A globally convergent difference-of-convex algorithmic framework and application to log-determinant optimization problems

Xin Jiang

Institute for Data, Intelligent Systems, and Computation
Industrial and Systems Engineering Department
Lehigh University

Modeling and Optimization: Theory and Applications (MOPTA)

August 17, 2023

Difference-of-convex (DC) programming

consider the class of difference-of-convex (DC) optimization problems

$$\begin{array}{ll} \text{minimize} & f(x) = g(x) - h(x) \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

- g, h are closed, convex, and continuously differentiable
- different assumptions can be posed on $\mathcal{C}$
- assume optimum is attained at $x^\star$, with finite optimal value $f^\star$

Applications: some problems have an equivalent DC reformulation

- problems with a concave objective
- some bilevel optimization problems
- some nonconvex regularizers have DC reformulation or relaxation

Lipp & Boyd (2016), de Oliveira (2020)

Difference-of-convex algorithm (DCA)

the difference-of-convex algorithm (DCA) is a conceptually simple method

$$x^{(k+1)} \in \operatorname*{argmin}_{x \in \mathcal{C}} \left(g(x) - \left(h(x^{(k)}) + \langle \nabla h(x^{(k)}), x - x^{(k)} \rangle\right)\right)$$

it has been studied under various names

- a special case of the majorization–minimization (MM) algorithm
- nonsmooth extension exists ($\nabla h(x^{(k)})$ is replaced with a subgradient of h)
- also known as the convex–concave procedure (CCCP)

most research focuses on $\mathcal{C}$ is the entire space or defined by DC functions

Properties and convergence results

- monotonicity of function values: $f(x^{(k+1)}) \leq f(x^{(k)})$ for all $k \in \mathbb{N}$
- DCA converges to a first-order stationary point with an $O(1/k)$ rate

Tao and Souad (1986)
Yuille and Rangarajan (2003), Sriperumbudur and Lanckriet (2009), Smola et al. (2015)

Motivation and contributions

Running example from network information theory

$$\begin{array}{ll} \text{minimize} & -\log\det(X+\Sigma_1)+\lambda\log\det(X+\Sigma_2) \\ \text{subject to} & 0 \preceq X \preceq C \end{array}$$

with variable $X \in \mathbb{S}^n$; data $\Sigma_1, \Sigma_2 \in \mathbb{S}^n_{++}$, $C \in \mathbb{S}^n_+$, $\lambda > 1$

- the problem is nonconvex as $\lambda > 1$
- the problem has a unique global optimum (Lau, Nair, and Yao (2022))

Motivation and contributions

Running example from network information theory

$$\begin{array}{ll} \text{minimize} & -\log\det(X+\Sigma_1)+\lambda\log\det(X+\Sigma_2) \\ \text{subject to} & 0 \preceq X \preceq C \end{array}$$

with variable $X \in \mathbb{S}^n$; data $\Sigma_1, \Sigma_2 \in \mathbb{S}^n_{++}$, $C \in \mathbb{S}^n_{+}$, $\lambda > 1$

- the problem is nonconvex as $\lambda > 1$
- the problem has a unique global optimum (Lau, Nair, and Yao (2022))

Contributions

- Global linear convergence of DCA under generalized PL conditions
- Subproblem solver: primal–dual proximal methods with Bregman distances
- Application to several problems in various fields

Outline

Two interpretations of DCA
- DCA from Frank–Wolfe algorithm
- DCA from Bregman proximal point algorithm

Convergence of DCA to global optimum

Bregman PDHG as subproblem solver

Applications and numerical results

Frank–Wolfe algorithm

consider the canonical optimization problem

$$\begin{array}{ll} \text{minimize} & \psi(z) \\ \text{subject to} & z \in \mathcal{D}, \end{array}$$

where $\mathcal{D}$ is closed and convex, and ψ is continuously differentiable

Frank–Wolfe algorithm takes the following iterations

$$\begin{aligned} \hat{z} &\in \operatorname*{argmin}_{z \in \mathcal{D}} \big(\langle \nabla \psi(z^{(k)}), z - z^{(k)} \rangle \big) \\ z^{(k+1)} &= (1 - \theta_k) z^{(k)} + \theta_k \hat{z}, \end{aligned}$$

where $\theta_k \in [0, 1]$ can be chosen via various techniques

- if ψ is convex or concave, FW converges with an $O(1/k)$ rate
- if ψ is nonconvex, FW converges to a stationary point with rate $O(1/\sqrt{k})$

DCA from FW algorithm

- the DC program can be rewritten as

$$\begin{array}{ll} \text{minimize} & t - h(x) \\ \text{subject to} & g(x) + \delta_{\mathcal{C}}(x) \leq t \end{array}$$

with variables $x \in \mathbb{R}^d$ and $t \in \mathbb{R}$

Yurtsever and Sra (2022)

DCA from FW algorithm

- the DC program can be rewritten as

$$\begin{array}{ll} \text{minimize} & t - h(x) \\ \text{subject to} & g(x) + \delta_{\mathcal{C}}(x) \le t \end{array}$$

  with variables $x \in \mathbb{R}^d$ and $t \in \mathbb{R}$
- the $\hat{z}$-update in FW method linearizes the objective

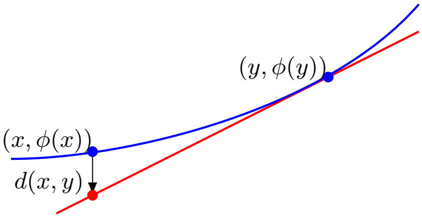
$$\begin{aligned} \hat{z} \in & \operatorname*{argmin}_{z=(x,t)\in\mathcal{D}} \ \langle \nabla\psi(z^{(k)}), z - z^{(k)} \rangle \\ &= \operatorname*{argmin}_{(x,t)\in\mathcal{D}} \Big(t - \langle \nabla h(x^{(k)}), x - x^{(k)} \rangle\Big) \\ &= \operatorname*{argmin}_{x\in\mathcal{C}} \Big(g(x) - \langle \nabla h(x^{(k)}), x - x^{(k)} \rangle\Big), \end{aligned}$$

  where $\psi(x,t) = t - h(x)$ is concave
- it can be shown that $\theta_k = 1$ is valid in this case
- previous $O(1/k)$ convergence result applies

Yurtsever and Sra (2022)

Bregman distance (generalized distance)

$$d_\phi(x,y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x-y\rangle$$



- ϕ is the *kernel function*
- ϕ is convex and continuously differentiable on $\mathrm{int}(\mathrm{dom}\,\phi)$

other properties of ϕ may be required; *e.g.*, strict convexity implies

$$d_\phi(x,y) = 0 \quad \Longrightarrow \quad x = y$$

Bregman (1967), Censor and Zenios (1997)

Bregman proximal point algorithm (BPPA)

BPPA minimizes a closed convex function ψ via the iterations

$$x^{(k+1)} = \underset{x}{\operatorname{argmin}} \Big(\psi(x) + \tfrac{1}{\alpha_k} d_\phi(x, x^{(k)}) \Big)$$

- assume the subproblem has a unique solution at every iteration

Censor and Zenios (1992), Auslender and Teboulle (2006), Tseng (2008)
Faust et al. (2023)

Bregman proximal point algorithm (BPPA)

BPPA minimizes a closed convex function ψ via the iterations

$$x^{(k+1)} = \operatorname*{argmin}_x \big(\psi(x) + \tfrac{1}{\alpha_k} d_\phi(x, x^{(k)}) \big)$$

- assume the subproblem has a unique solution at every iteration

DCA from BPPA

- consider again the DC program

$$\text{minimize} \quad \psi(x) = g(x) + \delta_{\mathcal{C}}(x) - h(x)$$

- BPPA follows the iterations (take $\phi = h$ and $\alpha_k = 1$ for all $k \in \mathbb{N}$)

$$\begin{aligned} x^{(k+1)} &= \operatorname{argmin} \big(\psi(x) + d_h(x, x^{(k)}) \big) \\ &= \operatorname*{argmin}_{x \in \mathcal{C}} \big(g(x) - h(x) + h(x) - h(x^{(k)}) - \langle \nabla h(x^{(k)}), x - x^{(k)} \rangle \big) \\ &= \operatorname*{argmin}_{x \in \mathcal{C}} \big(g(x) - h(x^{(k)}) - \langle \nabla h(x^{(k)}), x - x^{(k)} \rangle \big) \end{aligned}$$

Censor and Zenios (1992), Auslender and Teboulle (2006), Tseng (2008)
Faust et al. (2023)

Outline

Polyak–Łojasiewicz (PL) inequality

a function $\psi\colon \mathbb{R}^n \to \mathbb{R}$ is said to satisfy PL inequality on a set $\mathcal{D}$ if

$$\exists \mu > 0 \quad \text{s.t.} \ \ \psi(x) - \psi^\star \leq \tfrac{1}{2\mu}\|\xi\|_2^2, \ \ \text{for all } x \in \mathcal{D} \text{ and } \xi \in \operatorname{conv}(\widehat{\partial}\psi(x)),$$

where $\widehat{\partial}\psi(x)$ is the regular subdifferential of ψ

- existence of $\widehat{\partial}\psi$ requires ψ to be locally Lipschitz continuous
- for differentiable ψ, PL inequality reduces to $\psi(x) - \psi^\star \leq \frac{1}{2\mu}\|\nabla\psi(x)\|_2^2$

Polyak (1963), Łojasiewicz (1963), Abbaszadehpeivasti, de Klerk, and Zamani (2023)

Polyak–Łojasiewicz (PL) inequality

a function $\psi\colon \mathbb{R}^n \to \mathbb{R}$ is said to satisfy PL inequality on a set $\mathcal{D}$ if

$$\exists \mu > 0 \quad \text{s.t. } \psi(x) - \psi^\star \le \tfrac{1}{2\mu}\|\xi\|_2^2, \text{ for all } x \in \mathcal{D} \text{ and } \xi \in \operatorname{conv}(\widehat{\partial}\psi(x)),$$

where $\widehat{\partial}\psi(x)$ is the regular subdifferential of ψ

- existence of $\widehat{\partial}\psi$ requires ψ to be locally Lipschitz continuous
- for differentiable ψ, PL inequality reduces to $\psi(x) - \psi^\star \le \frac{1}{2\mu}\|\nabla\psi(x)\|_2^2$

Global linear convergence of DCA assume for the DC program

- $\mathcal{C} = \mathbb{R}^d$, g and h are (globally) Lipschitz continuous with $L_g, L_h > 0$
- f satisfies PL inequality on $\mathcal{D} = \{x \mid f(x) \le f(x_0)\}$

then for all $k \in \mathbb{N}$,

$$f(x^{(k+1)}) - f^\star \le \left(\frac{1 - \mu/L_g}{1 + \mu/L_h}\right)\left(f(x^{(k)}) - f^\star\right)$$

Polyak (1963), Łojasiewicz (1963), Abbaszadehpeivasti, de Klerk, and Zamani (2023)

Generalized PL condition

Generalized PL condition for DC programs there exists $\mu, r \in \mathbb{R}_{++}$ s.t.

$$\mu(f(x) - f^\star) \leq d_{h^*}(\nabla g(x) + y, \nabla h(x)), \quad \text{for all } x \in \mathcal{C}, y \in N_{\mathcal{C}}(x) \cap \mathcal{B}(r),$$

where $N_{\mathcal{C}}(x)$ is the normal cone of $\mathcal{C}$ at x, and $\mathcal{B}(r) = \{y \mid \|y\|_2 \leq r\}$

- DC program is formulated as an unconstrained problem with objective

$$\psi(x) = f(x) + \delta_{\mathcal{C}}(x) = g(x) + \delta_{\mathcal{C}}(x) - h(x)$$

- Euclidean distance in PL inequality is generalized to a Bregman distance

$$\|\xi\|_2^2 = \|\nabla g(x) + y - \nabla h(x)\|_2^2 \quad \Longrightarrow \quad d_{h^*}(\nabla g(x) + y, \nabla h(x))$$

Faust et al. (2023): a simpler version of this condition (with $\mathcal{C} = \mathbb{R}^d$ and more assumptions on g, h)
Yao and **Jiang** (2023)

Generalized PL condition

Generalized PL condition for DC programs there exists $\mu, r \in \mathbb{R}_{++}$ s.t.

$$\mu(f(x) - f^\star) \leq d_{h^*}(\nabla g(x) + y, \nabla h(x)), \quad \text{for all } x \in \mathcal{C}, y \in N_{\mathcal{C}}(x) \cap \mathcal{B}(r),$$

where $N_{\mathcal{C}}(x)$ is the normal cone of $\mathcal{C}$ at x, and $\mathcal{B}(r) = \{y \mid \|y\|_2 \leq r\}$

- DC program is formulated as an unconstrained problem with objective

$$\psi(x) = f(x) + \delta_{\mathcal{C}}(x) = g(x) + \delta_{\mathcal{C}}(x) - h(x)$$

- Euclidean distance in PL inequality is generalized to a Bregman distance

$$\|\xi\|_2^2 = \|\nabla g(x) + y - \nabla h(x)\|_2^2 \quad \Longrightarrow \quad d_{h^*}(\nabla g(x) + y, \nabla h(x))$$

Global linear convergence of DCA

$$f(x^{(k+1)}) - f^\star \leq \frac{1}{1+\mu}(f(x^{(k)}) - f^\star)$$

Faust et al. (2023): a simpler version of this condition (with $\mathcal{C} = \mathbb{R}^d$ and more assumptions on g, h)
Yao and **Jiang** (2023)

Outline

DCA for running example

consider the running example

$$\begin{array}{ll} \text{minimize} & -\log\det(X+\Sigma_1) + \lambda \log\det(X+\Sigma_2) \\ \text{subject to} & 0 \preceq X \preceq C \end{array}$$

with variable $X \in \mathbb{S}^n$; data $\Sigma_1, \Sigma_2, C \in \mathbb{S}^n_{++}$, and $\lambda > 1$

- DCA takes the iterations

$$X^{(k+1)} = \underset{0 \preceq X \preceq C}{\operatorname{argmin}} \Big(-\log\det(X+\Sigma_1) + \langle (X^{(k)}+\Sigma_2)^{-1}, X \rangle \Big)$$

- at each DCA iteration, one solves the convex subproblem of the form

$$\begin{array}{ll} \text{minimize} & -\log\det(X+\Sigma_1) + \langle A, X \rangle \\ \text{subject to} & 0 \preceq X \preceq C \end{array}$$

  with variable $X \in \mathbb{S}^n$ and data $\Sigma_1, A \in \mathbb{S}^n_{++}$

Bregman primal–dual hybrid gradient method

consider the canonical convex problem

$$\text{minimize} \quad F(u) + G(\mathcal{A}u),$$

where F, G are convex, (potentially) nonsmooth, and $\mathcal{A}$ is a linear operator

Bregman PDHG

$$\begin{aligned} u^{(k+1)} &= \operatorname*{argmin}_u \Big(F(u) + \langle v^{(k)}, \mathcal{A}u \rangle + \tfrac{1}{\tau} d_{\phi_\mathrm{p}}(u, u^{(k)}) \Big) \\ \overline{u}^{(k+1)} &= u^{(k+1)} + \theta(u^{(k+1)} - u^{(k)}) \\ v^{(k+1)} &= \operatorname*{argmin}_v \Big(G^*(v) - \langle v, \mathcal{A}\overline{u}^{(k+1)} \rangle + \tfrac{1}{\sigma} d_{\phi_\mathrm{d}}(v, v^{(k)}) \Big) \end{aligned}$$

where ϕ_p, ϕ_d are two kernel functions, σ, τ, and θ are stepsizes

Chambolle and Pock (2016)

Discussion on Bregman PDHG

Potential benefits of Bregman distances in PDHG

1. make the generalized proximal mapping easier to compute
2. "preconditioning": use a more accurate model of $F(u)$ around $u^{(k)}$

goal of 1 is to reduce cost per iteration
goal of 2 is to reduce number of iterations

Malitsky & Pock (2018), Yazdandoost Hamedani & Aybat (2021), **Jiang** & Vandenberghe (2022)

Discussion on Bregman PDHG

Potential benefits of Bregman distances in PDHG

1. make the generalized proximal mapping easier to compute
2. "preconditioning": use a more accurate model of $F(u)$ around $u^{(k)}$

goal of 1 is to reduce cost per iteration
goal of 2 is to reduce number of iterations

Requirements

- the minimizer in u (and v) update exists and is unique
- ϕ_{p}, ϕ_{d} are two strongly convex Bregman kernels

$$d_{\mathrm{p}}(u, u') \geq \frac{1}{2}\|u - u'\|_{\mathrm{p}}^2, \qquad d_{\mathrm{d}}(v, v') \geq \frac{1}{2}\|v - v'\|_{\mathrm{d}}^2$$

- stepsizes must satisfy $\sigma\tau\|\mathcal{A}\|^2 \leq 1$, where

$$\|\mathcal{A}\| = \sup_{u \neq 0, v \neq 0} \frac{\langle v, \mathcal{A}u \rangle}{\|v\|_{\mathrm{d}} \|u\|_{\mathrm{p}}}$$

- line search techniques are developed to adaptively choose the stepsizes

Malitsky & Pock (2018), Yazdandoost Hamedani & Aybat (2021), **Jiang** & Vandenberghe (2022)

Bregman PDHG as subproblem solver

apply Bregman PDHG to the subproblem

$$\text{minimize} \quad -\log\det(X+\Sigma_1) + \langle A, X\rangle + \delta_{\mathbb{S}^n_+}(X) + \delta_{\{X \mid X \preceq C\}}(X)$$

- take $\phi_{\mathrm{d}} = \frac{1}{2}\|\cdot\|_F^2$, dual update involves PSD projection
- take $\phi_{\mathrm{p}}(X) = -\log\det(X+\Sigma_1)$, primal update involves the problem

$$\begin{array}{ll} \text{minimize} & -(1+\frac{1}{\tau})\log\det(X+\Sigma_1) + \langle B, X\rangle \\ \text{subject to} & X \succeq 0 \end{array}$$

  with variable $X \in \mathbb{S}^n$ and data $\Sigma_1, B \in \mathbb{S}^n_{++}$
- this problem has a closed-form solution

$$X^\star = \Sigma_1^{1/2} Q \zeta(\Lambda) Q^T \Sigma_1^{1/2}, \quad \text{where } \zeta(\gamma) = \max\{(1-\gamma)/\gamma, 0\}$$

  and $\Sigma_1^{1/2} B \Sigma_1^{1/2} = Q\Lambda Q^T$ is the eigen-decomposition

A general algorithmic framework for DC programming

$$\begin{array}{ll} \text{minimize} & f(x) = g(x) - h(x) \\ \text{subject to} & x \in \mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2 \end{array}$$

- g, h are differentiable, and strongly convex on $\mathcal{C}$
- $\mathcal{C}_1$, $\mathcal{C}_2$ are bounded, convex; projection on $\mathcal{C}_1$, $\mathcal{C}_2$ is much easier than on $\mathcal{C}$
- recall the DCA iteration

$$x^{(k+1)} = \operatorname*{argmin}_{x \in \mathcal{C}_1 \cap \mathcal{C}_2} \left(g(x) - \langle \nabla h(x^{(k)}), x \rangle \right)$$

Bregman PDHG as subproblem solver

- reformulate the DCA subproblem as minimizing $F + G \circ \mathcal{A}$ with

$$F = g - \langle \nabla h(x^{(k)}), \cdot \rangle + \delta_{\mathcal{C}_1}, \quad G = \delta_{\mathcal{C}_2}, \quad \mathcal{A} = \mathrm{Id}$$

- with $\phi_{\mathrm{p}} = g$, primal PDHG update reduces to a Bregman projection

$$u^{(t+1)} = \operatorname*{argmin}_{u \in \mathcal{C}_1} d_g(u, \tilde{u}),$$

  where $\tilde{u}$ depends on data and previous iterates
  (t is PDHG iteration counter while k is DCA counter)

Yao and **Jiang** (2023)

Outline

Two interpretations of DCA

Convergence of DCA to global optimum

Bregman PDHG as subproblem solver

Applications and numerical results

Numerical results for running example

n	algo	num. of DCA iter.	num. of inner iter.	runtime (in sec.)	runtime per DCA iter.
500	DCA-PDHG (Breg.)	9.5	1735	3.63×10^2	38.23
	DCA-PDHG (Euc.)	9.5	2046	3.81×10^2	40.09
	DCA-MOSEK	8.9	76	1.02×10^3	108.1
1000	DCA-PDHG (Breg.)	13.6	1324	1.73×10^3	127.2
	DCA-PDHG (Euc.)	13.6	1684	2.20×10^3	162.4
	DCA-MOSEK	13.2	96	9.87×10^3	726.3

- results are averaged over 10 synthetic datasets
- DCA-PDHG (Euc.) uses Euclidean PDHG as subproblem solver
  each PDHG iteration involves two `eigens` and solving n quadratic systems
- DCA-MOSEK uses the interior-point-method-based solver MOSEK

Example: Gaussian broadcast channel

$$\begin{array}{ll} \text{minimize} & -\beta \log\det(X+Y+\Sigma_2) + \alpha \log\det(X+Y+\Sigma_1) \\ & -\log\det(X+\Sigma_1) + \lambda \log\det(X+\Sigma_2) \\ \text{subject to} & X+Y \preceq C, \quad X \succeq 0, \quad Y \succeq 0 \end{array}$$

with variables $X, Y \in \mathbb{S}^n$; data $\Sigma_1, \Sigma_2, C \in \mathbb{S}^n_{++}$, $\alpha \in [0,1]$, $\beta > 0$, $\lambda > 1$

- the objective satisfies the generalized PL condition
- PDHG iteration has a closed-form expression, and is dominated by `eigen`

n	algo	num. of DCA iter.	num. of inner iter.	runtime (in sec.)	runtime per DCA iter.
	DCA-PDHG (Breg.)	10.2	1273	5.63×10^2	56.07
500	DCA-PDHG (Euc.)	10.2	1496	5.71×10^2	75.83
	DCA-MOSEK	9.8	93	2.32×10^3	225.1
	DCA-PDHG (Breg.)	12.4	1468	3.50×10^3	281.9
1000	DCA-PDHG (Euc.)	12.4	1632	4.08×10^3	313.3
	DCA-MOSEK	-	-	-	-

Example: generalized Brascamp–Lieb inequality

this problem generalizes the computation of Brascamp–Lieb constant

$$
\begin{array}{ll}
\text{minimize} & -\sum_{i=1}^{p} \beta_i \log\det X_i + \sum_{j=1}^{q} \alpha_j \log\det\big(\sum_{i=1}^{p} A_{ij} X_i A_{ij}^T + \rho I_{m_j}\big) \\
\text{subject to} & 0 \preceq X_i \preceq C_i, \quad i = 1, \ldots, p
\end{array}
$$

with variable $X_i \in \mathbb{S}^{n_i}$; and data $A_{ij} \in \mathbb{R}^{m_j \times n_i}$, $C_i \in \mathbb{S}_+^{n_i}$, $\alpha \in \mathbb{R}_+^q$, $\beta \in \mathbb{R}_+^p$

- its optimum computes the optimal constant for a family of inequalities
- it covers the well-known Brascamp–Lieb inequality (with $\mathbf{1}^T \alpha = 1$)

$$
f_{\mathsf{BL}}(X) = -\log\det X + \sum_{j=1}^{q} \alpha_j \log\det(A_j X A_j^T)
$$

- this problem satisfies the generalized PL condition

Bregman PDHG as subproblem solver

- in DCA subproblem, the variables $\{X_i\}$ are separable
- PDHG update has a closed-form expression, and is dominated by `eigen`

Numerical results

n	algo	num. of DCA iter.	num. of inner iter.	runtime (in sec.)	runtime per DCA iter.
	DCA-PDHG (Breg.)	14.7	1157.9	9.98×10^2	64.21
500	DCA-PDHG (Euc.)	14.7	1297.5	1.14×10^3	70.42
	DCA-MOSEK	13.9	85.2	5.36×10^4	364.8
	DCA-PDHG (Breg.)	14.2	1048.7	5.74×10^3	412.6
1000	DCA-PDHG (Euc.)	14.2	1362.6	6.52×10^3	468.7
	DCA-MOSEK	-	-	-	-

- results are averaged over 10 synthetic datasets ($p = q = 3$, $n_i = n$)
- Bregman PDHG takes fewer iterations and has cheaper per-iteration cost
- IPM-based solver has much more expensive per-iteration complexity

Summary

New convergence results for DCA

- generalized PL condition for DC programs with set constraints
- convergence to global optimum with linear rate

Bregman PDHG as subproblem solver

- split the constraint set into $\mathcal{C}_1$ and $\mathcal{C}_2$
- primal distance generated by g
- primal PDHG update is Bregman projection on a simple convex set

Applications in network information theory

- generalized PL condition is satisfied
- each PDHG iteration has closed-form expression
- per-iteration cost is comparable to eigen-decomposition