Bregman proximal methods for semidefinite optimization

Lieven Vandenberghe

joint work with Xin Jiang and Hsiao-Han Chao

Department of Electrical and Computer Engineering, UCLA

One World Optimization Seminar

March 8, 2021

Semidefinite program (SDP)

$$\begin{array}{ll} \text{minimize} & \operatorname{tr}(CX) \\ \text{subject to} & \operatorname{tr}(A_i X) = b_i, \quad i = 1, \ldots, m \\ & X \succeq 0 \end{array}$$

X is a symmetric $n \times n$ matrix; $X \succeq 0$ means X is positive semidefinite

Interior-point methods

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

Proximal splitting methods

- exploiting structure in linear equality constraints is easier
- require eigenvalue decompositions for projections on positive semidefinite cone

Sparse semidefinite programs

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

- relaxations of combinatorial graph optimization problems
- semidefinite relaxations of polynomial optimization problems

Example: relaxation of maximum-cut problem

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

L is weighted graph Laplacian

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

Nonnegative trigonometric polynomials

$$F_x(\omega) = x_0 + \sum_{k=1}^{n} (x_k e^{-\mathrm{i}k\omega} + \bar{x}_k e^{\mathrm{i}k\omega}) \geq 0 \quad \text{for all } \omega \qquad (\mathrm{i} = \sqrt{-1})$$

- coefficients x form a semidefinite-representable convex cone K
- dual cone K^* is cone of positive semidefinite Toeplitz matrices

Applications

- source of many SDP applications in signal processing since 1990s
- recent applications to superresolution, grid-free compressed sensing
- SDP formulations extend to matrix polynomials, rational (Popov) functions, ...

Complexity: convex optimization over K or K^*

- general-purpose interior-point SDP solvers: $O(n^4)$ per iteration
- customized interior-point solvers: $O(n^3)$ per iteration
- proximal splitting methods: $O(n^3)$ per iteration (for projection on p.s.d. cone)

Outline

1. Proximal methods with generalized (Bregman) distances
2. Itakura–Saito distance for nonnegative trigonometric polynomials
3. Logarithmic barrier distance for sparse p.s.d. completable matrices

Proximal mapping

Proximal mapping: for closed convex function f

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

if f is the indicator of a closed convex set C, this is the Euclidean projection on C

Proximal algorithms

- proximal point method: $x_{k+1} = \mathrm{prox}_{\tau f}(x_k)$
- proximal gradient method for minimizing $f(x) + g(x)$, with g differentiable:

$$\begin{aligned} x_{k+1} &= \mathrm{prox}_{\tau f}(x_k - \tau \nabla g(x_k)) \\ &= \operatorname*{argmin}_x \left(f(x) + g(x_k) + \langle \nabla g(x_k), x - x_k \rangle + \frac{1}{2\tau}\|x - x_k\|_2^2 \right) \end{aligned}$$

- splitting methods: ADMM, Douglas–Rachford splitting, Spingarn's method
- primal–dual methods: primal–dual hybrid gradient (Chambolle–Pock) method

Proximal algorithms with generalized distances

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

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

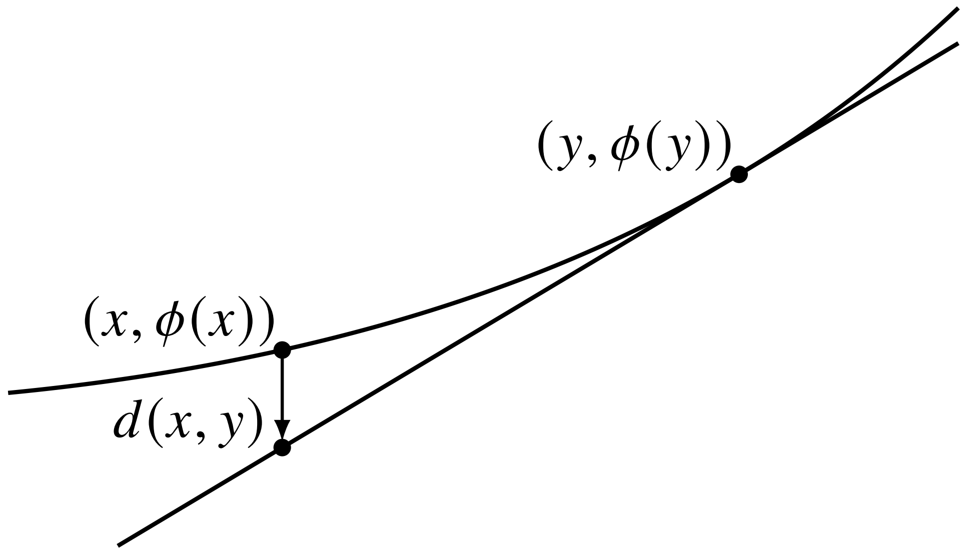
Potential benefits

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

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

Bregman distance

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



- ϕ is the *kernel function*, convex and continuously differentiable on $\mathrm{int}\,(\mathrm{dom}\,\phi)$
- we define the domain of d as $\mathrm{dom}\,d = \mathrm{dom}\,\phi \times \mathrm{int}(\mathrm{dom}\,\phi)$
- domain of ϕ may include its boundary or a subset of its boundary

other properties of ϕ may be required

[Censor and Zenios 1997]

Generalized proximal mapping

- proximal mapping of f for Bregman distance d

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

- for $d(x, y) = \frac{1}{2}\|x - y\|_2^2$, this is the standard proximal mapping

$$\begin{aligned} \text{prox}_f^d(y, a) &= \underset{x}{\text{argmin}} \left(f(x) + \langle a, x \rangle + \frac{1}{2}\|x - y\|_2^2 \right) \\ &= \underset{x}{\text{argmin}} \left(f(x) + \frac{1}{2}\|x - y + a\|_2^2 \right) \\ &= \text{prox}_f(y - a) \end{aligned}$$

Requirements

- minimizer x exists and is unique for all $y \in \text{int}\,(\text{dom}\,\phi)$ and all a
- minimizer x is in interior of $\text{dom}\,\phi$
- minimizer is inexpensive to compute

Example: relative entropy

$$d(x,y) = \sum_{i=1}^{n} (x_i \log(x_i/y_i) - x_i + y_i), \qquad \mathbf{dom}\, d = \mathbf{R}_+^n \times \mathbf{R}_{++}^n$$

- the Bregman distance for

$$\phi(x) = \sum_{i=1}^{n} x_i \log x_i, \qquad \mathbf{dom}\, \phi = \mathbf{R}_+^n$$

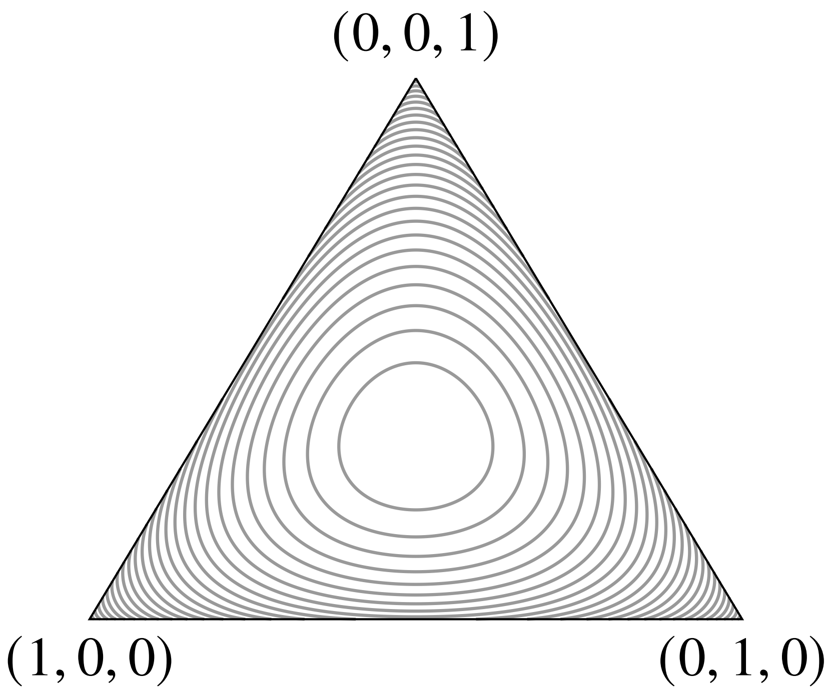
- generalized projection (proximal operator for indicator) on $H = \{x \mid \mathbf{1}^T x = 1\}$

$$\operatorname*{argmin}_{\mathbf{1}^T x = 1} \left(a^T x + d(x,y)\right) = \frac{1}{\sum_{j=1}^{n} y_j e^{-a_j}} \begin{bmatrix} y_1 e^{-a_1} \\ \vdots \\ y_n e^{-a_n} \end{bmatrix}$$

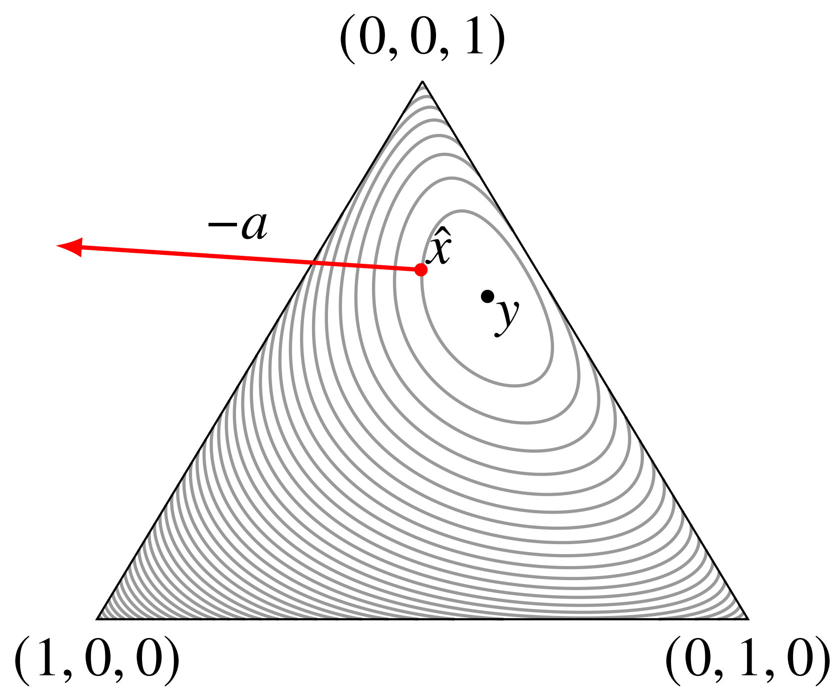
used in entropic proximal point method, exponential method of multipliers

Example: relative entropy

Contour lines of $\phi(x)$



Contour lines of $d(x, y)$



$$\hat{x} = \underset{\mathbf{1}^T x = 1}{\operatorname{argmin}} \left(a^T x + d(x, y)\right)$$

Outline

1. Proximal methods with generalized (Bregman) distances
2. **Itakura–Saito distance for nonnegative trigonometric polynomials**
3. Logarithmic barrier distance for sparse p.s.d. completable matrices

Cone of nonnegative trigonometric polynomials

- F_x is a trigonometric polynomial with coefficients x_k (real for simplicity)

$$F_x(\omega) = x_0 + 2x_1 \cos \omega + \cdots + 2x_n \cos n\omega$$

- K is the convex cone

$$K = \{x \in \mathbf{R}^{n+1} \mid F_x(\omega) \geq 0 \ \forall \omega\}$$

we consider optimization problems that include constraints

$$x \in K, \quad x_0 = 1$$

equality $x_0 = 1$ normalizes F_x:

$$\frac{1}{2\pi} \int_0^{2\pi} F_x(\omega) d\omega = 1$$

Semidefinite representation of K and dual cone K^*

$$K = \{D(X) \mid X \in \mathbf{S}^{n+1},\ X \succeq 0\}, \qquad K^* = \{y \in \mathbf{R}^{n+1} \mid T(y) \succeq 0\}$$

- $D : \mathbf{S}^{n+1} \to \mathbf{R}^{n+1}$ maps symmetric matrix X to vector of diagonal sums

$$D(X) = \begin{bmatrix} X_{00} + X_{11} + \cdots + X_{nn} \\ X_{01} + X_{12} + \cdots + X_{n-1,n} \\ \vdots \\ X_{0,n-1} + X_{1n} \\ X_{0n} \end{bmatrix}$$

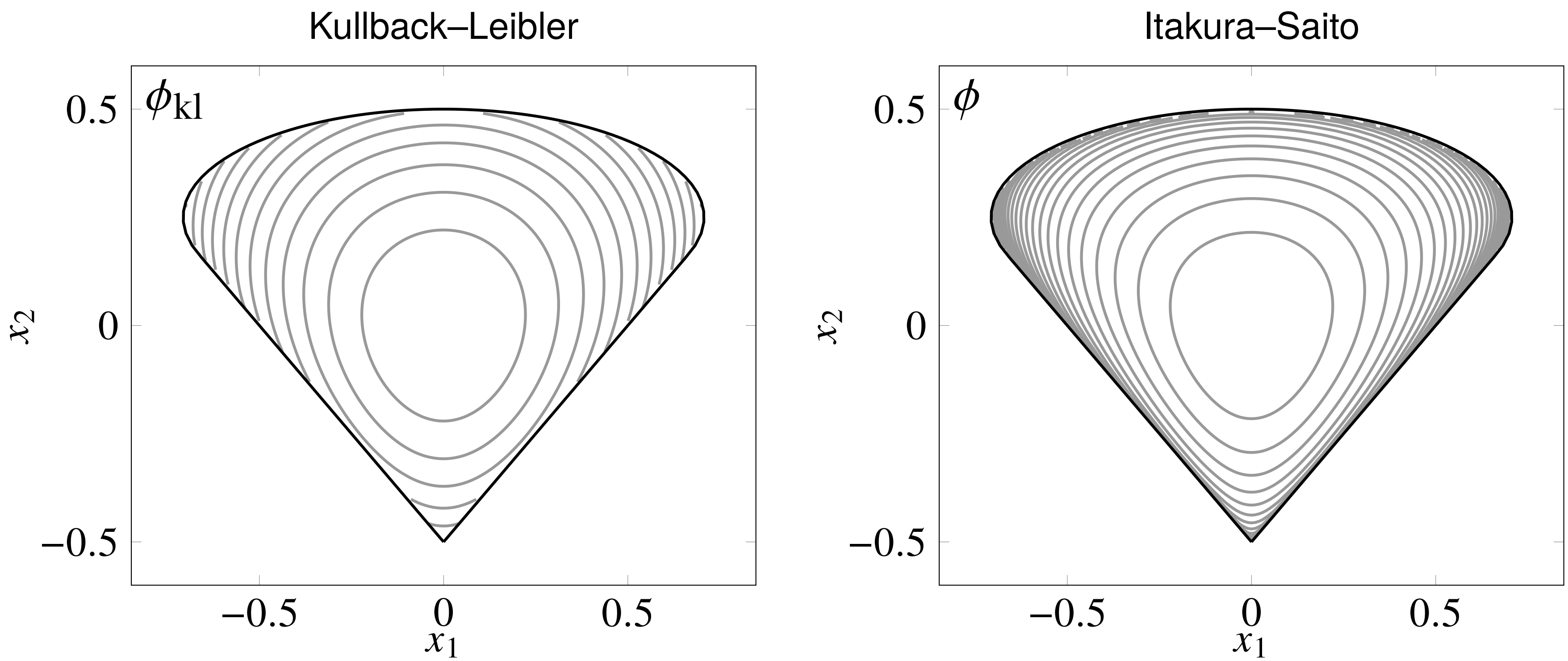
- $T : \mathbf{R}^{n+1} \to \mathbf{S}^{n+1}$ maps vector $(y_0, \ldots, y_n)$ to the symmetric Toeplitz matrix

$$T(y) = \begin{bmatrix} y_0 & y_1 & \cdots & y_{n-1} & y_n \\ y_1 & y_0 & \cdots & y_{n-2} & y_{n-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_{n-1} & y_{n-2} & \cdots & y_0 & y_1 \\ y_n & y_{n-1} & \cdots & y_1 & y_0 \end{bmatrix}$$

Kernel functions

kernels for Kullback–Leibler distance and Itakura–Saito distance

$$\phi_{\mathrm{kl}}(x) = \frac{1}{2\pi}\int_0^{2\pi} F_x(\omega)\log F_x(\omega)\,d\omega \qquad\qquad \phi(x) = -\frac{1}{2\pi}\int_0^{2\pi} \log F_x(\omega)\,d\omega$$



- plots show contour lines on section $\{x \in K \mid x_0 = 1\}$
- ϕ is *essentially smooth*; ϕ_{kl} is not

Semidefinite representation of entropy kernel ϕ

$$\begin{array}{ll} \text{minimize (over } X) & -\log X_{00} \\ \text{subject to} & D(X) = x \\ & X \succeq 0 \end{array}$$

- for $x \in K \setminus \{0\}$, optimal value is

$$\phi(x) = -\frac{1}{2\pi} \int_0^{2\pi} \log F_x(\omega) d\omega$$

- optimal X has rank one:

$$X = bb^T, \qquad \phi(x) = -2 \log b_0$$

- b is minimum-phase spectral factor ($b_0 + b_1 z^{-1} + \cdots + b_n z^{-n} \neq 0$ for $|z| > 1$)
- b is efficiently computed by spectral factorization of x: solve quadratic equation

$$D(bb^T) = x$$

Dual of semidefinite representation of ϕ

$$\text{maximize (over } y) \quad -\psi(y) - \langle x, y \rangle + 1$$

- convex function ψ is defined as

$$\psi(y) = \log(e^T T(y)^{-1} e), \qquad \mathbf{dom}\,\psi = \{y \mid T(y) \succ 0\}$$

  where $e = (1, 0, \ldots, 0)$

- by duality, optimal value is $\phi(x)$
- optimal y is $y = -\nabla\phi(x)$, and related to primal solution $X = bb^T$ as

$$T(y)b = e$$

  y can be computed from spectral factor b by reverse Levinson algorithm

Itakura–Saito distance and projection

$$d(x,y) = \frac{1}{2\pi}\int_0^{2\pi} \left(\frac{F_x(\omega)}{F_y(\omega)} - \log\frac{F_x(\omega)}{F_y(\omega)} - 1\right) d\omega$$

- proposed in 1970s as spectral distance measure in speech processing
- generalized projection on hyperplane $H = \{x \mid x_0 = 1\}$:

$$\begin{aligned}
\mathrm{prox}^{d}_{\delta_H}(y,a) &= \operatorname*{argmin}_{x_0=1} \left(\langle a, x\rangle + d(x,y)\right) \\
&= \operatorname*{argmin}_{x_0=1} \left(\langle c, x\rangle + \phi(x)\right) \qquad \text{(where } c = a - \nabla\phi(y)\text{)}
\end{aligned}$$

- dual problem (scalar variable λ is multiplier for constraint $x_0 = 1$)

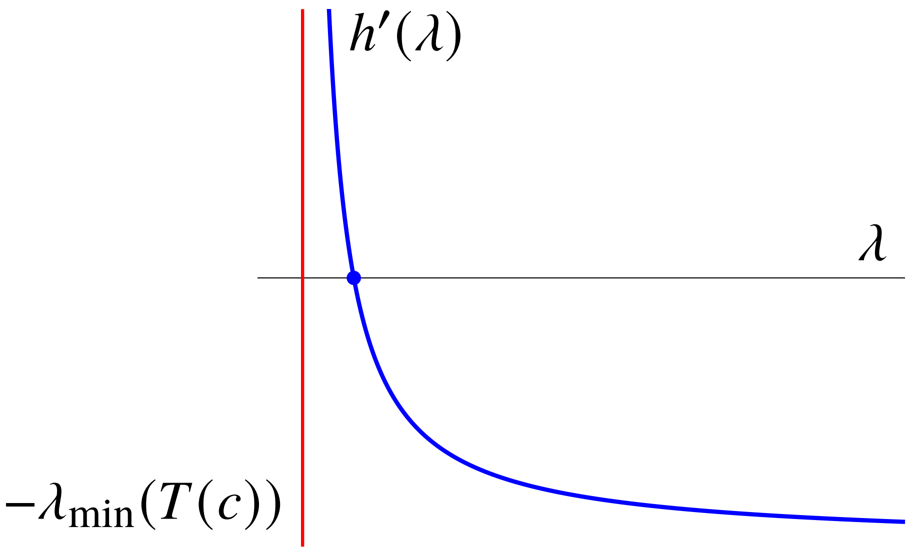
$$\text{maximize} \quad -\log\left(e^T (T(c) + \lambda I)^{-1} e\right) - \lambda$$

$e^T (T(c) + \lambda I)^{-1} e$ is the 1st element of the inverse of Toeplitz matrix $T(c) + \lambda I$

Computing Itakura–Saito projection

solve dual problem for λ, for example, by Newton's method

$$\text{maximize} \quad h(\lambda) = -\log(e^T(T(c)+\lambda I)^{-1}e) - \lambda$$

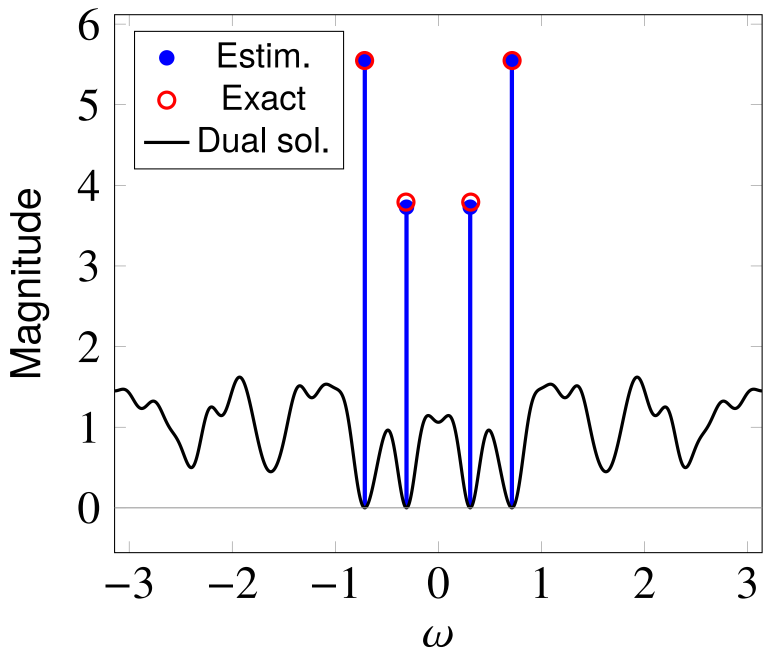


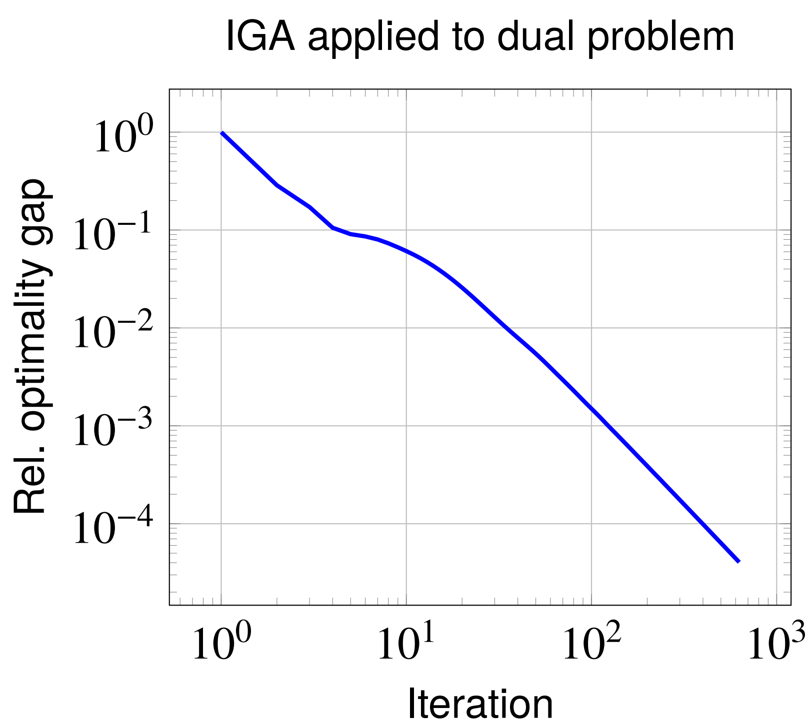
- at each Newton step, factorize positive definite Toeplitz matrix $T(c)+\lambda I$
- complexity: $O(n^2)$ with Levinson algorithm, $O(n(\log n)^2)$ with superfast solvers
- from optimal λ, compute solution $x = (1/b_0)D(bb^T)$ where $b = (T(c)+\lambda I)^{-1}e$

Covariance estimation

$$\begin{array}{ll} \text{minimize (over } y, s) & \|T(y) + sI - R\|_F^2 + \gamma \operatorname{tr}(T(y)) \\ \text{subject to} & T(y) \succeq 0 \end{array}$$

- estimate parameters in signal model $v(t) = \sum_{k=1}^{\rho} c_k e^{\mathrm{i}\omega_k t}$ + white noise
- fit covariance $T(y) + sI$: low-rank p.s.d. Toeplitz plus multiple of identity
- R is sample covariance matrix ($n + 1 = 30$ in the example)





IGA: proximal gradient algorithm with Bregman distances

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C \end{array}$$

C a convex set; f convex with Lipschitz continuous gradient

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|$$

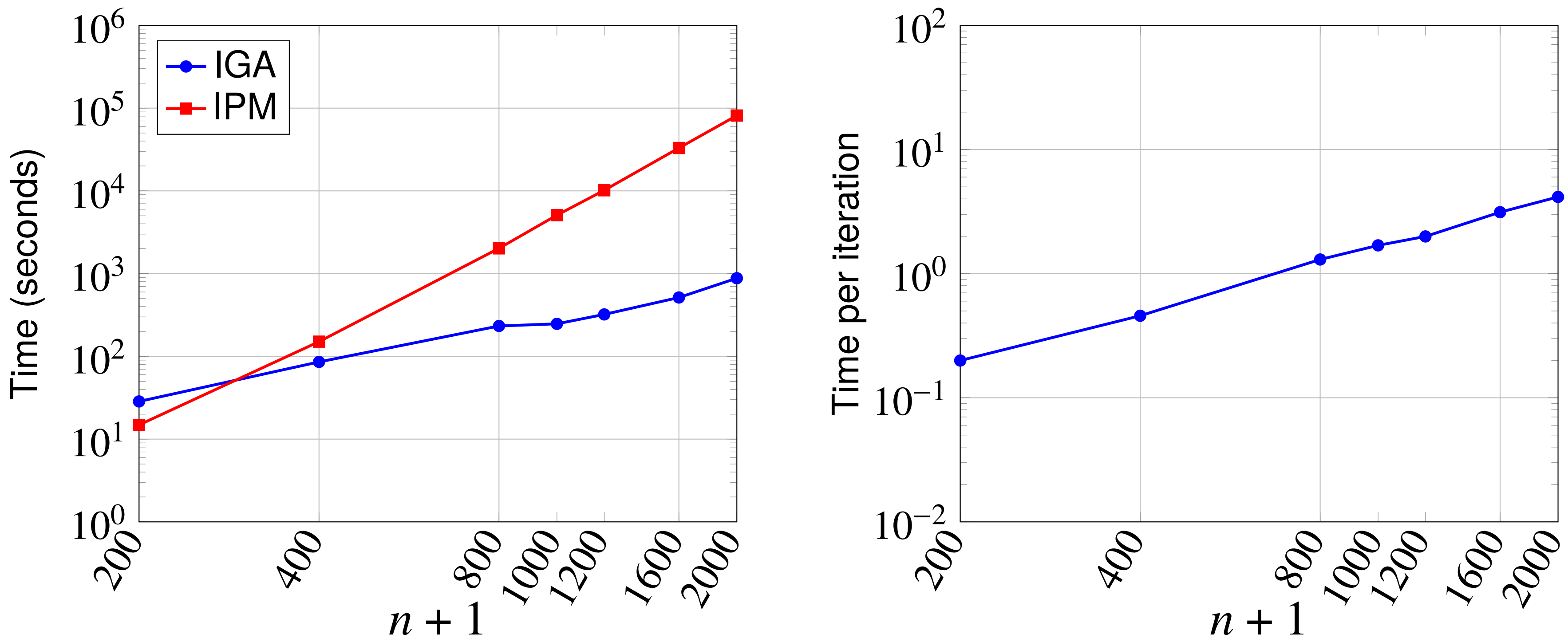
Improved gradient algorithm (IGA) [Auslender and Teboulle 2006]

$$\begin{aligned} y_{k+1} &= (1-\theta_k)x_k + \theta_k v_k \\ v_{k+1} &= \operatorname*{argmin}_{x \in C} \left(\langle \nabla f(y_{k+1}), x \rangle + \frac{1}{\tau_k} d(x, v_k) \right) \\ x_{k+1} &= (1-\theta_k)x_k + \theta_k v_{k+1} \end{aligned}$$

- Bregman extension version of Nesterov fast gradient projection method
- we assume Bregman kernel is strongly convex: $d(x, y) \geq \frac{1}{2}\|x - y\|^2$
- θ_k, τ_k determined by line search; does not require knowledge of L

Euclidean projection

$$\begin{array}{ll} \text{minimize} & \sum_{k=0}^{n} (x_k - a_k)^2 \\ \text{subject to} & x \in K, \quad x_0 = 1 \end{array}$$



- IPM is SDPT3/SeDuMi via CVX; IGA is Auslender–Teboulle algorithm
- number of IGA iterations is 100–200 to reach relative accuracy 10^{-4}
- about 10 Newton steps per projection; Toeplitz solver is Levinson algorithm

Outline

1. Proximal methods with generalized (Bregman) distances
2. Itakura–Saito distance for nonnegative trigonometric polynomials
3. **Logarithmic barrier distance for sparse p.s.d. completable matrices**

Sparse semidefinite program

$$\begin{array}{ll} \text{minimize} & \mathrm{tr}(CX) \\ \text{subject to} & \mathrm{tr}(A_i X) = b_i, \quad i = 1, \ldots, m \\ & X \succeq 0 \end{array}$$

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

Equivalent conic linear program

$$\begin{array}{ll} \text{minimize} & \mathrm{tr}(CX) \\ \text{subject to} & \mathrm{tr}(A_i X) = b_i, \quad i = 1, \ldots, m \\ & X \in K \end{array}$$

- variable X is a *sparse* matrix with sparsity pattern E (notation: $\mathbf{S}^n_E$)
- K is cone of matrices in $\mathbf{S}^n_E$ that have a positive semidefinite completion

Centering problem

Logarithmic barrier

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

- dual cone K^* is cone of positive semidefinite matrices in $\mathbf{S}^n_E$
- ϕ is conjugate barrier of log-det barrier $\phi_*(S) = -\log\det S$ for K^*

Centering problem

$$\begin{array}{ll} \text{minimize} & \operatorname{tr}(CX) + \mu\phi(X) \\ \text{subject to} & \operatorname{tr}(A_i X) = b_i, \quad i = 1, \ldots, m \end{array}$$

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

Bregman distance generated by barrier kernel

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

- optimal $\hat{S}_X$ is inverse of maximum determinant pos. definite completion of X

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

- gradient $\nabla\phi(X) = -\hat{S}_X$
- for chordal E: efficient algorithms for computing $\hat{S}_X$ given X
- complexity is comparable with sparse Cholesky factorization with pattern E

Distance

$$\begin{aligned} d(X,Y) &= \phi(X) - \phi(Y) - \operatorname{tr}\left(\nabla\phi(Y)(X-Y)\right) \\ &= -\log\det(\hat{S}_Y \hat{S}_X^{-1}) + \operatorname{tr}(\hat{S}_Y \hat{S}_X^{-1}) + n \end{aligned}$$

the relative entropy (Kullback–Leibler divergence) between $\hat{S}_Y$ and $\hat{S}_X$

Bregman proximal operator for centering problem

$$\begin{array}{ll} \text{minimize} & \operatorname{tr}(CX) + \mu\phi(X) \\ \text{subject to} & \operatorname{tr}(A_iX) = b_i, \quad i = 1, \ldots, m \\ & \operatorname{tr} X = 1 \end{array}$$

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

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

- Bregman proximal operator $\operatorname{prox}^d_{\tau f}(Y, \tau D)$ for centering objective

$$\begin{aligned} \hat{X} &= \operatorname*{argmin}_X \left(f(X) + \operatorname{tr}(DX) + \frac{1}{\tau} d(X, Y) \right) \\ &= \operatorname*{argmin}_{\operatorname{tr} X = 1} \left(\operatorname{tr}(BX) + \phi(X) \right) \qquad \text{where } B = \frac{1}{1 + \mu\tau}(\tau(D + C) + \hat{S}_Y) \in \mathbf{S}^n_E \end{aligned}$$

- dual problem (scalar variable λ is multiplier for $\operatorname{tr} X = 1$):

$$\text{maximize} \quad \log\det(B + \lambda I) - \lambda$$

Algorithm for Bregman proximal operator

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

- use Newton's method to find unique solution λ of the nonlinear equation

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

- 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

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

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

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

Maximum-cut problem

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

- compute approximate solution on central path (parameter $\mu = 0.001/n$)
- Bregman variant of primal–dual hybrid gradient algorithm [Chambolle & Pock 2016]
- four problems from SDPLIB, four graphs from SuiteSparse matrix 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 (centering 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

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

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

- Bregman variant of primal–dual hybrid gradient (Chambolle–Pock) method [Chambolle & Pock 2016]
- parameters θ_k, σ_k, τ_k can be determined by line search
- does not require knowledge of norm of A or strong convexity constant of ϕ

Summary

Bregman proximal methods for two classes of SDP-representable constraints

Nonnegative trigonometric polynomials

- Itakura–Saito distance
- cost of generalized projection is roughly $O(n^2)$

Positive semidefinite completable sparse matrices

- distance generated by logarithmic barrier
- prox-operator for centering objective
- cost roughly on the same order as sparse Cholesky factorization

References

1. H.-H. Chao, L. Vandenberghe, IEEE Trans. Signal Processing, 2018.
2. X. Jiang, L. Vandenberghe, submitted, 2021.