Convex Optimization

Lecture 1: Introduction

A. Mathematical Optimization

(1) (mathematical) optimization problem

$$\text{minimize } f_0(x)$$
$$\text{subject to } f_i(x) \le b_i \quad i=1,\dots,m$$

$x = (x_1, \dots, x_n)$: optimization variables

$f_0 : \mathbb{R}^n \to \mathbb{R}$: objective function

$f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \dots, m$ constraint functions

optimal solution: x^* has smallest value of f_0 among all vectors that satisfy the constraints.

(2)

B. Least-squares and Linear Programming

(1) <u>Least-squares</u>

$$\text{minimize } \|Ax - b\|_2^2$$

(2) solving least-squres problems

* analytical solution: $x^* = (A^TA)^{-1}A^Tb$
* computation time proportional to n^2k ($A \in \mathbb{R}^{k \times n}$); less if structured

(3) <u>Linear Programming</u>

$$\text{minimize } c^Tx$$
$$\text{subject to } a_i^Tx \le b_i, \quad i=1,\dots,m$$

(4) solving linear programming

* no analytical formula for solution
* computation time proportional to n^2m if $m \ge n$; less with structure

C. Convex optimization

(1) convex optimization problem

$$\text{minimize } f_0(x)$$
$$\text{subject to } f_i(x) \le b_i, \quad i = 1, \dots, m$$

* objective and constraint functions are convex:

$$f_i(\alpha x + \beta y) \le \alpha f_i(x) + \beta f_i(y)$$

if $\alpha + \beta = 1$, $\alpha \ge 0$, $\beta \ge 0$

(2) solving convex optimization problems

* no analytical solution
* computation time (roughly) proportional to $\max\{n^3, n^2 m, F\}$, where F is cost of evaluating f_i's and their first and second derivatives.

D. Nonlinear optimization

(1) <u>local optimization methods</u> (nonlinear programming)

* find a point that minimizes f_0 among feasible points near it.
* fast and can handle large problems.
* require initial guess.
* provide no info about distance to global optimum.

(2) <u>global optimization methods</u>

* find the global solution
* worst-case complexity grows exponentially with problem size.

Lecture 2: Convex Sets

A. Affine and convex sets

(1) <u>Affine set</u>: contains the line through any two distinct points in the set.

(2) <u>Convex set</u>: contains line segment between any two points in the set.

(3) <u>Convex combination</u> of $x_1 \dots x_k$: any point x of the form

$$x = \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_k x_k$$

with $\theta_1 + \dots + \theta_k = 1$, $\theta_i \geq 0$

(4) <u>Convex hull conv S</u>: set of all convex combinations of points in S.

(5) <u>conic (nonnegative) combination</u> of x_1 and x_2: any point of the form

$$x = \theta_1 x_1 + \theta_2 x_2 \quad \text{with } \theta_i \geq 0, \theta_2 \geq 0$$

(6) Convex cone: set that contains all conic combinations of points in the set.

B. Some important examples

(1) <u>Hyperplane</u>: set of the form $\{x \mid a^T x = b\}$ $(a \neq 0)$

(2) <u>Halfspace</u>: set of the form $\{x \mid a^T x \leq b\}$ $(a \neq 0)$

a is the normal vector

convex

affine and convex

(3) <u>Euclidean balls</u> with center x_c and radius r:

$$B(x_c, r) = \{x \mid \|x - x_c\|_2 \leq r\} = \{x_c + ru \mid \|u\|_2 \leq 1\}$$

(4) <u>Ellipsoid</u>: set of the form

$$\{x \mid (x - x_c)^T P^{-1} (x - x_c) \leq 1\}$$

with $P \in S^n_{++}$ (i.e. P symmetric positive definite)

semiaxis of ellipsoid are the eigenvectors of P, radius are square roots of eigenvalues

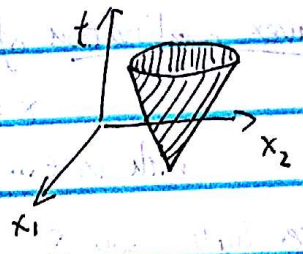
other representation: $\{x_c + Au \mid \|u\|_2 \le 1\}$ with A square and nonsingular

(5) <u>Norm</u>: a function $\|\cdot\|$ that satisfies

* $\|x\| \ge 0$, $\|x\| = 0$ iff $x = 0$
* $\|tx\| = |t|\,\|x\|$ for $t \in \mathbb{R}$
* $\|x+y\| \le \|x\| + \|y\|$

(6) <u>Norm Cone</u>: $\{(x,t) \mid \|x\| \le t\}$

convex



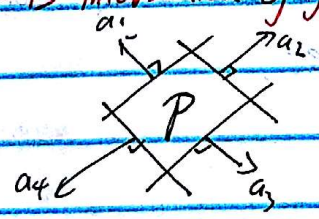
(7) <u>Polyhedra</u>: solution set of finitely many linear inequalities and equalities

componentwise inequality

$Ax \preceq b \qquad Cx = d.$

$A \in \mathbb{R}^{m \times n}$, $C \in \mathbb{R}^{p \times n}$,

polyhedron is intersection of finite number of halfspaces and hyperplanes



(8) Positive semidefinite cone

* S^n is set of symmetric $n \times n$ matrices
* $S^n_+ = \{X \in S^n \mid X \succeq 0\}$: positive semidefinite $n \times n$ matrices

$$X \in S^n_+ \iff z^T X z \ge 0 \text{ for all } z$$

* $S^n_{++} = \{X \in S^n \mid X \succ 0\}$: positive definite $n \times n$ matrices

C. Operations that preserve convexity

(1) <u>Intersection</u>: the intersection of (any number of) ~~cove~~ convex sets is convex

(2) <u>Affine function</u>

Suppose $f: \mathbb{R}^n \to \mathbb{R}^m$ is affine ($f(x) = Ax + b$ with $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$).

* the image of a convex set under f is convex

$$S \subseteq \mathbb{R}^n \text{ convex} \Rightarrow f(S) = \{f(x) \mid x \in S\} \text{ convex}$$

* the inverse image $f^{-1}(C)$ of a convex set under f is convex

$$C \subseteq \mathbb{R}^m \text{ convex} \Rightarrow f^{-1}(C) = \{x \in \mathbb{R}^n \mid f(x) \in C\} \text{ convex}$$

Examples:

* scaling, translation, projection
* solution set of linear matrix inequality $\{x \mid x_1 A_1 + \cdots x_m A_m \preceq B\}$ (with $A_i, B \in S^p$)
* hyperbolic cone $\{x \mid x^T P x \le (c^T x)^2, c^T x \ge 0\}$ (with $P \in S^n_+$)

(3) <u>Perspective and linear-fractional function</u>

<u>perspective function</u> $P: \mathbb{R}^{n+1} \to \mathbb{R}^n$:

$$P(x, t) = x/t \qquad \mathrm{dom}\, P = \{(x,t) \mid t > 0\}$$

images and inverse images of convex sets under $\begin{cases}\text{perspective} \\ \text{linear-fractional}\end{cases}$ are convex

linear-fractional function $f: \mathbb{R}^n \to \mathbb{R}^m$:

$$f(x) = \frac{Ax + b}{c^T x + d} \qquad \mathrm{dom} f = \{x \mid c^T x + d > 0\}$$

D. Generalized inequalities.

(1) a convex cone $K \subseteq \mathbb{R}^n$ is a <u>proper cone</u> if

* K is closed (contains its boundary)
* K is solid (has nonempty interior)
* K is pointed (contains no line)

(2) <u>generalized inequality</u> : defined by a proper cone K:

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

$$x \prec_K y \iff y - x \in \operatorname{int} K$$ → interior of the cone

$\preceq_K$ is not in general a linear ordering: we can have $x \not\preceq_K y$ and $y \not\preceq_K x$ (incomparable)

(3) ~~minimu~~ <u>minimum elements</u>

$$x \in S, \forall y \in S \Rightarrow x \preceq_K y$$

(4) <u>minimal elements</u>

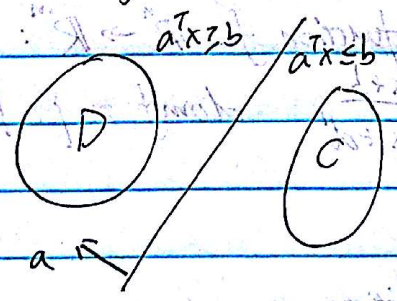
$$x \in S, \forall y \in S, y \preceq_K x \Rightarrow y = x$$

E. Separating and supporting hyperplanes

(1) <u>Separating hyperplane theorem</u>

if C and D are disjoint convex sets, then there exists $a \neq 0$, b such that

$$a^T x \le b \text{ for } x \in C, \quad a^T x \ge b \text{ for } x \in D$$



the hyperplane $\{x \mid a^T x = b\}$ separates C and D

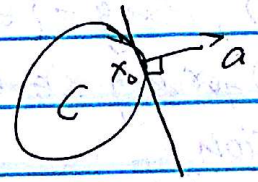
(2) <u>Supporting hyperplane theorem</u>

supporting hyperplane to set C at boundary point x_0:

$$\{x \mid a^T x = a^T x_0\}$$

where $a \neq 0$ and $a^T x \le a^T x_0$ for all $x \in C$

supporting hyperplane theorem: if C is convex, then there exists a supporting hyperplane at every boundary point of C.



F. Dual cones and generalized inequalities

(1) dual cone of a cone K:

$$K^* = \{y \mid y^T x \geq 0 \text{ for all } x \in K\}$$

dual cones of proper cones are proper, hence define generalized inequalities:

$$y \succeq_{K^*} 0 \iff y^T x \geq 0 \text{ for all } x \succeq_K 0$$

(2) <u>Minimum element</u> w.r.t. $\preceq_K$

x is ~~minimal~~ minimum element of S iff for all $\lambda \succ_{K^*} 0$, x is the unique minimizer of $\lambda^T z$ over S.

(3) <u>Minimal element</u> w.r.t. $\preceq_K$

* if x minimizes $\lambda^T z$ over S for some $\lambda \succ_{K^*} 0$, then x is minimal.
* if x is a minimal element of a convex set S, then there exists a nonzero $\lambda \succeq_{K^*} 0$ such that x minimizes $\lambda^T z$ over S.

Lecture 3: Convex Function

A. Basic properties and examples

(1) Definition: $f: \mathbb{R}^n \to \mathbb{R}$ is convex if dom f is a convex set and

$$f(\theta x + (1-\theta) y) \leq \theta f(x) + (1-\theta) f(y)$$

for all $x, y \in$ dom f, $0 \leq \theta \leq 1$

(2) Restriction of a convex function to a line

$f: \mathbb{R}^n \to \mathbb{R}$ is convex if and only if the function $g: \mathbb{R} \to \mathbb{R}$

$$g(t) = f(x + tv), \quad \mathrm{dom}\, g = \{t \mid x + tv \in \mathrm{dom}\, f\}$$

is convex ~~(in t)~~ (in t) for any $x \in \mathrm{dom}\, f$, $v \in \mathbb{R}^n$

(3) <u>Extended-value extension</u>

$\tilde{f}$ of f is

$$\tilde{f}(x) = f(x), \; x \in \mathrm{dom}\, f, \quad \tilde{f}(x) = \infty, \; x \notin \mathrm{dom}\, f$$

(4) <u>First-order condition</u>

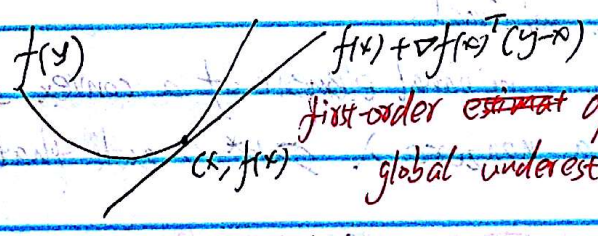
f is differentiable if $\mathrm{dom}\, f$ is open and the gradient

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \cdots, \frac{\partial f(x)}{\partial x_n} \right)$$

exists at each $x \in \mathrm{dom}\, f$

<u>First-order condition</u>: differentiable f with convex domain is convex iff

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y \in \mathrm{dom}\, f$$



first-order ~~estimat~~ approximation of f is global underestimator

(5) ~~Sec~~ <u>Second-order condition</u>

f is twice differentiable if $\mathrm{dom}\, f$ is open and the Hessian $\nabla^2 f(x) \in S^n$,

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i, j = 1, \ldots, n.$$

exists at each $x \in \mathrm{dom}\, f$

~~2-nd~~ <u>2nd-order condition</u> for twice differentiable f with convex domain

* f is convex iff $\nabla^2 f(x) \succeq 0$ for all $x \in \mathrm{dom}\, f$
* if $\nabla^2 f(x) \succ 0$ for all $x \in \mathrm{dom}\, f$, then f is strictly convex

(6) Epigraph and sublevel set

<u>α-sublevel set</u> of $f: \mathbb{R}^n \to \mathbb{R}$

$$C_\alpha = \{x \in \mathrm{dom}\, f \mid f(x) \le \alpha\}$$

sublevel sets of convex functions are convex

<u>epigraph</u> of $f: \mathbb{R}^n \to \mathbb{R}$

$$\mathrm{epi}\, f = \{(x, t) \in \mathbb{R}^{n+1} \mid x \in \mathrm{dom}\, f,\ f(x) \le t\}$$

f is convex iff epi f is a convex set

(7) <u>Jensen's inequality</u>

basic inequality: if f is convex, then for $0 \le \theta \le 1$

$$f(\theta x + (1-\theta) y) \le \theta f(x) + (1-\theta) f(y)$$

extension: if f is convex, then

$$f(E z) \le E f(z)$$

for any random variable z

if takes infinite sums of variables, it turns to be

B. Operations that preserve convexity

(1) Positive weighted sum & composition with affine function

<u>nonnegative multiple</u>: αf is convex if f is convex, $\alpha \ge 0$

<u>sum</u>: $f_1 + f_2$ convex if f_1, f_2 convex

<u>composition with affine function</u>: $f(Ax+b)$ is convex if f is convex

(2) <u>Pointwise maximum</u>: if $f_1, \dots, f_m$ are convex, then

$$f(x) = \max\{f_1(x), \dots, f_m(x)\} \text{ is convex.}$$

(3) <u>Pointwise Supremum</u>: if $f(x,y)$ is convex in x for each $y \in A$, then

$$g(x) = \sup_{y \in A} f(x,y) \quad \text{is convex}$$

(4) <u>Composition with scalar functions</u>

composition of $g: \mathbb{R}^n \to \mathbb{R}$ and $h: \mathbb{R} \to \mathbb{R}$:

$$f(x) = h(g(x))$$

extented-value extension

$\tilde{h}$ tilde

f is convex if $\begin{cases} g \text{ convex}, h \text{ convex}, \tilde{h} \text{ nondecreasing} \\ g \text{ concave}, h \text{ convex}, \tilde{h} \text{ nonincreasing} \end{cases}$

* proof (for $n=1$) (differentiable g, h)

$$f''(x) = h''(g(x))g'(x)^2 + h'(g(x))g''(x)$$

* note: monotonicity must hold for extended-value extension $\tilde{h}$.

(5) <u>Vector Composition</u>

composition of $g: \mathbb{R}^n \to \mathbb{R}^k$ and $h: \mathbb{R}^k \to \mathbb{R}$

$$f(x) = h(g(x)) = h(g_1(x), g_2(x), \dots, g_k(x))$$

f is convex if $\begin{cases} g_i \text{ convex}, h \text{ convex}, \tilde{h} \text{ nondecreasing in each argument} \\ g_i \text{ concave}, h \text{ convex}, \tilde{h} \text{ nonincreasing in each argument} \end{cases}$

proof (for $n=1$, differential g, h)

$$f''(x) = g'(x)^T \nabla^2 h(g(x)) g'(x) + \nabla h(g(x))^T g''(x)$$

(6) <u>Minimization</u>

if $f(x,y)$ is convex in (x,y) and C is a convex set, then $g(x) = \inf_{y \in C} f(x,y)$ is convex

(7) <u>Perspective</u>

the perspective of a function $f: \mathbb{R}^n \to \mathbb{R}$ is the function

$$g: \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}, \quad g(x,t) = tf(x/t),$$

$$\mathrm{dom}\, g = \{(x,t) \mid x/t \in \mathrm{dom}\, f,\ t>0\}$$

g is convex if f is convex

C. The Conjugate Function

(1) the <u>conjugate</u> of a function f is

$$f^*(y) = \sup_{x \in \mathrm{dom} f} (y^T x - f(x))$$

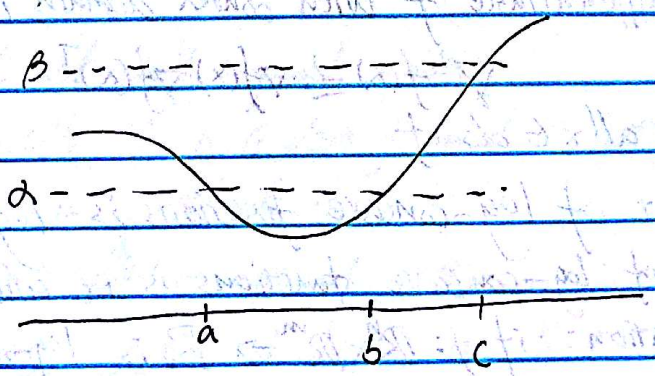
f^* is convex (even if f is not convex)

D. Quasiconvex Functions

(1) $f: \mathbb{R}^n \to \mathbb{R}$ is <u>quasiconvex</u> if $\mathrm{dom} f$ is convex and sublevel sets S_α

$$S_\alpha = \{x \in \mathrm{dom} f \mid f(x) \le \alpha\}$$

are convex for all α



* f is ~~quasiconvex if~~ quasiconcave if $-f$ is quasiconvex
* f is quasilinear if it is quasiconvex and quasiconcave

(2) ~~internal rate of return~~ Properties

* modified Jensen inequality: for quasiconvex f

$$0 \le \theta \le 1 \Rightarrow f(\theta x + (1-\theta) y) \le \max\{f(x), f(y)\}$$

* first-order condition: differentiable f with cvx domain is quasiconvex iff

$$f(y) \le f(x) \Rightarrow \nabla f(x)^T (y-x) \le 0$$

* sums of quasiconvex functions are not necessarily quasiconvex

E. Log-concave and log-convex functions

(1) a positive function f is log-concave if $\log f$ is concave:

$$f(\theta x + (1-\theta)y) \ge f(x)^{\theta} f(y)^{1-\theta} \quad \text{for } 0 \le \theta \le 1$$

f is log-convex if $\log f$ is convex

* powers: x^a on $\mathbb{R}_{++}$ is log-convex for $a \le 0$, ~~log-co~~ log-concave for $a \ge 0$

* many common probability densities are log-concave, e.g. normal: $f(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2}(x-\bar{x})^T \Sigma^{-1} (x-\bar{x})}$

(2) Properties of log-concave functions

* twice differentiable f with convex domain is log-concave iff

$$\nabla^2 f(x) \preceq \nabla f(x) \nabla f(x)^T$$

for all $x \in \mathrm{dom} f$

* product of log-concave functions is log-concave
* sum of log-concave functions is not alway log-concave
* integration: if $f: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is log-concave, then

$$g(x) = \int f(x, y)\, dy$$

is log-concave

F. Convexity with respect to generalized inequalities

(1) $f: \mathbb{R}^n \to \mathbb{R}^m$ is K-convex if dom f is convex, and

$$f(\theta x + (1-\theta)y) \preceq_K \theta f(x) + (1-\theta) f(y)$$

for $x, y \in \text{dom } f$, $0 \le \theta \le 1$

Lecture 4: Convex Optimization Problems

A. Optimization problem in standard form

(1)
$$\begin{aligned} &\text{minimize} && f_0(x) \\ &\text{subject to} && f_i(x) \le 0, \quad i = 1, \dots, m \\ & && h_i(x) = 0, \quad i = 1, \dots, p \end{aligned}$$

* $x \in \mathbb{R}^n$ is the optimization variable
* $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the objective or cost function
* $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \dots, m$, are the inequality constraint functions
* $h_i : \mathbb{R}^n \to \mathbb{R}$ are the equality constraint functions.

(2) <u>Optimal Value</u>: $p^* = \inf \{ f_0(x) \mid f_i(x) \le 0, i = 1, \dots m, h_i(x) = 0, i = 1 \dots p \}$

* $p^* = \infty$ if problem is infeasible (no x satisfies the constraints)
* $p^* = -\infty$ if problem is unbounded below

(3) Optimal and locally optimal points

* x is feasible if $x \in \text{dom } f_0$ and it satisfies the constraints
* a feasible x is <u>optimal</u> if $f_0(x) = p^*$
* X_{opt} is the set of optimal points
* x is <u>locally optimal</u> if there is an $R > 0$ such that x is optimal for

$$\begin{aligned} &\text{minimize (over } z) && f_0(z) \\ &\text{subject to} && f_i(z) \le 0, i = 1, \dots m, \quad h_i(z) = 0, i = 1 \dots p \\ & && \|z - x\|_2 \le R \end{aligned}$$

(4) <u>Implicit Constraints</u>: the standard form optimization problem has an implicit constraint:

$$x \in D = \bigcap_{i=0}^{m} \text{dom } f_i \cap \bigcap_{i=1}^{p} \text{dom } h_i$$

* D is called as the domain of the problem
* the constraints $f_i(x) \le 0$, $h_i(x) = 0$ are the explicit constraints
* a problem is unconstrained if it has no explicit constraints ($m = p = 0$)

(5) <u>feasibility problem</u>:

$$\begin{aligned} &\text{find } x \\ &\text{subject to } f_i(x) \le 0,\ i = 1,\dots,m,\quad h_i(x) = 0,\ i = 1..p. \end{aligned}$$

B. Convex Optimization Problems

(1) standard form:

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0,\ i = 1\dots m \\ & a_i^T x = b_i,\ i = 1,\dots,p \end{aligned}$$

* f_0, f_i, ($i = 1..m$) are convex; equality constraints are affine
* problem is quasiconvex if f_0 is quasiconvex (and f_i convex)
* ~~of~~ often written as

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0 \quad i = 1\dots m \\ & Ax = b \end{aligned}$$

important property: feasible set of a convex optimization problem is convex

(2) Local and global optima

any locally optimal point of a convex problem is (globally) optimal

~~proof:~~

(3) Optimality criterion for differentiable f_0

x is optimal iff it is feasible and

$$\nabla f_0(x)^T (y - x) \ge 0 \text{ for all feasible } y$$

(4) <u>unconstrained problem</u>: x is optimal iff $x \in \mathrm{dom} f_0$, $\nabla f_0(x) = 0$

(5) ~~equ~~ <u>equality constrained problem</u>: minimize $f_0(x)$ s.t. $Ax = b$

x is optimal iff there exists a ν such that

$$x \in \mathrm{dom} f_0, \quad Ax = b, \quad \nabla f_0(x) + A^T \nu = 0$$

(6) <u>minimization over nonnegative orthant</u>:

$$\text{minimize } f_0(x) \quad \text{s.t. } x \succeq 0$$

x is optimal iff

$$x \in \mathrm{dom} f_0, \quad x \succeq 0, \quad \begin{cases} \nabla f_0(x)_i \geq 0 & x_i = 0 \\ \nabla f_0(x)_i = 0 & x_i > 0 \end{cases}$$

(7) <u>Equivalent convex problems</u>

two problem are (informally) equivalent if the solution of one is readily obtained from the solution of the other, and vice-versa

some common transformations that preserve convexity:

* eliminating equality constraints

$$\begin{aligned} &\text{minimize } f_0(x) \\ &\text{s.t. } f_i(x) \leq 0, \quad i = 1 \ldots m \\ &\qquad Ax = b \end{aligned}$$

is equivalent to

$$\begin{aligned} &\text{minimize (over } z) \ f_0(Fz + x_0) \\ &\text{s.t. } f_i(Fz + x_0) \leq 0, \quad i = 1 \ldots m \end{aligned}$$

where F and x_0 are such that

$$Ax = b \iff x = Fz + x_0 \text{ for some } z$$

* introducing equality constraints

$$\begin{aligned} &\text{minimize } f_0(A_0 x + b_0) \\ &\text{s.t. } f_i(A_i x + b_i) \leq 0, \quad i = 1 \ldots m \end{aligned}$$

is equivalent to

$$\begin{aligned} &\text{minimize (over } x, y) \ f_0(y_0) \\ &\text{s.t. } f_i(y_i) \leq 0, \ i = 1 \ldots m; \quad y_i = A_i x + b_i, \ i = 0, 1 \ldots m. \end{aligned}$$

* introducing slack variables for linear inequalities

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{s.t.} \quad & a_i^T x \le b_i, \quad i=1..m \end{aligned}$$

is equivalent to

$$\begin{aligned} \text{minimize (over } x, s) \quad & f_0(x) \\ \text{s.t.} \quad & a_i^T x + s_i = b_i, \quad i=1..m \\ & s_i \ge 0, \quad i=1..m \end{aligned}$$

* epigraph form: standard form convex problem is equivalent to

$$\begin{aligned} \text{minimize (over } x, t) \quad & t \\ \text{s.t.} \quad & f_0(x) - t \le 0 \\ & f_i(x) \le 0, \quad i=1...m \\ & Ax = b \end{aligned}$$

* minimizing over some variables

$$\begin{aligned} \text{minimize} \quad & f_0(x_1, x_2) \\ \text{s.t.} \quad & f_i(x_1) \le 0, \quad i=1..m \end{aligned}$$

is equivalent to

$$\begin{aligned} \text{minimize} \quad & \tilde{f}_0(x_1) \\ \text{subject to} \quad & f_i(x_1) \le 0 \quad i=1..m \end{aligned}$$

where $\tilde{f}_0(x_1) = \inf_{x_2} f_0(x_1, x_2)$

C. Quasiconvex optimization

(1)

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, \quad i=1...m \\ & Ax = b \end{aligned}$$

with $f_0: \mathbb{R}^n \to \mathbb{R}$ quasiconvex, $f_1 \dots f_m$ convex

~~can hav~~ can have locally optimal ~~solutio~~ points that are not (globally) optimal

(2) Convex representation of sublevel sets of f_0

if f_0 is quasiconvex, there exists a family of functions ϕ_t such that:

* $\phi_t(x)$ is convex in x for fixed t
* t-sublevel set of f_0 is 0-sublevel set of ϕ_t, i.e.

$$f_0(x) \le t \iff \phi_t(x) \le 0$$

(3) quasiconvex ~~optima~~ optimization via convex feasibility problems

$$\phi_t(x) \le 0, \quad f_i(x) \le 0, \quad i = 1..m, \quad Ax = b \quad (I)$$

* for fixed t, a convex feasibility problem in x
* if feasible, we can conclude that $t \ge p^*$; if infeasible, $t \le p^*$

Bisection method for quasiconvex optimization

given $l \le p^*$, $u \ge p^*$, tolerance $\epsilon > 0$

repeat

$t := (l+u)/2$

solve the convex feasibility problem (I)

if (I) is feasible, $u := t$;

else $l := t$.

until $u - l \le \epsilon$.

D. Linear Optimization

(1) Linear Program

$$\begin{aligned} \text{minimize} \quad & c^T x + d \\ \text{s.t.} \quad & Gx \preceq h \\ & Ax = b \end{aligned}$$

* convex problem with affine objective and constraint functions
* feasible set is a polyhedron

(2) <u>piecewise-linear minimization</u> to LP

$$\text{minimize} \quad \max_{i=1..m} (a_i^T x + b_i)$$

equivalent to an LP

$$\begin{aligned} \text{minimize} \quad & t \\ \text{s.t.} \quad & a_i^T x + b_i \le t, \quad i=1..m \end{aligned}$$

(3) <u>Linear-fractional Program</u>

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{s.t.} \quad & Gx \preceq h \\ & Ax = b \end{aligned}$$

$$f_0(x) = \frac{c^T x + d}{e^T x + f} \qquad \mathrm{dom}\, f_0(x) = \{x \mid e^T x + f > 0\}$$

* a quasiconvex optimization problem; can be solved by bisection
* also equivalent to the LP (variables y, z)

$$\begin{aligned} \text{minimize} \quad & c^T y + dz \\ \text{s.t.} \quad & Gy \preceq hz \\ & Ay = bz \\ & e^T y + fz = 1 \\ & z \ge 0 \end{aligned}$$

(4) <u>generalized linear-fractional program</u>

$$f_0(x) = \max_{i=1..r} \frac{c_i^T x + d_i}{e_i^T x + f_i}, \qquad \mathrm{dom}\, f_0(x) = \{x \mid e_i^T x + f_i > 0, i=1..r\}$$

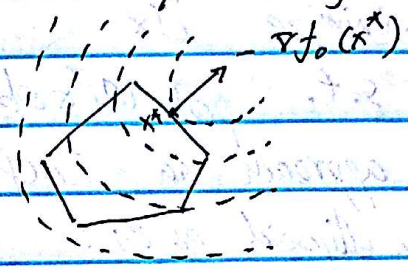
* a quasiconvex optimization problem; can be solved by bisection

E. Quadratic Optimization

(1)

$$\begin{aligned} \text{minimize} \quad & (\tfrac{1}{2}) x^T P x + q^T x + r \\ \text{s.t.} \quad & Gx \preceq h \\ & Ax = b \end{aligned}$$

* $P \in S_+^n$, so objective is convex quadratic
* minimize a convex quadratic function over a polyhedron



(2) <u>Quadratically constrained quadratic program</u> (QCQP)

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2} x^T P_0 x + q_0^T x + r_0 \\ \text{s.t.} \quad & \tfrac{1}{2} x^T P_i x + q_i^T x + r_i \le 0, \quad i = 1..m \\ & Ax = b \end{aligned}$$

* $P_i \in S_+^n$; objective and constraints are convex quadratic
* if $P_1, \dots, P_m \in S_{++}^n$, feasible region is intersection of m ellipsoids and an affine set

(3) <u>Second-order cone programming</u> (SOCP)

$$\begin{aligned} \text{minimize} \quad & f^T x \\ \text{s.t.} \quad & \|A_i x + b_i\|_2 \le c_i^T x + d_i, \quad i = 1..m \\ & Fx = g \end{aligned}$$

$(A_i \in \mathbb{R}^{n_i \times n}, F \in \mathbb{R}^{p \times n})$

* inequalities are called second-order cone (SOC) constraints: $(A_i x + b_i, c_i^T x + d_i) \in$ second-order cone in $\mathbb{R}^{n_i+1}$
* for $n_i = 0$, reduces to an LP; if $c_i = 0$, reduces to a QCQP
* more general than QCQP and LP.

(4) <u>Robust Linear Programming</u>

~~then~~ the parameters in optimization problems are often uncertain.
two common approaches to handling uncertainty (in a_i, for simplicity)

* deterministic model: constraints must hold for all $a_i \in \mathcal{E}_i$

$$\begin{aligned} \text{minimize} \quad & c^T x \\ \text{s.t.} \quad & a_i^T x \le b_i \text{ for all } a_i \in \mathcal{E}_i, \quad i = 1..m \end{aligned}$$

* stochastic model: a_i is random variable; constraints must hold with probability η.

$$\text{minimize} \quad c^T x$$
$$\text{s.t.} \quad \text{prob}(a_i^T x \le b_i) \ge \eta, \quad i=1,\dots,m$$

(5) deterministic approach via SOCP

* choose an ellipsoid as $\mathcal{E}_i$:

$$\mathcal{E}_i = \{\bar{a}_i + P_i u \mid \|u\|_2 \le 1\}, \quad (\bar{a}_i \in \mathbb{R}^n, P_i \in \mathbb{R}^{n\times n})$$

center is $\bar{a}_i$, semi-axes determined by singular values/vectors of P

* ~~ob~~ robust LP

$$\text{minimize} \quad c^T x$$
$$\text{s.t.} \quad a_i^T x \le b_i, \ \forall a_i \in \mathcal{E}_i, \quad i=1,\dots,m$$

is equivalent to SOCP

$$\text{minimize} \quad c^T x$$
$$\text{s.t.} \quad \bar{a}_i^T x + \|P_i^T x\|_2 \le b_i, \quad i=1,\dots,m$$

(follow from ~~$\sup_{\|u\|_2\le 1}$~~ $\sup_{\|u\|_2\le 1} (\bar{a}_i + P_i u)^T x = \bar{a}_i^T x + \|P_i^T x\|_2$)

(6) stochastic approach via SOCP

* assume a_i is Gaussian with mean $\bar{a}_i$, covariance Σ_i ($a_i \sim N(\bar{a}_i, \Sigma_i)$)
* $a_i^T x$ is Gaussian r.v. with mean $\bar{a}_i^T x$, variance $x^T \Sigma_i x$; hence

$$\text{prob}(a_i^T x \le b_i) = \Phi\left(\frac{b_i - \bar{a}_i^T x}{\|\Sigma_i^{1/2} x\|_2}\right)$$

where $\Phi(x) = \left(\frac{1}{\sqrt{2\pi}}\right) \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$ is CDF of $N(0,1)$

* robust LP: $\text{minimize} \quad c^T x$

$$\text{s.t.} \quad \text{prob}(a_i^T x \le b_i) \ge \eta, \quad i=1,\dots,m$$

with $\eta \ge \frac{1}{2}$ is equivalent to the SOCP

$$\text{minimize} \quad c^T x$$
$$\text{s.t.} \quad \bar{a}_i^T x + \Phi^{-1}(\eta) \cdot \|\Sigma_i^{1/2} x\|_2 \le b_i, \quad i=1,\dots,m$$

F. Geometric Programming

(1) <u>monomial function</u>

$$f(x) = c x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}, \quad \mathrm{dom}\, f = \mathbb{R}^n_{++}$$

with $c > 0$; exponent a_i can be any real number

(2) <u>posynomial function</u>: sum of monomials

$$f(x) = \sum_{k=1}^{K} c_k x_1^{a_{1k}} x_2^{a_{2k}} \cdots x_n^{a_{nk}}, \quad \mathrm{dom}\, f = \mathbb{R}^n_{++}$$

(3) <u>geometric program</u> (GP)

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 1, \quad i = 1, \dots, m \\ & h_i(x) = 1, \quad i = 1, \dots, p \end{aligned}$$

with f_i posynomial, h_i monomial

(4) Geometric program in convex form

change variables to $y_i = \log x_i$, and take logarithm of cost and constraints

* monomial $f(x) = c x_1^{a_1} \cdots x_n^{a_n}$ transforms to

$$\log f(e^{y_1}, \dots, e^{y_n}) = a^T y + b \qquad (b = \log c)$$

* posynomial $f(x) = \sum_{k=1}^{K} c_k x_1^{a_{1k}} x_2^{a_{2k}} \cdots x_n^{a_{nk}}$ ~~trans~~ transforms to

$$\log f(e^{y_1}, \dots, e^{y_n}) = \log\left(\sum_{k=1}^{K} e^{a_k^T y + b_k}\right) \qquad (b_k = \log c_k)$$

* geometric program transforms to convex problem

$$\begin{aligned} \text{minimize} \quad & \log\left(\sum_{k=1}^{K} \exp(a_{0k}^T y + b_{0k})\right) \\ \text{s.t.} \quad & \log\left(\sum_{k=1}^{K} \exp(a_{ik}^T y + b_{ik})\right) \le 0, \quad i = 1 \dots m \\ & G y + d = 0 \end{aligned}$$

G. <u>Generalized inequality constraints</u>

(1) convex problem with generalized inequality constraints

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{s.t.} \quad & \cancel{f_i(x) \preceq_{K_i} 0}, \; f_i(x) \preceq_{K_i} 0, \; i = 1 \dots m, \quad Ax = b \end{aligned}$$

* $f_0: \mathbb{R}^n \to \mathbb{R}$ convex; $f_i: \mathbb{R}^n \to \mathbb{R}^{k_i}$, K_i-convex w.r.t. proper cone K_i

* same properties as standard convex problem (convex feasible set, local optimum is global, etc.)

(2) conic form problem: special case with affine objective and constraints

$$\begin{aligned} \text{minimize} \quad & c^T x \\ \text{s.t.} \quad & Fx + g \preceq_K 0 \\ & Ax = b \end{aligned}$$

extends linear programming ($K = \mathbb{R}^m_+$) to nonpolyhedral cones

H. Semidefinite programming (SDP)

(1)

$$\begin{aligned} \text{minimize} \quad & c^T x \\ \text{s.t.} \quad & x_1 F_1 + x_2 F_2 + \cdots x_n F_n + G \preceq 0 \\ & Ax = b \end{aligned}$$

with $F_i, G \in \mathbb{S}^k$

* inequality constraint is called linear matrix inequality (LMI)

* includes problems with multiple LMI constraints: e.g.

$$x_1 \hat{F}_1 + \cdots x_n \hat{F}_n + \hat{G} \preceq 0, \quad x_1 \tilde{F}_1 + \cdots + x_n \tilde{F}_n + \tilde{G} \preceq 0$$

is equivalent to single LMI

$$x_1 \begin{bmatrix} \hat{F}_1 & 0 \\ 0 & \tilde{F}_1 \end{bmatrix} + \cdots x_n \begin{bmatrix} \hat{F}_n & 0 \\ 0 & \tilde{F}_n \end{bmatrix} + \begin{bmatrix} \hat{G} & 0 \\ 0 & \tilde{G} \end{bmatrix} \preceq 0$$

(2) LP and equivalent SDP

LP: minimize $c^T x$
s.t $Ax \preceq b$

SDP: min. $c^T x$
s.t. $\mathrm{diag}(Ax - b) \preceq 0$

(3) SOCP and equivalent SDP

SOCP: $\min.\ f^T x$

$\quad$ s.t. $\|A_i x + b_i\|_2 \le c_i^T x + d_i,\ i = 1..m$

SDP: $\min.\ f^T x$

$\quad$ s.t. $\begin{bmatrix} (c_i^T x + d_i) I & A_i x + b_i \\ (A_i x + b_i)^T & c_i^T x + d_i \end{bmatrix} \succeq 0,\ i = 1..m$

(4) <u>Eigenvalue minimization</u>

$$\min.\ \lambda_{max}(A(x))$$

where $A(x) = A_0 + x_1 A_1 + \cdots + x_n A_n$ (with given $A_i \in \mathbb{S}^k$)

equivalent SDP $\quad \min.\ t$

$$\text{s.t. } A(x) \preceq tI$$

* variables $x \in \mathbb{R}^n$, $t \in \mathbb{R}$
* follows from $\lambda_{max}(A) \le t \iff A \preceq tI$

(5) <u>Matrix Norm minimization</u>

$$\min.\ \|A(x)\|_2 = (\lambda_{max}(A(x)^T A(x)))^{\frac{1}{2}}$$

where $A(x) = A_0 + x_1 A_1 + \cdots x_n A_n$ (with given $A_i \in \mathbb{R}^{p \times q}$)

equivalent SDP

$$\min.\ t$$

$$\text{s.t. } \begin{bmatrix} tI & A(x) \\ A(x)^T & tI \end{bmatrix} \succeq 0$$

* variables $x \in \mathbb{R}^n$, $t \in \mathbb{R}$
* constraint follows from

$$\|A\|_2 \le t \iff A^T A \preceq t^2 I,\ t \ge 0$$

$$\iff \begin{bmatrix} tI & A \\ A^T & tI \end{bmatrix} \succeq 0$$

mini. (w.r.t. K) $f_0(x)$

s.t. $f_i(x) \le 0$, $i=1..m$

$h_i(x) = 0$, $i=1..p$

vector objective $f_0 : \mathbb{R}^n \to \mathbb{R}^q$, minimized w.r.t. proper cone $K \subseteq \mathbb{R}^q$

(2) convex vector optimization problem

mini. (w.r.t. K) $f_0(x)$

s.t. $f_i(x) \le 0$ $i=1..m$

$Ax = b$

with f_0 K-convex, $f_1 .. f_m$ convex

(3) Optimal and Pareto optimal points

set of achievable objective values

$$O = \{ f_0(x) \mid x \text{ feasible} \}$$

* feasible x is optimal if $f_0(x)$ is the minimum value of O
* feasible x is Pareto optimal if $f_0(x)$ is a minimal value of O

(4) Multicriterion optimization

Vector optimization problem with $K = \mathbb{R}^q_+$

$$f_0(x) = (F_1(x), \ldots, F_q(x))$$

* q different objectives F_i; roughly speaking we want all F_i's to be small
* feasible x^* is optimal if

$$y \text{ feasible} \Rightarrow f_0(x^*) \preceq f_0(y)$$

if there exists an optimal point, the objectives are noncompeting

* feasible x^{po} is Pareto optimal if

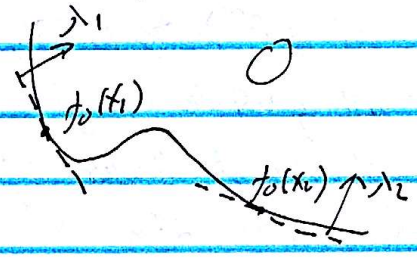
$$y \text{ feasible}, \ f_0(y) \preceq f_0(x^{po}) \Rightarrow f_0(x^{po}) = f_0(y)$$

if there are multiple Pareto optimal values, there is a trade-off between the objectives.

(5) <u>Scalarization</u>: to find Pareto optimal points: choose $\lambda \succ_{K^*} 0$ and solve scalar problem

$$\begin{aligned} \text{min.} \quad & \lambda^T f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, \ i=1..m \\ & h_i(x) = 0, \ i=1..p \end{aligned}$$

if x is optimal for scalar problem, then it is Pareto-optimal for vector optimization problem



for convex vector optimization problems, can find (almost) all Pareto optimal points by varying $\lambda \succ_{K^*} 0$

(6) Scalarization for multicriterion problems

to find Pareto optimal points, min. positive weighted sum

$$\lambda^T f_0(x) = \lambda_1 F_1(x) + \cdots \lambda_q F_q(x)$$

Lecture 5 Duality

A. ~~Introduction~~ Lagrange dual problem

(1) <u>Lagrangian</u>

standard form problem (not necessarily convex)

$$\begin{aligned} \text{min.} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, \ i=1,\dots,m \\ & h_i(x) = 0, \ i=1,\dots,p \end{aligned}$$

variable $x \in \mathbb{R}^n$, domain D, optimal value p^*

<u>Lagrangian</u>: $L: \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$, with $\operatorname{dom} L = D \times \mathbb{R}^m \times \mathbb{R}^p$

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \boxed{\lambda_i} f_i(x) + \sum_{i=1}^{p} \boxed{\nu_i} h_i(x)$$

Lagrange multiplier

(2) Lagrange dual function: $g: \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu)$$
$$= \inf_{x \in D} \Big(f_0(x) + \underbrace{\sum_{i=1}^{m} \lambda_i f_i(x)}_{\leq 0} + \underbrace{\sum_{i=1}^{p} \nu_i h_i(x)}_{0}\Big)$$

g is ~~convex~~ concave, can be $-\infty$ for some λ, ν

(3) lower bound property: if $\lambda \succeq 0$, then $g(\lambda, \nu) \leq p^*$

proof: if $\tilde{x}$ is feasible and $\lambda \succeq 0$, then

$$f_0(\tilde{x}) \geq L(\tilde{x}, \lambda, \nu) \geq \inf_{x \in D} L(x, \lambda, \nu) = g(\lambda, \nu)$$

minimize over all feasible $\tilde{x}$ gives $p^* \geq g(\lambda, \nu)$

(4) Least-norm solution of linear equations

$$\text{min.} \quad x^T x$$
$$\text{s.t.} \quad Ax = b$$

dual function:

* Lagrangian is $L(x, \nu) = x^T x + \nu^T (Ax - b)$
* to minimize L over x, set gradient equal to zero:

$$\nabla_x L(x, \nu) = 2x + A^T \nu = 0 \Rightarrow x = -\tfrac{1}{2} A^T \nu$$

* plug in in L to obtain g:

$$g(\nu) = L(-\tfrac{1}{2} A^T \nu, \nu) = -\tfrac{1}{4} \nu^T A A^T \nu - b^T \nu$$

a concave function of ν

Lower bound property: $p^* \geq -\tfrac{1}{4} \nu^T A A^T \nu - b^T \nu$ for all ν

(5) Standard form LP

$$\text{min.} \quad c^T x$$
$$\text{s.t.} \quad Ax = b, \quad x \succeq 0$$

dual ~~form~~ function

* Lagrangian is

$$L(x, \lambda, \nu) = c^T x + \nu^T (Ax - b) - \lambda^T x$$
$$= -b^T \nu + (c + A^T \nu - \lambda)^T x$$

* L is affine in x, hence

$$g(\lambda,\nu) = \inf_x L(x,\lambda,\nu) = \begin{cases} -b^T\nu & A^T\nu - \lambda + c = 0 \\ -\infty & \text{otherwise} \end{cases}$$

g is linear on affine domain $\{(\lambda,\nu) \mid A^T\nu - \lambda + c = 0\}$, hence concave

Lower bound property: $p^* \geq -b^T\nu$ if $A^T\nu + c \succeq 0$

(6) Equality constrained norm minimization

$$\begin{aligned} \text{mini.} \quad & \|x\| \\ \text{s.t.} \quad & Ax = b \end{aligned}$$

dual function:

$$g(\nu) = \inf_x (\|x\| - \nu^T Ax + b^T\nu) = \begin{cases} b^T\nu & \|A^T\nu\|_* \leq 1 \\ -\infty & \text{otherwise} \end{cases}$$

where $\|v\|_* = \sup_{\|u\|\leq 1} u^T v$ is dual norm of $\|\cdot\|$

Lower bound property: $p^* \geq b^T\nu$ if $\|A^T\nu\|_* \leq 1$

(7) Lagrange dual and conjugate function

$$\begin{aligned} \text{mini.} \quad & f_0(x) \\ \text{s.t.} \quad & Ax \preceq b, \quad Cx = d \end{aligned}$$

dual function:

$$g(\lambda,\nu) = \inf_{x \in \mathrm{dom}\, f_0} (f_0(x) + (A^T\lambda + C^T\nu)^T x - b^T\lambda - d^T\nu)$$

$$= -f_0^*(-A^T\lambda - C^T\nu) - b^T\lambda - d^T\nu$$

* recall definition of conjugate $f^*(y) = \sup_{x \in \mathrm{dom}\, f} (y^T x - f(x))$
* simplifies derivation of dual if conjugate of f_0 is known

(8) The dual problem

<u>Lagrange dual problem</u>

$$\begin{aligned} \text{maxi.} \quad & g(\lambda,\nu) \\ \text{s.t.} \quad & \lambda \succeq 0 \end{aligned}$$

* finds best lower bound on p^*, obtained from Lagrange dual function
* a convex optimization problem; optimal value denoted $\boxed{d^*}$ $p^* \geq d^*$
* λ, ν are dual feasible if $\lambda \succeq 0$, $(\lambda, \nu) \in \mathrm{dom}\, g$
* often simplified by making implicit constraint $(\lambda, \nu) \in \mathrm{dom}\, g$ explicit.

B. Weak and Strong duality

(1) weak duality: $d^* \leq p^*$
* always holds (for convex and nonconvex problems)
* can be used to find nontrivial lower bounds for difficult problems

(2) strong duality: $d^* = p^*$
* does not hold in general
* (usually) holds for convex problems
* conditions that guarantee strong duality in convex problems are called <u>constraint qualifications</u>

(3) Slater's constraint qualification

strong duality holds for a convex problem

$$\begin{aligned} \min.\ & f_0(x) \\ \text{s.t.}\ & f_i(x) \leq 0,\ i = 1..m \\ & Ax = b \end{aligned}$$

if it is strictly feasible, i.e.,

~~* also guarantees~~ $\exists x \in \mathrm{int}\, D$: $f_i(x) < 0$, $i = 1..m$, $Ax = b$

* also guarantees that the dual optimum is attained (if $p^* > -\infty$)
* can be sharpened: ~~e.g., can replace int D with relint D~~
* there exists many other types of constraint qualifications

(4) Inequality form LP

primal problem: $\min.\ c^T x$, s.t. $Ax \preceq b$

dual function: $g(\lambda) = \inf_x \left((c + A^T\lambda)^T x - b^T\lambda \right) = \begin{cases} -b^T\lambda & A^T\lambda + c = 0 \\ -\infty & \text{otherwise} \end{cases}$

dual problem: $\max.\ -b^T\lambda$

s.t. $A^T\lambda + c = 0$, $\lambda \succeq 0$

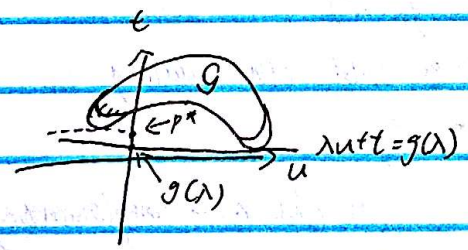
* from Slater's ~~const~~ condition: $p^* = d^*$ if $A\tilde{x} \prec b$ for some $\tilde{x}$
* in fact, $p^* = d^*$ except when primal and dual are infeasible

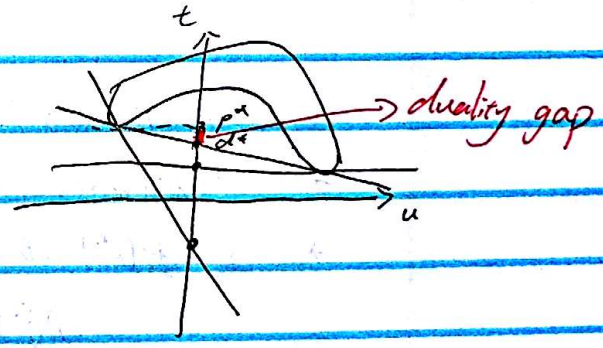
C. Geometric interpretation

(1) for simplicity, consider problem with one constraint $f_1(x) \le 0$

<u>interpretation of dual function</u>:

$$g(\lambda) = \inf_{(u,t) \in G} (t + \lambda u), \quad \text{where } G = \{(f_1(x), f_0(x)) \mid x \in D\}$$

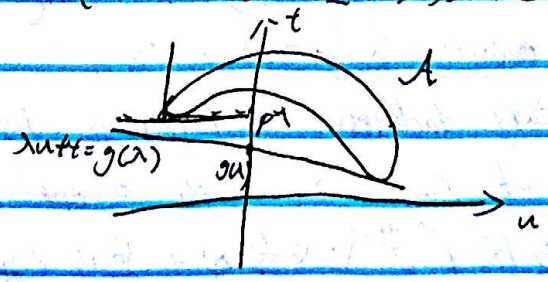


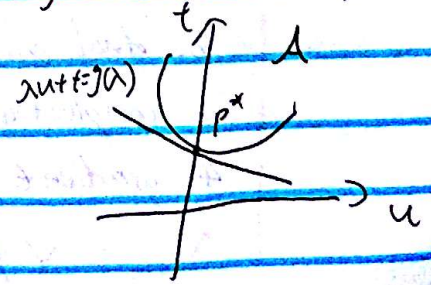


* $\lambda u + t = g(\lambda)$ is (non-vertical) supporting hyperplane to G
* hyperplane intersects t-axis at $t = g(\lambda)$

(2) epigraph variation: same interpretation if G is replaced with

$$A = \{(u,t) \mid f_1(x) \le u,\ f_0(x) \le t \text{ for some } x \in D\}$$





strong duality

* holds if there is a non-vertical supporting hyperplane to $\mathcal{A}$ at $(0, p^*)$
* for convex problem, $\mathcal{A}$ is convex, hence has support hyperplane at $(0, p^*)$
* Slater's condition: if there exist $(\tilde{u}, \tilde{t}) \in \mathcal{A}$ with $\tilde{u} < 0$, then supporting hyperplanes at $(0, p^*)$ must be non-vertical

D. Optimality conditions

(1) Complementary slackness

assume strong duality holds, x^* is primal optimal, (λ^*, ν^*) is dual optimal:

$$f_0(x^*) = g(\lambda^*, \nu^*) = \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p \nu_i^* h_i(x) \right)$$

$$\leq f_0(x^*) + \underbrace{\sum_{i=1}^m \lambda_i^* f_i(x^*)}_{\leq 0} + \underbrace{\sum_{i=1}^p \nu_i^* h_i(x^*)}_{=0}$$

$$= f_0(x^*)$$

hence, the two inequalities hold with equality

* x^* minimizes $L(x, \lambda^*, \nu^*)$
* $\lambda_i^* f_i(x^*) = 0$ for $i = 1, \dots, m$ (known as complementary slackness):

$$\lambda_i^* > 0 \Rightarrow f_i(x^*) = 0, \quad f_i(x^*) < 0 \Rightarrow \lambda_i^* = 0$$

(2) KKT conditions

the following four conditions are called KKT conditions (for a problem with differentiable f_i, h_i):

1. primal constraints: $f_i(x) \leq 0$, $i = 1 \dots m$, $h_i(x) = 0$, $i = 1 \dots p$
2. dual constraints: $\lambda \succeq 0$
3. complementary slackness: $\lambda_i f_i(x) = 0$, $i = 1 \dots m$
4. gradient of Lagrangian with respect to x vanishes:

$$\nabla_x L = \nabla f_0(x) + \sum_{i=1}^m \lambda_i \nabla f_i(x) + \sum_{i=1}^p \nu_i \nabla h_i(x) = 0$$

E. Perturbation and sensitivity analysis

(1) perturbed problem and its dual

$$\begin{aligned} \text{mini. } & f_0(x) \\ \text{s.t. } & f_i(x) \le u_i \quad i=1..m \\ & h_i(x) = v_i \quad i=1..p \end{aligned} \qquad \begin{aligned} \text{maxi. } & g(\lambda, \nu) - u^T\lambda - v^T\nu \\ \text{s.t. } & \lambda \succeq 0 \end{aligned}$$

* x is primal variable; u, v are parameters
* $p^*(u, v)$ is optimal value as a function of u, v
* we are interested in information about $p^*(u, v)$ that we can obtain from the solution of the unperturbed problem and its dual

Lecture 6 Approximation and fitting

A. Norm Approximation

(1)

$$\text{mini. } \|Ax - b\|$$

$A \in \mathbb{R}^{m \times n}$ with $m \ge n$, $\|\cdot\|$ is a norm on $\mathbb{R}^m$

interpretations of solution $x^* = \operatorname{argmin}_x \|Ax - b\|$:

* geometric: Ax^* is point in $R(A)$ closest to b
* estimation: linear measurement model

$$y = Ax + v$$

y are measurements, x is unknown, v is measurement error

given $y = b$, best guess of x is x^*

* optimal design: x are design variables (input), Ax is result (output)

x^* is design that best approximates desired result b.

(2) Penalty function approximation

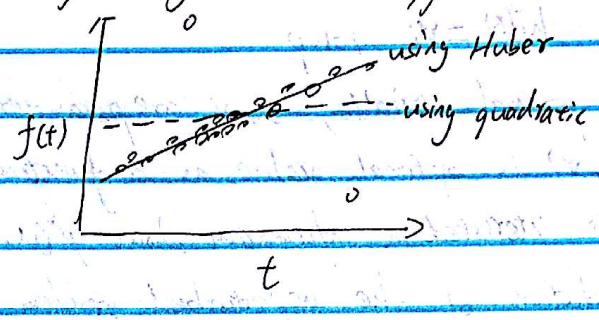
$$\begin{aligned} \text{mini. } & \phi(r_1) + \dots + \phi(r_m) \\ \text{s.t. } & r = Ax - b \end{aligned}$$

$A \in \mathbb{R}^{m \times n}$, $\phi: \mathbb{R} \to \mathbb{R}$ is a convex penalty function

(3) Huber penalty function (with parameter M as threshold)

$$\phi_{hub}(u) = \begin{cases} u^2 & |u| \le M \\ M(2|u| - M) & |u| > M \end{cases}$$

linear growth for large u makes approximation less sensitive to outliers



B. Least-norm problems

(1)
$$\begin{aligned} \text{min.} \quad & \|x\| \\ \text{s.t.} \quad & Ax = b \end{aligned}$$

$A \in \mathbb{R}^{m \times n}$ with $m \le n$, $\|\cdot\|$ is a norm on $\mathbb{R}^n$

interpretations of solution $x^* = \operatorname{argmin}_{Ax=b} \|x\|$:

* geometric: x^* is point in affine set $\{x \mid Ax = b\}$ with minimum distance to 0
* estimation: $b = Ax$ are (perfect) measurements of x; x^* is smallest ('most plausible') estimate consistent with measurements
* design: x are design variables (inputs); b are required results (outputs) x^* is smallest ('most efficient') design that satisfies requirements.

(2) least-penalty problem

$$\begin{aligned} \text{min.} \quad & \phi(x_1) + \cdots + \phi(x_n) \\ \text{s.t.} \quad & Ax = b \end{aligned}$$

$\phi : \mathbb{R} \to \mathbb{R}$ is convex penalty function

C. Regularized approximation

(1) $\text{mini.}\ (\text{w.r.t. } \mathbb{R}^2_+)\quad (\|Ax-b\|, \|x\|)$

$A \in \mathbb{R}^{m\times n}$, norms on $\mathbb{R}^m$ and $\mathbb{R}^n$ can be different

interpretation: find good approximation $Ax \approx b$ with small x

* estimation: linear measurement model $y = Ax + v$, with prior knowledge that $\|x\|$ is small
* optimal design: small x is cheaper or more efficient, or the linear model $y = Ax$ is only valid for small x
* robust approximation: good approximation $Ax \approx b$ with small x is less sensitive to errors in A than good approximation with large x.

(2) Scalarized problem

$$\text{mini.}\ \|Ax-b\| + \gamma\|x\|$$

* solution for $\gamma > 0$ traces out optimal trade-off curve
* other common method: $\text{mini.}\ \|Ax-b\|^2 + \delta\|x\|^2$ with $\delta > 0$

(3) Tikhonov regularization

$$\text{mini.}\ \|Ax-b\|_2^2 + \delta\|x\|_2^2 \rightarrow \text{Ridge Regression}$$

can be solved as a least-squares problem

$$\text{mini.}\ \left\| \begin{bmatrix} A \\ \sqrt{\delta} I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2$$

solution $x^* = (A^TA + \delta I)^{-1}A^Tb$

(4) Signal reconstruction

$$\text{mini.}\ (\text{w.r.t. } \mathbb{R}^2_+)\quad (\|\hat{x} - x_{cor}\|_2, \phi(\hat{x}))$$

* $x \in \mathbb{R}^n$ is unknown ~~sig~~ signal
* $x_{cor} = x + v$ is (known) corrupted version of x, with additive noise v
* variable $\hat{x}$ (reconstructed signal) is estimation of x
* $\phi: \mathbb{R}^n \rightarrow \mathbb{R}$ is regularization function or smoothing objective

D. Robust approximation

(1) $\min. \|Ax-b\|$ with uncertain A

two approaches:

* stochastic: assume A is random, mini. $E\|Ax-b\|$
* worst-case: set $\mathcal{A}$ of possible values of A, mini.

$$\sup_{A\in\mathcal{A}} \|Ax-b\| \Rightarrow \text{minimax fitting}$$

tractable only in special cases (certain norms $\|\cdot\|$, distributions set $\mathcal{A}$)

Lecture 18: Statistical Estimation

A. Maximum likelihood estimation

(1) Parametric distribution estimation

* distribution estimation problem: estimate probability density $p(y)$ of a random variable from observed values.
* parametric distribution estimation: choose from a family of densities $P_x(y)$, indexed by a parameter x

(2) Maximum likelihood estimation

$$\max. \text{ (over } x) \quad \log P_x(y)$$

* y is observed value
* $l(x) = \log P_x(y)$ is called log-likelihood function
* can add constraints $x \in C$ explicitly, or define $P_x(y)=0$ for $x \notin C$
* a convex optimization problem if $\log P_x(y)$ is concave in x for fixed y.

(3) Linear measurements with IID noise

linear measurement model: $y_i = a_i^T x + v_i$, $i=1..m$

* $x \in \mathbb{R}^n$ is vector of unknow parameters

* v_i is IID measurement noise, with density $p(z)$
* y_i is measurement: $y \in \mathbb{R}^m$ has density $p_x(y) = \prod_{i=1}^{m} p(y_i - a_i^T x)$

(4) maximum likelihood estimate: any solution x of

$$\text{maximize } l(x) = \sum_{i=1}^{m} \log p(y_i - a_i^T x)$$

(5) Logistic regression

random variable $y \in \{0,1\}$ with distribution

$$p = \text{prob}(y=1) = \frac{\exp(a^T u + b)}{1 + \exp(a^T u + b)}$$

* a, b are parameters; $u \in \mathbb{R}^n$ are (observable) explanatory variables
* estimation problem: estimate a, b from m observations (u_i, y_i)

(6) log-likelihood function (for $y_1 = \dots = y_k = 1$, $y_{k+1} = \dots = y_m = 0$):

$$l(a,b) = \log\left(\prod_{i=1}^{k} \frac{\exp(a^T u_i + b)}{1+\exp(a^T u_i + b)} \prod_{i=k+1}^{m} \frac{1}{1+\exp(a^T u_i + b)}\right)$$

$$= \sum_{i=1}^{k} (a^T u_i + b) - \sum_{i=1}^{m} \log(1 + \exp(a^T u_i + b))$$

concave in a, b.

B. Optimal detector design

(1) (Binary) hypothesis testing

detection (hypothesis testing) problem: give observation of a random variable $X \in \{1, \dots, n\}$, choose between:

* hypothesis 1: X was generated by distribution $p = (p_1, \dots, p_n)$
* hypothesis 2: . . . . . $q = (q_1, \dots q_n)$

(2) randomized detector

* a nonnegative matrix $T \in \mathbb{R}^{2 \times n}$, with $\mathbb{1}^T T = \mathbb{1}^T$
* if we observe $X = k$, we choose hypothesis 1 with probability t_{1k}, hypothesis 2 with probability t_{2k}
* if all elements of T are 0 or 1, it is called a deterministic detector

(3) detection probability matrix:

$$D = [T_p \ \ T_q] = \begin{bmatrix} 1-P_{fp} & P_{fn} \\ P_{fp} & 1-P_{fn} \end{bmatrix}$$

* P_{fp} is probability of selecting hypothesis 2 if X is generated by distribution 1 (false positive)
* P_{fn} is probability of selecting hypothesis 1 if X is generated by 2 (false negative)

(4) multicriterion formulation of detector design

$$\begin{aligned} \text{mini. (w.r.t. } \mathbb{R}_+^2) \quad & (P_{fp}, P_{fn}) = ((T_p)_2, (T_q)_1) \\ \text{s.t.} \quad & t_{1k} + t_{2k} = 1, \quad k=1,\dots,n \\ & t_{ik} \geq 0, \quad i=1,2, \quad k=1\dots n \end{aligned}$$

(5) Scalarization (with weight $\lambda > 0$)

$$\begin{aligned} \text{mini.} \quad & (T_p)_2 + \lambda (T_q)_1 \\ \text{s.t.} \quad & t_{1k} + t_{2k} = 1, \ t_{ik} \geq 0, \ i=1,2, \ k=1\dots n \end{aligned}$$

(6) an LP with a simple analytical solution

$$(t_{1k}, t_{2k}) = \begin{cases} (1,0) & p_k \geq \lambda q_k \\ (0,1) & p_k < \lambda q_k \end{cases}$$

* a deterministic detector, given by a likelihood ratio test
* if $p_k = \lambda q_k$ for some k, any value $0 \leq t_{1k} \leq 1$, $t_{1k} = 1 - t_{2k}$ is optimal (i.e.

*

(6) minimax detector

$$\begin{aligned} \text{mini.} \quad & \max\{P_{fp}, P_{fn}\} = \max\{(T_p)_2, (T_q)_1\} \\ \text{s.t.} \quad & t_{1k} + t_{2k} = 1, \ t_{ik} \geq 0, \ i=1,2, \ k=1\dots n \end{aligned}$$

C. Experiment Design

(1) m linear measurements $y_i = a_i^T x + w_i$, $i=1\dots m$ of unknown $x \in \mathbb{R}^n$

* measurement errors w_i are IID $\mathcal{N}(0,1)$

* ML (least-squares) estimate is

$$\hat{x} = \left(\sum_{i=1}^{m} a_i a_i^T\right)^{-1} \sum_{i=1}^{m} y_i a_i$$

* error $e = \hat{x} - x$ has zero mean and covariance

$$E = Eee^T = \left(\sum_{i=1}^{m} a_i a_i^T\right)^{-1}$$

confidence ellipsoids are given by $\{x \mid (x-\hat{x})^T E^{-1} (x-\hat{x}) \le \beta\}$

<u>experiment design</u>: choose $a_i \in \{v_1, \dots v_p\}$ (a set of possible test vector) to make E 'small'.

(2) vector optimization formulation

$$\text{min}_{\cdot} \ (\text{w.r.t. } S_+^n) \quad E = \left(\sum_{k=1}^{p} m_k v_k v_k^T\right)^{-1}$$

$$\text{s.t.} \quad m_k \ge 0, \quad m_1 + \dots + m_p = m$$

$$m_k \in \mathbb{Z}$$

* variables are m_k (# of vectors a_i equal to v_k)

* difficult in general, due to integer constraint

(3) relaxed ~~ep~~ experiment design

assume $m \gg p$, use $\lambda_k = m_k/m$ as (continuous) real variable

$$\text{min}_{\cdot} \ (\text{w.r.t. } S_+^n) \quad E = \left(\frac{1}{m}\right)\left(\sum_{k=1}^{p} \lambda_k v_k v_k^T\right)^{-1}$$

$$\text{s.t.} \quad \lambda \succeq 0, \quad \mathbb{1}^T \lambda = 1$$

* common scalarizations: ~~minimiz~~ minimize $\log\det E$, $\mathrm{tr}(E)$, $\lambda_{max}(E)$

* can add other convex constraints,

(4) D-optimal design

Lecture 8: Geometric problems

A. Extremal volume ellipsoids

(1) Minimum volume ellipsoid around a set

<u>Löwner-John ellipsoid</u> of a set C: minimum volume ellipsoid $\mathcal{E}$ s.t.

$$C \subseteq \mathcal{E}$$

* parameterize $\mathcal{E}$ as $\mathcal{E} = \{v \mid \|Av+b\|_2 \le 1\}$; w.l.o.g. assume $A \in \mathbb{S}^n_{++}$

* vol $\mathcal{E}$ is proportional to $\det A^{-1}$; to compute minimum volume ellipsoid,

volume of the ellipsoid

$$\begin{aligned} \text{min.} \ (\text{over } A, b) \quad & \log \det A^{-1} \\ \text{s.t.} \quad & \sup_{v \in C} \|Av+b\|_2 \le 1 \end{aligned}$$

convex, but evaluating the constraint can be hard (for general C)

finite set $C = \{x_1, \dots x_m\}$: $\begin{aligned} \text{min.} \ (\text{over } A, b) \quad & \log \det A^{-1} \\ \text{s.t.} \quad & \|Ax_i + b\|_2 \le 1, \ i = 1 \dots m \end{aligned}$

also gives Löwner-John ellipsoid for polyhedron $\operatorname{conv}\{x_1, \dots x_m\}$

(2) Maximum volume inscribed ellipsoid

maximum volume ellipsoid $\mathcal{E}$ inside a convex set $C \subseteq \mathbb{R}^n$

* parametrize $\mathcal{E}$ as $\mathcal{E} = \{Bu+d \mid \|u\|_2 \le 1\}$; w.l.o.g. assume $B \in \mathbb{S}^n_{++}$

* vol $\mathcal{E}$ is proportional to $\det B$; can compute $\mathcal{E}$ by solving

$$\begin{aligned} \text{maxi.} \quad & \log \det B \\ \text{s.t.} \quad & \sup_{\|u\|_2 \le 1} I_C(Bu+d) \le 0 \end{aligned}$$

(where $I_C(x) = 0$ for $x \in C$ and $I_C(x) = \infty$ for $x \notin C$)

convex, but evaluating the constraint can be hard (for general C)

polyhedron $\{x \mid a_i^T x \le b_i, \ i = 1 \dots m\}$:

$$\begin{aligned} \text{maxi.} \quad & \log \det B \\ \text{s.t.} \quad & \|Ba_i\|_2 + a_i^T d \le b_i, \ i = 1 \dots m \end{aligned}$$

(constraint follows from $\sup_{\|u\|_2 \le 1} a_i^T(Bu+d) = \|Ba_i\|_2 + a_i^T d$)

B. Centering

(1) ~~Analytical~~ Analytic center of a set of inequalities

~~the analytic center of set of convex inequalities and linear~~ equations

$$\begin{aligned} f_i(x) &\le 0, \quad i = 1 \dots m \\ Fx &= g \end{aligned}$$

log barrier function

is defined as the optimal point of

$$\begin{aligned} \text{min.} \quad & -\sum_{i=1}^{m} \log(-f_i(x)) \\ \text{s.t.} \quad & Fx = g \end{aligned}$$

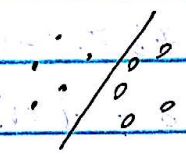
* more easily computed than MVE or Chebyshev center
* not just a property of the feasible set: two sets of inequalities can describe the same set, but have different analytic centers.

C. Classification

(1) Linear discrimination

separate two sets of points $\{x_1, \dots x_N\}$ $\{y_1 \dots y_M\}$ by a hyperplane:

$$a^T x_i + b > 0, \quad i = 1..N, \qquad a^T y_i + b < 0, \quad i = 1..M$$



homogeneous in a, b; hence equivalent to

$$a^T x_i + b \geq 1, \quad i = 1..N, \qquad a^T y_i + b \leq -1, \quad i = 1..M$$

a set of linear inequalities in a, b.

(2) Robust linear discrimination (maximum margin classification)

(Euclidean) distance between hyperplanes

$$\mathcal{H}_1 = \{z \mid a^T z + b = 1\} \qquad \mathcal{H}_2 = \{z \mid a^T z + b = -1\}$$

is $\text{dist}(\mathcal{H}_1, \mathcal{H}_2) = \dfrac{2}{\|a\|_2}$

to separate two sets of points by maximum margin,

$$\begin{aligned} \text{min.} \quad & \tfrac{1}{2}\|a\|_2 \\ \text{s.t.} \quad & a^T x_i + b \geq 1 \quad i = 1..N \\ & a^T y_i + b \leq -1 \quad i = 1..M \end{aligned}$$

(after squaring objective) a QP in a, b.

(3) Lagrange dual of maximum margin separation problem

$$\begin{aligned} \text{max.} \quad & \mathbb{1}^T \lambda + \mathbb{1}^T u \\ \text{s.t.} \quad & 2\left\|\sum_{i=1}^{N} \lambda_i x_i - \sum_{i=1}^{M} u_i y_i\right\|_2 \leq 1 \\ & \mathbb{1}^T \lambda = \mathbb{1}^T u, \quad \lambda \geq 0, \quad u \geq 0 \end{aligned}$$

from duality, optimal value is inverse of maximum margin of separation

interpretation

* change variables to $\theta_i = \lambda_i / \mathbb{1}^T\lambda$, $\gamma_i = u_i / \mathbb{1}^T u$, $t = 1/(\mathbb{1}^T\lambda + \mathbb{1}^T u)$
* invert objective to minimize $1/(\mathbb{1}^T\lambda + \mathbb{1}^T u) = t$

$$\begin{aligned} \text{min.} \quad & t \\ \text{s.t.} \quad & \left\| \sum_{i=1}^{N} \theta_i x_i - \sum_{i=1}^{M} \gamma_i y_i \right\|_2 \le t \\ & \theta \succeq 0, \ \mathbb{1}^T\theta = 1, \ \gamma \succeq 0, \ \mathbb{1}^T\gamma = 1 \end{aligned}$$

optimal value is distance between convex hulls

(4) Approximate linear separation of non-separable sets

$$\begin{aligned} \text{min.} \quad & \mathbb{1}^T u + \mathbb{1}^T v \\ \text{s.t.} \quad & a^T x_i + b \ge 1 - u_i \quad i = 1..N \\ & a^T y_i + b \le -1 + v_i \quad i = 1..M \\ & u \succeq 0, \ v \succeq 0 \end{aligned}$$

* an LP in a, b, u, v
* at optimum, $u_i = \max\{0, 1 - a^T x_i - b\}$, $v_i = \max\{0, 1 + a^T y_i + b\}$
* can be interpreted as a heuristic for minimizing # of misclassified points.

(5) support vector classifier

$$\begin{aligned} \text{min.} \quad & \|a\|_2 + \gamma(\mathbb{1}^T u + \mathbb{1}^T v) \\ \text{s.t.} \quad & a^T x_i + b \ge 1 - u_i, \quad i = 1..N \\ & a^T y_i + b \le -1 + v_i, \quad i = 1..M \\ & u \succeq 0, \ v \succeq 0 \end{aligned}$$

produces point on trade-off curve between inverse of margin $\frac{2}{\|a\|_2}$ and classification error, measured by total slack $\mathbb{1}^T u + \mathbb{1}^T v$

(6) Nonlinear discrimination

separate two sets of points by a nonlinear function.

$f(x_i) > 0$, $i = 1..N$, $\quad f(y_i) < 0$, $i = 1..M$

* choose a linearly parametrized family of functions

$$f(z) = \theta^T F(z)$$

$F = (F_1 \dots F_k) : \mathbb{R}^n \to \mathbb{R}^k$ are basis functions

* solve a set of linear inequalities in θ:

$$\theta^T F(x_i) \geq 1, \quad i = 1..N; \quad \theta^T F(y_i) \leq -1, \quad i = 1..M$$

(7) quadratic discrimination: $f(z) = z^T P z + q^T z + r$

$$x_i^T P x_i + q^T x_i + r \geq 1, \qquad y_i^T P y_i + q^T y_i + r \leq -1$$

can add additional constraints (e.g. $P \succeq -I$ to separate by an ellipsoid)

(8) polynomial discrimination: $F(z)$ are all monomials up to a given degree

Lecture 9: Unconstrained minimization

A. Terminology and assumptions

B. Gradient descent method

(1) Descent methods

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

* other notations: $x^+ = x + t\Delta x$, $\quad x := x + t\Delta x$
* 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)

(2) General descent method

given a starting point $x \in \mathrm{dom}\, f$.

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.

(3) Line search types

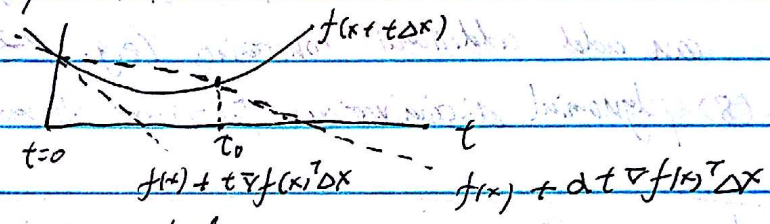
<u>exact line search</u>: $t = \underset{t>0}{\operatorname{argmin}} \; f(x + t\Delta x)$

<u>backtracking line search</u> (with parameters $\alpha \in (0, \frac{1}{2})$, $\beta \in (0,1)$)

* 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 $t \le t_0$



(4) Gradient descent method

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

<u>Alg</u>: given a starting point $x \in \operatorname{dom} f$

<u>repeat</u>: 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~~ until: stopping criterion is satisfied.

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

$$f(x^{(k)}) - p^* \le c^k (f(x^{(0)}) - p^*)$$

$c \in (0,1)$ depends on m, $x^{(0)}$, line search type

* very simple; but very slow; rarely used in practice.

C. Steepest descent method

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

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

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

(2) (unnormalized) steepest descent direction

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

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

(3) steepest descent method.

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

D. Newton's ~~ste~~ method

(1) Newton step

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

interpretations:

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

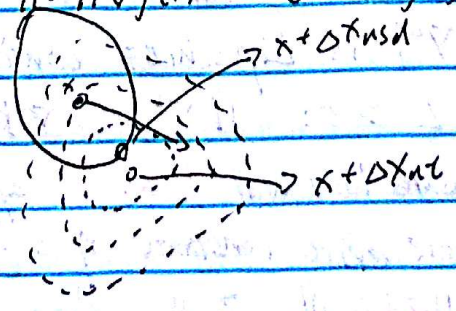
$$\hat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$$

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

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

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

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



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

(2) Newton decrement

$$\lambda(x) = (\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x))^{\frac{1}{2}}$$

a measure of the proximity of x to x^*

properties

* gives an estimate of $f(x) - p^*$, using quadratic approximation

  $\hat{f}$: $f(x) - \inf_y \hat{f}(y) = \frac{1}{2}\lambda(x)^2$

* equal to the norm of the Newton step in the quadratic Hessian norm: $\lambda(x) = (\Delta x_{nt}^T \nabla^2 f(x) \Delta x_{nt})^{\frac{1}{2}}$

* directional derivative in the Newton direction: $\nabla f(x)^T \Delta x_{nt} = -\lambda(x)^2$

* affine invariant

(3) Newton's method

given a starting point $x \in \mathrm{dom}\, f$, tolerance $\varepsilon > 0$

repeat: ① Compute the Newton step and decrement

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

$$\lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$$

② stopping criterion, quit if $\lambda^2/2 \le \varepsilon$

③ Line Search. Choose step size t by backtracking line search

④ update. $x := x + t\Delta x_{nt}$

(4) Classical convergence analysis

assumptions

* f strongly convex on S with constant m
* $\nabla^2 f$ is Lipschitz continuous on S, with constant $L > 0$: $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \le L\|x-y\|_2$

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

Outline: there exist constants $\eta \in (0, \frac{m^2}{L})$, $\gamma > 0$ such that

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

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

E. Self concordant functions

(1) definition

* convex $f: \mathbb{R} \to \mathbb{R}$ is self-concordant if $|f'''(x)| \le 2 f''(x)^{\frac{3}{2}}$ for all $x \in \mathrm{dom}\, f$

* $f: \mathbb{R}^n \to \mathbb{R}$ is self-concordant if $g(t) = f(x+tv)$ is ~~self~~ self-concordant for all $x \in \mathrm{dom}\, f$, $v \in \mathbb{R}^n$.

(2) Convergence analysis for self-concordant functions

summary: there exist constants $\eta \in (0, \frac{1}{2}]$, $\gamma > 0$ such that

* if $\lambda(x) > \eta$ then $f(x^{(k+1)}) - f(x^{(k)}) \le -\gamma$

* if $\lambda(x) \le \eta$ then $2\lambda(x^{(k+1)}) \le (2\lambda(x^{(k)}))^2$

(η and γ only depend on backtracking parameters α, β)

(3) complexity bound: ~~no~~ # of Newton iterations bounded by

$$\frac{f(x^{(0)}) - p^*}{\gamma} + \log_2 \log_2 \left(\frac{1}{\varepsilon}\right)$$

Lecture 10: Equality constrained minimization

A. Equality constrained minimization

(1)
$$\begin{aligned} \text{min.} \quad & f(x) \\ \text{s.t.} \quad & Ax = b \end{aligned}$$

* f convex, twice ~~continou~~ continuously differentiable

* $A \in \mathbb{R}^{p \times n}$ with rank $A = p$

* we assume p^* is finite and attained

(2) optimality conditions: x^* is optimal iff there exists a ν^* such that

$$\nabla f(x^*) + A^T \nu^* = 0, \quad Ax^* = b$$

(3) equality constrained quadratic minimization (with $P \in \mathbb{S}_+^n$)

$$\begin{aligned} \text{min.} \quad & \frac{1}{2} x^T P x + q^T x + r \\ \text{s.t.} \quad & Ax = b \end{aligned}$$

optimality condition: $\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \nu^* \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}$

* coefficient matrix is called KKT matrix
* KKT matrix is nonsingular if and only if

$$Ax = 0,\ x \neq 0 \Rightarrow x^T P x > 0$$

* equivalent condition for nonsingularity: $P + A^T A \succ 0$

B. Eliminating equality constraints

(1) represent solution of $\{x \mid Ax = b\}$ as

$$\{x \mid Ax = b\} = \{Fz + \hat{x} \mid z \in \mathbb{R}^{n-p}\}$$

* $\hat{x}$ is (any) particular solution
* range of $F \in \mathbb{R}^{n \times (n-p)}$ is nullspace of A (rank $F = n-p$ and $AF = 0$)

reduced or eliminated problem: $\min.\ f(Fz + \hat{x})$

* an unconstrained problem with variable $z \in \mathbb{R}^{n-p}$
* from solution z^*, obtain x^* and ν^* as

$$x^* = Fz^* + \hat{x},\quad \nu^* = -(AA^T)^{-1} A \nabla f(x^*)$$

C. Newton's method with equality constraints

(1) Newton step Δx_{nt} of f at feasible x is given by solution v of

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}$$

<u>interpretations</u>

* Δx_{nt} solves second order approximation (with variable v)

$$\begin{aligned} \min.\ & \hat{f}(x+v) = f(x) + \nabla f(x)^T v + \tfrac{1}{2} v^T \nabla^2 f(x) v \\ \text{s.t.}\ & A(x+v) = b \end{aligned}$$

* Δx_{nt} equations follow from linearizing optimality conditions

$$\nabla f(x+v) + A^T w \approx \nabla f(x) + \nabla^2 f(x) v + A^T w = 0,\quad A(x+v) = b$$

(2) Newton decrement

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

properties

* gives an estimate of $f(x) - p^*$ using quadratic approximation $\hat{f}$:

$$f(x) - \inf_{Ay=b} \hat{f}(y) = \frac{1}{2}\lambda(x)^2$$

* directional derivative in Newton direction:

$$\frac{d}{dt} f(x + t\Delta x_{nt})\Big|_{t=0} = -\lambda(x)^2$$

* in general, $\lambda(x) \neq (\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x))^{\frac{1}{2}}$

(3) Newton's method with equality constraints

given starting point $x \in \mathrm{dom}\, f$ with $Ax = b$, tolerance $\varepsilon > 0$

repeat ① compute the newton step and decrement Δx_{nt}, $\lambda(x)$

② stopping criterion. quit if $\lambda^2/2 \leq \varepsilon$

③ Line search. Choose step size t by backtracking line search.

④ Update. $x := x + t\Delta x_{nt}$

D. Infeasible start Newton method

(1) Newton step at infeasible points

linearizing optimality conditions at infeasible x (with $x \in \mathrm{dom}\, f$)

gives
$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{nt} \\ w \end{bmatrix} = -\begin{bmatrix} \nabla f(x) \\ Ax - b \end{bmatrix}$$

primal-dual interpretation

* write optimality condition as $r(y) = 0$, where

$$y = (x, \nu), \quad r(y) = (\nabla f(x) + A^T\nu, Ax - b)$$

* linearizing $r(y) = 0$ gives $r(y + \Delta y) \approx r(y) + Dr(y)\Delta y = 0$.

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{nt} \\ \Delta \nu_{nt} \end{bmatrix} = -\begin{bmatrix} \nabla f(x) + A^T\nu \\ Ax - b \end{bmatrix}$$

same as (1) with $w = \hat{\nu} + \Delta\nu_{nt}$

(2) Infeasible start Newton method

given starting point $x \in \mathrm{dom}\, f$, ν, tolerance $\varepsilon > 0$, $\alpha \in (0, \frac{1}{2})$, $\beta \in (0, 1)$.

repeat. ① compute primal and dual Newton steps Δx_{nt}, $\Delta \nu_{nt}$

② Backtracking line search on $\|r\|_2$

$t := 1$

while $\|r(x + t\Delta x_{nt}, \nu + t\Delta\nu_{nt})\|_2 > (1 - \alpha t)\|r(x, \nu)\|_2$, $t := \beta t$

③ Update. $x := x + t\Delta x_{nt}$, $\nu := \nu + t\Delta \nu_{nt}$.

until $Ax = b$ and $\|r(x,\nu)\|_2 \le \varepsilon$

* not a descent method: $f(x^{(k+1)}) > f(x^{(k)})$ is possible
* directional derivative of $\|r(y)\|_2$ in direction $\Delta y = (\Delta x_{nt}, \Delta \nu_{nt})$

is $$\frac{d}{dt}\|r(y + t\Delta y)\|_2 \Big|_{t=0} = -\|r(y)\|_2$$

Lecture 11: Interior-point methods

A. Inequality constrained minimization

(1) $$\min. \ f_0(x)$$
$$\text{s.t.} \ f_i(x) \le 0, \ i = 1..m, \quad Ax = b$$

* f_i convex, twice continuously differentiable
* $A \in \mathbb{R}^{p \times n}$ with rank $A = p$
* we assume p^* is finite and attained
* we assume problem is strictly feasible: there exists $\tilde{x}$ with

$$\tilde{x} \in \operatorname{dom} f_0, \ f_i(\tilde{x}) < 0, \ i = 1..m, \ A\tilde{x} = b$$

hence, strong duality holds and dual optimum is attained.

B. Logarithmic barrier function and central path

(1) $$\min. \ f_0(x) + \sum_{i=1}^{m} I_-(f_i(x))$$
$$\text{s.t.} \ Ax = b$$

where $I_-(u) = 0$ if $u \le 0$, $I_-(u) = \infty$ otherwise

(2) approximation via logarithmic barrier

$$\min. \ f_0(x) - \frac{1}{t}\sum_{i=1}^{m} \log(-f_i(x))$$
$$\text{s.t.} \ Ax = b$$

* an equality constrained problem
* for $t > 0$, $-(\frac{1}{t})\log(-u)$ is a smooth approximation of I_-

* approximation improves as $t \to \infty$

(3) logarithmic barrier function

$$\phi(x) = -\sum_{i=1}^{m} \log(-f_i(x)), \quad \mathrm{dom}\,\phi = \{x \mid f_1(x) < 0, \ldots f_m(x) < 0\}$$

* convex (follows from composition rules)
* twice continuously differentiable, with derivatives

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

$$\nabla^2 \phi(x) = \sum_{i=1}^{m} \frac{1}{f_i(x)^2} \nabla f_i(x) \nabla f_i(x)^T + \sum_{i=1}^{m} \frac{1}{-f_i(x)} \nabla^2 f_i(x)$$

C. Barrier method

<u>given</u> strictly feasible x, $t := t^{(0)} > 0$, $\mu > 1$, tolerance $\varepsilon > 0$

<u>repeat</u> ① Centering step. Compute $x^*(t)$ by mini. $t f_0 + \phi$, s.t. $Ax = b$

② Update $x := x^*(t)$

③ Stopping criterion. <u>quit</u> if $\frac{m}{t} < \varepsilon$

④ increase t, $t := \mu t$.

* terminates with $f_0(x) - p^* \le \varepsilon$ (stopping criterion follows from $f_0(x^*(t)) - p^* \le \frac{m}{t}$)
* centering usually done using Newton's method, starting at current x
* choice of μ involves a trade-off: large μ means fewer outer iterations, more inner (Newton) iterations; typical values: $\mu = 10 \sim 20$
* several heuristics for choice of $t^{(0)}$.

D. Feasibility and phase I methods.

(1) basic phase I method

$$\begin{aligned} \text{mini. (over } x, s) \quad & s \\ \text{s.t.} \quad & f_i(x) \le s, \quad i = 1 \ldots m \\ & Ax = b \end{aligned}$$

* if x, s feasible, with $s < 0$, then x is strictly feasible

* if optimal ~~rate~~ value $\bar{p}^*$ of basic phase I is positive, then problem (2) original is infeasible
* if $\bar{p}^* = 0$ and attained, then problem is feasible (but not strictly)
* if $\bar{p}^* = 0$ and not attained, then problem (2) is infeasible.

(2) sum of infeasibilities phase I method

$$\begin{aligned} \text{min.} \quad & \mathbb{1}^T s \\ \text{s.t.} \quad & s \geq 0, \quad f_i(x) \leq s_i, \quad i = 1..m \\ & Ax = b. \end{aligned}$$

for infeasible problems, produces a solution that satisfies ~~ver~~ many more inequalities than basic phase I method.