Primal–dual proximal methods with Bregman distances

with applications to sparse SDP

Xin Jiang

Department of Electrical and Computer Engineering
University of California, Los Angeles

joint work with Lieven Vandenberghe

EUROPT Workshop on Continuous Optimization
Toulouse (virtual), July 7–9, 2021

Problem formulation

$$\text{minimize} \quad f(x) + g(Ax) + h(x)$$

- f, g and h are closed convex functions
- f and g have simple proximal operators, h is differentiable and L-smooth
- three-operator splitting algorithms: Condat–Vũ, PD3O, PDDY

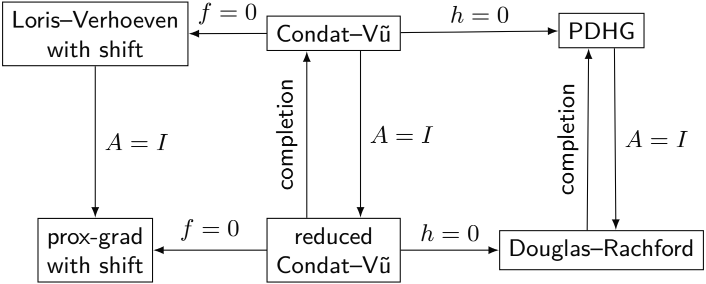
Algorithms for special cases

- $g = 0$: proximal gradient
- $h = 0$: ADMM, PDHG (Chambolle-Pock), Douglas–Rachford, *etc.*
- $f = 0$: Loris–Verhoeven (a.k.a., PDFP^{2}O, PAPC)
- $A = I$: Davis–Yin

Condat (2013), Vũ (2013), Yan (2018), Salim et al. (2020)
Boyd et al. (2010), Chambolle and Pock (2011, 2016)
Loris and Verhoeven (2011), Chen et al. (2013), Drori et al. (2015), Davis and Yin (2015)

Condat–Vũ three-operator splitting method

$$\text{minimize} \quad f(x) + g(Ax) + h(x)$$



- "completion": PDHG with DRS applied to a reformulation
- "with shift" means the gradient of h is evaluated at a shifted point

similar scheme also exists for PD3O and PDDY

"completion" trick: O'Connor and Vandenberghe (2018), Condat et al. (2020)

Proximal mapping

Proximal mapping: for closed convex function f

$$\mathrm{prox}_f(x) = \underset{y}{\mathrm{argmin}} \left(f(y) + \frac{1}{2}\|x - y\|_2^2 \right)$$

Generalized proximal mapping: use a generalized distance $d(x, y)$

$$\mathrm{prox}_f^{\phi}(y, a) = \underset{x}{\mathrm{argmin}} \left(f(x) + \langle a, x \rangle + d(x, y) \right)$$

for example, in proximal gradient method for minimizing $g(x) + h(x)$:

$$x_{k+1} = \underset{x}{\mathrm{argmin}} \left(h(x) + g(x_k) + \langle \nabla g(x_k), x - x_k \rangle + \frac{1}{\tau} {\color{red} d(x, x_k)} \right)$$

Potential benefits

1. "preconditioning": use a more accurate model of $g(x)$ around x_k
2. make the generalized proximal mapping easier to compute

Outline

Proximal methods with generalized distances

Applications to sparse semidefinite programs (SDPs)

Outline

Proximal methods with generalized distances

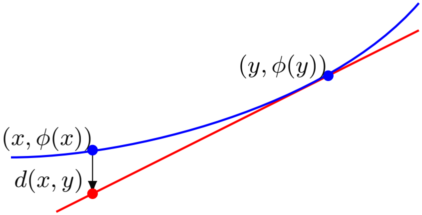
- Bregman proximal operator
- Bregman first-order splitting methods
- Line search in Bregman proximal methods

Applications to sparse semidefinite programs (SDPs)

Bregman distance (generalized distance)

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

where the kernel ϕ is convex, differentiable on its interior domain



Examples

- squared Euclidean distance: $\phi(x) = \frac{1}{2}\|x\|_2^2$ and $d(x, y) = \frac{1}{2}\|x - y\|_2^2$
- relative entropy:

$$\phi(x) = \sum_{i=1}^{n} x_i \log x_i, \qquad d(x, y) = \sum_{i=1}^{n} (x_i \log(x_i/y_i) - x_i + y_i)$$

Bregman (1967), Censor and Zenios (1997)

Generalized proximal mapping: difficulties

$$\operatorname{prox}_f^\phi(y, a) = \operatorname*{argmin}_x \big(f(x) + \langle a, x\rangle + d(x, y)\big)$$

- no simple condition for existence and uniqueness of minimizer x
- no simple analog to Moreau decomposition

$$\operatorname{prox}_{\tau f}(x) + \tau \operatorname{prox}_{\tau^{-1} f^*}(x/\tau) = x$$

- no simple extension of affine composition rule: if $g(x) = f(Ax + b)$

$$\operatorname{prox}_g(x) = x - \alpha A^T\big(Ax + b - \operatorname{prox}_{\alpha^{-1} f}(Ax + b)\big),$$

when $AA^T = (1/\alpha)I$

Bregman Condat–Vũ algorithm

$$\text{minimize} \quad f(x) + g(Ax) + h(x)$$

Algorithm

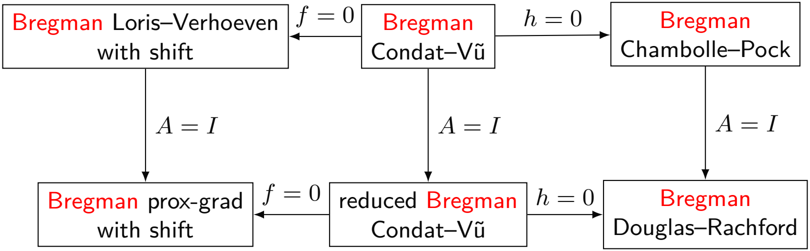
$$\begin{aligned} x^{(k+1)} &= \operatorname{prox}_{\tau f}^{\phi_{\mathrm{p}}}\big(x^{(k)}, \tau(A^T z^{(k)} + \nabla h(x^{(k)}))\big) \\ z^{(k+1)} &= \operatorname{prox}_{\sigma g^*}^{\phi_{\mathrm{d}}}\big(z^{(k)}, -\sigma A(2x^{(k+1)} - x^{(k)})\big) \end{aligned}$$

- ϕ_{p}, ϕ_{d} are two kernels, normalized to have strong convexity parameter 1
- step sizes must satisfy $\sigma\tau\|A\|^2 + \tau L \leq 1$
- Euclidean distance is usually used in dual space, especially when $g = \delta_{\{b\}}$
- when Euclidean distance is used in primal and dual spaces, it reduces to

$$\begin{aligned} x^{(k+1)} &= \operatorname{prox}_{\tau f}\big(x^{(k)} - \tau(A^T z^{(k)} + \nabla h(x^{(k)}))\big) \\ z^{(k+1)} &= \operatorname{prox}_{\sigma g^*}\big(z^{(k)} + \sigma A(2x^{(k+1)} - x^{(k)})\big) \end{aligned}$$

Condat (2013), Vũ (2013), Chambolle and Pock (2016)

Connections between Bregman proximal methods



- "completion" trick may not be applicable in Bregman case
- similar scheme also exists for Bregman PD3O
- it is still unclear how to extend PDDY to Bregman distance

Line search

Step sizes in Bregman proximal methods

Condat–Vũ	$\sigma\tau\|A\|^2 + \tau L \leq 1$
PD3O and PDDY	$\sigma\tau\|A\|^2 \leq 1, \;\; \tau L \leq 1$

- w.l.o.g., we normalize ϕ_{p} and ϕ_{d} to have the strong convexity parameter 1
- matrix norm is difficult to estimate or bound accurately

Bregman Condat–Vũ with line search

$$
\begin{aligned}
\bar{z}_{k+1} &= z_k + \theta_k(z_k - z_{k-1}) \\
x_{k+1} &= \operatorname{prox}_{\tau_k f}^{\phi_{\mathrm{p}}}\big(x_k, \tau_k(A^T \bar{z}_{k+1} + \nabla h(x_k))\big) \\
z_{k+1} &= \operatorname{prox}_{\sigma_k g^*}^{\phi_{\mathrm{d}}}\big(z_k, -\sigma_k(2x_{k+1} - x_k)\big)
\end{aligned}
$$

parameters θ_k, τ_k, and σ_k are determined adaptively by line search

Malitsky and Pock (2018), Jiang and Vandenberghe (2020)

Outline

Proximal methods with generalized distances

Applications to sparse semidefinite programs (SDPs)
- Generalized proximal operator with log-barrier distance
- Numerical experiments

Semidefinite programs (SDPs)

$$\begin{array}{ll} \text{minimize} & \operatorname{tr}(CX) \\ \text{subject to} & \mathcal{A}(X) = b \\ & X \in \mathbf{S}^n_+ \end{array} \qquad\qquad \begin{array}{ll} \text{maximize} & \langle b, y \rangle \\ \text{subject to} & \mathcal{A}^*(y) + S = C \\ & S \in \mathbf{S}^n_+ \end{array}$$

$\mathcal{A}\colon \mathbf{S}^n \to \mathbf{R}^m$ is a linear mapping, and $\mathcal{A}^*$ is its adjoint

first-order proximal methods (ADMM, primal–dual hybrid gradient, ...)

- exploit structure in linear constraints is straightforward
- require eigenvalue decompositions for projections on PSD cones

large SDPs often have sparse coefficient matrices $C, A_1, \ldots, A_m$

- applications related to graphs, Euclidean distance geometry
- relaxations of nonconvex quadratic and polynomial optimization

Sparse semidefinite programs

$$\begin{array}{ll} \text{minimize} & \mathrm{tr}(CX) \\ \text{subject to} & \mathcal{A}(X) = b, \ X \in \mathbf{S}^n_+ \end{array} \qquad \begin{array}{ll} \text{maximize} & \langle b, y \rangle \\ \text{subject to} & \mathcal{A}^*(y) + S = C, \ S \in \mathbf{S}^n_+ \end{array}$$

- $C, A_1, \ldots, A_m$ are sparse with common sparsity pattern E
- without loss of generality, assume E is *chordal* (a filled Cholesky pattern)
- optimal X is typically dense, even for sparse coefficients

Equivalent conic linear program

$$\begin{array}{ll} \text{minimize} & \mathrm{tr}(CX) \\ \text{subject to} & \mathcal{A}(X) = b, \ X \in K \end{array} \qquad \begin{array}{ll} \text{maximize} & \langle b, y \rangle \\ \text{subject to} & \mathcal{A}^*(y) + S = C, \ S \in K^* \end{array}$$

- variable X is a sparse matrix with pattern E (notation: $\mathbf{S}^n_E$)
- primal cone is set of matrices in $\mathbf{S}^n_E$ with PSD completion: $K = \Pi_E(\mathbf{S}^n_+)$
- dual cone is the set of sparse PSD matrices in $\mathbf{S}^n_E$: $K^* = \mathbf{S}^n_+ \cap \mathbf{S}^n_E$

Fujisawa, Kojima, Nakata (1997)

Centering problem

Logarithmic barrier

- ϕ is the conjugate of log-det barrier $\phi_*(S) = -\log\det S$ for K^*:

$$\phi(X) = \sup_{S \in \operatorname{int} K^*} (-\operatorname{tr}(XS) + \log\det S)$$

- for chordal E: efficient algorithms for computing ϕ, ϕ_*, $\nabla\phi$, $\nabla\phi_*$, *etc.*
- cost is about the same as sparse Cholesky factorization with pattern E

Centering problem

$$\begin{array}{ll} \text{minimize} & \operatorname{tr}(CX) + \mu\phi(X) \\ \text{subject to} & \mathcal{A}(X) = b \\ & \operatorname{tr} X = 1 \end{array}$$

- solutions for $\mu > 0$ form the central path of the SDP
- optimal X is (μn)-suboptimal for the SDP
- can be solved by Bregman PDHG

Bregman proximal operator for the centering objective

- centering objective, restricted to $\operatorname{tr} X = 1$

$$f(X) = \operatorname{tr}(CX) + \mu\phi(X) + \delta_H(X), \qquad \text{where } H = \{X \mid \operatorname{tr} X = 1\}$$

- prox-operator $\widehat{X} = \operatorname{prox}_f^{\phi}(Y, A)$, using Bregman distance generated by ϕ

$$\begin{array}{ll} \text{minimize} & \operatorname{tr}(BX) + \phi(X) \\ \text{subject to} & \operatorname{tr} X = 1 \end{array}$$

  where $B \in \mathbf{S}^n_E$ depends on Y, A, C and μ
- use Newton's method to find unique solution $\hat{\lambda}$ of the nonlinear equation

$$\operatorname{tr}((B + \lambda I)^{-1}) = 1 \qquad (\text{with } B + \lambda I \succ 0)$$

- for chordal sparsity patterns E, efficient algorithms exist for computing

  $$g(\lambda) = \operatorname{tr}((B{+}\lambda I)^{-1}), \quad g'(\lambda) = -\operatorname{tr}((B{+}\lambda I)^{-2}), \quad \widehat{X} = \Pi_E((B{+}\hat{\lambda} I)^{-1})$$

  from sparse Cholesky factorization of $B + \lambda I$

complexity $\approx$ # Newton iterations $\times$ cost of sparse Cholesky factorization

Maximum-cut problem

$$\begin{array}{ll} \text{maximize} & \mathrm{tr}(LX) \\ \text{subject to} & \mathrm{diag}(X) = \mathbf{1},\ X \succeq 0 \end{array}$$

- compute approximate solution on central path (parameter $\mu = 0.001/n$)
- four problems from SDPLIB, four graphs from SuiteSparse collection

	n	PDHG iterations	time per Cholesky factorization	Newton steps per prox-evaluation	time per PDHG iteration
maxG51	1000	267	0.05	2.45	0.12
maxG32	2000	240	0.12	1.56	0.18
maxG55	5000	249	0.29	2.10	0.58
maxG60	7000	279	0.60	2.55	1.22
barth4	6019	346	0.42	3.57	1.55
tuma2	12992	375	0.48	4.36	1.89
biplane-9	21701	287	0.95	2.58	2.12
c-67	57975	378	0.76	3.58	3.56

SDP relaxation of graph partitioning

$$\begin{array}{ll} \text{minimize} & \mathrm{tr}(P^T L P X) \\ \text{subject to} & \mathrm{diag}(P X P^T) = \mathbf{1}, \quad X \succeq 0 \end{array}$$

- columns of P are sparse basis of $\{x \mid \mathbf{1}^T x = 0\}$
- Bregman PDHG for centering problem (parameter $\mu = 0.001/n$)
- four problems from SDPLIB, four graphs from SuiteSparse collection

	n	PDHG iterations	time per Cholesky factorization	Newton steps per prox-evaluation	time per PDHG iteration
gpp100	100	305	0.01	2.43	0.02
gpp124-1	124	392	0.01	2.00	0.02
gpp250-1	250	365	0.01	2.65	0.03
gpp500-1	500	394	0.02	3.01	0.07
delaunay_n10	1024	403	0.37	4.36	1.76
delaunay_n11	2048	420	0.48	4.70	2.54
delaunay_n12	4096	367	0.60	4.43	3.05
delaunay_n13	8192	375	1.02	4.42	4.98

Summary

Bregman primal–dual first-order methods for

$$\text{minimize} \quad f(x) + g(Ax) + h(x)$$

- main steps are matrix–vector products with A, A^T, $\mathrm{prox}_f^{\phi_\mathrm{p}}$, and $\mathrm{prox}_{g^*}^{\phi_\mathrm{d}}$
- algorithm parameters are fixed or determined by line search

Applications to centering problem in sparse SDP

- distance generated by logarithmic barrier
- new, efficient algorithm for prox-operator of centering objective
- cost of prox-evaluation is comparable to sparse Cholesky factorization