Bregman primal–dual first-order method and application to sparse semidefinite optimization

Xin Jiang (UCLA)

joint work with Lieven Vandenberghe

INFORMS Annual Meeting

November 7–13, 2020

Semidefinite programs (SDPs)

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

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

Interior-point methods

- general-purpose implementations for dense problems do not scale well
- each iteration involves computations with complexity m^3, m^2n^2, nm^3
- customization to exploit problem structure is difficult

Proximal splitting methods (ADMM, primal–dual hybrid gradient, ...)

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

Sparse semidefinite programs

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

Example: relaxation of maximum-cut problem

$$\begin{array}{ll} \text{maximize} & \mathrm{tr}(LX) \\ \text{subject to} & X_{ii} = 1, \quad i = 1, \ldots, m \\ & X \succeq 0 \end{array}$$

- complexity of general-purpose interior-point solver: $O(n^4)$ per iteration
- customized interior-point solver: $O(n^3)$ per iteration
- proximal method: $O(n^3)$ per iteration (projection on PSD cone)

Semidefinite programs (SDPs)

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

Interior-point methods

- general-purpose implementations for dense problems do not scale well
- customization to exploit problem structure is difficult

Proximal splitting methods (ADMM, primal–dual hybrid gradient, ...)

- exploiting 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

Outline

Proximal methods with generalized distances

Logarithmic barrier distance for sparse PSD completable matrices

Outline

Proximal methods with generalized distances

- Bregman proximal operator
- Bregman primal–dual hybrid gradient (PDHG) methods

Logarithmic barrier distance for sparse PSD completable matrices

Proximal mapping

Proximal mapping: for closed convex function f

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

Generalized proximal mapping

- use a generalized distance $d(x,y)$ instead of $(1/2)\|x-y\|_2^2$
- for example, in proximal gradient method of minimizing $f(x) + g(x)$:

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \left(f(x) + g(x_k) + \langle \nabla g(x_k), x - x_k\rangle + (1/\tau)d(x,y)\right)$$

Potential benefits

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

goals: 1 is to reduce number of iterations; 2 is to reduce cost per iteration

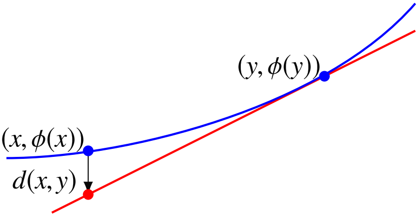
Bregman distance

Kernel function: ϕ convex, differentiable on its interior domain

Bregman distance (generalized distance)

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

with domain $\operatorname{dom} d = \operatorname{dom}\phi \times \operatorname{int}\operatorname{dom}\phi$



Bregman (1967), Censor and Zenios (1997)

Generalized proximal mapping

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

Requirements for minimizer x:

- existence in $\mathrm{int}(\mathrm{dom}\,\phi)$ and uniqueness for all $y \in \mathrm{int}(\mathrm{dom}\,\phi)$ and all a

Examples

- squared Euclidean distance: $\mathrm{prox}_f^d(y,a) = \mathrm{prox}_f(y-a)$
- f is indicator for $\{x \in \mathbf{R}_+^n \mid \mathbf{1}^T x = 1\}$ and d the relative entropy

$$\mathrm{prox}_f^d(y,a)_i = \frac{y_i e^{-a_i}}{\sum_{j=1}^n y_j e^{-a_j}}, \quad \text{for } i = 1,\ldots,n$$

used in entropic proximal point method, exponential method of multipliers

Primal–dual hybrid gradient (PDHG) method

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array}$$

f is a closed convex function

Algorithm

$$\begin{aligned} y_{k+1} &= z_k + \theta_k(z_k - z_{k-1}) \\ x_{k+1} &= \underset{x}{\operatorname{argmin}} \left(f(x) + y_{k+1}^T A x + \frac{1}{\tau_k} d(x, x_k)\right) \\ z_{k+1} &= z_k + \sigma_k(Ax_{k+1} - b) \end{aligned}$$

- x-update is evaluation of Bregman proximal operator
- parameters θ_k, τ_k, and σ_k are fixed or determined by line search
- Bregman variant of primal–dual hybrid gradient (Chambolle–Pock) method [Chambolle & Pock (2016)]

Outline

Proximal methods with generalized distances

Logarithmic barrier distance for sparse PSD completable matrices

- Generalized proximal operator with log-barrier distance
- Numerical experiments

Sparse semidefinite program

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

- $C, A_1, \ldots, A_m$ 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

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

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

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

- optimal $\hat{S}_X$ is (sparse) inverse of max-det PSD completion of X

$$\phi(X) = \log\det \hat{S}_X - n, \qquad \nabla\phi(X) = -\hat{S}_X$$

- for chordal E: efficient algorithms for computing $\hat{S}_X$ given X
- cost is about the same as sparse Cholesky factorization with pattern E

Centering problem

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

- solutions for $\mu > 0$ form the central path of the SDP
- optimal X is (μn)-suboptimal for the SDP

Bregman proximal operator for the centering problem

we formulate a Bregman proximal method for the centering problem

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

- centering objective, restricted to $\mathrm{tr}\, X = 1$ (alternatively, $\mathrm{tr}\, X \leq 1$)

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

- use Bregman distance generated by ϕ

$$\begin{aligned} \hat{X} = \mathrm{prox}_f^d(Y, D) &= \underset{X}{\mathrm{argmin}} \left(f(X) + \mathrm{tr}(DX) + (1/\tau)d(X, Y)\right) \\ &= \underset{\mathrm{tr}\, X = 1}{\mathrm{argmin}} \left(\mathrm{tr}(BX) + \phi(X)\right) \end{aligned}$$

where $B = (\tau(D + C) + \hat{S}_Y)/(1 + \mu\tau) \in \mathbf{S}_E^n$

Algorithm for Bregman proximal operator

$$\begin{array}{ll} \text{minimize} & \operatorname{tr}(BX) + \phi(X) \\ \text{subject to} & \operatorname{tr} X = 1 \end{array} \qquad \begin{array}{ll} \text{maximize} & \log\det(B + \lambda I) - \lambda \end{array}$$

- dual variable $\lambda \in \mathbf{R}$ is multiplier for $\operatorname{tr} X = 1$
- use Newton's method to find unique solution λ of the nonlinear equation

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

- from λ, compute solution $\hat{X}$ as projection $\Pi_E((B + \lambda I)^{-1})$ on $\mathbf{S}^n_E$
- for chordal sparsity patterns E, efficient algorithms exist for computing

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

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

complexity ≈ # Newton iterations × cost of sparse Cholesky factorization

Maximum-cut problem

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

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

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

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

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

Summary

Bregman primal–dual first-order method for

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array}$$

- main steps are matrix–vector products with A, A^T and $\mathrm{prox}_f^d(x, a)$
- 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 is comparable with cost of sparse Cholesky factorization

Conversion methods for sparse SDPs

Interior point methods for converted SDPs

- Schur complement systems may be easier to solve
- effective when all the maximal cliques are small

First-order methods for converted SDPs

- examples are DRS, ADMM, dual coordinate descent, *etc.*
- each step involves evaluation of prox-operator (or projection)
- bottleneck: eigenvalue decompositions for projections onto PSD cone

Drawbacks of conversion methods

- may require large number of consistency constraints
- constructing a feasible solution for original SDP is not trivial

Fukuda et al. (2001), Nakata et al. (2003), `SDPA`; Zheng et al. (2017, 2019), `CDCS`, *etc.*

Non-symmetric interior point methods

recall the non-symmetric formulation of sparse SDPs

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

- solve by primal, dual, non-symmetric primal-dual interior point methods
- require efficient evaluation of logarithmic barrier and its derivative
- bottleneck: solving Schur complement system

Burer (2003), Srijuntongsiri et al. (2004), Andersen et al. (2011), Zhang (2018); SMCP