Optimization Methods in Machine Learning

Katya Scheinberg

Lehigh University

katyas@lehigh.edu

Primal Semidefinite Programming Problem

$$\begin{array}{ll} \min & \operatorname{trace}(CX), \\ \text{s.t.} & \operatorname{trace}(A_i X) = b_i, \;\; i = 1, \ldots, m \\ & X \in \mathbf{S}^n \;\; X \succeq 0 \\ & C, A_i \in \mathbf{S}^n, b \in \mathbf{R}^m. \end{array}$$

SDP cone $K = \{x \in \mathbf{S}^n : \; X \succeq 0\}$ - self dual.

$$\max_{y, S \succeq 0} \min_{X} L(X, y, S) =$$

$$\operatorname{trace}(CX) - \sum_{i=1}^{m} y_i(\operatorname{trace}(A_i X) - b_i) - \operatorname{trace}(SX)$$

Primal Semidefinite Programming Problem

$$\begin{array}{ll} \min & \operatorname{trace}(CX), \\ \text{s.t.} & \operatorname{trace}(A_i X) = b_i, \ \ i = 1, \ldots, m \\ & X \in \mathbf{S}^n \ \ X \succeq 0 \\ & C, A_i \in \mathbf{S}^n, b \in \mathbf{R}^m. \end{array}$$

SDP cone $K = \{x \in \mathbf{S}^n : \ X \succeq 0\}$ - self dual.

Dual Semidefinite Programming Problem

$$\begin{array}{ll} \max & b^T y, \\ \text{s.t.} & \sum_{i=1}^{m} y_i A_i + S = C \\ & S \succeq 0. \end{array}$$

Duality gap and complementarity

$$A \bullet B = \mathrm{trace}(AB)$$

$$b^T y = \sum_i (A_i \bullet D) y_i = (\sum_i y_i A_i) \bullet D = C \bullet S - S \bullet X$$

Duality Gap:

$$S \bullet X \geq 0$$

$X \bullet S = 0$ at optimality (given Slater condition)

$X \bullet S = 0, X \succeq 0, S \succeq 0 \Rightarrow XS = SX = 0.$

HW: prove the last statement

Complementarity of eignevalues

Assume $\bar{X}$ and $\bar{S}$ are optimal $\Rightarrow \bar{X}\bar{S} = \bar{S}\bar{X} = 0 \Rightarrow \bar{X}$ and $\bar{S}$ commute, $\Rightarrow$

$$\bar{X} = \bar{Q}\bar{\Lambda}\bar{Q}^T,$$
$$\bar{S} = \bar{Q}\bar{W}\bar{Q}^T,$$

$$\bar{Q}\bar{Q}^T = I,$$

$$\bar{\Lambda} = \begin{bmatrix} \bar{\lambda}_1 & & \\ & \ddots & \\ & & \bar{\lambda}_n \end{bmatrix}, \quad \bar{W} = \begin{bmatrix} \bar{w}_1 & & \\ & \ddots & \\ & & \bar{w}_n \end{bmatrix}.$$

Columns of $\bar{Q}$ - orthonormal basis of **eigenvectors** of $\bar{X}$ and $\bar{S}$.
$\bar{\lambda}_i$, $\bar{w}_i$, $i = 1, \ldots, n$ - **eigenvalues** of $\bar{X}$ and $\bar{S}$, respectively.

$$\bar{X}\bar{S} = 0 \Rightarrow \bar{\lambda}_i\bar{w}_i = 0, \quad i = 1, \ldots, n-\textbf{complementarity condition}$$

Complementarity of eigenvalues

$$\bar{\Lambda} = \begin{bmatrix} \bar{\lambda}_1 & & & & & \\ & \ddots & & & & \\ & & \bar{\lambda}_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} \qquad \bar{W} = \begin{bmatrix} \bar{0} & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & w_{n-s+1} & & \\ & & & & \ddots & \\ & & & & & w_n \end{bmatrix}$$

$$\text{rank}\bar{X} = r, \;\; \text{rank}\bar{S} = s,$$

from **complementarity** $\Rightarrow r + s \leq n$.

If $r + s = n$ – $\bar{X}$ and $\bar{S}$ **are strictly complementary**.

Convex QP with linear equality constraints.

$$\begin{aligned} \min \quad & x^\top Q x + c^\top x, \\ \text{s.t.} \quad & Ax = b, \\ & A \in \mathbf{R}^{m \times n}, b \in \mathbf{R}^m, Q \succeq 0. \end{aligned}$$

$$L(x, y) = x^\top Q x + c^\top x - y^\top (Ax - b)$$

Optimality conditions

$$\nabla_x L(x, y) \quad = \quad \begin{aligned} & Qx + c - y^\top A = 0, \\ & Ax = b. \end{aligned}$$

Closed form solution via solving a linear system

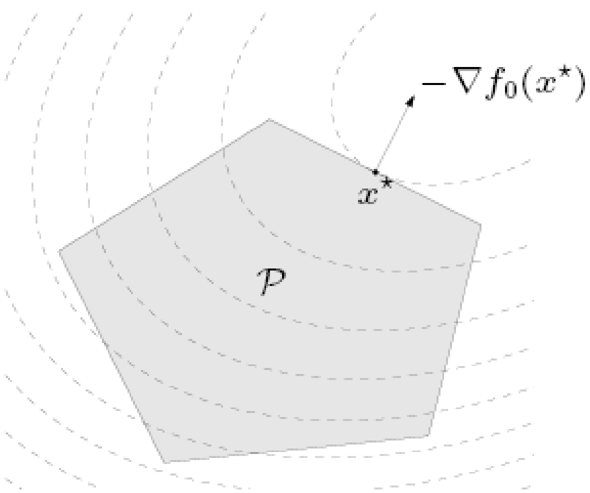
Convex QP with linear inequality constraints.

$$\begin{aligned} \min \quad & x^\top Q x + c^\top x, \\ \text{s.t.} \quad & Ax = b, \\ & x \geq 0, \end{aligned}$$

$$L(x, y) = x^\top Q x + c^\top x - y^\top (Ax - b)$$

Optimality conditions

$$\begin{aligned} & Qx + c - y^\top A - s = 0, \\ & Ax = b, \\ & s_i x_i = 0 \end{aligned}$$



No closed form solution

Convex Quadratically Constrained Quadratic Problems

$$\begin{aligned} \min \quad & x^\top Q_0 x + c_0^\top x, \\ \text{s.t.} \quad & x^\top Q_i x + c_i^\top x \leq b_i, \;\; i = 1 \ldots, m \\ & Q_i \succeq 0 \;\; i = 0 \ldots, m \end{aligned}$$

Nonlinear Constraints, linear objective:

$$\begin{aligned} \min \quad & t \\ & x^\top Q_0 x + c_0^\top x \leq t \\ \text{s.t.} \quad & x^\top Q_i x + c_i^\top x \leq b_i, \;\; i = 1 \ldots, m \\ & Q_i \succeq 0 \;\; i = 0 \ldots, m \end{aligned}$$

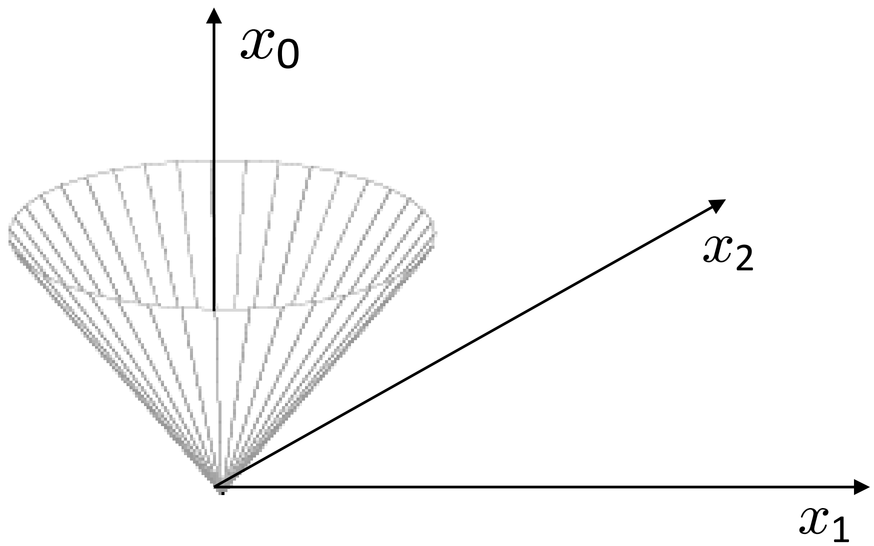
Feasible set can be described as a convex cone $\cap$ affine set

Second Order Cone

$$x = (x_0, x_1, \ldots, x_n),\ \bar{x} = (x_1, \ldots, x_n)$$

K$\in$ R^{n+1} is a second order cone:

$$x \in K \ \Leftrightarrow x \geq_K 0 \ \Leftrightarrow x^0 \geq ||\bar{x}||,$$



Discovering SOCP cone

A convex quadratic constraint: $x^\top Q x + c^\top x \le b, Q \succeq 0 \;\Leftrightarrow\; Q = LL^\top$

Factorize and rewrite: $$x^\top L L^\top x + c^\top L^{-\top} L^\top x \le b$$

Norm constraint $$||L^\top x + \frac{1}{2}L^{-1}c||^2 \le b - \frac{1}{4}c^\top L^{-\top} L c$$

More general form $$||Ax + b|| \le c^\top x + d$$

Variable substitution $$y = Ax + b \text{ and } t = c^\top x + d$$

SOCP: $$||y|| \le t,\; (y, t) \in K$$

Second Order Cone Programming

$$\begin{array}{ll} \min & {c_1}^\top x_1 + {c_2}^\top x_2 + \ldots + {c_N}^\top x_N \\ \text{s.t.} & A_1 x_1 + A_2 x_2 + \ldots + A_N x_N = b, \\ & x_i \geq_{K_i} 0, \end{array}$$

$$x_i = (x_i^0, \bar{x}_i),\ x_i \geq_{K_i} 0 \Leftrightarrow x_i^0 \geq ||\bar{x}_i||$$

$$\begin{array}{ll} \max & b^\top y \\ \text{s.t.} & {A_i}^\top y + s_i = c_i, \quad i = 1, \ldots, N \\ & s_i \geq_{K_i} 0, \end{array}$$

$A_i \in \mathbf{R}^{m \times n_i}$, $c_i \in \mathbf{R}^{n_i}$, $x_i \in \mathbf{R}^{n_i}$, $s_i \in \mathbf{R}^{n_i}$, $i = 1, \ldots, N$, $b \in \mathbf{R}^m$ $y \in \mathbf{R}^m$.
$A = [A_1, A_2, \ldots, A_N]$, $x = ({x_1}^\top, {x_2}^\top, \ldots, {x_N}^\top)^\top$ and $s = ({s_1}^\top, {s_2}^\top, \ldots, {s_N}^\top)^\top$.

Complementarity Conditions

$$\begin{aligned} x_i^0 s_i^0 + \bar{x}_i{}^\top \bar{s}_i &= 0 \quad i = 1, \ldots, N \\ s_i^0 \bar{x}_i + x_i^0 \bar{s}_i &= 0, \quad i = 1, \ldots, N \end{aligned}$$

If we define an “arrow-shaped” matrix $\mathbf{Arr}(x_i)$ as

$$\mathbf{Arr}(x_i) = \begin{bmatrix} x_i^0 & x_i^1 & \ldots & x_i^{n_i} \\ x_i^1 & x_i^0 & & \\ \vdots & & \ddots & \\ x_i^{n_i} & & & x_i^0 \end{bmatrix},$$

and the block diagonal matrix $\mathbf{Arr}(x)$ as

$$\mathbf{Arr}(x) = \begin{bmatrix} \mathbf{Arr}(x_1) & & & \\ & \mathbf{Arr}(x_2) & & \\ & & \ddots & \\ & & & \mathbf{Arr}(x_N) \end{bmatrix},$$

then the complementarity conditions can be expressed as

$$\mathbf{Arr}(x)s = \mathbf{Arr}(s)x = \mathbf{Arr}(x)\mathbf{Arr}(s)e_0 = 0,$$

where

$$e^{0^T} = (e_1^{0^T}, e_2^{0^T}, \ldots, e_N^{0^T}) \equiv (\underbrace{1, 0, \ldots, 0}_{n_1}, \underbrace{1, 0, \ldots, 0}_{n_2}, \ldots, \underbrace{1, 0, \ldots, 0}_{n_N})^\top.$$

Formulating SOCPs

Rotated SOCP cone

$$K_r = \{x = (x_0, x_1, \bar{x}) \in \mathbf{R}^{n+2} : \, x_0 x_1 \geq \|\bar{x}\|^2, x_1, x_0 \geq 0\}$$

Equivalent to SOCP cone

$$x_0 x_1 \geq \|\bar{x}\|^2 \;\Leftrightarrow\; \left\| \begin{array}{c} 2\bar{x} \\ x_0 - x_1 \end{array} \right\| \leq x_0 + x_1$$

Example: $\min_x \sum_{i=1}^{m} \frac{1}{{a_i}^\top x + b_i}, \; {a_i}^\top x + b_i > 0, \forall i = 1, \ldots, m.$

$$\begin{array}{ll} \min & \displaystyle\sum_{i=1}^{m} u_i \\ & v_i = {a_i}^\top x + b_i, \; i = 0 \ldots, m \\ \text{s.t.} & 1 \leq u_i v_i, \quad i = 1 \ldots, m \\ & u_i \geq 0 \; i = 0 \ldots, m \end{array}$$

Unconstrained Optimization

Traditional methods

- Gradient descent
- Newton method
- Quazi-Newton method
- Conjugate gradient method

Unconstrained minimization

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

- f convex, twice continuously differentiable (hence $\mathbf{dom}\, f$ open)
- we assume optimal value $p^\star = \inf_x f(x)$ is attained (and finite)

unconstrained minimization methods

- produce sequence of points $x^{(k)} \in \mathbf{dom}\, f$, $k = 0, 1, \ldots$ with

$$f(x^{(k)}) \rightarrow p^\star$$

- can be interpreted as iterative methods for solving optimality condition

$$\nabla f(x^\star) = 0$$

Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Strong convexity and implications

f is strongly convex on S if there exists an $m > 0$ such that

$$\nabla^2 f(x) \succeq mI \qquad \text{for all } x \in S$$

implications

- for $x, y \in S$,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|x - y\|_2^2$$

  hence, S is bounded

- $p^\star > -\infty$, and for $x \in S$,

$$f(x) - p^\star \leq \frac{1}{2m} \|\nabla f(x)\|_2^2$$

  useful as stopping criterion (if you know m)

Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Descent methods

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)} \quad \text{with } f(x^{(k+1)}) < f(x^{(k)})$$

- other notations: $x^+ = x + t\Delta x$, $x := x + t\Delta x$
- Δx is the *step*, or *search direction*; t is the *step size*, or *step length*
- from convexity, $f(x^+) < f(x)$ implies $\nabla f(x)^T \Delta x < 0$ (*i.e.*, Δx is a *descent direction*)

General descent method.

given a starting point $x \in \textbf{dom}\, f$.
repeat
1. Determine a descent direction Δx.
2. *Line search.* Choose a step size $t > 0$.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Line search types

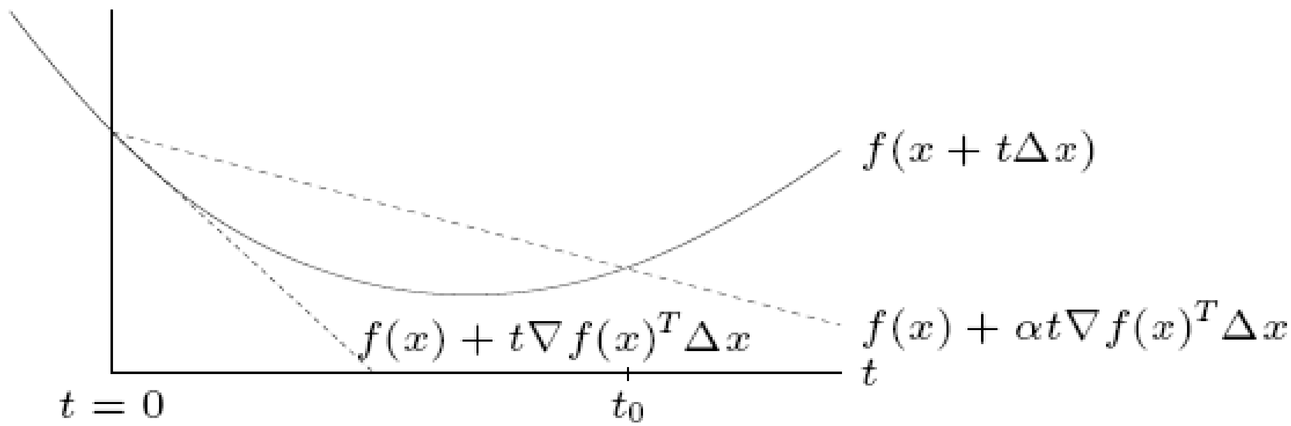
exact line search: $t = \operatorname{argmin}_{t>0} f(x + t\Delta x)$

backtracking line search (with parameters $\alpha \in (0, 1/2)$, $\beta \in (0, 1)$)

- starting at $t = 1$, repeat $t := \beta t$ until

$$f(x + t\Delta x) < f(x) + \alpha t \nabla f(x)^T \Delta x$$

- graphical interpretation: backtrack until $t \leq t_0$



Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Gradient descent method

general descent method with $\Delta x = -\nabla f(x)$

---

given a starting point $x \in \mathbf{dom}\, f$.
repeat
1. $\Delta x := -\nabla f(x)$.
2. *Line search.* Choose step size t via exact or backtracking line search.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

- stopping criterion usually of the form $\|\nabla f(x)\|_2 \leq \epsilon$
- convergence result: for strongly convex f,

$$f(x^{(k)}) - p^\star \leq c^k (f(x^{(0)}) - p^\star)$$

  $c \in (0,1)$ depends on m, $x^{(0)}$, line search type
- very simple, but often very slow; rarely used in practice

Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

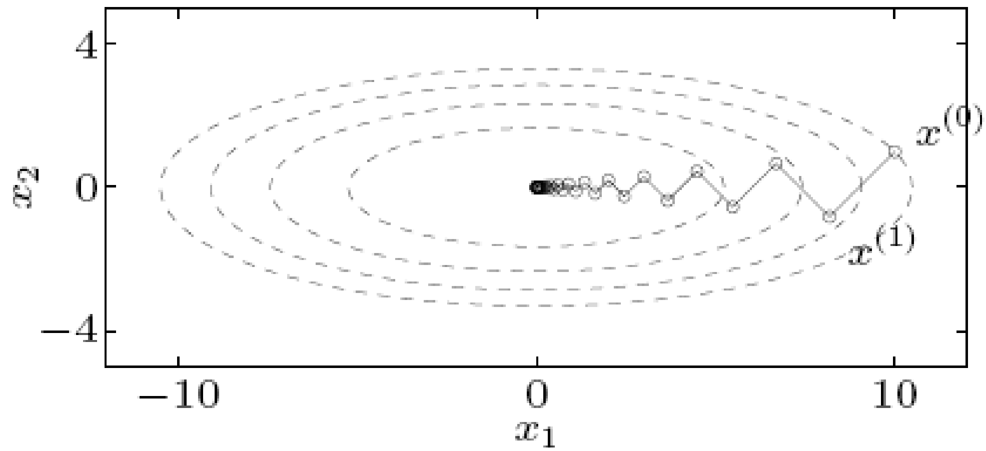
quadratic problem in $\mathbf{R}^2$

$$f(x) = (1/2)(x_1^2 + \gamma x_2^2) \qquad (\gamma > 0)$$

with exact line search, starting at $x^{(0)} = (\gamma, 1)$:

$$x_1^{(k)} = \gamma \left(\frac{\gamma - 1}{\gamma + 1}\right)^k, \qquad x_2^{(k)} = \left(-\frac{\gamma - 1}{\gamma + 1}\right)^k$$

- very slow if $\gamma \gg 1$ or $\gamma \ll 1$
- example for $\gamma = 10$:



Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Steepest descent method

normalized steepest descent direction (at x, for norm $\|\cdot\|$):

$$\Delta x_{\mathrm{nsd}} = \mathrm{argmin}\{\nabla f(x)^T v \mid \|v\| = 1\}$$

interpretation: for small v, $f(x+v) \approx f(x) + \nabla f(x)^T v$;
direction Δx_{nsd} is unit-norm step with most negative directional derivative

(unnormalized) steepest descent direction

$$\Delta x_{\mathrm{sd}} = \|\nabla f(x)\|_* \Delta x_{\mathrm{nsd}}$$

satisfies $\nabla f(x)^T \Delta_{\mathrm{sd}} = -\|\nabla f(x)\|_*^2$

steepest descent method

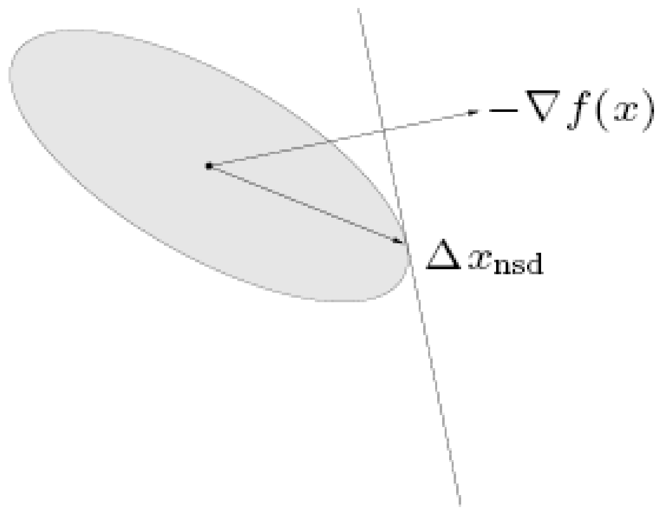
- general descent method with $\Delta x = \Delta x_{\mathrm{sd}}$
- convergence properties similar to gradient descent

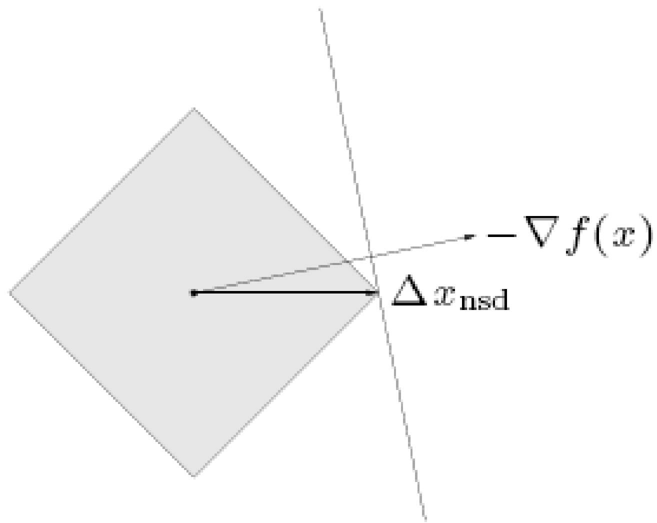
Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

examples

- Euclidean norm: $\Delta x_{\mathrm{sd}} = -\nabla f(x)$
- quadratic norm $\|x\|_P = (x^T P x)^{1/2}$ ($P \in \mathbf{S}^n_{++}$): $\Delta x_{\mathrm{sd}} = -P^{-1}\nabla f(x)$
- ℓ_1-norm: $\Delta x_{\mathrm{sd}} = -(\partial f(x)/\partial x_i)e_i$, where $|\partial f(x)/\partial x_i| = \|\nabla f(x)\|_\infty$

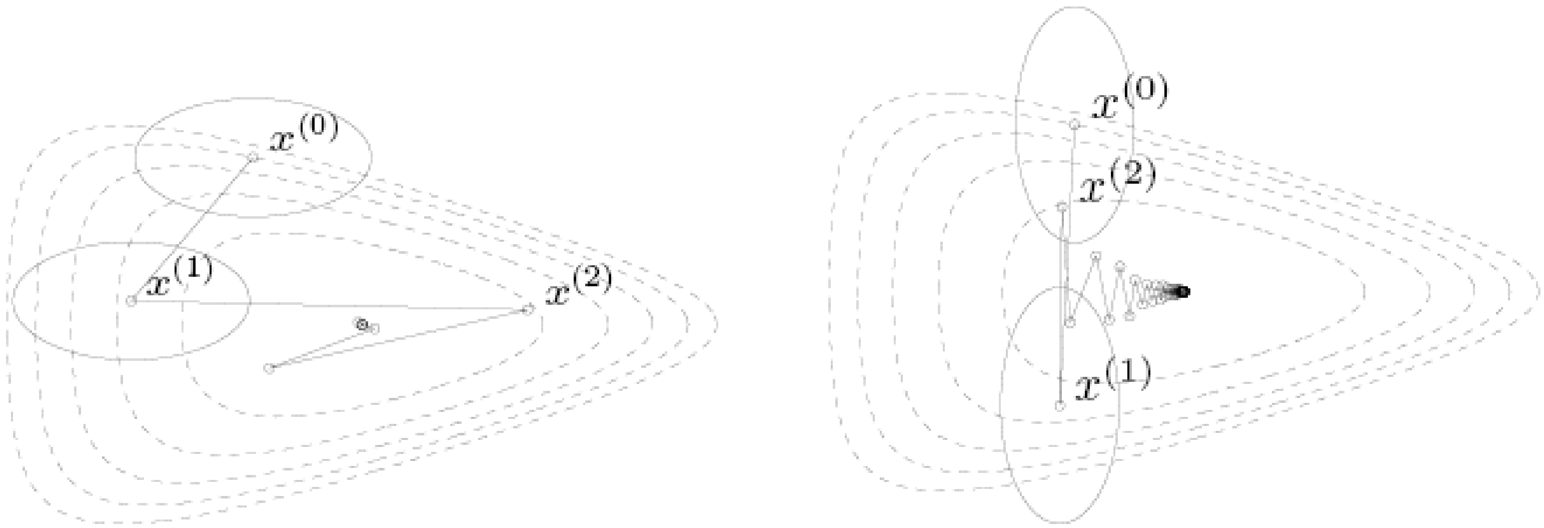
unit balls and normalized steepest descent directions for a quadratic norm and the ℓ_1-norm:





Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

choice of norm for steepest descent



- steepest descent with backtracking line search for two quadratic norms
- ellipses show $\{x \mid \|x - x^{(k)}\|_P = 1\}$
- equivalent interpretation of steepest descent with quadratic norm $\|\cdot\|_P$: gradient descent after change of variables $\bar{x} = P^{1/2}x$

shows choice of P has strong effect on speed of convergence

Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Newton step

$$\Delta x_{\rm nt} = -\nabla^2 f(x)^{-1}\nabla f(x)$$

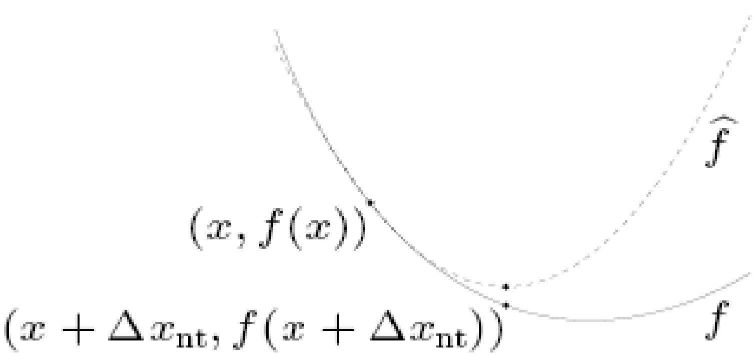
interpretations

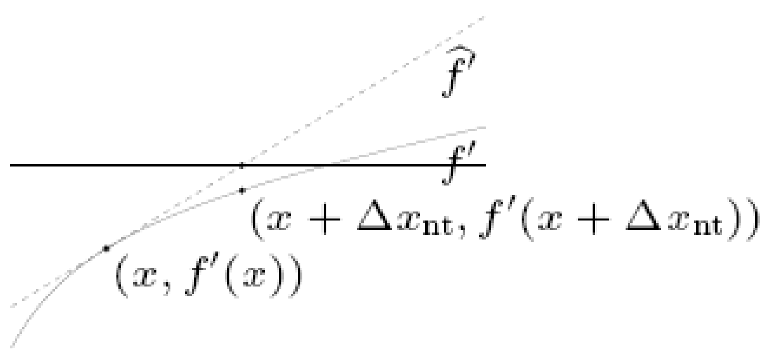
- $x + \Delta x_{\rm nt}$ minimizes second order approximation

$$\widehat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

- $x + \Delta x_{\rm nt}$ solves linearized optimality condition

$$\nabla f(x+v) \approx \nabla \widehat{f}(x+v) = \nabla f(x) + \nabla^2 f(x) v = 0$$

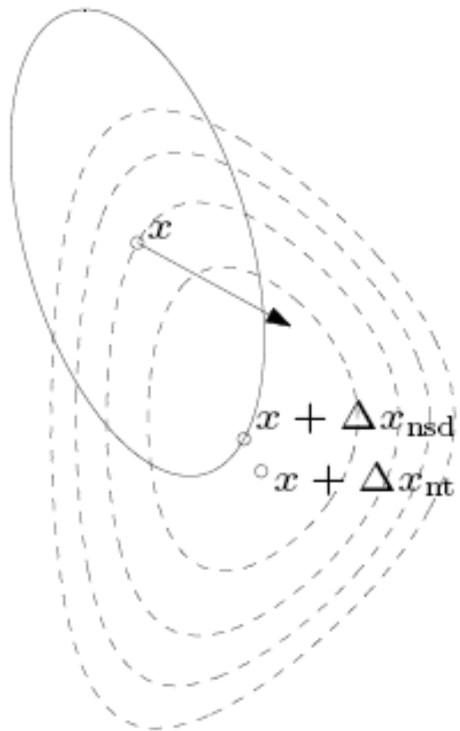




Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

- Δx_{nt} is steepest descent direction at x in local Hessian norm

$$\|u\|_{\nabla^2 f(x)} = \left(u^T \nabla^2 f(x) u\right)^{1/2}$$



dashed lines are contour lines of f; ellipse is $\{x + v \mid v^T \nabla^2 f(x) v = 1\}$

arrow shows $-\nabla f(x)$

Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Newton's method

---

given a starting point $x \in \mathbf{dom}\, f$, tolerance $\epsilon > 0$.
repeat

1. *Compute the Newton step and decrement.*
   $\Delta x_{\mathrm{nt}} := -\nabla^2 f(x)^{-1} \nabla f(x); \quad \lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$.
2. *Stopping criterion.* **quit** if $\lambda^2/2 \leq \epsilon$.
3. *Line search.* Choose step size t by backtracking line search.
4. *Update.* $x := x + t\Delta x_{\mathrm{nt}}$.

---

affine invariant, *i.e.*, independent of linear changes of coordinates:

Newton iterates for $\tilde{f}(y) = f(Ty)$ with starting point $y^{(0)} = T^{-1}x^{(0)}$ are

$$y^{(k)} = T^{-1}x^{(k)}$$

Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Classical convergence analysis

assumptions

- f strongly convex on S with constant m
- $\nabla^2 f$ is Lipschitz continuous on S, with constant $L > 0$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L\|x - y\|_2$$

(L measures how well f can be approximated by a quadratic function)

outline: there exist constants $\eta \in (0, m^2/L)$, $\gamma > 0$ such that

- if $\|\nabla f(x)\|_2 \geq \eta$, then $f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma$
- if $\|\nabla f(x)\|_2 < \eta$, then

$$\frac{L}{2m^2}\|\nabla f(x^{(k+1)})\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2\right)^2$$

damped Newton phase $(\|\nabla f(x)\|_2 \geq \eta)$

- most iterations require backtracking steps
- function value decreases by at least γ
- if $p^\star > -\infty$, this phase ends after at most $(f(x^{(0)}) - p^\star)/\gamma$ iterations

quadratically convergent phase $(\|\nabla f(x)\|_2 < \eta)$

- all iterations use step size $t = 1$
- $\|\nabla f(x)\|_2$ converges to zero quadratically: if $\|\nabla f(x^{(k)})\|_2 < \eta$, then

$$\frac{L}{2m^2}\|\nabla f(x^l)\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^k)\|_2\right)^{2^{l-k}} \leq \left(\frac{1}{2}\right)^{2^{l-k}}, \qquad l \geq k$$

conclusion: number of iterations until $f(x) - p^\star \leq \epsilon$ is bounded above by

$$\frac{f(x^{(0)}) - p^\star}{\gamma} + \log_2 \log_2(\epsilon_0/\epsilon)$$

- γ, ϵ_0 are constants that depend on m, L, $x^{(0)}$
- second term is small (of the order of 6) and almost constant for practical purposes
- in practice, constants m, L (hence γ, ϵ_0) are usually unknown
- provides qualitative insight in convergence properties (*i.e.*, explains two algorithm phases)

Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Self-concordance

shortcomings of classical convergence analysis

- depends on unknown constants (m, L, ...)
- bound is not affinely invariant, although Newton's method is

convergence analysis via self-concordance (Nesterov and Nemirovski)

- does not depend on any unknown constants
- gives affine-invariant bound
- applies to special class of convex functions ('self-concordant' functions)
- developed to analyze polynomial-time interior-point methods for convex optimization

Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Self-concordant functions

definition

- convex $f : \mathbf{R} \to \mathbf{R}$ is self-concordant if $|f'''(x)| \leq 2f''(x)^{3/2}$ for all $x \in \mathbf{dom}\, f$
- $f : \mathbf{R}^n \to \mathbf{R}$ is self-concordant if $g(t) = f(x + tv)$ is self-concordant for all $x \in \mathbf{dom}\, f$, $v \in \mathbf{R}^n$

examples on R

- linear and quadratic functions
- negative logarithm $f(x) = -\log x$
- negative entropy plus negative logarithm: $f(x) = x \log x - \log x$

affine invariance: if $f : \mathbf{R} \to \mathbf{R}$ is s.c., then $\tilde{f}(y) = f(ay + b)$ is s.c.:

$$\tilde{f}'''(y) = a^3 f'''(ay + b), \qquad \tilde{f}''(y) = a^2 f''(ay + b)$$

Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Self-concordant calculus

properties

- preserved under positive scaling $\alpha \geq 1$, and sum
- preserved under composition with affine function
- if g is convex with $\textbf{dom}\, g = \mathbf{R}_{++}$ and $|g'''(x)| \leq 3g''(x)/x$ then

$$f(x) = \log(-g(x)) - \log x$$

  is self-concordant

examples: properties can be used to show that the following are s.c.

- $f(x) = -\sum_{i=1}^m \log(b_i - a_i^T x)$ on $\{x \mid a_i^T x < b_i,\ i = 1, \ldots, m\}$
- $f(X) = -\log\det X$ on $\mathbf{S}_{++}^n$
- $f(x) = -\log(y^2 - x^T x)$ on $\{(x, y) \mid \|x\|_2 < y\}$

Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Interior Point Methods

Interior Point Methods: a history

- Ellipsoid Method, Nemirovskii, 1970's. No complexity result.
- Polynomial Ellipsoid Method for LP, Khachian 1979. Not practical.
- Karmarkar's method, 1984, first "efficient" interior point method.
- Primal-dual path following methods and others late 1980's. Very efficent practical methods.
- Extensions to other classes of convex problems. Early 1990's.
- General theory of interior point methods, self-concordant barriers, Nesterov and Nemirovskii, 1990's.

Self-concordant barrier

$$
\begin{aligned}
\min \quad & c^T x - \mu B_K(x), \\
\text{s.t.} \quad & Ax = b, \\
& x \in \mathbf{R}^n \; x \succ_K 0 \\
& A \in \mathbf{R}^{m \times n}, b \in \mathbf{R}^m.
\end{aligned}
$$

Log barrier for LP

$$
\begin{aligned}
\min \quad & c^T x - \mu \sum_{i=1}^{n} \log x_i, \\
\text{s.t.} \quad & Ax = b, \\
& x \in \mathbf{R}^n \; x > 0 \\
& A \in \mathbf{R}^{m \times n}, b \in \mathbf{R}^m.
\end{aligned}
$$

Log-barrier for SDP

$$\begin{array}{ll} \min & \operatorname{trace}(CX) - \mu \log \det X, \\ \text{s.t.} & \operatorname{trace}(A_i X) = b_i, \;\; i = 1, \ldots, m \\ & X \in \mathbf{S}^n \;\; X \succ 0 \\ & C, A_i \in \mathbf{S}^n, b \in \mathbf{R}^m. \end{array}$$

Log barrier for SOCP

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} {c_i}^\top x_i - \mu \sum_{i=1}^{N} \log((x_i^0)^2 - \|\bar{x}_i\|^2) \\ \text{s.t.} \quad & A_1 x_1 + A_2 x_2 + \ldots + A_N x_N = b, \\ & x_i >_{K_i} 0, \end{aligned}$$

Primal Linear Programming Problem

$$\begin{aligned} \min \quad & c^T x, \\ \text{s.t.} \quad & Ax = b, \\ & x \in \mathbf{R}^n \; x \geq 0 \end{aligned}$$

Dual Linear Programming Problem

$$\begin{aligned} \max \quad & b^T y, \\ \text{s.t.} \quad & A^T y + s = c \\ & s \geq 0 \end{aligned}$$

Optimality (KKT) conditions

$$
\begin{aligned}
&Ax = b \\
&A^\top y + s = c, \\
&x_i s_i = 0, \quad \forall i \\
&x, s \geq 0
\end{aligned}
$$

$x_i s_i = 0 \ \forall i$ - complementarity, $x_i + s_i > 0 \ \forall i$ - strict complementarity.

Central Path

Consider the following "barrier" problem

$$\min c^\top x - \mu \sum_i \ln x_i \quad \text{s.t. } Ax = b,$$

Solution for a given μ

$$(x(\mu), y(\mu), s(\mu))$$

As $\mu \to 0$,

$$(x(\mu), y(\mu), s(\mu)) \to (x^*, y^*, s^*)$$

Apply Newton method to the (self-concordant) barrier problem (i.e. to its optimality conditions)

Apply one or two steps of Newton method for a given μ and then reduce μ

KKT conditions for primal central path

$$\min c^\top x - \mu \sum_i \ln x_i \quad \text{s.t. } Ax = b,$$

$$\begin{aligned} Ax &= b \\ A^\top y + \mu X^{-1} e &= c \\ x, s &> 0 \end{aligned}$$

(where $X = \text{diag}(x)$ and $e = (1, \ldots, 1)^\top$).

$$\begin{aligned} Ax &= b \\ A^\top y + s &= c \\ s &= \mu X^{-1} e \\ x, s &> 0 \end{aligned}$$

Central Path

Consider the following optimization problem

$$\min c^\top x - \mu \sum_i \ln x_i \quad \text{s.t. } Ax = b,$$

Solution for a given μ

$$(x(\mu), y(\mu), s(\mu))$$

As $\mu \to 0$,

$$(x(\mu), y(\mu), s(\mu)) \to (x^*, y^*, s^*)$$

Optimality conditions for the barrier problem ⇒

$$Ax = b$$
$$A^\top y + s = c,$$
$$s_i = \frac{\mu}{x_i}, \quad \forall i$$
$$x, s \geq 0$$

⇒ Apply Newton method to the system of nonlinear equations

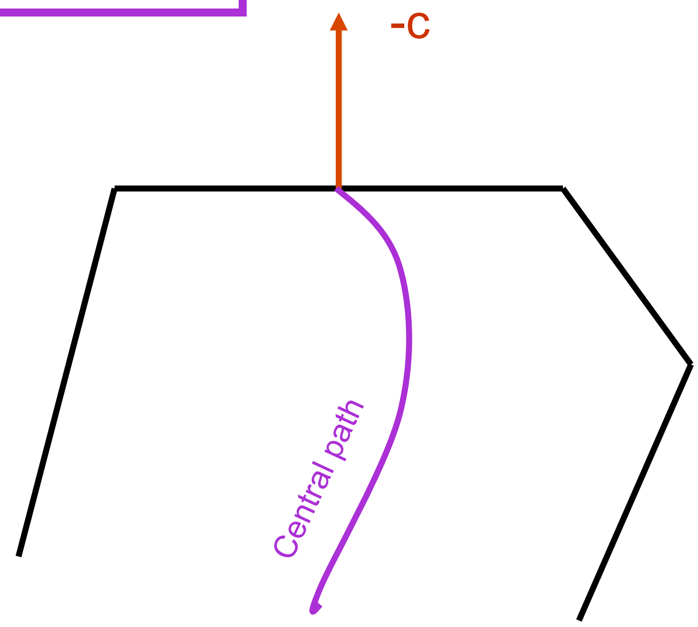
Central Path

$$A x = b$$

$$A^\top y + s = c,$$

$$s_i = \frac{\mu}{x_i}, \quad \forall i$$

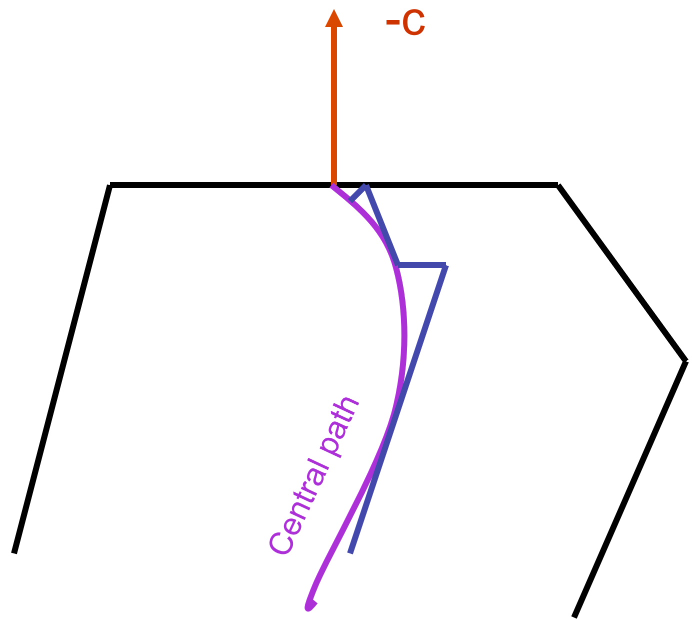
$$x, s \geq 0$$



It exists iff there is nonempty interior
for the primal and dual problems.

Interior point methods, the main idea

- Each point on the central path can be approximated by applying Newton method to the perturbed KKT system.
- Start at some point near the central path for some value of μ, reduce μ.
- Make one or more Newton steps toward the solution with the new value of μ.
- Keep driving μ to 0, always staying close to the solutions of the central path.
- This prevents the iterates from getting trapped near the boundary and keeps them nicely central.


-c
Central path

KKT conditions for dual and primal-dual central paths

$$\max b^\top y + \mu \sum_i \ln s_i \quad \text{s.t. } A^\top y + s = c,$$

$$\begin{aligned} Ax &= b \\ A^\top y + s &= c \\ x &= \mu S^{-1} e \\ x, s &> 0 \end{aligned}$$

(where $S = \text{diag}(s)$ and $e = (1, \ldots, 1)^\top$).

$$\begin{aligned} Ax &= b \\ A^\top y + s &= c \\ Xs &= \mu e \\ x, s &> 0 \end{aligned}$$

Newton step

$$A\Delta x = b - Ax$$
$$A^\top \Delta y + \Delta s = c - A^\top y - s$$

$$\Delta s = -\mu X^{-2} \Delta x$$

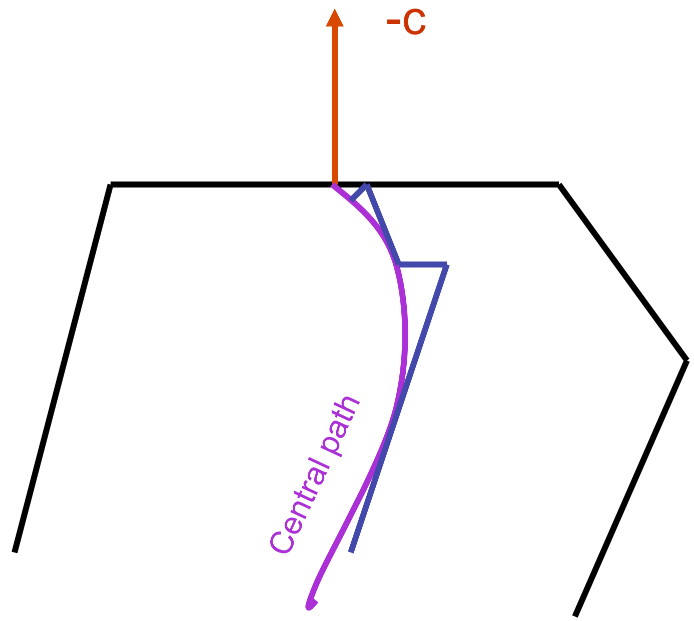
Primal method

$$X\Delta s + S\Delta x = \mu e - Xs$$

Primal-dual method

$$\Delta x = \mu S^{-2} e$$

Dual method



Predictor-Corrector steps

$$A\Delta x = b - Ax$$
$$A^\top \Delta y + \Delta s = c - A^\top y - s$$
$$X\Delta s + S\Delta x = \sigma\mu e - Xs$$

$\sigma = 0$ for predictor step and $\sigma > 0$ for corrector step.

Solve the system of linear equations twice with the same matrix

Predictor-Corrector steps

$$
\begin{aligned}
A\Delta x &= b - Ax \\
A^\top \Delta y + \Delta s &= c - A^\top y - s \\
\Delta s &= \sigma\mu X^{-1}e - Se - X^{-1}S\Delta x
\end{aligned}
$$

$$\Downarrow$$

$$
\begin{aligned}
A\Delta x &= b - Ax \\
A^\top \Delta y - X^{-1}S\Delta x &= c - A^\top y - s - \sigma\mu X^{-1}e + Se
\end{aligned}
$$

$$
\begin{bmatrix} -D & A^\top \\ A & 0 \end{bmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} r_x \\ r_y \end{pmatrix}
$$

Augmented system

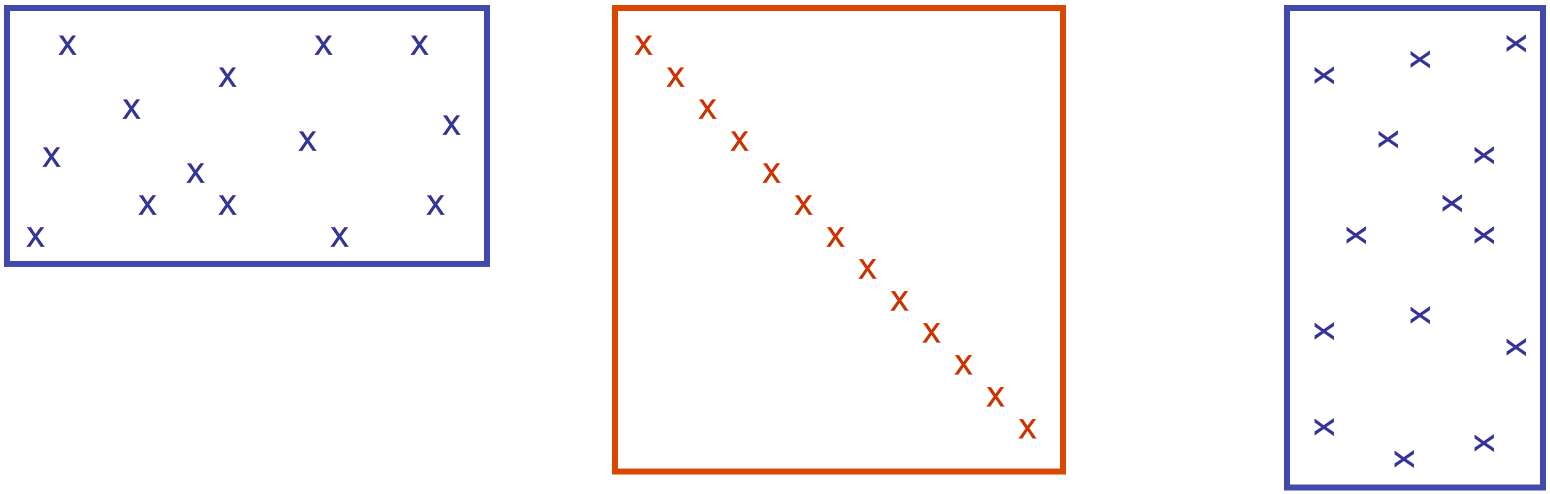
$$D = X^{-1}S \ (\text{or } D = S^{-2} \text{ or } D = X^{-2}).$$

Solving the augmented system

$$\begin{bmatrix} -D & A \\ A^\top & 0 \end{bmatrix} \begin{pmatrix} \Delta y \\ \Delta s \end{pmatrix} = \begin{pmatrix} r_y \\ r_s \end{pmatrix}$$

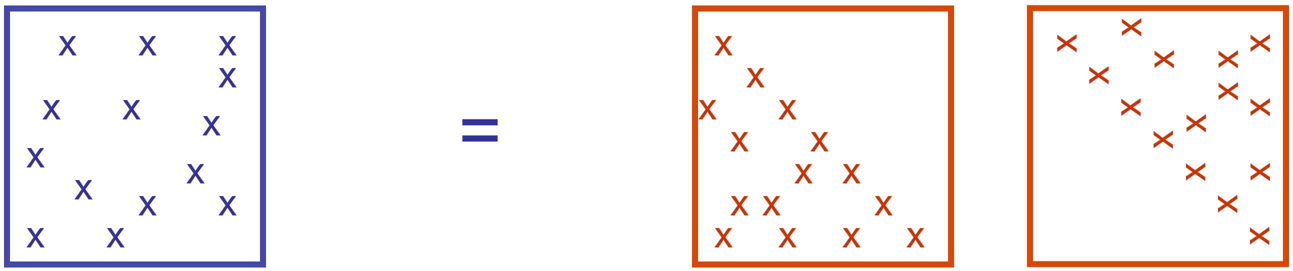
Schur complement system: $AD^{-1}A^\top \Delta y = r.$

Normal equation



Cholesky Factorization

$$AD^{-1}A^\top = LL^\top.$$



- Numerically very stable!
- The sparsity pattern of L remains the same at each iteration
- Depends on sparsity pattern of A and ordering of rows of A
- Can compute the pattern in advance (symbolic factorization)
- The work for each factorization depends on sparsity pattern, can be as little as O(n) if very sparse and as much as O(n^3) (if dense).

Complexity per iteration

- At each iteration form and factorize $AD^{-1}A^\top$, where D is diagonal and G is fixed.

- $A \in \mathbf{R}^{m\times n}$ hence factorizing $AD^{-1}A^\top$ is $O(m^3)$, in general.

- The sparsity structure of $AD^{-1}A^\top$ and its factors is the same at all iterations.

- The work to form $AD^{-1}A^\top \sim$ # of nonzeros in $AD^{-1}A^\top$. The work to factorize $\sim$ # of nonzeros in the Cholesky factor.

Complexity and performance

- Theoretical complexity: $O(\sqrt{n}L)$ iterations for short step methods and $O(nL)$ iteration for long step methods. In practice everyone uses long step methods.
- In practice almost always < 50 iterations, independent of the size.
- In case of multiple solutions converges to the center of the optimal face, not to a vertex.
- Never attains the the exact solution! For LP there are polynomial crossover techniques to obtain an exact vertex from the approximate (central) solution.
- Does not benefit from warm start (not much, anyway)

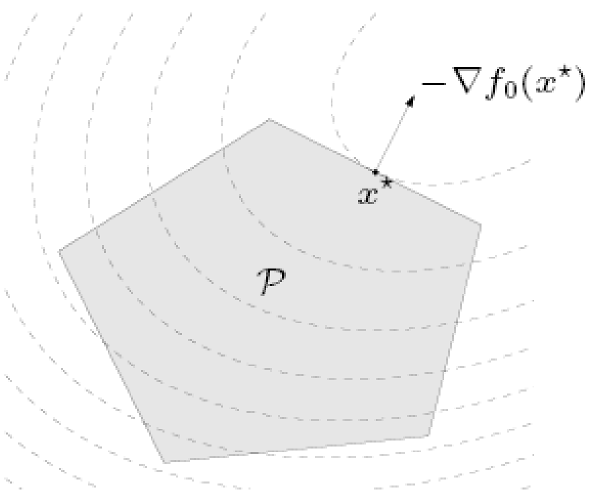
Convex QP with linear inequality constraints.

$$\begin{aligned} \min \quad & x^\top Q x + c^\top x, \\ \text{s.t.} \quad & Ax = b, \\ & x \geq 0, \end{aligned}$$

$$L(x, y) = x^\top Q x + c^\top x - y^\top (Ax - b) - s^\top x$$

Optimality conditions

$$\begin{aligned} & Qx + c - y^\top A - s = 0, \\ & Ax = b, \\ & s_i x_i = 0 \\ & x, s \geq 0 \end{aligned}$$



Interior Point method

Consider the following optimization problem

$$\min \frac{1}{2} x^\top Q x + c^\top x - \mu \sum_i \ln x_i \quad \text{s.t. } Ax = b,$$

$(x(\mu), y(\mu), s(\mu))$ is the central path.

$$\begin{aligned}
&Ax = b \\
&-Qx + A^\top y + s = c \\
&s = \mu X^{-1} \\
&x, s > 0
\end{aligned}$$

or

$$Xs = \mu e$$

Perturb complementarity conditions in a uniform way

Newton Step

$$\begin{aligned} S\Delta x + X\Delta s &= \mu e - Xs \\ A\Delta x &= r_p \\ -Q\Delta x + A^\top \Delta y + \Delta s &= r_d \end{aligned}$$

Augmented system

$$\begin{aligned} A\Delta x &= r_p \\ A^\top \Delta y - (X^{-1}S + Q)\Delta x &= r_d - X^{-1}(\mu e - Xs) \end{aligned}$$

Normal Equation (Schur Complement System)

$$A(X^{-1}S + Q)^{-1}A^\top \Delta y = r$$

Complexity per iteration

- At each iteration form and factorize $(Q + D)$ and $A(Q + D)^{-1}A^\top$, where D is diagonal and G is fixed.

- $A \in \mathbf{R}^{m \times n}$ hence factorizing $(Q + D)$ is $O(n^3)$ and factorizing $A(Q + D)^{-1}A^\top$ is $O(m^3)$, in general.

- The sparsity structure of $A(Q + D)^{-1}A^\top$ and its factors is the same at all iterations.

- The work to form $A(Q + D)^{-1}A^\top \sim$ # of nonzeros in $A(Q + D)^{-1}A^\top$. The work to factorize $\sim$ # of nonzeros in the Cholesky factor. Same for factorizing $Q + D$.

Primal Semidefinite Programming Problem

$$\begin{array}{ll} \min & \operatorname{trace}(CX), \\ \text{s.t.} & \operatorname{trace}(A_i X) = b_i,\ \ i = 1, \ldots, m \\ & X \in \mathbf{S}^n\ \ X \succeq 0 \\ & C, A_i \in \mathbf{S}^n, b \in \mathbf{R}^m. \end{array}$$

SDP cone $K = \{x \in \mathbf{S}^n :\ X \succeq 0\}$ - self dual.

Dual Semidefinite Programming Problem

$$\begin{array}{ll} \max & b^T y, \\ \text{s.t.} & \sum_{i=1}^{m} y_i A_i + S = C \\ & S \succeq 0. \end{array}$$

Duality gap and complementarity

$$A \bullet B = \mathrm{trace}(AB)$$

$$b^T y = \sum_i (A_i \bullet D) y_i = (\sum_i y_i A_i) \bullet D = C \bullet S - S \bullet X$$

Duality Gap:

$$S \bullet X \geq 0$$

Complementarity:

$$XS = SX = 0.$$

Central Path

$$(PCP) \qquad \begin{array}{rl} \min & C \bullet X - \mu(\ln \det X) \\ \text{s.t.} & A_i \bullet X = b_i, \quad i = 1, \ldots, m \\ & X \succ 0 \end{array}$$

Central Path exists **iff** both primal and dual problems have interior solutions

Optimality conditions for (PCP):

$$L(X, y) = C \bullet X - \mu(\ln \det X) - \sum_{i=1}^{m} y_i (A_i \bullet X - b_i)$$

$$\nabla_X L(X, y) = C - \mu X^{-1} - \sum_{i=1}^{m} y_i A_i = 0.$$

Central Path

$C \bullet X - \mu(\ln \det X)$ is strictly convex for $\mu > 0$ thus the solution for (PCP) is unique and satisfies:

$$
(CP) \qquad \begin{aligned}
& S = \mu X^{-1} \\
& A_i \bullet X = b_i, \quad i = 1, \ldots, m \\
& \sum_{i=1}^{m} y_i A_i + S = C, \\
& X, S \succ 0
\end{aligned}
$$

$$
X(\mu) \text{ and } S(\mu) \text{ satisfy (CP)} \;\Rightarrow\; S(\mu) = \mu X(\mu)^{-1} \;\Rightarrow
$$

$$
X(\mu) \bullet S(\mu) = \mu n.
$$

$$
\mu \to 0 \Rightarrow S(\mu) \bullet X(\mu) \to 0.
$$

Central Path

Dual CP $$X = \mu S^{-1}$$

Primal-Dual CP $$XS = \mu I$$

Symmetric Primal-Dual $$\frac{1}{2}(XS + SX) = \mu I$$

Computing a step

Newton step

$$
\begin{aligned}
&X\Delta S + \Delta X S = \mu I - XS \\
&A_i \bullet \Delta X = b_i - A_i \bullet X, \quad i = 1, \dots, m \\
&\sum_{i=1}^{m} \Delta y_i A_i + \Delta S = C - \sum_{i=1}^{m} y_i A_i + S, \\
&X, S \succ 0
\end{aligned}
$$

$$\Delta X + X \Delta S S^{-1} = \mu S^{-1} - X$$

To symmetrize: $\Delta X = -\frac{1}{2}(X \Delta S S^{-1} + S^{-1} \Delta S X) + \mu S^{-1} - X$

Computing a step

The system to solve on each step

$$\begin{bmatrix} -M & A \\ A^\top & 0 \end{bmatrix} \begin{pmatrix} \Delta y \\ \Delta X \end{pmatrix} = \begin{pmatrix} r_y \\ r_x \end{pmatrix}$$

$M = \frac{1}{2}(X \otimes S^{-1} + S^{-1} \otimes X)$

(Kronecker product $A \otimes B = \{A_{ij} B_{kl}\}_{(ijkl)}$)

For dual direction $M = S^{-1} \otimes S^{-1}$.

Cholesky factorization

The normal equaltion matrix to factorize on each step

$$AM^{-1}A^{\top}$$

$M = \frac{1}{2}(X \otimes S^{-1} + S^{-1} \otimes X)$ - $n^2 \times n^2$ almost dense matrix

$M = \frac{1}{2}(S \otimes S)$- $n^2 \times n^2$ sparse (maybe) matrix

$M = \frac{1}{2}(W \otimes W)$ - $n^2 \times n^2$ dense matrix
(W is a symmetric scaling matrix such as $WXW = S$ - Nesterov-Todd).

Each iteration may require $O(n^6)$ operations and $O(n^4)$ memory.

Second Order Cone Programming

$$\begin{array}{ll} \min & {c_1}^\top x_1 + {c_2}^\top x_2 + \ldots + {c_N}^\top x_N \\ \text{s.t.} & A_1 x_1 + A_2 x_2 + \ldots + A_N x_N = b, \\ & x_i \geq_{K_i} 0, \end{array}$$

$$x_i = (x_i^0, \bar{x}_i),\ x_i \geq_{K_i} 0 \Leftrightarrow x_i^0 \geq ||\bar{x}_i||$$

$$\begin{array}{ll} \max & b^\top y \\ \text{s.t.} & {A_i}^\top y + s_i = c_i, \quad i = 1, \ldots, N \\ & s_i \geq_{K_i} 0, \end{array}$$

$A_i \in \mathbf{R}^{m \times n_i}$, $c_i \in \mathbf{R}^{n_i}$, $x_i \in \mathbf{R}^{n_i}$, $s_i \in \mathbf{R}^{n_i}$, $i = 1, \ldots, N$, $b \in \mathbf{R}^m$ $y \in \mathbf{R}^m$.
$A = [A_1, A_2, \ldots, A_N]$, $x = ({x_1}^\top, {x_2}^\top, \ldots, {x_N}^\top)^\top$ and $s = ({s_1}^\top, {s_2}^\top, \ldots, {s_N}^\top)^\top$.

Complementarity Conditions

$$\begin{aligned} x_i^0 s_i^0 + \bar{x}_i^\top \bar{s}_i &= 0 \quad i = 1, \ldots, N \\ s_i^0 \bar{x}_i + x_i^0 \bar{s}_i &= 0, \quad i = 1, \ldots, N \end{aligned}$$

If we define an "arrow-shaped" matrix $\mathbf{Arr}(x_i)$ as

$$\mathbf{Arr}(x_i) = \begin{bmatrix} x_i^0 & x_i^1 & \ldots & x_i^{n_i} \\ x_i^1 & x_i^0 & & \\ \vdots & & \ddots & \\ x_i^{n_i} & & & x_i^0 \end{bmatrix},$$

and the block diagonal matrix $\mathbf{Arr}(x)$ as

$$\mathbf{Arr}(x) = \begin{bmatrix} \mathbf{Arr}(x_1) & & & \\ & \mathbf{Arr}(x_2) & & \\ & & \ddots & \\ & & & \mathbf{Arr}(x_N) \end{bmatrix},$$

then the complementarity conditions can be expressed as

$$\mathbf{Arr}(x)s = \mathbf{Arr}(s)x = \mathbf{Arr}(x)\mathbf{Arr}(s)e_0 = 0,$$

where

$$e^{0^T} = (e_1^{0^T}, e_2^{0^T}, \ldots, e_N^{0^T}) \equiv (\underbrace{1, 0, \ldots, 0}_{n_1}, \underbrace{1, 0, \ldots, 0}_{n_2}, \ldots, \underbrace{1, 0, \ldots, 0}_{n_N})^\top.$$

Log-barrier formulation

$$\begin{aligned} \min \quad & c^\top x + \mu \sum_{i=1}^{N} \ln((x_i^0)^2 - \|\bar{x}_i\|^2) \\ \text{s.t.} \quad & Ax = b, \\ & x_i \geq_{K_i} 0, \end{aligned}$$

Perturbed optimality conditions

$$\begin{aligned} x_i^0 s_i^0 + \bar{x}_i^\top \bar{s}_i &= \mu \quad i = 1, \ldots, N \\ s_i^0 \bar{x}_i + x_i^0 \bar{s}_i &= 0, \quad i = 1, \ldots, N \end{aligned}$$

The optimality conditions

$$\begin{aligned} &Ax = b \\ &A^\top y + s = c \\ &\mathbf{Arr}(x)s = \mathbf{Arr}(s)x = \mathbf{Arr}(x)\mathbf{Arr}(s)e_0 = \mu e_0, \end{aligned}$$

where

$$e^{0^T} = (e_1^{0^T}, e_2^{0^T}, \ldots, e_N^{0^T}) \equiv (\underbrace{1, 0, \ldots, 0}_{n_1}, \underbrace{1, 0, \ldots, 0}_{n_2}, \ldots, \underbrace{1, 0, \ldots, 0}_{n_N})^\top.$$

Newton step

$$\mathbf{Arr}(x)\Delta s + \mathbf{Arr}(s)\Delta x = \mu e_0 - \mathbf{Arr}(x)\mathbf{Arr}(s)e_0,$$
$$A\Delta x = b - Ax,$$
$$A^\top \Delta y + \Delta s = c - A^\top y - s$$

$$\begin{bmatrix} -F & A \\ A^\top & 0 \end{bmatrix} \begin{pmatrix} \Delta y \\ \Delta x \end{pmatrix} = \begin{pmatrix} r_y \\ r_s \end{pmatrix}$$

$$F = \mathbf{Arr}(x)^{-1}\mathbf{Arr}(s),\ F^{-1} = \mathbf{Arr}(s)^{-1}\mathbf{Arr}(x),$$

$$(\mathbf{Arr}(x_i))^{-1} = \frac{1}{\gamma^2(x_i)} \begin{bmatrix} x_i^0 & -\bar{x}_i{}^\top \\ -\bar{x}_i & \frac{\gamma^2(x_i)}{x_0} I - \bar{x}_i \bar{x}_i{}^\top \end{bmatrix},$$

$$\gamma(x_i) = \sqrt{(x_i^0)^2 - \|\bar{x}_i\|^2}.$$

Optimization methods for convex problems

- Interior Point methods
  - Best iteration complexity O(log(1/ϵ)), in practice <50.
  - Worst per-iteration complexity (sometimes prohibitive)
- Active set methods
  - Exponential complexity in theory, often linear in practice.
  - Better per iteration complexity.
- Gradient based methods
  - $O(1/\sqrt{\epsilon})$ or O(1/ϵ) iterations
  - Matrix/vector multiplication per iteration
- Nonsmooth gradient based methods
  - O(1/ϵ) or O(1/ϵ^2) iterations
  - Matrix/vector multiplication per iteration
- Block coordinate descent
  - Iteration complexity ranges from unknown to similar to FOMs.
  - Per iteration complexity can be constant.

Homework

1.

Given a matrix $M = \begin{bmatrix} M_{11} & \dots & M_{1m} \\ \vdots & \ddots & \vdots \\ M_{n1} & \dots & M_{nm} \end{bmatrix} \in \mathbf{R}^{n\times m}$ prove

- $\|M\|_2 = \sigma_{max}$ - where σ_{max} is the largest singular value of M.
- $\|M\|_1 = \max_j \sum_{i=1}^{n} |M_{ij}|$ - matrix l_1-norm
- $\|M\|_\infty = \max_i \sum_{j=1}^{m} |M_{ij}|$ - l_∞-norm

2. Let cone $K = \{(x,t) : ||x||_1 \leq t\}$. Prove that $K^* = \{(x,t) : ||x||_\infty \leq t\}$.

3. Prove for two symmetric matrices X and S that if $\text{trace}(XS) = 0$, $X \succeq 0$ and $S \succeq 0$ then $XS = SX = 0$. .