k\rangle\right\} \\ &= \operatorname{prox}_{\frac{1}{L}G}(\boldsymbol{x}_k - \tfrac{1}{L}\nabla F(\boldsymbol{x}_k))\end{aligned}$$

DCA convergence theorem implies

$$\frac{1}{K}\sum_{k=1}^{K} \max_{\boldsymbol{x}\in\mathbb{R}^m} \left\{\langle \nabla F(\boldsymbol{x}_k),\, \boldsymbol{x}_k - \boldsymbol{x}\rangle + G(\boldsymbol{x}_k) - G(\boldsymbol{x}) - \tfrac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|^2\right\} \leq \frac{\phi(\boldsymbol{x}_1) - \phi_\star}{K}$$

$$\implies \frac{1}{K}\sum_{k=1}^{K} \frac{L}{2}\|\boldsymbol{x}_k - \operatorname{prox}_{\frac{1}{L}G}(\boldsymbol{x}_k - \tfrac{1}{L}\nabla F(\boldsymbol{x}_k))\|^2 \leq \frac{\phi(\boldsymbol{x}_1) - \phi_\star}{K}$$

EM via DCA

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^m} \quad \mathcal{L}(\boldsymbol{\theta}) := -\sum_{x\in\mathcal{X}} \log P(x|\boldsymbol{\theta})$$

P is defined through hidden variables: $P(x|\boldsymbol{\theta}) = \sum_y P(x,y|\boldsymbol{\theta})$.

Consider the natural exponential family of distributions, defined by

$$P(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta}) = \frac{\varphi(\boldsymbol{x},\boldsymbol{y})e^{\langle \boldsymbol{\theta},\, T(\boldsymbol{x},\boldsymbol{y})\rangle}}{\sum_{\boldsymbol{x},\boldsymbol{y}} \varphi(\boldsymbol{x},\boldsymbol{y})e^{\langle \boldsymbol{\theta},\, T(\boldsymbol{x},\boldsymbol{y})\rangle}},$$

EM can be viewed as DCA applied to

$$f(\boldsymbol{\theta}) = |\mathcal{X}| \log\Big(\sum_{\boldsymbol{x},\boldsymbol{y}} \varphi(\boldsymbol{x},\boldsymbol{y})e^{\langle \boldsymbol{\theta},\, T(\boldsymbol{x},\boldsymbol{y})\rangle}\Big)$$

$$h(\boldsymbol{\theta}) = \sum_{x\in\mathcal{X}} \log\Big(\sum_{\boldsymbol{y}} \varphi(\boldsymbol{x},\boldsymbol{y})e^{\langle \boldsymbol{\theta},\, T(\boldsymbol{x},\boldsymbol{y})\rangle}\Big)$$

These terms are smooth, owing to log-sum-exp structure.

EM via DCA

DCA convergence theorem implies

$$\frac{1}{K}\sum_{k=1}^{K}\max_{\boldsymbol{\theta}\in\mathbb{R}^m}\left\{f(\boldsymbol{\theta}_k)-f(\boldsymbol{\theta})-\langle\nabla h(\boldsymbol{\theta}_k),\,\boldsymbol{\theta}_k-\boldsymbol{\theta}\rangle\right\}\leq\frac{\mathcal{L}(\boldsymbol{\theta}_1)-\mathcal{L}_\star}{K}$$

$$\implies\frac{1}{K}\sum_{k=1}^{K}\max_{\boldsymbol{\theta}\in\mathbb{R}^m}\left\{\langle\underbrace{\nabla f(\boldsymbol{\theta}_k)-\nabla h(\boldsymbol{\theta}_k)}_{\nabla\mathcal{L}(\boldsymbol{\theta}_k)},\,\boldsymbol{\theta}_k-\boldsymbol{\theta}\rangle-\tfrac{L}{2}\|\boldsymbol{\theta}_k-\boldsymbol{\theta}\|^2\right\}\leq\frac{\mathcal{L}(\boldsymbol{\theta}_1)-\mathcal{L}_\star}{K}$$

$$\iff\frac{1}{K}\sum_{k=1}^{K}\frac{1}{2L}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2\leq\frac{\mathcal{L}(\boldsymbol{\theta}_1)-\mathcal{L}_\star}{K}$$

⚠ This rate is consistent with the existing guarantees, as provided by Kumar and Schmidth (2017) and Kunstner et al. (2020).

DCA is FW in disguise

DCA is FW

Proposition

DCA is equivalent to the FW method applied to the following epigraph reformulation of (DC)*:*

$$\min_{x,\,t} \quad t - h(x), \qquad f(x) \le t.$$

$$\underbrace{\min_{x,\,t}}_{\omega} \quad \underbrace{t - h(x)}_{\psi(\omega)}, \qquad \underbrace{f(x) \le t}_{\omega \in \mathcal{M}}.$$

The linear minimization step of FW amounts to

$$(x_k^*, t_k^*) \in \operatorname*{argmin}_{x,\,t} \; t - \langle \nabla h(x_k), x \rangle \quad \text{s.t.} \quad f(x) \le t. \tag{2}$$

Eliminate t in (2) and get

$$x_k^* \in \operatorname*{argmin}_{x} \; f(x) - \langle \nabla h(x_k), x \rangle.$$

FW update becomes

$$x_{k+1} = (1 - \eta_k)x_k + \eta_k x_k^* = x_k^*$$

Curvature constant

The convergence of FW crucially depends on the *curvature constant* of ψ over $\mathcal{M}$, defined as follows:

$$C_\psi := \sup_{\substack{\omega,\hat{\omega}\in\mathcal{M};\ \eta\in[0,1] \\ \bar{\omega}=(1-\eta)\omega+\eta\hat{\omega}}} \frac{2}{\eta^2}\Big(\psi(\bar{\omega}) - \psi(\omega) - \langle\nabla\psi(\omega),\, \bar{\omega}-\omega\rangle\Big).$$

Equivalently, for all $\omega, \hat{\omega} \in \mathcal{M};\ \ \eta \in [0,1],\ \ \text{and } \bar{\omega} = (1-\eta)\omega + \eta\hat{\omega};$

$$\psi(\bar{\omega}) \leq \psi(\omega) + \langle\nabla\psi(\omega),\, \bar{\omega}-\omega\rangle + \frac{1}{2}\eta^2 C_\psi$$

Special cases:

- If ψ is L-smooth and $\mathcal{M}$ has a bounded diameter D, then $C_\psi \leq LD^2$.
- If ψ is concave, then $C_\psi = 0$.

Guarantees of FW

Theorem (Jaggi, 2013, Theorem 1)

*Suppose ψ is **convex** and C_ψ is bounded. Then, the sequence $\boldsymbol{\omega}_k$ generated by FW converges to a solution in objective value:*

$$\psi(\boldsymbol{\omega}_k) - \psi_\star \leq \frac{2C_\psi}{k+1}.$$

Theorem (Lacoste-Julien, 2016, Theorem 1)

*Suppose C_ψ is bounded but ψ is **non-convex**. Then,*

$$\frac{1}{k}\sum_{\tau=1}^{k} \max_{\boldsymbol{\omega}\in\mathcal{M}} \langle \nabla\psi(\boldsymbol{\omega}_\tau), \boldsymbol{\omega}_\tau - \boldsymbol{\omega}\rangle \leq \frac{\max\{C_\psi, 2(\psi(\boldsymbol{\omega}_1) - \psi_\star)\}}{\sqrt{k}}$$

Theorem (Journée et al., 2010, Theorem 1)

*Suppose ψ is **concave**. Then,*

$$\frac{1}{k}\sum_{\tau=1}^{k} \max_{\boldsymbol{\omega}\in\mathcal{M}} \langle \nabla\psi(\boldsymbol{\omega}_\tau), \boldsymbol{\omega}_\tau - \boldsymbol{\omega}\rangle \leq \frac{\psi(\boldsymbol{\omega}_1) - \psi_\star}{k}.$$

Beyond DC decompositions

Classical DCA focuses on a convex – convex decomposition.

FW viewpoint suggests a broader range of decompositions.

$$\min_{\boldsymbol{x},\, t} \quad t - h(\boldsymbol{x}), \qquad f(\boldsymbol{x}) \leq t.$$

If h is non-convex:

- full step $\eta = 1$ is no longer valid in general,
- the rate reduces to $\mathcal{O}(1/\sqrt{k})$.

The extended algorithm is

$$\boldsymbol{s}_k \in \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^m} \hat{\phi}(\boldsymbol{x}; \boldsymbol{x}_k)$$
$$\boldsymbol{x}_{k+1} = (1 - \eta_k)\boldsymbol{x}_k + \eta_k \boldsymbol{s}_k$$

Subgradient Method via DCA

$$\min_{\boldsymbol{x}\in\mathbb{R}^m} \quad F(\boldsymbol{x})$$

Add and subtract a strongly convex function:

$$\min_{\boldsymbol{x}\in\mathbb{R}^m} \quad \underbrace{\tfrac{1}{2\alpha}\|\boldsymbol{x}\|^2}_{f(\boldsymbol{x})} - \underbrace{\left(\tfrac{1}{2\alpha}\|\boldsymbol{x}\|^2 - F(\boldsymbol{x})\right)}_{h(\boldsymbol{x})}$$

h can be nonconvex, since F is convex but nonsmooth.

Let $\boldsymbol{u}_k \in \partial F(x_k)$. DCA update is

$$\boldsymbol{s}_k = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^m} \left\{ \tfrac{1}{2\alpha}\|\boldsymbol{x}\|^2 - \left\langle \tfrac{1}{\alpha}\boldsymbol{x}_k - \boldsymbol{u}_k,\, \boldsymbol{x} - \boldsymbol{x}_k \right\rangle \right\} = \boldsymbol{x}_k - \alpha\boldsymbol{u}_k$$

$$\boldsymbol{x}_{k+1} = (1-\eta_k)\boldsymbol{x}_k + \eta_k \boldsymbol{s}_k = \boldsymbol{x}_k - \alpha\eta_k \boldsymbol{u}_k$$

A common step-size for nonconvex FW is $\eta_k = 1/\sqrt{k}$.

Dynamic Decompositions

Classical DCA focuses on a fixed decomposition

$$\phi(\boldsymbol{x}) := f(\boldsymbol{x}) - h(\boldsymbol{x}).$$

The analysis allows for a dynamic decomposition

$$\phi(\boldsymbol{x}) := f_k(\boldsymbol{x}) - h_k(\boldsymbol{x}).$$

⚠ $\mathrm{gap}_{\mathrm{DC}}(\boldsymbol{x}_k) := \mathrm{gap}_{\mathrm{DC}}^{f_k,h_k}(\boldsymbol{x}_k)$ depends on the decompositions.

We get the guarantees in terms of $\dfrac{1}{K}\displaystyle\sum_{k=1}^{K} \mathrm{gap}_{\mathrm{DC}}^{f_k,h_k}(\boldsymbol{x}_k)$.

Non-Euclidean GM via DCA

$$\min_{\boldsymbol{x}\in\mathbb{R}^m} \quad F(\boldsymbol{x})$$

Add and subtract squared norms centered at $\boldsymbol{x}_k$:

$$\min_{\boldsymbol{x}\in\mathbb{R}^m} \quad \underbrace{\tfrac{1}{2\alpha}\|\boldsymbol{x}-\boldsymbol{x}_k\|^2}_{f_k(\boldsymbol{x})} - \underbrace{\left(\tfrac{1}{2\alpha}\|\boldsymbol{x}-\boldsymbol{x}_k\|^2 - F(\boldsymbol{x})\right)}_{h_k(\boldsymbol{x})}$$

DCA update is

$$\boldsymbol{s}_k = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^m} \left\{\tfrac{1}{2\alpha}\|\boldsymbol{x}-\boldsymbol{x}_k\|^2 - \tfrac{1}{\alpha}\langle\nabla F(\boldsymbol{x}_k),\, \boldsymbol{x}-\boldsymbol{x}_k\rangle\right\}$$

$$\boldsymbol{x}_{k+1} = (1-\eta_k)\boldsymbol{x}_k + \eta_k \boldsymbol{s}_k.$$

Revisiting Frank-Wolfe for Structured Nonconvex Problems

FW for DC functions

$$\min_{\boldsymbol{x}\in\mathcal{D}} \quad \phi(\boldsymbol{x}) := f(\boldsymbol{x}) - h(\boldsymbol{x}),$$

- $\mathcal{D} \subset \mathbb{R}^m$ is a convex and compact set with diameter D,
- f and h are lower semi-continuous convex functions,
- f is L_f-smooth.

DCFW Algorithm

Idea: Apply DCA, and solve the convex subproblems with FW.

$$\boldsymbol{x}_{t+1} \in \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x}) - \langle \boldsymbol{u}_t, \boldsymbol{x}\rangle, \qquad \boldsymbol{u}_t \in \partial h(\boldsymbol{x}_t).$$

Input: initial point $\boldsymbol{x}_1 \in \mathcal{D}$, target accuracy $\epsilon > 0$
for $t = 1, 2, \ldots$ **do**
  Choose $\boldsymbol{u}_t \in \partial h(\boldsymbol{x}_t)$
  Initialize $\boldsymbol{X}_{t,1} = \boldsymbol{x}_t$
  for $k = 1, 2, \ldots$ **do**
    $\boldsymbol{S}_{t,k} = \arg\min_{\boldsymbol{x}\in\mathcal{D}} \langle \nabla f(\boldsymbol{X}_{t,k}) - \boldsymbol{u}_t,\ \boldsymbol{x}\rangle$
    $\boldsymbol{D}_{t,k} = \boldsymbol{S}_{t,k} - \boldsymbol{X}_{t,k}$
    if $-\langle \nabla f(\boldsymbol{X}_{t,k}) - \boldsymbol{u}_t,\ \boldsymbol{D}_{t,k}\rangle \le \epsilon/2$ **then**
      Set $\boldsymbol{x}_{t+1} = \boldsymbol{X}_{t,k}$ and **break**
    end if
    $\boldsymbol{X}_{t,k+1} = \boldsymbol{X}_{t,k} + \eta_{t,k} \cdot \boldsymbol{D}_{t,k}$ // $\eta_{t,k} = 2/(k+1)$ or line-search
  end for
end for

DCFW guarantees

Theorem

DCFW generates a sequence of solutions that satisfy

$$\min_{0 \leq \tau \leq t} \text{gap}_{\text{DC}}(\boldsymbol{x}_\tau) \leq \epsilon$$

within $\mathcal{O}(1/\epsilon)$ outer iterations, with a total of $\mathcal{O}(1/\epsilon^2)$ inner iterations.

**Intuition:* DCA converges at $\mathcal{O}(1/t)$ rate. Each DCA step solves a smooth convex problem via FW with $\mathcal{O}(1/k)$ rate.

Oracle Complexity to reach ϵ-accuracy:

- $\mathcal{O}(1/\epsilon^2)$ linear minimization oracles,
- $\mathcal{O}(1/\epsilon^2)$ gradient calls for f,
- $\mathcal{O}(1/\epsilon)$ (sub)gradient calls for h.

Conditional gradient sliding

DC decomposition is not unique — choosing a suitable decomposition can reduce oracle complexity.

Example: Since f is L_f-smooth, $\frac{L_f}{2}\|x\|^2 - f(x)$ is convex.
Consider DCFW on the following reformulation of the problem:

$$\min_{x \in \mathcal{D}} \underbrace{\frac{L_f}{2}\|x\|^2}_{f'(x)} - \underbrace{\left(h(x) + \frac{L_f}{2}\|x\|^2 - f(x)\right)}_{h'(x)}$$

Benefit: Gradients of f are now accessed less frequently.

Oracle Complexity to reach ϵ-accuracy:

- $\mathcal{O}(1/\epsilon^2)$ linear minimization oracle calls,
- $\mathcal{O}(1/\epsilon)$ gradient calls for f,
- $\mathcal{O}(1/\epsilon)$ (sub)gradient calls for h.

When the domain is strongly convex

The inner FW loop can be faster if both the surrogate and the feasible region have curvature.

- Strongly convex sets contain a small ball around every chord.
- Examples: ℓ_p-norm balls and Schatten-p balls for $1 < p \leq 2$.

Theorem

If f is strongly convex and $\mathcal{D}$ is strongly convex, then DCFW reaches

$$\min_{0 \leq \tau \leq t} \text{gap}_{\text{DC}}(\boldsymbol{x}_\tau) \leq \epsilon$$

within $\mathcal{O}(1/\epsilon)$ outer iterations and $\mathcal{O}(1/\epsilon^{3/2})$ linear minimization oracle calls.

Reason: FW has an $\mathcal{O}(1/k^2)$ rate for smooth strongly convex objectives over strongly convex sets.

Case study: QAP

Relax-and-round for the quadratic assignment problem:

$$\min_{\boldsymbol{X}} \langle \boldsymbol{A}^\top \boldsymbol{X}, \boldsymbol{X}\boldsymbol{B} \rangle \quad \text{s.t.} \quad \boldsymbol{X} \in [0,1]^{n \times n},\ \boldsymbol{X}\mathbf{1} = \boldsymbol{X}^\top \mathbf{1} = \mathbf{1}.$$

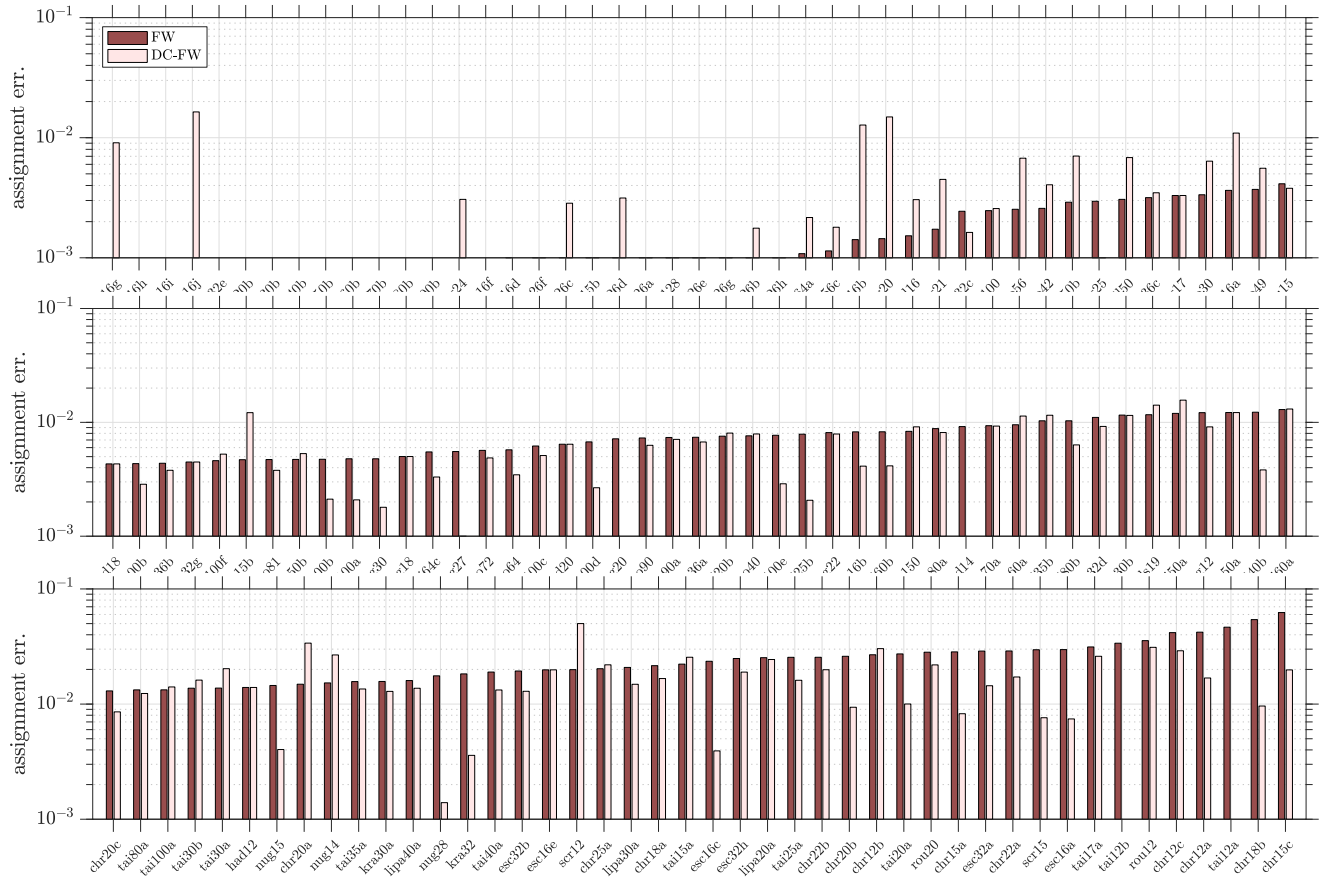
The feasible set is the Birkhoff polytope; the linear minimization oracle is a linear assignment problem.

DCFW uses the decomposition

$$f(\boldsymbol{X}) = \tfrac{1}{4}\|\boldsymbol{A}^\top \boldsymbol{X} + \boldsymbol{X}\boldsymbol{B}\|_{\mathrm{F}}^2, \qquad h(\boldsymbol{X}) = \tfrac{1}{4}\|\boldsymbol{A}^\top \boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\|_{\mathrm{F}}^2.$$

QAPLIB instances	DCFW better	FW better	tie
134 total	73	43	18

QAP numerical result



Assignment error after rounding. Instances are ordered from best to worst performance of FW. DCFW also achieved a lower average assignment error: 0.0085 vs. 0.0112.

Block Coordinate DCA

Motivation: QBO

Quadratic Binary Optimization:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \lambda\|\boldsymbol{x}\|_1 + \lambda d \quad \text{subject to} \quad \boldsymbol{x} \in \{-1,1\}^d$$

Convex hull relaxation:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \lambda\|\boldsymbol{x}\|_1 + \lambda d \quad \text{subject to} \quad \|\boldsymbol{x}\|_\infty \leq 1$$

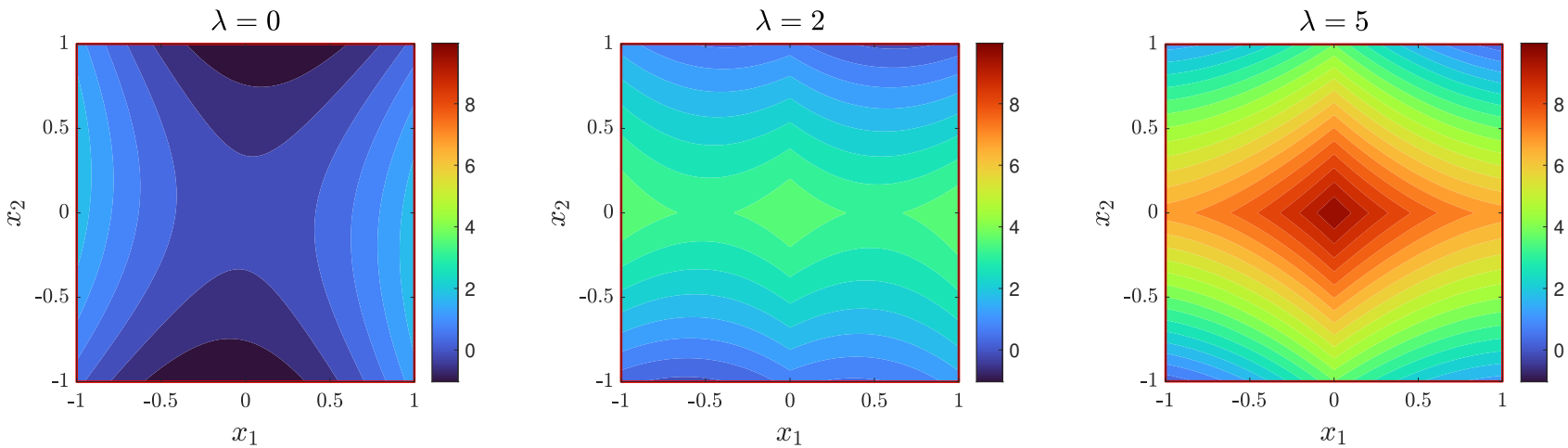


Figure: Illustration of level curves with increasing values of λ.

Negative ℓ_1 regularization promotes solutions closer to the binary set. We use $\boldsymbol{Q} = \begin{bmatrix} 2 & -0.25 \\ -0.25 & -1 \end{bmatrix}$.

Motivation: Challenge

$$\min_{x \in \mathbb{R}^d} \; x^\top Q x - \lambda \|x\|_1 \quad \text{subject to} \quad \|x\|_\infty \leq 1$$

How to decompose into f and h?

Approach I:

Split $Q = Q_P + Q_N$ (positive and negative semidefinite parts) and

$$f(x) = x^\top Q_P x + \iota_{\{\|x\|_\infty \leq 1\}}(x), \quad h(x) = -x^\top Q_N x + \lambda \|x\|_1.$$

Challenge:

- Requires eigendecomposition of Q; expensive for large d.
- Q_P and Q_N may lose useful structural properties (e.g., sparsity).

Approach II:

Consider nonconvex f by choosing

$$f(x) = x^\top Q x + \iota_{\{\|x\|_\infty \leq 1\}}(x), \quad h(x) = \lambda \|x\|_1.$$

Challenge:

Motivation

Coordinate descent idea: Minimize over one coordinate (or block coordinate) at a time, fixing all other coordinates.

Notation:

- $x_i \in \mathbb{R}^{d_i}$ represents the i^{th} coordinate block of $\boldsymbol{x}$.
- $\boldsymbol{x}_i \in \mathbb{R}^d$ is an extension of x_i (other coordinates padded with zeros).
- $\bar{\boldsymbol{x}}_i \in \mathbb{R}^d$ is the complement of $\boldsymbol{x}_i$, containing zeros in the i^{th} block and ensuring $\boldsymbol{x}_i + \bar{\boldsymbol{x}}_i = \boldsymbol{x}$.

Subproblem becomes:

$$x_{i_k}^{k+1} \leftarrow \underset{x_{i_k} \in [-1,1]}{\operatorname{argmin}} \quad (\bar{\boldsymbol{x}}_{i_k}^k + \boldsymbol{x}_{i_k})^\top \boldsymbol{Q} (\bar{\boldsymbol{x}}_{i_k}^k + \boldsymbol{x}_{i_k}) - \lambda \langle \operatorname{sign}(\boldsymbol{x}^k), \bar{\boldsymbol{x}}_{i_k}^k + \boldsymbol{x}_{i_k} \rangle$$

This reduces to a simple quadratic subproblem:

$$x_i^{k+1} \leftarrow \underset{x_i \in [-1,1]}{\operatorname{argmin}} \quad a x_i^2 + b x_i + c,$$

- $a = q_{ii}$ (q_{ii} is the ith diagonal entry of $\boldsymbol{Q}$)
- $b = 2\boldsymbol{q}_i^\top \bar{\boldsymbol{x}}_i^k - \lambda \operatorname{sign}(x_i^k)$ ($\boldsymbol{q}_i$ is the ith column of $\boldsymbol{Q}$)

Problem template

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \quad \phi(\boldsymbol{x}) := f(\boldsymbol{x}) + g(\boldsymbol{x}) - h(\boldsymbol{x}), \tag{DC}$$

- f, g and h are lower semi-continuous functions. h is convex.
- f is L-smooth, g can be nonsmooth (can take $+\infty$).
- g is (block) coordinate-wise separable: $g(\boldsymbol{x}) = \sum_{i=1}^{n} g_i(x_i)$.
- We assume that the optimal value $\phi^\star$ is finite.

Block coordinate DCA

1. Start from a feasible initial point $\boldsymbol{x}^1 \in \mathbb{R}^d$.
2. For $k = 1, \ldots, K$, update the estimate as follows:
   a. Choose a block i_k uniformly at random from $\{1, \ldots, n\}$.
   b. Keep all other coordinates fixed, except i_k, i.e., $\bar{\boldsymbol{x}}_{i_k}^{k+1} = \bar{\boldsymbol{x}}_{i_k}^{k}$.
   c. Update the i_k^{th} block by minimizing the surrogate objective along these coordinates:

$$x_{i_k}^{k+1} = \underset{x_{i_k} \in \mathbb{R}^{d_i}}{\operatorname{argmin}} \hat{\phi}(\boldsymbol{x}, \boldsymbol{x}^k) \quad \text{subj.to} \quad \boldsymbol{x} = \boldsymbol{x}_{i_k} + \bar{\boldsymbol{x}}_{i_k}^{k}$$

Let us denote the surrogate objective restricted to i^{th} block by

$$\begin{aligned}\hat{\phi}_i(x_i, \boldsymbol{x}^k) &:= \hat{\phi}(\boldsymbol{x}_i + \bar{\boldsymbol{x}}_i^k, \boldsymbol{x}^k) \\ &= f(\boldsymbol{x}_i + \bar{\boldsymbol{x}}_i^k) + g_i(x_i) - g_i(x_i^k) + g(\boldsymbol{x}^k) - h(\boldsymbol{x}^k) - \langle u_i^k, x_i - x_i^k \rangle.\end{aligned}$$

where u_i^k is the i_k^{th} block of a subgradient $\boldsymbol{u}^k \in \partial h(\boldsymbol{x}^k)$.

(Modified) Gap function

Definition (DC Gap)

We measure convergence in terms of the following gap function:

$$\mathrm{gap}_{\mathrm{DC}}(\boldsymbol{y}) = \max_{\boldsymbol{x}\in\mathbb{R}^d} \min_{\boldsymbol{u}\in\partial h(\boldsymbol{y})} \{f(\boldsymbol{y}) - f(\boldsymbol{x}) + g(\boldsymbol{y}) - g(\boldsymbol{x}) - \langle \boldsymbol{u}, \boldsymbol{y}-\boldsymbol{x}\rangle\}.$$

Definition (BDC Gap)

Consider the following gap function:

$$\mathrm{gap}^L_{\mathrm{BDC}}(\boldsymbol{y}) = \max_{\boldsymbol{x}\in\mathbb{R}^d} \min_{\boldsymbol{u}\in\partial h(\boldsymbol{y})} \left\{\langle \nabla f(\boldsymbol{y}) - \boldsymbol{u}, \boldsymbol{y}-\boldsymbol{x}\rangle + g(\boldsymbol{y}) - g(\boldsymbol{x}) - \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2\right\}$$

Lemma

- $\mathrm{gap}^L_{\mathrm{BDC}}(\boldsymbol{y}) \geq 0$ *for all* $\boldsymbol{y}\in\mathbb{R}^d$,
- $\mathrm{gap}^L_{\mathrm{BDC}}(\boldsymbol{y}) = 0$ *if and only if* $\boldsymbol{y}$ *is a critical point:*

$$(\nabla f(\boldsymbol{y}) + \partial g(\boldsymbol{y})) \cap \partial h(\boldsymbol{y}) \neq \emptyset.$$

If h is differentiable, then any such $\boldsymbol{y}$ is a first-order stationary point.

Convergence guarantees of BDCA

Theorem

Suppose $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^K$ *is a sequence generated by BDCA, where each subproblem is solved approximately to satisfy*

$$\hat{\phi}_{i_k}(x_{i_k}^{k+1}, \boldsymbol{x}^k) - \min_{x_{i_k} \in \mathbb{R}^{d_i}} \hat{\phi}_{i_k}(x_{i_k}, \boldsymbol{x}^k) \leq \frac{\epsilon}{2}$$

for some target accuracy $\epsilon > 0$*. Then,*

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\left[\text{gap}_{\text{BDC}}^{L}(\boldsymbol{x}^k)\right] \leq \frac{n}{K}\left(\phi(\boldsymbol{x}^1) - \phi^\star\right) + \frac{\epsilon}{2}.$$

Numerical experiment (methods)

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \lambda\|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \|\boldsymbol{x}\|_\infty \leq 1$$

We consider 3 methods:

(Method 1) BDCA with nonconvex quadratic f:

$$f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}, \quad g(\boldsymbol{x}) = \iota_{\{\|\boldsymbol{x}\|_\infty \leq 1\}}(\boldsymbol{x}), \quad h(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1$$

(Method 2) DCA with eigensplitting:

$$f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{Q}_P\boldsymbol{x} + \iota_{\{\|\boldsymbol{x}\|_\infty \leq 1\}}(\boldsymbol{x}), \quad h(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1 - \boldsymbol{x}^\top \boldsymbol{Q}_N\boldsymbol{x}$$

(Method 3) BDCA with quadratic majorization:

$$f(\boldsymbol{x}) = \tfrac{L}{2}\|\boldsymbol{x}\|^2, \quad g(\boldsymbol{x}) = \iota_{\{\|\boldsymbol{x}\|_\infty \leq 1\}}(\boldsymbol{x}), \quad h(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1 + \tfrac{L}{2}\|\boldsymbol{x}\|^2 - \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$$

where $L = 2\|\boldsymbol{Q}\|$.

Numerical experiment (results)

Experimental setup:

- $\lambda = \|Q\|_F/\sqrt{d}$
- Initialization: same Gaussian *iid* vector for all methods
- Stop when gap reaches 10^{-6}
- GSet graph instances (67 datasets, d ranging from 800 to 10′000)

Table: Rounded objective values and running times for different methods on the largest 5 GSet instances.

Dataset	Size (d)	OBJ			TIME (s)		
		M1	M2	M3	M1	M2	M3
G63	7000	**-107010.03**	-61763	**-107010.03**	**117.79**	4218.65	354.19
G64	7000	**-49222.03**	-48250.03	-48670.03	**43.75**	1535.71	365.12
G65	8000	-32372	-32156	**-32532**	**17.69**	3468.09	32.90
G66	9000	**-36720**	-36360	-36704	**25.39**	1011.99	55.51
G67	10000	**-40428**	-40052	-40402	**33.96**	6298.46	63.50

⚠ Out of 67 problems, M1 finds the best solution in 56 instances, M2 in 22 instances, and M3 in 32 instances.

References

- Yurtsever, A. & Sra, S. (2022). CCCP is Frank-Wolfe in disguise.
- Maskan, H. Halvachi, P., Sra, S., & Yurtsever, A. (2024). Randomized block coordinate DC algorithm.
- Maskan, H., Hou, Y., Sra, S., & Yurtsever, A. (2025). Revisiting Frank-Wolfe for structured nonconvex optimization.
- Pham Dinh, T., & Souad, E.B. (1986). Algorithms for solving a class of nonconvex optimization problems: methods of subgradient.
- Frank, M., & Wolfe, P. (1956). An algorithm for quadratic programming.
- Jaggi, M. (2013). Revisiting Frank-Wolfe: Projection-free sparse convex optimization.
- Lacoste-Julien, S. (2016). Convergence rate of Frank-Wolfe for non-convex objectives.
- Journée, M., Nesterov, Y., Richtárik, P., & Sepulchre, R. (2010). Generalized Power Method for Sparse Principal Component Analysis.
- Kumar, R., & Schmidt, M. (2017) Convergence rate of expectation maximization.
- Kunstner, F., Kumar, R., & Schmidt, M. (2020) Homeomorphic-invariance of EM: Non-asymptotic converfence in KL divergence for exponential families via mirror descent.