# Convex Optimization and Applications

**Lin Xiao**

Center for the Mathematics of Information
California Institute of Technology

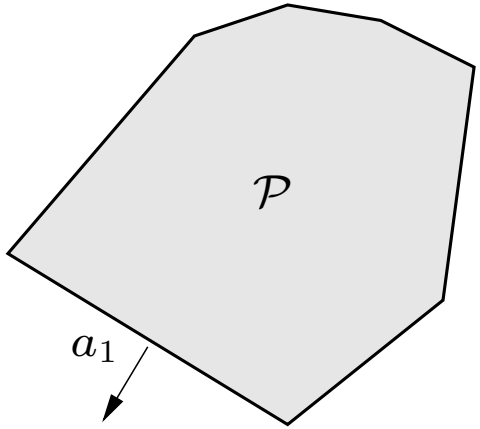
*Acknowledgment:*

**Stephen Boyd** (Stanford) and **Lieven Vandenberghe** (UCLA)
and their research groups

CS286a: Mathematics of Information seminar, 10/15/04

# Two problems

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




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

**Problem 2:** find maximum volume ellipsoid $\subseteq \mathcal{P}$

are these (computationally) difficult? or easy?

problem 1 is **very difficult**

- in practice
- in theory (NP-hard)

problem 2 is **very easy**

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

# Moral

**very difficult** and **very easy** problems can look **quite similar**

. . . unless we are trained to recognize the difference

# Linear program (LP)

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & a_i^T x \leq b_i, \quad i = 1, \ldots, m \end{array}$$

$c,\ a_i \in \mathbf{R}^n$ are parameters; $x \in \mathbf{R}^n$ is variable

- **easy** to solve, in theory and practice
- can solve dense problems with $n = 1000$ vbles, $m = 10000$ constraints easily; far larger for sparse or structured problems

# Polynomial minimization

$$\text{minimize} \quad p(x)$$

$p$ is polynomial of degree $d$; $x \in \mathbf{R}^n$ is variable

- except for special cases (e.g., $d = 2$) this is a **very difficult problem**
- even sparse problems with size $n = 20$, $d = 10$ are essentially intractable
- all algorithms known to solve this problem require effort exponential in $n$

# Moral

- a problem can **appear*** **hard**, but **be easy**
- a problem can **appear*** **easy**, but **be hard**

* if we are not trained to recognize them

# What makes a problem easy or hard?

classical view:

- **linear** is easy
- **nonlinear** is hard(er)

# What makes a problem easy or hard?

emerging (and correct) view:

**. . . the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity.**

*— R. Rockafellar, SIAM Review 1993*

# Convex optimization

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_1(x) \leq 0, \ldots, f_m(x) \leq 0, \quad Ax = b \end{array}$$

$x \in \mathbf{R}^n$ is optimization variable; $f_i : \mathbf{R}^n \to \mathbf{R}$ are **convex**:

$$f_i(\lambda x + (1-\lambda)y) \leq \lambda f_i(x) + (1-\lambda) f_i(y)$$

for all $x$, $y$, $0 \leq \lambda \leq 1$

- includes least-squares, linear programming, maximum volume ellipsoid in polyhedron, and **many others**
- convex problems are **fundamentally tractable**

# Example: Robust LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \mathbf{Prob}(a_i^T x \leq b_i) \geq \eta, \quad i = 1, \ldots, m \end{array}$$

coefficient vectors $a_i$ IID, $\mathcal{N}(\overline{a}_i, \Sigma_i)$; $\eta$ is required reliability

- for fixed $x$, $a_i^T x$ is $\mathcal{N}(\overline{a}_i^T x, x^T \Sigma_i x)$
- so for $\eta = 50\%$, robust LP reduces to LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \overline{a}_i^T x \leq b_i, \quad i = 1, \ldots, m \end{array}$$

  and so is easily solved

- what about other values of $\eta$, *e.g.*, $\eta = 10\%$? $\eta = 90\%$?

constraint $\mathbf{Prob}(a_i^T x \leq b_i) \geq \eta$ equivalent to

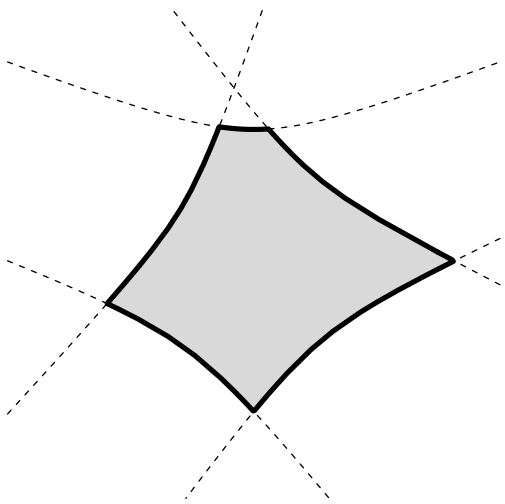
$$\bar{a}_i^T x + \Phi^{-1}(\eta)\|\Sigma_i^{1/2} x\|_2 - b_i \leq 0$$

$\Phi$ is CDF of unit Gaussian

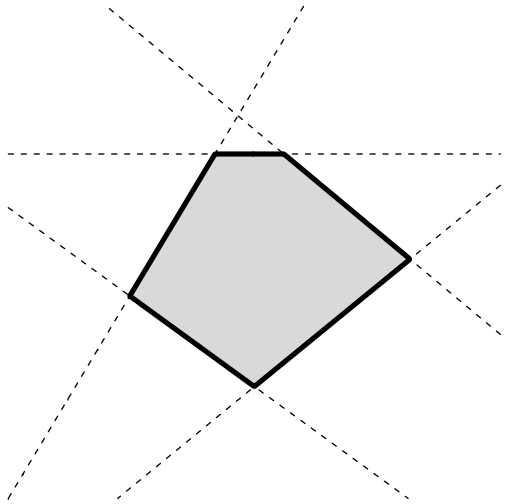
is LHS a convex function?

# Hint

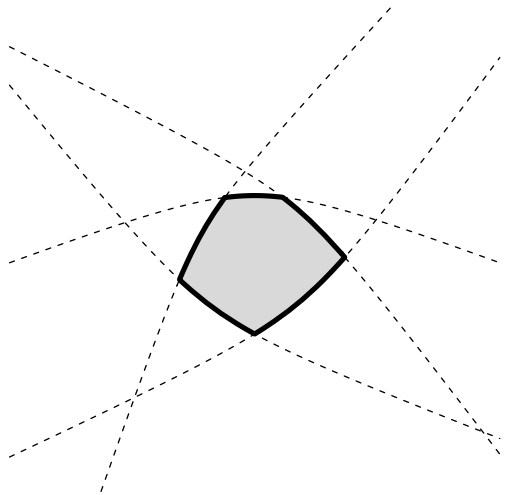
$$\{x \mid \mathbf{Prob}(a_i^T x \leq b_i) \geq \eta,\ i = 1, \ldots, m\}$$



$\eta = 10\%$



$\eta = 50\%$



$\eta = 90\%$

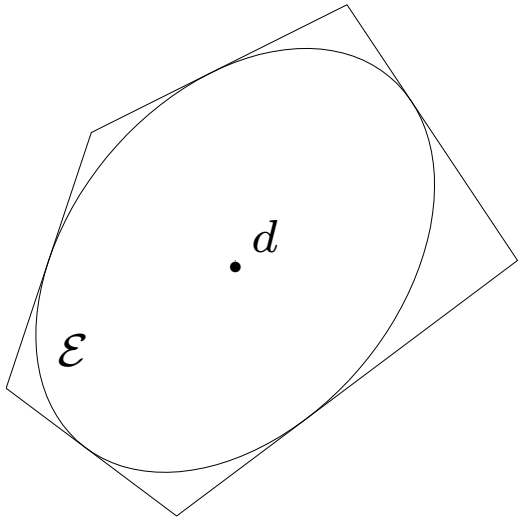
# That's right

robust LP with reliability $\eta = 90\%$ is convex, and **very easily solved**

robust LP with reliability $\eta = 10\%$ is not convex, and **extremely difficult**

# Maximum volume ellipsoid in polyhedron

- polyhedron: $\mathcal{P} = \{x \mid a_i^T x \leq b_i,\ i = 1, \ldots, m\}$
- ellipsoid: $\mathcal{E} = \{By + d \mid \|y\| \leq 1\}$, with $B = B^T \succ 0$




**maximum volume** $\mathcal{E} \subseteq \mathcal{P}$, as convex problem in variables $B$, $d$:

$$\begin{array}{ll} \text{maximize} & \log\det B \\ \text{subject to} & B = B^T \succ 0, \quad \|Ba_i\| + a_i^T d \leq b_i, \quad i = 1, \ldots, m \end{array}$$

# Moral

- it's not easy to recognize convex functions and convex optimization problems
- **huge benefit**, though, when we do

# Convex Analysis and Optimization

# 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)

- primal-dual interior-point (IP) methods
  *extremely efficient, handle nonlinear large scale problems, polynomial-time complexity results, software implementations*

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

- extension to generalized inequalities
  *semidefinite, cone programming*

# Applications and uses

- lots of applications
  *control, combinatorial optimization, signal processing, circuit design, communications, machine learning . . .*

- robust optimization
  *robust versions of LP, least-squares, other problems*

- relaxations and randomization
  *provide bounds, heuristics for solving hard ($e.g.$, combinatorial optimization) problems*

# 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, . . .*
- 1988: Nesterov & Nemirovsky's self-concordance analysis
- 1989–: LMIs and semidefinite programming in control
- 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, . . .*

# New Standard Convex Problem Classes

# Some new standard convex problem classes

- second-order cone program (SOCP)
- geometric program (GP) (and entropy problems)
- semidefinite program (SDP)

all these new problem classes have

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

# Second-order cone program

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

$$\begin{array}{ll} \text{minimize} & c_0^T x \\ \text{subject to} & \|A_i x + b_i\|_2 \leq c_i^T x + d_i, \quad i = 1, \ldots, m \end{array}$$

with variable $x \in \mathbf{R}^n$

- includes LP and QP as special cases
- nondifferentiable when $A_i x + b_i = 0$
- new IP methods can solve (almost) as fast as LPs

# Example: robust linear program

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & \mathbf{Prob}(a_i^T x \leq b_i) \geq \eta, \quad i = 1, \ldots, m \end{array}$$

where $a_i \sim \mathcal{N}(\bar{a}_i, \Sigma_i)$

equivalent to

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

where $\Phi$ is (unit) normal CDF

robust LP is an SOCP for $\eta \geq 0.5$ ($\Phi^{-1}(\eta) \geq 0$)

# Geometric program (GP)

**log-sum-exp** function:

$$\mathbf{lse}(x) = \log\left(e^{x_1} + \cdots + e^{x_n}\right)$$

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

**geometric program**:

$$\begin{array}{ll} \text{minimize} & \mathbf{lse}(A_0 x + b_0) \\ \text{subject to} & \mathbf{lse}(A_i x + b_i) \leq 0, \quad i = 1, \ldots, m \end{array}$$

$A_i \in \mathbf{R}^{m_i \times n}$, $b_i \in \mathbf{R}^{m_i}$; variable $x \in \mathbf{R}^n$

# Posynomial form geometric program

$x = (x_1, \ldots, x_n)$: vector of **positive** variables

function $f$ of form

$$f(x) = \sum_{k=1}^{t} c_k x_1^{\alpha_{1k}} x_2^{\alpha_{2k}} \cdots x_n^{\alpha_{nk}}$$

with $c_k \geq 0$, $\alpha_{ik} \in \mathbf{R}$, is called **posynomial**

like polynomial, but

- coefficients must be positive
- exponents can be fractional or negative

# Posynomial form geometric program

**posynomial form GP:**

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 1, \quad i = 1, \ldots, m \end{array}$$

$f_i$ are posynomial; $x_i > 0$ are variables

to convert to (convex form) GP, express as

$$\begin{array}{ll} \text{minimize} & \log f_0(e^y) \\ \text{subject to} & \log f_i(e^y) \leq 0, \quad i = 1, \ldots, m \end{array}$$

objective and constraints have form $\mathbf{lse}(A_i y + b_i)$

# Entropy problems

**unnormalized negative entropy** is convex function

$$-\mathbf{entr}(x) = \sum_{i=1}^{n} x_i \log(x_i/\mathbf{1}^T x)$$

defined for $x_i \geq 0$, $\mathbf{1}^T x > 0$

**entropy problem:**

$$\begin{array}{ll} \text{minimize} & -\mathbf{entr}(A_0 x + b_0) \\ \text{subject to} & -\mathbf{entr}(A_i x + b_i) \leq 0, \quad i = 1, \ldots, m \end{array}$$

$A_i \in \mathbf{R}^{m_i \times n}$, $b_i \in \mathbf{R}^{m_i}$

# Solving GPs (and entropy problems)

- GP and entropy problems are **duals** (if we solve one, we solve the other)
- new IP methods can solve large scale GPs (and entropy problems) almost as fast as LPs
- applications in many areas:
  - information theory, statistics
  - communications, wireless power control
  - digital and analog circuit design

# Generalized inequalities

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

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

**convex optimization problem with generalized inequalities:**

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

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

$$f_i(\lambda x + (1-\lambda)y) \preceq_{K_i} \lambda f_i(x) + (1-\lambda) f_i(y)$$

# Semidefinite program

**semidefinite program** (SDP):

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & x_1 A_1 + \cdots + x_n A_n \preceq B \end{array}$$

- $B,\ A_i$ are symmetric matrices; variable is $x \in \mathbf{R}^n$
- $\preceq$ is matrix inequality; constraint is **linear matrix inequality** (LMI)
- SDP can be expressed as convex problem as

$$\lambda_{\max}(B - x_1 A_1 \cdots - x_n A_n) \leq 0$$

  or handled directly as **cone problem**

# Semidefinite 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
  - communications and information theory
  - quantum computing
  - algebraic geometry
  - machine learning

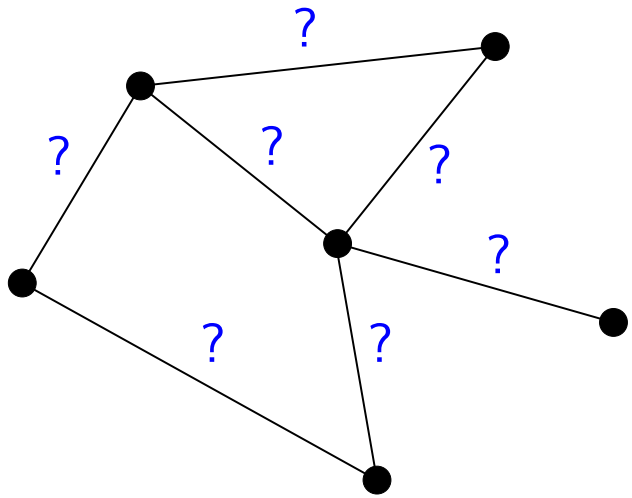
## Continuous time symmetric Markov chain on a graph

- connected graph $\mathcal{G} = (\mathcal{V},\ \mathcal{E})$, $\mathcal{V} = \{1, \ldots, n\}$, no self-loops
- Markov process on $\mathcal{V}$, with symmetric rate matrix $Q$
  - $Q_{ij} = 0$ for $(i, j) \notin \mathcal{E}$, $i \neq j$
  - $Q_{ij} \geq 0$ for $(i, j) \in \mathcal{E}$
  - $Q_{ii} = -\sum_{j \neq i} Q_{ij}$
- eigenvalues of $Q$ ordered as $0 = \lambda_1(Q) > \lambda_2(Q) \geq \cdots \geq \lambda_n(Q)$
- state distribution given by $\pi(t) = e^{tQ}\pi(0)$
- distribution converges to uniform with rate determined by $\lambda_2$

$$\|\pi(t) - \mathbf{1}/n\|_{\mathrm{tv}} \leq (\sqrt{n}/2)e^{\lambda_2(Q)t}$$

# Fastest mixing Markov chain (FMMC) problem

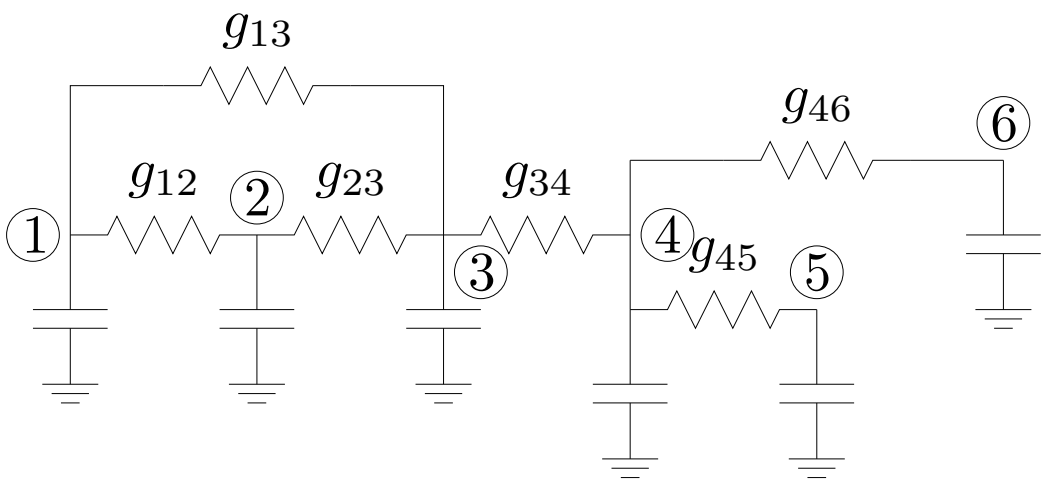
$$
\begin{array}{ll}
\text{minimize} & \lambda_2(Q) \\
\text{subject to} & Q\mathbf{1} = 0, \quad Q = Q^T \\
& Q_{ij} = 0 \text{ for } (i,j) \notin \mathcal{E},\ i \neq j \\
& Q_{ij} \geq 0 \text{ for } (i,j) \in \mathcal{E} \\
& \sum_{(i,j)\in\mathcal{E}} d_{ij}^2 Q_{ij} \leq 1
\end{array}
$$



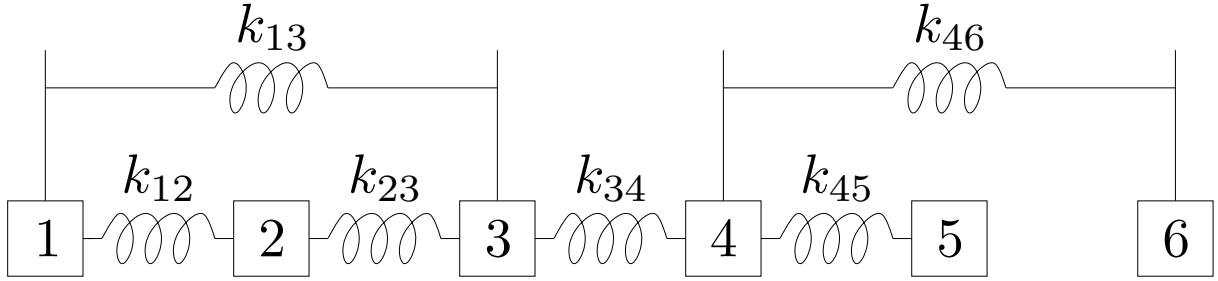

- variable is matrix $Q$; problem data is graph, constants $d_{ij} = d_{ij}$ on $\mathcal{E}$
- need to add constraint on rates since $\lambda_2$ is homogeneous
- optimal $Q$ gives fastest diffusion process on graph
  (subject to rate constraint)

# Fast diffusion processes

electrical




mechanical



# Swap objective and constraint

- $\lambda_2(Q)$ and $\sum_{(i,j)\in\mathcal{E}} d_{ij}^2 Q_{ij}$ are homogeneous
- hence, can just as well minimize weighted rate sum $\sum_{(i,j)\in\mathcal{E}} d_{ij}^2 Q_{ij}$ subject to bound on $\lambda_2(Q)$

$$
\begin{array}{ll}
\text{minimize} & \sum_{(i,j)\in\mathcal{E}} d_{ij}^2 Q_{ij} \\
\text{subject to} & Q\mathbf{1} = 0, \quad Q = Q^T \\
& Q_{ij} = 0 \text{ for } (i,j) \notin \mathcal{E},\ i \neq j \\
& Q_{ij} \geq 0 \text{ for } (i,j) \in \mathcal{E} \\
& \lambda_2(Q) \leq -1
\end{array}
$$

# SDP formulation of FMMC

using $Q\mathbf{1} = 0$, we have

$$\lambda_2(Q) \leq -1 \iff Q - \mathbf{1}\mathbf{1}^T/n \preceq -I$$

so FMMC reduces to SDP

$$\begin{array}{ll} \text{minimize} & \sum_{(i,j)\in\mathcal{E}} d_{ij}^2 Q_{ij} \\ \text{subject to} & Q\mathbf{1} = 0, \quad Q = Q^T \\ & Q_{ij} = 0 \text{ for } (i,j) \notin \mathcal{E},\ i \neq j \\ & Q_{ij} \geq 0 \text{ for } (i,j) \in \mathcal{E} \\ & Q - \mathbf{1}\mathbf{1}^T/n \preceq -I \end{array}$$

hence: can solve efficiently, duality theory, . . .

# Robust Optimization

# Robust optimization problem

robust optimization problem:

$$\begin{array}{ll} \text{minimize} & \max_{a \in \mathcal{A}} f_0(a, x) \\ \text{subject to} & \max_{a \in \mathcal{A}} f_i(a, x) \leq 0, \quad i = 1, \ldots, m \end{array}$$

- $x$ is optimization variable
- $a$ is **uncertain parameter**
- $\mathcal{A}$ is parameter set ($e.g.$, box, polyhedron, ellipsoid)

heuristics (detuning, sensitivity penalty term) have been used since beginning of optimization

# Robust optimization via convex optimization

*El Ghaoui, Ben Tal & Nemirovsky, . . .*

- robust versions of LP, QP, SOCP problems, with ellipsoidal or polyhedral uncertainty, can be formulated as SDPs (or simpler)
- other robust problems ($e.g.$, SDP) intractable, but there are good convex approximations

# Robust least-squares

robust LS problem with uncertain parameters $v_1, \ldots, v_p$

$$\text{minimize} \quad \sup_{\|v\| \leq 1} \|(A_0 + v_1 A_1 + \cdots + v_p A_p)x - b\|^2$$

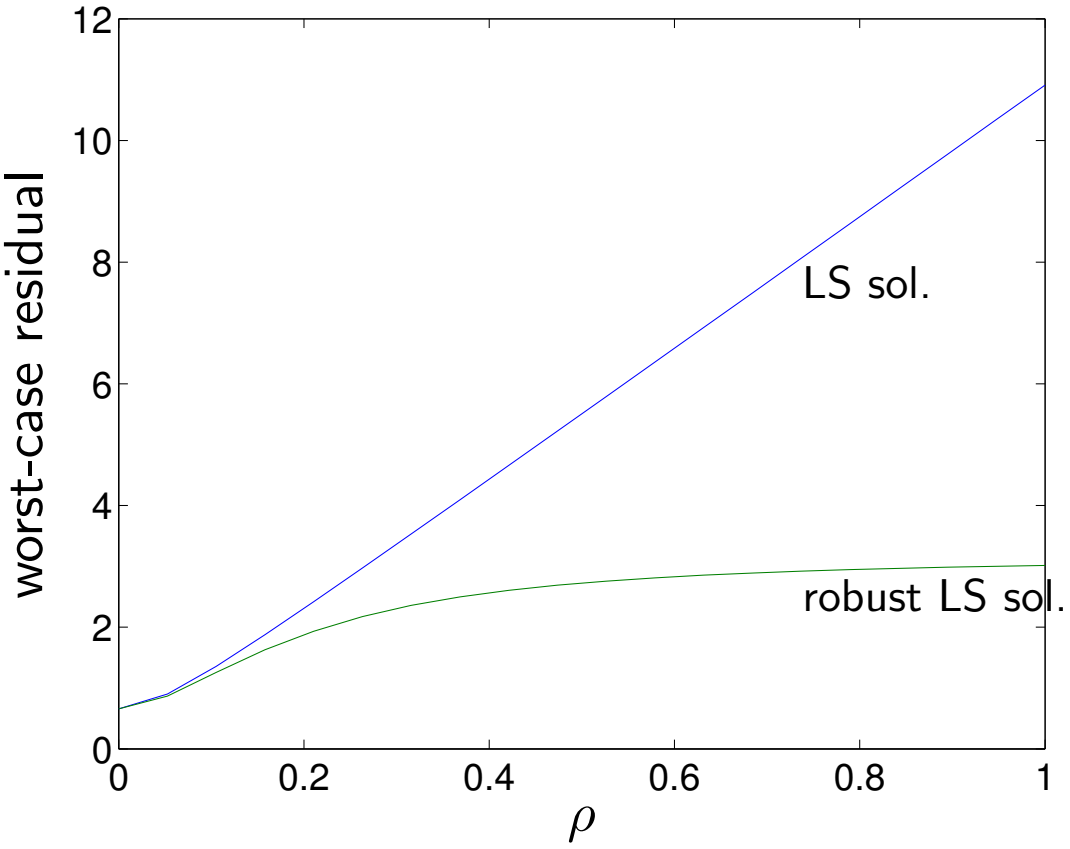
**equivalent SDP** (variables $x$, $t_1$, $t_2$):

$$\begin{array}{ll} \text{minimize} & t_1 + t_2 \\ \text{subject to} & \begin{bmatrix} I & P(x) & q(x) \\ P(x)^T & t_1 I & 0 \\ q(x)^T & 0 & t_2 \end{bmatrix} \succeq 0 \end{array}$$

where

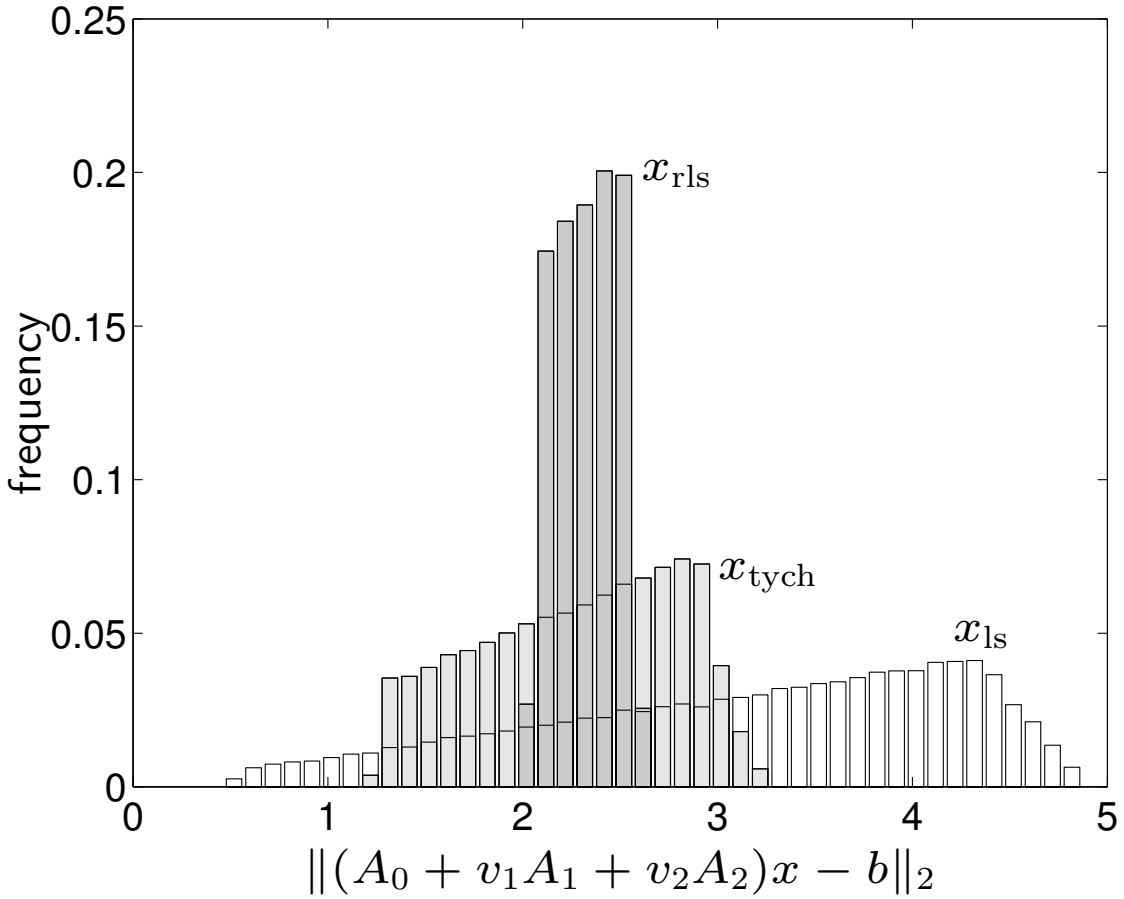
$$P(x) = \begin{bmatrix} A_1 x & A_2 x & \cdots & A_p x \end{bmatrix} \in \mathbf{R}^{m \times p}, \qquad q(x) = A_0 x - b$$

**example:** minimize $\sup_{\|v\| \leq \rho} \|(A_0 + v_1 A_1 + v_2 A_2)x - b\|^2$




$(\|A_0\| = 10,\ \|A_1\| = \|A_2\| = 1)$

**example:** minimize $\sup_{\|v\|\leq 1} \|(A_0 + v_1A_1 + v_2A_2)x - b\|^2$




distribution assuming $v$ is uniformly distributed

# Relaxations & Randomization

# Relaxations & randomization

convex optimization is increasingly used

- to find good bounds for hard (i.e., nonconvex) problems, via **relaxation**
- as a heuristic for finding good suboptimal points, often via **randomization**

# Example: Boolean least-squares

**Boolean least-squares problem**:

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|^2 \\ \text{subject to} & x_i^2 = 1, \quad i = 1, \ldots, n \end{array}$$

- basic problem in digital communications
- could check all $2^n$ possible values of $x$ . . .
- an NP-hard problem, and very hard in practice
- many heuristics for approximate solution

# Boolean least-squares as matrix problem

$$\begin{aligned} \|Ax-b\|^2 &= x^TA^TAx - 2b^TAx + b^Tb \\ &= \mathbf{Tr}\, A^TAX - 2b^TA^Tx + b^Tb \end{aligned}$$

where $X = xx^T$

hence can express BLS as

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}\, A^TAX - 2b^TAx + b^Tb \\ \text{subject to} & X_{ii} = 1, \quad X \succeq xx^T, \quad \text{rank}(X) = 1 \end{array}$$

. . . 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 $X$, $x$)

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}\, A^T A X - 2b^T A^T x + b^T b \\ \text{subject to} & X_{ii} = 1, \quad \begin{bmatrix} X & x \\ x^T & 1 \end{bmatrix} \succeq 0 \end{array}$$

- optimal value of SDP gives **lower bound** for BLS
- if optimal matrix is rank one, we're done

# Interpretation via randomization

- can think of variables $X$, $x$ in SDP relaxation as defining a normal distribution $z \sim \mathcal{N}(x, X - xx^T)$, with $\mathbf{E}\, z_i^2 = 1$
- SDP objective is $\mathbf{E}\, \|Az - b\|^2$

suggests randomized method for BLS:

- find $X^\star$, $x^\star$, optimal for SDP relaxation
- generate $z$ from $\mathcal{N}(x^\star, X^\star - x^\star x^{\star T})$
- take $x = \mathbf{sgn}(z)$ as approximate solution of BLS
  (can repeat many times and take best one)

# Example

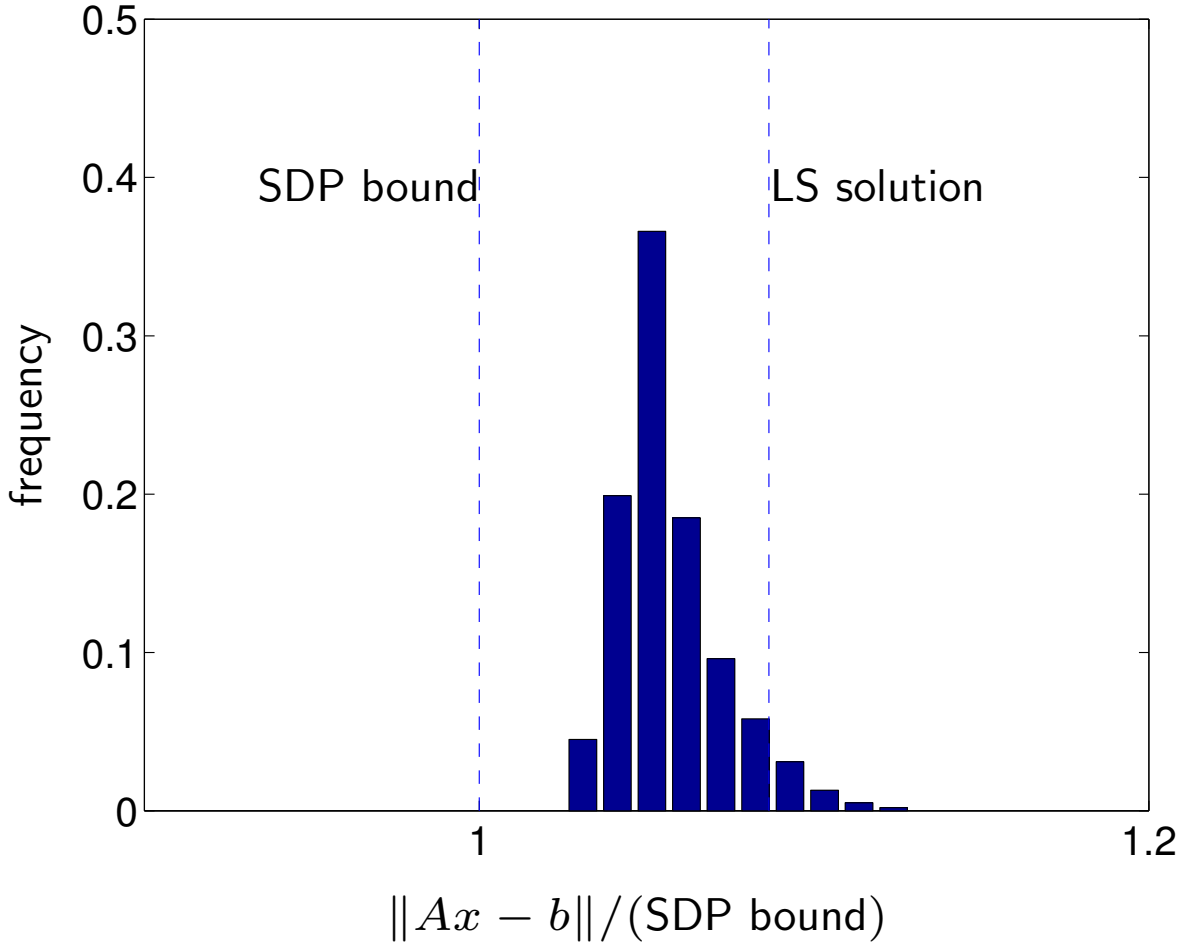
- (randomly chosen) parameters $A \in \mathbf{R}^{150\times 100}$, $b \in \mathbf{R}^{150}$
- $x \in \mathbf{R}^{100}$, so feasible set has $2^{100} \approx 10^{30}$ points

**LS approximate solution:** minimize $\|Ax - b\|$ s.t. $\|x\|^2 = n$, then round

yields objective 8.7% over SDP relaxation bound

**randomized method:** (using SDP optimal distribution)

- best of 20 samples: $3.1\%$ over SDP bound
- best of 1000 samples: $2.6\%$ over SDP bound


0.5
0.4
0.3
0.2
0.1
0
frequency
SDP bound
LS solution
1
1.2
$\|Ax - b\|/(\text{SDP bound})$

# Calculus of Convex Functions

# Approach

- basic examples or atoms
- calculus rules or transformations that preserve convexity

# Convex functions: Basic examples

- $x^p$ for $p \geq 1$ or $p \leq 0$; $-x^p$ for $0 \leq p \leq 1$
- $e^x$, $-\log x$, $x \log x$
- $x^T x$; $x^T x/y$ (for $y > 0$); $(x^T x)^{1/2}$
- $\|x\|$ (any norm)
- $\max(x_1, \ldots, x_n)$, $\log(e^{x_1} + \cdots + e^{x_n})$
- $\log \Phi(x)$ ($\Phi$ is Gaussian CDF)
- $\log \det X^{-1}$ (for $X \succ 0$)

# Calculus rules

- convexity preserved under sums, nonnegative scaling
- if $f$ cvx, then $g(x) = f(Ax + b)$ cvx
- pointwise sup: if $f_\alpha$ cvx for each $\alpha \in A$, then $g(x) = \sup_{\alpha \in A} f_\alpha(x)$ cvx
- minimization: if $f(x, y)$ cvx, then $g(x) = \inf_y f(x, y)$ cvx
- composition rules: if $h$ cvx & increasing, $f$ cvx, then $g(x) = h(f(x))$ cvx
- perspective transformation: if $f$ cvx, then $g(x, t) = tf(x/t)$ cvx for $t > 0$

. . . and many, many others

# More examples

- $\lambda_{\max}(X)$ (for $X = X^T$)
- $f(x) = x_{[1]} + \cdots + x_{[k]}$ (sum of largest $k$ elements of $x$)
- $-\sum_{i=1}^{m} \log(-f_i(x))$ (on $\{x \mid f_i(x) < 0\}$; $f_i$ cvx)
- $f(x) = \log \mathbf{Prob}(x + z \in C)$ ($C$ convex, $z \sim \mathcal{N}(0, \Sigma)$)
- $x^T Y^{-1} x$ is cvx in $(x, Y)$ for $Y = Y^T \succ 0$

# Duality

# Lagrangian and dual function

**primal** problem:

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \end{array}$$

**Lagrangian**: $L(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x)$

**dual function**: $g(\lambda) = \inf_x L(x, \lambda)$

# Lower bound property

- for any primal feasible $x$ and any $\lambda \geq 0$, we have $f_0(x) \leq g(\lambda)$
- hence for any $\lambda \geq 0$, $g(\lambda) \leq p^\star$ (optimal value of primal problem)
- **duality gap** of $x$, $\lambda$ is defined as $f_0(x) - g(\lambda)$
  (gap is nonnegative; bounds suboptimality of $x$)

# Dual problem

find best lower bound on $p^\star$:

$$\begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \lambda_i \geq 0, \quad i = 1, \ldots, m \end{array}$$

$p^\star$ is optimal value of primal, and $d^\star$ is optimal value of dual

- **weak duality**: even when primal not convex, $d^\star \leq p^\star$
- for convex primal problem, we have **strong duality**: $d^\star = p^\star$ (provided technical condition holds)

# Example: linear program

primal:

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \end{array}$$

dual:

$$\begin{array}{ll} \text{maximize} & b^T \lambda \\ \text{subject to} & A^T \lambda + c = 0, \quad \lambda \succeq 0 \end{array}$$

# Example: unconstrained geometric program

primal:

$$\text{minimize} \quad \log\left(\sum_{i=1}^m \exp(a_i^T x + b_i)\right)$$

dual:

$$\begin{array}{ll} \text{maximize} & b^T\nu - \sum_{i=1}^m \nu_i \log \nu_i \\ \text{subject to} & \mathbf{1}^T\nu = 1, \quad A^T\nu = 0, \quad \nu \succeq 0 \end{array}$$

. . . an entropy maximization problem

# Example: duality between FMMC and MVU

FMMC problem

$$\begin{array}{ll} \text{minimize} & \sum_{(i,j)\in\mathcal{E}} d_{ij}^2 Q_{ij} \\ \text{subject to} & Q\mathbf{1} = 0, \quad Q = Q^T \\ & Q_{ij} = 0 \text{ for } (i,j) \notin \mathcal{E},\ i \neq j \\ & Q_{ij} \geq 0 \text{ for } (i,j) \in \mathcal{E} \\ & Q - \mathbf{1}\mathbf{1}^T/n \preceq -I \end{array}$$

dual of FMMC problem

$$\begin{array}{ll} \text{maximize} & \mathbf{Tr}\, X \\ \text{subject to} & X_{ii} + X_{jj} - X_{ij} - X_{ji} \leq d_{ij}^2, \quad (i,j) \in \mathcal{E} \\ & X\mathbf{1} = 0, \quad X \succeq 0 \end{array}$$

# FMMC dual as maximum-variance unfolding

- use variables $x_1, \ldots, x_n$, with $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} [x_1 \cdots x_n]$

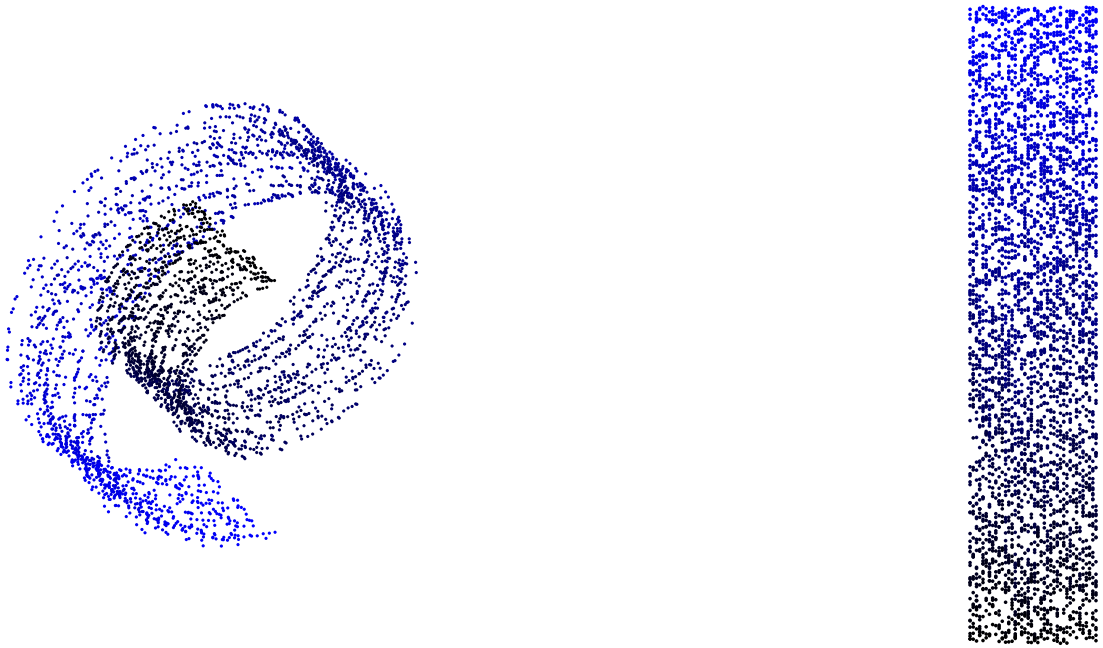
- dual FMMC problem becomes

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^n \|x_i\|^2 \\ \text{subject to} & \sum_i x_i = 0, \quad \|x_i - x_j\| \leq d_{ij}, \quad (i,j) \in \mathcal{E} \end{array}$$

- position $n$ points in $\mathbf{R}^n$ to maximize variance, respecting local distance constraints, *i.e.*, **maximum-variance unfolding problem**

# Semidefinite embedding

- similar to **semidefinite embedding** for unsupervised learning of manifolds (which has distance equality constraints)
  (Weinberger, Saul 2003)



- **surprise:** fastest diffusion on graph, max-variance unfolding are duals

# Interior-Point Methods

# Interior-point methods

- handle linear and **nonlinear** convex problems *Nesterov & Nemirovsky*
- based on Newton's method applied to 'barrier' functions that trap $x$ in **interior** of feasible region (hence the name IP)
- worst-case complexity theory: # Newton steps $\sim \sqrt{\text{problem size}}$
- in practice: # Newton steps between 20 & 50 (!)
  — over wide range of problem dimensions, type, and data
- 1000 variables, 10000 constraints feasible on PC; far larger if structure is exploited
- readily available (commercial and noncommercial) packages

# Log barrier

for convex problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \end{array}$$

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

$$x^\star(t) = \operatorname*{argmin}_x \left(t f_0(x) + \phi(x)\right), \qquad t \geq 0$$

- $x^\star(t)$ is strictly feasible, *i.e.*, $f_i(x) < 0$
- $x^\star(t)$ can be computed by, *e.g.*, Newton's method
- intuition suggests $x^\star(t)$ converges to optimal as $t \to \infty$
- using duality can prove $x^\star(t)$ is $m/t$-suboptimal

# Central path & duality

from

$$\nabla_x(tf_0(x^\star) + \phi(x^\star)) = t\nabla f_0(x^\star) + \sum_{i=1}^{m} \frac{1}{-f_i(x^\star)} \nabla f_i(x^\star) = 0$$

we find that

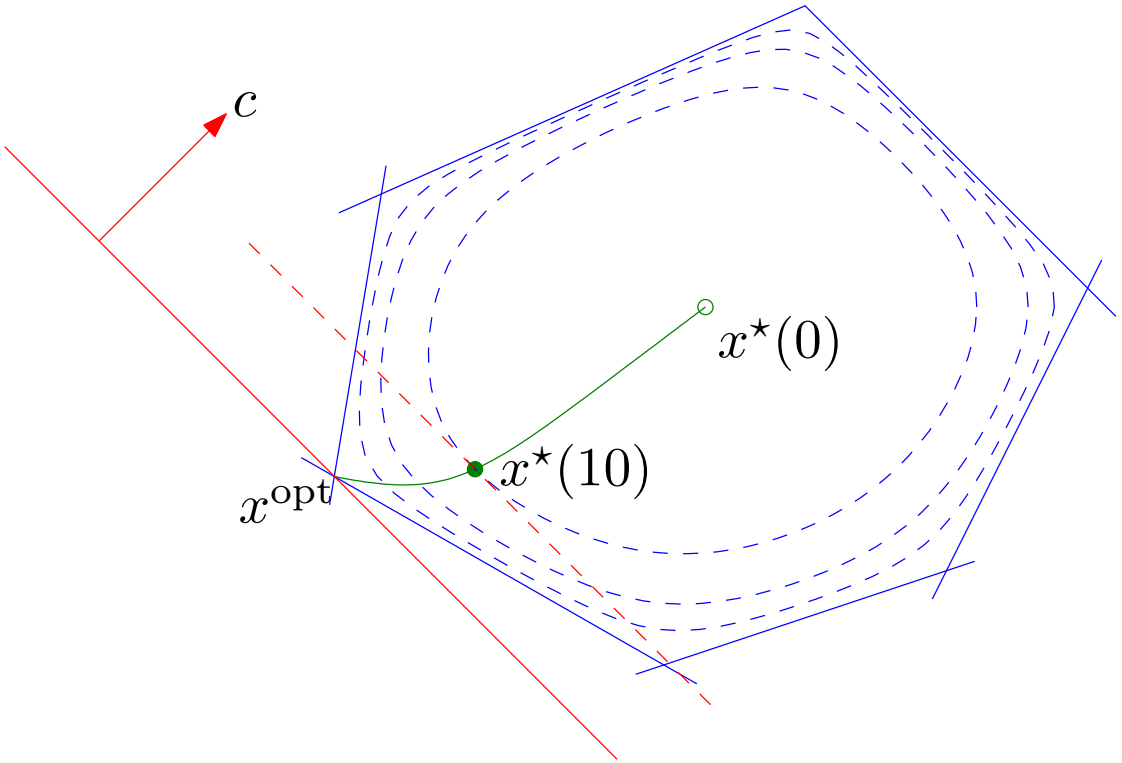
$$\lambda_i^\star(t) = \frac{1}{-tf_i(x^\star)}, \quad i = 1, \ldots, m$$

is dual feasible, with $g(\lambda^\star(t)) = f_0(x^\star) - m/t$

- duality gap associated with pair $x^\star(t),\ \lambda^\star(t)$ is $m/t$
- hence, $x^\star(t)$ is $m/t$-suboptimal

# Example: central path for LP

$$x^\star(t) = \operatorname{argmin}_x \left( tc^T x - \sum_{i=1}^6 \log(b_i - a_i^T x) \right)$$



# Barrier method

a.k.a. **path-following method**

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

**repeat**

1. compute $x := x^\star(t)$

   (using Newton's method, starting from $x$)

2. **exit if** $m/t <$ tol

3. $t := \mu t$

duality gap reduced by $\mu$ each outer iteration

# Trade-off in choice of $\mu$

large $\mu$ means

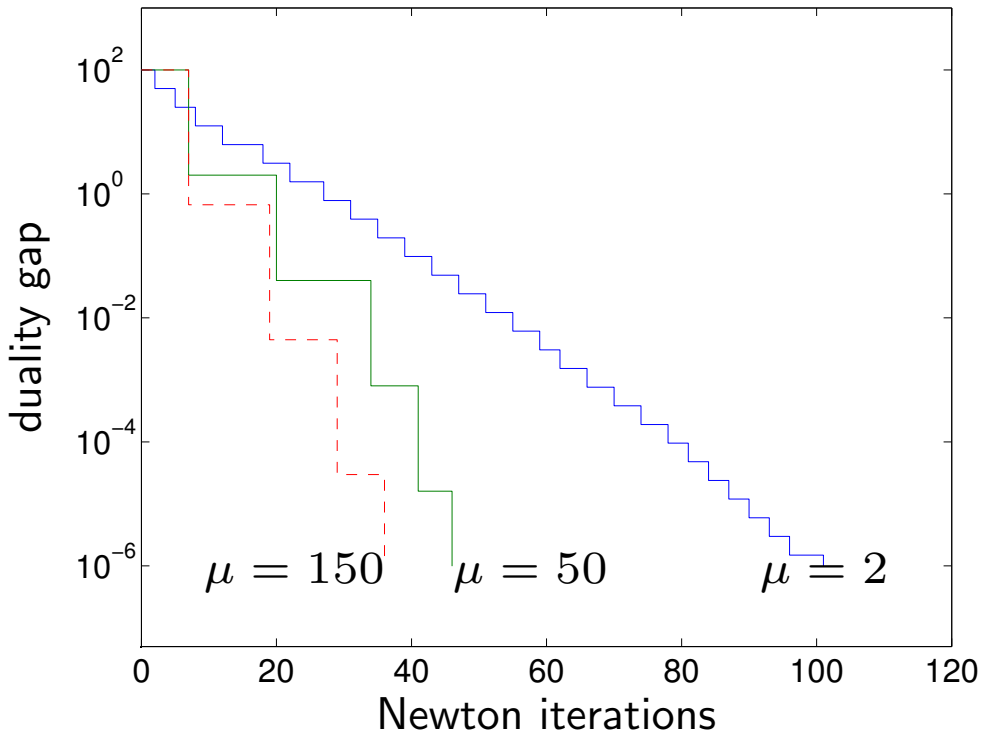
- fast duality gap reduction (fewer outer iterations), but
- many Newton steps to compute $x^\star(t^+)$
  (more Newton steps per outer iteration)

total effort measured by total number of Newton steps

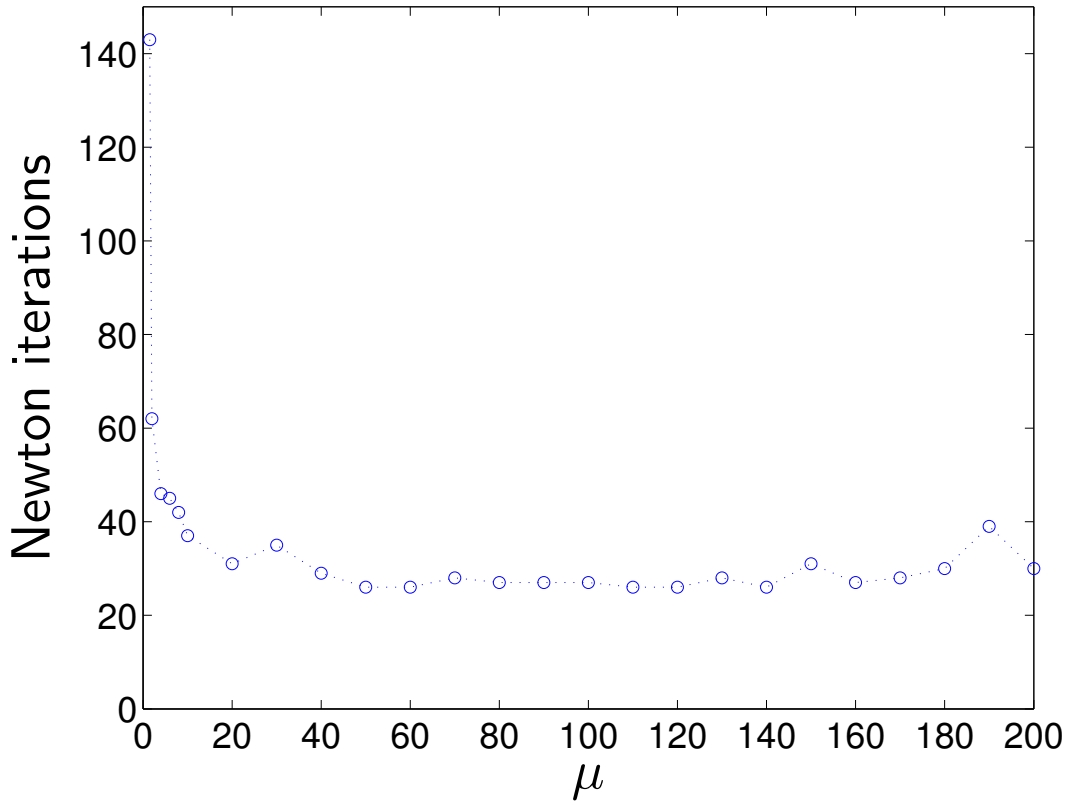
# Typical example

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

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

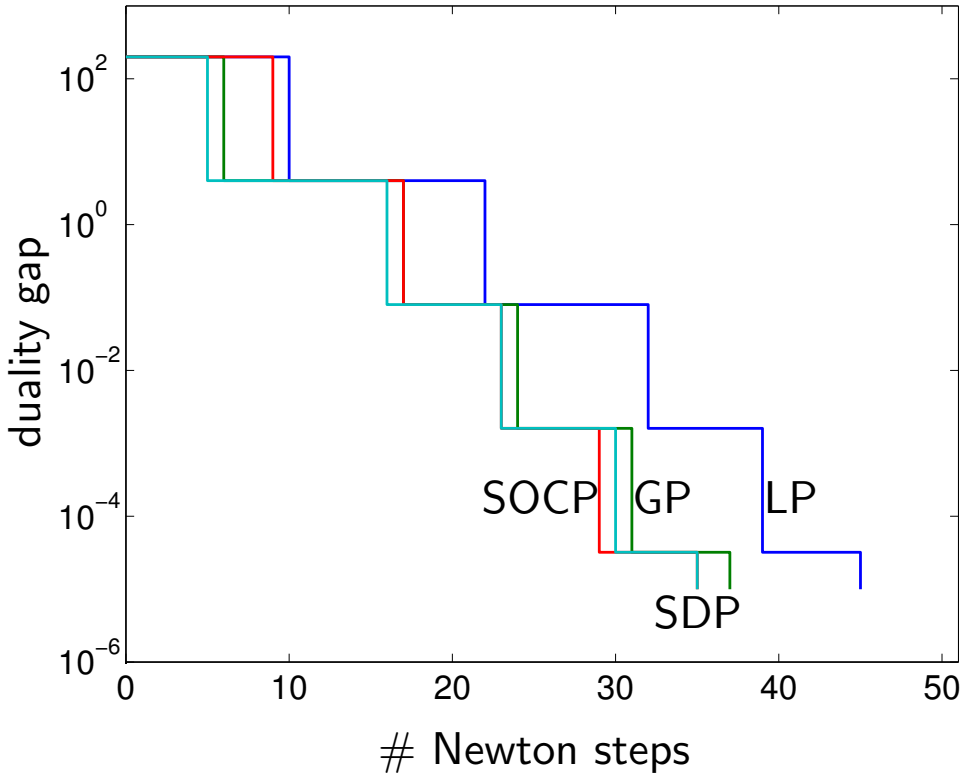


# Effect of $\mu$




barrier method works well for $\mu$ in large range

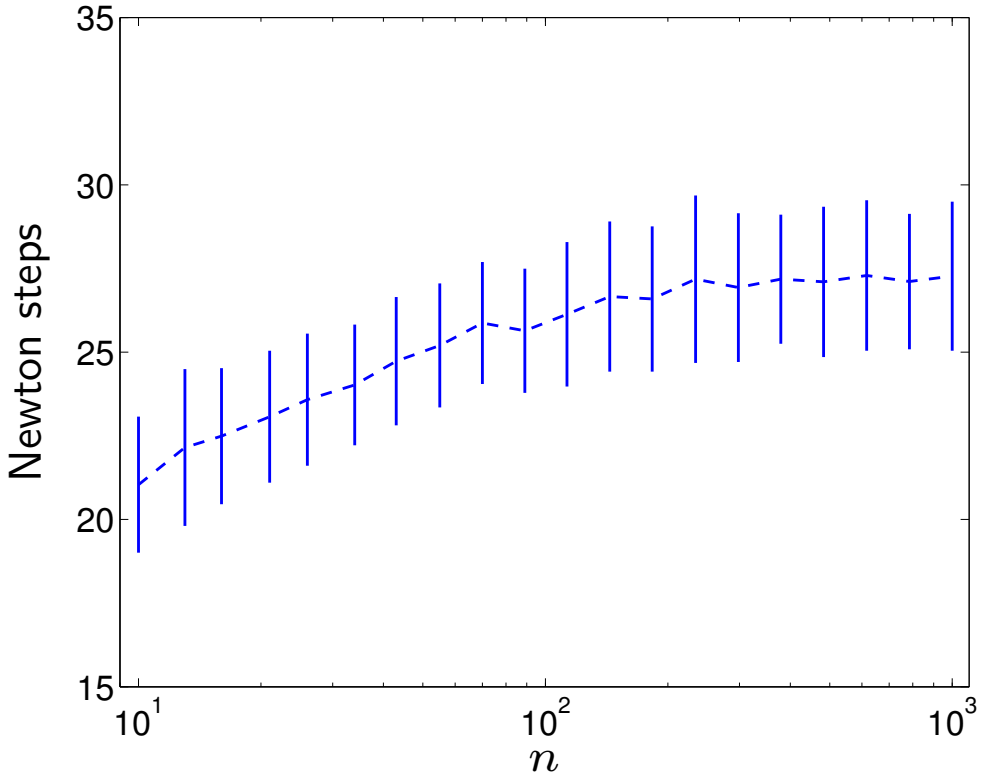
# Typical convergence of IP method




LP, GP, SOCP, SDP with 100 variables

# Typical effort versus problem dimensions

- LPs with $n$ vbles, $2n$ constraints
- $100$ instances for each of $20$ problem sizes
- avg & std dev shown



# Complexity analysis

- based on **self-concordance** (Nesterov & Nemirovsky, 1988)
- for any choice of $\mu$, #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 #steps $O(\sqrt{m} \log 1/\epsilon)$
- in any case, IP methods work extremely well in practice

# Computational effort per Newton step

- Newton step effort dominated by solving linear equations to find primal-dual search direction

- equations inherit structure from underlying problem ($e.g.$, sparsity, symmetry, toeplitz, circulant, Hankel, Kronecker)

- equations same as for least-squares problem of similar size and structure

conclusion:

we can solve a **convex problem** with about the same effort as solving **20–50 least-squares problems**

# Other interior-point methods

more sophisticated IP algorithms

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

typical performance: $20$ – $50$ Newton steps (!)
— over wide range of problem dimensions, problem type, and problem data

# Exploiting structure

**sparsity**

- well developed, since late 1970s
- direct (sparse factorizations) and iterative methods (CG, LSQR)
- standard in general purpose LP, QP, GP, SOCP implementations
- can solve problems with $10^5$, $10^6$ vbles, constraints
  (depending on sparsity pattern)

**symmetry**

- reduce number of variables
- reduce size of matrices (particularly helpful for SDP)

# A return to subgradient-type methods

- very large-scale problems: $10^5 - 10^7$ variables
- applications: medical imaging, shape design of mechanical structures, machine learning and data mining, . . .
- IPM out of consideration (even single iteration prohibitive); can use only first-order information (function values and subgradients)
- a reasonable algorithm should have
  - computational effort per iteration at most linear in design dimension
  - potential to obtain (and willing to accept) medium-accuracy solutions
  - error reduction factor essentially independent of problem dimension

*Ben Tal & Nemirovsky, Nesterov, . . .*

# Conclusions

# Conclusions

convex optimization

- theory fairly mature; practice has advanced tremendously last decade
- qualitatively different from general nonlinear programming
- cost only $30\times$ more than least-squares, but far more expressive
- **lots of applications** still to be discovered

# Some references

- Convex Optimization, *Boyd & Vandenberghe, 2004*
  `www.stanford.edu/~boyd/cvxbook.html`
  (pdf of full text on web)

- Introductory Lectures on Convex Optimization, *Nesterov, 2003*

- Lectures on Modern Convex Optimization, *Ben Tal & Nemirovsky, 2001*

- Interior-point Polynomial Algorithms in Convex Programming,
  *Nesterov & Nemirovsky, 1994*

- Linear Matrix Inequalities in System and Control Theory,
  *Boyd, El Ghaoui, Feron, & Balakrishnan, 1994*