Advances in Convex Optimization: Theory, Algorithms, and Applications

Stephen Boyd
Electrical Engineering Department
Stanford University

(joint work with Lieven Vandenberghe, UCLA)

ISIT 02

ISIT 02 Lausanne 7/3/02
Two problems

polytope $\mathcal{P}$ described by linear inequalities, $a_i^T x \leq b_i, \; i = 1, \ldots, L$

Problem 1a: find minimum volume ellipsoid $\supseteq \mathcal{P}$

Problem 1b: find maximum volume ellipsoid $\subseteq \mathcal{P}$

are these (computationally) difficult? or easy?
problem 1a is **very difficult**

- in practice
- in theory (NP-hard)

problem 1b is **very easy**

- in practice (readily solved on small computer)
- in theory (polynomial complexity)
Two more problems

find capacity of discrete memoryless channel, subject to constraints on input distribution

**Problem 2a:** find channel capacity, subject to:
no more than 30% of the probability is concentrated on any 10% of the input symbols

**Problem 2b:** find channel capacity, subject to:
at least 30% of the probability is concentrated on 10% of the input symbols

are problems 2a and 2b (computationally) difficult? or easy?
problem 2a is very easy in practice & theory

problem 2b is very difficult\textsuperscript{1}

\textsuperscript{1}I’m almost sure
Moral

very difficult and very easy problems can look quite similar

... unless you’re trained to recognize the difference
Outline

• what’s new in convex optimization

• some new standard problem classes

• generalized inequalities and semidefinite programming

• interior-point algorithms and complexity analysis
Convex optimization problems

minimize \( f_0(x) \)
subject to \( f_1(x) \leq 0, \ldots, f_L(x) \leq 0, \quad Ax = b \)

- \( x \in \mathbb{R}^n \) is optimization variable
- \( f_i : \mathbb{R}^n \to \mathbb{R} \) are convex, i.e., for all \( x, y, 0 \leq \lambda \leq 1 \),
  \[
  f_i(\lambda x + (1 - \lambda)y) \leq \lambda f_i(x) + (1 - \lambda)f_i(y)
  \]

examples:
- linear & (convex) quadratic programs
- problem 1b & 2a (if formulated properly)
Convex analysis & optimization

nice properties of convex optimization problems known since 1960s

- local solutions are global
- duality theory, optimality conditions
- simple solution methods like alternating projections

convex analysis well developed by 1970s (Rockafellar)

- separating & supporting hyperplanes
- subgradient calculus
What’s new (since 1990 or so)

• powerful primal-dual interior-point methods
  
extremely efficient, handle nonlinear large scale problems

• polynomial-time complexity results for interior-point methods
  
based on self-concordance analysis of Newton’s method

• extension to generalized inequalities
  
semidefinite & maxdet programming

• new standard problem classes
  
generalizations of LP, with theory, algorithms, software

• lots of applications
  
control, combinatorial optimization, signal processing,
  circuit design, . . .
Recent history

- (1984–97) interior-point methods for LP
  - (1984) Karmarkar’s interior-point LP method
  - theory (Ye, Renegar, Kojima, Todd, Monteiro, Roos, . . . )
  - practice (Wright, Mehrotra, Vanderbei, Shanno, Lustig, . . . )
- (1989–) semidefinite programming in control
  (Boyd, El Ghaoui, Balakrishnan, Feron, Scherer, . . . )
- (1990–) semidefinite programming in combinatorial optimization
  (Alizadeh, Goemans, Williamson, Lovasz & Schrijver, Parrilo, . . . )
- (1994) interior-point methods for nonlinear convex problems
  (Nesterov & Nemirovsky, Overton, Todd, Ye, Sturm, . . . )
- (1997–) robust optimization (Ben Tal, Nemirovsky, El Ghaoui, . . . )
Some new standard (convex) problem classes

- second-order cone programming (SOCP)
- semidefinite programming (SDP), maxdet programming
- (convex form) geometric programming (GP)

for these new problem classes we have

- complete duality theory, similar to LP
- good algorithms, and robust, reliable software
- wide variety of new applications
Second-order cone programming

**second-order cone program** (SOCP) has form

\[
\begin{align*}
\text{minimize} & \quad c_0^T x \\
\text{subject to} & \quad \|A_i x + b_i\|_2 \leq c_i^T x + d_i, \quad i = 1, \ldots, m \\
& \quad F x = g
\end{align*}
\]

- variable is \( x \in \mathbb{R}^n \)
- includes LP as special case \((A_i = 0, b_i = 0)\), QP \((c_i = 0)\)
- nondifferentiable when \( A_i x + b_i = 0 \)
- new IP methods can solve (almost) as fast as LPs
Robust linear programming

robust linear program:

\[
\begin{align*}
\text{minimize} & \quad c^T x \\
\text{subject to} & \quad a_i^T x \leq b_i \quad \text{for all} \quad a_i \in \mathcal{E}_i
\end{align*}
\]

- **ellipsoid** \( \mathcal{E}_i = \{ \bar{a}_i + F_i p \mid \|p\|_2 \leq 1 \} \) describes **uncertainty** in constraint vectors \( a_i \)
- \( x \) must satisfy constraints for all possible values of \( a_i \)
- can extend to uncertain \( c \) & \( b_i \), correlated uncertainties . . .
Robust LP as SOCP

robust LP is

\[
\begin{align*}
\text{minimize} & \quad c^T x \\
\text{subject to} & \quad \bar{a}_i^T x + \sup \{(F_ip)^T x \mid \|p\|_2 \leq 1\} \leq b_i
\end{align*}
\]

which is the same as

\[
\begin{align*}
\text{minimize} & \quad c^T x \\
\text{subject to} & \quad \bar{a}_i^T x + \|F_i^T x\|_2 \leq b_i
\end{align*}
\]

- an SOCP (hence, readily solved)
- term \(\|F_i^T x\|_2\) is extra margin required to accommodate uncertainty in \(a_i\)
Stochastic robust linear programming

\[
\begin{align*}
\text{minimize} & \quad c^T x \\
\text{subject to} & \quad \text{Prob}(a_i^T x \leq b_i) \geq \eta, \quad i = 1, \ldots, m
\end{align*}
\]

where \( a_i \sim \mathcal{N}(\bar{a}_i, \Sigma_i) \), \( \eta \geq 1/2 \) (\( c \) and \( b_i \) are fixed)

\( i.e. \), each constraint must hold with probability at least \( \eta \)

equivalent to SOCP

\[
\begin{align*}
\text{minimize} & \quad c^T x \\
\text{subject to} & \quad \bar{a}_i^T x + \Phi^{-1}(\eta) \| \Sigma_i^{1/2} x \|_2 \leq 1, \quad i = 1, \ldots, m
\end{align*}
\]

where \( \Phi \) is CDF of \( \mathcal{N}(0, 1) \) random variable
Geometric programming

**log-sum-exp** function:

\[ \text{lse}(x) = \log (e^{x_1} + \cdots + e^{x_n}) \]

...a smooth **convex** approximation of the max function

**geometric program** (GP), with variable \(x \in \mathbb{R}^n\):

\[
\begin{align*}
\text{minimize} & \quad \text{lse}(A_0 x + b_0) \\
\text{subject to} & \quad \text{lse}(A_i x + b_i) \leq 0, \quad i = 1, \ldots, m
\end{align*}
\]

where \(A_i \in \mathbb{R}^{m_i \times n}, b_i \in \mathbb{R}^{m_i}\)

new IP methods can solve large scale GPs (almost) as fast as LPs
Dual geometric program

dual of geometric program is an **unnormalized entropy problem**

maximize \[ \sum_{i=0}^{m} (b_i \nu_i + \text{entr}(\nu_i)) \]
subject to \[ \nu_i \geq 0, \quad i = 0, \ldots, m, \quad 1^T \nu_0 = 1, \]
\[ \sum_{i=0}^{m} A_i^T \nu_i = 0 \]

- dual variables are \( \nu_i \in \mathbb{R}^{m_i}, i = 0, \ldots, m \)
- (unnormalized) entropy is
\[
\text{entr}(\nu) = - \sum_{i=1}^{n} \nu_i \log \frac{\nu_i}{1^T \nu}
\]

- GP is closely related to problems involving entropy, KL divergence
Example: DMC capacity problem

\( x \in \mathbb{R}^n \) is distribution of input; \( y \in \mathbb{R}^m \) is distribution of output
\( P \in \mathbb{R}^{m \times n} \) gives conditional probabilities: \( y = Px \)

**primal channel capacity problem:**

\[
\begin{align*}
\text{maximize} & \quad -c^T x + \text{entr}(y) \\
\text{subject to} & \quad x \succeq 0, \quad 1^T x = 1, \quad y = Px
\end{align*}
\]

where \( c_j = -\sum_{i=1}^m p_{ij} \log p_{ij} \)

**dual channel capacity problem** is a simple GP:

\[
\begin{align*}
\text{minimize} & \quad \text{lse}(u) \\
\text{subject to} & \quad c + P^T u \succeq 0
\end{align*}
\]
Generalized inequalities

with proper convex cone $K \subseteq \mathbb{R}^k$ we associate generalized inequality

$$x \leq_K y \iff y - x \in K$$

convex optimization problem with generalized inequalities:

$$\begin{align*}
\text{minimize} & \quad f_0(x) \\
\text{subject to} & \quad f_1(x) \leq_{K_1} 0, \ldots, f_L(x) \leq_{K_L} 0, \quad Ax = b
\end{align*}$$

$f_i : \mathbb{R}^n \to \mathbb{R}^{k_i}$ are $K_i$-convex: for all $x, y, 0 \leq \lambda \leq 1$,

$$f_i(\lambda x + (1 - \lambda)y) \leq_{K_i} \lambda f_i(x) + (1 - \lambda)f_i(y)$$
Semidefinite program

semidefinite program (SDP):

\[
\begin{align*}
\text{minimize} & \quad c^T x \\
\text{subject to} & \quad A_0 + x_1 A_1 + \cdots + x_n A_n \preceq 0, \quad Cx = d
\end{align*}
\]

- \( A_i = A_i^T \in \mathbb{R}^{m \times m} \)
- inequality is matrix inequality, \( i.e., K \) is positive semidefinite cone
- single constraint, which is affine (hence, matrix convex)
Maxdet problem

extension of SDP: maxdet problem

\[
\begin{align*}
\text{minimize} & \quad c^T x - \log \det_+(G_0 + x_1 G_1 + \cdots + x_m G_m) \\
\text{subject to} & \quad A_0 + x_1 A_1 + \cdots + x_n A_n \preceq 0, \quad Cx = d
\end{align*}
\]

- \( x \in \mathbb{R}^n \) is variable

- \( A_i = A_i^T \in \mathbb{R}^{m \times m}, \ G_i = G_i^T \in \mathbb{R}^{p \times p} \)

- \( \det_+(Z) = \begin{cases} 
\det Z & \text{if } Z \succ 0 \\
0 & \text{otherwise}
\end{cases} \)
Semidefinite & maxdet programming

- nearly complete duality theory, similar to LP

- interior-point algorithms that are efficient in theory & practice

- applications in many areas:
  - control theory
  - combinatorial optimization & graph theory
  - structural optimization
  - statistics
  - signal processing
  - circuit design
  - geometrical problems
  - algebraic geometry
Chebyshev bounds

generalized Chebyshev inequalities: lower bounds on

\[ \text{Prob}(X \in C) \]

- \( X \in \mathbb{R}^n \) is a random variable with \( \mathbb{E} X = a, \mathbb{E} X X^T = S \)
- \( C \) is an open polyhedron \( C = \{ x \mid a_i^T x < b_i, \ i = 1, \ldots, m \} \)

cf. classical Chebyshev inequality on \( \mathbb{R} \)

\[ \text{Prob}(X < 1) \geq \frac{1}{1 + \sigma^2} \]

if \( \mathbb{E} X = 0, \mathbb{E} X^2 = \sigma^2 \)
Chebyshev bounds via SDP

minimize \[ 1 - \sum_{i=1}^{m} \lambda_i \]

subject to \[ a_i^T z_i \geq b_i \lambda_i, \quad i = 1, \ldots, m \]

\[ \sum_{i=1}^{m} \begin{bmatrix} Z_i & z_i \\ z_i^T & \lambda_i \end{bmatrix} \preceq \begin{bmatrix} S & a \\ a^T & 1 \end{bmatrix} \]

\[ \begin{bmatrix} Z_i & z_i \\ z_i^T & \lambda_i \end{bmatrix} \succeq 0, \quad i = 1, \ldots, m \]

- an SDP with variables \( Z_i = Z_i^T \in \mathbb{R}^{n \times n}, \ z_i \in \mathbb{R}^n, \) and \( \lambda_i \in \mathbb{R} \)
- optimal value is a (sharp) lower bound on \( \text{Prob}(X \in C) \)
- can construct a distribution with \( \mathbf{E} X = a, \ \mathbf{E} X X^T = S \) that attains the lower bound
Detection example

\[ x = s + v \]

- \( x \in \mathbb{R}^n \): received signal
- \( s \): transmitted signal \( s \in \{s_1, s_2, \ldots, s_N\} \) (one of \( N \) possible symbols)
- \( v \): noise with \( \mathbb{E} v = 0, \mathbb{E} vv^T = I \) (but otherwise unknown distribution)

**Detection problem**: given observed value of \( x \), estimate \( s \)
example \((n = 2, \ N = 7)\)

- detector selects symbol \(s_k\) closest to received signal \(x\)
- correct detection if \(s_k + v\) lies in the Voronoi region around \(s_k\)
**example:** bound on probability of correct detection of $s_1$ is 0.205

solid circles: distribution with probability of correct detection 0.205
Boolean least-squares

$x \in \{-1, 1\}^n$ is transmitted; we receive $y = Ax + v$, $v \sim \mathcal{N}(0, I)$

ML estimate of $x$ found by solving boolean least-squares problem

\[
\begin{align*}
\text{minimize} & \quad \|Ax - y\|^2 \\
\text{subject to} & \quad x_i^2 = 1, \quad i = 1, \ldots, n
\end{align*}
\]

- could check all $2^n$ possible values of $x$ . . .
- an NP-hard problem
- many heuristics for approximate solution
Boolean least-squares as matrix problem

\[ \|Ax - y\|^2 = x^T A^T Ax - 2y^T A^T x + y^T y \]
\[ = \text{Tr} A^T AX - 2y^T A^T x + y^T y \]

where \( X = xx^T \)

hence can express BLS as

\[
\begin{align*}
\text{minimize} \quad & \text{Tr} A^T AX - 2y^T A^T x + y^T y \\
\text{subject to} \quad & X_{ii} = 1, \quad X \succeq xx^T, \quad \text{rank}(X) = 1
\end{align*}
\]

\ldots still a very hard problem.
SDP relaxation for BLS

ignore rank one constraint, and use

$$X \succeq xx^T \iff \begin{bmatrix} X & x \\ x^T & 1 \end{bmatrix} \succeq 0$$

to obtain SDP relaxation (with variables $X, x$)

minimize $\text{Tr} A^T A X - 2y^T A^T x + y^T y$

subject to $X_{ii} = 1, \begin{bmatrix} X & x \\ x^T & 1 \end{bmatrix} \succeq 0$

- optimal value of SDP gives lower bound for BLS
- if optimal matrix is rank one, we’re done
Stochastic interpretation and heuristic

- suppose $X, x$ are optimal for SDP relaxation
- generate $z$ from normal distribution $\mathcal{N}(x, X - xx^T)$
- take $x = \text{sgn}(z)$ as approximate solution of BLS
  (can repeat many times and take best one)
Interior-point methods

- handle linear and nonlinear convex problems (Nesterov & Nemirovsky)
- based on Newton’s method applied to ‘barrier’ functions that trap $x$ in \textit{interior} of feasible region (hence the name IP)
- worst-case complexity theory: $\# \text{ Newton steps} \sim \sqrt{\text{problem size}}$
- in practice: $\# \text{ Newton steps}$ between 5 & 50 (!)
- 1000s variables, 10000s constraints feasible on PC; far larger if structure is exploited
Log barrier

for convex problem

minimize $f_0(x)$

subject to $f_i(x) \leq 0$, $i = 1, \ldots, m$

we define logarithmic barrier as

$$\phi(x) = - \sum_{i=1}^{m} \log(-f_i(x))$$

- $\phi$ is convex, smooth on interior of feasible set
- $\phi \to \infty$ as $x$ approaches boundary of feasible set
Central path

central path is curve

\[ x^*(t) = \arg\min_x \left( t f_0(x) + \phi(x) \right), \quad t \geq 0 \]

- \( x^*(t) \) is strictly feasible, \textit{i.e.}, \( f_i(x) < 0 \)
- \( x^*(t) \) can be computed by, \textit{e.g.}, Newton’s method
- intuition suggests \( x^*(t) \) converges to optimal as \( t \to \infty \)
- using duality can prove \( x^*(t) \) is \( m/t \)-suboptimal
Example: central path for LP

\[ x^*(t) = \arg\min_x \left( tc^T x - \sum_{i=1}^{6} \log(b_i - a_i^T x) \right) \]
Barrier method

a.k.a. path-following method

given strictly feasible $x$, $t > 0$, $\mu > 1$

repeat

1. compute $x := x^*(t)$
   (using Newton’s method, starting from $x$)

2. exit if $m/t < \text{tol}$

3. $t := \mu t$

duality gap reduced by $\mu$ each outer iteration
Trade-off in choice of $\mu$

large $\mu$ means

- fast duality gap reduction (fewer outer iterations), but

- many Newton steps to compute $x^*(t^+)$
  (more Newton steps per outer iteration)

total effort measured by total number of Newton steps
Typical example

GP with $n = 50$ variables, $m = 100$ constraints, $m_i = 5$

- wide range of $\mu$ works well
- very typical behavior (even for large $m$, $n$)
Effect of $\mu$

barrier method works well for $\mu$ in large range
Typical effort versus problem dimensions

- LPs with $n = 2m$ variables, $m$ constraints
- 100 instances for each of 20 problem sizes
- avg & std dev shown
Other interior-point methods

more sophisticated IP algorithms

- primal-dual, incomplete centering, infeasible start
- use same ideas, \textit{e.g.}, central path, log barrier
- readily available (commercial and noncommercial packages)

typical performance: 10 – 50 Newton steps (!)
— over wide range of problem dimensions, problem type, and problem data
Complexity analysis of Newton’s method

- classical result: if $|f'''|$ small, Newton’s method converges fast
- classical analysis is local, and coordinate dependent
- need analysis that is global, and, like Newton’s method, coordinate invariant
Self-concordance

Self-concordant function $f$ (Nesterov & Nemirovsky, 1988): when restricted to any line,

$$|f'''(t)| \leq 2f''(t)^{3/2}$$

- $f$ SC $\iff \tilde{f}(z) = f(Tz)$ SC, for $T$ nonsingular (i.e., SC is coordinate invariant)

- a large number of common convex functions are SC

$$x \log x - \log x, \quad \log \det X^{-1}, \quad - \log(y^2 - x^T x), \quad \ldots$$
Complexity analysis of Newton’s method for self-concordant functions

for self-concordant function $f$, with minimum value $f^*$,

• **Theorem:** #Newton steps to minimize $f$, starting from $x$:

$$
\text{#steps} \leq 11(f(x) - f^*) + 5
$$

• **Empirically:** #steps $\approx 0.6(f(x) - f^*) + 5$

note absence of unknown constants, problem dimension, etc.
Complexity of path-following algorithm

- to compute $x^*(\mu t)$ starting from $x^*(t)$,

$$\#\text{steps} \leq 11m(\mu - 1 - \log \mu) + 5$$

using N&N’s self-concordance theory, duality to bound $f^*$

- number of outer steps to reduce duality gap by factor $\alpha$: $\lceil \log \alpha / \log \mu \rceil$

- total number of Newton steps bounded by product,

$$\left\lfloor \frac{\log \alpha}{\log \mu} \right\rfloor (11m(\mu - 1 - \log \mu) + 5)$$

... captures trade-off in choice of $\mu$
Complexity analysis conclusions

- for any choice of $\mu$, $\#\text{steps is } O(m \log 1/\epsilon)$, where $\epsilon$ is final accuracy
- to optimize complexity bound, can take $\mu = 1 + 1/\sqrt{m}$, which yields $\#\text{steps } O(\sqrt{m} \log 1/\epsilon)$
- in any case, IP methods work extremely well in practice
Conclusions

since 1985, lots of advances in theory & practice of convex optimization

• complexity analysis

• semidefinite programming, other new problem classes

• efficient interior-point methods & software

• lots of applications
Some references

- *Semidefinite Programming*, SIAM Review 1996
- *Determinant Maximization with Linear Matrix Inequality Constraints*, SIMAX 1998
- *Applications of Second-order Cone Programming*, LAA 1999
Shameless promotion

*Convex Optimization*, Boyd & Vandenberghe

- to be published 2003

- pretty good draft available at Stanford EE364 (UCLA EE236B) class web site as course reader