LECTURE SLIDES ON

CONVEX ANALYSIS AND OPTIMIZATION

BASED ON 6.253 CLASS LECTURES AT THE

MASS. INSTITUTE OF TECHNOLOGY

CAMBRIDGE, MASS

SPRING 2014

BY DIMITRI P. BERTSEKAS

http://web.mit.edu/dimitrib/www/home.html

Based on the books

1) "Convex Optimization Theory," Athena Scientific, 2009
2) "Convex Optimization Algorithms," Athena Scientific, 2014 (in press)

Supplementary material (solved exercises, etc) at

http://www.athenasc.com/convexduality.html

LECTURE 1
AN INTRODUCTION TO THE COURSE

LECTURE OUTLINE

- The Role of Convexity in Optimization
- Duality Theory
- Algorithms and Duality
- Course Organization

HISTORY AND PREHISTORY

- Prehistory: Early 1900s - 1949.
  - Caratheodory, Minkowski, Steinitz, Farkas.
  - Properties of convex sets and functions.

- Fenchel - Rockafellar era: 1949 - mid 1980s.
  - Duality theory.
  - Minimax/game theory (von Neumann).
  - (Sub)differentiability, optimality conditions, sensitivity.

- Modern era - Paradigm shift: Mid 1980s - present.
  - Nonsmooth analysis (a theoretical/esoteric direction).
  - Algorithms (a practical/high impact direction).
  - A change in the assumptions underlying the field.

OPTIMIZATION PROBLEMS

- Generic form:

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

Cost function $f : \Re^n \mapsto \Re$, constraint set C, e.g.,

$$\begin{aligned} C = X \cap \big\{x \mid h_1(x) = 0, \ldots, h_m(x) = 0\big\} \\ \cap \big\{x \mid g_1(x) \leq 0, \ldots, g_r(x) \leq 0\big\} \end{aligned}$$

- Continuous vs discrete problem distinction

- Convex programming problems are those for which f and C are convex
  - They are continuous problems
  - They are nice, and have beautiful and intuitive structure

- However, convexity permeates all of optimization, including discrete problems

- Principal vehicle for continuous-discrete connection is duality:
  - The dual of a discrete problem is continuous/convex
  - The dual provides info for the solution of the discrete primal (e.g., lower bounds, etc)

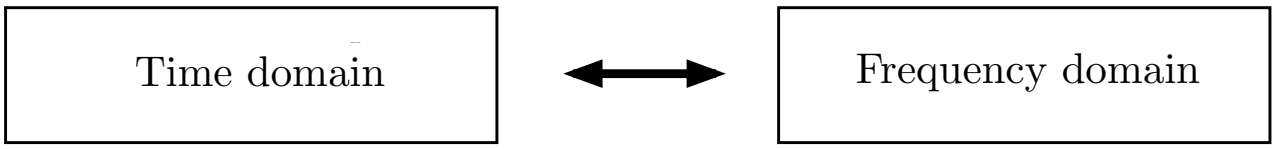
WHY IS CONVEXITY SO SPECIAL?

- A convex function has no local minima that are not global
- A nonconvex function can be "convexified" while maintaining the optimality of its global minima
- A convex set has nice "shape":
  - Nonempty relative interior
  - Connected
  - Has feasible directions at any point
- A polyhedral convex set is characterized in terms of a finite set of extreme points and extreme directions
- A real-valued convex function is continuous and has nice differentiability properties
- Closed convex cones are self-dual with respect to polarity
- Convex, lower semicontinuous functions are self-dual with respect to conjugacy
- Many important problems are convex!!

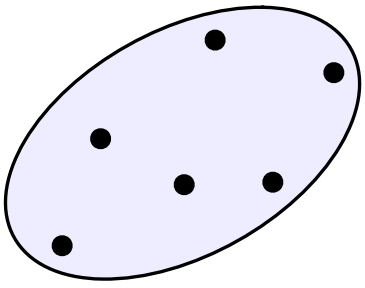
DUALITY

- Two different views of the same object.

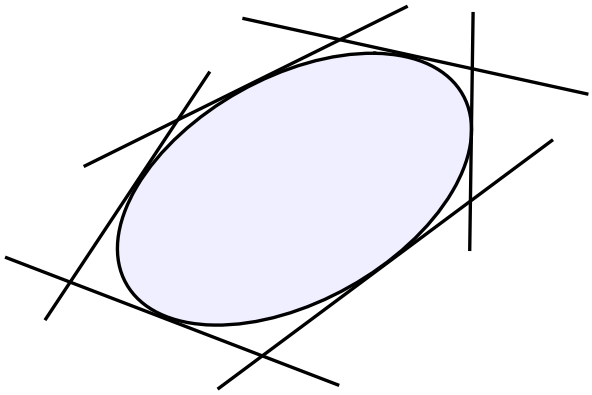
- Example: Dual description of signals.



- Dual description of **closed** convex sets



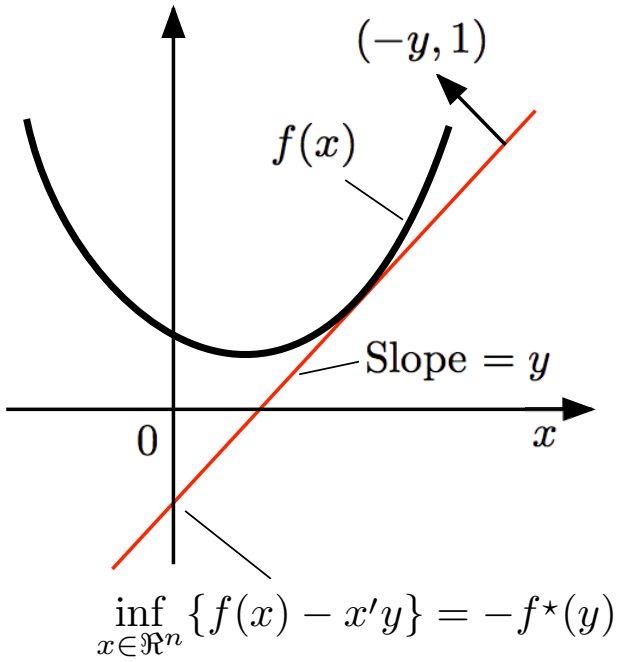
A union of points



An intersection of halfspaces

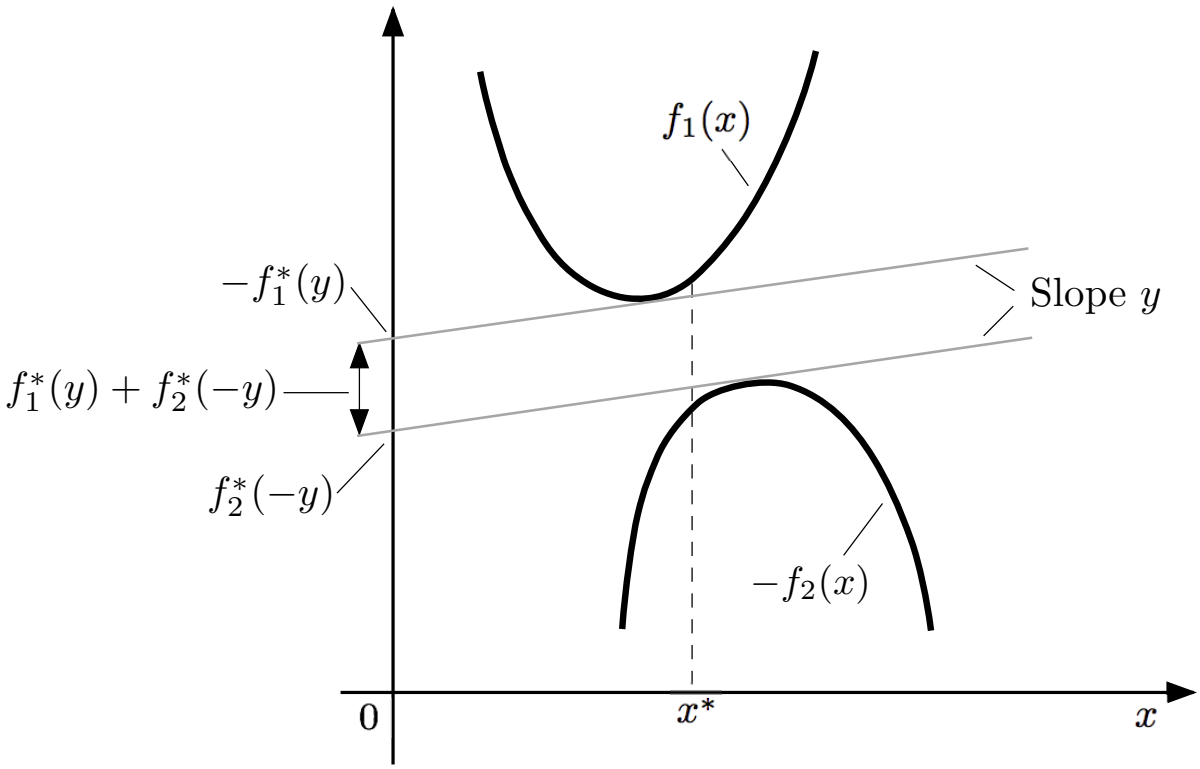
DUAL DESCRIPTION OF CONVEX FUNCTIONS

- Define a closed convex function by its epigraph.
- Describe the epigraph by hyperplanes.
- Associate hyperplanes with crossing points (the conjugate function).



Primal Description	Dual Description
Values $f(x)$	Crossing points $f^*(y)$

FENCHEL PRIMAL AND DUAL PROBLEMS



Primal Problem Description
Vertical Distances

Dual Problem Description
Crossing Point Differentials

- Primal problem:

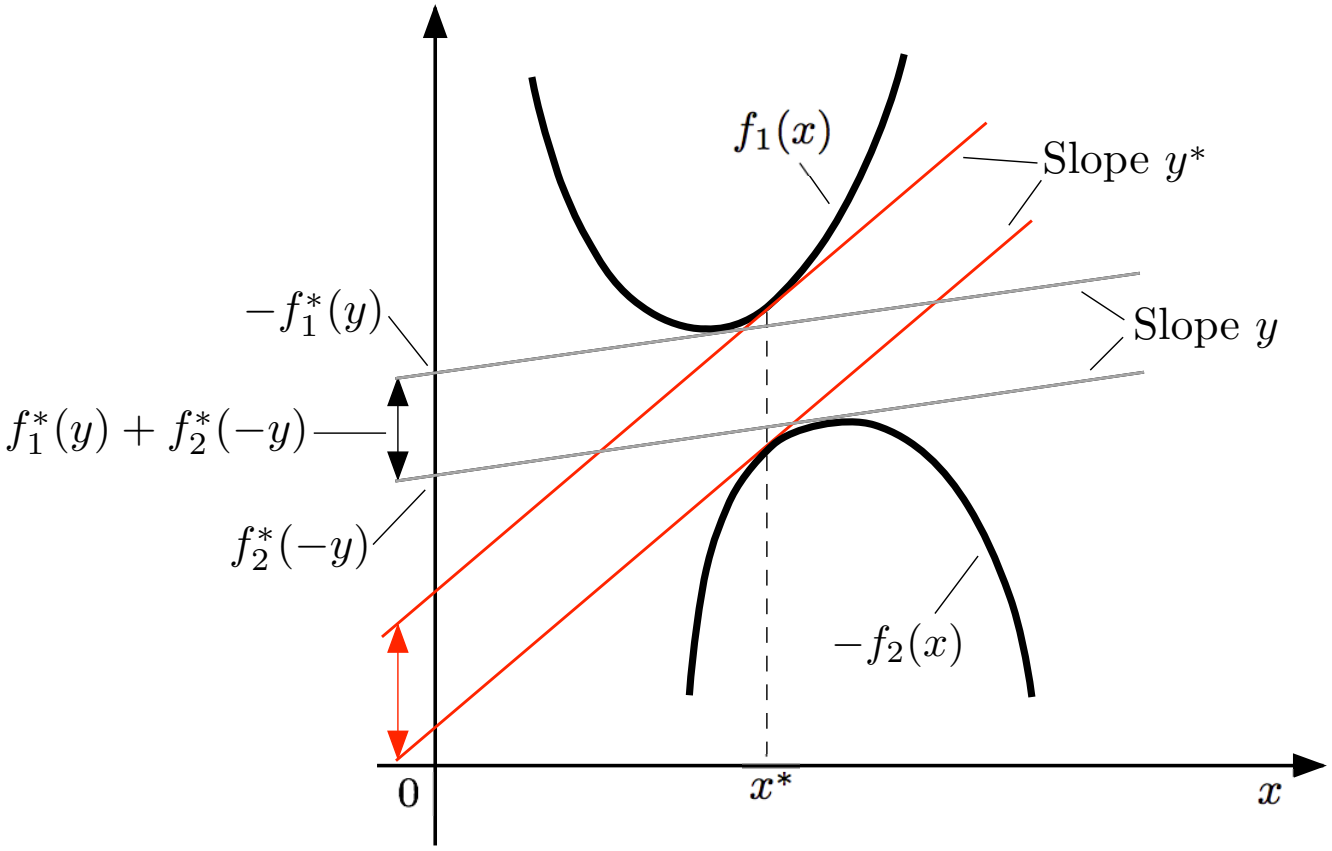
$$\min_x \big\{ f_1(x) + f_2(x) \big\}$$

- Dual problem:

$$\max_y \big\{ -f_1^*(y) - f_2^*(-y) \big\}$$

where f_1^* and f_2^* are the conjugates

FENCHEL DUALITY



$$\min_x \{f_1(x) + f_2(x)\} = \max_y \{ -f_1^\star(y) - f_2^\star(-y)\}$$

- Under favorable conditions (convexity):
  - The optimal primal and dual values are equal
  - The optimal primal and dual solutions are related

A MORE ABSTRACT VIEW OF DUALITY

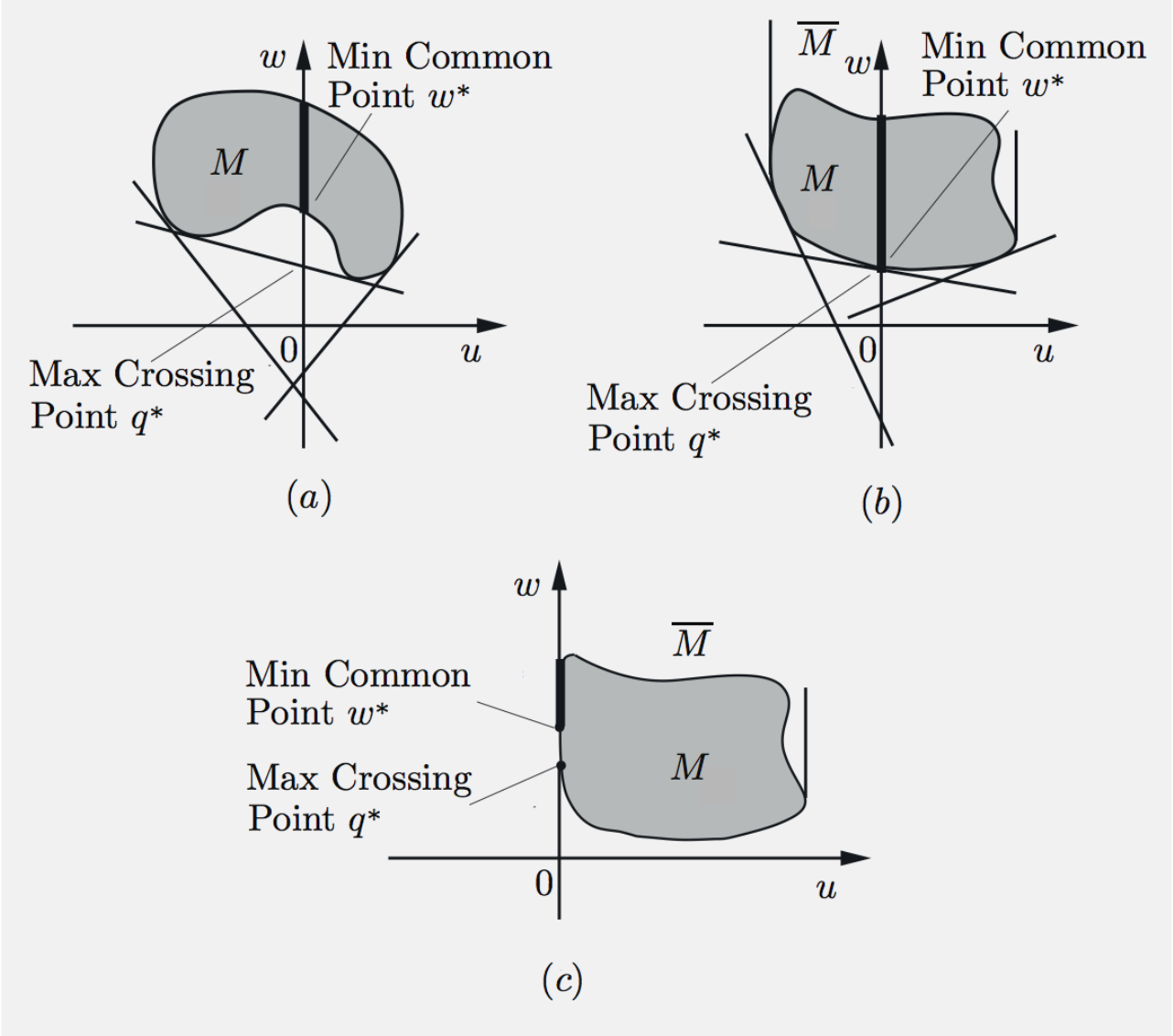
- Despite its elegance, the Fenchel framework is somewhat indirect.

- From duality of set descriptions, to
  - duality of functional descriptions, to
  - duality of problem descriptions.

- A more direct approach:
  - Start with a set, then
  - Define two simple prototype problems dual to each other.

- Skip the functional descriptions
  - A simpler, less constrained framework

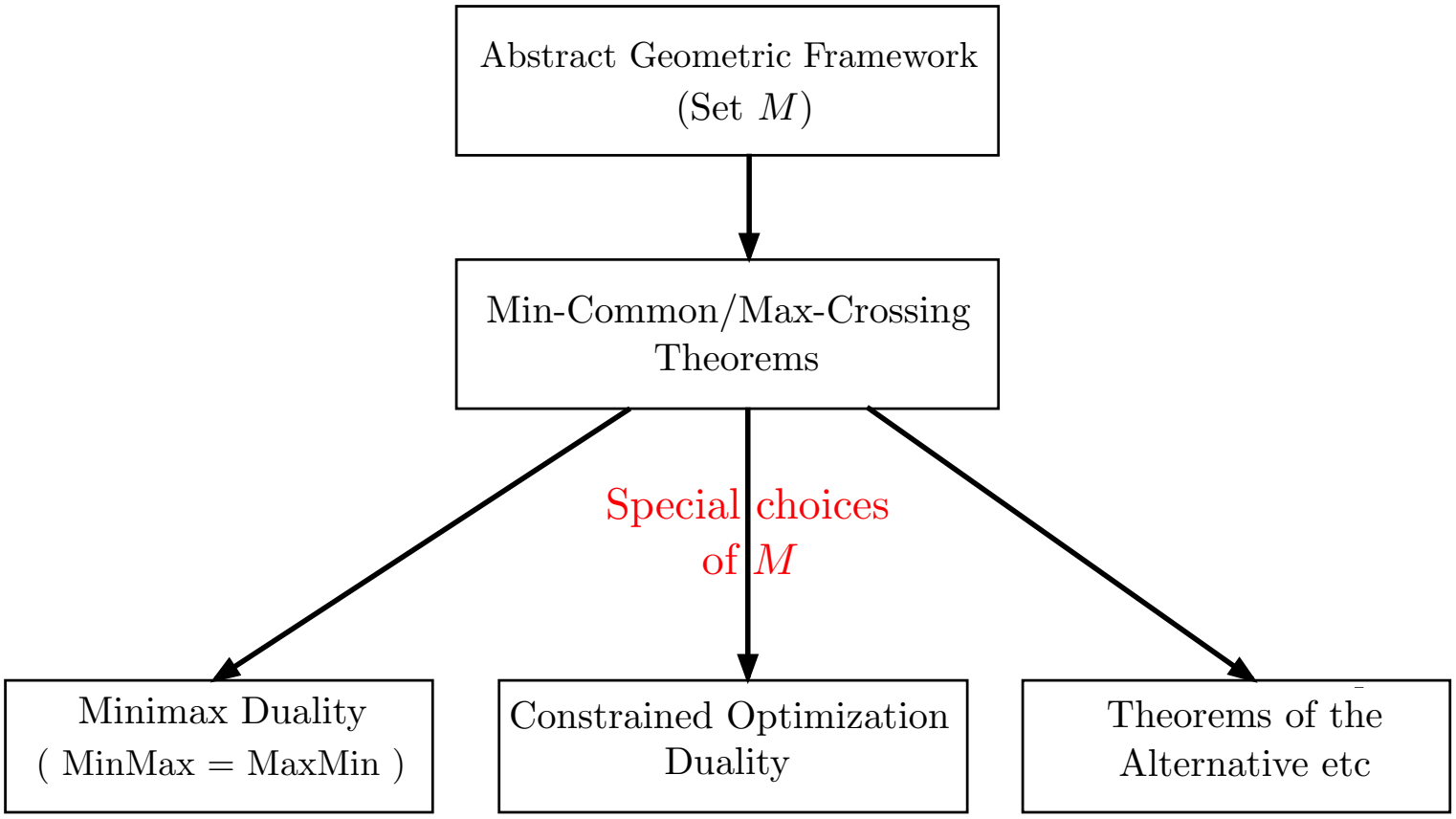
MIN COMMON/MAX CROSSING DUALITY



• All of duality theory and all of (convex/concave) minimax theory can be developed/explained in terms of this one figure.

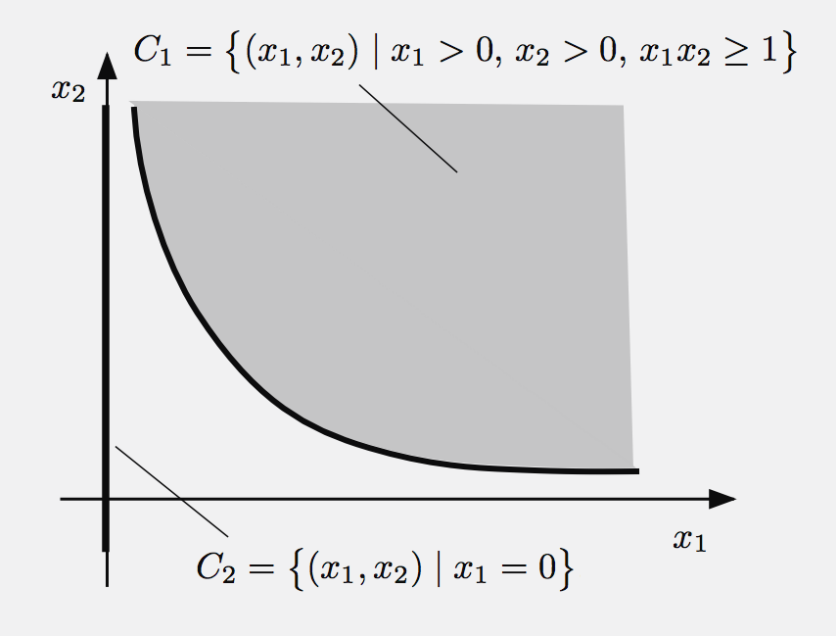
• The machinery of convex analysis is needed to flesh out this figure, and to rule out the exceptional/pathological behavior shown in (c).

ABSTRACT/GENERAL DUALITY ANALYSIS



EXCEPTIONAL BEHAVIOR

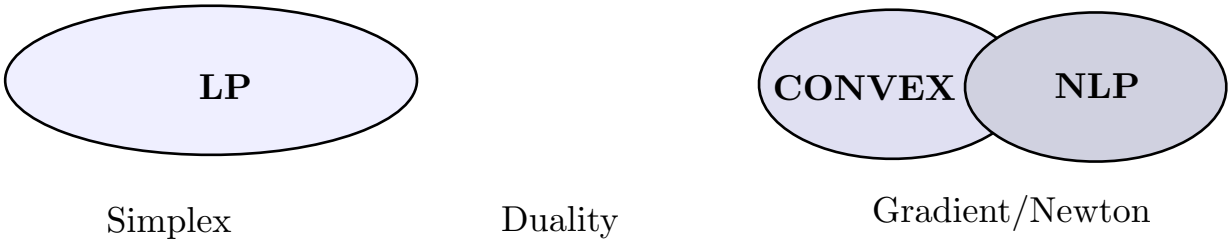
- If convex structure is so favorable, what is the source of exceptional/pathological behavior?
- **Answer:** Some common operations on convex sets do not preserve some basic properties.
- **Example:** A linearly transformed closed convex set need not be closed (if it is not polyhedral).
  - Also the vector sum of two closed convex sets need not be closed.



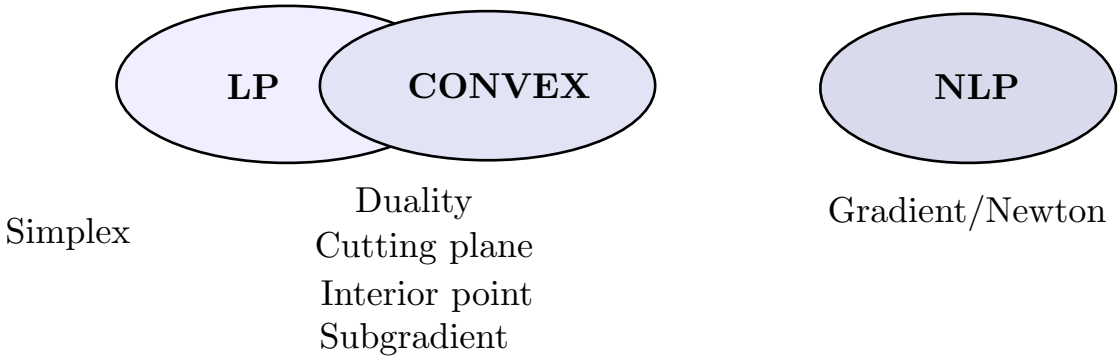
- This is a major reason for the analytical difficulties in convex analysis and pathological behavior in convex optimization (and the favorable character of polyhedral sets).

MODERN VIEW OF CONVEX OPTIMIZATION

- Traditional view: Pre 1990s
  - LPs are solved by simplex method
  - NLPs are solved by gradient/Newton methods
  - Convex programs are special cases of NLPs



- Modern view: Post 1990s
  - LPs are often solved by nonsimplex/convex methods
  - Convex problems are often solved by the same methods as LPs
  - "Key distinction is not Linear-Nonlinear but Convex-Nonconvex" (Rockafellar)



THE RISE OF THE ALGORITHMIC ERA

- Convex programs and LPs connect around
  - Duality
  - Large-scale piecewise linear problems
- Synergy of:
  - Duality
  - Algorithms
  - Applications
- New problem paradigms with rich applications
- Duality-based decomposition
  - Large-scale resource allocation
  - Lagrangian relaxation, discrete optimization
  - Stochastic programming
- Conic programming
  - Robust optimization
  - Semidefinite programming
- Machine learning
  - Support vector machines
  - l_1 regularization/Robust regression/Compressed sensing

METHODOLOGICAL TRENDS

- New methods, renewed interest in old methods.
  - Subgradient/incremental methods
  - Polyhedral approximation/cutting plane methods
  - Regularization/proximal methods
  - Interior point methods
  - Incremental methods

- Renewed emphasis on complexity analysis
  - Nesterov, Nemirovski, and others ...
  - "Optimal algorithms" (e.g., extrapolated gradient methods)

- Emphasis on interesting (often duality-related) large-scale special structures
  - Separable problems
  - Cost functions consisting of a large number of additive components
  - Many constraints

COURSE OUTLINE

- We will follow closely the textbooks
  - Bertsekas, "Convex Optimization Theory," Athena Scientific, 2009
  - Bertsekas, "Convex Optimization Algorithms," Athena Scientific, 2014 (in press)
- Additional book references:
  - Rockafellar, "Convex Analysis," 1970.
  - Boyd and Vanderbergue, "Convex Optimization," Cambridge U. Press, 2004. (On-line)
  - Bertsekas, Nedic, and Ozdaglar, "Convex Analysis and Optimization," Ath. Scientific, 2003.
- Topics :
  - **Basic Convexity:** Ch. 1 (Theory book).
  - **Convexity and Optimization:** Ch. 3.
  - **Geometric Duality Framework:** Ch. 4.
  - **Duality, Opt. Conditions:** Sect. 5.1-5.3.
  - **Overview of Problem Structures and Algorithms:** Ch. 1 (Alg. Book).
  - **Subgradient Methods:** Ch. 2.
  - **Polyhedral Approx. Methods:** Ch. 3.
  - **Proximal Methods:** Ch. 4.
  - **Additional Methods/Variants:** Ch. 5.

WHAT TO EXPECT FROM THIS COURSE

- Requirements: Homework (50%); term paper on mutually agreed subject (50%). (Midterm ?)
- We aim:
  - To develop insight and deep understanding of a fundamental optimization topic
  - To treat with mathematical rigor an important branch of methodological research, and to provide an account of the state of the art in the field
  - To get an understanding of the merits, limitations, and characteristics of the rich set of available algorithms
- Mathematical level:
  - Prerequisites are linear algebra (preferably abstract) and real analysis (a course in each)
  - Proofs will matter ... but the rich geometry of the subject helps guide the mathematics
- Applications:
  - They are many and pervasive ... but don't expect much in this course.
  - You can do your term paper on an application area

A NOTE ON THESE SLIDES

- These slides are a teaching aid, not a text
- Don't expect a rigorous mathematical development
- The statements of theorems are fairly precise, but the proofs are not
- Many proofs have been omitted or greatly abbreviated
- Figures are meant to convey and enhance understanding of ideas, not to express them precisely
- The omitted proofs and a fuller discussion can be found in the textbooks and supplementary material
- One further note: The present set of slides differs from slides for this class from earlier years in that it has a considerably stronger focus on algorithms.

LECTURE 2

LECTURE OUTLINE

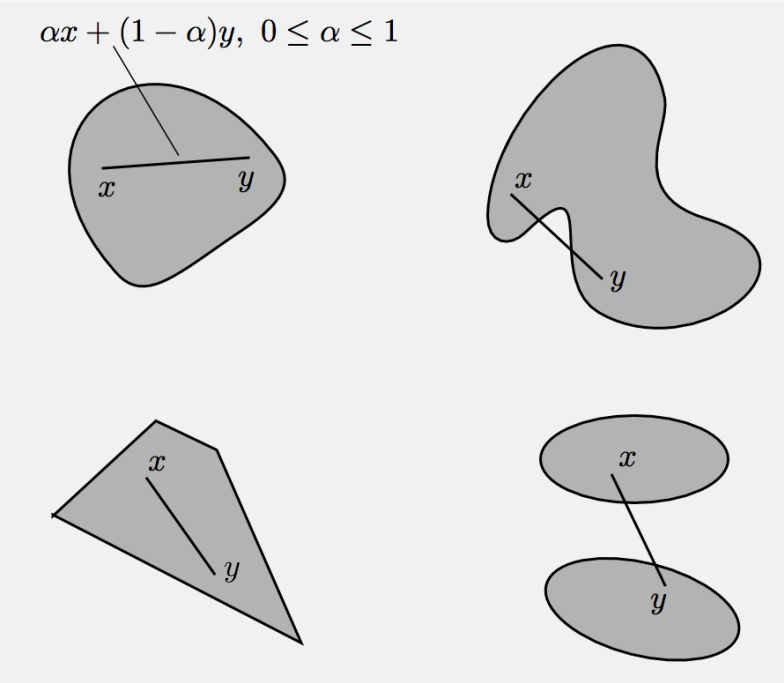
- Convex sets and functions
- Epigraphs
- Closed convex functions
- Recognizing convex functions

Reading: Section 1.1

SOME MATH CONVENTIONS

- All of our work is done in $\Re^n$: space of n-tuples $x = (x_1, \ldots, x_n)$
- All vectors are assumed column vectors
- "$'$" denotes transpose, so we use x' to denote a row vector
- $x'y$ is the inner product $\sum_{i=1}^{n} x_i y_i$ of vectors x and y
- $\|x\| = \sqrt{x'x}$ is the (Euclidean) norm of x. We use this norm almost exclusively
- See Appendix A of the textbook for an overview of the linear algebra and real analysis background that we will use. Particularly the following:
  - Definition of sup and inf of a set of real numbers
  - Convergence of sequences (definitions of lim inf, lim sup of a sequence of real numbers, and definition of lim of a sequence of vectors)
  - Open, closed, and compact sets and their properties
  - Definition and properties of differentiation

CONVEX SETS



- A subset C of $\Re^n$ is called convex if

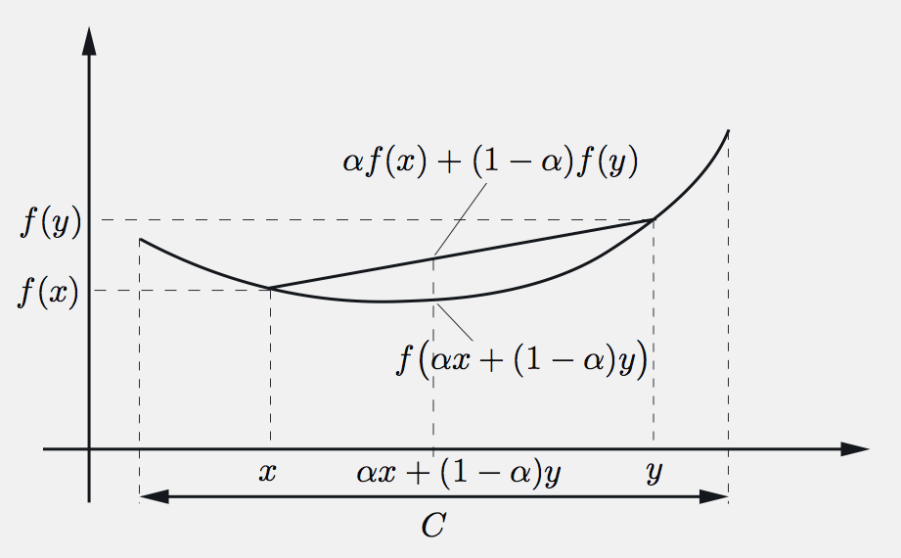
$$\alpha x + (1-\alpha)y \in C, \qquad \forall\ x, y \in C,\ \forall\ \alpha \in [0,1]$$

- Operations that preserve convexity
  - Intersection, scalar multiplication, vector sum, closure, interior, linear transformations
- Special convex sets:
  - Polyhedral sets: Nonempty sets of the form

$$\{x \mid a_j' x \le b_j,\ j = 1, \ldots, r\}$$

    (always convex, closed, not always bounded)
  - Cones: Sets C such that $\lambda x \in C$ for all $\lambda > 0$ and $x \in C$ (not always convex or closed)

CONVEX FUNCTIONS



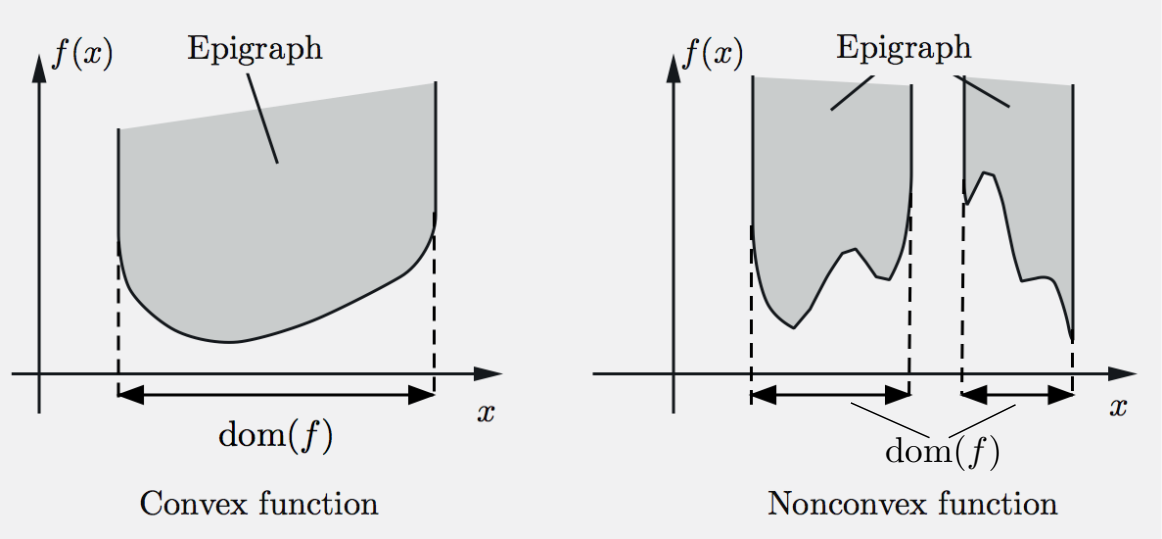
- Let C be a convex subset of $\Re^n$. A function $f : C \mapsto \Re$ is called convex if for all $\alpha \in [0, 1]$

$$f\big(\alpha x+(1-\alpha)y\big) \leq \alpha f(x)+(1-\alpha)f(y), \quad \forall\, x, y \in C$$

If the inequality is strict whenever $a \in (0, 1)$ and $x \neq y$, then f is called strictly convex over C.

- If f is a convex function, then all its level sets $\{x \in C \mid f(x) \leq \gamma\}$ and $\{x \in C \mid f(x) < \gamma\}$, where γ is a scalar, are convex.

EXTENDED REAL-VALUED FUNCTIONS



- The epigraph of a function $f : X \mapsto [-\infty, \infty]$ is the subset of $\Re^{n+1}$ given by

$$\mathrm{epi}(f) = \big\{(x, w) \mid x \in X,\, w \in \Re,\, f(x) \leq w\big\}$$

- The effective domain of f is the set

$$\mathrm{dom}(f) = \big\{x \in X \mid f(x) < \infty\big\}$$

- We say that f is convex if $\mathrm{epi}(f)$ is a convex set. If $f(x) \in \Re$ for all $x \in X$ and X is convex, the definition "coincides" with the earlier one.

- We say that f is closed if $\mathrm{epi}(f)$ is a closed set.

- We say that f is lower semicontinuous at a vector $x \in X$ if $f(x) \leq \liminf_{k\to\infty} f(x_k)$ for every sequence $\{x_k\} \subset X$ with $x_k \to x$.

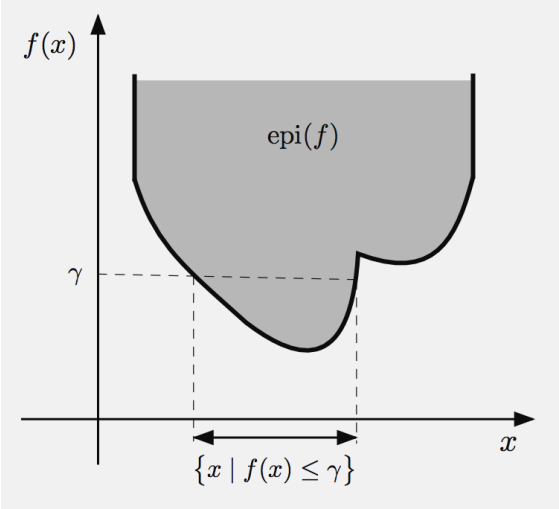
CLOSEDNESS AND SEMICONTINUITY I

• Proposition: For a function $f : \Re^n \mapsto [-\infty, \infty]$, the following are equivalent:

(i) $V_\gamma = \{x \mid f(x) \leq \gamma\}$ is closed for all $\gamma \in \Re$.

(ii) f is lower semicontinuous at all $x \in \Re^n$.

(iii) f is closed.



• (ii) $\Rightarrow$ (iii): Let $\{(x_k, w_k)\} \subset \text{epi}(f)$ with $(x_k, w_k) \to (\overline{x}, \overline{w})$. Then $f(x_k) \leq w_k$, and

$$f(\overline{x}) \leq \liminf_{k\to\infty} f(x_k) \leq \overline{w} \quad \text{so } (\overline{x}, \overline{w}) \in \text{epi}(f)$$

• (iii) $\Rightarrow$ (i): Let $\{x_k\} \subset V_\gamma$ and $x_k \to \overline{x}$. Then $(x_k, \gamma) \in \text{epi}(f)$ and $(x_k, \gamma) \to (\overline{x}, \gamma)$, so $(\overline{x}, \gamma) \in \text{epi}(f)$, and $\overline{x} \in V_\gamma$.

• (i) $\Rightarrow$ (ii): If $x_k \to \overline{x}$ and $f(\overline{x}) > \gamma > \liminf_{k\to\infty} f(x_k)$ consider subsequence $\{x_k\}_\mathcal{K} \to \overline{x}$ with $f(x_k) \leq \gamma$
- contradicts closedness of V_γ.

CLOSEDNESS AND SEMICONTINUITY II

- Lower semicontinuity of a function is a "domain-specific" property, but closeness is not:
  - If we change the domain of the function without changing its epigraph, its lower semicontinuity properties may be affected.
  - Example: Define $f : (0,1) \to [-\infty, \infty]$ and $\hat{f} : [0,1] \to [-\infty, \infty]$ by

$$f(x) = 0, \qquad \forall\, x \in (0,1),$$

$$\hat{f}(x) = \begin{cases} 0 & \text{if } x \in (0,1), \\ \infty & \text{if } x = 0 \text{ or } x = 1. \end{cases}$$

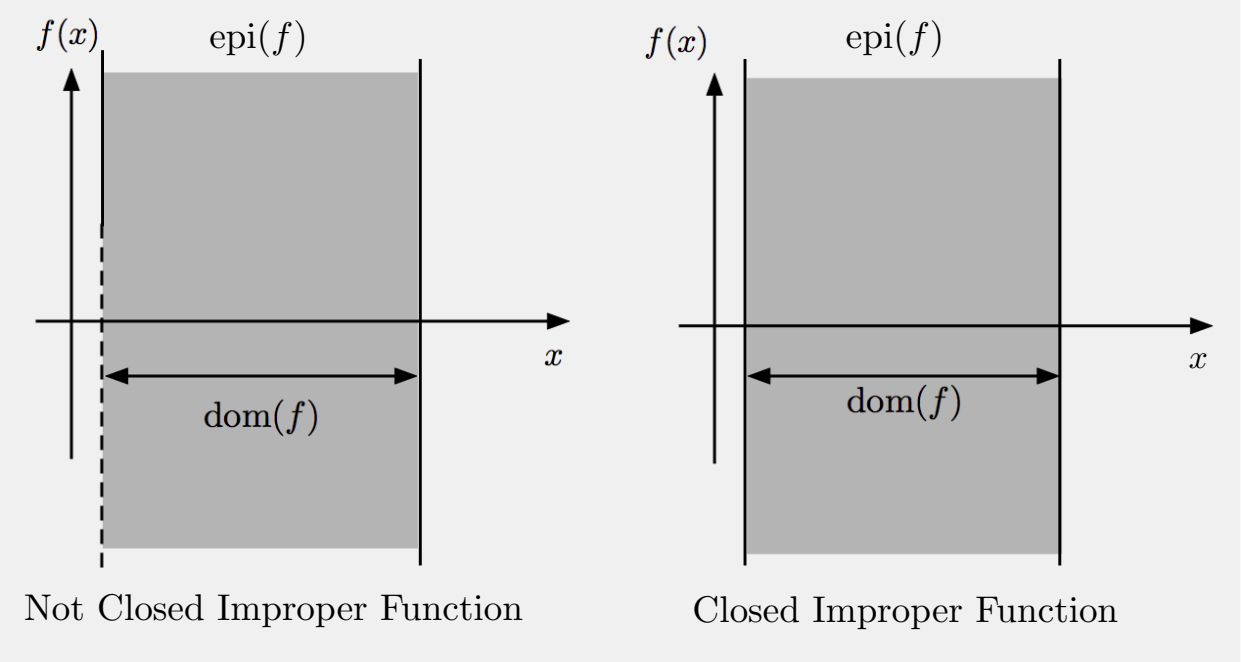
    Then f and $\hat{f}$ have the same epigraph, and both are not closed. But f is lower-semicontinuous at all x of its domain while $\hat{f}$ is not.

- Note that:
  - If f is lower semicontinuous at all $x \in \text{dom}(f)$, it is not necessarily closed
  - If f is closed, $\text{dom}(f)$ is not necessarily closed

- Proposition: Let $f : X \mapsto [-\infty, \infty]$ be a function. If $\text{dom}(f)$ is closed and f is lower semicontinuous at all $x \in \text{dom}(f)$, then f is closed.

PROPER AND IMPROPER CONVEX FUNCTIONS

- We say that f is proper if $f(x) < \infty$ for at least one $x \in X$ and $f(x) > -\infty$ for all $x \in X$, and we will call f improper if it is not proper.
- Note that f is proper if and only if its epigraph is nonempty and does not contain a "vertical line."



- An improper closed convex function is very peculiar: it takes an infinite value (∞ or $-\infty$) at every point.

RECOGNIZING CONVEX FUNCTIONS

- Some important classes of elementary convex functions: Affine functions, positive semidefinite quadratic functions, norm functions, etc.

- Proposition: (a) The function $g : \Re^n \mapsto (-\infty, \infty]$ given by

$$g(x) = \lambda_1 f_1(x) + \cdots + \lambda_m f_m(x), \qquad \lambda_i > 0$$

is convex (or closed) if $f_1, \ldots, f_m$ are convex (respectively, closed).

(b) The function $g : \Re^n \mapsto (-\infty, \infty]$ given by

$$g(x) = f(Ax)$$

where A is an $m \times n$ matrix is convex (or closed) if f is convex (respectively, closed).

(c) Consider $f_i : \Re^n \mapsto (-\infty, \infty]$, $i \in I$, where I is any index set. The function $g : \Re^n \mapsto (-\infty, \infty]$ given by

$$g(x) = \sup_{i \in I} f_i(x)$$

is convex (or closed) if the f_i are convex (respectively, closed).

LECTURE 3

LECTURE OUTLINE

- Differentiable Convex Functions
- Convex and Affine Hulls
- Caratheodory's Theorem

Reading: Sections 1.1, 1.2

DIFFERENTIABLE FUNCTIONS

- Let $f : \Re^n \mapsto \Re$ be some function. We define ith partial derivative of f at $x \in \Re^n$, by

$$\frac{\partial f}{\partial x_i}(x) = \lim_{\alpha \to 0} \frac{f(x + \alpha e_i) - f(x)}{\alpha},$$

where e_i is the ith unit vector (assuming the limit exists).

- The gradient of f at x is the column vector

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}$$

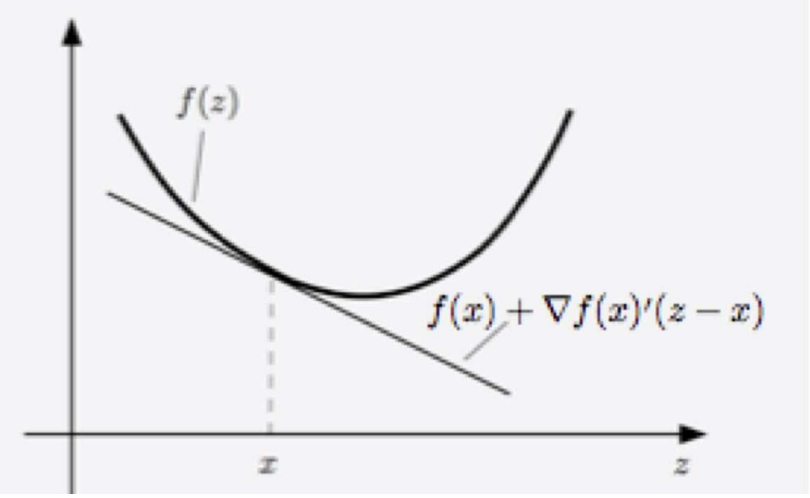
- f is called differentiable at x if $\nabla f(x)$ exists and satisfies for all $d \in \Re^n$

$$f(x + \alpha d) = f(x) + \alpha \nabla f(x)' d + o(|\alpha|), \quad \forall\ \alpha \in \Re$$

- $o(\cdot)$ Notation: $o(\|y\|)$ is a function $h : \Re^m \mapsto \Re$ s.t. for all $\{y_k\} \subset \Re^m$ with $y_k \to 0$ and $y_k \neq 0$ for all k,

$$\lim_{k \to \infty} \frac{h(y_k)}{\|y_k\|} = 0$$

DIFFERENTIABLE CONVEX FUNCTIONS



- Basic Characterization: Linear approximation based on $\nabla f(x)$ underestimates f
- Proposition: Let $C \subset \Re^n$ be a convex set and let $f : \Re^n \mapsto \Re$ be differentiable over $\Re^n$.

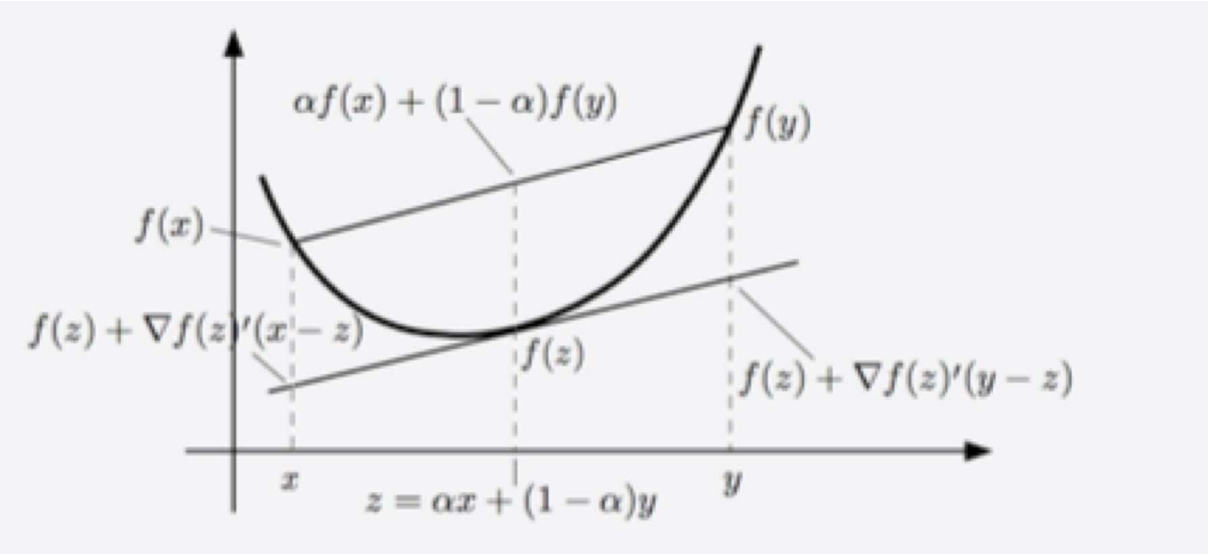
(a) The function f is convex over C iff

$$f(z) \geq f(x) + \nabla f(x)'(z - x), \qquad \forall\ x, z \in C$$

(gradient inequality for convex functions)

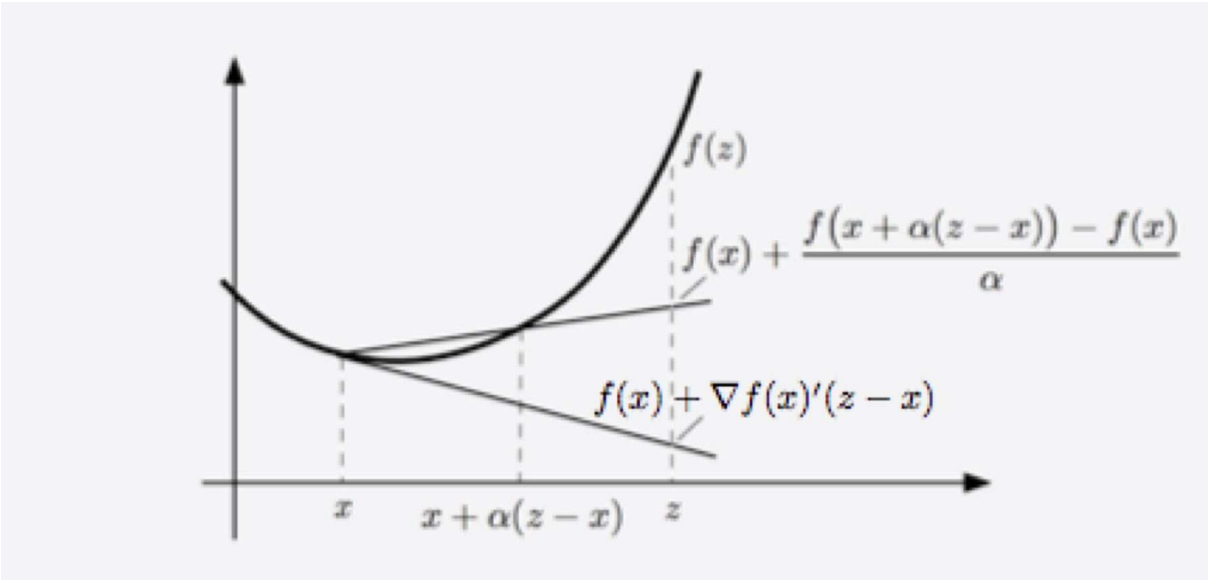
(b) If the inequality is strict whenever $x \neq z$, then f is strictly convex over C.

PROOF IDEAS



Proof that

$f(z) \geq f(x) + \nabla f(x)'(z - x), \ \forall\, x, z \quad \Rightarrow \quad f$ is convex



Proof that

f is convex $\quad \Rightarrow \quad f(z) \geq f(x) + \nabla f(x)'(z - x), \ \forall\, x, z$

OPTIMALITY CONDITION

• Let C be a nonempty convex subset of $\Re^n$ and let $f : \Re^n \mapsto \Re$ be convex and differentiable. Then:

$$x^* \in \arg\min_{x \in C} f(x) \quad \Leftrightarrow \quad \nabla f(x^*)'(x - x^*) \geq 0, \ \forall\, x \in C$$

Proof: Let the condition on the right hold. Then

$$f(x) \geq f(x^*) + \nabla f(x^*)'(x - x^*) \geq f(x^*), \quad \forall\, x \in C,$$

so x^* minimizes f over C.

Converse: Assume the contrary, i.e., x^* minimizes f over C and $\nabla f(x^*)'(x - x^*) < 0$ for some $x \in C$. By differentiation, we have

$$\lim_{\alpha \downarrow 0} \frac{f\big(x^* + \alpha(x - x^*)\big) - f(x^*)}{\alpha} = \nabla f(x^*)'(x - x^*) < 0$$

so $f\big(x^* + \alpha(x - x^*)\big)$ decreases strictly for sufficiently small $\alpha > 0$, contradicting the optimality of x^*. **Q.E.D.**

PROJECTION THEOREM

- Let C be a nonempty closed convex set in $\Re^n$.

  (a) For every $z \in \Re^n$, there exists a unique minimum of

  $$f(x) = \|z - x\|^2$$

  over all $x \in C$ (called the projection of z on C).

  (b) x^* is the projection of z if and only if

  $$(x - x^*)'(z - x^*) \leq 0, \qquad \forall\ x \in C$$

Proof: (a) f is strictly convex and has compact level sets.

(b) This is just the necessary and sufficient optimality condition

$$\nabla f(x^*)'(x - x^*) \geq 0, \qquad \forall\ x \in C.$$

TWICE DIFFERENTIABLE CONVEX FNS

- Let C be a convex subset of $\Re^n$ and let $f : \Re^n \mapsto \Re$ be twice continuously differentiable.

  (a) If $\nabla^2 f(x)$ is positive semidefinite for all $x \in C$, then f is convex over C.

  (b) If $\nabla^2 f(x)$ is positive definite for all $x \in C$, then f is strictly convex over C.

  (c) If C is open and f is convex over C, then $\nabla^2 f(x)$ is positive semidefinite for all $x \in C$.

Proof: (a) By mean value theorem, for $x, y \in C$

$$f(y) = f(x) + \nabla f(x)'(y-x) + \tfrac{1}{2}(y-x)'\nabla^2 f\Big(x+\alpha(y-x)\Big)(y-x)$$

for some $\alpha \in [0,1]$. Using the positive semidefiniteness of $\nabla^2 f$, we obtain

$$f(y) \geq f(x) + \nabla f(x)'(y-x), \qquad \forall\ x, y \in C$$

This is the gradient inequality, so f is convex.

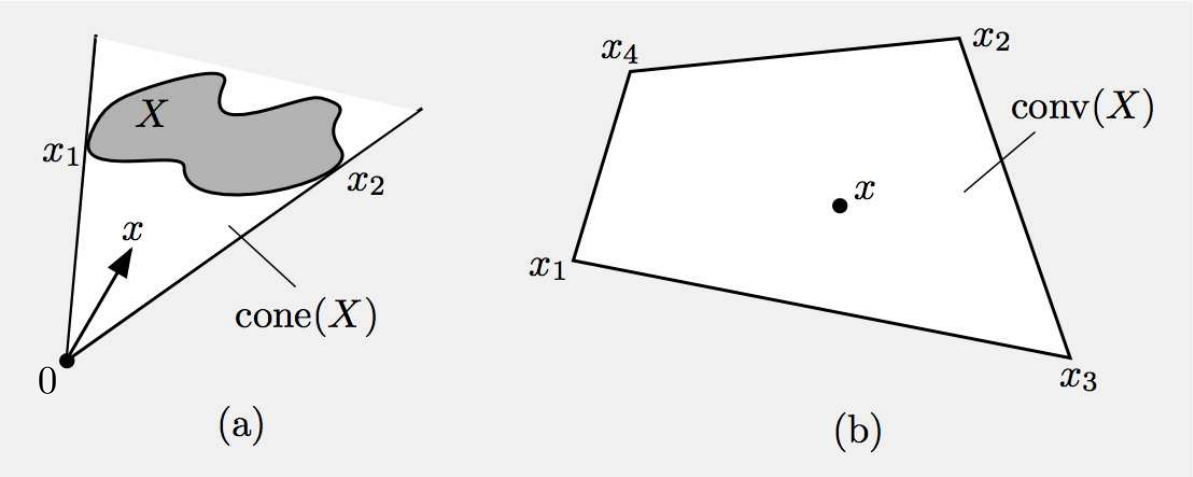
(b) Similar to (a), $f(y) > f(x) + \nabla f(x)'(y-x)$ for all $x, y \in C$ with $x \neq y$, and we use the gradient inequality result.

(c) By contradiction ... similar.

CONVEX AND AFFINE HULLS

- Given a set $X \subset \Re^n$:
- A convex combination of elements of X is a vector of the form $\sum_{i=1}^m \alpha_i x_i$, where $x_i \in X$, $\alpha_i \geq 0$, and $\sum_{i=1}^m \alpha_i = 1$.
- The convex hull of X, denoted $\text{conv}(X)$, is the intersection of all convex sets containing X. (Can be shown to be equal to the set of all convex combinations from X).
- The affine hull of X, denoted $\text{aff}(X)$, is the intersection of all affine sets containing X (an affine set is a set of the form $\overline{x} + S$, where S is a subspace).
- A nonnegative combination of elements of X is a vector of the form $\sum_{i=1}^m \alpha_i x_i$, where $x_i \in X$ and $\alpha_i \geq 0$ for all i.
- The cone generated by X, denoted $\text{cone}(X)$, is the set of all nonnegative combinations from X:
  - It is a convex cone containing the origin.
  - It need not be closed!
  - If X is a finite set, $\text{cone}(X)$ is closed (nontrivial to show!)

CARATHEODORY'S THEOREM



- Let X be a nonempty subset of $\Re^n$.

  (a) Every $x \neq 0$ in cone(X) can be represented as a positive combination of vectors $x_1, \ldots, x_m$ from X that are linearly independent (so $m \leq n$).

  (b) Every $x \in \text{conv}(X)$ can be represented as a convex combination of vectors $x_1, \ldots, x_m$ from X with $m \leq n + 1$.

PROOF OF CARATHEODORY'S THEOREM

(a) Let $x \neq 0$ belong to cone(X), and let m be the smallest integer such that $x = \sum_{i=1}^{m} \alpha_i x_i$, where $\alpha_i > 0$ and $x_i \in X$, $i = 1, \ldots, m$.

If the x_i were linearly dependent, there would exist $\lambda_1, \ldots, \lambda_m$, with

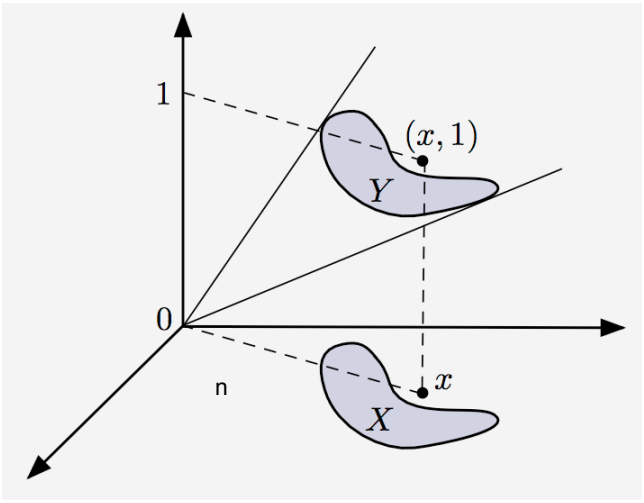
$$\sum_{i=1}^{m} \lambda_i x_i = 0$$

and at least one of the λ_i is positive. We have

$$x = \sum_{i=1}^{m} (\alpha_i - \overline{\gamma}\lambda_i) x_i,$$

where $\overline{\gamma}$ is the largest γ such that $\alpha_i - \gamma\lambda_i \geq 0$ for all i. This represents x as a positive combination of fewer than m vectors of X – a contradiction. Therefore, $x_1, \ldots, x_m$, are linearly independent.

(b) Apply part (a) to $Y = \{(x, 1) \mid x \in X\}$.



AN APPLICATION OF CARATHEODORY

- The convex hull of a closed set need not be closed! But ...
- The convex hull of a compact set is compact.

Proof: Let X be compact. We take a sequence in $\text{conv}(X)$ and show that it has a convergent subsequence whose limit is in $\text{conv}(X)$.

By Caratheodory, a sequence in $\text{conv}(X)$ can be expressed as $\left\{\sum_{i=1}^{n+1} \alpha_i^k x_i^k\right\}$, where for all k and i, $\alpha_i^k \geq 0$, $x_i^k \in X$, and $\sum_{i=1}^{n+1} \alpha_i^k = 1$. Since the sequence

$$\left\{(\alpha_1^k, \ldots, \alpha_{n+1}^k, x_1^k, \ldots, x_{n+1}^k)\right\}$$

is bounded, it has a limit point

$$\left\{(\alpha_1, \ldots, \alpha_{n+1}, x_1, \ldots, x_{n+1})\right\},$$

which must satisfy $\sum_{i=1}^{n+1} \alpha_i = 1$, and $\alpha_i \geq 0$, $x_i \in X$ for all i.

The vector $\sum_{i=1}^{n+1} \alpha_i x_i$ belongs to $\text{conv}(X)$ and is a limit point of $\left\{\sum_{i=1}^{n+1} \alpha_i^k x_i^k\right\}$, showing that $\text{conv}(X)$ is compact. **Q.E.D.**

LECTURE 4

LECTURE OUTLINE

- Relative interior and closure
- Algebra of relative interiors and closures
- Directions of recession

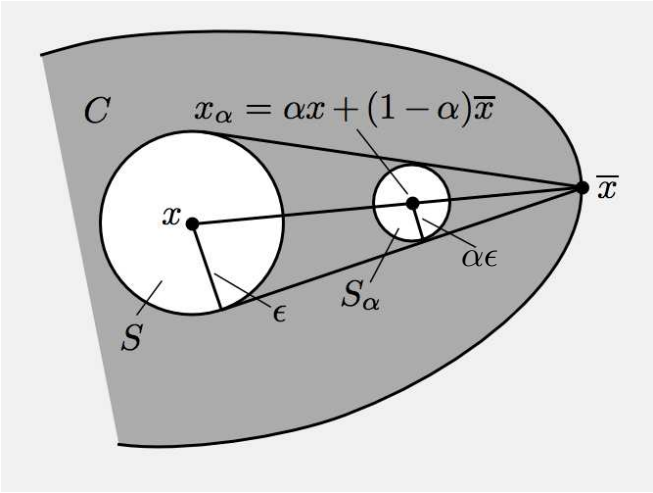
Reading: Section 1.3.1 and Section 1.4 up to (but not including) Section 1.4.1

Two key facts about convex sets:

- A convex set has nonempty interior (when viewed relative to its affine hull)
- A convex set has nice behavior "at ∞": If a closed convex set contains a half line that starts at one of its points, it contains every translation that starts at another one of its points

RELATIVE INTERIOR

- x is a relative interior point of C, if x is an interior point of C relative to aff(C).
- ri(C) denotes the relative interior of C, i.e., the set of all relative interior points of C.
- Line Segment Principle: If C is a convex set, $x \in \text{ri}(C)$ and $\overline{x} \in \text{cl}(C)$, then all points on the line segment connecting x and $\overline{x}$, except possibly $\overline{x}$, belong to ri(C).



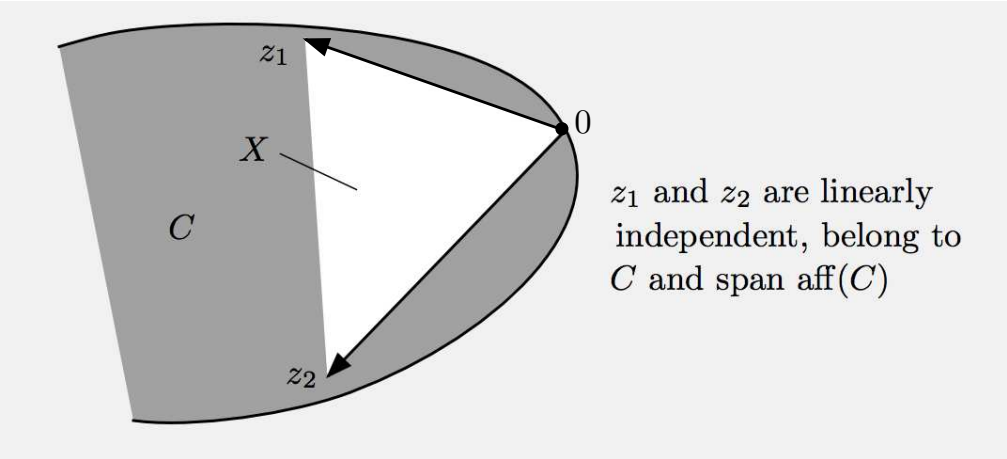
- Proof of case where $\overline{x} \in C$: See the figure.
- Proof of case where $\overline{x} \notin C$: Take sequence $\{x_k\} \subset C$ with $x_k \to \overline{x}$. Argue as in the figure.

ADDITIONAL MAJOR RESULTS

- Let C be a nonempty convex set.

  (a) ri(C) is a nonempty convex set, and has the same affine hull as C.

  (b) Prolongation Lemma: $x \in \text{ri}(C)$ if and only if every line segment in C having x as one endpoint can be prolonged beyond x without leaving C.



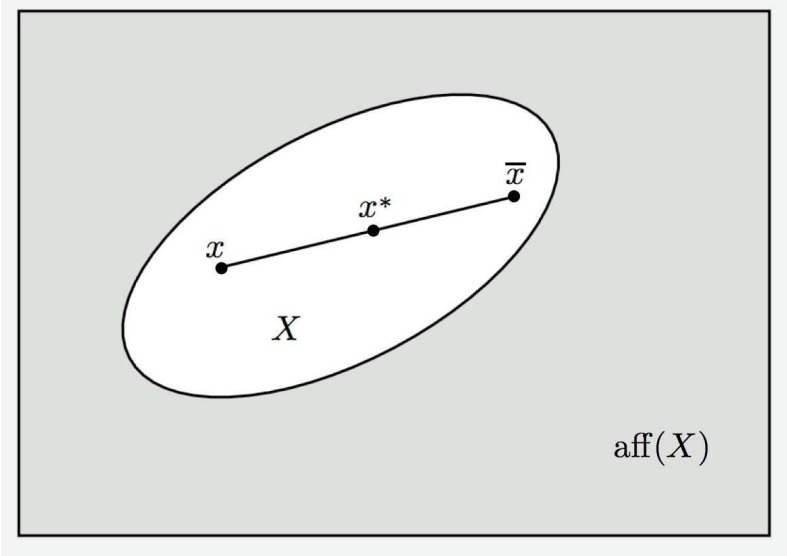
Proof: (a) Assume $0 \in C$. Choose m linearly independent vectors $z_1, \ldots, z_m \in C$, where $m =$ dimension(aff(C)). Prove that $X \subset \text{ri}(C)$, where

$$X = \left\{ \sum_{i=1}^{m} \alpha_i z_i \;\middle|\; \sum_{i=1}^{m} \alpha_i < 1,\ \alpha_i > 0,\ i = 1, \ldots, m \right\}$$

(b) => is clear by the def. of rel. interior. Reverse: take any $\overline{x} \in \text{ri}(C)$; use Line Segment Principle.

OPTIMIZATION APPLICATION

• A concave function $f : \Re^n \mapsto \Re$ that attains its minimum over a convex set X at an $x^* \in \text{ri}(X)$ must be constant over X.



Proof: (By contradiction) Let $x \in X$ be such that $f(x) > f(x^*)$. Prolong beyond x^* the line segment x-to-x^* to a point $\overline{x} \in X$. By concavity of f, we have for some $\alpha \in (0, 1)$

$$f(x^*) \geq \alpha f(x) + (1 - \alpha) f(\overline{x}),$$

and since $f(x) > f(x^*)$, we must have $f(x^*) > f(\overline{x})$ - a contradiction. **Q.E.D.**

• Corollary: A linear function $f(x) = c'x$, $c \neq 0$, cannot attain a minimum at an interior point of a convex set.

CALCULUS OF REL. INTERIORS: SUMMARY

- The $\text{ri}(C)$ and $\text{cl}(C)$ of a convex set C "differ very little."
  - $\text{ri}(C) = \text{ri}\big(\text{cl}(C)\big), \qquad \text{cl}(C) = \text{cl}\big(\text{ri}(C)\big)$
  - Any point in $\text{cl}(C)$ can be approximated arbitrarily closely by a point in $\text{ri}(C)$.

- Relative interior and closure commute with Cartesian product.

- Relative interior commutes with image under a linear transformation and vector sum, but closure does not.

- Neither relative interior nor closure commute with set intersection.

- "Good" operations: Cartesian product for both, and image for relative interior.

- "Bad" operations: Set intersection for both, and image for closure (need additional assumptions for equality).

CLOSURE VS RELATIVE INTERIOR

- Proposition:

  (a) We have $\text{cl}(C) = \text{cl}\big(\text{ri}(C)\big)$ and $\text{ri}(C) = \text{ri}\big(\text{cl}(C)\big)$.

  (b) Let $\overline{C}$ be another nonempty convex set. Then the following three conditions are equivalent:

    (i) C and $\overline{C}$ have the same rel. interior.

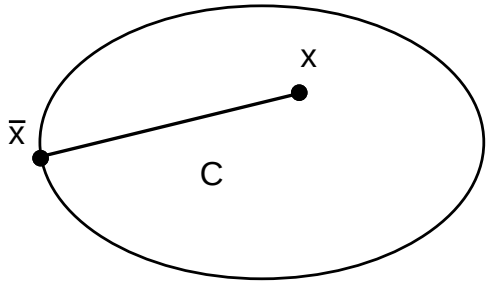
    (ii) C and $\overline{C}$ have the same closure.

    (iii) $\text{ri}(C) \subset \overline{C} \subset \text{cl}(C)$.

Proof: (a) Since $\text{ri}(C) \subset C$, we have $\text{cl}\big(\text{ri}(C)\big) \subset \text{cl}(C)$. Conversely, let $\overline{x} \in \text{cl}(C)$. Let $x \in \text{ri}(C)$. By the Line Segment Principle, we have

$$\alpha x + (1-\alpha)\overline{x} \in \text{ri}(C), \qquad \forall\ \alpha \in (0,1].$$

Thus, $\overline{x}$ is the limit of a sequence that lies in $\text{ri}(C)$, so $\overline{x} \in \text{cl}\big(\text{ri}(C)\big)$.



The proof of $\text{ri}(C) = \text{ri}\big(\text{cl}(C)\big)$ is similar.

LINEAR TRANSFORMATIONS

• Let C be a nonempty convex subset of $\Re^n$ and let A be an $m \times n$ matrix.

(a) We have $A \cdot \mathrm{ri}(C) = \mathrm{ri}(A \cdot C)$.

(b) We have $A \cdot \mathrm{cl}(C) \subset \mathrm{cl}(A \cdot C)$. Furthermore, if C is bounded, then $A \cdot \mathrm{cl}(C) = \mathrm{cl}(A \cdot C)$.

Proof: (a) Intuition: Spheres within C are mapped onto spheres within $A \cdot C$ (relative to the affine hull).

(b) We have $A \cdot \mathrm{cl}(C) \subset \mathrm{cl}(A \cdot C)$, since if a sequence $\{x_k\} \subset C$ converges to some $x \in \mathrm{cl}(C)$ then the sequence $\{Ax_k\}$, which belongs to $A \cdot C$, converges to Ax, implying that $Ax \in \mathrm{cl}(A \cdot C)$.

To show the converse, assuming that C is bounded, choose any $z \in \mathrm{cl}(A \cdot C)$. Then, there exists $\{x_k\} \subset C$ such that $Ax_k \to z$. Since C is bounded, $\{x_k\}$ has a subsequence that converges to some $x \in \mathrm{cl}(C)$, and we must have $Ax = z$. It follows that $z \in A \cdot \mathrm{cl}(C)$. **Q.E.D.**

Note that in general, we may have

$$A \cdot \mathrm{int}(C) \neq \mathrm{int}(A \cdot C), \qquad A \cdot \mathrm{cl}(C) \neq \mathrm{cl}(A \cdot C)$$

VECTOR SUMS AND INTERSECTIONS

- Let C_1 and C_2 be nonempty convex sets.

(a) We have

$$\text{ri}(C_1 + C_2) = \text{ri}(C_1) + \text{ri}(C_2),$$

$$\text{cl}(C_1) + \text{cl}(C_2) \subset \text{cl}(C_1 + C_2)$$

If one of C_1 and C_2 is bounded, then

$$\text{cl}(C_1) + \text{cl}(C_2) = \text{cl}(C_1 + C_2)$$

(b) We have

$$\text{ri}(C_1) \cap \text{ri}(C_2) \subset \text{ri}(C_1 \cap C_2), \quad \text{cl}(C_1 \cap C_2) \subset \text{cl}(C_1) \cap \text{cl}(C_2)$$

If $\text{ri}(C_1) \cap \text{ri}(C_2) \neq \emptyset$, then

$$\text{ri}(C_1 \cap C_2) = \text{ri}(C_1) \cap \text{ri}(C_2), \quad \text{cl}(C_1 \cap C_2) = \text{cl}(C_1) \cap \text{cl}(C_2)$$

Proof of (a): $C_1 + C_2$ is the result of the linear transformation $(x_1, x_2) \mapsto x_1 + x_2$.

- Counterexample for (b):

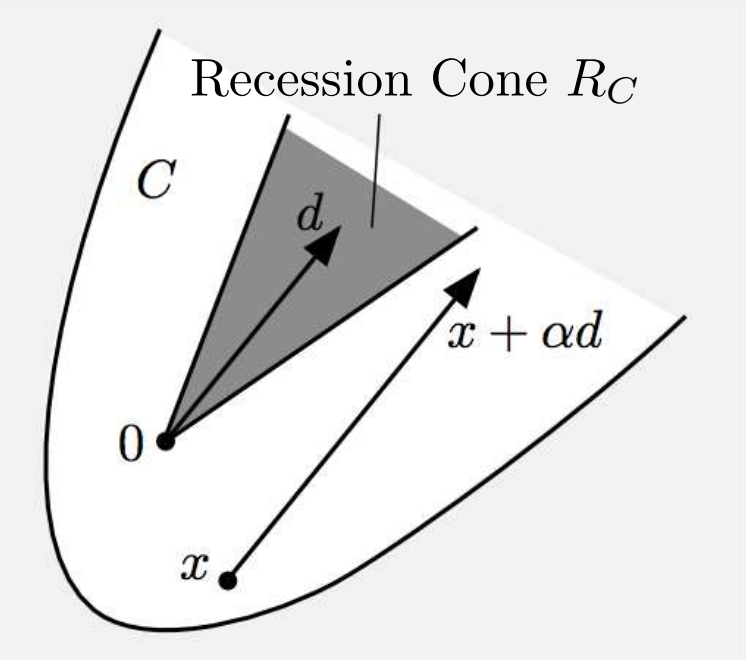
$$C_1 = \{x \mid x \leq 0\}, \qquad C_2 = \{x \mid x \geq 0\}$$

$$C_1 = \{x \mid x < 0\}, \qquad C_2 = \{x \mid x > 0\}$$

RECESSION CONE OF A CONVEX SET

- Given a nonempty convex set C, a vector d is a direction of recession if starting at **any** x in C and going indefinitely along d, we never cross the relative boundary of C to points outside C:

$$x + \alpha d \in C, \qquad \forall\ x \in C, \quad \forall\ \alpha \geq 0$$



- Recession cone of C (denoted by R_C): The set of all directions of recession.

- R_C is a cone containing the origin.

RECESSION CONE THEOREM

- Let C be a nonempty closed convex set.

  (a) The recession cone R_C is a closed convex cone.

  (b) A vector d belongs to R_C if and only if there exists some vector $x \in C$ such that $x + \alpha d \in C$ for all $\alpha \geq 0$.

  (c) C is compact if and only if $R_C = \{0\}$.

  (d) If D is another closed convex set such that $C \cap D \neq \emptyset$, we have

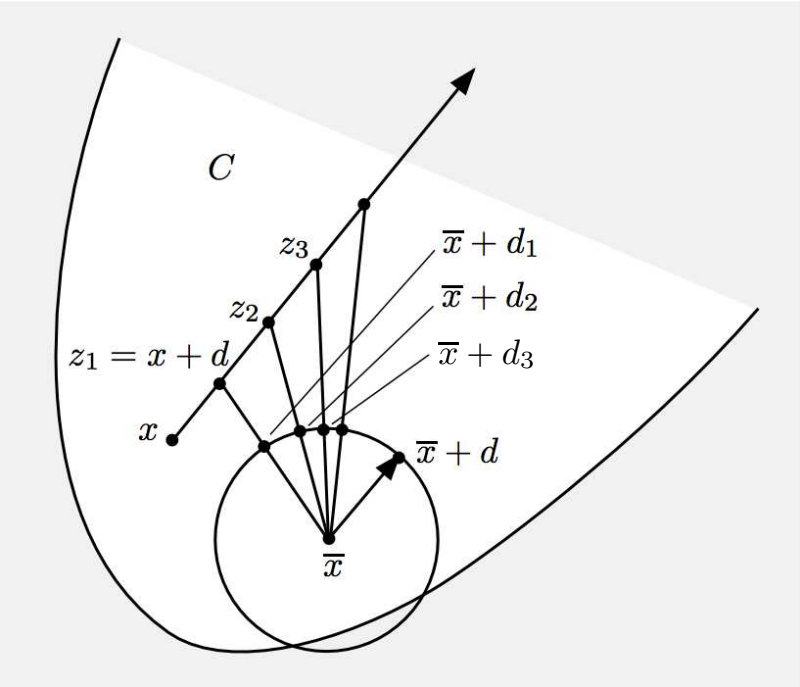
$$R_{C \cap D} = R_C \cap R_D$$

  More generally, for any collection of closed convex sets C_i, $i \in I$, where I is an arbitrary index set and $\cap_{i \in I} C_i$ is nonempty, we have

$$R_{\cap_{i \in I} C_i} = \cap_{i \in I} R_{C_i}$$

- Note an important fact: A nonempty intersection of closed sets $\cap_{i \in I} C_i$ is compact if and only if $\cap_{i \in I} R_{C_i} = \{0\}$.

PROOF OF PART (B)



• Let $d \neq 0$ be such that there exists a vector $x \in C$ with $x + \alpha d \in C$ for all $\alpha \geq 0$. We fix $\overline{x} \in C$ and $\alpha > 0$, and we show that $\overline{x} + \alpha d \in C$. By scaling d, it is enough to show that $\overline{x} + d \in C$.

For $k = 1, 2, \ldots$, let

$$z_k = x + kd, \qquad d_k = \frac{(z_k - \overline{x})}{\|z_k - \overline{x}\|}\|d\|$$

We have

$$\frac{d_k}{\|d\|} = \frac{\|z_k - x\|}{\|z_k - \overline{x}\|}\frac{d}{\|d\|} + \frac{x - \overline{x}}{\|z_k - \overline{x}\|}, \quad \frac{\|z_k - x\|}{\|z_k - \overline{x}\|} \to 1, \quad \frac{x - \overline{x}}{\|z_k - \overline{x}\|} \to 0,$$

so $d_k \to d$ and $\overline{x} + d_k \to \overline{x} + d$. Use the convexity and closedness of C to conclude that $\overline{x} + d \in C$.

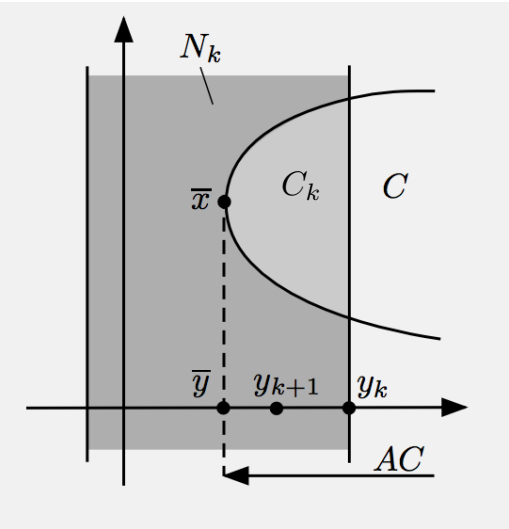
APPLICATION: CLOSURE OF $A \cdot C$

- Let C be a nonempty closed convex, and let A be a matrix with nullspace $N(A)$. Then AC is closed if $R_C \cap N(A) = \{0\}$.

Proof: Let $\{y_k\} \subset AC$ with $y_k \to \overline{y}$. Define the nested sequence $C_k = C \cap N_k$, where

$$N_k = \{x \mid Ax \in W_k\}, \quad W_k = \left\{z \mid \|z-\overline{y}\| \leq \|y_k-\overline{y}\|\right\}$$

We have $R_{N_k} = N(A)$, $R_{C_k} = R_C \cap N(A) = \{0\}$, so C_k is compact, and $\{C_k\}$ has nonempty intersection. **Q.E.D.**



- A special case: $C_1 + C_2$ is closed if C_1, C_2 are closed and one of the two is compact. [Write $C_1+C_2 = A(C_1 \times C_2)$, where $A(x_1, x_2) = x_1+x_2$.]

- Related theorem: $A \cdot C$ is closed if C is polyhedral. Can be shown by a more refined method (see the text), or by other methods.

LECTURE 5

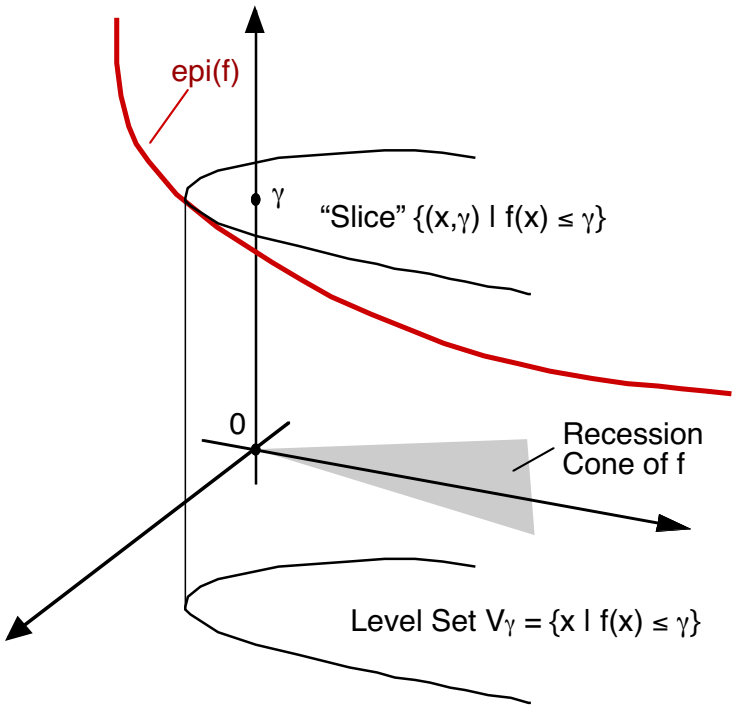
LECTURE OUTLINE

- Directions of recession of convex functions
- Local and global minima
- Existence of optimal solutions

Reading: Sections 1.4.1, 3.1, 3.2

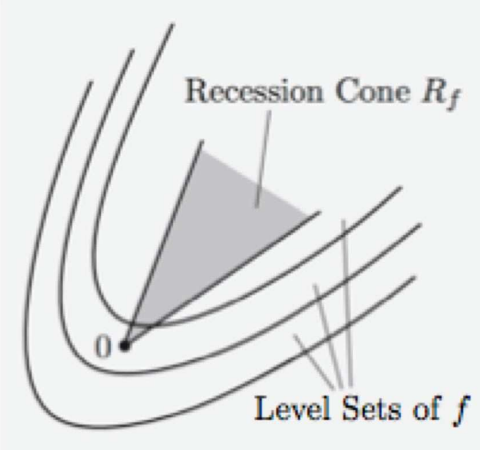
DIRECTIONS OF RECESSION OF A FN

- We aim to characterize directions of monotonic decrease of convex functions.
- Some basic geometric observations:
  - The "horizontal directions" in the recession cone of the epigraph of a convex function f are directions along which the level sets are unbounded.
  - All the nonempty level sets $\{x \mid f(x) \leq \gamma\}$ are unbounded along these same directions.
  - f is monotonically nonincreasing along these directions.
- These are the directions of recession of f.



RECESSION CONE OF LEVEL SETS

• Proposition: Let $f : \Re^n \mapsto (-\infty, \infty]$ be a closed proper convex function and consider the level sets $V_\gamma = \{x \mid f(x) \leq \gamma\}$, where γ is a scalar. Then:



(a) All the nonempty level sets V_γ have the same recession cone, denoted R_f, and called the recession cone of f:

$$R_{V_\gamma} = R_f = \{d \mid (d, 0) \in R_{\text{epi}(f)}\}$$

(b) If one nonempty level set V_γ is compact, then all level sets are compact.

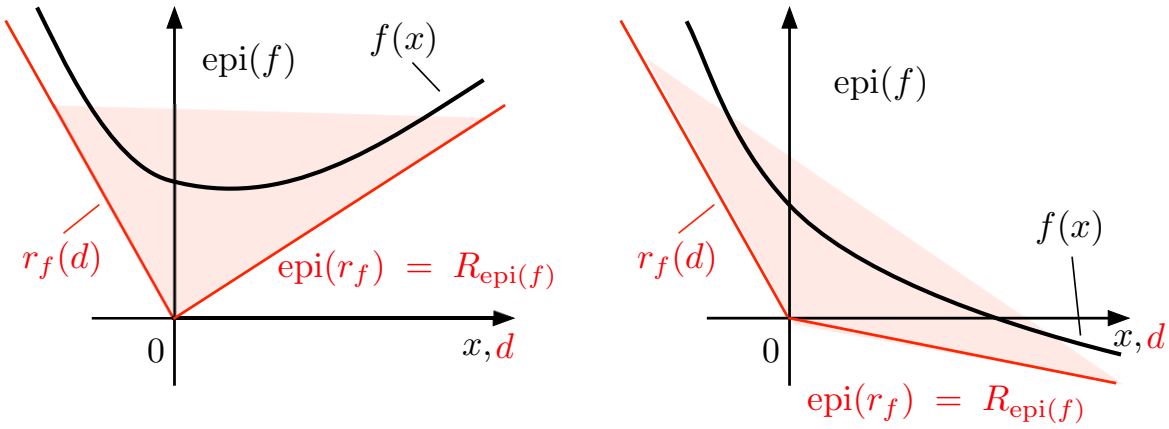
Proof: (a) Just translate to math the fact that

R_{V_γ} = the "horizontal" directions of recession of $\text{epi}(f)$

(b) This is the case where $R_{V_\gamma} = \{(0, 0)\}$ for all γ such that V_γ is nonempty.

RECESSION FUNCTION

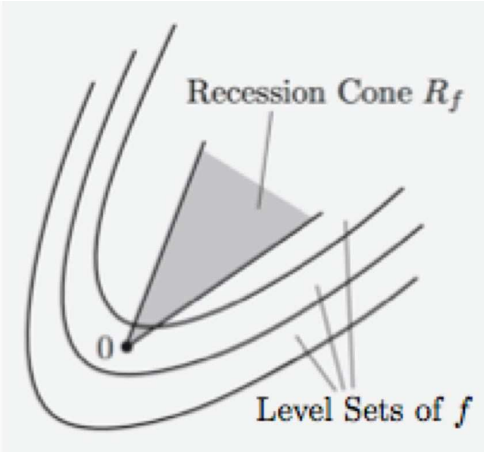
- Recession fn of closed proper convex f: Function $r_f : \Re^n \mapsto (-\infty, \infty]$ whose epigraph is $R_{\text{epi}(f)}$.



- We have

$$R_f = \{d \mid (d, 0) \in R_{\text{epi}(f)}\} = \{d \mid r_f(d) \leq 0\}$$

This is the set of all directions along which f does not increase.

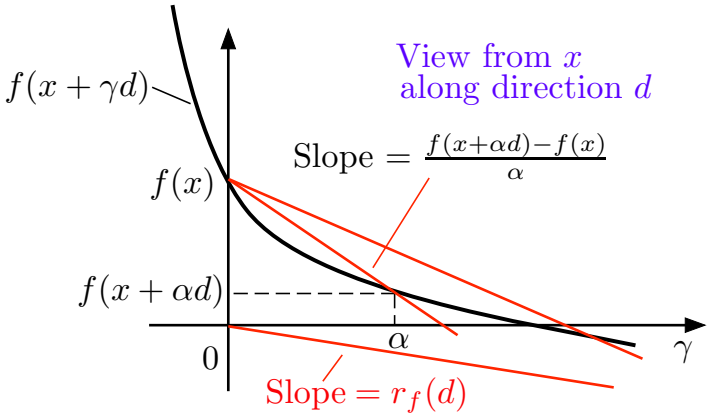


RECESSION FUNCTION & ASYMPTOTIC SLOPES

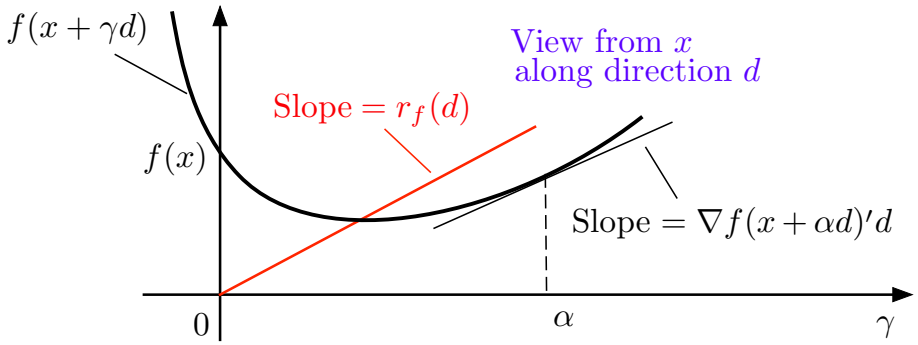
- It can be seen that for all $x \in \mathrm{dom}(f)$, $d \in \Re^n$,

$$r_f(d) = \sup_{\alpha>0} \frac{f(x+\alpha d) - f(x)}{\alpha} = \lim_{\alpha\to\infty} \frac{f(x+\alpha d) - f(x)}{\alpha}$$

$r_f(d)$ is the "asymptotic slope" of f along d



- f differentiable: $r_f(d) = \lim_{\alpha\to\infty} \nabla f(x+\alpha d)'d$

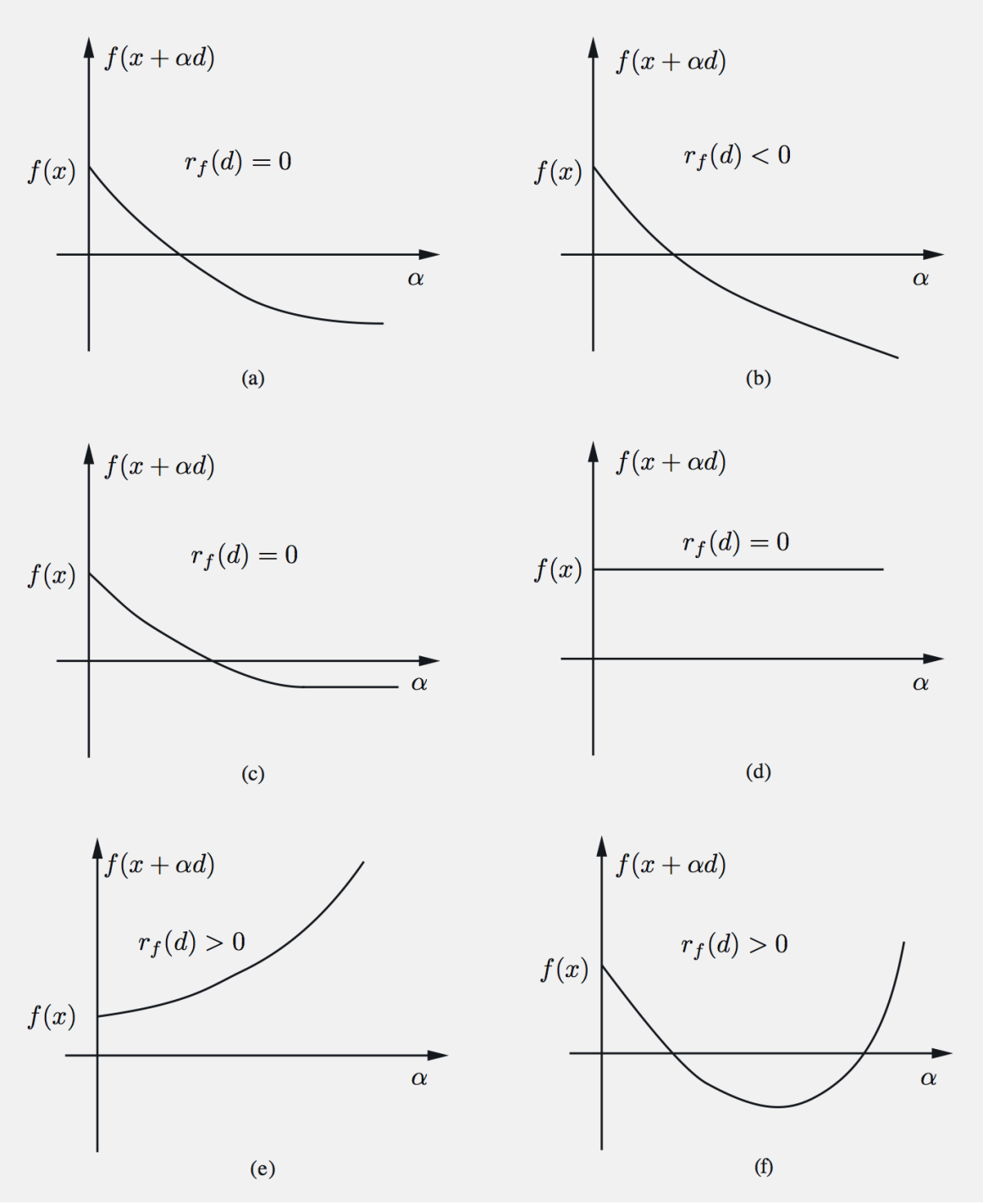


- Calculus of recession functions:

$$r_{f_1+\cdots+f_m}(d) = r_{f_1}(d) + \cdots + r_{f_m}(d),$$

$$r_{\sup_{i\in I} f_i}(d) = \sup_{i\in I} r_{f_i}(d)$$

DESCENT BEHAVIOR OF A CONVEX FN



- y is a direction of recession in (a)-(d).
- This behavior is independent of the starting point x, as long as $x \in \text{dom}(f)$.

EXAMPLE: POS. SEMIDEFINITE FUNCTIONS

- Consider

$$f(x) = x'Qx + a'x + b$$

where Q: positive semidefinite symmetric, $a \in \Re^n$, $b \in \Re$.

- Recession cone:

$$R_f = \{d \mid Qd = 0,\ a'd \leq 0\}$$

- Constancy space (set of directions along which f is constant):

$$L_f = (R_f) \cap (-R_f) = \{d \mid Qd = 0,\ a'd = 0\}$$

- Recession function:

$$r_f(d) = \begin{cases} a'd & \text{if } d \in N(Q), \\ \infty & \text{if } d \notin N(Q). \end{cases}$$

- $R_f = L_f = \{0\}$ if and only if Q is positive definite.

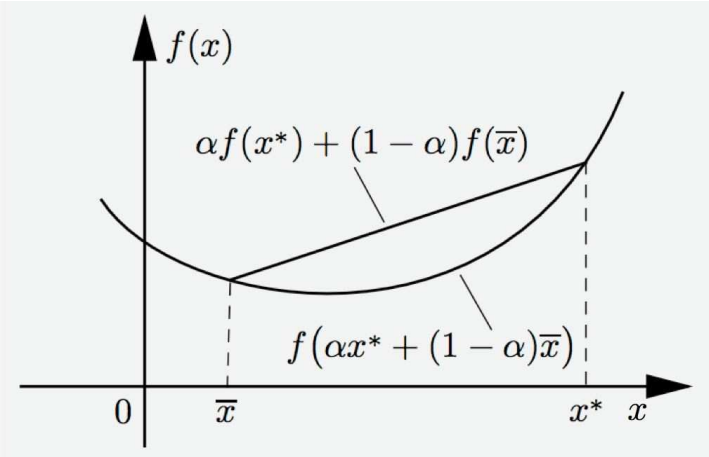
LOCAL AND GLOBAL MINIMA

- Consider minimizing $f : \Re^n \mapsto (-\infty, \infty]$ over a set $X \subset \Re^n$.
- x is feasible if $x \in X \cap \text{dom}(f)$.
- x^* is a (global) minimum of f over X if x^* is feasible and $f(x^*) = \inf_{x \in X} f(x)$.
- x^* is a local minimum of f over X if x^* is a minimum of f over a set $X \cap \{x \mid \|x - x^*\| \leq \epsilon\}$.

Proposition: If X is convex and f is convex, then:

(a) A local minimum of f over X is also a global minimum of f over X.

(b) If f is strictly convex, then there exists at most one global minimum of f over X.



EXISTENCE OF OPTIMAL SOLUTIONS

- The set of minima of a proper $f : \Re^n \mapsto (-\infty, \infty]$ is the intersection of its nonempty level sets.

- The set of minima of f is nonempty and compact if all the level sets of f are compact.

- (An Extension of the) Weierstrass' Theorem: The set of minima of f over X is nonempty and compact if X is closed, f is lower semicontinuous over X, and one of the following conditions holds:

  (1) X is bounded.

  (2) Some set $\{x \in X \mid f(x) \leq \gamma\}$ is nonempty and bounded.

  (3) If $\{x_k\} \subset X$ and $\|x_k\| \to \infty$, then

$$\lim_{k\to\infty} f(x_k) = \infty.$$

Proof: The function $\hat{f}$ given by

$$\hat{f}(x) = \begin{cases} f(x) & \text{if } x \in X, \\ \infty & \text{if } x \notin X, \end{cases}$$

is closed and has compact level sets under any of (1)-(3). **Q.E.D.**

EXISTENCE OF SOLUTIONS - CONVEX CASE

- Weierstrass' Theorem specialized to convex functions: Let X be a closed convex subset of $\Re^n$, and let $f : \Re^n \mapsto (-\infty, \infty]$ be closed convex with $X \cap \text{dom}(f) \neq \emptyset$. The set of minima of f over X is nonempty and compact if and only if X and f have no common nonzero direction of recession.

Proof: Let $f^* = \inf_{x \in X} f(x)$ and note that $f^* < \infty$ since $X \cap \text{dom}(f) \neq \emptyset$. Let $\{\gamma_k\}$ be a scalar sequence with $\gamma_k \downarrow f^*$, and consider the sets

$$V_k = \big\{x \mid f(x) \leq \gamma_k\big\}.$$

Then the set of minima of f over X is

$$X^* = \cap_{k=1}^{\infty}(X \cap V_k).$$

The sets $X \cap V_k$ are nonempty and have $R_X \cap R_f$ as their common recession cone, which is also the recession cone of X^*, when $X^* \neq \emptyset$. It follows that X^* is nonempty and compact if and only if $R_X \cap R_f = \{0\}$. **Q.E.D.**

EXISTENCE OF SOLUTION, SUM OF FNS

- Let $f_i : \Re^n \mapsto (-\infty, \infty]$, $i = 1, \ldots, m$, be closed proper convex such that the function

$$f = f_1 + \cdots + f_m$$

is proper. Assume that a single f_i satisfies $r_{f_i}(d) = \infty$ for all $d \neq 0$. Then the set of minima of f is nonempty and compact.

- **Proof:** We have $r_f(d) = \infty$ for all $d \neq 0$ since $r_f(d) = \sum_{i=1}^m r_{f_i}(d)$. Hence f has no nonzero directions of recession. **Q.E.D.**

- Example of application: If one of the f_i is positive definite quadratic.
  - The set of minima of $f = f_1 + \cdots + f_m$ is nonempty and compact.
  - f has a unique minimum because the positive definite quadratic is strictly convex, which makes f strictly convex.
- The conclusion also holds for $f = \max\{f_1, \ldots, f_m\}$.

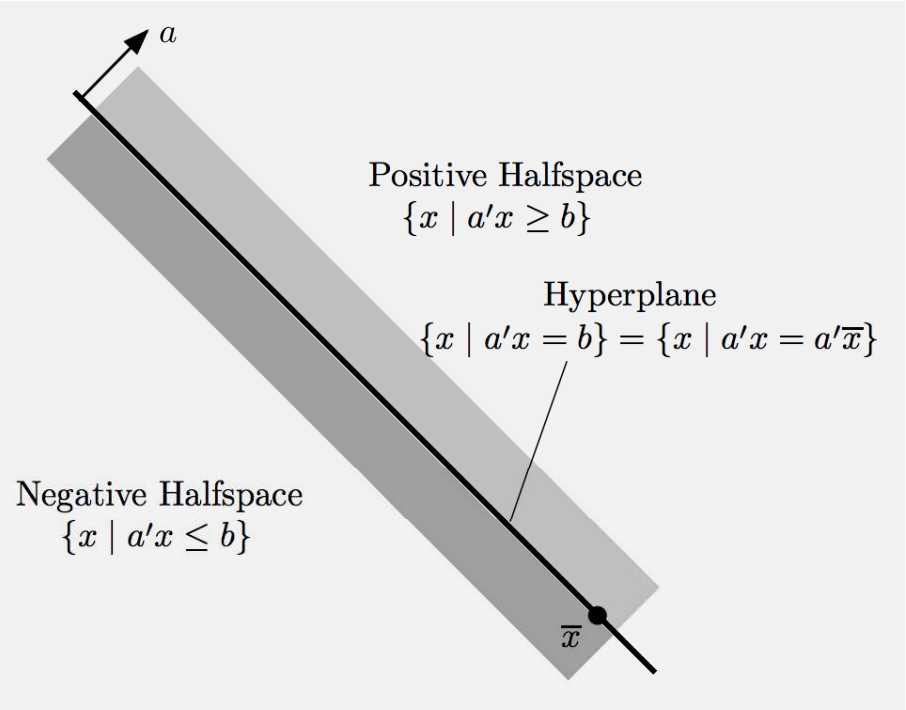
LECTURE 6

LECTURE OUTLINE

- Hyperplanes
- Supporting and Separating Hyperplane Theorems
- Strict Separation
- Proper Separation
- Nonvertical Hyperplanes

Reading: Section 1.5

HYPERPLANES



- A hyperplane is a set of the form $\{x \mid a'x = b\}$, where a is nonzero vector in $\Re^n$ and b is a scalar.

- We say that two sets C_1 and C_2 are separated by a hyperplane $H = \{x \mid a'x = b\}$ if each lies in a different closed halfspace associated with H, i.e.,

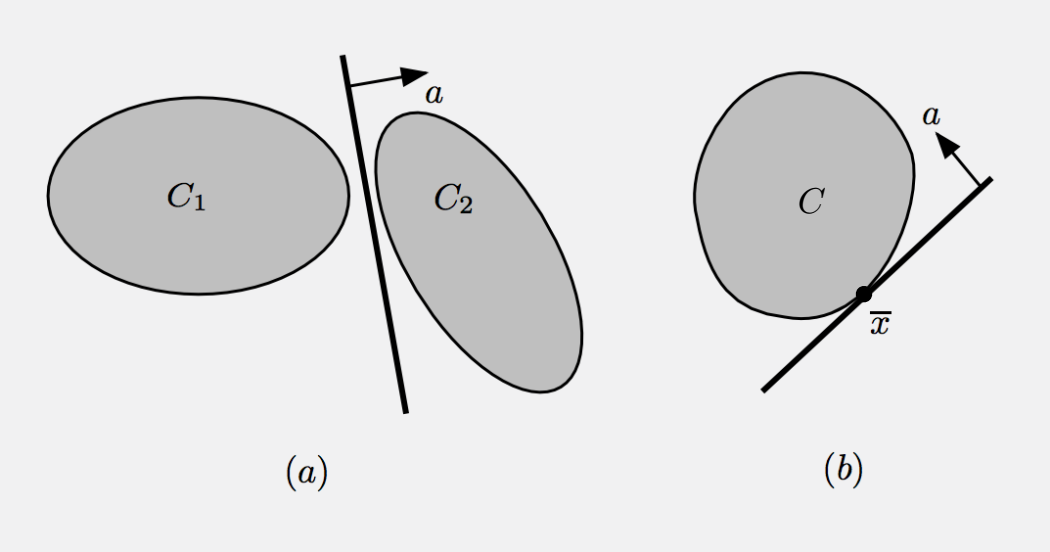
$$\text{either} \quad a'x_1 \leq b \leq a'x_2, \qquad \forall\, x_1 \in C_1,\ \forall\, x_2 \in C_2,$$

$$\text{or} \quad a'x_2 \leq b \leq a'x_1, \qquad \forall\, x_1 \in C_1,\ \forall\, x_2 \in C_2$$

- If $\overline{x}$ belongs to the closure of a set C, a hyperplane that separates C and the singleton set $\{\overline{x}\}$ is said be supporting C at $\overline{x}$.

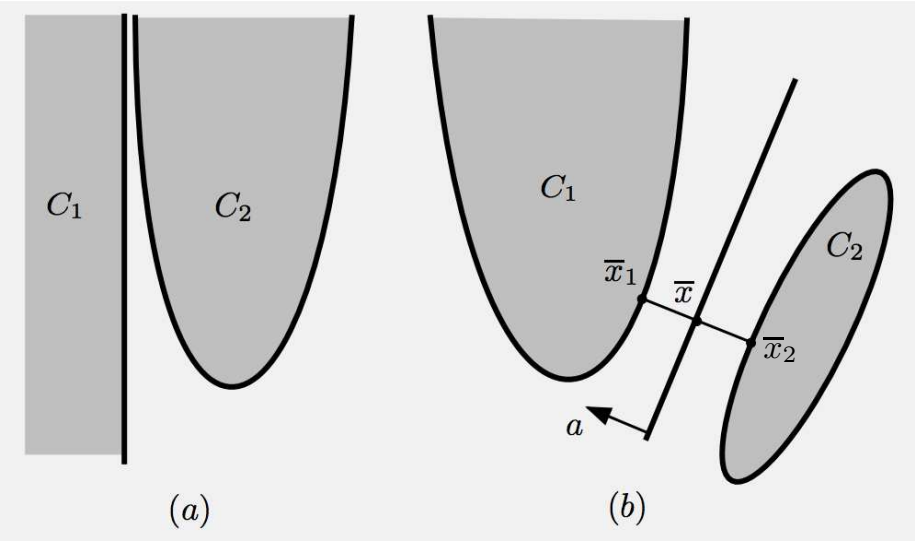
VISUALIZATION

- Separating and supporting hyperplanes:



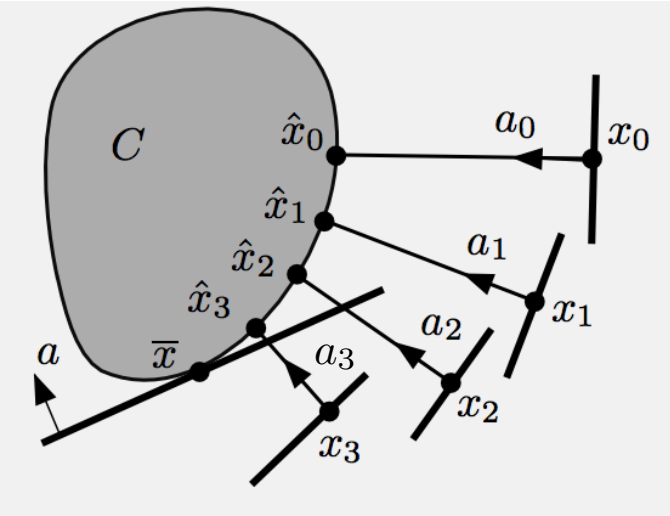
- A separating $\{x \mid a'x = b\}$ that is disjoint from C_1 and C_2 is called strictly separating:

$$a'x_1 < b < a'x_2, \qquad \forall\, x_1 \in C_1,\ \forall\, x_2 \in C_2$$



SUPPORTING HYPERPLANE THEOREM

• Let C be convex and let $\overline{x}$ be a vector that is not an interior point of C. Then, there exists a hyperplane that passes through $\overline{x}$ and contains C in one of its closed halfspaces.



Proof: Take a sequence $\{x_k\}$ that does not belong to $\text{cl}(C)$ and converges to $\overline{x}$. Let $\hat{x}_k$ be the projection of x_k on $\text{cl}(C)$. We have for all $x \in \text{cl}(C)$

$$a_k' x \geq a_k' x_k, \qquad \forall\ x \in \text{cl}(C),\ \forall\ k = 0, 1, \ldots,$$

where $a_k = (\hat{x}_k - x_k)/\|\hat{x}_k - x_k\|$. Let a be a limit point of $\{a_k\}$, and take limit as $k \to \infty$. **Q.E.D.**

SEPARATING HYPERPLANE THEOREM

• Let C_1 and C_2 be two nonempty convex subsets of $\Re^n$. If C_1 and C_2 are disjoint, there exists a hyperplane that separates them, i.e., there exists a vector $a \neq 0$ such that

$$a'x_1 \leq a'x_2, \qquad \forall\ x_1 \in C_1,\ \forall\ x_2 \in C_2.$$

Proof: Consider the convex set

$$C_1 - C_2 = \{x_2 - x_1 \mid x_1 \in C_1,\ x_2 \in C_2\}$$

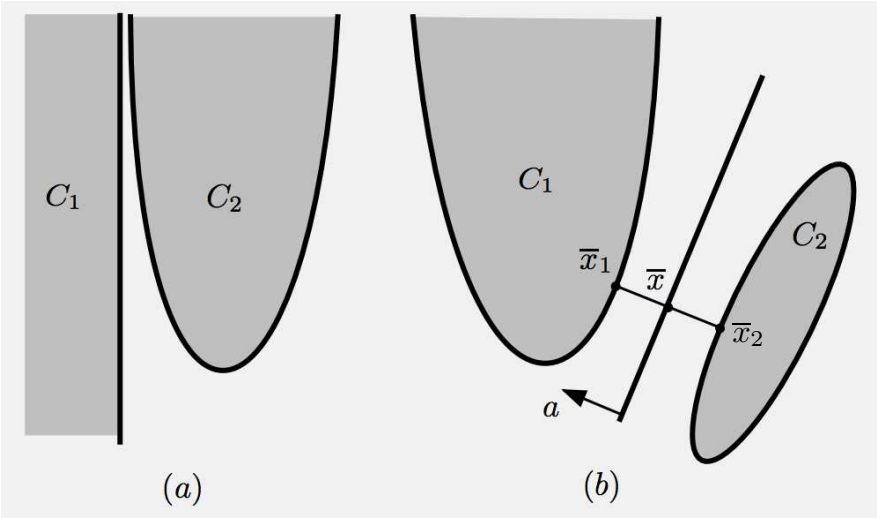
Since C_1 and C_2 are disjoint, the origin does not belong to $C_1 - C_2$, so by the Supporting Hyperplane Theorem, there exists a vector $a \neq 0$ such that

$$0 \leq a'x, \qquad \forall\ x \in C_1 - C_2,$$

which is equivalent to the desired relation. **Q.E.D.**

STRICT SEPARATION THEOREM

• Strict Separation Theorem: Let C_1 and C_2 be two disjoint nonempty convex sets. If C_1 is closed, and C_2 is compact, there exists a hyperplane that strictly separates them.



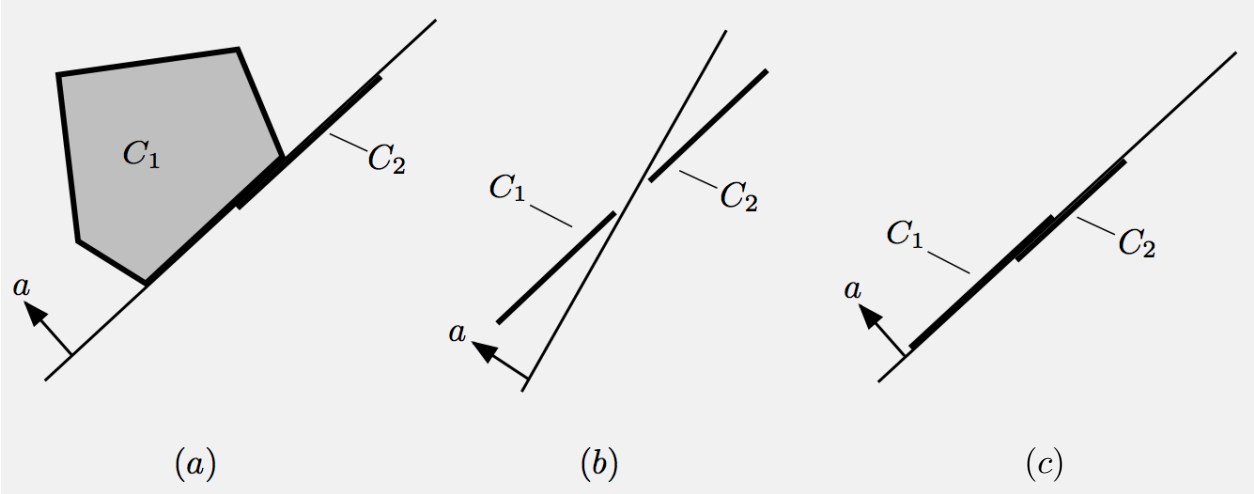
Proof: (Outline) Consider the set $C_1 - C_2$. Since C_1 is closed and C_2 is compact, $C_1 - C_2$ is closed. Since $C_1 \cap C_2 = \emptyset$, $0 \notin C_1 - C_2$. Let $\overline{x}_1 - \overline{x}_2$ be the projection of 0 onto $C_1 - C_2$. The strictly separating hyperplane is constructed as in (b).

• Note: Any conditions that guarantee closedness of $C_1 - C_2$ guarantee existence of a strictly separating hyperplane. However, there may exist a strictly separating hyperplane without $C_1 - C_2$ being closed.

ADDITIONAL THEOREMS

• Fundamental Characterization: The closure of the convex hull of a set $C \subset \Re^n$ is the intersection of the closed halfspaces that contain C. (Proof uses the strict separation theorem.)

• We say that a hyperplane properly separates C_1 and C_2 if it separates C_1 and C_2 and does not fully contain both C_1 and C_2.



• Proper Separation Theorem: Let C_1 and C_2 be two nonempty convex subsets of $\Re^n$. There exists a hyperplane that properly separates C_1 and C_2 if and only if

$$\text{ri}(C_1) \cap \text{ri}(C_2) = \emptyset$$

PROPER POLYHEDRAL SEPARATION

- Recall that two convex sets C and P such that

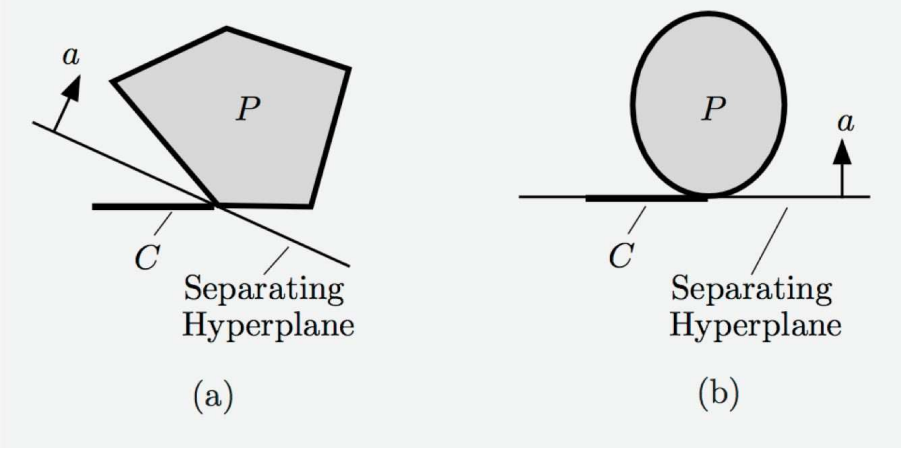
$$\text{ri}(C) \cap \text{ri}(P) = \emptyset$$

can be properly separated, i.e., by a hyperplane that does not contain both C and P.

- If P is polyhedral and the slightly stronger condition

$$\text{ri}(C) \cap P = \emptyset$$

holds, then the properly separating hyperplane can be chosen so that it does not contain the non-polyhedral set C while it may contain P.



On the left, the separating hyperplane can be chosen so that it does not contain C. On the right where P is not polyhedral, this is not possible.

NONVERTICAL HYPERPLANE THEOREM

- Let C be a nonempty convex subset of $\Re^{n+1}$ that contains no vertical lines. Then:

  (a) C is contained in a closed halfspace of a nonvertical hyperplane, i.e., there exist $\mu \in \Re^n$, $\beta \in \Re$ with $\beta \neq 0$, and $\gamma \in \Re$ such that $\mu' u + \beta w \geq \gamma$ for all $(u, w) \in C$.

  (b) If $(\overline{u}, \overline{w}) \notin \text{cl}(C)$, there exists a nonvertical hyperplane strictly separating $(\overline{u}, \overline{w})$ and C.

Proof: Note that $\text{cl}(C)$ contains no vert. line [since C contains no vert. line, $\text{ri}(C)$ contains no vert. line, and $\text{ri}(C)$ and $\text{cl}(C)$ have the same recession cone]. So we just consider the case: C closed.

(a) C is the intersection of the closed halfspaces containing C. If all these corresponded to vertical hyperplanes, C would contain a vertical line.

(b) There is a hyperplane strictly separating $(\overline{u}, \overline{w})$ and C. If it is nonvertical, we are done, so assume it is vertical. "Add" to this vertical hyperplane a small ϵ-multiple of a nonvertical hyperplane containing C in one of its halfspaces as per (a).

LECTURE 7

LECTURE OUTLINE

- Convex conjugate functions
- Conjugacy theorem
- Support functions and polar cones
- Examples

Reading: Section 1.6

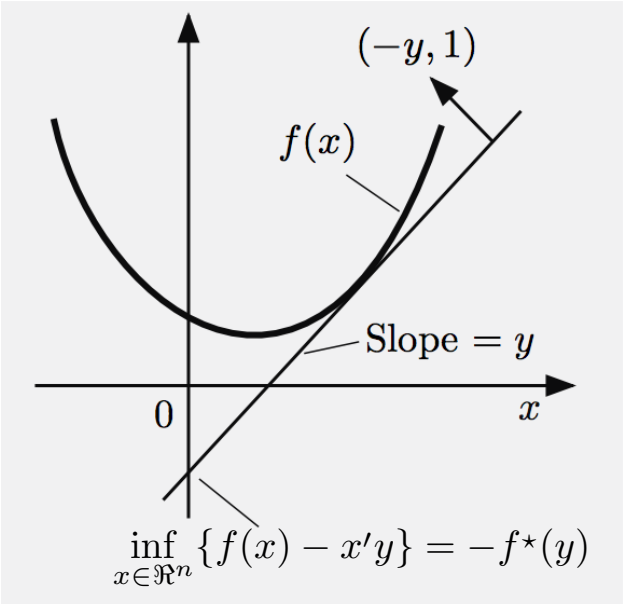
CONJUGATE CONVEX FUNCTIONS

- Consider a function f and its epigraph

Nonvertical hyperplanes supporting epi(f)

$\mapsto$ Crossing points of vertical axis

$$f^\star(y) = \sup_{x \in \Re^n} \{x'y - f(x)\}, \qquad y \in \Re^n.$$

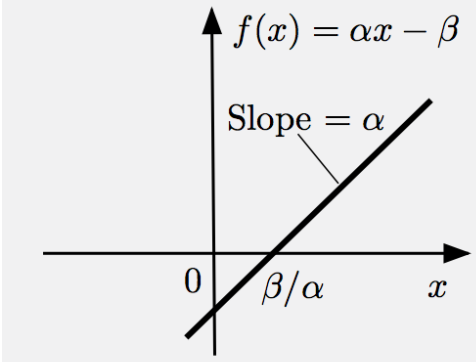


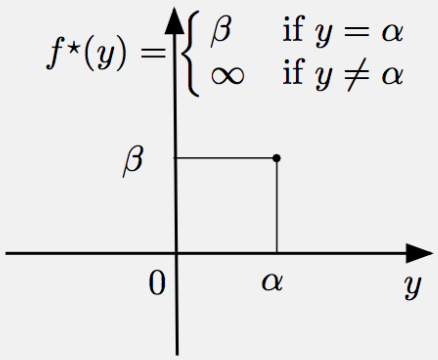
- For any $f : \Re^n \mapsto [-\infty, \infty]$, its conjugate convex function is defined by

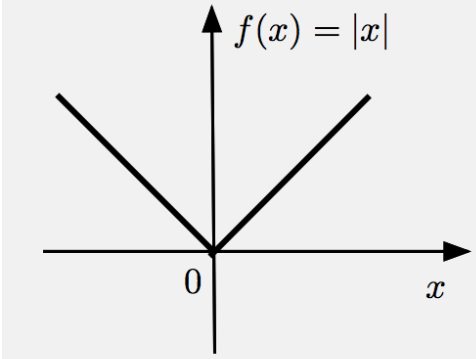
$$f^\star(y) = \sup_{x \in \Re^n} \{x'y - f(x)\}, \qquad y \in \Re^n$$

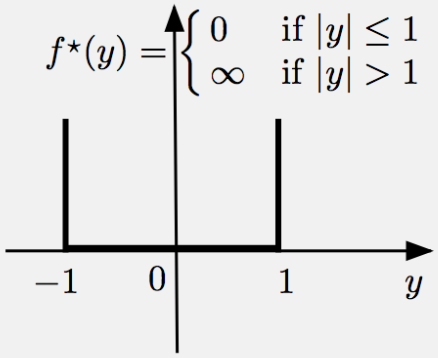
EXAMPLES

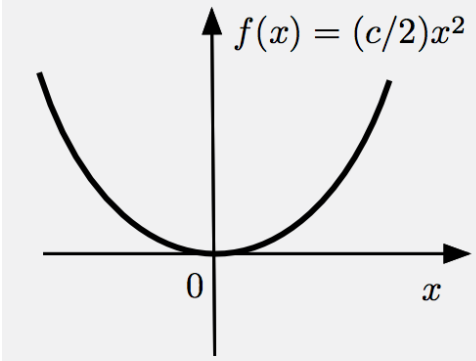
$$f^\star(y) = \sup_{x \in \Re^n} \{x'y - f(x)\}, \qquad y \in \Re^n$$

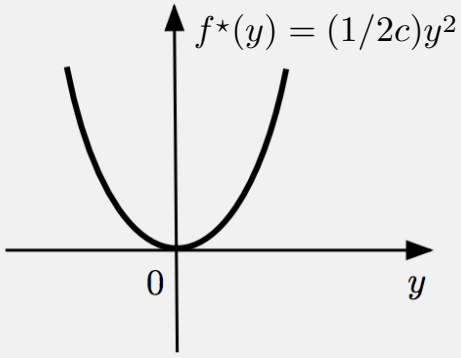












CONJUGATE OF CONJUGATE

- From the definition

$$f^\star(y) = \sup_{x \in \Re^n} \big\{ x'y - f(x) \big\}, \qquad y \in \Re^n,$$

note that $f^\star$ is convex and closed.

- Reason: $\text{epi}(f^\star)$ is the intersection of the epigraphs of the linear functions of y

$$x'y - f(x)$$

as x ranges over $\Re^n$.

- Consider the conjugate of the conjugate:

$$f^{\star\star}(x) = \sup_{y \in \Re^n} \big\{ y'x - f^\star(y) \big\}, \qquad x \in \Re^n.$$

- $f^{\star\star}$ is convex and closed.

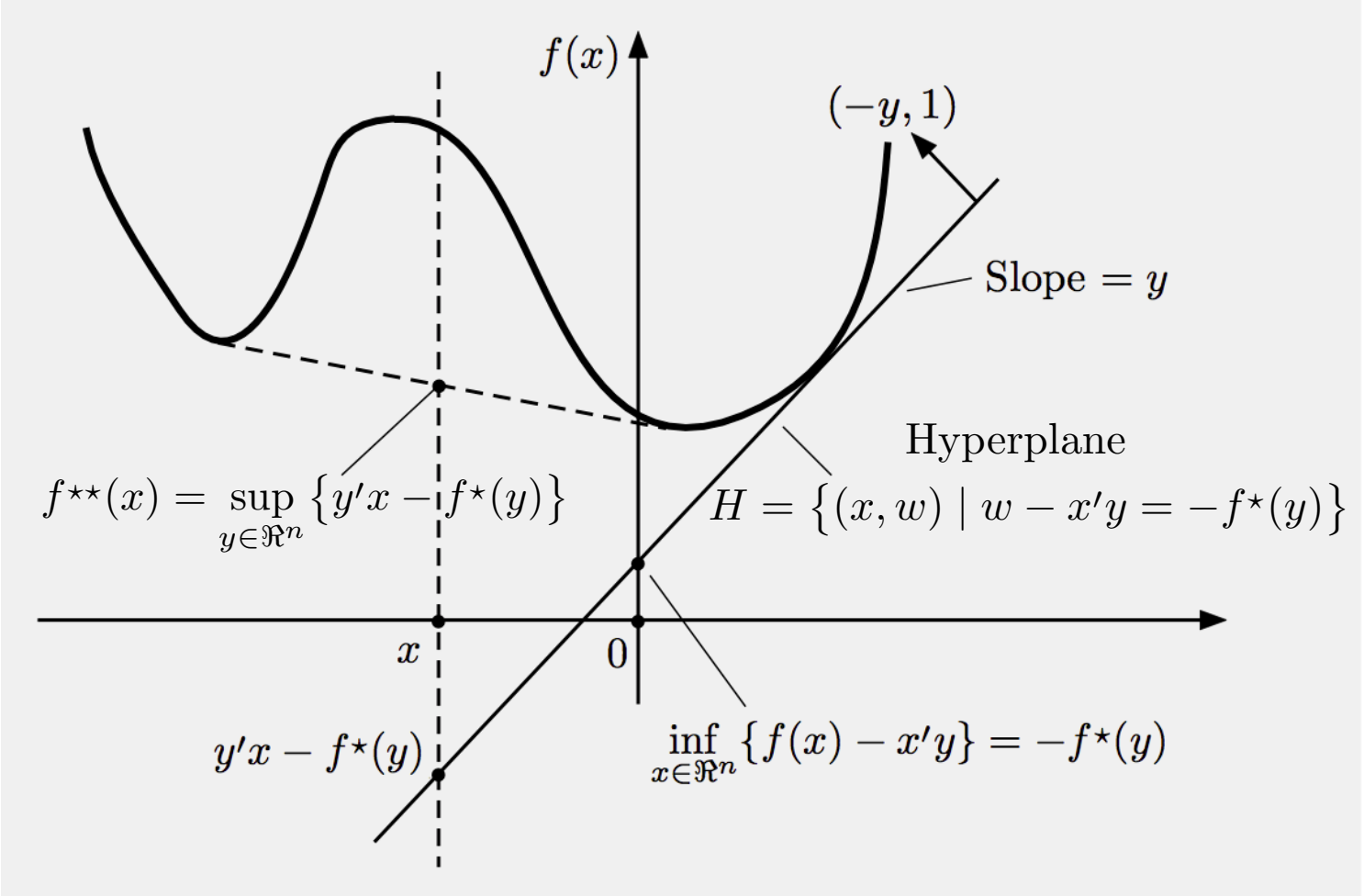
- Important fact/Conjugacy theorem: If f is closed proper convex, then $f^{\star\star} = f$.

CONJUGACY THEOREM - VISUALIZATION

$$f^\star(y) = \sup_{x \in \Re^n} \{x'y - f(x)\}, \qquad y \in \Re^n$$

$$f^{\star\star}(x) = \sup_{y \in \Re^n} \{y'x - f^\star(y)\}, \qquad x \in \Re^n$$

- If f is closed convex proper, then $f^{\star\star} = f$.



CONJUGACY THEOREM

• Let $f : \Re^n \mapsto (-\infty, \infty]$ be a function, let $\check{\text{cl}}\, f$ be its convex closure, let $f^\star$ be its convex conjugate, and consider the conjugate of $f^\star$,

$$f^{\star\star}(x) = \sup_{y \in \Re^n} \left\{ y'x - f^\star(y) \right\}, \qquad x \in \Re^n$$

(a) We have

$$f(x) \geq f^{\star\star}(x), \qquad \forall\ x \in \Re^n$$

(b) If f is closed proper and convex, then

$$f(x) = f^{\star\star}(x), \qquad \forall\ x \in \Re^n$$

(c) If f is convex, then properness of any one of f, $f^\star$, and $f^{\star\star}$ implies properness of the other two.

(d) If $\check{\text{cl}}\, f(x) > -\infty$ for all $x \in \Re^n$, then

$$\check{\text{cl}}\, f(x) = f^{\star\star}(x), \qquad \forall\ x \in \Re^n$$

PROOF OF CONJUGACY THEOREM (A), (B)

- (a) For all x, y, we have $f^\star(y) \geq y'x - f(x)$, implying that $f(x) \geq \sup_y\{y'x - f^\star(y)\} = f^{\star\star}(x)$.

- (b) By contradiction. Assume there is $(x, \gamma) \in \text{epi}(f^{\star\star})$ with $(x, \gamma) \notin \text{epi}(f)$. There exists a non-vertical hyperplane with normal $(y, -1)$ that strictly separates (x, γ) and $\text{epi}(f)$. (The vertical component of the normal vector is normalized to -1.) Thus we have for some $c \in \Re$

$$y'z - w < c < y'x - \gamma, \qquad \forall\ (z, w) \in \text{epi}(f)$$

Since $\gamma \geq f^{\star\star}(x)$ and $\big(z, f(z)\big) \in \text{epi}(f)$,

$$y'z - f(z) < c < y'x - f^{\star\star}(x), \qquad \forall\ z \in \text{dom}(f).$$

Hence

$$f^\star(y) = \sup_{z \in \Re^n} \big\{y'z - f(z)\big\} \leq c < y'x - f^{\star\star}(x),$$

contradicting the fact $f^{\star\star}(x) = \sup_{y \in \Re^n} \big\{y'x - f^\star(y)\big\}$. Thus, $\text{epi}(f^{\star\star}) \subset \text{epi}(f)$, which implies that $f(x) \leq f^{\star\star}(x)$ for all $x \in \Re^n$. This, together with part (a), shows that $f^{\star\star}(x) = f(x)$ for all x.

A COUNTEREXAMPLE

• A counterexample (with closed convex but improper f) showing the need to assume properness in order for $f = f^{\star\star}$:

$$f(x) = \begin{cases} \infty & \text{if } x > 0, \\ -\infty & \text{if } x \leq 0. \end{cases}$$

We have

$$f^\star(y) = \infty, \qquad \forall\ y \in \Re^n,$$

$$f^{\star\star}(x) = -\infty, \qquad \forall\ x \in \Re^n.$$

But

$$\check{\text{cl}}\, f = f,$$

so $\check{\text{cl}}\, f \neq f^{\star\star}$.

A FEW EXAMPLES

- l_p and l_q norm conjugacy, where $\frac{1}{p} + \frac{1}{q} = 1$

$$f(x) = \frac{1}{p}\sum_{i=1}^{n} |x_i|^p, \qquad f^\star(y) = \frac{1}{q}\sum_{i=1}^{n} |y_i|^q$$

- Conjugate of a strictly convex quadratic

$$f(x) = \frac{1}{2}x'Qx + a'x + b,$$

$$f^\star(y) = \frac{1}{2}(y-a)'Q^{-1}(y-a) - b.$$

- Conjugate of a function obtained by invertible linear transformation/translation of a function p

$$f(x) = p\big(A(x-c)\big) + a'x + b,$$

$$f^\star(y) = q\big((A')^{-1}(y-a)\big) + c'y + d,$$

where q is the conjugate of p and $d = -(c'a + b)$.

LECTURE 8

LECTURE OUTLINE

- Review of conjugate convex functions
- Polar cones and Farkas' Lemma
- Min common/max crossing duality
- Weak duality
- Special cases

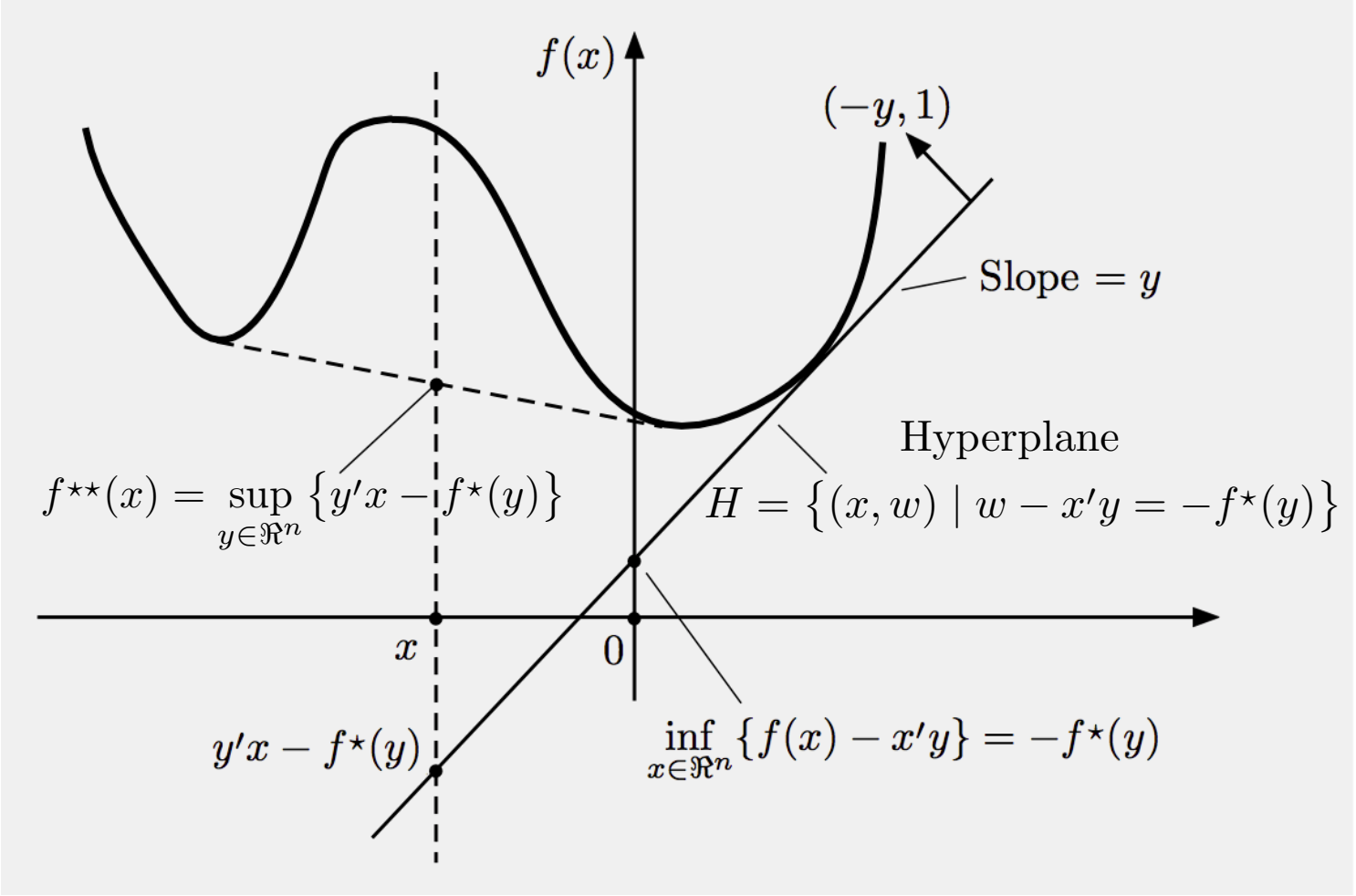
Reading: Sections 1.6, 4.1, 4.2

CONJUGACY THEOREM

$$f^\star(y) = \sup_{x \in \Re^n} \big\{x'y - f(x)\big\}, \qquad y \in \Re^n$$

$$f^{\star\star}(x) = \sup_{y \in \Re^n} \big\{y'x - f^\star(y)\big\}, \qquad x \in \Re^n$$

- If f is closed convex proper, then $f^{\star\star} = f$.
- More generally, $\text{epi}(f^{\star\star}) = \text{cl}\big(\text{conv}(\text{epi}(f))\big)$.



SUPPORT FUNCTIONS

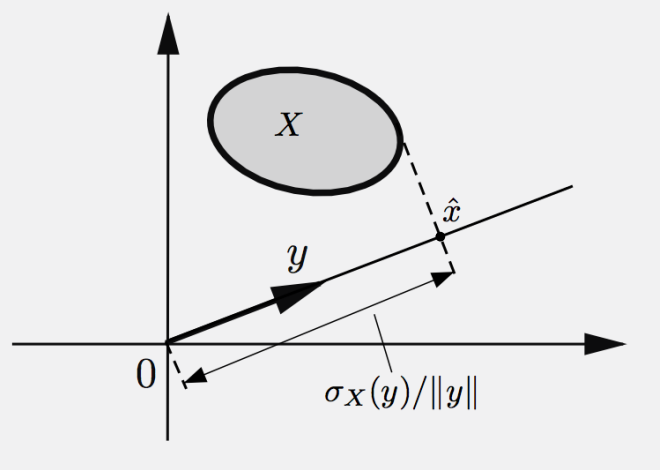
- Conjugate of indicator function δ_X of set X

$$\sigma_X(y) = \sup_{x \in X} y'x$$

is called the support function of X.

- To determine $\sigma_X(y)$ for a given vector y, we project the set X on the line determined by y, we find $\hat{x}$, the extreme point of projection in the direction y, and we scale by setting

$$\sigma_X(y) = \|\hat{x}\| \cdot \|y\|$$



- epi(σ_X) is a closed convex cone.

- X, conv(X), cl(X), and cl$\big($conv$(X)\big)$ have the same support function (by the conjugacy theorem).

SUPPORT FN OF A CONE - POLAR CONE

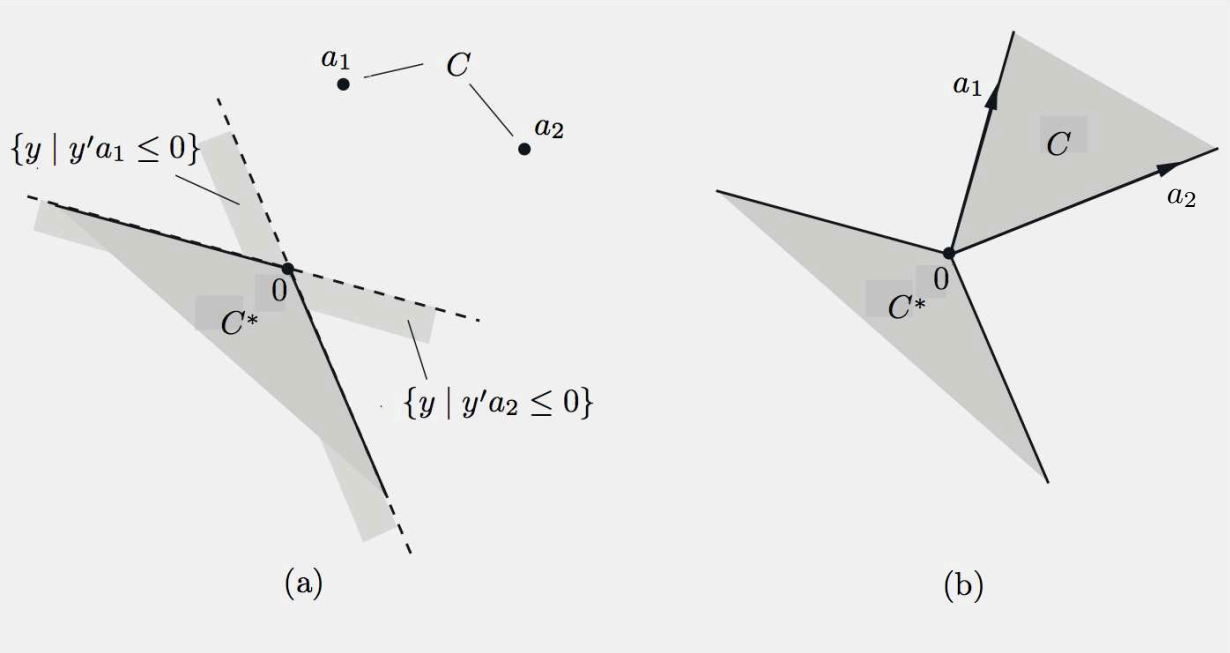
- If C is a cone,

$$\sigma_C(y) = \sup_{x \in C} y'x = \begin{cases} 0 & \text{if } y'x \leq 0, \ \forall \ x \in C, \\ \infty & \text{otherwise} \end{cases}$$

i.e., σ_C is the indicator function δ_{C^*} of the polar cone of C, given by

$$C^* = \{y \mid y'x \leq 0, \ \forall \ x \in C\}$$

- By the Conjugacy Theorem the polar cone of C^* is $\text{cl}\big(\text{conv}(C)\big)$. This is the Polar Cone Theorem.



POLYHEDRAL CONES - FARKAS' LEMMA

- Polyhedral Cone Duality: Let $a_1, \ldots, a_r$ be vectors in $\Re^n$. Then $C = \text{cone}\big(\{a_1, \ldots, a_r\}\big)$ is a closed convex cone, so we have $(C^*)^* = C$, where

$$C^* = \{x \mid A'x \geq 0\}, \qquad C = \{A\mu \mid \mu \geq 0\} \quad (*)$$

and A is the $n \times r$ matrix $A = [a_1 \cdots a_r]$.

Proof: C is obtained by applying A to the nonnegative orthant, and Prop. 1.4.13 of the text shows as a special case that linearly transformed polyhedral sets are closed, implying that C is closed. For other proofs that C is closed, see the internet-posted Ch. 1 and Ch. 2 exercises.

- Farkas' Lemma deals with existence of solutions of systems of linear equations and inequalities.

- Farkas' Lemma (pure inequality case): Let A be an $r \times n$ matrix and $c \in \Re^r$. We have

$$c'x \leq 0, \qquad \forall\ x \text{ such that } A'x \leq 0$$

if and only if there exists $\mu \geq 0$ such that $A\mu = c$.

Proof: Let C and C^* be as in (*). The first assertion can be written as $c \in (C^*)^*$, while the second assertion can be written as $c \in C$. Use the Polar Cone Theorem equation $(C^*)^* = C$. **Q.E.D.**

LAGRANGE MULTIPLIERS

- Consider the problem

$$\min_{a_j' x \leq b_j,\ j=1,\ldots,r} f(x)$$

where $f : \Re^n \mapsto \Re$ is convex and differentiable. A feasible vector x^* is an optimal solution if and only if there exist scalars $\mu_1, \ldots, \mu_r \geq 0$ such that

$$\nabla f(x^*) + \sum_{j=1}^{r} \mu_j a_j = 0, \quad \mu_j (a_j' x^* - b_j) = 0, \ \ \forall\, j \ \ (*)$$

Proof: If x^* is optimal, then

$$\nabla f(x^*)'(x - x^*) \geq 0, \quad \text{for all feasible } x$$

from which

$$\nabla f(x^*)' y \geq 0 \quad \text{for all } y \text{ with } a_j' y \leq 0,\ \forall\, j \in J(x^*),$$

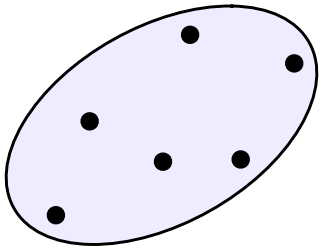
where $J(x^*) = \{j \mid a_j' x^* = b_j\}$. Applying Farkas' Lemma, we have that $-\nabla f(x^*) = \sum_{j \in J(x^*)} \mu_j a_j$ for some $\mu_j \geq 0$, $j \in J(x^*)$. Letting $\mu_j = 0$ for $j \notin J(x^*)$, we obtain (*).

Conversely, if (*) holds, x^* minimizes $f(x) + \sum_{j=1}^{r} \mu_j (a_j' x - b_j)$, so for all feasible x,

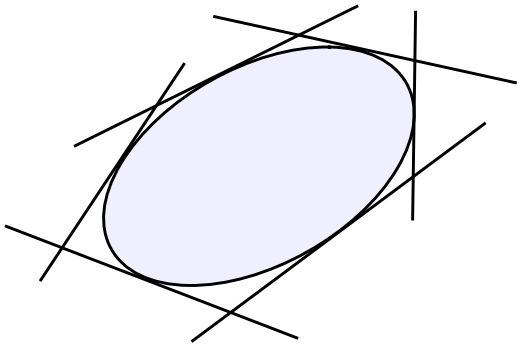
$$f(x^*) \leq f(x) + \sum_{j=1}^{r} \mu_j (a_j' x - b_j) \leq f(x)$$

EXTENDING DUALITY CONCEPTS

- From dual descriptions of closed convex sets

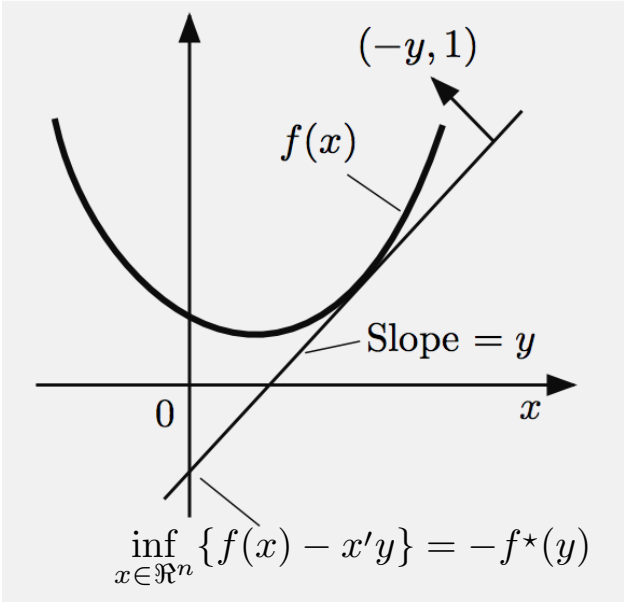


A union of points



An intersection of halfspaces

- To dual descriptions of closed convex functions (applying set duality to epigraphs)



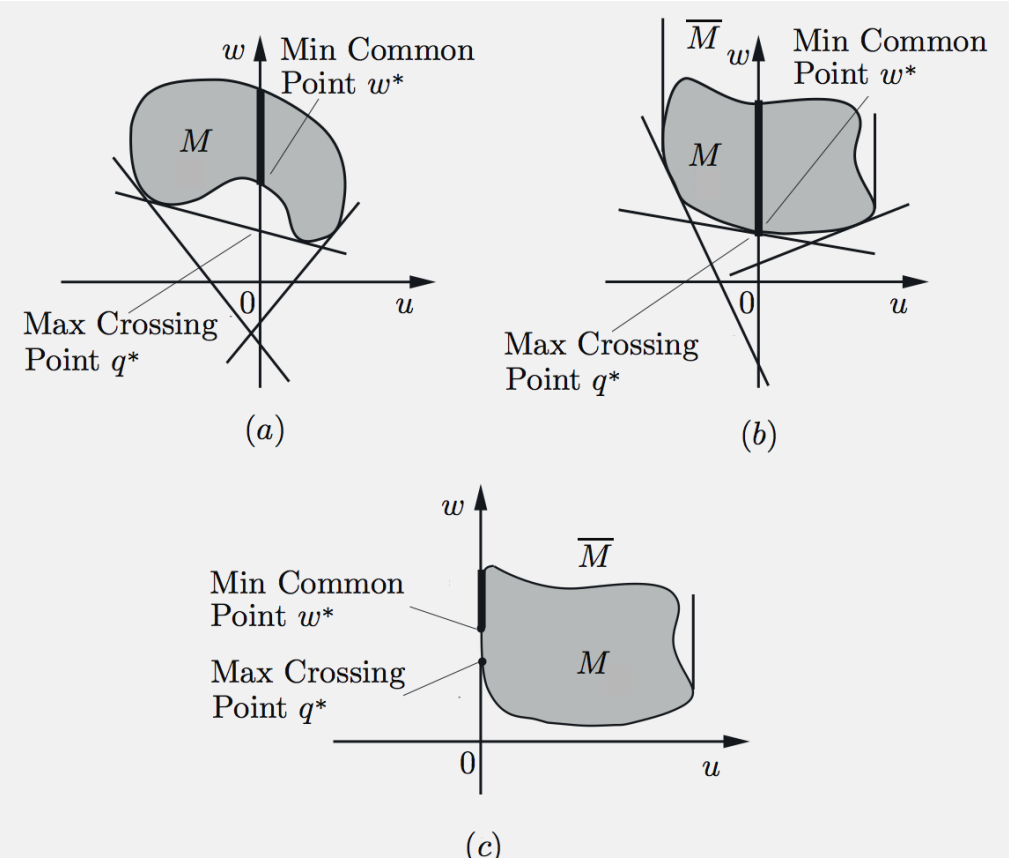
- We now go to dual descriptions of problems, by applying conjugacy constructions to a simple generic geometric optimization problem

MIN COMMON / MAX CROSSING PROBLEMS

- We introduce a pair of fundamental problems:
- Let M be a nonempty subset of $\Re^{n+1}$

(a) Min Common Point Problem: Consider all vectors that are common to M and the $(n+1)$st axis. Find one whose $(n+1)$st component is minimum.

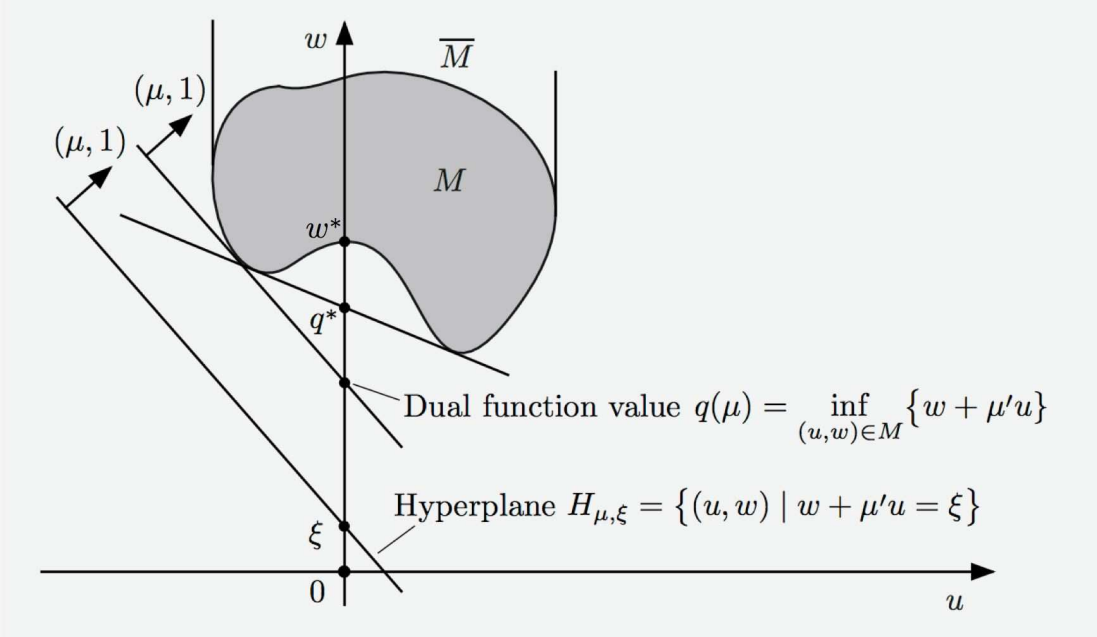
(b) Max Crossing Point Problem: Consider nonvertical hyperplanes that contain M in their "upper" closed halfspace. Find one whose crossing point of the $(n+1)$st axis is maximum.



MATHEMATICAL FORMULATIONS

- Optimal value of min common problem:

$$w^* = \inf_{(0,w)\in M} w$$



- Math formulation of max crossing problem: Focus on hyperplanes with normals $(\mu, 1)$ whose crossing point ξ satisfies

$$\xi \leq w + \mu' u, \qquad \forall\ (u, w) \in M$$

Max crossing problem is to maximize ξ subject to $\xi \leq \inf_{(u,w)\in M} \{w + \mu' u\}$, $\mu \in \Re^n$, or

$$\text{maximize} \quad q(\mu) \stackrel{\triangle}{=} \inf_{(u,w)\in M} \{w + \mu' u\}$$

$$\text{subject to} \quad \mu \in \Re^n$$

GENERIC PROPERTIES – WEAK DUALITY

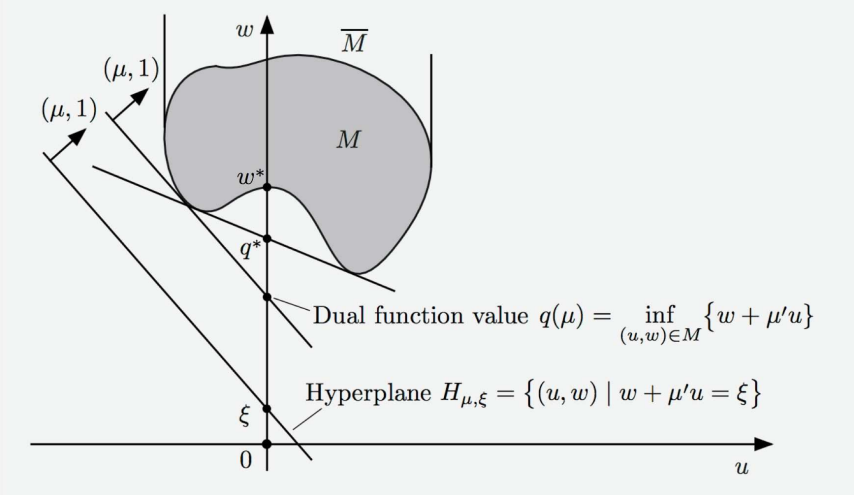
- Min common problem

$$\inf_{(0,w)\in M} w$$

- Max crossing problem

$$\text{maximize } q(\mu) \stackrel{\triangle}{=} \inf_{(u,w)\in M} \{w + \mu' u\}$$

$$\text{subject to } \mu \in \Re^n$$



- Note that q is concave and upper-semicontinuous (inf of linear functions).
- Weak Duality: For all $\mu \in \Re^n$

$$q(\mu) = \inf_{(u,w)\in M} \{w + \mu' u\} \leq \inf_{(0,w)\in M} w = w^*,$$

so maximizing over $\mu \in \Re^n$, we obtain $q^* \leq w^*$.

- We say that strong duality holds if $q^* = w^*$.

CONNECTION TO CONJUGACY

- An important special case:

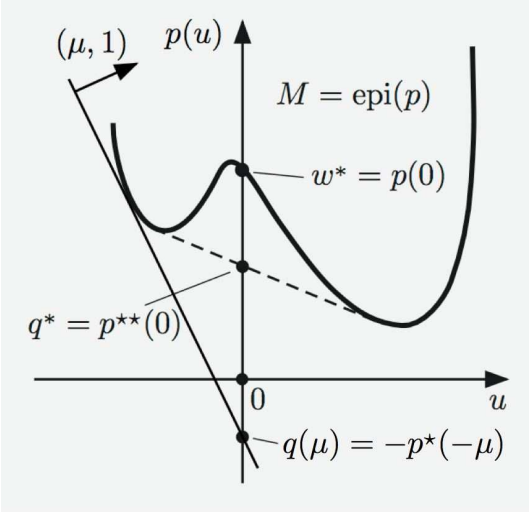
$$M = \text{epi}(p)$$

where $p : \Re^n \mapsto [-\infty, \infty]$. Then $w^* = p(0)$, and

$$q(\mu) = \inf_{(u,w) \in \text{epi}(p)} \{w + \mu' u\} = \inf_{\{(u,w) \mid p(u) \leq w\}} \{w + \mu' u\},$$

and finally

$$q(\mu) = \inf_{u \in \Re^m} \big\{p(u) + \mu' u\big\}$$



- Thus, $q(\mu) = -p^\star(-\mu)$ and

$$q^* = \sup_{\mu \in \Re^n} q(\mu) = \sup_{\mu \in \Re^n} \big\{0 \cdot (-\mu) - p^\star(-\mu)\big\} = p^{\star\star}(0)$$

so $q^* = w^*$ if p is closed, proper, convex.

GENERAL OPTIMIZATION DUALITY

- Consider minimizing a function $f : \Re^n \mapsto [-\infty, \infty]$.
- Let $F : \Re^{n+r} \mapsto [-\infty, \infty]$ be a function with

$$f(x) = F(x, 0), \qquad \forall\ x \in \Re^n$$

- Consider the perturbation function

$$p(u) = \inf_{x \in \Re^n} F(x, u)$$

and the MC/MC framework with $M = \text{epi}(p)$

- The min common value w^* is

$$w^* = p(0) = \inf_{x \in \Re^n} F(x, 0) = \inf_{x \in \Re^n} f(x)$$

- The dual function is

$$q(\mu) = \inf_{u \in \Re^r} \big\{p(u) + \mu' u\big\} = \inf_{(x,u) \in \Re^{n+r}} \big\{F(x, u) + \mu' u\big\}$$

so $q(\mu) = -F^\star(0, -\mu)$, where $F^\star$ is the conjugate of F, viewed as a function of (x, u).

- We have

$$q^* = \sup_{\mu \in \Re^r} q(\mu) = -\inf_{\mu \in \Re^r} F^\star(0, -\mu) = -\inf_{\mu \in \Re^r} F^\star(0, \mu),$$

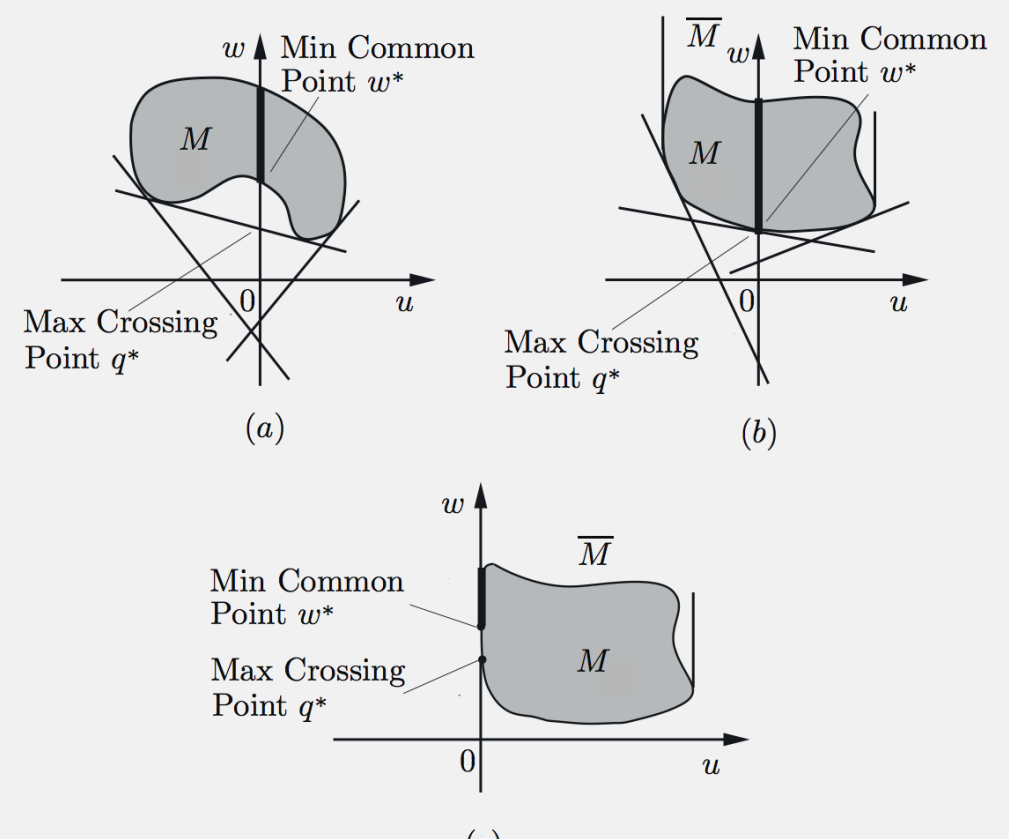
$$w^* = \inf_{x \in \Re^n} F(x, 0)$$

LECTURE 9

LECTURE OUTLINE

- Min Common/Max Crossing duality for constrained optimization
- Min Common/Max Crossing duality for minimax and zero-sum games
- Min Common/Max Crossing duality theorems
- Strong duality conditions and existence of dual optimal solutions

Reading: Sections 4.1, 4.2



REVIEW OF THE MC/MC FRAMEWORK

- Given set $M \subset \Re^{n+1}$,

$$w^* = \inf_{(0,w)\in M} w, \quad q^* = \sup_{\mu\in\Re^n} q(\mu) \stackrel{\triangle}{=} \inf_{(u,w)\in M} \{w+\mu' u\}$$

- Weak Duality: $q^* \leq w^*$ (always holds)
- Strong Duality: $q^* = w^*$ (requires that M have some convexity structure, among other conditions)
- Important special case: $M = \text{epi}(p)$. Then $w^* = p(0)$, $q^* = p^{\star\star}(0)$, so we have $w^* = q^*$ if p is closed, proper, convex.
- Some applications:
  - Constrained optimization: $\min_{x\in X,\, g(x)\leq 0} f(x)$, with $p(u) = \inf_{x\in X,\, g(x)\leq u} f(x)$
  - Other optimization problems: Fenchel and conic optimization
  - Minimax problems, 0-sum games
  - Subgradient theory
  - Useful theorems related to optimization: Farkas' lemma, theorems of the alternative

CONSTRAINED OPTIMIZATION

- Minimize $f : \Re^n \mapsto \Re$ over the set

$$C = \big\{x \in X \mid g(x) \leq 0\big\},$$

where $X \subset \Re^n$ and $g : \Re^n \mapsto \Re^r$.

- Introduce a "perturbed constraint set"

$$C_u = \big\{x \in X \mid g(x) \leq u\big\}, \qquad u \in \Re^r,$$

and the function

$$F(x,u) = \begin{cases} f(x) & \text{if } x \in C_u, \\ \infty & \text{otherwise,} \end{cases}$$

which satisfies $F(x,0) = f(x)$ for all $x \in C$.

- Consider the perturbation function

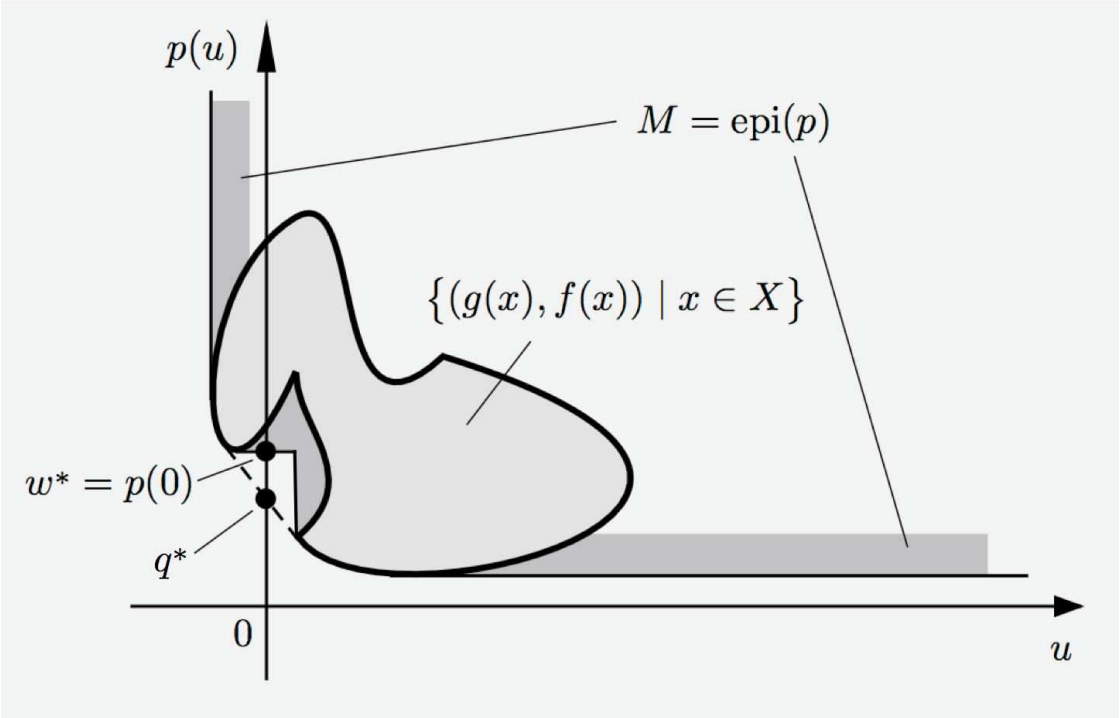
$$p(u) = \inf_{x \in \Re^n} F(x,u) = \inf_{x \in X,\, g(x) \leq u} f(x),$$

and the MC/MC framework with $M = \text{epi}(p)$.

CONSTR. OPT. - PRIMAL AND DUAL FNS

- Perturbation function (or primal function)

$$p(u) = \inf_{x \in X,\, g(x) \le u} f(x),$$



- Let $L(x, \mu) = f(x) + \mu' g(x)$ be the Lagrangian function. Then

$$\begin{aligned} q(\mu) &= \inf_{u \in \Re^r} \left\{ p(u) + \mu' u \right\} = \inf_{u \in \Re^r} \left\{ \inf_{x \in X,\, g(x) \le u} f(x) + \mu' u \right\} \\ &= \inf_{u \in \Re^r,\, x \in X,\, g(x) \le u} \left\{ f(x) + \mu' u \right\} = \inf_{x \in X} \left\{ f(x) + \mu' g(x) \right\} \\ &= \begin{cases} \inf_{x \in X} L(x, \mu) & \text{if } \mu \ge 0, \\ -\infty & \text{otherwise} \end{cases} \end{aligned}$$

LINEAR PROGRAMMING DUALITY

- Consider the linear program

$$\begin{aligned} &\text{minimize} \quad c'x \\ &\text{subject to} \quad a_j'x \geq b_j, \quad j = 1, \ldots, r, \end{aligned}$$

where $c \in \Re^n$, $a_j \in \Re^n$, and $b_j \in \Re$, $j = 1, \ldots, r$.

- For $\mu \geq 0$, the dual function has the form

$$\begin{aligned} q(\mu) &= \inf_{x \in \Re^n} L(x, \mu) \\ &= \inf_{x \in \Re^n} \left\{ c'x + \sum_{j=1}^{r} \mu_j (b_j - a_j'x) \right\} \\ &= \begin{cases} b'\mu & \text{if } \sum_{j=1}^{r} a_j \mu_j = c, \\ -\infty & \text{otherwise} \end{cases} \end{aligned}$$

- Thus the dual problem is

$$\begin{aligned} &\text{maximize} \quad b'\mu \\ &\text{subject to} \quad \sum_{j=1}^{r} a_j \mu_j = c, \quad \mu \geq 0 \end{aligned}$$

MINIMAX PROBLEMS

Given $\phi : X \times Z \mapsto \Re$, where $X \subset \Re^n$, $Z \subset \Re^m$ consider

$$\begin{aligned} &\text{minimize} && \sup_{z \in Z} \phi(x, z) \\ &\text{subject to} && x \in X \end{aligned}$$

or

$$\begin{aligned} &\text{maximize} && \inf_{x \in X} \phi(x, z) \\ &\text{subject to} && z \in Z \end{aligned}$$

- Some important contexts:
  - Constrained optimization duality theory
  - Zero sum game theory
- We always have

$$\sup_{z \in Z} \inf_{x \in X} \phi(x, z) \leq \inf_{x \in X} \sup_{z \in Z} \phi(x, z)$$

- Key question: When does equality hold?

RELATION TO CONSTRAINED OPTIMIZATION

- For the problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in X, \qquad g(x) \leq 0 \end{aligned}$$

introduce the Lagrangian function

$$L(x, \mu) = f(x) + \mu' g(x)$$

- Write the primal problem as

$$\min_{x \in X} \sup_{\mu \geq 0} L(x, \mu) = \begin{cases} f(x) & \text{if } g(x) \leq 0, \\ \infty & \text{otherwise} \end{cases}$$

- Write the dual problem as

$$\max_{\mu \geq 0} \inf_{x \in X} L(x, \mu)$$

- Key duality question: Is it true that

$$\inf_{x \in \Re^n} \sup_{\mu \geq 0} L(x, \mu) = w^* \overset{?}{=} q^* = \sup_{\mu \geq 0} \inf_{x \in \Re^n} L(x, \mu)$$

ZERO SUM GAMES

- Two players: 1st chooses $i \in \{1, \ldots, n\}$, 2nd chooses $j \in \{1, \ldots, m\}$.
- If i and j are selected, the 1st player gives a_{ij} to the 2nd.
- Mixed strategies are allowed: The two players select probability distributions

$$x = (x_1, \ldots, x_n), \qquad z = (z_1, \ldots, z_m)$$

over their possible choices.

- Probability of (i, j) is $x_i z_j$, so the expected amount to be paid by the 1st player

$$x' A z = \sum_{i,j} a_{ij} x_i z_j$$

where A is the $n \times m$ matrix with elements a_{ij}.

- Each player optimizes his choice against the worst possible selection by the other player. So
  - 1st player minimizes $\max_z x' A z$
  - 2nd player maximizes $\min_x x' A z$

MINIMAX MC/MC FRAMEWORK - SUMMARY

- Introduce perturbation fn $p : \Re^m \mapsto [-\infty, \infty]$

$$p(u) = \inf_{x \in X} \sup_{z \in Z} \big\{ \phi(x, z) - u'z \big\}, \qquad u \in \Re^m$$

- We have

$$w^* = p(0) = \inf_{x \in X} \sup_{z \in Z} \phi(x, z)$$

- Assume that Z is convex, and $-\phi(x, \cdot) : Z \mapsto \Re$ is closed and convex, viewed as a function of $z \in Z$ for every fixed $x \in X$.

- The dual function can be shown to be

$$q(\mu) = \inf_{x \in X} \phi(x, \mu), \qquad \forall\ \mu \in \Re^m,$$

so

$$w^* = \inf_{x \in X} \sup_{z \in Z} \phi(x, z), \qquad q^* = \sup_{z \in Z} \inf_{x \in X} \phi(x, z)$$

- Apply the MC/MC framework with $M = \text{epi}(p)$. We have $\inf_{x \in X} \sup_{z \in Z} \phi(x, z) = \sup_{z \in Z} \inf_{x \in X} \phi(x, z)$ if p is convex, closed, and proper.

DUALITY THEOREMS

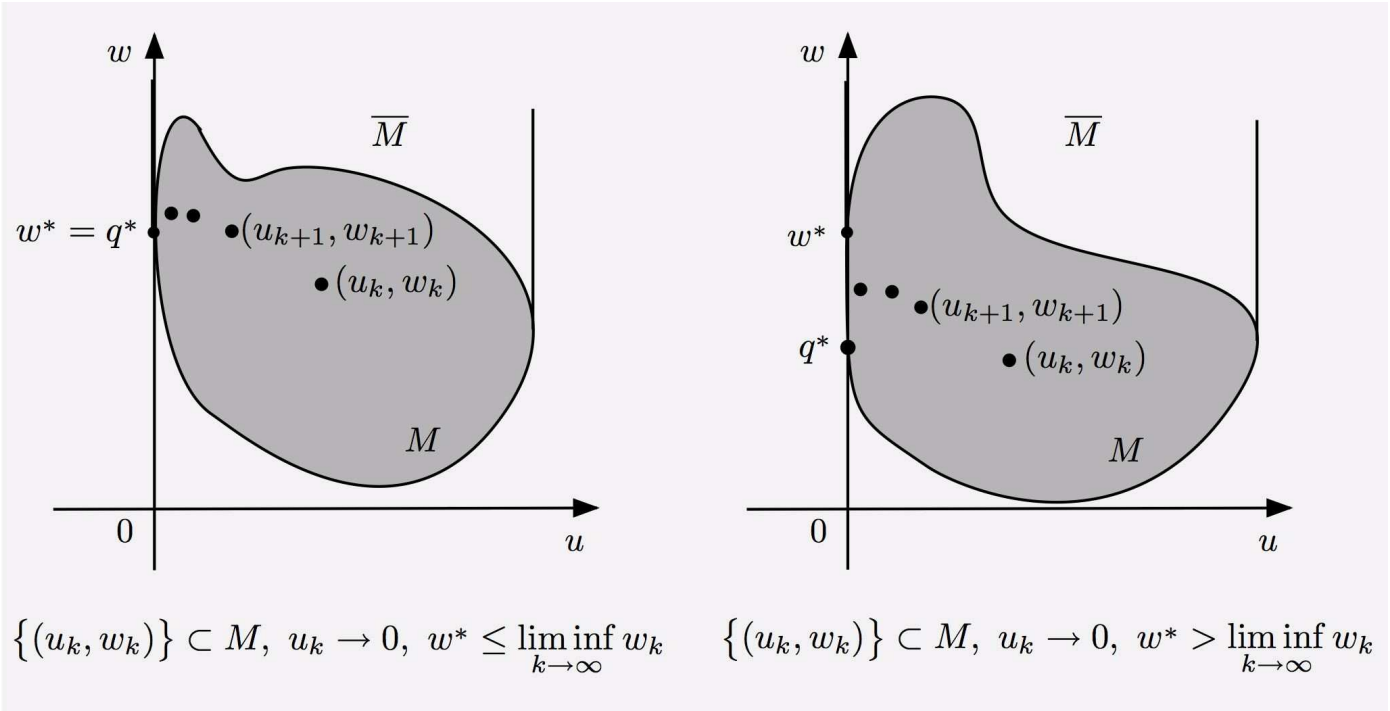
- Assume that $w^* < \infty$ and that the set

$$\overline{M} = \big\{(u, w) \mid \text{there exists } \overline{w} \text{ with } \overline{w} \leq w \text{ and } (u, \overline{w}) \in M\big\}$$

is convex.

- Min Common/Max Crossing Theorem I: We have $q^* = w^*$ if and only if for every sequence $\big\{(u_k, w_k)\big\} \subset M$ with $u_k \to 0$, there holds

$$w^* \leq \liminf_{k \to \infty} w_k.$$



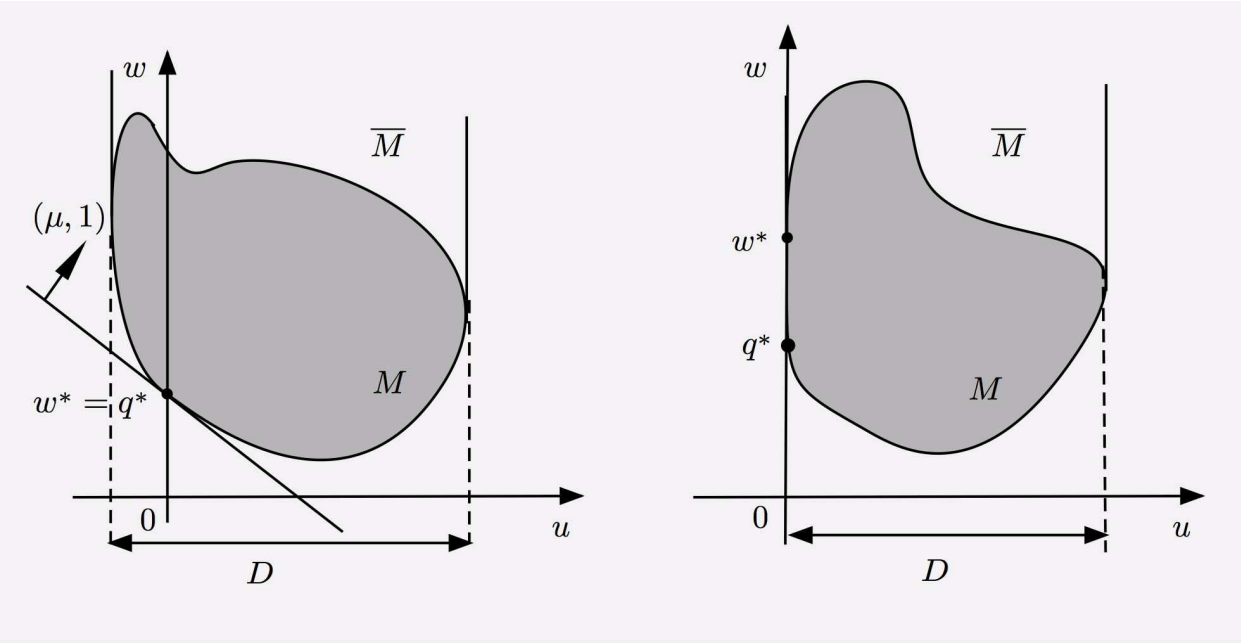
- Corollary: If $M = \text{epi}(p)$ where p is closed proper convex and $p(0) < \infty$, then $q^* = w^*$.

DUALITY THEOREMS (CONTINUED)

• Min Common/Max Crossing Theorem II: Assume in addition that $-\infty < w^*$ and that

$$D = \{u \mid \text{there exists } w \in \Re \text{ with } (u, w) \in \overline{M}\}$$

contains the origin in its relative interior. Then $q^* = w^*$ and there exists μ such that $q(\mu) = q^*$.



• Furthermore, the set $\{\mu \mid q(\mu) = q^*\}$ is nonempty and compact if and only if D contains the origin in its interior.

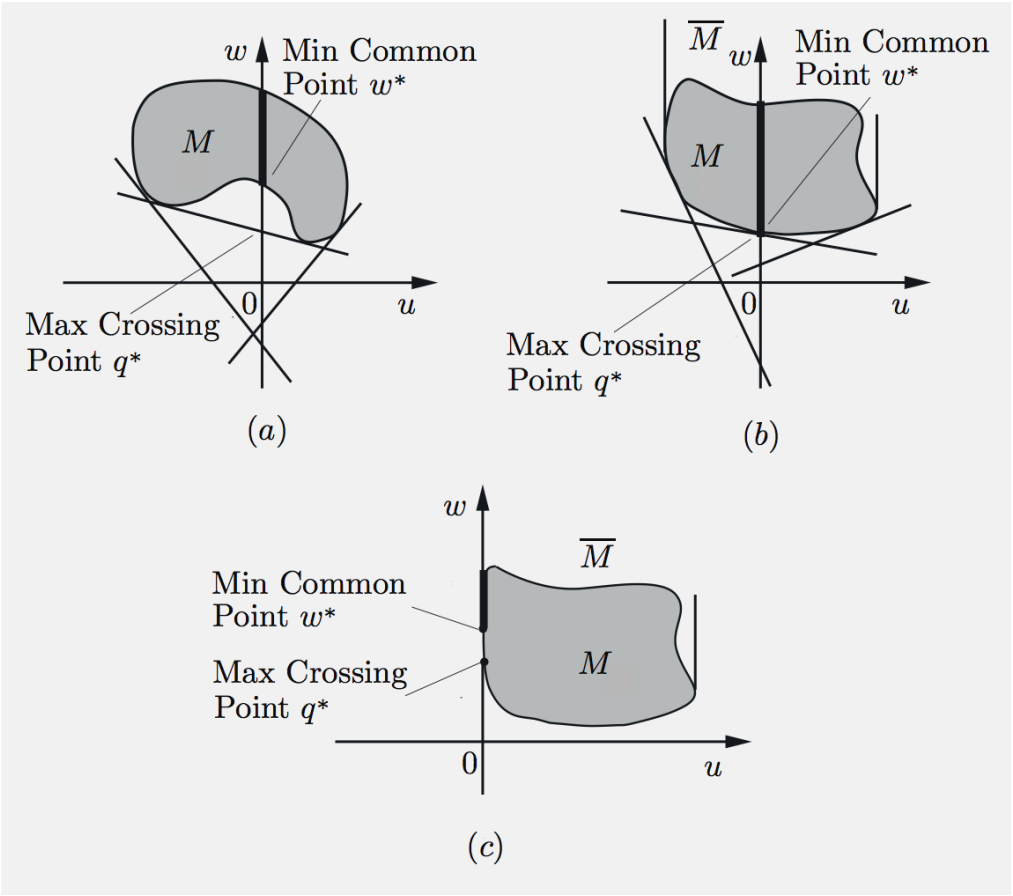
• Min Common/Max Crossing Theorem III: This is a more refined version of Theorem II and involves polyhedral assumptions; see the text.

LECTURE 10

LECTURE OUTLINE

- Strong duality for MC/MC
- Existence of dual optimal solutions
- Nonlinear Farkas' Lemma
- Convex Programming

Reading: Sections 4.3, 4.4, 5.1



REVIEW OF THE MC/MC FRAMEWORK

- Given a set $M \subset \Re^{n+1}$,

$$w^* = \inf_{(0,w) \in M} w$$

$$q^* = \sup_{\mu \in \Re^n} q(\mu) \stackrel{\triangle}{=} \inf_{(u,w) \in M} \{w + \mu' u\}$$

- Weak Duality: $q^* \leq w^*$ (always holds)
- Strong Duality: $q^* = w^*$
- Duality theorems deal with conditions under which:
  - $q^* = w^*$
  - The dual problem or the primal problem have an optimal solution
  - Necessary and sufficient conditions under which a pair of primal and dual variables are optimal for the primal and dual problems, respectively.
- We will address the first two questions in the general MC/MC setting.
- We will address the third question in specific settings, such as constrained optimization duality, Fenchel duality, conic duality, etc.

DUALITY THEOREM I

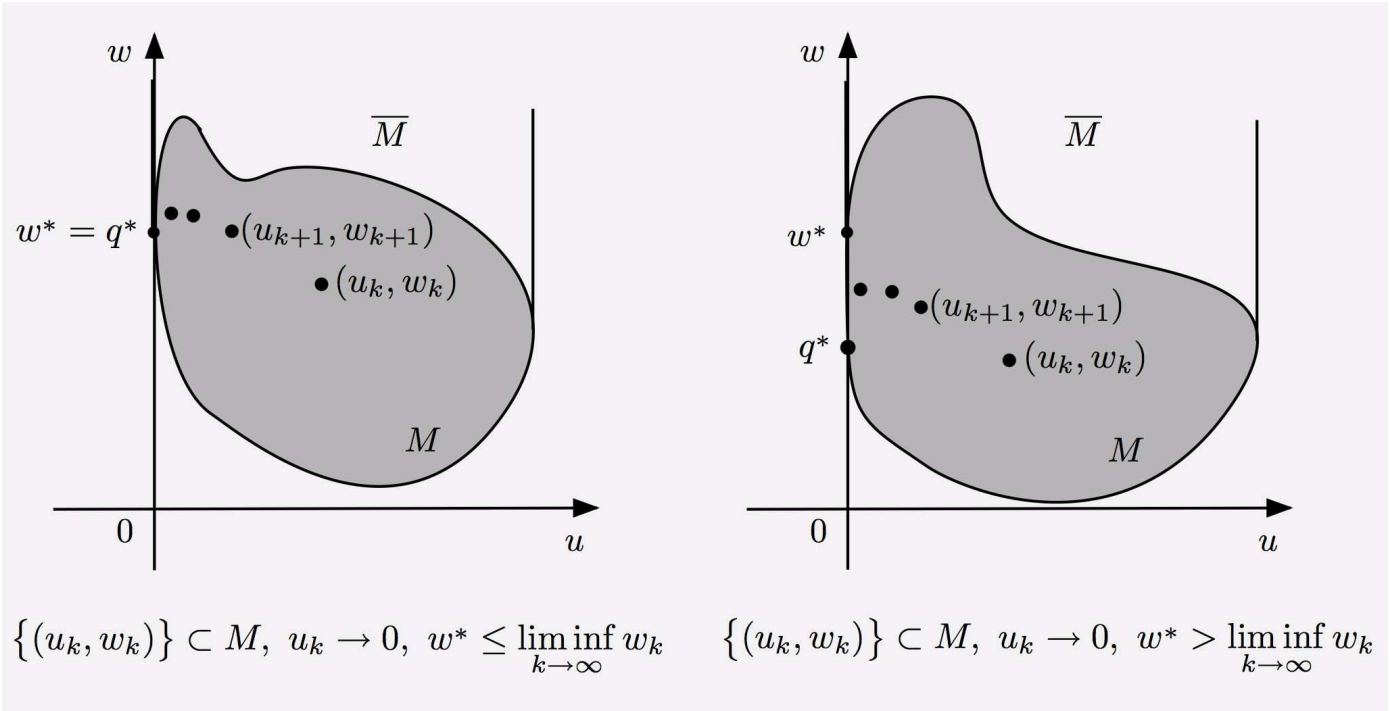
- Assume that $w^* < \infty$ and that the set

$$\overline{M} = \big\{(u, w) \mid \text{there exists } \overline{w} \text{ with } \overline{w} \leq w \text{ and } (u, \overline{w}) \in M\big\}$$

is convex.

- Min Common/Max Crossing Theorem I: We have $q^* = w^*$ if and only if for every sequence $\big\{(u_k, w_k)\big\} \subset M$ with $u_k \to 0$, there holds

$$w^* \leq \liminf_{k \to \infty} w_k.$$



PROOF OF THEOREM I

- Assume that $q^* = w^*$. Let $\{(u_k, w_k)\} \subset M$ be such that $u_k \to 0$. Then,

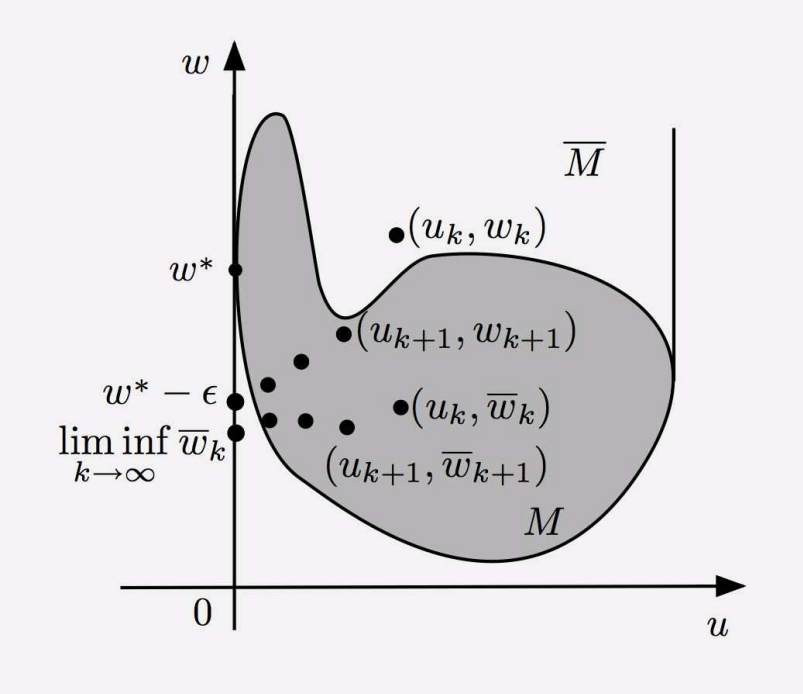
$$q(\mu) = \inf_{(u,w)\in M} \{w+\mu' u\} \leq w_k + \mu' u_k, \quad \forall\, k,\ \forall\, \mu \in \Re^n$$

Taking the limit as $k \to \infty$, we obtain $q(\mu) \leq \liminf_{k\to\infty} w_k$, for all $\mu \in \Re^n$, implying that

$$w^* = q^* = \sup_{\mu\in\Re^n} q(\mu) \leq \liminf_{k\to\infty} w_k$$

Conversely, assume that for every sequence $\{(u_k, w_k)\} \subset M$ with $u_k \to 0$, there holds $w^* \leq \liminf_{k\to\infty} w_k$. If $w^* = -\infty$, then $q^* = -\infty$, by weak duality, so assume that $-\infty < w^*$. Steps:

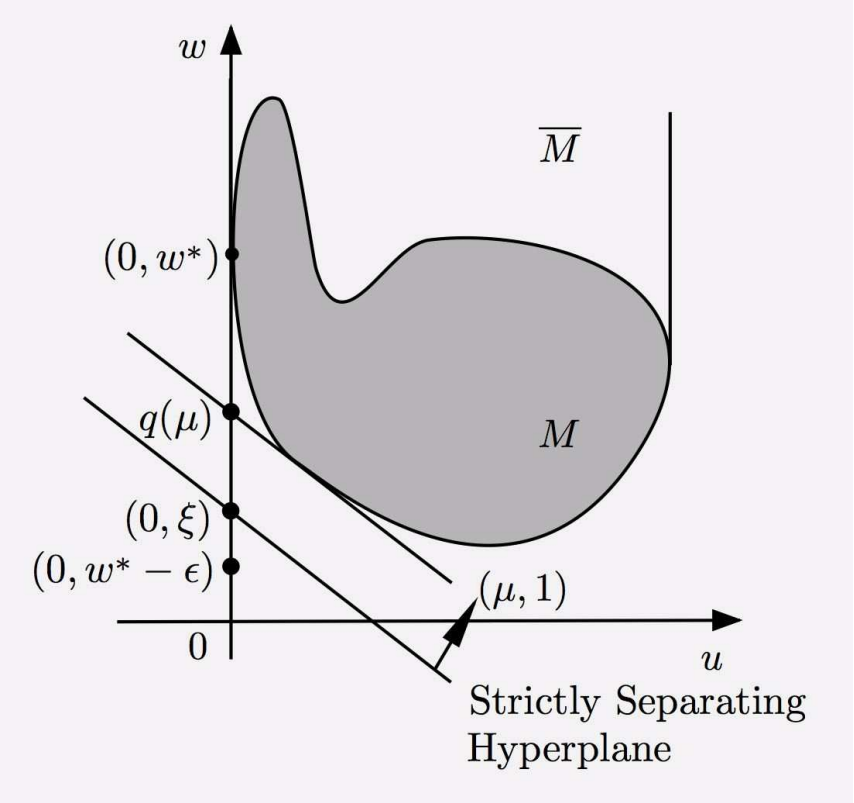
- Step 1: $(0, w^* - \epsilon) \notin \text{cl}(\overline{M})$ for any $\epsilon > 0$.



PROOF OF THEOREM I (CONTINUED)

• Step 2: $\overline{M}$ does not contain any vertical lines. If this were not so, $(0, -1)$ would be a direction of recession of $\text{cl}(\overline{M})$. Because $(0, w^*) \in \text{cl}(\overline{M})$, the entire halfline $\{(0, w^* - \epsilon) \mid \epsilon \geq 0\}$ belongs to $\text{cl}(\overline{M})$, contradicting Step 1.

• Step 3: For any $\epsilon > 0$, since $(0, w^* - \epsilon) \notin \text{cl}(\overline{M})$, there exists a nonvertical hyperplane strictly separating $(0, w^* - \epsilon)$ and $\overline{M}$. This hyperplane crosses the $(n+1)$st axis at a vector $(0, \xi)$ with $w^* - \epsilon \leq \xi \leq w^*$, so $w^* - \epsilon \leq q^* \leq w^*$. Since ϵ can be arbitrarily small, it follows that $q^* = w^*$.

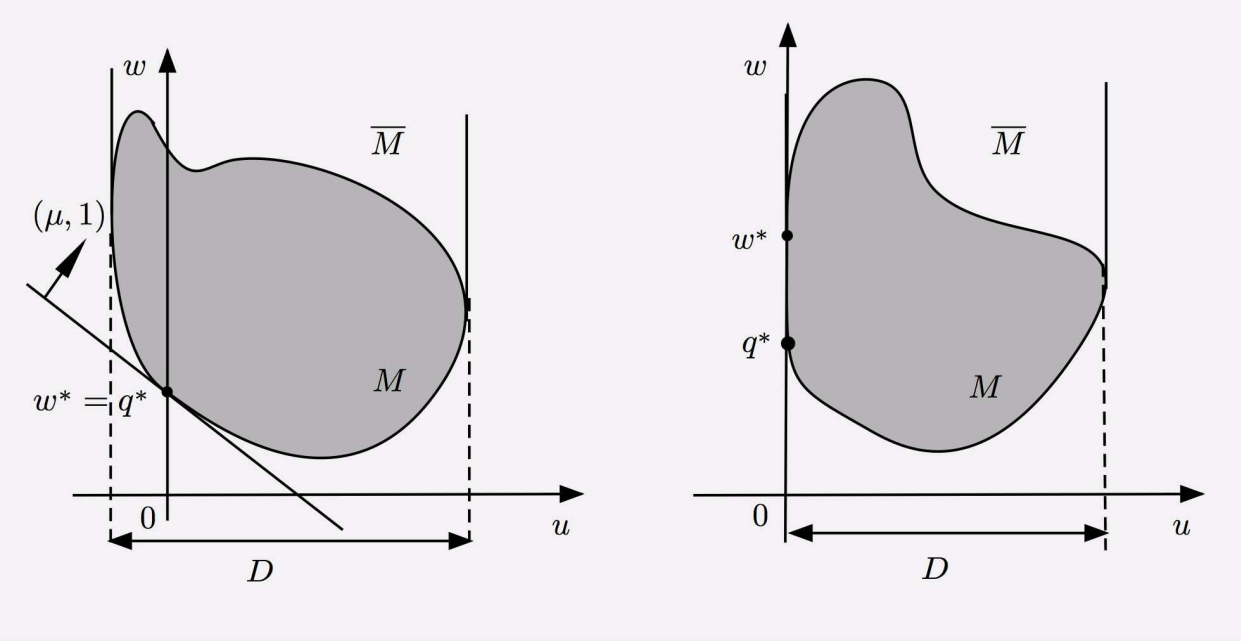


DUALITY THEOREM II

• Min Common/Max Crossing Theorem II: Assume in addition that $-\infty < w^*$ and that

$$D = \{u \mid \text{there exists } w \in \Re \text{ with } (u, w) \in \overline{M}\}$$

contains the origin in its relative interior. Then $q^* = w^*$ and there exists μ such that $q(\mu) = q^*$.



• Furthermore, the set $\{\mu \mid q(\mu) = q^*\}$ is nonempty and compact if and only if D contains the origin in its interior.

PROOF OF THEOREM II

• Hyperplane Separation Argument: Note that $(0, w^*)$ is not a relative interior point of $\overline{M}$. Therefore, by the Proper Separation Theorem, there is a hyperplane that passes through $(0, w^*)$, contains $\overline{M}$ in one of its closed halfspaces, but does not fully contain $\overline{M}$, i.e., for some $(\mu, \beta) \neq (0, 0)$

$$\beta w^* \leq \mu' u + \beta w, \qquad \forall\ (u, w) \in \overline{M}, \qquad (*)$$

$$\beta w^* < \sup_{(u,w)\in\overline{M}} \{\mu' u + \beta w\} \qquad (**)$$

We will show that the hyperplane is nonvertical.

• Since for any $(\overline{u}, \overline{w}) \in M$, the set $\overline{M}$ contains the halfline $\big\{(\overline{u}, w) \mid \overline{w} \leq w\big\}$, it follows that $\beta \geq 0$. If $\beta = 0$, then from (*), $0 \leq \mu' u$ for all $u \in D$. Since $0 \in \text{ri}(D)$ by assumption, we must have $\mu' u = 0$ for all $u \in D$ (by Prop. 1.3.4 of the text) a contradiction of (**). Therefore, $\beta > 0$, and we can assume that $\beta = 1$. It follows from (*) that

$$w^* \leq \inf_{(u,w)\in\overline{M}} \{\mu' u + w\} = q(\mu) \leq q^*$$

Since the inequality $q^* \leq w^*$ holds always, we must have $q(\mu) = q^* = w^*$.

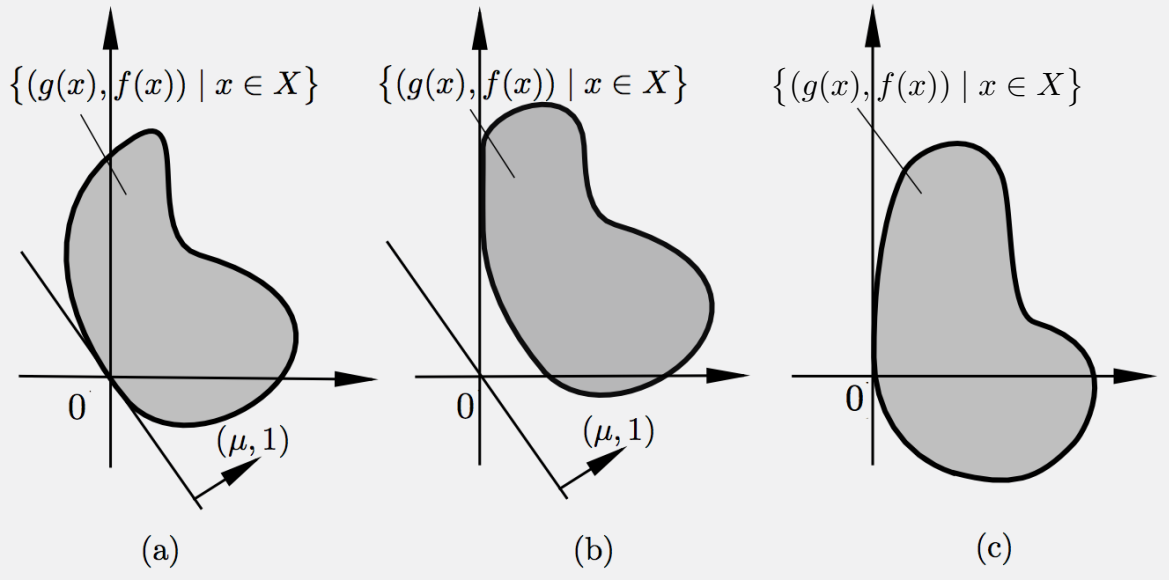
NONLINEAR FARKAS' LEMMA

• Let $X \subset \Re^n$, $f : X \mapsto \Re$, and $g_j : X \mapsto \Re$, $j = 1, \ldots, r$, be convex. Assume that

$$f(x) \geq 0, \qquad \forall\ x \in X \text{ with } g(x) \leq 0$$

Assume there exists a vector $\overline{x} \in X$ such that $g_j(\overline{x}) < 0$ for all $j = 1, \ldots, r$. Then there exists $\mu \geq 0$ such that

$$f(x) + \mu' g(x) \geq 0, \qquad \forall\ x \in X$$



• The lemma asserts the existence of a nonvertical hyperplane in $\Re^{r+1}$, with normal $(\mu, 1)$, that passes through the origin and contains the set

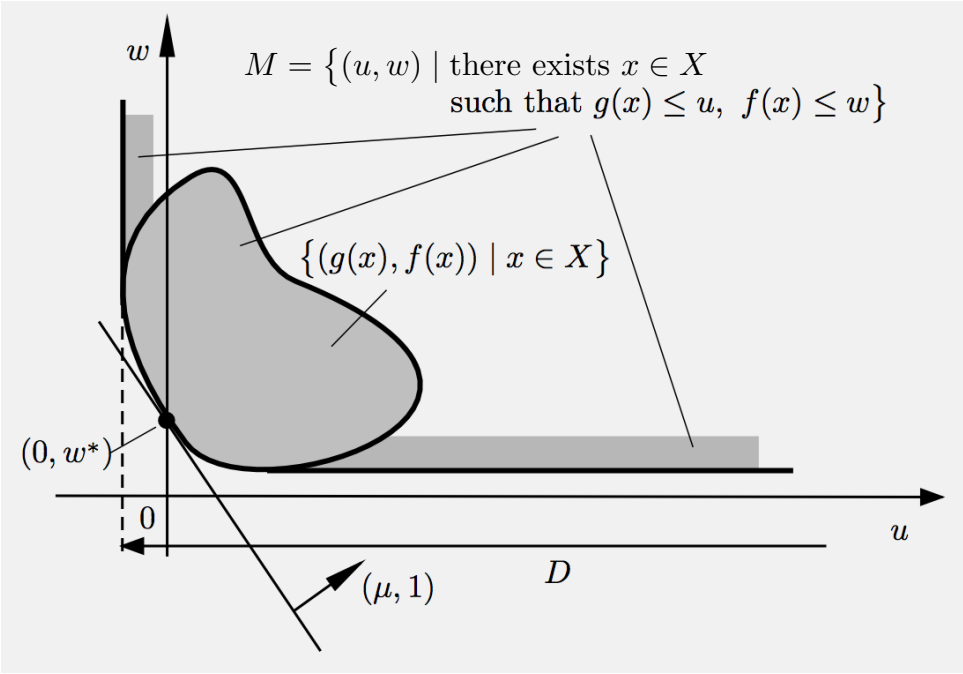
$$\Big\{\big(g(x), f(x)\big) \mid x \in X\Big\}$$

in its positive halfspace.

PROOF OF NONLINEAR FARKAS' LEMMA

- Apply MC/MC to

$$M = \big\{(u,w) \mid \text{there is } x \in X \text{ s. t. } g(x) \le u,\ f(x) \le w\big\}$$



- M is equal to $\overline{M}$ and is the union of positive orthants translated to points $\big(g(x), f(x)\big)$, $x \in X$.
- Since X, f, and g_j are convex, M is convex (requires a proof).
- MC/MC Theorem II applies: we have

$$D = \big\{u \mid \text{there exists } w \in \Re \text{ with } (u,w) \in M\big\}$$

and $0 \in \text{int}(D)$, because $\big(g(\overline{x}), f(\overline{x})\big) \in M$.

CONVEX PROGRAMMING

Consider the problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in X,\ g_j(x) \leq 0,\ j = 1, \ldots, r, \end{aligned}$$

where $X \subset \Re^n$ is convex, and $f : X \mapsto \Re$ and $g_j : X \mapsto \Re$ are convex. Assume f^*: finite.

- Recall the connection with the max crossing problem in the MC/MC framework where $M = \text{epi}(p)$ with

$$p(u) = \inf_{x \in X,\, g(x) \leq u} f(x)$$

- Consider the Lagrangian function

$$L(x, \mu) = f(x) + \mu' g(x),$$

the dual function

$$q(\mu) = \begin{cases} \inf_{x \in X} L(x, \mu) & \text{if } \mu \geq 0, \\ -\infty & \text{otherwise} \end{cases}$$

and the dual problem of maximizing $\inf_{x \in X} L(x, \mu)$ over $\mu \geq 0$.

STRONG DUALITY TH. - SLATER CONDITION

- Assume that f^* is finite, and there exists $\overline{x} \in X$ such that $g(\overline{x}) < 0$. Then $q^* = f^*$ and the set of optimal solutions of the dual problem is nonempty and compact.

Proof: Replace $f(x)$ by $f(x) - f^*$ so that $f(x) - f^* \geq 0$ for all $x \in X$ w/ $g(x) \leq 0$. Apply Nonlinear Farkas' Lemma. Then, there exist $\mu_j^* \geq 0$, s.t.

$$f^* \leq f(x) + \sum_{j=1}^{r} \mu_j^* g_j(x), \qquad \forall\ x \in X$$

- It follows that

$$f^* \leq \inf_{x \in X} \left\{ f(x) + \mu^{*\prime} g(x) \right\} \leq \inf_{x \in X,\ g(x) \leq 0} f(x) = f^*.$$

Thus equality holds throughout, and we have

$$f^* = \inf_{x \in X} \left\{ f(x) + \sum_{j=1}^{r} \mu_j^* g_j(x) \right\} = q(\mu^*)$$

NONL. FARKAS' L. - POLYHEDRAL ASSUM.

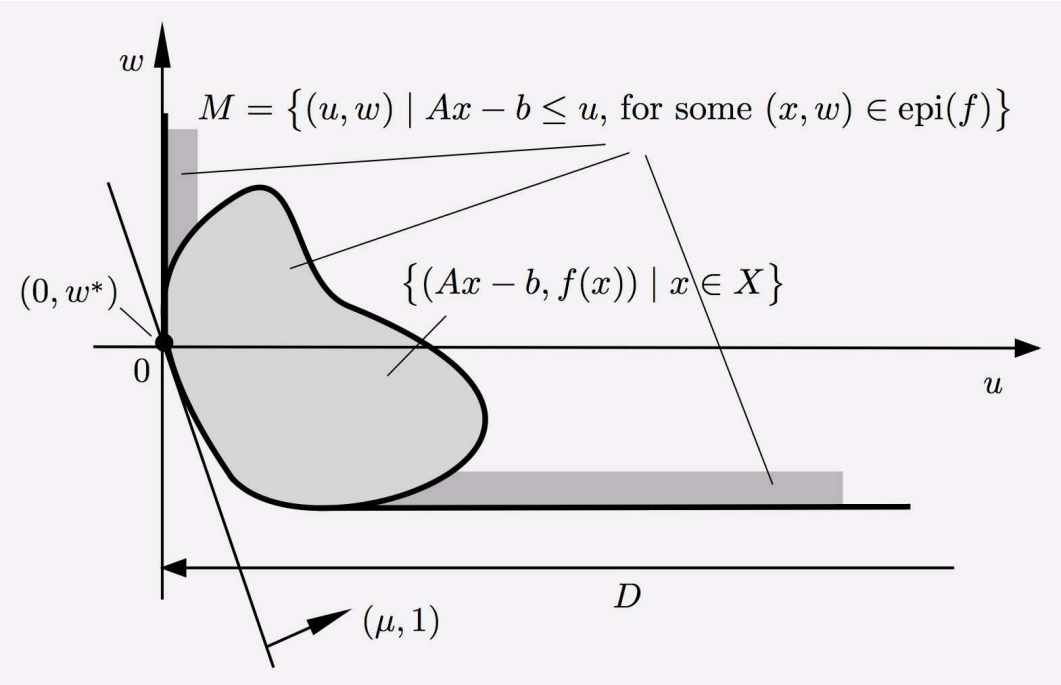
• Let $X \subset \Re^n$ be convex, and $f : X \mapsto \Re$ and $g_j : \Re^n \mapsto \Re$, $j = 1, \ldots, r$, be linear so $g(x) = Ax - b$ for some A and b. Assume that

$$f(x) \geq 0, \qquad \forall\ x \in X \text{ with } Ax - b \leq 0$$

and that there exists a vector $\overline{x} \in \text{ri}(X)$ such that $A\overline{x} - b \leq 0$. Then there exists $\mu \geq 0$ such that

$$f(x) + \mu'(Ax - b) \geq 0, \ \forall\ x \in X$$

Proof: This is an application of MC/MC Theorem III (next slide), which involves polyhedral assumptions; see the text for proof and analysis.



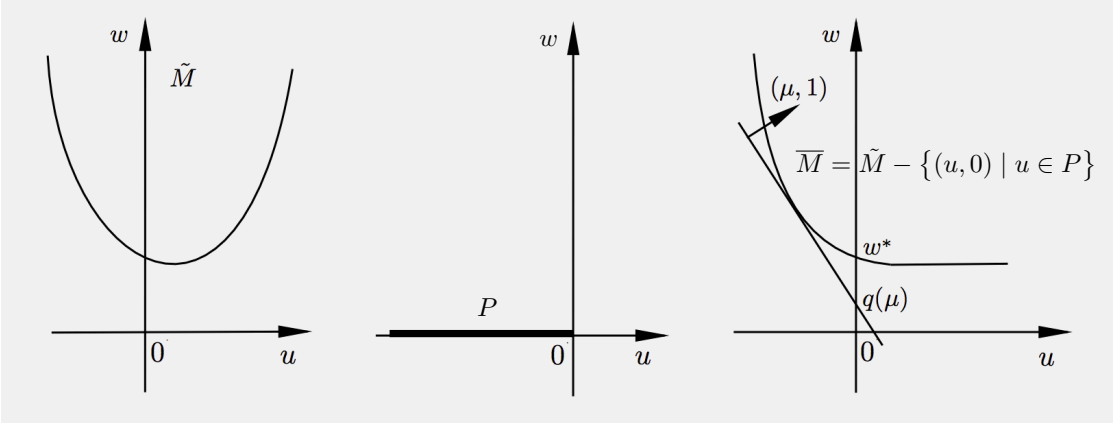
MC/MC TH. III - POLYHEDRAL

• Consider the MC/MC problems, and assume that $-\infty < w^*$ and:

(1) $\overline{M}$ is a "horizontal translation" of $\tilde{M}$ by $-P$,

$$\overline{M} = \tilde{M} - \{(u, 0) \mid u \in P\},$$

where P: polyhedral and $\tilde{M}$: convex.



(2) We have $\text{ri}(\tilde{D}) \cap P \neq \emptyset$, where

$$\tilde{D} = \{u \mid \text{there exists } w \in \Re \text{ with } (u, w) \in \tilde{M}\}$$

Then $q^* = w^*$, there is a max crossing solution, and all max crossing solutions $\overline{\mu}$ satisfy $\overline{\mu}' d \leq 0$ for all $d \in R_P$.

• Compare with Th. II: Since $D = \tilde{D} - P$, the condition $0 \in \text{ri}(D)$ of Th. II is $\text{ri}(\tilde{D}) \cap \text{ri}(P) \neq \emptyset$. Proof is similar, but uses the polyhedral proper separation theorem.

STRONG DUALITY - POLYHEDRAL CONSTR.

- Assume that f^* is finite, the functions g_j, $j = 1, \ldots, r$, are affine, and one of the following two conditions holds:

  (1) X is polyhedral.

  (2) There exists $\overline{x} \in \text{ri}(X)$ such that $g(\overline{x}) \leq 0$.

Then $q^* = f^*$ and the set of optimal solutions of the dual problem is nonempty.

Proof: Replace $f(x)$ by $f(x) - f^*$ so that $f(x) - f^* \geq 0$ for all $x \in X$ w/ $g(x) \leq 0$. Apply Nonlinear Farkas' Lemma for polyhedral assumptions.

- Note: For the special case where:

  (a) There exists an optimal primal solution x^*

  (b) $X = \Re^n$

we have already proved that there exists a Lagrange multiplier vector (a dual optimal solution) using the Polar Cone Theorem, which is the same as the linear version of Farkas' Lemma.

The sharper version given here shows that strong duality holds even if there is no optimal primal solution, and X is nonpolyhedral.

LECTURE 11

LECTURE OUTLINE

- Review of Convex Programming Duality
- Optimality Conditions
- Fenchel Duality

Reading: Sections 5.3.1, 5.3.2, 5.3.3, 5.3.5

CONVEX PROGRAMMING DUALITY REVIEW

Strong Duality Theorem: Consider the problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in X,\ g_j(x) \leq 0,\ j = 1, \ldots, r, \end{aligned}$$

where $X \subset \Re^n$ is convex, and $f : X \mapsto \Re$ and $g_j : X \mapsto \Re$ are convex.

- Assume that f^* is finite, and that one of the following two conditions holds:
  - (1) There exists $\overline{x} \in X$ such that $g(\overline{x}) < 0$.
  - (2) The functions g_j, $j = 1, \ldots, r$, are affine, and there exists $\overline{x} \in \text{ri}(X)$ such that $g(\overline{x}) \leq 0$.

Then $q^* = f^*$ and the set of optimal solutions of the dual problem is nonempty. Under condition (1) this set is also compact.

- Important remaining questions:
  - Optimality conditions for (x^*, μ^*) to be an optimal primal and dual solution pair.
  - Extensions to the case of mixed (linear) equality constraints, and mixture of linear and convex inequality constraints.
  - Extension to the Fenchel duality framework.

COUNTEREXAMPLE I

- Strong Duality Counterexample: Consider

$$\begin{aligned} &\text{minimize} \quad f(x) = e^{-\sqrt{x_1 x_2}} \\ &\text{subject to} \quad x_1 \leq 0, \qquad x \in X = \{x \mid x \geq 0\} \end{aligned}$$

Here $f^* = 1$ and f is convex (its Hessian is > 0 in the interior of X). The dual function is

$$q(\mu) = \inf_{x \geq 0} \left\{ e^{-\sqrt{x_1 x_2}} + \mu x_1 \right\} = \begin{cases} 0 & \text{if } \mu \geq 0, \\ -\infty & \text{otherwise,} \end{cases}$$

(when $\mu \geq 0$, the expression in braces is nonnegative for $x \geq 0$ and can approach zero by taking $x_1 \to 0$ and $x_1 x_2 \to \infty$). Thus $q^* = 0$.

- The relative interior assumption is violated.

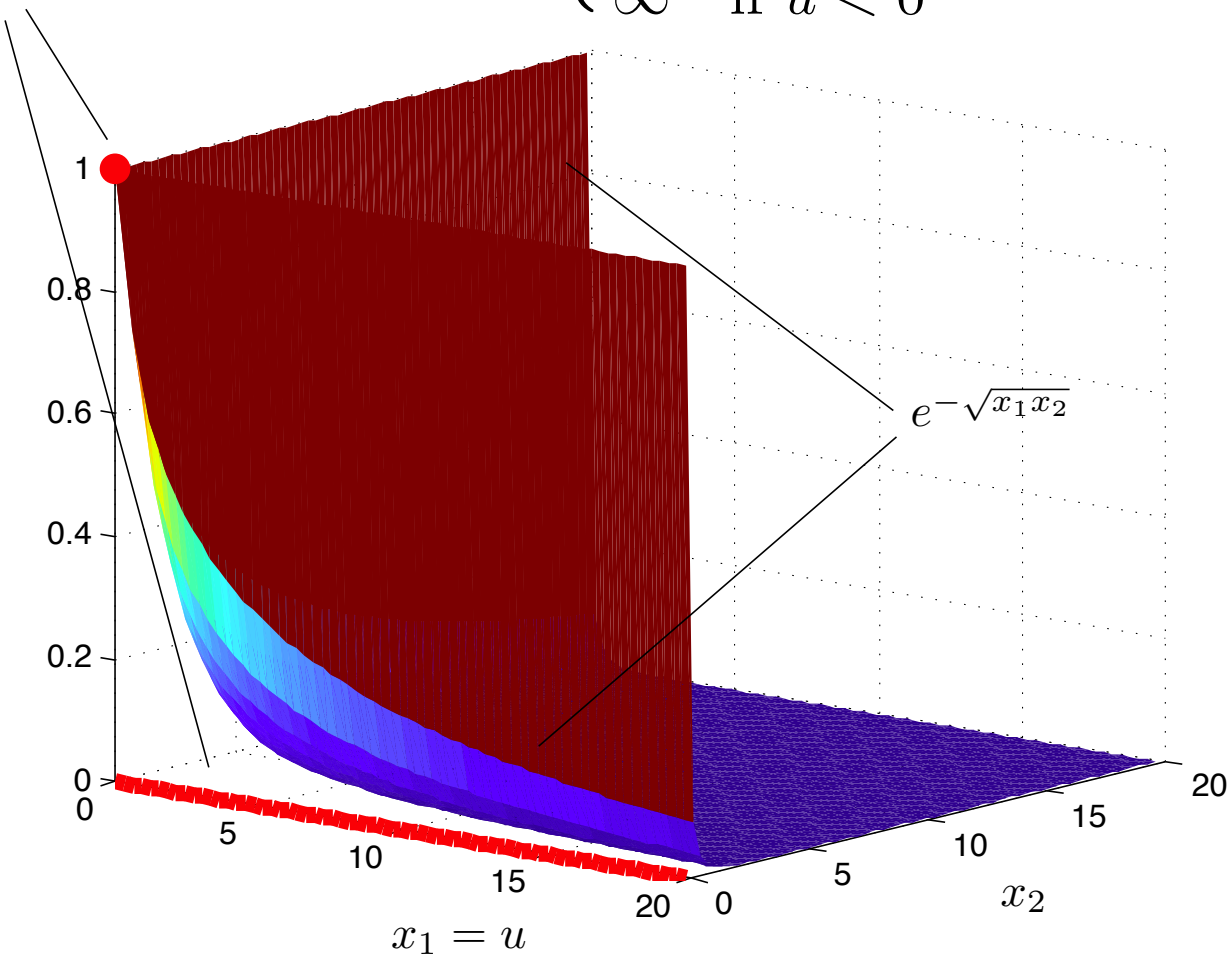
- As predicted by the corresponding MC/MC framework, the perturbation function

$$p(u) = \inf_{x_1 \leq u,\, x \geq 0} e^{-\sqrt{x_1 x_2}} = \begin{cases} 0 & \text{if } u > 0, \\ 1 & \text{if } u = 0, \\ \infty & \text{if } u < 0, \end{cases}$$

is not lower semicontinuous at $u = 0$.

COUNTEREXAMPLE I VISUALIZATION

$$p(u) = \inf_{x_1 \le u,\, x \ge 0} e^{-\sqrt{x_1 x_2}} = \begin{cases} 0 & \text{if } u > 0 \\ 1 & \text{if } u = 0 \\ \infty & \text{if } u < 0 \end{cases}$$



- Connection with counterexample for preservation of closedness under partial minimization.

COUNTEREXAMPLE II

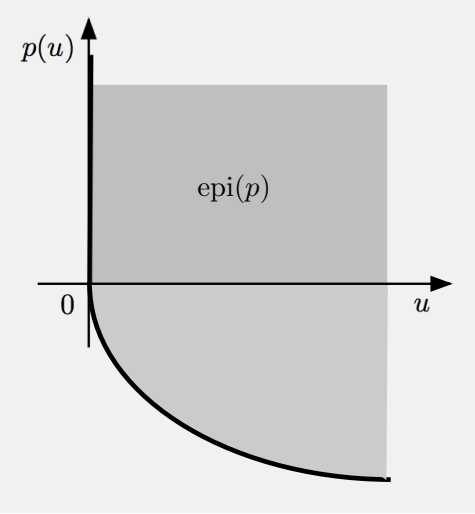
- Existence of Dual Solutions Counterexample: Let $X = \Re$, $f(x) = x$, $g(x) = x^2$. Then $x^* = 0$ is the only feasible/optimal solution, and we have

$$q(\mu) = \inf_{x \in \Re} \{x + \mu x^2\} = -\frac{1}{4\mu}, \qquad \forall\ \mu > 0,$$

and $q(\mu) = -\infty$ for $\mu \leq 0$, so that $q^* = f^* = 0$. However, there is no $\mu^* \geq 0$ such that $q(\mu^*) = q^* = 0$, and the dual problem has no optimal solution.

- Here the perturbation function is

$$p(u) = \inf_{x^2 \leq u} x = \begin{cases} -\sqrt{u} & \text{if } u \geq 0, \\ \infty & \text{if } u < 0. \end{cases}$$



QUADRATIC PROGRAMMING DUALITY

- Consider the quadratic program

$$\begin{aligned} &\text{minimize} \quad \tfrac{1}{2}x'Qx + c'x \\ &\text{subject to} \quad Ax \leq b, \end{aligned}$$

where Q is positive definite.

- If f^* is finite, then $f^* = q^*$ and there exist both primal and dual optimal solutions, since the constraints are linear.

- Calculation of dual function:

$$q(\mu) = \inf_{x \in \Re^n} \{\tfrac{1}{2}x'Qx + c'x + \mu'(Ax - b)\}$$

The infimum is attained for $x = -Q^{-1}(c + A'\mu)$, and, after substitution and calculation,

$$q(\mu) = -\tfrac{1}{2}\mu' AQ^{-1}A'\mu - \mu'(b + AQ^{-1}c) - \tfrac{1}{2}c'Q^{-1}c$$

- The dual problem, after a sign change, is

$$\begin{aligned} &\text{minimize} \quad \tfrac{1}{2}\mu' P\mu + t'\mu \\ &\text{subject to} \quad \mu \geq 0, \end{aligned}$$

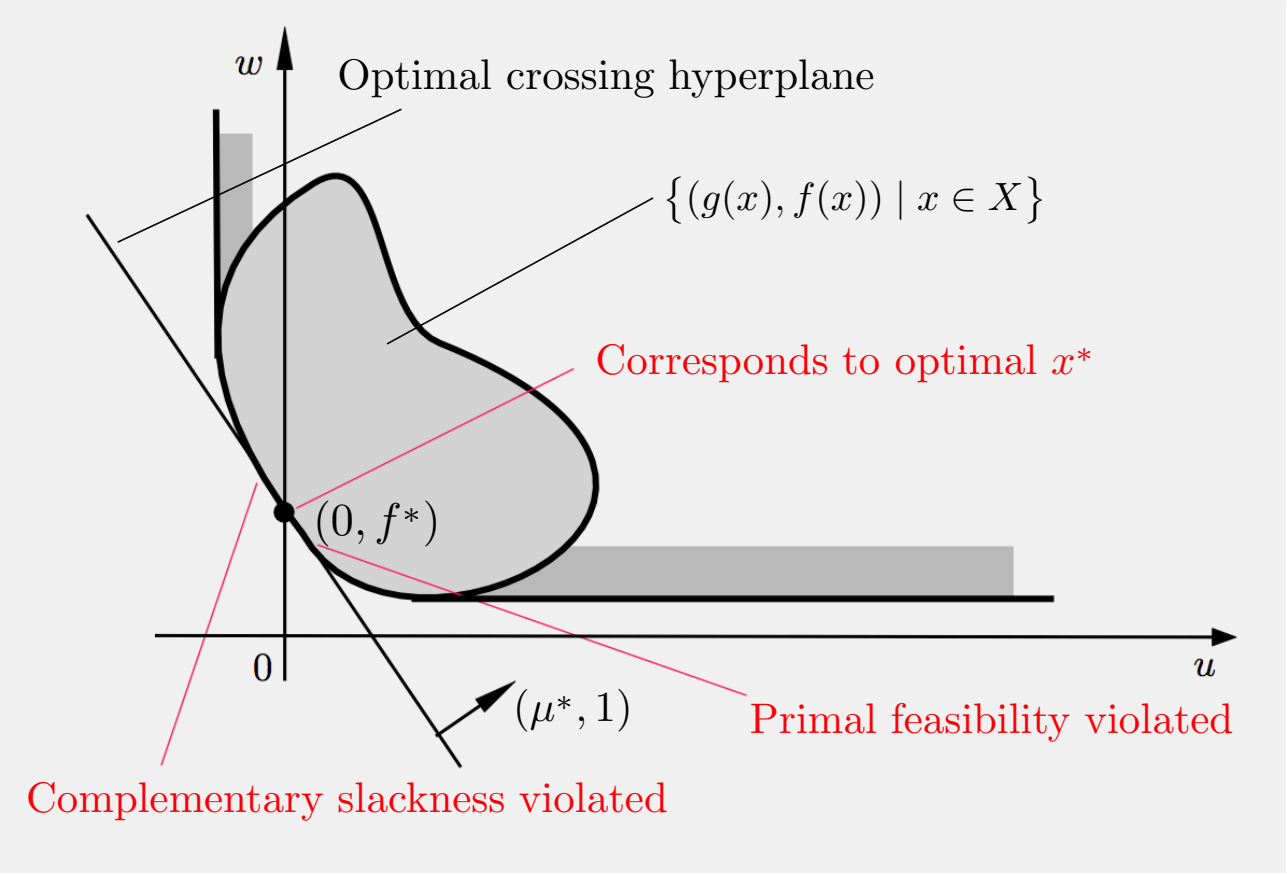
where $P = AQ^{-1}A'$ and $t = b + AQ^{-1}c$.

OPTIMALITY CONDITIONS - VISUALIZATION

- Suppose we have dual optimal μ^* and no duality gap

$$q(\mu^*) = q^* = f^*$$

- How do we find primal optimal x^*?



- We look for x that correspond to points on the max crossing hyperplane, i.e., $\overline{x}$ such that

$$L(\overline{x}, \mu^*) = \inf_{x \in X} L(x, \mu^*) = q(\mu^*) = q^* = f^*$$

OPTIMALITY CONDITIONS

• We have $q^* = f^*$, and the vectors x^* and μ^* are optimal solutions of the primal and dual problems, respectively, iff x^* is feasible, $\mu^* \geq 0$, and

$$x^* \in \arg\min_{x \in X} L(x, \mu^*), \qquad \mu_j^* g_j(x^*) = 0, \quad \forall\ j. \tag{*}$$

Proof: If $q^* = f^*$, and x^*, μ^* are optimal, then

$$f^* = q^* = q(\mu^*) = \inf_{x \in X} L(x, \mu^*) \leq L(x^*, \mu^*)$$

$$= f(x^*) + \sum_{j=1}^{r} \mu_j^* g_j(x^*) \leq f(x^*),$$

where the last inequality follows from $\mu_j^* \geq 0$ and $g_j(x^*) \leq 0$ for all j. Hence equality holds throughout above, and (*) holds.

Conversely, if x^*, μ^* are feasible, and (*) holds,

$$q(\mu^*) = \inf_{x \in X} L(x, \mu^*) = L(x^*, \mu^*)$$

$$= f(x^*) + \sum_{j=1}^{r} \mu_j^* g_j(x^*) = f(x^*),$$

so $q^* = f^*$, and x^*, μ^* are optimal. **Q.E.D.**

QUADRATIC PROGRAMMING OPT. COND.

For the quadratic program

$$\begin{aligned} &\text{minimize} \quad \tfrac{1}{2}x'Qx + c'x \\ &\text{subject to} \quad Ax \leq b, \end{aligned}$$

where Q is positive definite, (x^*, μ^*) is a primal and dual optimal solution pair if and only if:

- Primal and dual feasibility holds:

$$Ax^* \leq b, \qquad \mu^* \geq 0$$

- Lagrangian optimality holds [x^* minimizes $L(x, \mu^*)$ over $x \in \Re^n$]. This yields

$$x^* = -Q^{-1}(c + A'\mu^*)$$

- Complementary slackness holds [$(Ax^* - b)'\mu^* = 0$]. It can be written as

$$\mu_j^* > 0 \qquad \Rightarrow \qquad a_j'x^* = b_j, \quad \forall\ j = 1, \ldots, r,$$

where a_j' is the jth row of A, and b_j is the jth component of b.

LINEAR EQUALITY CONSTRAINTS

- The problem is

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in X, \quad g(x) \leq 0, \quad Ax = b, \end{aligned}$$

where X is convex, $g(x) = \big(g_1(x), \ldots, g_r(x)\big)'$, $f : X \mapsto \Re$ and $g_j : X \mapsto \Re$, $j = 1, \ldots, r$, are convex.

- Convert the constraint $Ax = b$ to $Ax \leq b$ and $-Ax \leq -b$, with corresponding dual variables $\lambda^+ \geq 0$ and $\lambda^- \geq 0$.

- The Lagrangian function is

$$f(x) + \mu' g(x) + (\lambda^+ - \lambda^-)'(Ax - b),$$

and by introducing a dual variable $\lambda = \lambda^+ - \lambda^-$, with no sign restriction, it can be written as

$$L(x, \mu, \lambda) = f(x) + \mu' g(x) + \lambda'(Ax - b).$$

- The dual problem is

$$\begin{aligned} &\text{maximize} \quad q(\mu, \lambda) \equiv \inf_{x \in X} L(x, \mu, \lambda) \\ &\text{subject to} \quad \mu \geq 0, \ \lambda \in \Re^m. \end{aligned}$$

DUALITY AND OPTIMALITY COND.

- Pure equality constraints:

  (a) Assume that f^*: finite and there exists $\overline{x} \in \text{ri}(X)$ such that $A\overline{x} = b$. Then $f^* = q^*$ and there exists a dual optimal solution.

  (b) $f^* = q^*$, and (x^*, λ^*) are a primal and dual optimal solution pair if and only if x^* is feasible, and

$$x^* \in \arg\min_{x \in X} L(x, \lambda^*)$$

Note: No complementary slackness for equality constraints.

- Linear and nonlinear constraints:

  (a) Assume f^*: finite, that there exists $\overline{x} \in X$ such that $A\overline{x} = b$ and $g(\overline{x}) < 0$, and that there exists $\tilde{x} \in \text{ri}(X)$ such that $A\tilde{x} = b$. Then $q^* = f^*$ and there exists a dual optimal solution.

  (b) $f^* = q^*$, and (x^*, μ^*, λ^*) are a primal and dual optimal solution pair if and only if x^* is feasible, $\mu^* \geq 0$, and

$$x^* \in \arg\min_{x \in X} L(x, \mu^*, \lambda^*), \quad \mu_j^* g_j(x^*) = 0, \quad \forall\, j$$

FENCHEL DUALITY FRAMEWORK

- Consider the problem

$$\begin{aligned} &\text{minimize} \quad f_1(x) + f_2(x) \\ &\text{subject to} \quad x \in \Re^n, \end{aligned}$$

where $f_1 : \Re^n \mapsto (-\infty, \infty]$ and $f_2 : \Re^n \mapsto (-\infty, \infty]$ are closed proper convex functions.

- Convert to the equivalent problem

$$\begin{aligned} &\text{minimize} \quad f_1(x_1) + f_2(x_2) \\ &\text{subject to} \quad x_1 = x_2, \quad x_1 \in \text{dom}(f_1),\ x_2 \in \text{dom}(f_2) \end{aligned}$$

- The dual function is

$$\begin{aligned} q(\lambda) &= \inf_{x_1 \in \text{dom}(f_1),\ x_2 \in \text{dom}(f_2)} \Big\{ f_1(x_1) + f_2(x_2) + \lambda'(x_2 - x_1) \Big\} \\ &= \inf_{x_1 \in \Re^n} \Big\{ f_1(x_1) - \lambda' x_1 \Big\} + \inf_{x_2 \in \Re^n} \Big\{ f_2(x_2) + \lambda' x_2 \Big\} \end{aligned}$$

- Dual problem: $\max_\lambda \{-f_1^\star(\lambda) - f_2^\star(-\lambda)\} = -\min_\lambda \{-q(\lambda)\}$ or

$$\begin{aligned} &\text{minimize} \quad f_1^\star(\lambda) + f_2^\star(-\lambda) \\ &\text{subject to} \quad \lambda \in \Re^n, \end{aligned}$$

where $f_1^\star$ and $f_2^\star$ are the conjugates.

FENCHEL DUALITY THEOREM

- Consider the Fenchel problem $\min_{x\in\Re^n} f_1(x) + f_2(x)$:

  (a) If f^* is finite and $\mathrm{ri}\big(\mathrm{dom}(f_1)\big)\cap\mathrm{ri}\big(\mathrm{dom}(f_2)\big) \neq \emptyset$, then $f^* = q^*$ and there exists at least one dual optimal solution.

  (b) There holds $f^* = q^*$, and (x^*, λ^*) is a primal and dual optimal solution pair if and only if

$$x^* \in \arg\min_{x\in\Re^n} \big\{f_1(x) - x'\lambda^*\big\}, \quad x^* \in \arg\min_{x\in\Re^n} \big\{f_2(x) + x'\lambda^*\big\}$$

Proof: For strong duality use the equality constrained problem

$$\begin{aligned} &\text{minimize} \quad f_1(x_1) + f_2(x_2) \\ &\text{subject to} \quad x_1 = x_2, \quad x_1 \in \mathrm{dom}(f_1),\ x_2 \in \mathrm{dom}(f_2) \end{aligned}$$

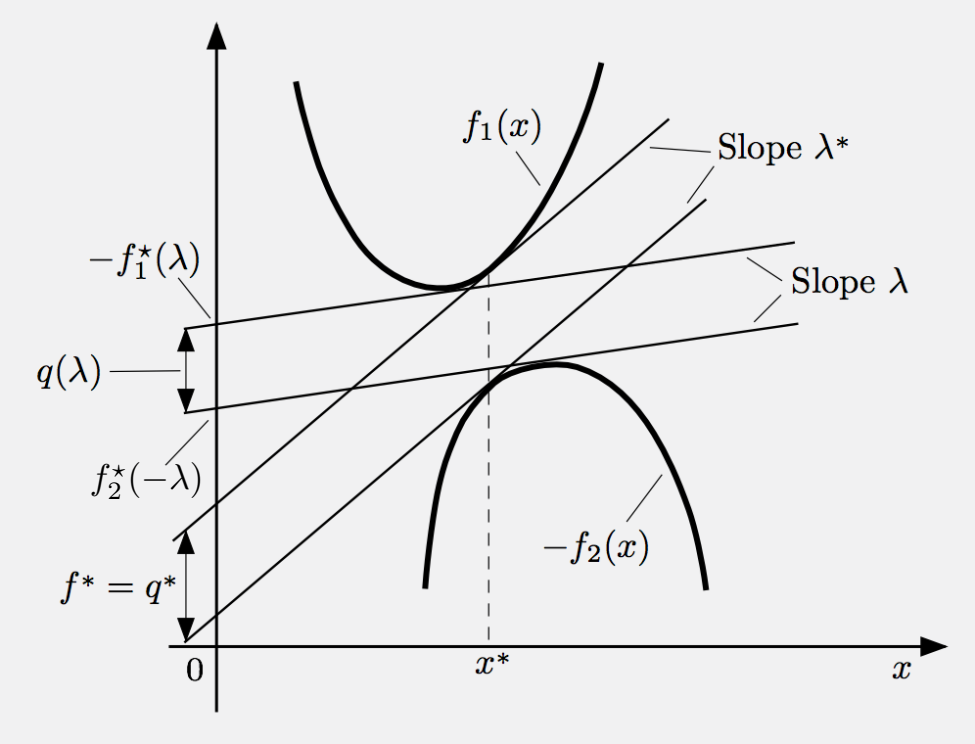
and the fact

$$\mathrm{ri}\big(\mathrm{dom}(f_1)\times\mathrm{dom}(f_2)\big) = \mathrm{ri}\big(\mathrm{dom}(f_1)\big)\times\big(\mathrm{dom}(f_2)\big)$$

to satisfy the relative interior condition.

For part (b), apply the optimality conditions (primal and dual feasibility, and Lagrangian optimality).

GEOMETRIC INTERPRETATION



• When $\mathrm{dom}(f_1) = \mathrm{dom}(f_2) = \Re^n$, and f_1 and f_2 are differentiable, the optimality condition is equivalent to

$$\lambda^* = \nabla f_1(x^*) = -\nabla f_2(x^*)$$

• By reversing the roles of the (symmetric) primal and dual problems, we obtain alternative criteria for strong duality: if q^* is finite and $\mathrm{ri}\big(\mathrm{dom}(f_1^\star)\big) \cap \mathrm{ri}\big(-\mathrm{dom}(f_2^\star)\big) \neq \emptyset$, then $f^* = q^*$ and there exists at least one primal optimal solution.

CONIC DUALITY

- Consider minimizing $f(x)$ over $x \in C$, where $f : \Re^n \mapsto (-\infty, \infty]$ is a closed proper convex function and C is a closed convex cone in $\Re^n$.
- We apply Fenchel duality with the definitions

$$f_1(x) = f(x), \qquad f_2(x) = \begin{cases} 0 & \text{if } x \in C, \\ \infty & \text{if } x \notin C. \end{cases}$$

The conjugates are

$$f_1^\star(\lambda) = \sup_{x \in \Re^n} \left\{ \lambda' x - f(x) \right\}, \; f_2^\star(\lambda) = \sup_{x \in C} \lambda' x = \begin{cases} 0 & \text{if } \lambda \in C^*, \\ \infty & \text{if } \lambda \notin C^*, \end{cases}$$

where $C^* = \{\lambda \mid \lambda' x \leq 0, \; \forall \; x \in C\}$.

- The dual problem is

$$\begin{aligned} &\text{minimize} \quad f^\star(\lambda) \\ &\text{subject to} \quad \lambda \in \hat{C}, \end{aligned}$$

where $f^\star$ is the conjugate of f and

$$\hat{C} = \{\lambda \mid \lambda' x \geq 0, \; \forall \; x \in C\}.$$

$\hat{C}$ is called the dual cone. ($-\hat{C}$ is the polar cone.)

CONIC DUALITY THEOREM

- Assume that the optimal value of the primal conic problem is finite, and that

$$\mathrm{ri}\big(\mathrm{dom}(f)\big) \cap \mathrm{ri}(C) \neq \emptyset.$$

Then, there is no duality gap and the dual problem has an optimal solution.

- Using the symmetry of the primal and dual problems, we also obtain that there is no duality gap and the primal problem has an optimal solution if the optimal value of the dual conic problem is finite, and

$$\mathrm{ri}\big(\mathrm{dom}(f^\star)\big) \cap \mathrm{ri}(\hat{C}) \neq \emptyset.$$

LECTURE 12

LECTURE OUTLINE

- We transition from theory to algorithms
- The next two lectures provide:
  - An overview of interesting/challenging large-scale convex problem structures
  - An overview of fundamental algorithmic ideas for large-scale convex programming
- Problem Structures
  - Separable problems
  - Integer/discrete problems – Branch-and-bound
  - Large sum problems
  - Problems with many constraints
- Conic Programming
  - Second Order Cone Programming
  - Semidefinite Programming

SEPARABLE PROBLEMS

- Consider the problem

$$\text{minimize} \quad \sum_{i=1}^{m} f_i(x_i)$$

$$\text{s. t.} \quad \sum_{i=1}^{m} g_{ji}(x_i) \leq 0, \quad j = 1, \ldots, r, \quad x_i \in X_i, \quad \forall\, i$$

where $f_i : \Re^{n_i} \mapsto \Re$ and $g_{ji} : \Re^{n_i} \mapsto \Re$ are given functions, and X_i are given subsets of $\Re^{n_i}$.

- Form the dual problem

$$\text{maximize} \quad \sum_{i=1}^{m} q_i(\mu) \equiv \sum_{i=1}^{m} \inf_{x_i \in X_i} \left\{ f_i(x_i) + \sum_{j=1}^{r} \mu_j g_{ji}(x_i) \right\}$$

$$\text{subject to} \quad \mu \geq 0$$

- Important point: The calculation of the dual function has been decomposed into m simpler minimizations.

- Another important point: If X_i is a discrete set (e.g., $X_i = \{0, 1\}$), the dual optimal value is a lower bound to the optimal primal value. It is still useful in a branch-and-bound scheme.

LARGE SUM PROBLEMS

- Consider cost function of the form

$$f(x) = \sum_{i=1}^{m} f_i(x), \qquad m \text{ is very large}$$

- Dual cost of a separable problem.

- Data analysis/machine learning. x is parameter vector of a model; each f_i corresponds to error between data and output of the model.
  - Least squares problems (f_i quadratic).
  - ℓ_1-regularization (least squares plus ℓ_1 penalty):

$$\min_x \sum_{j=1}^{m} (a_j' x - b_j)^2 + \gamma \sum_{i=1}^{n} |x_i|$$

    The nondifferentiable penalty tends to set a large number of components of x to 0.
  - Maximum likelihood estimation.

- Min of an expected value $E\{F(x,w)\}$, where w is a random variable taking a finite but very large number of values w_i, $i = 1, \ldots, m$, with corresponding probabilities π_i. A special case: Stochastic programming.

- Special type of algorithms, called incremental apply (they operate on a single f_i at a time).

PROBLEMS WITH MANY CONSTRAINTS

- Problems of the form

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad a_j' x \leq b_j, \quad j = 1, \ldots, r, \end{aligned}$$

where r: very large.

- One possibility is a penalty function approach: Replace problem with

$$\min_{x \in \Re^n} f(x) + c \sum_{j=1}^{r} P(a_j' x - b_j)$$

where $P(\cdot)$ is a scalar penalty function satisfying $P(t) = 0$ if $t \leq 0$, and $P(t) > 0$ if $t > 0$, and c is a positive penalty parameter.

- Examples:
  - The quadratic penalty $P(t) = \big(\max\{0, t\}\big)^2$.
  - The nondifferentiable penalty $P(t) = \max\{0, t\}$.

- Another possibility: Initially discard some of the constraints, solve a less constrained problem, and later reintroduce constraints that seem to be violated at the optimum (outer approximation).

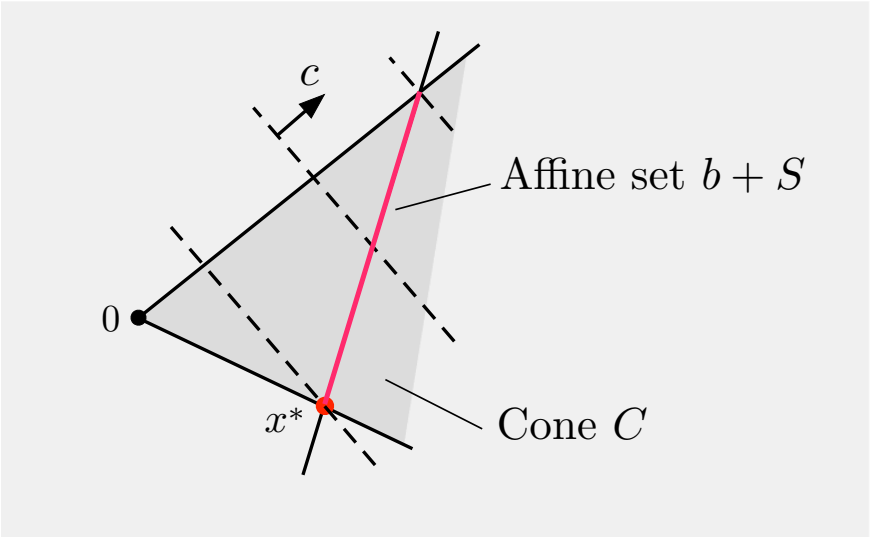
- Also inner approximation of the constraint set.

CONIC PROBLEMS

- A conic problem is to minimize a convex function $f : \Re^n \mapsto (-\infty, \infty]$ subject to a cone constraint.

- The most useful/popular special cases:
  - Linear-conic programming
  - Second order cone programming
  - Semidefinite programming

involve minimization of linear function $c'x$ over intersection of an affine set $b + S$ and a cone C.



- Can be analyzed as a special case of Fenchel duality.

- There are many interesting applications of conic problems, including in discrete optimization.

PROBLEM RANKING IN INCREASING PRACTICAL DIFFICULTY

- Linear and (convex) quadratic programming.
  - Favorable special cases (e.g., network flows).
- Second order cone programming.
- Semidefinite programming.
- Convex programming.
  - Favorable special cases (e.g., network flows, monotropic programming, geometric programming).
- Nonlinear/nonconvex/continuous programming.
  - Favorable special cases (e.g., twice differentiable, quasi-convex programming).
  - Unconstrained.
  - Constrained.
- Discrete optimization/Integer programming.
  - Favorable special cases.

CONIC DUALITY

- Consider minimizing $f(x)$ over $x \in C$, where $f : \Re^n \mapsto (-\infty, \infty]$ is a closed proper convex function and C is a closed convex cone in $\Re^n$.

- We apply Fenchel duality with the definitions

$$f_1(x) = f(x), \qquad f_2(x) = \begin{cases} 0 & \text{if } x \in C, \\ \infty & \text{if } x \notin C. \end{cases}$$

The conjugates are

$$f_1^\star(\lambda) = \sup_{x \in \Re^n} \left\{ \lambda' x - f(x) \right\}, \; f_2^\star(\lambda) = \sup_{x \in C} \lambda' x = \begin{cases} 0 & \text{if } \lambda \in C^*, \\ \infty & \text{if } \lambda \notin C^*, \end{cases}$$

where $C^* = \{\lambda \mid \lambda' x \leq 0, \, \forall \, x \in C\}$ is the polar cone of C.

- The dual problem is $\min_\lambda \left\{ f_1^\star(\lambda) + f_2^\star(-\lambda) \right\}$, or

$$\begin{aligned} &\text{minimize} \quad f^\star(\lambda) \\ &\text{subject to} \quad \lambda \in \hat{C}, \end{aligned}$$

where $f^\star$ is the conjugate of f and $\hat{C}$ is the dual cone ($= -C^*$, negative polar cone)

$$\hat{C} = \{\lambda \mid \lambda' x \geq 0, \, \forall \, x \in C\}$$

LINEAR-CONIC PROBLEMS

- Let f be affine, $f(x) = c'x$, with $\mathrm{dom}(f)$ being an affine set, $\mathrm{dom}(f) = b + S$, where S is a subspace.

- The primal problem is

$$\begin{aligned} &\text{minimize} \quad c'x \\ &\text{subject to} \quad x - b \in S, \quad x \in C. \end{aligned}$$

- The conjugate is

$$\begin{aligned} f^\star(\lambda) &= \sup_{x-b\in S} (\lambda - c)'x = \sup_{y\in S} (\lambda - c)'(y + b) \\ &= \begin{cases} (\lambda - c)'b & \text{if } \lambda - c \in S^\perp, \\ \infty & \text{if } \lambda - c \notin S^\perp, \end{cases} \end{aligned}$$

so the dual problem can be written as

$$\begin{aligned} &\text{minimize} \quad b'\lambda \\ &\text{subject to} \quad \lambda - c \in S^\perp, \quad \lambda \in \hat{C}. \end{aligned}$$

- The primal and dual have the same form.

- If C is closed, the dual of the dual yields the primal.

SPECIAL LINEAR-CONIC FORMS

$$\min_{Ax=b,\ x\in C} c'x \quad \Longleftrightarrow \quad \max_{c-A'\lambda\in\hat{C}} b'\lambda,$$

$$\min_{Ax-b\in C} c'x \quad \Longleftrightarrow \quad \max_{A'\lambda=c,\ \lambda\in\hat{C}} b'\lambda,$$

where $x \in \Re^n$, $\lambda \in \Re^m$, $c \in \Re^n$, $b \in \Re^m$, $A : m \times n$.

- Proof of first relation: Let $\overline{x}$ be such that $A\overline{x} = b$, and write the problem on the left as

$$\text{minimize} \quad c'x$$

$$\text{subject to} \quad x - \overline{x} \in \mathrm{N}(A), \quad x \in C$$

- The dual conic problem is

$$\text{minimize} \quad \overline{x}'\mu$$

$$\text{subject to} \quad \mu - c \in \mathrm{N}(A)^\perp, \quad \mu \in \hat{C}$$

- Using $\mathrm{N}(A)^\perp = \mathrm{Ra}(A')$, write the constraints as $c - \mu \in -\mathrm{Ra}(A') = \mathrm{Ra}(A')$, $\mu \in \hat{C}$, or

$$c - \mu = A'\lambda, \qquad \mu \in \hat{C}, \qquad \text{for some } \lambda \in \Re^m$$

- Change variables $\mu = c - A'\lambda$, write the dual as

$$\text{minimize} \quad \overline{x}'(c - A'\lambda)$$

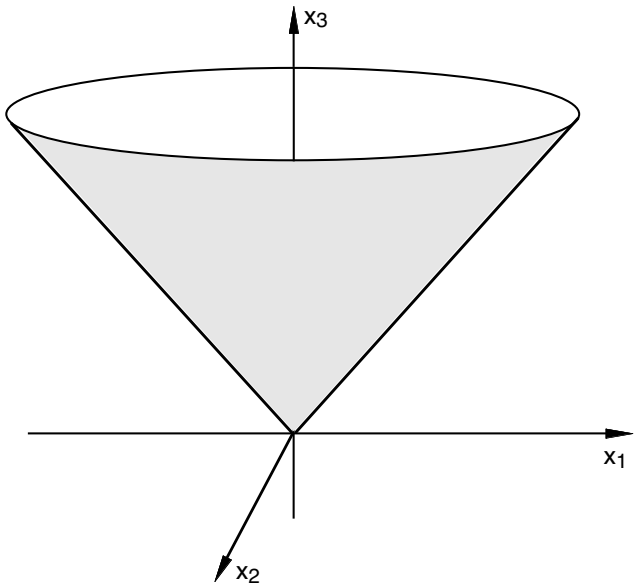
$$\text{subject to} \quad c - A'\lambda \in \hat{C}$$

discard the constant $\overline{x}'c$, use the fact $A\overline{x} = b$, and change from min to max.

SOME EXAMPLES

- Nonnegative Orthant: $C = \{x \mid x \geq 0\}$
- The Second Order Cone: Let

$$C = \left\{ (x_1, \ldots, x_n) \mid x_n \geq \sqrt{x_1^2 + \cdots + x_{n-1}^2} \right\}$$



- The Positive Semidefinite Cone: Consider the space of symmetric $n \times n$ matrices, viewed as the space $\Re^{n^2}$ with the inner product

$$< X, Y >= \text{trace}(XY) = \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} y_{ij}$$

Let C be the cone of matrices that are positive semidefinite.

- All these are self-dual, i.e., $C = -C^* = \hat{C}$.

SECOND ORDER CONE PROGRAMMING

- Second order cone programming is the linear-conic problem

$$\begin{aligned} &\text{minimize} \quad c'x \\ &\text{subject to} \quad A_i x - b_i \in C_i, \ i = 1, \ldots, m, \end{aligned}$$

where c, b_i are vectors, A_i are matrices, b_i is a vector in $\Re^{n_i}$, and

$$C_i : \text{ the second order cone of } \Re^{n_i}$$

- The cone here is

$$C = C_1 \times \cdots \times C_m$$

and the constraints $A_i x - b_i \in C_i, \ i = 1, \ldots, m,$ can be lumped into a single constraint

$$Ax - b \in C$$

SECOND ORDER CONE DUALITY

- Using the generic duality form

$$\min_{Ax-b\in C} c'x \qquad \Longleftrightarrow \qquad \max_{A'\lambda=c,\ \lambda\in\hat{C}} b'\lambda,$$

and self duality of C, the dual problem is

$$\text{maximize} \quad \sum_{i=1}^{m} b_i'\lambda_i$$

$$\text{subject to} \quad \sum_{i=1}^{m} A_i'\lambda_i = c, \quad \lambda_i \in C_i,\ i=1,\ldots,m,$$

where $\lambda = (\lambda_1, \ldots, \lambda_m)$.

- The duality theory is no more favorable than the one for linear-conic problems.

- There is no duality gap if there exists a feasible solution in the interior of the 2nd order cones C_i.

- Generally, 2nd order cone problems can be recognized from the presence of norm or convex quadratic functions in the cost or the constraint functions.

- There are many applications.

EXAMPLE: ROBUST LINEAR PROGRAMMING

$$\begin{aligned}&\text{minimize} \quad c'x\\ &\text{subject to} \quad a_j'x \le b_j, \quad \forall\ (a_j,b_j) \in T_j, \quad j=1,\ldots,r,\end{aligned}$$

where $c \in \Re^n$, and T_j is a given subset of $\Re^{n+1}$.

- We convert the problem to the equivalent form

$$\begin{aligned}&\text{minimize} \quad c'x\\ &\text{subject to} \quad g_j(x) \le 0, \qquad j=1,\ldots,r,\end{aligned}$$

where $g_j(x) = \sup_{(a_j,b_j)\in T_j}\{a_j'x - b_j\}$.

- For the special choice where T_j is an ellipsoid,

$$T_j = \left\{(\overline{a}_j + P_j u_j, \overline{b}_j + q_j'u_j) \mid \|u_j\| \le 1,\ u_j \in \Re^{n_j}\right\}$$

we can express $g_j(x) \le 0$ in terms of a SOC:

$$\begin{aligned}g_j(x) &= \sup_{\|u_j\|\le 1} \left\{(\overline{a}_j + P_j u_j)'x - (\overline{b}_j + q_j'u_j)\right\}\\ &= \sup_{\|u_j\|\le 1} (P_j'x - q_j)'u_j + \overline{a}_j'x - \overline{b}_j,\\ &= \|P_j'x - q_j\| + \overline{a}_j'x - \overline{b}_j.\end{aligned}$$

Thus, $g_j(x) \le 0$ iff $(P_j'x - q_j, \overline{b}_j - \overline{a}_j'x) \in C_j$, where C_j is the SOC of $\Re^{n_j+1}$.

LECTURE 13

LECTURE OUTLINE

- A taxonomy of algorithms for convex optimization
  - Iterative descent
  - Approximation
- A brief overview of approximation algorithms
- Focus on cost function descent
  - Gradient and subgradient methods
  - Gradient projection
  - Newton's method
- Incremental methods

APPROXIMATION

- Problem: Minimize convex $f : \Re^n \mapsto \Re$ over a closed convex set X.
- Approximation approach: Generate $\{x_k\}$ by

$$x_{k+1} \in \arg\min_{x \in X_k} F_k(x),$$

where:

F_k is a function that approximates f

X_k is a set that approximates X

- F_k and X_k may depend on the prior iterates $x_0, \ldots, x_k$, and other parameters.
- Key ideas:
  - Minimization of F_k over X_k should be easier than minimization of f over X
  - x_k should be a good starting point for obtaining x_{k+1}
  - Approximation of f by F_k and/or X by X_k should improve as k increases
- Major types of approximation algorithms:
  - Polyhedral approximation
  - Penalty, proximal, interior point methods
  - Smoothing

ITERATIVE DESCENT

- Generate $\{x_k\}$ such that

$$\phi(x_{k+1}) < \phi(x_k) \qquad \text{iff } x_k \text{ is not optimal}$$

- ϕ is a merit function (also called Lyapounov function)
  - Measures progress towards optimality
  - Is minimized only at optimal points, i.e.,

$$\arg\min_{x \in X} \phi(x) = \arg\min_{x \in X} f(x)$$

- Examples:

$$\phi(x) = f(x), \qquad \phi(x) = \inf_{x^* : \text{optimal}} \|x - x^*\|$$

- In some cases, iterative descent may be the primary idea, but modifications or approximations are introduced:
  - To make the method tolerant of random or nonrandom errors.
  - To make the method suitable for distributed asynchronous computation.

FOCUS ON COST FUNCTION DESCENT

- Consider the unconstrained problem: Minimize $f : \Re^n \mapsto \Re$ over $x \in \Re^n$.
- Generate $\{x_k\}$ by

$$x_{k+1} = x_k + \alpha_k d_k, \qquad k = 0, 1, \ldots$$

where d_k is a descent direction at x_k, i.e.,

$$f(x_k + \alpha d_k) < f(x_k), \qquad \forall \alpha \in (0, \bar{\alpha}]$$

- Many ways to choose the stepsize α_k.
- Sometimes a descent direction is used but the descent condition $f(x_k + \alpha_k d_k) < f(x_k)$ may not be strictly enforced in all iterations.
- Cost function descent is used primarily for differentiable f, with

$$d_k = -S_k \nabla f(x_k)$$

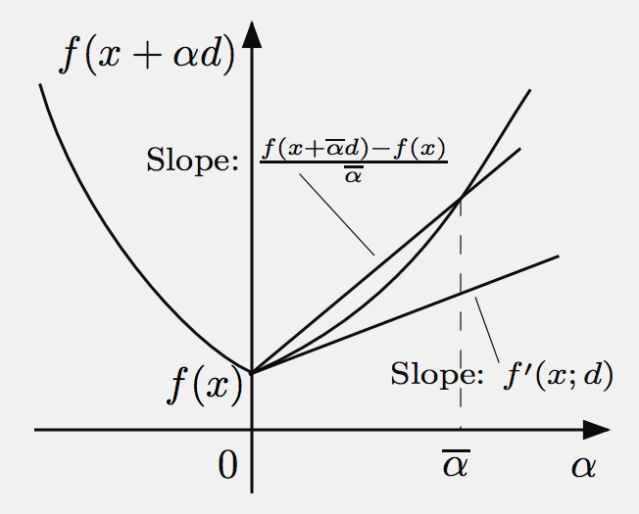
where S_k is positive definite (scaling) matrix.

- Encounters serious theoretical difficulties for nondifferentiable f.

DIRECTIONAL DERIVATIVES

- Directional derivative of a proper convex f:

$$f'(x;d) = \lim_{\alpha \downarrow 0} \frac{f(x+\alpha d) - f(x)}{\alpha}, \quad x \in \text{dom}(f), \; d \in \Re^n$$



- The ratio $\frac{f(x+\alpha d)-f(x)}{\alpha}$ is monotonically nonincreasing as $\alpha \downarrow 0$ and converges to $f'(x;d)$.

- d is a descent direction at x, i.e.,

$f(x+\alpha d) < f(x)$, for all $\alpha > 0$ sufficiently small

iff $f'(x;d) < 0$.

- If f is differentiable, $f'(x;d) = \nabla f(x)'d$, so if S is positive definite, $d = -S\nabla f(x)$ is a descent direction.

MANY ALGORITHMS BASED ON GRADIENT

- Consider unconstrained minimization of differentiable $f : \Re^n \mapsto \Re$ by

$$x_{k+1} = x_k - \alpha_k S_k \nabla f(x_k), \qquad k = 0, 1, \ldots$$

- Gradient or steepest descent method: $S_k = I$.
- Newton's method (fast local convergence):

$$S_k = \big(\nabla^2 f(x_k)\big)^{-1}$$

assuming $\nabla^2 f(x_k)$ is positive definite (otherwise modifications are needed).

- Many algorithms try to emulate Newton's method with less overhead (quasi-Newton, Gauss-Newton method, limited memory, conjugate direction, etc).
- Diagonal scaling: Choose S_k diagonal with inverse 2nd derivatives of f along the diagonal.
- Common stepsize rules:
  - Constant: $\alpha_k \equiv \alpha$
  - Diminishing: $\sum_{k=0}^{\infty} \alpha_k = \infty, \ \alpha_k \downarrow 0$
  - Minimization: $\alpha_k \in \arg\min_{\alpha>0} f(x + \alpha d)$

FAILURE FOR NONDIFFERENTIABLE COST

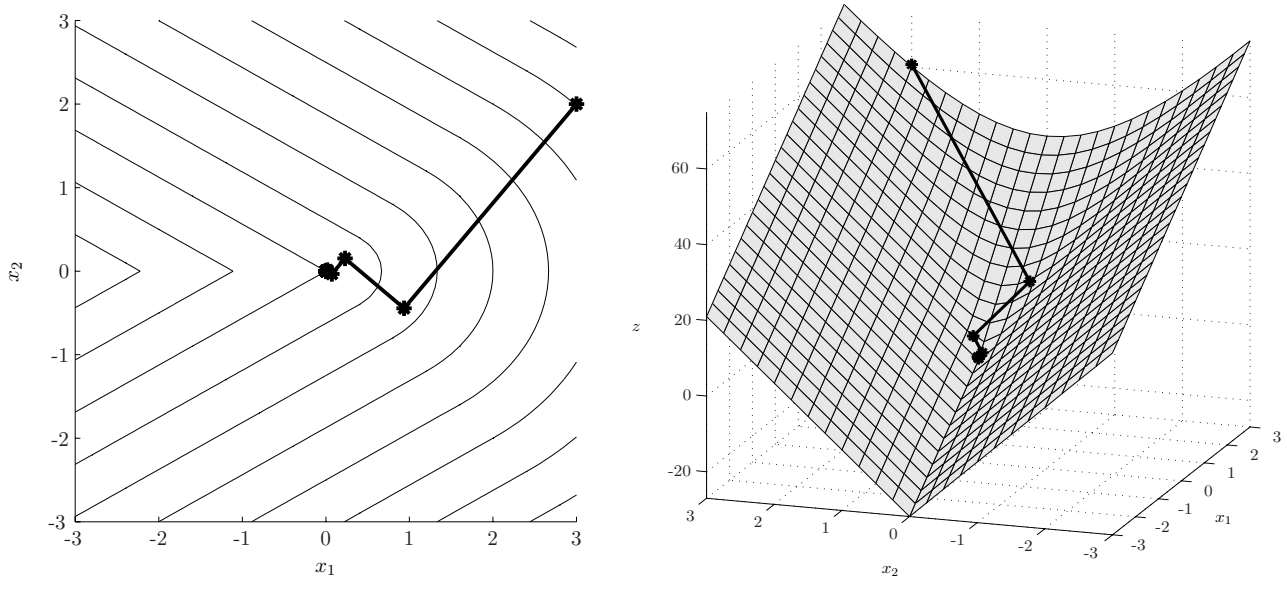
- Start with any $x_0 \in \Re^n$.
- Calculate d_k as the steepest descent direction at x_k

$$d_k = \arg \min_{\|d\|=1} f'(x_k; d)$$

and set

$$x_{k+1} = x_k + \alpha_k d_k$$

- Serious difficulties:
  - Computing d_k is nontrivial at points x_k where f is nondifferentiable.
  - Serious convergence issues due to discontinuity of steepest descent direction.
- Example with α_k determined by minimization along d_k: $\{x_k\}$ converges to nonoptimal point.



CONSTRAINED CASE: GRADIENT PROJECTION

- Problem: Minimization of differentiable $f : \Re^n \mapsto \Re$ over a closed convex set X.
- Cost function descent

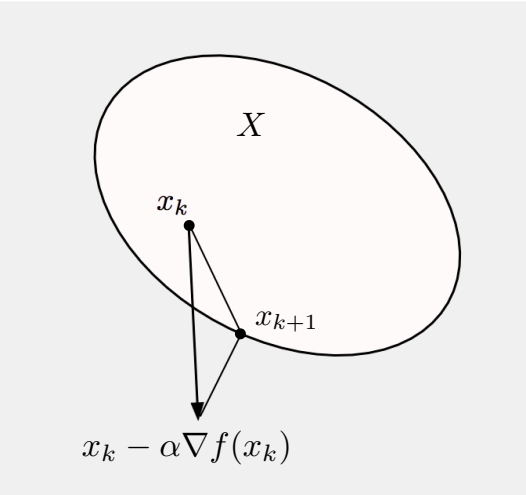
$$x_{k+1} = x_k + \alpha_k d_k$$

where d_k is a feasible descent direction at x_k: $x_k + \alpha d_k$ must belong to X for small enough $a > 0$.

- The gradient projection method:

$$x_{k+1} = P_X\big(x_k - \alpha_k \nabla f(x_k)\big)$$

where $\alpha_k > 0$ is a stepsize and $P_X(\cdot)$ denotes projection on X.



- Projection may be costly. Scaling is tricky.

SUBGRADIENT PROJECTION

- Problem: Minimization of nondifferentiable convex $f : \Re^n \mapsto \Re$ over a closed convex set X.

- Key notion: A subgradient of a convex function $f : \Re^n \mapsto \Re$ at a point x is a vector g such that

$$f(z) \geq f(x) + g'(z - x), \qquad \forall\ z \in \Re^n.$$

At points x where f is differentiable, $\nabla f(x)$ is the unique subgradient.

- Subgradient projection method:

$$x_{k+1} = P_X(x_k - \alpha_k g_k)$$

where g_k is an arbitrary subgradient at x_k.

- Does not attain cost function descent ... but has another descent property: at any nonoptimal point x_k, it satisfies for $a_k > 0$ small enough,

$$\text{dist}(x_{k+1}, X^*) < \text{dist}(x_k, X^*)$$

where X^* is the optimal solution set.

- Typically, a diminishing stepsize α_k is needed.

INCREMENTAL GRADIENT METHOD

- Problem: Minimization of $f(x) = \sum_{i=1}^{m} f_i(x)$ over a closed convex set X (f_i differentiable).

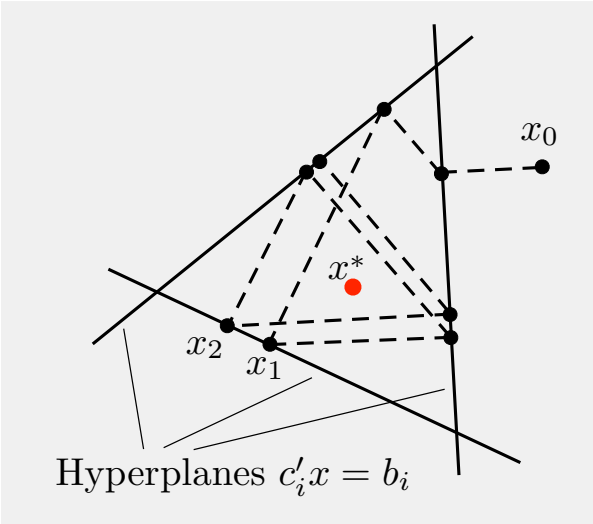
- Operates in cycles: If x_k is the vector obtained after k cycles, the vector x_{k+1} obtained after one more cycle is $x_{k+1} = \psi_{m,k}$, where $\psi_{0,k} = x_k$, and

$$\psi_{i,k} = P_X\big(\psi_{i-1,k} - \alpha_k \nabla f_{i,k}(\psi_{i-1,k})\big), \qquad i = 1, \ldots, m$$

- Example: The Kazcmarz method

$$\psi_{i,k} = \psi_{i-1,k} - \frac{1}{\|c_i\|^2}(c_i'\psi_{i-1,k} - b_i)c_i, \quad i = 1, \ldots, m,$$

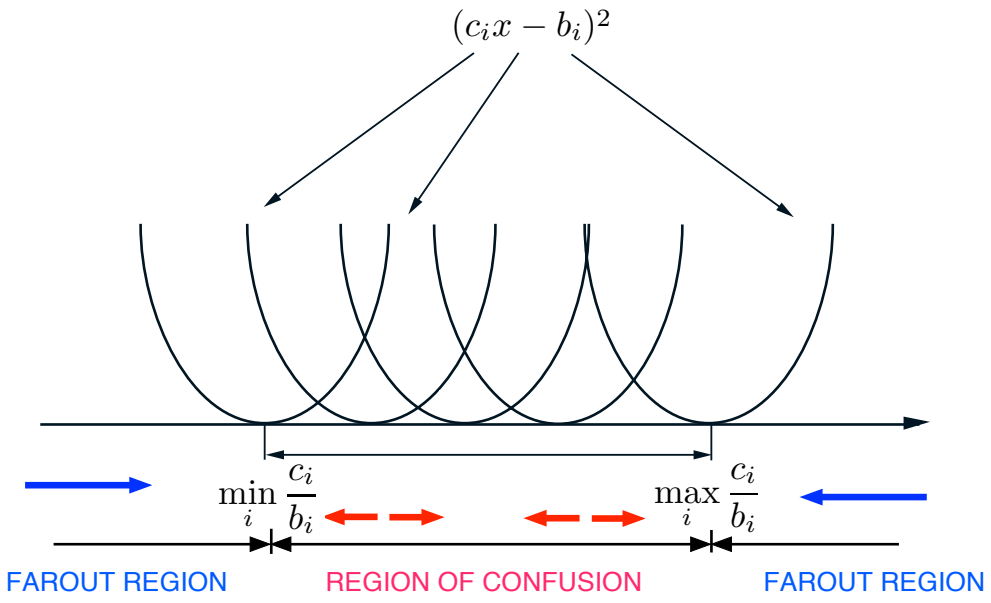
for the case $f_i(x) = \frac{1}{2\|c_i\|^2}(c_i'x - b_i)^2$



COMPARE W/ NONINCREMENTAL GRADIENT

- Two complementary performance issues:
  - Progress when far from convergence. Here the incremental method can be much faster.
  - Progress when close to convergence. Here the incremental method can be inferior.
- Example: Scalar case

$$f_i(x) = \tfrac{1}{2}(c_i x - b_i)^2, \qquad x \in \Re$$



- A diminishing stepsize is necessary for convergence (otherwise the method ends up oscillating within the region of confusion).
- Randomization of selection of component f_i is possible.

OTHER INCREMENTAL METHODS

- Aggregated gradient method:

$$x_{k+1} = P_X \left(x_k - \alpha_k \sum_{\ell=0}^{m-1} \nabla f_{i_{k-\ell}}(x_{k-\ell}) \right)$$

- Gradient method with momentum (heavy ball method):

$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k) + \beta_k (x_k - x_{k-1})$$

- Stochastic gradient method for $f(x) = E\{F(x,w)\}$ where w is a random variable, and $F(\cdot, w)$ is a convex function for each value of w:

$$x_{k+1} = P_X \big(x_k - \alpha_k \nabla F(x_k, w_k) \big)$$

where $\nabla F(x_k, w_k)$ is a "sampled" gradient.

- Incremental Newton method.

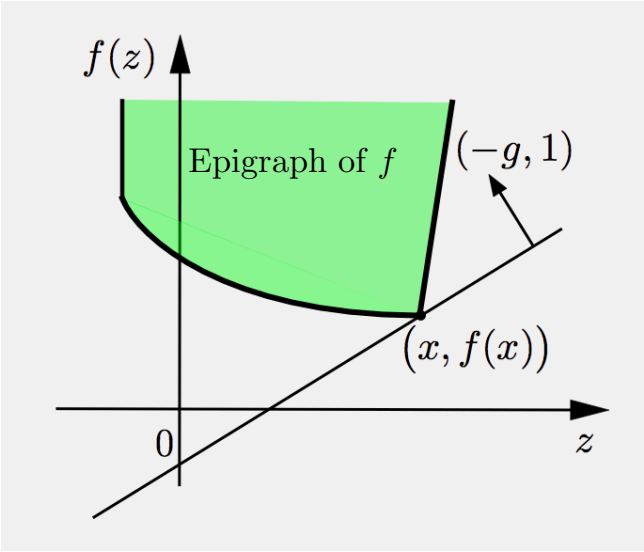
- Incremental Gauss-Newton method for least squares (extended Kalman filter).

LECTURE 14

LECTURE OUTLINE

- Subgradients of convex functions
- Sugradients of real-valued convex functions
- Properties of subgradients
- Computation of subgradients
- Reading:
  - Section 5.4 of Convex Optimization Theory (focus on extended real-valued convex functions)
  - Section 2.1 of Convex Optimization Algorithms (focus on real-valued convex functions)

SUBGRADIENTS



- Let $f : \Re^n \mapsto (-\infty, \infty]$ be a convex function. A vector $g \in \Re^n$ is a subgradient of f at a point $x \in \text{dom}(f)$ if

$$f(z) \geq f(x) + (z - x)'g, \qquad \forall\ z \in \Re^n$$

- Support Hyperplane Interpretation: g is a subgradient if and only if

$$f(z) - z'g \geq f(x) - x'g, \qquad \forall\ z \in \Re^n$$

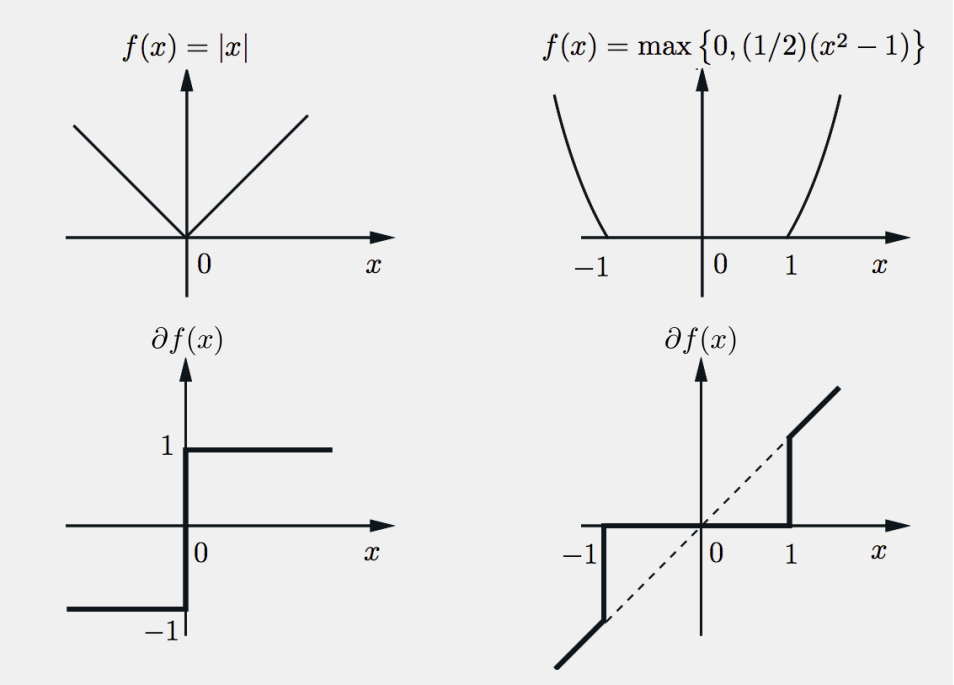
so g is a subgradient at x if and only if the hyperplane in $\Re^{n+1}$ that has normal $(-g, 1)$ and passes through $\big(x, f(x)\big)$ supports the epigraph of f.

- The set of all subgradients at x is the subdifferential of f at x, denoted $\partial f(x)$.

- x^* minimizes f if and only if $0 \in \partial f(x^*)$.

EXAMPLES OF SUBDIFFERENTIALS

- Some examples:



- If f is differentiable, then $\partial f(x) = \{\nabla f(x)\}$.

Proof: Clearly $\nabla f(x) \in \partial f(x)$. Conversely, if $g \in \partial f(x)$, then for all $\alpha \in \Re$ and $d \in \Re^n$,

$$\alpha g'd \leq f(x + \alpha d) - f(x) = \alpha \nabla f(x)'d + o(|\alpha|).$$

Let $d = \nabla f(x) - g$ to obtain

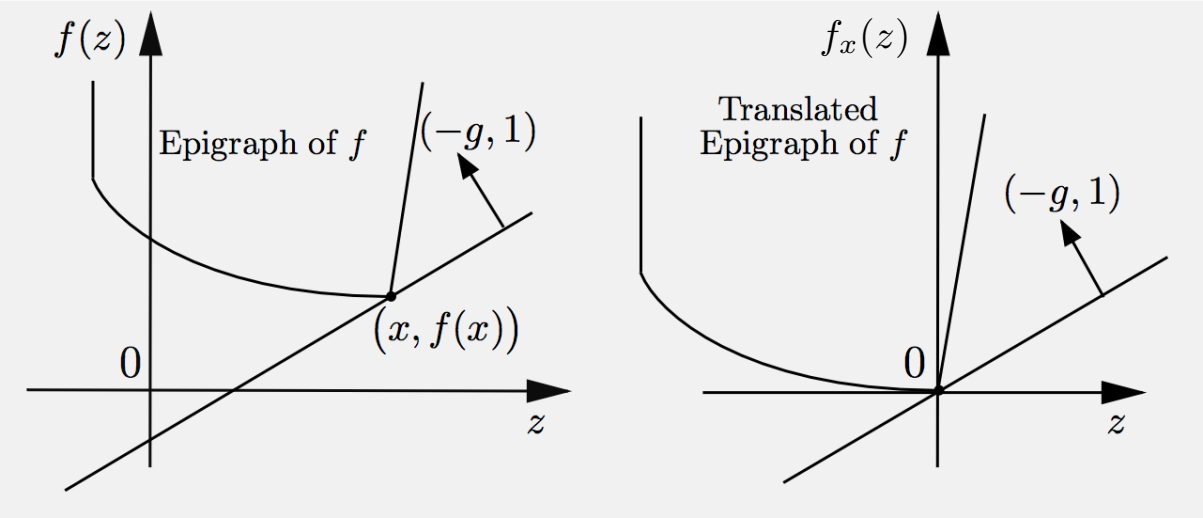
$$\|\nabla f(x) - g\|^2 \leq -o(|\alpha|)/\alpha, \qquad \forall\ \alpha < 0$$

Take $\alpha \uparrow 0$ to obtain $g = \nabla f(x)$.

EXISTENCE OF SUBGRADIENTS

- Let $f : \Re^n \mapsto (-\infty, \infty]$ be proper convex.
- Consider MC/MC with

$$M = \text{epi}(f_x), \qquad f_x(z) = f(x+z) - f(x)$$



- By 2nd MC/MC Duality Theorem, $\partial f(x)$ is nonempty if $x \in \text{ri}\big(\text{dom}(f)\big)$.
- If f is real-valued, $\partial f(x)$ is nonempty for all x
- For $x \notin \text{ri}\big(\text{dom}(f)\big)$, $\partial f(x)$ may be empty.

SUBGRADIENTS OF REAL-VALUED FUNCTIONS

- Let $f : \Re^n \mapsto \Re$ be a real-valued convex function, and let $X \subset \Re^n$ be compact.

  (a) The set $\cup_{x \in X} \partial f(x)$ is bounded.

  (b) f is Lipschitz over X, i.e., for all $x, z \in X$,

$$\big|f(x) - f(z)\big| \leq L \, \|x - z\|, \quad L = \sup_{g \in \cup_{x \in X} \partial f(x)} \|g\|.$$

Proof: (a) Assume the contrary, so there exist $\{x_k\} \subset X$, and unbounded $\{g_k\}$ with

$$g_k \in \partial f(x_k), \qquad 0 < \|g_k\| < \|g_{k+1}\|, \quad k = 0, 1, \ldots.$$

Let $d_k = g_k / \|g_k\|$. Since $g_k \in \partial f(x_k)$, we have

$$f(x_k + d_k) - f(x_k) \geq g_k' d_k = \|g_k\|$$

Since $\{x_k\}$ and $\{d_k\}$ are bounded, we assume they converge to some vectors. By continuity of f, the left-hand side is bounded, contradicting the unboundedness of $\{g_k\}$.

(b) If $g \in \partial f(x)$, then for all $x, z \in X$,

$$f(x) - f(z) \leq g'(x - z) \leq \|g\| \cdot \|x - z\| \leq L \, \|x - z\|$$

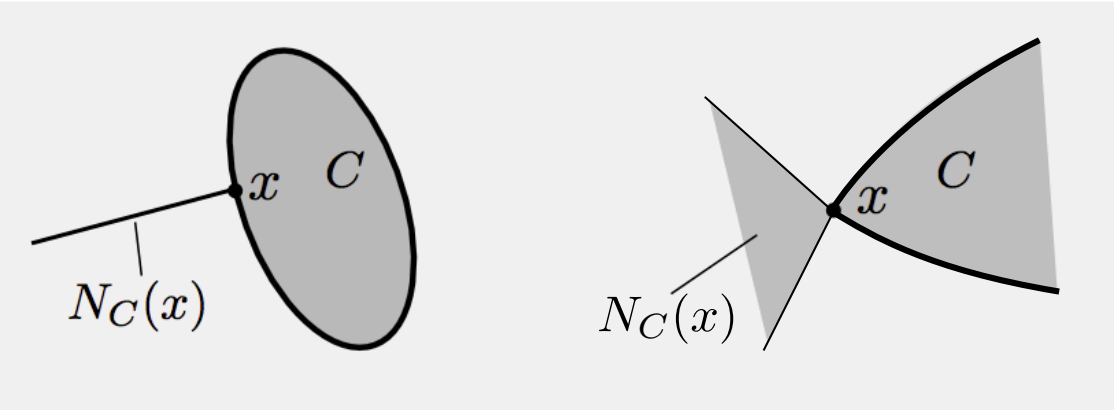
EXAMPLE: SUBDIFFERENTIAL OF INDICATOR

- Let C be a convex set, and δ_C be its indicator function.
- For $x \in C$, we have $g \in \partial\delta_C(x)$ iff

$$\delta_C(x) + g'(z - x) \leq \delta_C(z), \qquad \forall\ z \in C,$$

or equivalently $g'(z - x) \leq 0$ for all $z \in C$. Thus $\partial\delta_C(x)$ is the normal cone of C at x:

$$N_C(x) = \big\{g \mid g'(z - x) \leq 0,\ \forall\ z \in C\big\}.$$



CALCULUS OF SUBDIFFERENTIALS

• Chain Rule: Let $f : \Re^m \mapsto (-\infty, \infty]$ be convex, and A be a matrix. Consider $F(x) = f(Ax)$ and assume that F is proper. If

$$\text{Range}(A) \cap \text{ri}(\text{dom}(f)) \neq \emptyset,$$

then

$$\partial F(x) = A' \partial f(Ax), \qquad \forall\ x \in \Re^n.$$

• Subdifferential of a Sum: Let $f_i : \Re^n \mapsto (-\infty, \infty]$, $i = 1, \ldots, m$, be proper convex functions, and let

$$F = f_1 + \cdots + f_m.$$

Assume that $\cap_{1=1}^m \text{ri}\big(\text{dom}(f_i)\big) \neq \emptyset$. Then

$$\partial F(x) = \partial f_1(x) + \cdots + \partial f_m(x), \qquad \forall\ x \in \Re^n.$$

• Relative interior condition is needed as simple examples show.

• The relative interior conditions are automatically satisfied if the functions are real-valued.

• The relative interior conditions are unnecessary if the functions are polyhedral.

CONSTRAINED OPTIMALITY CONDITION

• Let $f : \Re^n \mapsto \Re$ and $X \subset \Re^n$ be convex. Then, a vector x^* minimizes f over X iff there exists $g \in \partial f(x^*)$ such that $-g$ belongs to the normal cone $N_X(x^*)$, i.e.,

$$g'(x - x^*) \geq 0, \qquad \forall\ x \in X.$$

Proof: x^* minimizes

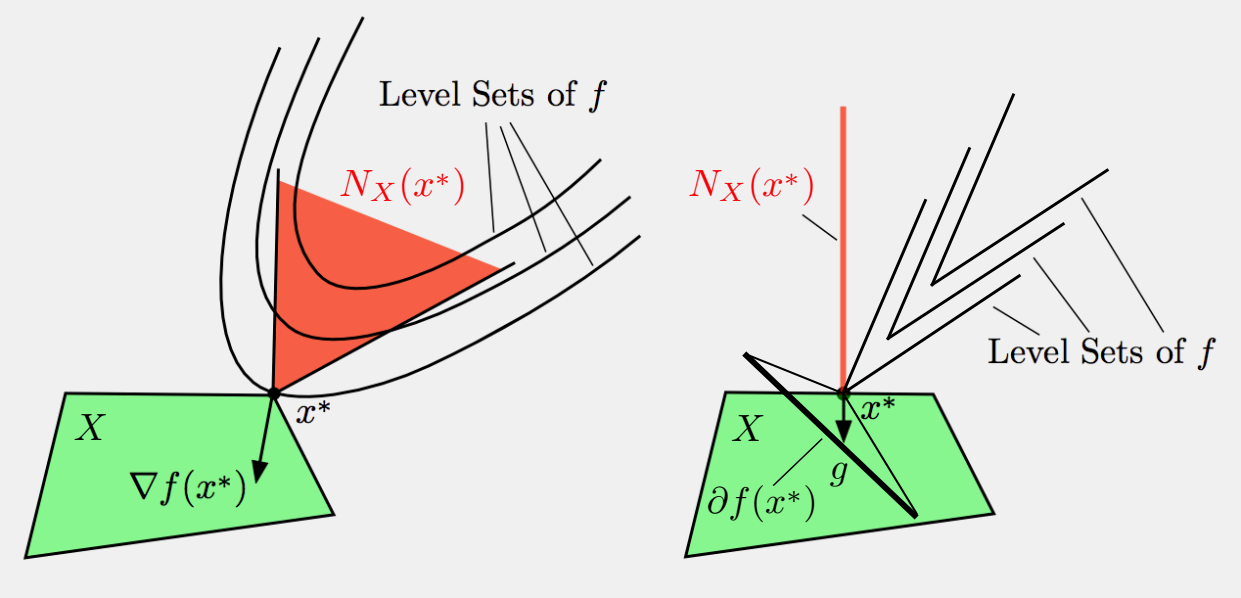
$$F(x) = f(x) + \delta_X(x)$$

if and only if $0 \in \partial F(x^*)$. Use the formula for subdifferential of sum to write

$$0 \in \partial F(x^*) = \partial f(x^*) + N_X(x^*)$$

Q.E.D.

ILLUSTRATION OF OPTIMALITY CONDITION



- In the figure on the left, f is differentiable and the optimality condition is

$$-\nabla f(x^*) \in N_X(x^*),$$

which is equivalent to

$$\nabla f(x^*)'(x - x^*) \geq 0, \qquad \forall\ x \in X.$$

- In the figure on the right, f is nondifferentiable, and the optimality condition is

$$-g \in N_X(x^*) \quad \text{for some } g \in \partial f(x^*).$$

DANSKIN'S THEOREM FOR MAX FUNCTIONS

- Let
$$f(x) = \max_{z \in Z} \phi(x, z),$$
where $x \in \Re^n$, $z \in \Re^m$, $\phi : \Re^n \times \Re^m \mapsto \Re$ is a function, Z is a compact subset of $\Re^m$, $\phi(\cdot, z)$ is convex and differentiable for each $z \in Z$, and $\nabla_x \phi(x, \cdot)$ is continuous on Z for each x. Then

$$\partial f(x) = \text{conv}\big\{\nabla_x \phi(x, z) \mid z \in Z(x)\big\}, \qquad x \in \Re^n,$$

where $Z(x)$ is the set of maximizing points

$$Z(x) = \left\{ \overline{z} \;\middle|\; \phi(x, \overline{z}) = \max_{z \in Z} \phi(x, z) \right\}$$

- Special case: $f(x) = \max\big\{\phi_1(x), \ldots, \phi_m(x)\big\}$ where ϕ_i are differentiable convex. Then

$$\partial f(x) = \text{conv}\big\{\nabla \phi_i(x) \mid i \in I(x)\big\},$$

where
$$I(x) = \big\{i \mid \phi_i(x) = f(x)\big\}$$

IMPORTANT ALGORITHMIC POINT

- Computing a single subgradient is often much easier than computing the entire subdifferential.
- Special case of dual functions: Consider

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in X, \quad g(x) \leq 0, \end{aligned}$$

where $f : \Re^n \mapsto \Re$, $g : \Re^n \mapsto \Re^r$, $X \subset \Re^n$. Consider the dual problem $\max_{\mu \geq 0} q(\mu)$, where

$$q(\mu) = \inf_{x \in X} \big\{ f(x) + \mu' g(x) \big\}.$$

For a given $\mu \geq 0$, suppose that x_μ minimizes the Lagrangian over $x \in X$,

$$x_\mu \in \arg\min_{x \in X} \big\{ f(x) + \mu' g(x) \big\}.$$

Then $-g(x_\mu)$ is a subgradient of the negative of the dual function $-q$ at μ.

- Verification: For all $\nu \in \Re^r$,

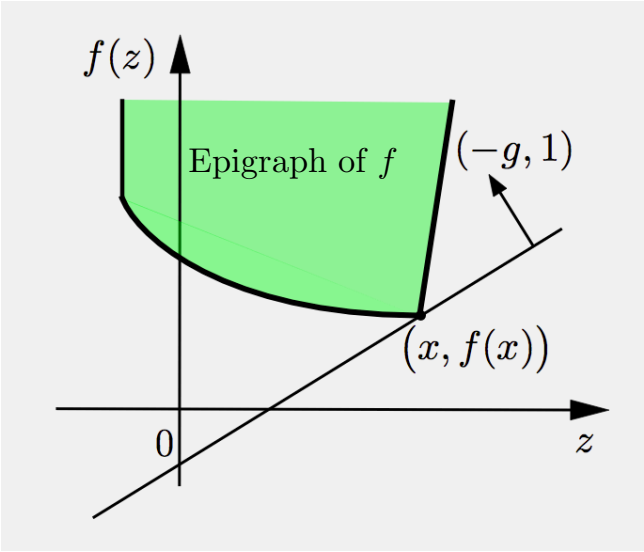
$$\begin{aligned} q(\nu) &= \inf_{x \in X} \big\{ f(x) + \nu' g(x) \big\} \leq f(x_\mu) + \nu' g(x_\mu) \\ &= f(x_\mu) + \mu' g(x_\mu) + (\nu - \mu)' g(x_\mu) = q(\mu) + (\nu - \mu)' g(x_\mu) \end{aligned}$$

LECTURE 15

LECTURE OUTLINE

- Overview of properties of subgradients
- Subgradient methods
- Convergence analysis
- Reading: Section 2.2 of Convex Optimization Algorithms

SUBGRADIENTS - REAL-VALUED FUNCTIONS



- Let $f : \Re^n \mapsto (-\infty, \infty]$ be a convex function. A vector $g \in \Re^n$ is a subgradient of f at a point $x \in \text{dom}(f)$ if

$$f(z) \geq f(x) + (z - x)'g, \qquad \forall\ z \in \Re^n$$

The set of subgradients at x is the subdifferential $\partial f(x)$.

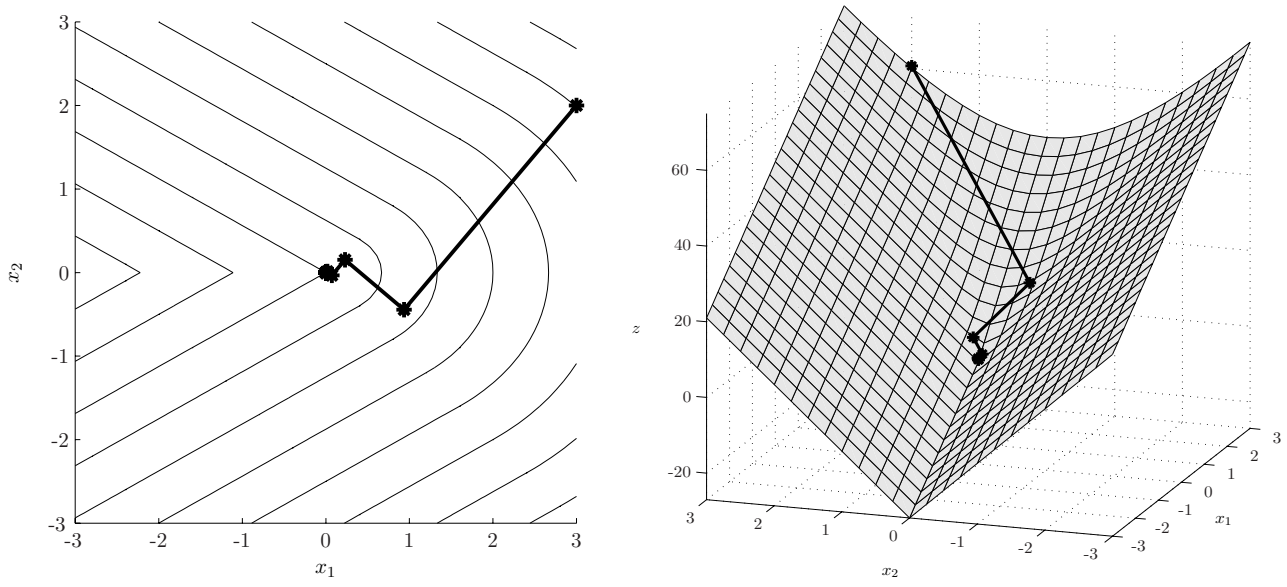
- If f is real-valued, $\partial f(x)$ is nonempty, convex, and compact for all x.

- $\cup_{x \in X} \partial f(x)$ is bounded if X is bounded.

- A single subgradient $g \in \partial f(x)$ is much easier to calculate than $\partial f(x)$ for many contexts involving dual functions and minimax.

MOTIVATION

- Consider minimization of convex f.
- Steepest descent at a point requires knowledge of the entire subdifferential at a point.
- Convergence failure of steepest descent



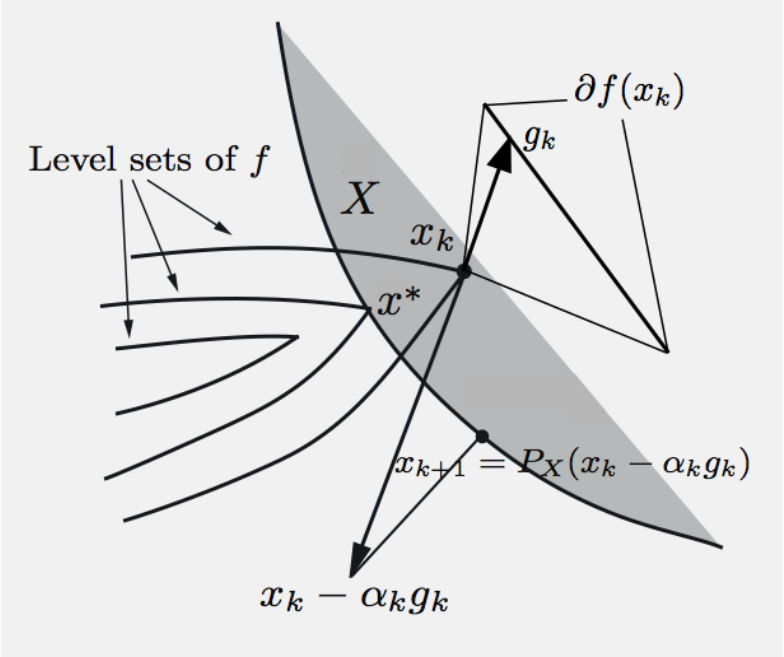
- Subgradient methods abandon the idea of computing the full subdifferential to effect cost function descent ...
- Move instead along the direction of an arbitrary subgradient

THE BASIC SUBGRADIENT METHOD

- Problem: Minimize convex function $f : \Re^n \mapsto \Re$ over a closed convex set X.
- Subgradient method:

$$x_{k+1} = P_X(x_k - \alpha_k g_k),$$

where g_k is any subgradient of f at x_k, α_k is a positive stepsize, and $P_X(\cdot)$ is projection on X.



- Key geometric fact: By subgradient inequality

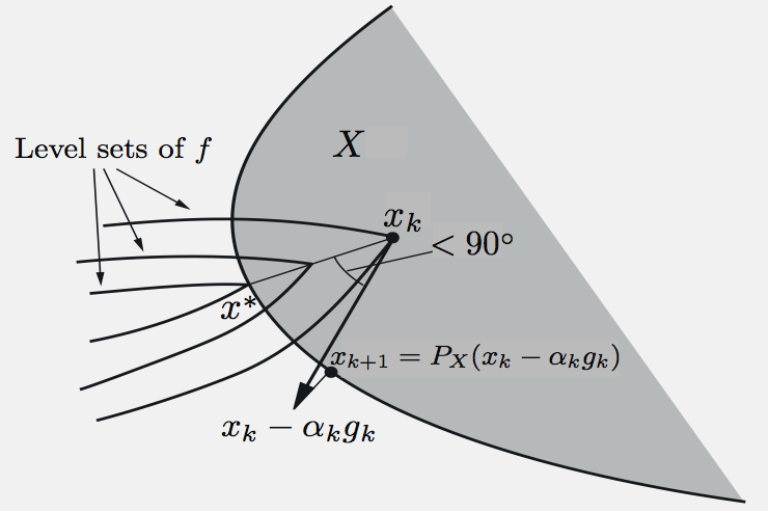
$$g_k'(y - x_k) < 0, \qquad \forall\ y \text{ such that } f(y) < f(x_k),$$

including all optimal y.

KEY PROPERTIES OF SUBGRADIENT METHOD

- For a small enough stepsize α_k, it reduces the Euclidean distance to the optimum.
- Nonexpansiveness of projection: For all x, y,

$$\|P_X(x) - P_X(y)\| \leq \|x - y\|$$



- Proposition: Let $\{x_k\}$ be generated by the subgradient method. Then, for all $y \in X$ and k:

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k \big(f(x_k) - f(y)\big) + \alpha_k^2 \|g_k\|^2$$

and if $f(y) < f(x_k)$,

$$\|x_{k+1} - y\| < \|x_k - y\|,$$

for all α_k such that $0 < \alpha_k < \dfrac{2\big(f(x_k) - f(y)\big)}{\|g_k\|^2}$.

PROOF

• Proof of nonexpansive property: From the projection theorem

$$\big(z - P_X(x)\big)'\big(x - P_X(x)\big) \leq 0, \qquad \forall\ z \in X,$$

from which $\big(P_X(y) - P_X(x)\big)'\big(x - P_X(x)\big) \leq 0$. Similarly, $\big(P_X(x) - P_X(y)\big)'\big(y - P_X(y)\big) \leq 0$.

Adding and using the Schwarz inequality,

$$\begin{aligned}\big\|P_X(y) - P_X(x)\big\|^2 &\leq \big(P_X(y) - P_X(x)\big)'(y - x)\\ &\leq \big\|P_X(y) - P_X(x)\big\| \cdot \|y - x\|\end{aligned}$$

Q.E.D.

• Proof of proposition: Since projection is nonexpansive, we obtain for all $y \in X$ and k,

$$\begin{aligned}\|x_{k+1} - y\|^2 &= \big\|P_X\left(x_k - \alpha_k g_k\right) - y\big\|^2\\ &\leq \|x_k - \alpha_k g_k - y\|^2\\ &= \|x_k - y\|^2 - 2\alpha_k g_k'(x_k - y) + \alpha_k^2\|g_k\|^2\\ &\leq \|x_k - y\|^2 - 2\alpha_k\big(f(x_k) - f(y)\big) + \alpha_k^2\|g_k\|^2,\end{aligned}$$

where the last inequality follows from the subgradient inequality. **Q.E.D.**

CONVERGENCE MECHANISM

- Assume constant stepsize: $\alpha_k \equiv \alpha$
- Assume that $\|g_k\| \leq c$ for some c and all k.

Then for all optimal x^*

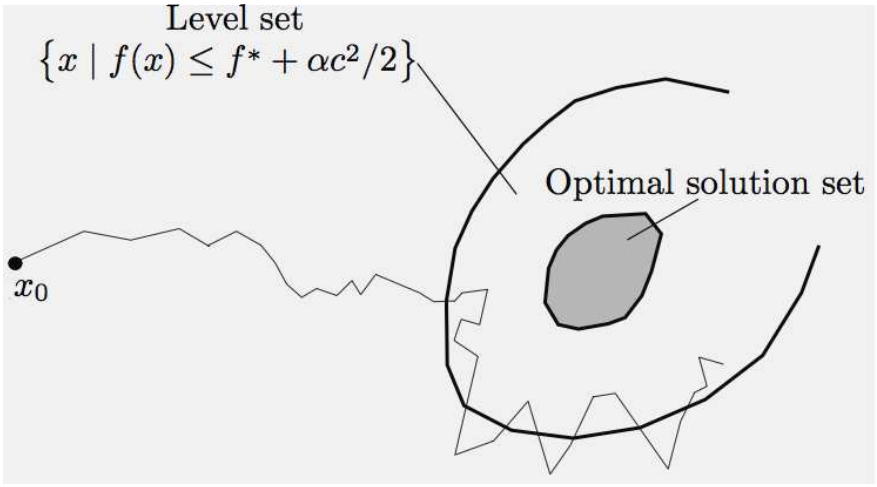
$$\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2-2\alpha\big(f(x_k)-f(x^*)\big)+\alpha^2c^2$$

Thus the distance to the optimum decreases if

$$0 < \alpha < \frac{2\big(f(x_k) - f(x^*)\big)}{c^2}$$

or equivalently, if x_k does not belong to the level set

$$\left\{ x \;\middle|\; f(x) < f(x^*) + \frac{\alpha c^2}{2} \right\}$$



STEPSIZE RULES

- Constant Stepsize: $\alpha_k \equiv \alpha$.
- Diminishing Stepsize: $\alpha_k \to 0$, $\sum_k \alpha_k = \infty$
- Dynamic Stepsize:

$$\alpha_k = \frac{f(x_k) - f_k}{c^2}$$

where f_k is an estimate of f^*:

  - If $f_k = f^*$, makes progress at every iteration. If $f_k < f^*$ it tends to oscillate around the optimum. If $f_k > f^*$ it tends towards the level set $\{x \mid f(x) \leq f_k\}$.
  - f_k can be adjusted based on the progress of the method.

- Example of dynamic stepsize rule:

$$f_k = \min_{0 \leq j \leq k} f(x_j) - \delta_k,$$

and δ_k (the "aspiration level of cost reduction") is updated according to

$$\delta_{k+1} = \begin{cases} \rho\delta_k & \text{if } f(x_{k+1}) \leq f_k, \\ \max\{\beta\delta_k, \delta\} & \text{if } f(x_{k+1}) > f_k, \end{cases}$$

where $\delta > 0$, $\beta < 1$, and $\rho \geq 1$ are fixed constants.

SAMPLE CONVERGENCE RESULTS

- Let $\overline{f} = \inf_{k \geq 0} f(x_k)$, and assume that for some c, we have

$$c \geq \sup\{\|g\| \mid g \in \partial f(x_k), \ k \geq 0\}$$

- Proposition: Assume that α_k is fixed at some positive scalar α. Then:

  (a) If $f^* = -\infty$, then $\overline{f} = f^*$.

  (b) If $f^* > -\infty$, then

$$\overline{f} \leq f^* + \frac{\alpha c^2}{2}.$$

- Proposition: If α_k satisfies

$$\lim_{k \to \infty} \alpha_k = 0, \qquad \sum_{k=0}^{\infty} \alpha_k = \infty,$$

then $\overline{f} = f^*$.

- Similar propositions for dynamic stepsize rules.
- Many variants ...

CONVERGENCE METHODOLOGY I

• Classical Contraction Mapping Theorem: Consider iteration $x_{k+1} = G(x_k)$, where $G : \Re^n \mapsto \Re^n$ is a contraction, i.e., for some $\rho < 1$

$$\big\|G(x) - G(y)\big\| \leq \rho\|x - y\|, \qquad \forall\ x, y \in \Re^n$$

where $\|\cdot\|$ is any norm. It converges to the unique fixed point of G.

• Can be used for gradient iterations with constant stepsize, but not subgradient iterations.

• Consider time varying contraction iteration $x_{k+1} = G_k(x_k)$, where

$$\big\|G_k(x)-G_k(y)\big\| \leq (1-\rho_k)\|x-y\|, \qquad \forall\ x, y \in \Re^n,$$

the G_k have a common fixed point, and

$$\rho_k \in (0, 1], \qquad \sum_{k=0}^{\infty} \rho_k = \infty$$

It converges to the unique common fixed point of G_k.

• Can be used for some time-varying gradient iterations, but not subgradient iterations.

CONVERGENCE METHODOLOGY II

• Supermartingale convergence (deterministic case): Let $\{Y_k\}$, $\{Z_k\}$, $\{W_k\}$, and $\{V_k\}$ be four nonnegative scalar sequences such that

$$Y_{k+1} \leq (1 + V_k)Y_k - Z_k + W_k, \qquad \forall\ k,$$

and

$$\sum_{k=0}^{\infty} W_k < \infty, \qquad \sum_{k=0}^{\infty} V_k < \infty$$

Then $\{Y_k\}$ converges and $\sum_{k=0}^{\infty} Z_k < \infty$.

• Supermartingale convergence (stochastic case): Let $\{Y_k\}$, $\{Z_k\}$, $\{W_k\}$, and $\{V_k\}$ be four nonnegative sequences of random variables, and let $\mathcal{F}_k$, $k = 0, 1, \ldots$, be sets of random variables such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k. Assume that

(1) For each k, Y_k, Z_k, W_k, and V_k are functions of the random variables in $\mathcal{F}_k$.

(2) $E\{Y_{k+1} \mid \mathcal{F}_k\} \leq (1 + V_k)Y_k - Z_k + W_k\ \forall\ k$

(3) There holds, $\sum_{k=0}^{\infty} W_k < \infty$, $\sum_{k=0}^{\infty} V_k < \infty$

Then $\{Y_k\}$ converges to some random variable Y, and $\sum_{k=0}^{\infty} Z_k < \infty$, with probability 1.

CONVERGENCE FOR DIMINISHING STEPSIZE

- Proposition: Assume that the optimal solution set X^* is nonempty, and that for some c and all k,

$$c^2 \left(1 + \min_{x^* \in X^*} \|x_k - x^*\|^2\right) \geq \sup\{\|g\|^2 \mid g \in \partial f(x_k)\}$$

and α_k satisfies

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

Then $\{x_k\}$ converges to an optimal solution.

Proof: Write for any optimal x^*

$$\begin{aligned} \|x_{k+1} - x^*\|^2 \leq (1 + \alpha_k^2 c^2)\|x_k - x^*\|^2 \\ - 2\alpha_k \big(f(x_k) - f(x^*)\big) \\ + \alpha_k^2 c^2 \end{aligned}$$

Use the supermartingale convergence theorem.

LECTURE 16

LECTURE OUTLINE

- Approximation approach for convex optimization algorithms:
- Cutting plane method
- Simplicial decomposition

- Reading: Section 6.4 of on-line Chapter 6 on algorithms

CUTTING PLANE METHOD

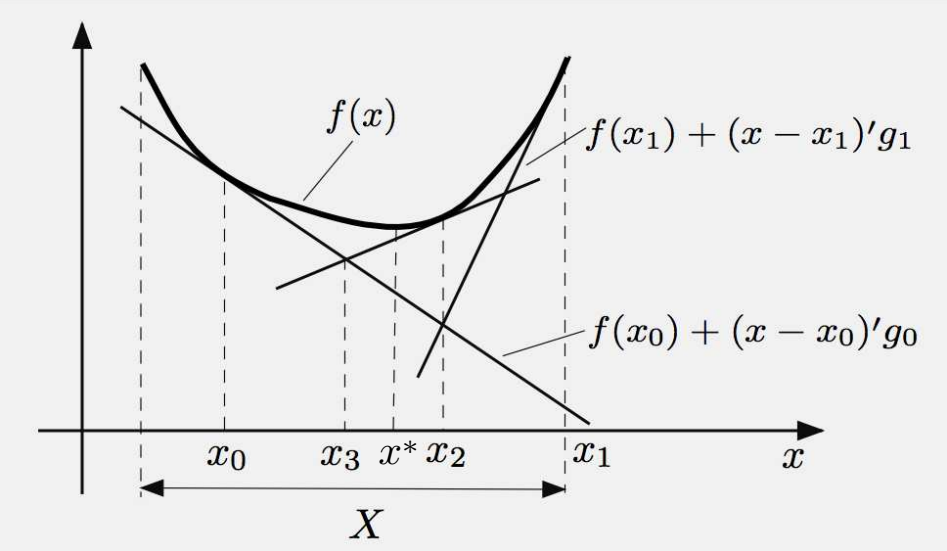
- Problem: Minimize $f : \Re^n \mapsto \Re$ subject to $x \in X$, where f is convex, and X is closed convex.
- Method: Start with any $x_0 \in X$. For $k \geq 0$, set

$$x_{k+1} \in \arg\min_{x \in X} F_k(x),$$

where

$$F_k(x) = \max\Big\{f(x_0)+(x-x_0)'g_0, \ldots, f(x_k)+(x-x_k)'g_k\Big\}$$

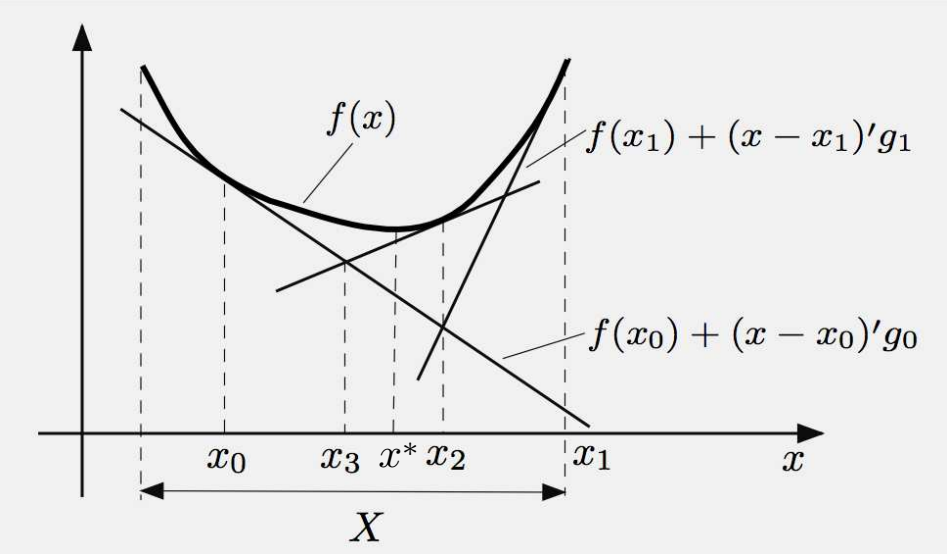
and g_i is a subgradient of f at x_i.



CONVERGENCE OF CUTTING PLANE METHOD

$$F_k(x) = \max\big\{f(x_0)+(x-x_0)'g_0, \ldots, f(x_k)+(x-x_k)'g_k\big\}$$

$$F_k(x_{k+1}) \leq F_k(x) \leq f(x), \qquad \forall\ x$$



- $F_k(x_k)$ increases monotonically with k, and all limit points of $\{x_k\}$ are optimal.

Proof: (Abbreviated) If $x_k \to \overline{x}$ then $F_k(x_k) \to f(\overline{x})$, [otherwise there would exist a hyperplane strictly separating epi(f) and $(\overline{x}, \lim_{k\to\infty} F_k(x_k))$]. This implies that

$$f(\overline{x}) \leq \lim_{k\to\infty} F_k(x) \leq f(x), \qquad \forall\ x \in X$$

Q.E.D.

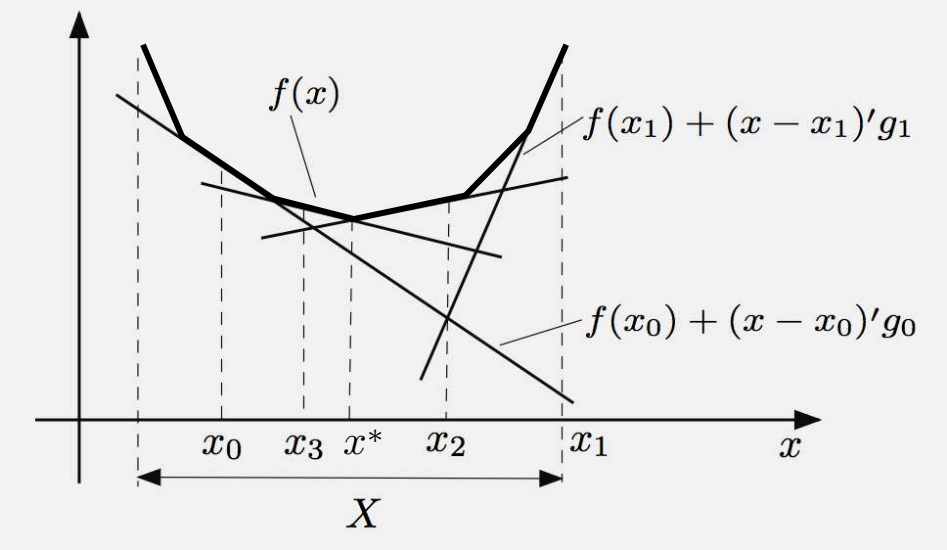
TERMINATION

- We have for all k

$$F_k(x_{k+1}) \leq f^* \leq \min_{i \leq k} f(x_i)$$

- Termination when $\min_{i \leq k} f(x_i) - F_k(x_{k+1})$ comes to within some small tolerance.

- For f polyhedral, we have finite termination with an exactly optimal solution.

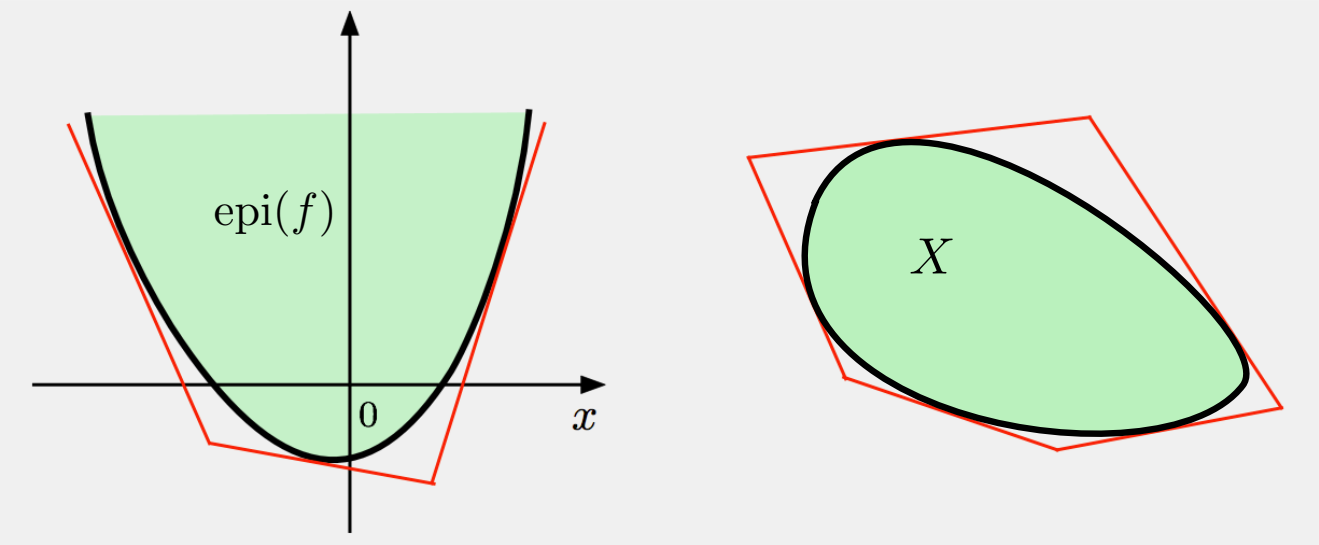


- Instability problem: The method can make large moves that deteriorate the value of f.

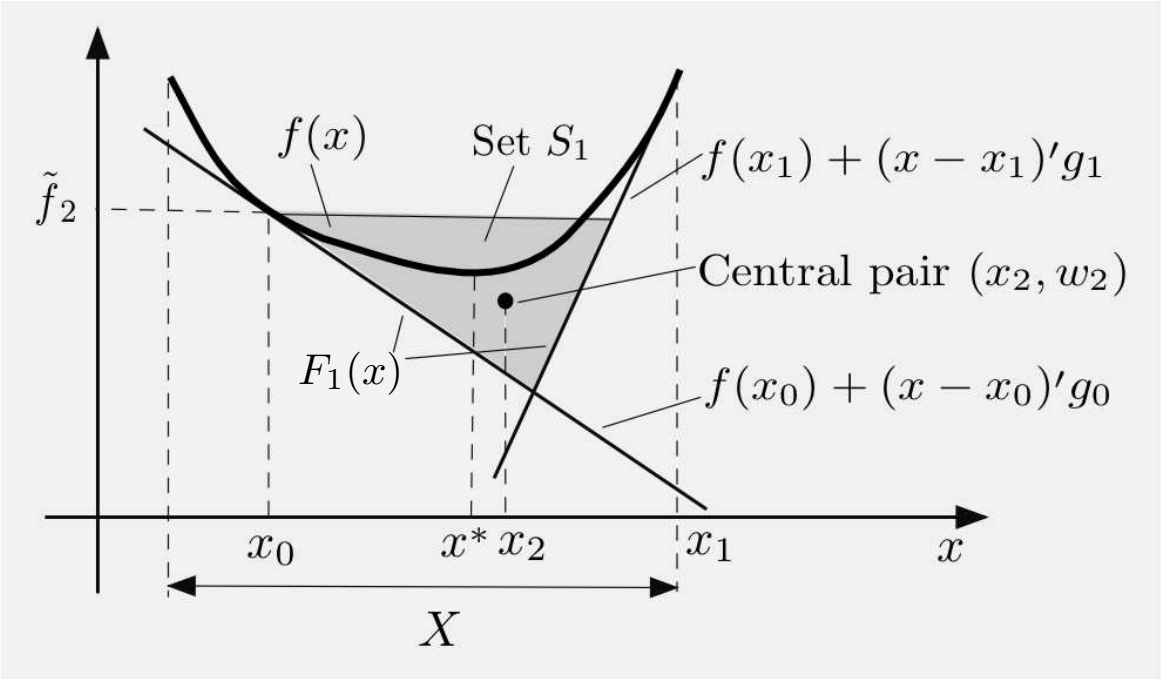
- Starting from the exact minimum it typically moves away from that minimum.

VARIANTS

- **Variant I:** Simultaneously with f, construct polyhedral approximations to X.



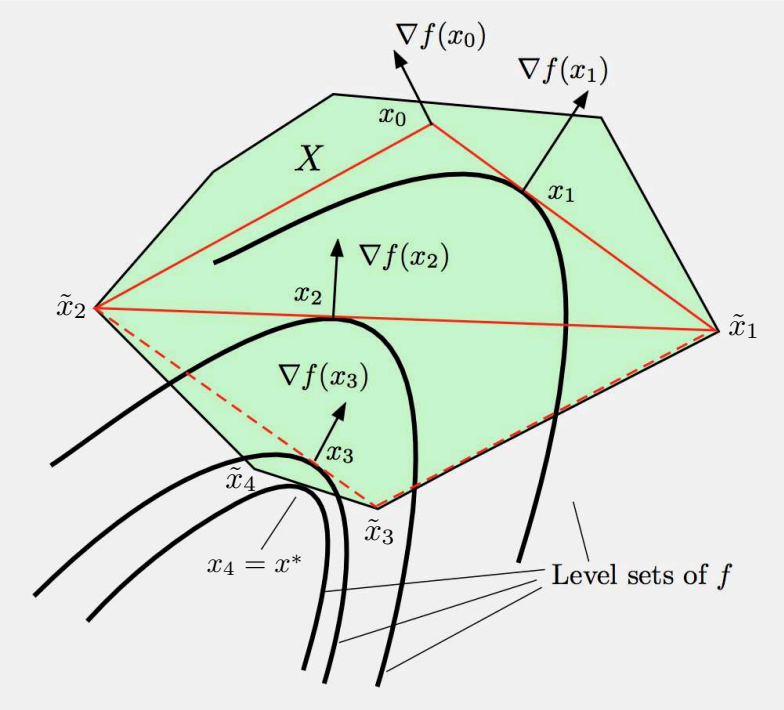
- **Variant II:** Central cutting plane methods



SIMPLICIAL DECOMPOSITION IDEAS

- Minimize a differentiable convex $f : \Re^n \mapsto \Re$ over bounded polyhedral constraint set X.

- Approximate X with a simpler inner approximating polyhedral set.



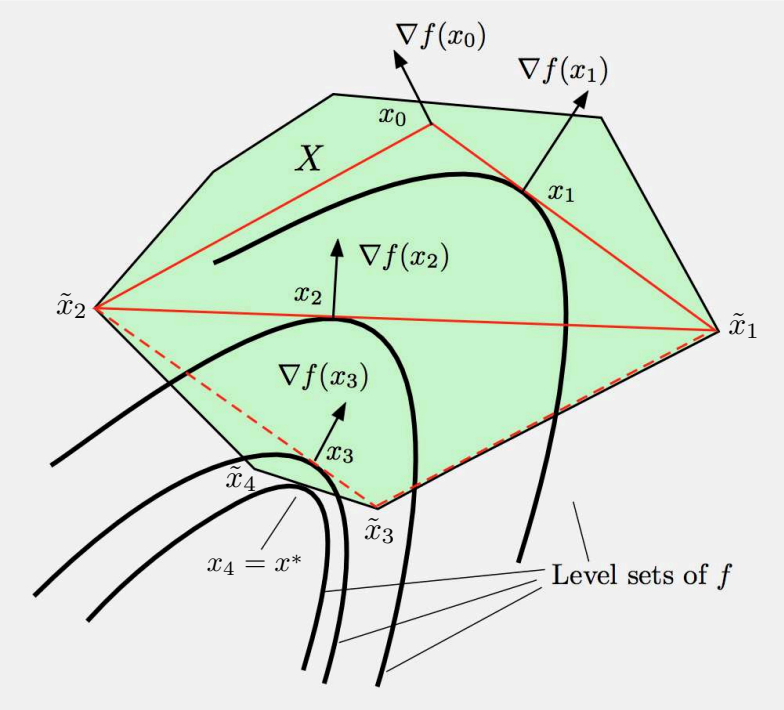
- Approximating problem (min over a simplex):

$$\text{minimize} \quad f\left(\sum_{j=1}^{k} \alpha_j \tilde{x}_j\right)$$

$$\text{subject to} \quad \sum_{j=1}^{k} \alpha_j = 1, \ \alpha_j \geq 0$$

- Construct a more refined problem by solving a linear minimization over the original constraint.

SIMPLICIAL DECOMPOSITION METHOD



- Given current iterate x_k, and finite set $X_k \subset X$ (initially $x_0 \in X$, $X_0 = \{x_0\}$).
- Let $\tilde{x}_{k+1}$ be extreme point of X that solves

$$\begin{aligned} &\text{minimize} \quad \nabla f(x_k)'(x - x_k) \\ &\text{subject to} \quad x \in X \end{aligned}$$

and add $\tilde{x}_{k+1}$ to X_k: $X_{k+1} = \{\tilde{x}_{k+1}\} \cup X_k$.

- Generate x_{k+1} as optimal solution of

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in \text{conv}(X_{k+1}). \end{aligned}$$

CONVERGENCE

- There are two possibilities for $\tilde{x}_{k+1}$:

  (a) We have

  $$0 \leq \nabla f(x_k)'(\tilde{x}_{k+1} - x_k) = \min_{x \in X} \nabla f(x_k)'(x - x_k)$$

  Then x_k minimizes f over X (satisfies the optimality condition)

  (b) We have

  $$0 > \nabla f(x_k)'(\tilde{x}_{k+1} - x_k)$$

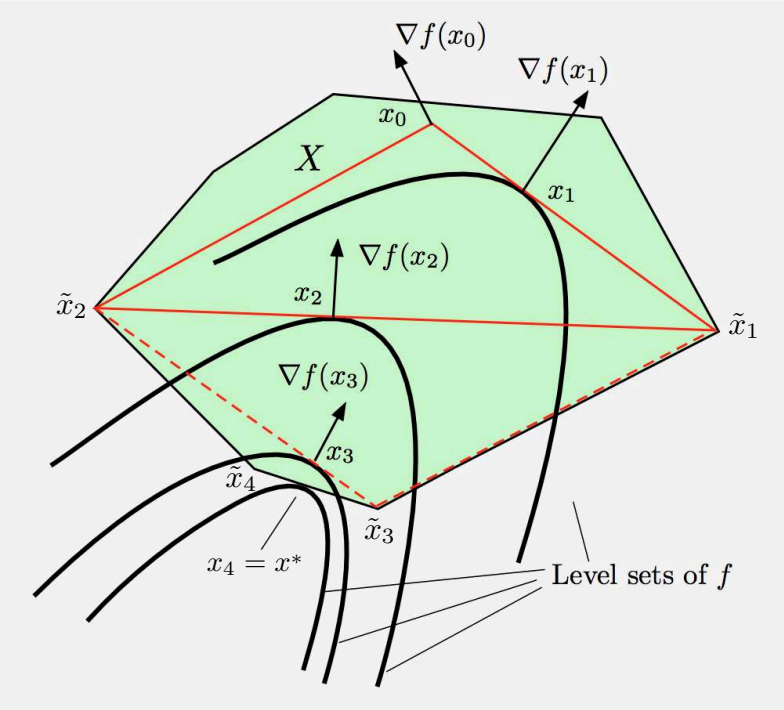
  Then $\tilde{x}_{k+1} \notin \text{conv}(X_k)$, since x_k minimizes f over $x \in \text{conv}(X_k)$, so that

  $$\nabla f(x_k)'(x - x_k) \geq 0, \qquad \forall\ x \in \text{conv}(X_k)$$

- Case (b) cannot occur an infinite number of times ($\tilde{x}_{k+1} \notin X_k$ and X has finitely many extreme points), so case (a) must eventually occur.

- The method will find a minimizer of f over X in a finite number of iterations.

COMMENTS ON SIMPLICIAL DECOMP.



- The method is appealing under two conditions:
  - Minimizing f over the convex hull of a relative small number of extreme points is much simpler than minimizing f over X.
  - Minimizing a linear function over X is much simpler than minimizing f over X.
- Important specialized applications relating to routing problems in data networks and transportation.

VARIANTS OF SIMPLICIAL DECOMP.

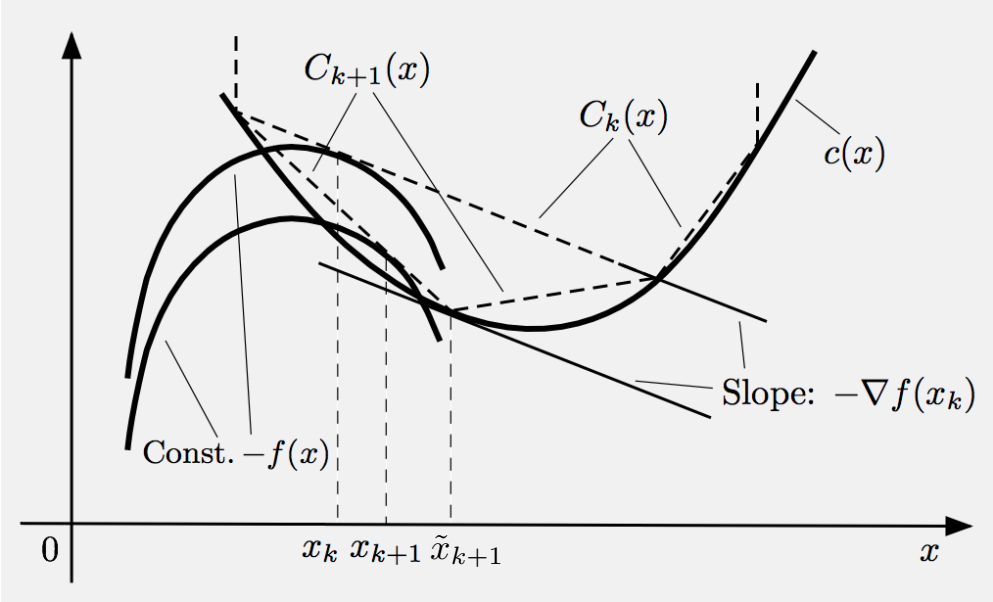
- Variant to remove the boundedness assumption on X (impose artificial constraints).
- Variant to enhance efficiency: Discard some of the extreme points that seem unlikely to "participate" in the optimal solution, i.e., all $\tilde{x}$ such that

$$\nabla f(x_{k+1})'(\tilde{x} - x_{k+1}) > 0$$

- Additional methodological enhancements:
  - Extension to X nonpolyhedral (method remains unchanged, but convergence proof is more complex)
  - Extension to f nondifferentiable (requires use of subgradients in place of gradients, and more sophistication)
  - Duality relation with cutting plane methods based on Fenchel duality.
- We will derive, justify, and extend these by showing that cutting plane and simplicial decomposition are special cases of two polyhedral approximation methods that are dual to each other (next lecture).

GENERALIZED SIMPLICIAL DECOMPOSITION

- Consider minimization of $f(x) + c(x)$, over $x \in \Re^n$, where f and c are closed proper convex
- Case where f is differentiable



- Form C_k: inner linearization of c [epi(C_k) is the convex hull of the halflines $\{(\tilde{x}_j, w) \mid w \geq f(\tilde{x}_j)\}$, $j = 1, \ldots, k$]. Find

$$x_k \in \arg\min_{x \in \Re^n} \{f(x) + C_k(x)\}$$

- Obtain $\tilde{x}_{k+1}$ such that

$$-\nabla f(x_k) \in \partial c(\tilde{x}_{k+1}),$$

and form $X_{k+1} = X_k \cup \{\tilde{x}_{k+1}\}$

NONDIFFERENTIABLE CASE

- Given C_k: inner linearization of c, obtain

$$x_k \in \arg\min_{x \in \Re^n} \{f(x) + C_k(x)\}$$

- Obtain a subgradient $g_k \in \partial f(x_k)$ such that

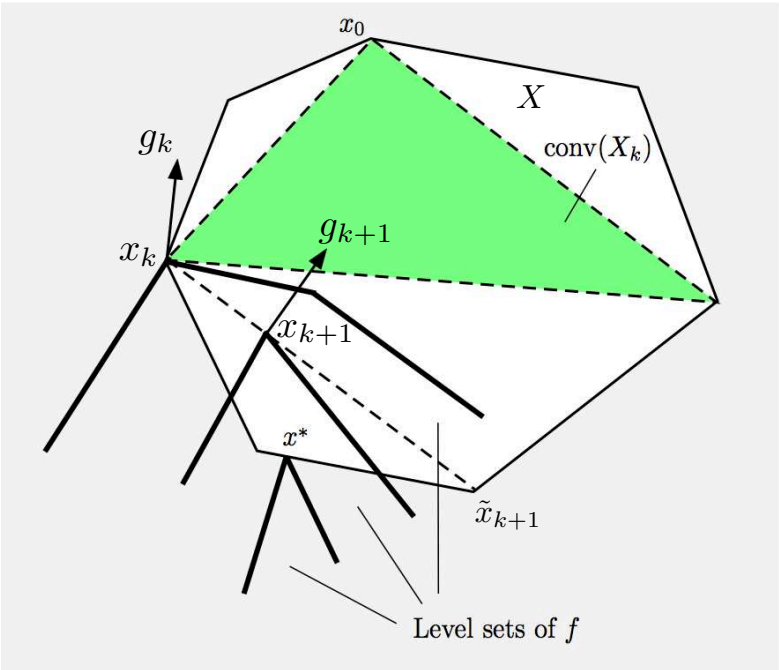
$$-g_k \in \partial C_k(x_k)$$

- Obtain $\tilde{x}_{k+1}$ such that

$$-g_k \in \partial c(\tilde{x}_{k+1}),$$

and form $X_{k+1} = X_k \cup \{\tilde{x}_{k+1}\}$

- Example: c is the indicator function of a polyhedral set



LECTURE 18

LECTURE OUTLINE

- Proximal algorithm
- Convergence
- Rate of convergence
- Extensions

****

Consider minimization of closed proper convex $f : \Re^n \mapsto (-\infty, +\infty]$ using a different type of approximation:

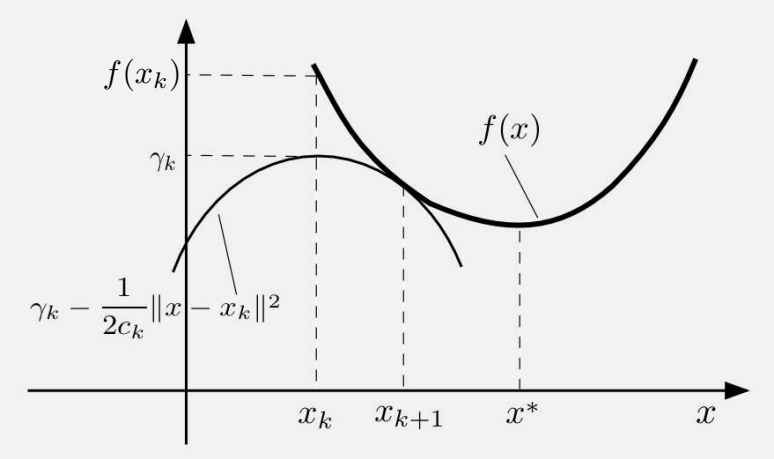
- Regularization in place of linearization
- Add a quadratic term to f to make it strictly convex and "well-behaved"
- Refine the approximation at each iteration by changing the quadratic term

PROXIMAL MINIMIZATION ALGORITHM

- A general algorithm for convex fn minimization

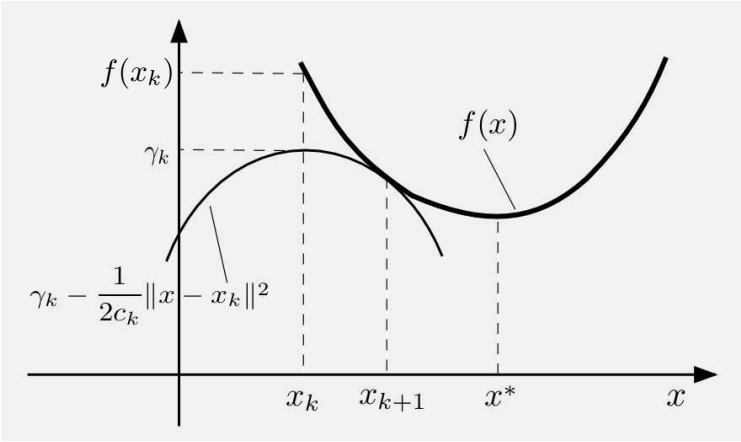
$$x_{k+1} \in \arg\min_{x \in \Re^n} \left\{ f(x) + \frac{1}{2c_k} \|x - x_k\|^2 \right\}$$

  - $f : \Re^n \mapsto (-\infty, \infty]$ is closed proper convex
  - c_k is a positive scalar parameter
  - x_0 is arbitrary starting point



- x_{k+1} exists because of the quadratic.
- Note it does not have the instability problem of cutting plane method
- If x_k is optimal, $x_{k+1} = x_k$.
- Main Convergence Theorem: If $\sum_k c_k = \infty$, $f(x_k) \to f^*$. Moreover $\{x_k\}$ converges to an optimal solution if one exists.

CONVERGENCE: SOME BASIC PROPERTIES



- Note the connection with Fenchel framework
- From subdifferential of sum formula (or Fenchel duality theorem)

$$(x_k - x_{k+1})/c_k \in \partial f(x_{k+1})$$

Note the similarity with the subgradient method $(x_k - x_{k+1})/c_k \in \partial f(x_k)$

- Cost improves:

$$f(x_{k+1}) + \frac{1}{2c_k}\|x_{k+1} - x_k\|^2 \leq f(x_k)$$

- Distance to the optimum improves:

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2c_k\big(f(x_{k+1}) - f(y)\big) - \|x_k - x_{k+1}\|^2$$

for all k and $y \in \Re^n$.

CONVERGENCE PROOF I

• Main Convergence Theorem: If $\sum_k c_k = \infty$, $f(x_k) \downarrow f^*$. Moreover $\{x_k\}$ converges to an optimal solution if one exists.

Proof: Have $f(x_k) \downarrow f_\infty \geq f^*$. For all y and k,

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2c_k\big(f(x_{k+1}) - f(y)\big)$$

By adding over $k = 0, \ldots, N$,

$$\|x_{N+1} - y\|^2 + 2\sum_{k=0}^{N} c_k\big(f(x_{k+1}) - f(y)\big) \leq \|x_0 - y\|^2,$$

so taking the limit as $N \to \infty$,

$$2\sum_{k=0}^{\infty} c_k\big(f(x_{k+1}) - f(y)\big) \leq \|x_0 - y\|^2 \qquad (*)$$

• Argue by contradiction: Assume $f_\infty > f^*$, and let $\hat{y}$ be such that $f_\infty > f(\hat{y}) > f^*$. Then

$$f(x_{k+1}) - f(\hat{y}) \geq f_\infty - f(\hat{y}) > 0.$$

Since $\sum_{k=0}^{\infty} c_k = \infty$, $(*)$ leads to a contradiction. Thus $f_\infty = f^*$.

CONVERGENCE PROOF II

• Assume $X^* \neq \emptyset$. We will show convergence to some $x^* \in X^*$. Applying

$$\|x_{k+1} - y\|^2 \leq \|x_k - y\|^2 - 2c_k\big(f(x_{k+1}) - f(y)\big)$$

with $y = x^* \in X^*$,

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2c_k\big(f(x_{k+1}) - f(x^*)\big), \qquad (**)$$

Thus $\|x_k - x^*\|^2$ is monotonically nonincreasing, so $\{x_k\}$ is bounded.

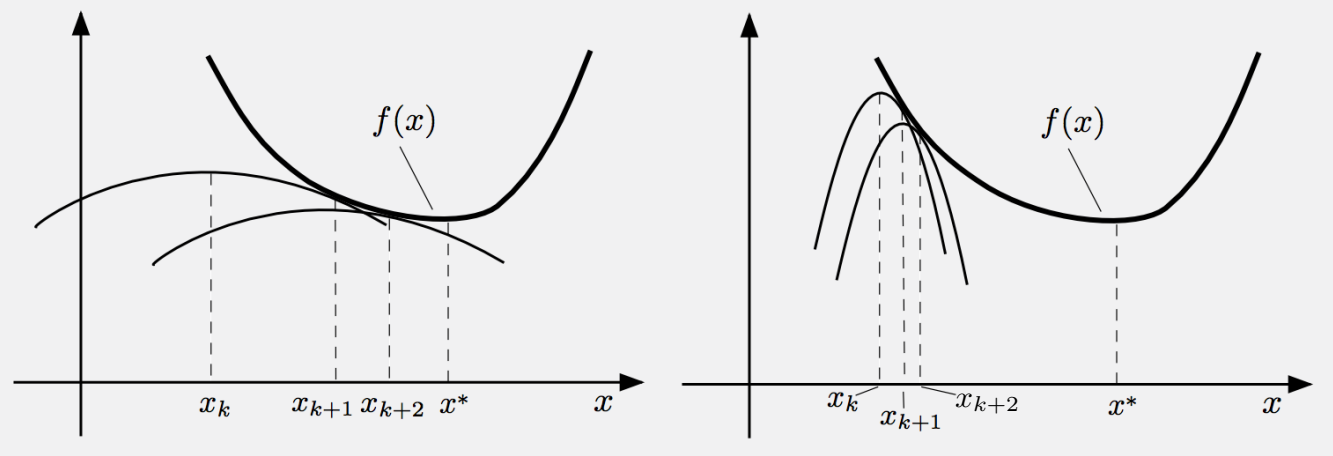
• If $\{x_k\}_{\mathcal{K}} \to \overline{z}$, the limit point $\overline{z}$ must belong to X^*, since $f(x_k) \downarrow f^*$, and f is closed, so

$$f(\overline{z}) \leq \liminf_{k \to \infty,\, k \in \mathcal{K}} f(x_k) = f^*$$

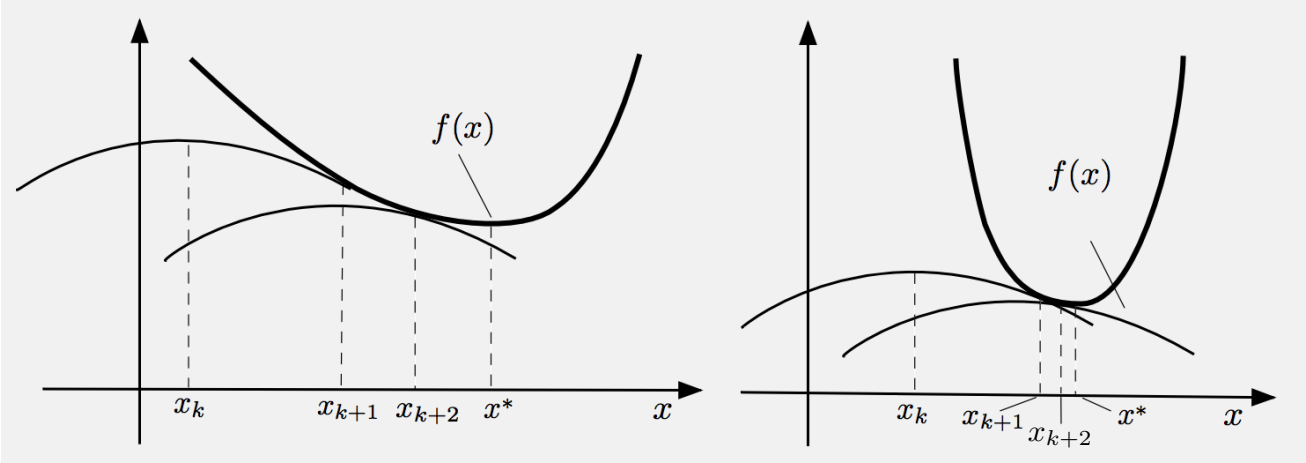
• By $(**)$, the distance of x_k to each limit point is monotonically nonincreasing, so $\{x_k\}$ must converge to a unique limit, which must be an element of X^*. **Q.E.D.**

RATE OF CONVERGENCE I

- Role of penalty parameter c_k:



- Role of growth properties of f near optimal solution set:



RATE OF CONVERGENCE II

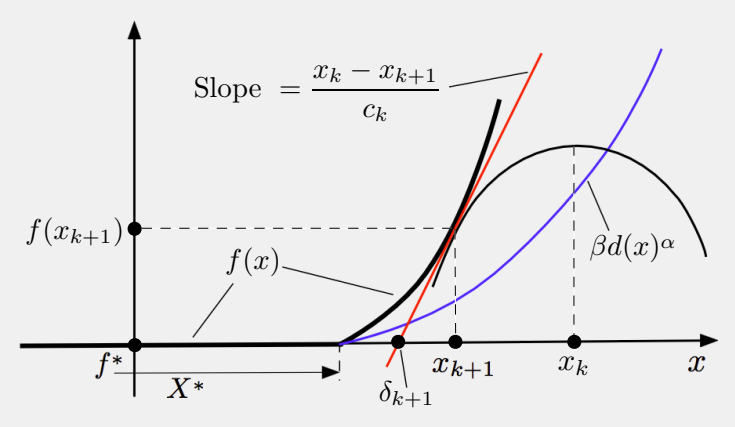
- Assume growth of order α from optimal solution set X^*, i.e., that for some $\beta > 0$, $\delta > 0$, and $\alpha \geq 1$,

$$f^* + \beta\big(d(x)\big)^\alpha \leq f(x), \quad \forall\, x \in \Re^n \ \text{ with } d(x) \leq \delta$$

where $d(x) = \min_{x^* \in X^*} \|x - x^*\|$

- Key property: For all k sufficiently large,

$$d(x_{k+1}) + \beta c_k \big(d(x_{k+1})\big)^{\alpha-1} \leq d(x_k)$$



- We have (in one dimension)

$$\begin{aligned}\beta\big(d(x_{k+1})\big)^\alpha &\leq f(x_{k+1}) - f^* \\ &= \frac{x_k - x_{k+1}}{c_k} \cdot (x_{k+1} - \delta_{k+1}) \\ &\leq \frac{d(x_k) - d(x_{k+1})}{c_k} \cdot d(x_{k+1})\end{aligned}$$

LINEAR AND SUPERLINEAR CONVERGENCE

- Use the key relation

$$d(x_{k+1}) + \beta c_k \big(d(x_{k+1})\big)^{\alpha-1} \leq d(x_k)$$

for various values of order of growth $\alpha \geq 1$.

- If $\alpha = 2$ and $\lim_{k\to\infty} c_k = \bar{c}$, then

$$\limsup_{k\to\infty} \frac{d(x_{k+1})}{d(x_k)} \leq \frac{1}{1+\beta\bar{c}}$$

linear convergence.

- If $1 < \alpha < 2$, then

$$\limsup_{k\to\infty} \frac{d(x_{k+1})}{\big(d(x_k)\big)^{1/(\alpha-1)}} < \infty$$

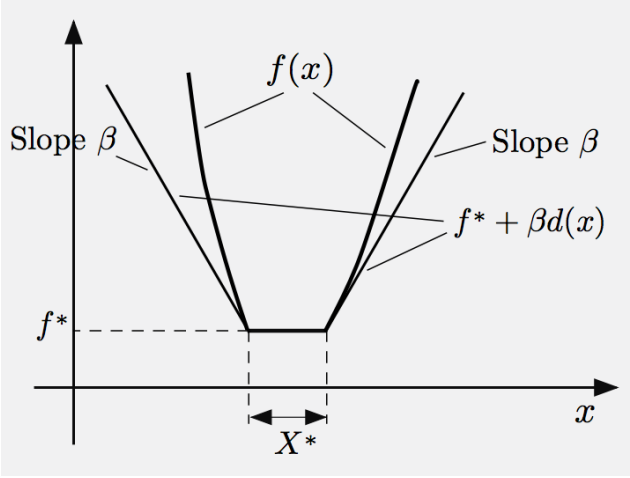
superlinear convergence.

FINITE CONVERGENCE

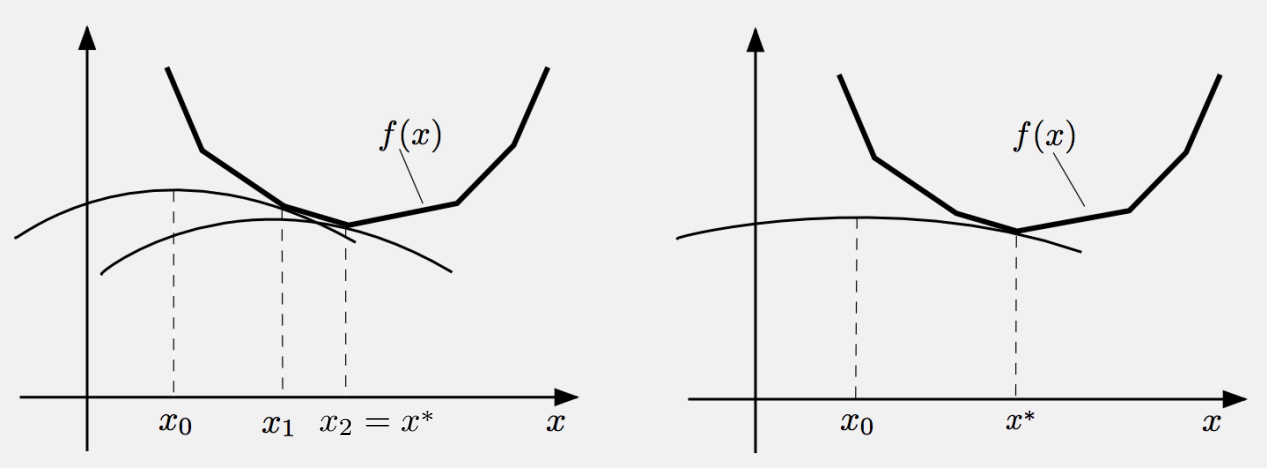
- Assume growth order $\alpha = 1$:

$$f^* + \beta d(x) \leq f(x), \qquad \forall\ x \in \Re^n$$

Can be shown to hold if f is polyhedral.

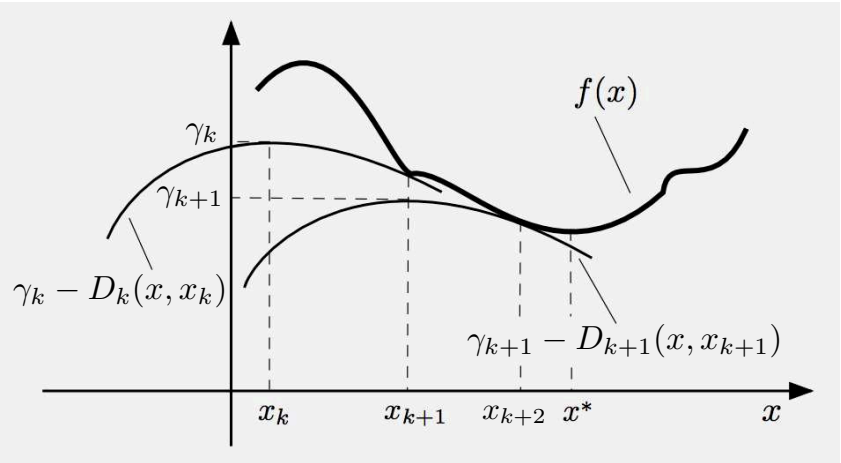


- Method converges finitely (in a single step for c_0 sufficiently large).



EXTENSIONS

- Combine with polyhedral approximation of f, to take advantage of finite convergence property.
  - Leads to bundle methods, which involve a mechanism to prevent the inherent instability of cutting plane method.
- Extension to more general problems:
  - Application to variational inequalities and games.
  - Application to finding a zero of a "maximally monotone multi-valued" mapping.
  - Allow nonconvex f (the theory is not clean and complete).
- Replace quadratic regularization by more general proximal term.



LECTURE 19

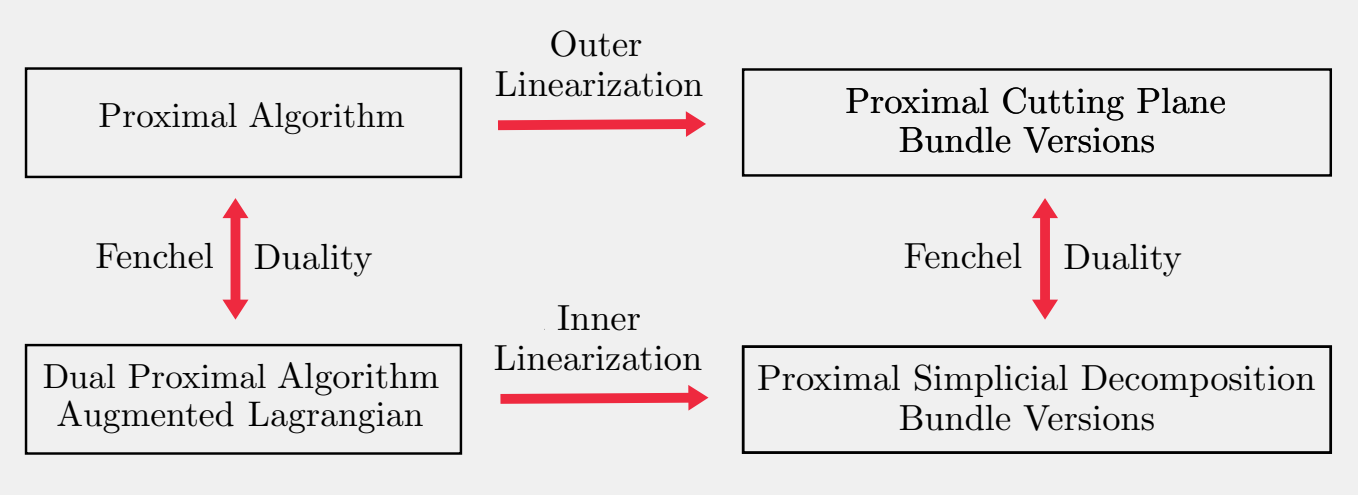
LECTURE OUTLINE

- Review of proximal algorithm
- Dual proximal algorithm
- Augmented Lagrangian methods
- Proximal cutting plane algorithm
- Bundle methods

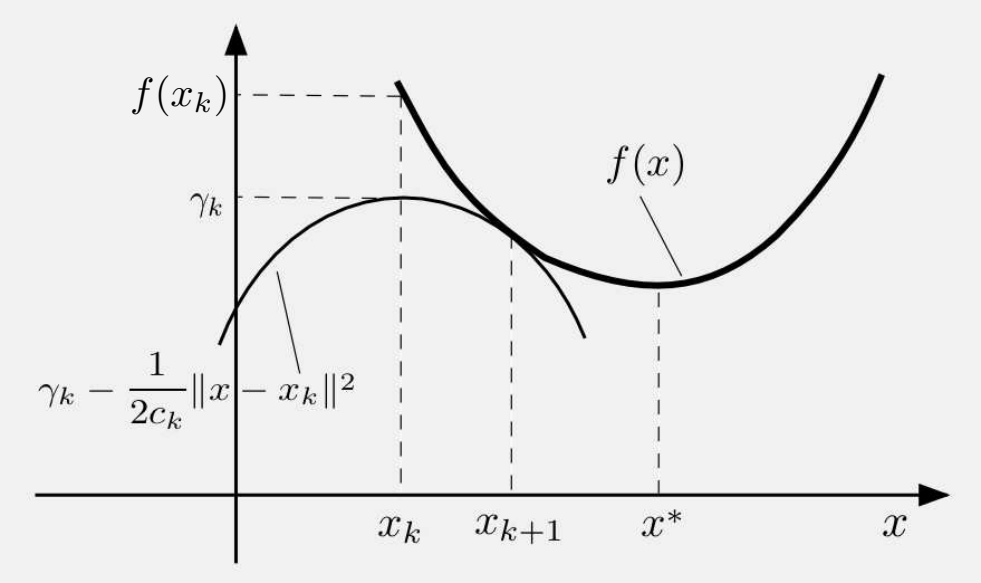
****

Start with proximal algorithm and generate other methods via:

  - Fenchel duality
  - Outer/inner linearization



RECALL PROXIMAL ALGORITHM



- Minimizes closed convex proper f:

$$x_{k+1} = \arg\min_{x \in \Re^n} \left\{ f(x) + \frac{1}{2c_k} \|x - x_k\|^2 \right\}$$

where x_0 is an arbitrary starting point, and $\{c_k\}$ is a positive parameter sequence.

- We have $f(x_k) \to f^*$. Also $x_k \to$ some minimizer of f, provided one exists.

- Finite convergence for polyhedral f.

- Each iteration can be viewed in terms of Fenchel duality.

REVIEW OF FENCHEL DUALITY

- Consider the problem

$$\begin{aligned} &\text{minimize} && f_1(x) + f_2(x) \\ &\text{subject to} && x \in \Re^n, \end{aligned}$$

where f_1 and f_2 are closed proper convex.

- Fenchel Duality Theorem:

  (a) If f^* is finite and $\text{ri}\big(\text{dom}(f_1)\big) \cap \text{ri}\big(\text{dom}(f_2)\big) \neq \emptyset$, then strong duality holds and there exists at least one dual optimal solution.

  (b) Strong duality holds, and (x^*, λ^*) is a primal and dual optimal solution pair if and only if

$$x^* \in \arg\min_{x \in \Re^n} \big\{ f_1(x) - x'\lambda^* \big\}, \quad x^* \in \arg\min_{x \in \Re^n} \big\{ f_2(x) + x'\lambda^* \big\}$$

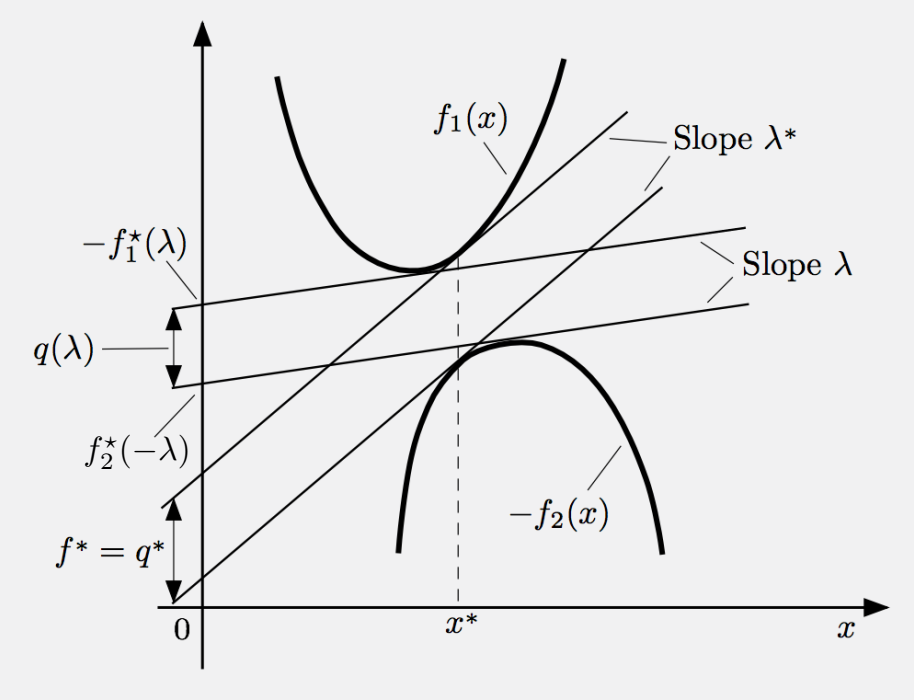
- By conjugate subgradient theorem, the last condition is equivalent to

$$\lambda^* \in \partial f_1(x^*) \qquad [\text{or equivalently } x^* \in \partial f_1^\star(\lambda^*)]$$

and

$$-\lambda^* \in \partial f_2(x^*) \qquad [\text{or equivalently } x^* \in \partial f_2^\star(-\lambda^*)]$$

GEOMETRIC INTERPRETATION



- The optimality condition is equivalent to

$$\lambda^* \in \partial f_1(x^*) \quad \text{and} \quad \lambda^* \in -\partial f_2(x^*); \quad \text{or}$$

$$x^* \in \partial f_1^\star(\lambda^*) \quad \text{and} \quad x^* \in \partial f_2^\star(-\lambda^*)$$

- More generally: Once we obtain one of x^* or λ^*, we can obtain the other by "differentiation"

$\lambda \in \partial f(x)$ $\Longleftrightarrow$ $x \in \arg\min_{z \in \Re^n} \{f(z) - z'\lambda\}$

$\Updownarrow$ Conjugate Subgradient Theorem $\Updownarrow$

$x \in \partial f^\star(\lambda)$ $\Longleftrightarrow$ $\lambda \in \arg\min_{z \in \Re^n} \{f^\star(z) - z'x\}$

DUAL PROXIMAL MINIMIZATION

- The proximal iteration can be written in the Fenchel form: $\min_x\{f_1(x) + f_2(x)\}$ with

$$f_1(x) = f(x), \qquad f_2(x) = \frac{1}{2c_k}\|x - x_k\|^2$$

- The Fenchel dual is

$$\begin{aligned} &\text{minimize} \quad f_1^\star(\lambda) + f_2^\star(-\lambda) \\ &\text{subject to} \quad \lambda \in \Re^n \end{aligned}$$

- We have $f_2^\star(-\lambda) = -x_k'\lambda + \frac{c_k}{2}\|\lambda\|^2$, so the dual problem is

$$\begin{aligned} &\text{minimize} \quad f^\star(\lambda) - x_k'\lambda + \frac{c_k}{2}\|\lambda\|^2 \\ &\text{subject to} \quad \lambda \in \Re^n \end{aligned}$$

where $f^\star$ is the conjugate of f.

- f_2 is real-valued, so no duality gap.

- Both primal and dual problems have a unique solution, since they involve a closed, strictly convex, and coercive cost function.

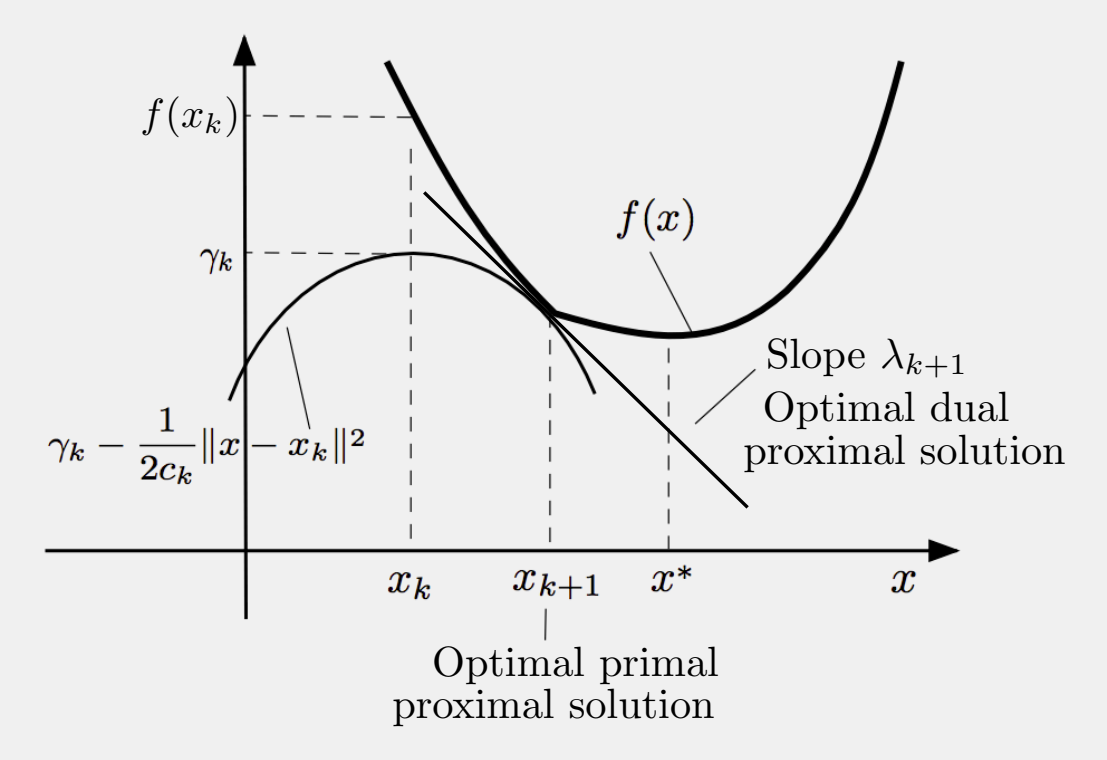
DUAL IMPLEMENTATION

- We can solve the Fenchel-dual problem instead of the primal at each iteration:

$$\lambda_{k+1} = \arg\min_{\lambda \in \Re^n} \left\{ f^\star(\lambda) - x_k' \lambda + \frac{c_k}{2} \|\lambda\|^2 \right\}$$

- Primal-dual optimal pair (x_{k+1}, λ_{k+1}) are related by the "differentiation" condition:

$$\lambda_{k+1} = \frac{x_k - x_{k+1}}{c_k}$$



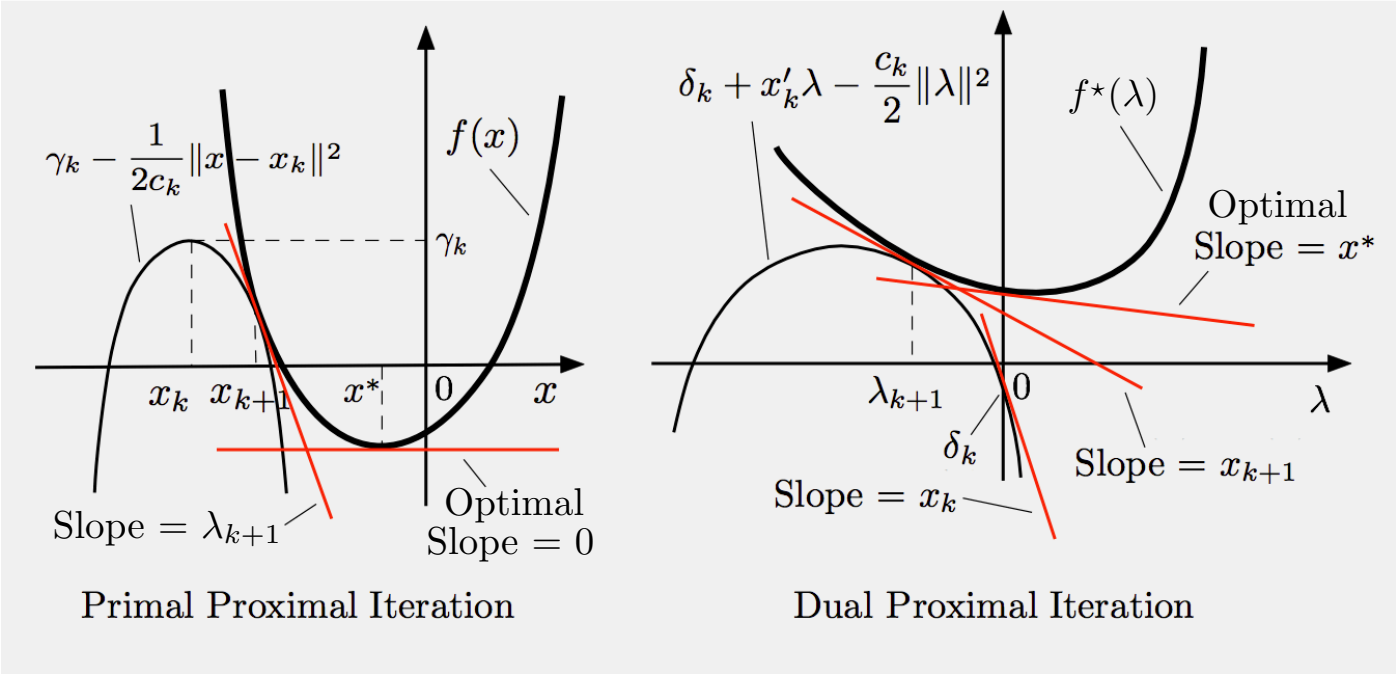
DUAL PROXIMAL ALGORITHM

- Obtain λ_{k+1} and x_{k+1} from

$$\lambda_{k+1} = \arg\min_{\lambda \in \Re^n} \left\{ f^\star(\lambda) - x_k'\lambda + \frac{c_k}{2}\|\lambda\|^2 \right\}$$

$$x_{k+1} = x_k - c_k\lambda_{k+1}$$

- As x_k converges to x^*, the dual sequence λ_k converges to 0 (a subgradient of f at x^*).



- The primal and dual algorithms generate identical sequences $\{x_k, \lambda_k\}$. Which one is preferable depends on whether f or its conjugate $f^\star$ has more convenient structure.

- Special case: The augmented Lagrangian method.

AUGMENTED LAGRANGIAN METHOD

- Consider the convex constrained problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in X, \quad Ax = b \end{aligned}$$

- Primal and dual functions:

$$p(u) = \inf_{\substack{x \in X \\ Ax - b = u}} f(x), \quad q(\lambda) = \inf_{x \in X} \left\{ f(x) + \lambda'(Ax - b) \right\}$$

- Assume p: closed, so (q, p) are "conjugate" pair.
- Primal and dual prox. algorithms for $\max_\lambda q(\lambda)$:

$$\lambda_{k+1} = \arg \max_{\lambda \in \Re^m} \left\{ q(\lambda) - \frac{1}{2c_k} \|\lambda - \lambda_k\|^2 \right\}$$

$$u_{k+1} = \arg \min_{u \in \Re^m} \left\{ p(u) + \lambda_k' u + \frac{c_k}{2} \|u\|^2 \right\}$$

Dual update: $\lambda_{k+1} = \lambda_k + c_k u_{k+1}$

- Implementation:

$$u_{k+1} = Ax_{k+1} - b, \qquad x_{k+1} \in \arg \min_{x \in X} L_{c_k}(x, \lambda_k)$$

where L_c is the Augmented Lagrangian function

$$L_c(x, \lambda) = f(x) + \lambda'(Ax - b) + \frac{c}{2} \|Ax - b\|^2$$

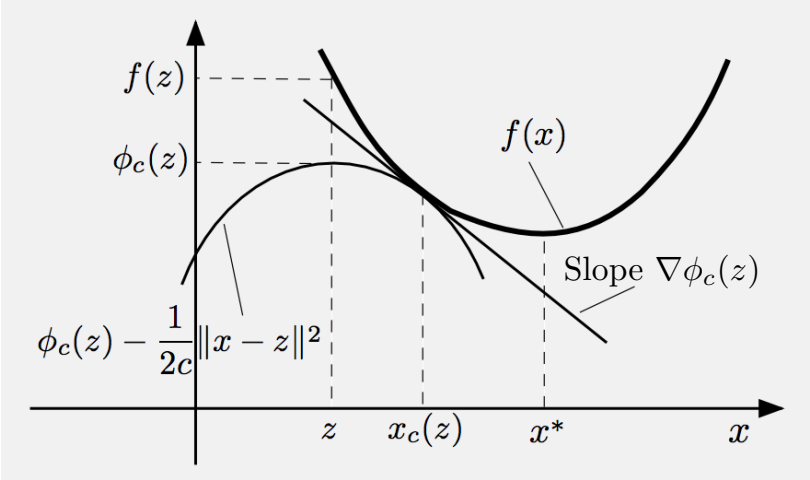
GRADIENT INTERPRETATION

- Back to the dual proximal algorithm and the dual update $\lambda_{k+1} = \frac{x_k - x_{k+1}}{c_k}$
- Proposition: λ_{k+1} can be viewed as a gradient,

$$\lambda_{k+1} = \frac{x_k - x_{k+1}}{c_k} = \nabla \phi_{c_k}(x_k),$$

where

$$\phi_c(z) = \inf_{x \in \Re^n} \left\{ f(x) + \frac{1}{2c}\|x - z\|^2 \right\}$$



- So the dual update $x_{k+1} = x_k - c_k \lambda_{k+1}$ can be viewed as a gradient iteration for minimizing $\phi_c(z)$ (which has the same minima as f).
- The gradient is calculated by the dual proximal minimization. Possibilities for faster methods (e.g., Newton, Quasi-Newton). Useful in augmented Lagrangian methods.

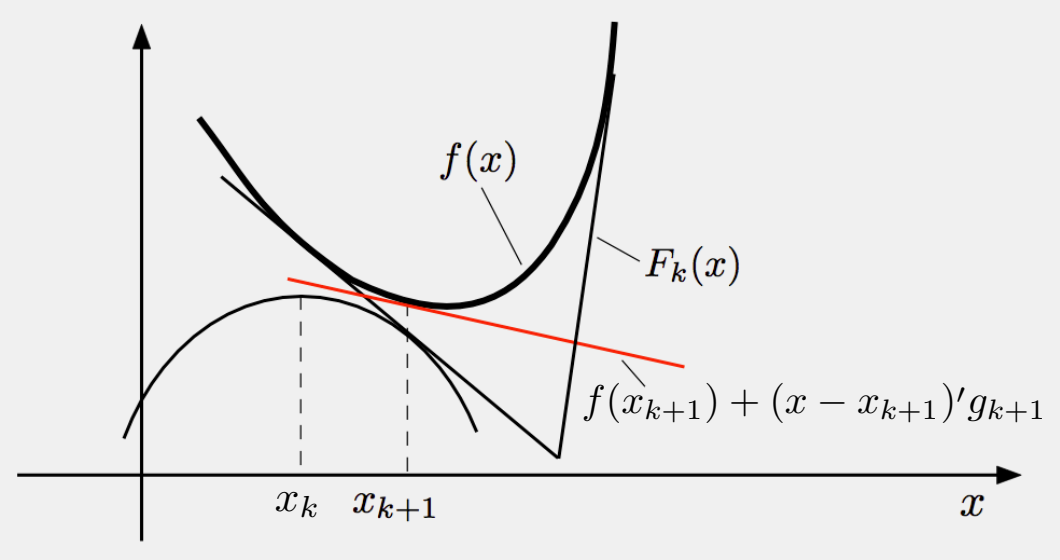
PROXIMAL CUTTING PLANE METHODS

- Same as proximal algorithm, but f is replaced by a cutting plane approximation F_k:

$$x_{k+1} \in \arg\min_{x \in X} \left\{ F_k(x) + \frac{1}{2c_k} \|x - x_k\|^2 \right\}$$

where

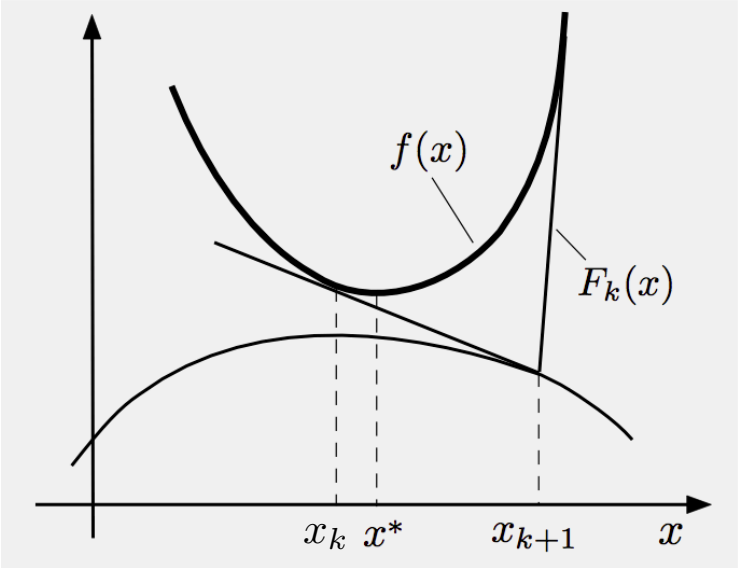
$$F_k(x) = \max\left\{ f(x_0) + (x - x_0)' g_0, \ldots, f(x_k) + (x - x_k)' g_k \right\}$$



- Main objective is to reduce instability ... but there are issues to contend with.

DRAWBACKS

- Stability issue:
  - For large enough c_k and polyhedral X, x_{k+1} is the exact minimum of F_k over X in a single minimization, so it is identical to the ordinary cutting plane method.



  - For small c_k convergence is slow.

- The number of subgradients used in F_k may become very large; the quadratic program may become very time-consuming.

- These drawbacks motivate algorithmic variants, called bundle methods.

BUNDLE METHODS I

- Replace f with a cutting plane approx. and change quadratic regularization more conservatively.
- A general form:

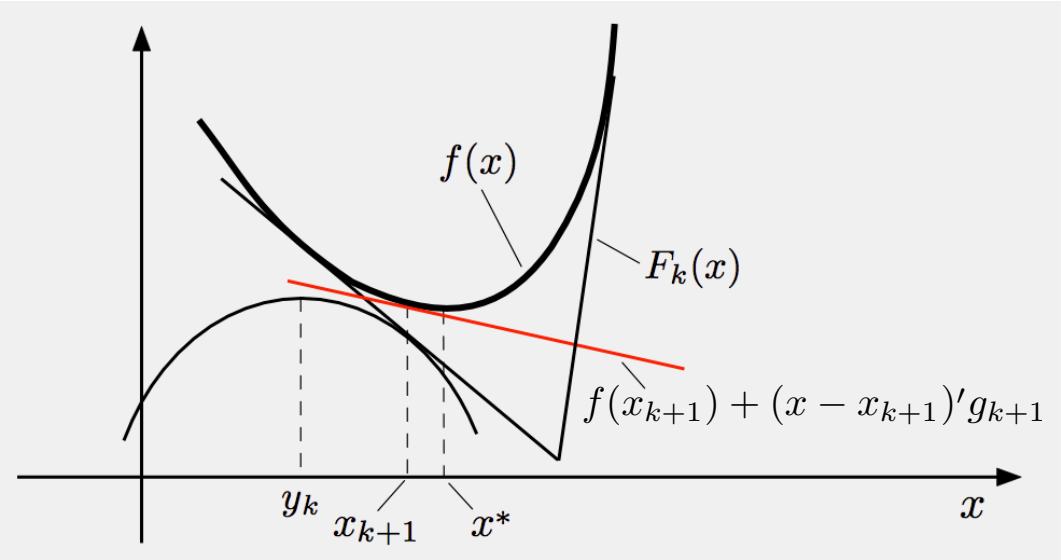
$$x_{k+1} \in \arg\min_{x \in X} \Big\{ F_k(x) + p_k(x) \Big\}$$

$$F_k(x) = \max\Big\{ f(x_0) + (x - x_0)' g_0, \ldots, f(x_k) + (x - x_k)' g_k \Big\}$$

$$p_k(x) = \frac{1}{2c_k} \|x - y_k\|^2$$

where c_k is a positive scalar parameter.

- We refer to $p_k(x)$ as the proximal term, and to its center y_k as the proximal center.



Change y_k in different ways => different methods.

BUNDLE METHODS II

- Allow a proximal center $y_k \neq x_k$:

$$x_{k+1} \in \arg\min_{x \in X} \Big\{ F_k(x) + p_k(x) \Big\}$$

$$F_k(x) = \max\Big\{ f(x_0) + (x - x_0)'g_0, \ldots, f(x_k) + (x - x_k)'g_k \Big\}$$

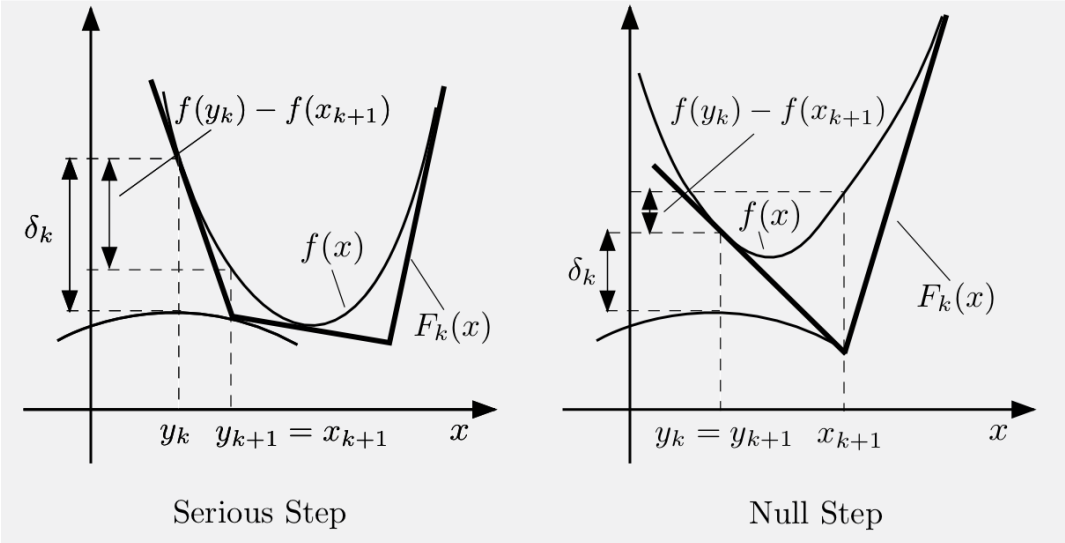
$$p_k(x) = \frac{1}{2c_k} \|x - y_k\|^2$$

- Null/Serious test for changing y_k

- Compare true cost f and proximal cost $F_k + p_k$ reduction in moving from y_k to x_{k+1}, i.e., for some fixed $\beta \in (0, 1)$

$$y_{k+1} = \begin{cases} x_{k+1} & \text{if } f(y_k) - f(x_{k+1}) \geq \beta \delta_k, \\ y_k & \text{if } f(y_k) - f(x_{k+1}) < \beta \delta_k, \end{cases}$$

$$\delta_k = f(y_k) - \Big(F_k(x_{k+1}) + p_k(x_{k+1}) \Big) > 0$$



PROXIMAL LINEAR APPROXIMATION

- Convex problem: Min $f : \Re^n \mapsto \Re$ over X.
- Proximal cutting plane method: Same as proximal algorithm, but f is replaced by a cutting plane approximation F_k:

$$x_{k+1} \in \arg\min_{x \in \Re^n} \left\{ F_k(x) + \frac{1}{2c_k} \|x - x_k\|^2 \right\}$$

$$\lambda_{k+1} = \frac{x_k - x_{k+1}}{c_k}$$

where $g_i \in \partial f(x_i)$ for $i \leq k$ and

$$F_k(x) = \max\Big\{ f(x_0) + (x - x_0)' g_0, \ldots, f(x_k) + (x - x_k)' g_k \Big\} + \delta_X(x)$$

- Proximal simplicial decomposition method (dual proximal implementation): Let $F_k^\star$ be the conjugate of F_k. Set

$$\lambda_{k+1} \in \arg\min_{\lambda \in \Re^n} \left\{ F_k^\star(\lambda) - x_k' \lambda + \frac{c_k}{2} \|\lambda\|^2 \right\}$$

$$x_{k+1} = x_k - c_k \lambda_{k+1}$$

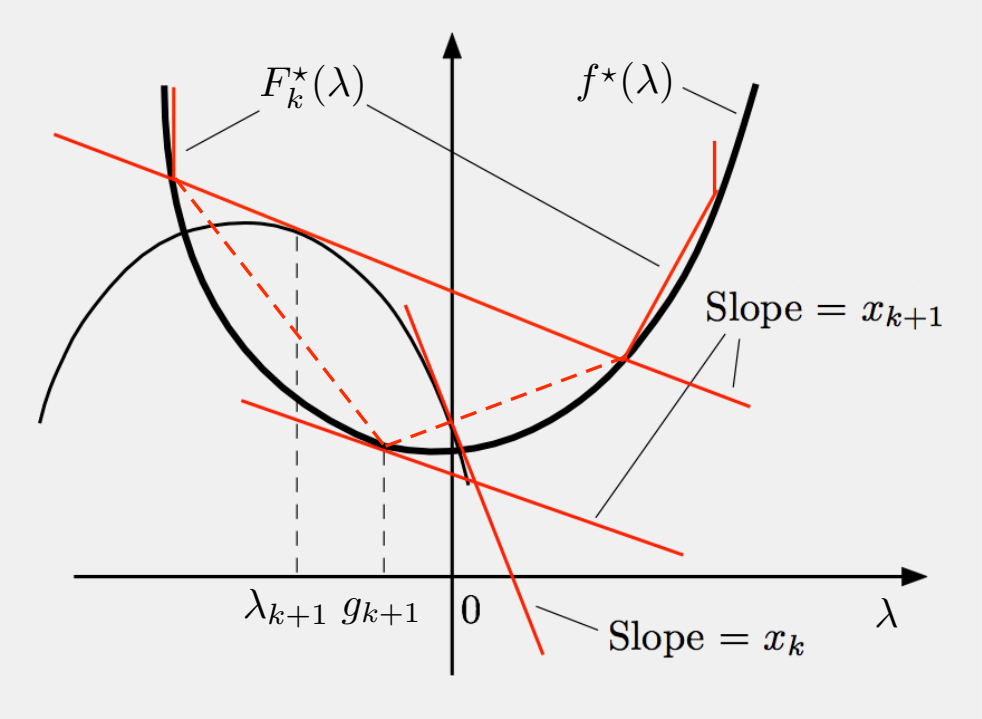
Obtain $g_{k+1} \in \partial f(x_{k+1})$, either directly or via

$$g_{k+1} \in \arg\max_{\lambda \in \Re^n} \left\{ x_{k+1}' \lambda - f^\star(\lambda) \right\}$$

- Add g_{k+1} to the outer linearization, or x_{k+1} to the inner linearization, and continue.

PROXIMAL SIMPLICIAL DECOMPOSITION

• It is a mathematical equivalent dual to the proximal cutting plane method.



• Here we use the conjugacy relation between outer and inner linearization.

• Versions of these methods where the proximal center is changed only after some "algorithmic progress" is made:

  – The outer linearization version is the (standard) bundle method.

  – The inner linearization version is an inner approximation version of a bundle method.

LECTURE 20

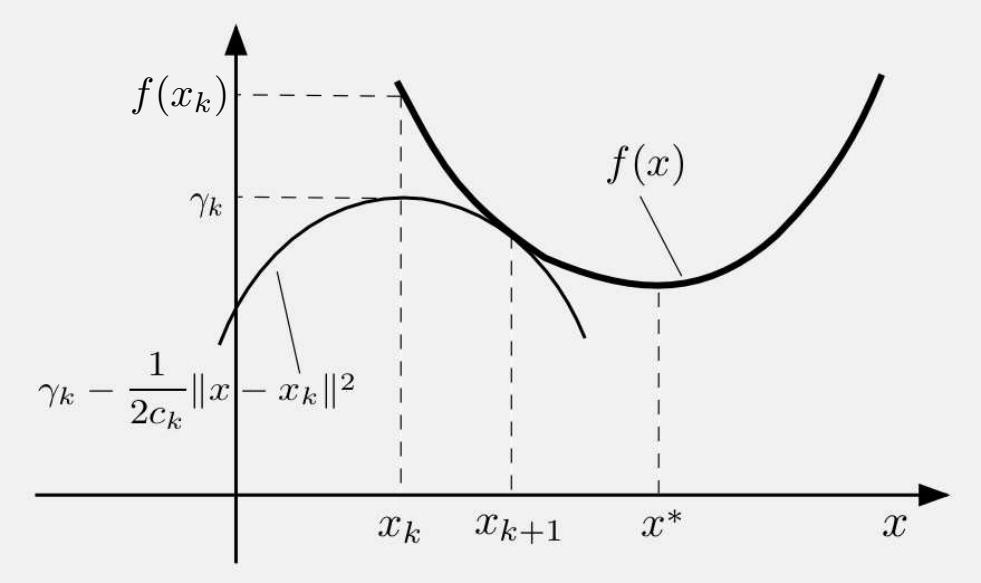
LECTURE OUTLINE

- Review of proximal and augmented Lagrangians
- Alternating direction methods of multipliers (ADMM)
- Applications of ADMM
- Extensions of proximal algorithm

********** References **********

- Bertsekas, D. P., and Tsitsiklis, J. N., 1989. Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, Englewood Cliffs, N. J.
- Eckstein, J., and Bertsekas, D. P., 1992. "On the Douglas-Rachford Splitting Method and the Proximal Point Algorithm for Maximal Monotone Operators," Math. Progr., Vol. 55, pp. 293-318.
- Eckstein, J., 2012. "Augmented Lagrangian and Alternating Direction Methods for Convex Optimization," Rutgers, Univ. Report
- Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J., 2011. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers, Now Publishers Inc.

RECALL PROXIMAL ALGORITHM



- Minimizes closed convex proper f:

$$x_{k+1} = \arg\min_{x \in \Re^n} \left\{ f(x) + \frac{1}{2c_k} \|x - x_k\|^2 \right\}$$

where x_0 is an arbitrary starting point, and $\{c_k\}$ is a positive parameter sequence.

- We have $f(x_k) \to f^*$. Also $x_k \to$ some minimizer of f, provided one exists.

- When applied with $f = -q$, where q is the dual function of a constrained optimization problem, we obtain the augmented Lagrangian method.

AUGMENTED LAGRANGIAN METHOD

- Consider the convex constrained problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in X, \quad Ax = b \end{aligned}$$

- Primal and dual functions:

$$p(u) = \inf_{\substack{x \in X \\ Ax - b = u}} f(x), \qquad q(\lambda) = \inf_{x \in X} \left\{ f(x) + \lambda'(Ax - b) \right\}$$

- Augmented Lagrangian function:

$$L_c(x, \lambda) = f(x) + \lambda'(Ax - b) + \frac{c}{2} \|Ax - b\|^2$$

- Augmented Lagrangian algorithm: Find

$$x_{k+1} \in \arg\min_{x \in X} L_{c_k}(x, \lambda_k)$$

and then set

$$\lambda_{k+1} = \lambda_k + c_k(Ax_{k+1} - b)$$

A DIFFICULTY WITH AUGM. LAGRANGIANS

- Consider the (Fenchel format) problem

$$\begin{aligned} &\text{minimize} \quad f_1(x) + f_2(z) \\ &\text{subject to} \quad x \in \Re^n,\ z \in \Re^m, \quad Ax = z, \end{aligned}$$

and its augmented Lagrangian function

$$L_c(x, z, \lambda) = f_1(x) + f_2(z) + \lambda'(Ax - z) + \frac{c}{2}\|Ax - z\|^2.$$

- The problem is separable in x and z, but $\|Ax - z\|^2$ couples x and x inconveniently.
- We may consider minimization by a block coordinate descent method:
  - Minimize $L_c(x, z, \lambda)$ over x, with z and λ held fixed.
  - Minimize $L_c(x, z, \lambda)$ over z, with x and λ held fixed.
  - Repeat many times, then update the multipliers, then repeat again.
- The ADMM does one minimization in x, then one minimization in z, before updating λ.

ADMM

- Start with some λ_0 and $c > 0$:

$$x_{k+1} \in \arg\min_{x \in \Re^n} L_c(x, z_k, \lambda_k),$$

$$z_{k+1} \in \arg\min_{z \in \Re^m} L_c(x_{k+1}, z, \lambda_k),$$

$$\lambda_{k+1} = \lambda_k + c(Ax_{k+1} - z_{k+1}).$$

- The penalty parameter c is kept constant in the ADMM (no compelling reason to change it).

- Strong convergence properties: $\{\lambda_k\}$ converges to optimal dual solution, and if $A'A$ is invertible, $\{x_k, z_k\}$ also converge to optimal primal solution.

- Big advantages:
  - x and z are decoupled in the minimization of $L_c(x, z, \lambda)$.
  - Very convenient for problems with special structures.
  - Has gained a lot of popularity for signal processing and machine learning problems.

- Not necessarily faster than augmented Lagrangian methods (many more iterations in λ are needed).

FAVORABLY STRUCTURED PROBLEMS I

- Additive cost problems:

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{m} f_i(x) \\ &\text{subject to} \quad x \in \cap_{i=1}^{m} X_i, \end{aligned}$$

where $f_i : \Re^n \mapsto \Re$ are convex functions and X_i are closed, convex sets.

- Feasibility problem: Given m closed convex sets $X_1, X_2, \ldots, X_m$ in $\Re^n$, find a point in $\cap_{i=1}^{m} X_i$.

- Problems involving ℓ_1 norms: A key fact is that proximal works well with ℓ_1. For any $\alpha > 0$ and $w = (w^1, \ldots, w^m) \in \Re^m$,

$$S(\alpha, w) \in \arg\min_{z \in \Re^m} \left\{ \|z\|_1 + \frac{1}{2\alpha} \|z - w\|^2 \right\},$$

is easily computed by the shrinkage operation:

$$S^i(\alpha, w) = \begin{cases} w^i - \alpha & \text{if } w^i > \alpha, \\ 0 & \text{if } |w^i| \leq \alpha, \\ w^i + \alpha & \text{if } w^i < -\alpha, \end{cases} \qquad i = 1, \ldots, m.$$

FAVORABLY STRUCTURED PROBLEMS II

- Basis pursuit:

$$\begin{aligned} &\text{minimize} && \|x\|_1 \\ &\text{subject to} && Cx = b, \end{aligned}$$

where $\|\cdot\|_1$ is the ℓ_1 norm in $\Re^n$, C is a given $m \times n$ matrix and b is a vector in $\Re^m$. Use $f_1 =$ indicator fn of $\{x \mid Cx = b\}$, and $f_2(z) = \|z\|_1$.

- ℓ_1 Regularization:

$$\begin{aligned} &\text{minimize} && f(x) + \gamma\|x\|_1 \\ &\text{subject to} && x \in \Re^n, \end{aligned}$$

where $f : \Re^n \mapsto (-\infty, \infty]$ is a closed proper convex function and γ is a positive scalar. Use $f_1 = f$, and $f_2(z) = \gamma\|z\|_1$.

- Least Absolute Deviations Problem:

$$\begin{aligned} &\text{minimize} && \|Cx - b\|_1 \\ &\text{subject to} && x \in \Re^n, \end{aligned}$$

where C is an $m \times n$ matrix, and $b \in \Re^m$ is a given vector. Use $f_1 = 0$, and $f_2(z) = \|z\|_1$.

SEPARABLE PROBLEMS I

- Consider a convex separable problem of the form

$$\text{minimize} \quad \sum_{i=1}^{m} f_i(x^i)$$

$$\text{subject to} \quad \sum_{i=1}^{m} A_i x^i = b, \qquad x^i \in X_i, \quad i = 1, \ldots, m,$$

- A plausible idea is the ADMM-like iteration

$$x_{k+1}^i \in \arg\min_{x^i \in X_i} L_c(x_{k+1}^1, \ldots, x_{k+1}^{i-1}, x^i, x_k^{i+1}, \ldots, x_k^m, \lambda_k),$$

$$\lambda_{k+1} = \lambda_k + c\left(\sum_{i=1}^{m} A_i x_{k+1}^i - b\right)$$

- For $m = 1$ it becomes the augmented Lagrangian method, for $m = 2$ it becomes the ADMM, and for $m > 2$ it maintains the attractive variable decoupling property of ADMM

- Unfortunately, it may not work for $m > 2$ (it does work but under restrictive assumptions)

- We will derive a similar but reliable version (a special case of ADMM for $m = 2$, from Bertsekas and Tsitsiklis 1989, Section 3.4).

SEPARABLE PROBLEMS II

- We reformulate the convex separable problem so it can be addressed by ADMM

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{m} f_i(x^i) \\ &\text{subject to} && A_i x^i = z^i, \quad x^i \in X_i, \qquad i = 1, \ldots, m, \\ & && \sum_{i=1}^{m} z^i = b, \end{aligned}$$

- The ADMM is given by

$$x_{k+1}^i \in \arg\min_{x^i \in X_i} \left\{ f_i(x^i) + (A_i x^i - z_k^i)' p_k^i + \frac{c}{2} \|A_i x^i - z_k^i\|^2 \right\},$$

$$z_{k+1} \in \arg\min_{\sum_{i=1}^m z^i = b} \left\{ \sum_{i=1}^{m} (A_i x_{k+1}^i - z^i)' p_k^i + \frac{c}{2} \|A_i x_{k+1}^i - z^i$$

$$p_{k+1}^i = p_k^i + c(A_i x_{k+1}^i - z_{k+1}^i),$$

where p_k^i is the multiplier of $A_i x^i = z^i$.

- A key fact is that all p_k^i, $i = 1, \ldots, m$, can be shown to be equal to a single vector λ_k, the multiplier of the constraint $\sum_{i=1}^m z^i = b$.

- This simplifies the algorithm.

PROXIMAL AS FIXED POINT ALGORITHM I

- Back to the proximal algorithm for minimizing closed convex $f : \Re^n \mapsto (-\infty, \infty]$.
- Proximal operator corresponding to c and f:

$$P_{c,f}(z) = \arg\min_{x \in \Re^n} \left\{ f(x) + \frac{1}{2c}\|x - z\|^2 \right\}, \qquad z \in \Re^n$$

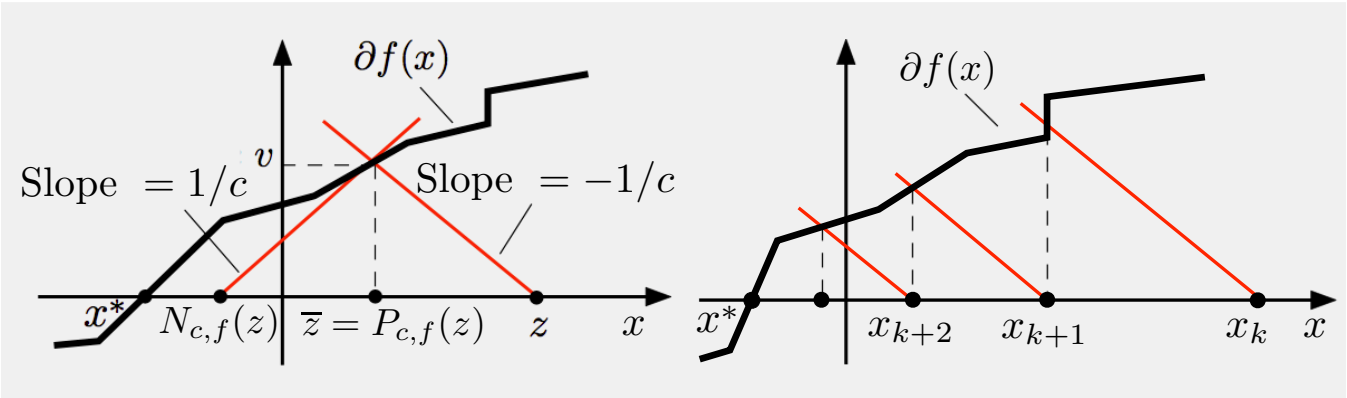
- The set of fixed points of $P_{c,f}$ coincides with the set of minima of f, and the proximal algorithm, written as

$$x_{k+1} = P_{c_k,f}(x_k),$$

may be viewed as a fixed point iteration.

- Decomposition:

$\overline{z} = P_{c,f}(z)$ iff $\overline{z} = z - cv$ for some $v \in \partial f(\overline{z})$



- Important mapping $N_{c,f}(z) = 2P_{c,f}(z) - z$

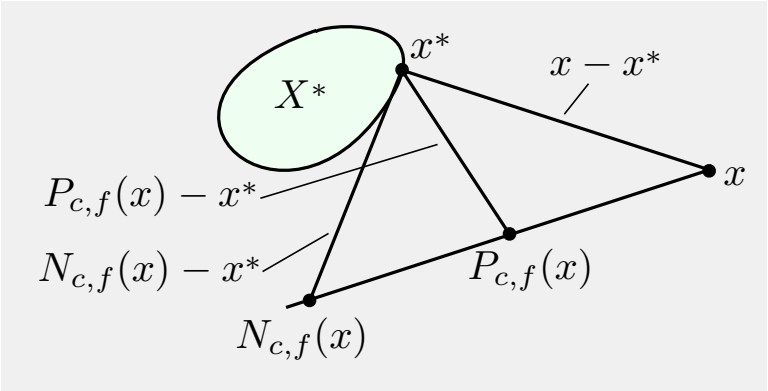
PROXIMAL AS FIXED POINT ALGORITHM II

- The mapping $N_{c,f} : \Re^n \mapsto \Re^n$ given by

$$N_{c,f}(z) = 2P_{c,f}(z) - z, \qquad z \in \Re^n,$$

is nonexpansive:

$$\big\|N_{c,f}(z_1) - N_{c,f}(z_2)\big\| \leq \|z_1 - z_2\|, \quad \forall\ z_1, z_2 \in \Re^n.$$



- The interpolated iteration

$$x_{k+1} = (1 - \alpha_k)x_k + \alpha_k N_{c,f}(x_k),$$

where $\alpha_k \in [\epsilon, 1-\epsilon]$ for some scalar $\epsilon > 0$, converges to a fixed point of $N_{c,f}$, provided $N_{c,f}$ has at least one fixed point.

- Extrapolation is more favorable

- ADMM and proximal belong to the same family of fixed point algorithms for finding a zero of a multivalued monotone operator (see refs).

LECTURE 21

LECTURE OUTLINE

- We enter a series of lectures on advanced topics
  - Gradient projection
  - Variants of gradient projection
  - Variants of proximal and combinations
  - Incremental subgradient and proximal methods
  - Coordinate descent methods
  - Interior point methods, etc
- Today's lecture on gradient projection
- Application to differentiable problems
- Iteration complexity issues

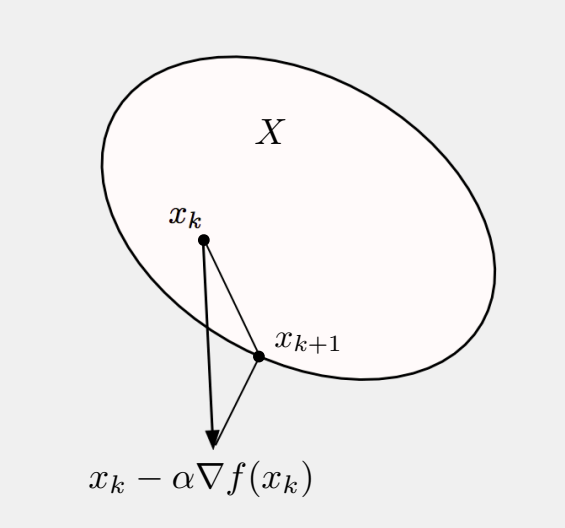
****

- Reference: The on-line chapter of the textbook

GRADIENT PROJECTION METHOD

- Let f be continuously differentiable, and X be closed convex.
- Gradient projection method:

$$x_{k+1} = P_X\big(x_k - \alpha_k \nabla f(x_k)\big)$$



- A specialization of subgradient method, but cost function descent comes into play
- $x_{k+1} - x_k$ is a feasible descent direction (by the projection theorem)
- $f(x_{k+1}) < f(x_k)$ if α_k: sufficiently small (unless x_k is optimal)
- α_k may be constant or chosen by cost descent-based stepsize rules

CONNECTION TO THE PROXIMAL ALGORITHM

- Linear approximation of f based on $\nabla f(x)$:

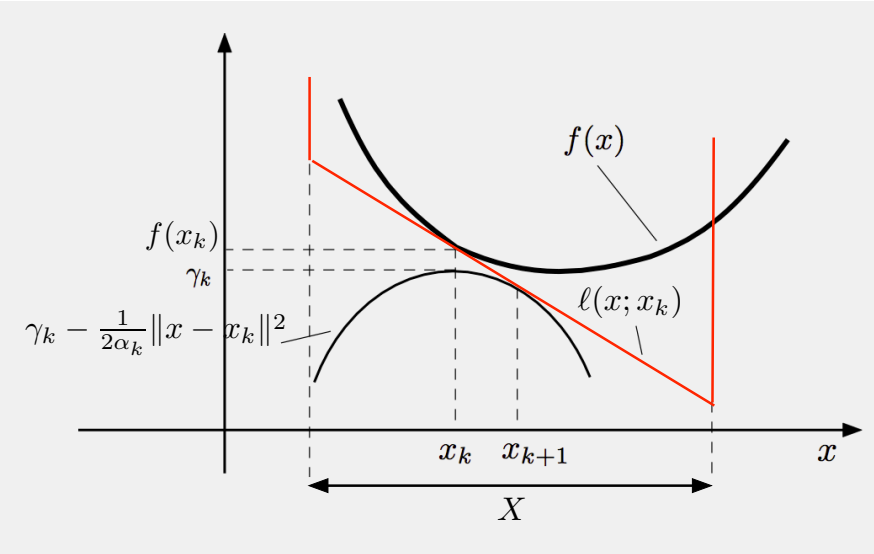
$$\ell(y;x) = f(x) + \nabla f(x)'(y-x), \qquad \forall\ x, y \in \Re^n$$

- For all $x \in X$ and $\alpha > 0$, we have

$$\frac{1}{2\alpha}\left\|y-\big(x-\alpha\nabla f(x)\big)\right\|^2 = \ell(y;x)+\frac{1}{2\alpha}\|y-x\|^2+\text{constant}$$

so

$$P_X\big(x-\alpha\nabla f(x)\big) \in \arg\min_{y\in X}\left\{\ell(y;x)+\frac{1}{2\alpha}\|y-x\|^2\right\}$$



- Three-term inequality holds: For all $y \in \Re^n$,

$$\|x_{k+1}-y\|^2 \le \|x_k-y\|^2-2\alpha_k\big(\ell(x_{k+1};x_k)-\ell(y;x_k)\big)-\|x_k-x_{k+1}\|^2$$

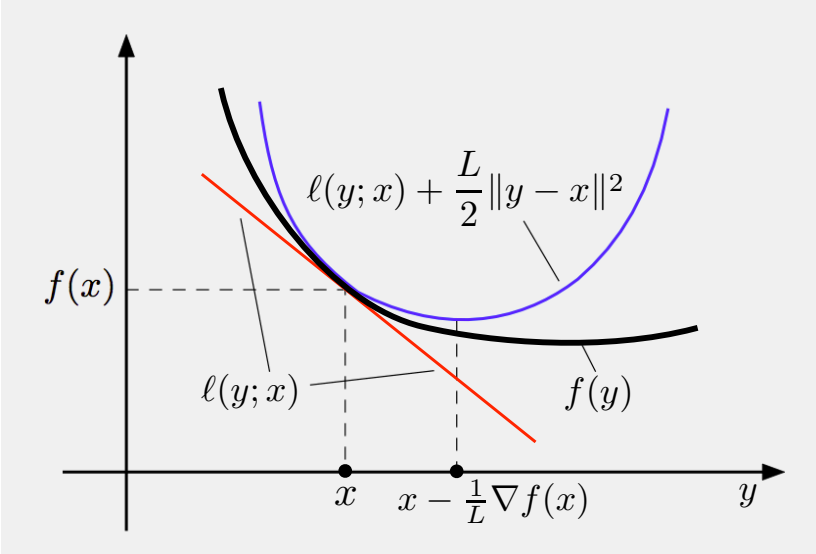
CONSTANT STEPSIZE - DESCENT LEMMA

- Consider constant α_k: $x_{k+1} = P_X\big(x_k - \alpha \nabla f(x_k)\big)$
- We need the gradient Lipschitz assumption

$$\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\,\|x - y\|, \qquad \forall\ x, y \in X$$

- Descent Lemma: For all $x, y \in X$,

$$f(y) \leq \ell(y; x) + \frac{L}{2}\|y - x\|^2$$



- Proof idea: The Lipschitz constant L serves as an upper bound to the "curvature" of f along directions, so $\frac{L}{2}\|y-x\|^2$ is an upper bound to $f(y) - \ell(y; x)$.

CONSTANT STEPSIZE - CONVERGENCE RESULT

- Assume the gradient Lipschitz condition, and $\alpha \in (0, 2/L)$ (no convexity of f). Then $f(x_k) \downarrow f^*$ and every limit point of $\{x_k\}$ is optimal.

Proof: From the projection theorem, we have

$$\big(x_k - \alpha \nabla f(x_k) - x_{k+1}\big)'(x - x_{k+1}) \leq 0, \qquad \forall\, x \in X,$$

so by setting $x = x_k$,

$$\nabla f(x_k)'(x_{k+1} - x_k) \leq -\frac{1}{\alpha} \left\|x_{k+1} - x_k\right\|^2$$

- Using this relation and the descent lemma,

$$\begin{aligned} f(x_{k+1}) &\leq \ell(x_{k+1}; x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) + \nabla f(x_k)'(x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &\leq f(x_k) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|x_{k+1} - x_k\|^2 \end{aligned}$$

so $\alpha \in (0, 2/L)$ reduces the cost function value.

- If $\alpha \in \big(0, 2/L\big)$ and $\overline{x}$ is the limit of a subsequence $\{x_k\}_{\mathcal{K}}$, then $f(x_k) \downarrow f(\overline{x})$, so $\|x_{k+1} - x_k\| \to 0$. This implies $P_X\big(\overline{x} - \alpha \nabla f(\overline{x})\big) = \overline{x}$. **Q.E.D.**

STEPSIZE RULES

- Eventually constant stepsize. Deals with the case of an unknown Lipschitz constant L. Start with some $\alpha > 0$, and keep using α as long as

$$f(x_{k+1}) \leq \ell(x_{k+1}; x_k) + \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2$$

is satisfied (this guarantees cost descent). When this condition is violated at some iteration, we reduce α by a certain factor, and repeat. (Satisfied once $\alpha \leq 1/L$, by the descent lemma.)

- A diminishing stepsize α_k, satisfying

$$\lim_{k\to\infty} \alpha_k = 0, \qquad \sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

Does not require Lipschitz condition or differentiability of f, only convexity of f.

- Stepsize reduction and line search rules - Armijo rules. These rules are based on cost function descent, and ensure that through some form of line search, we find α_k such that $f(x_{k+1}) < f(x_k)$, unless x_k is optimal. Do not require Lipschitz condition, only differentiability of f.

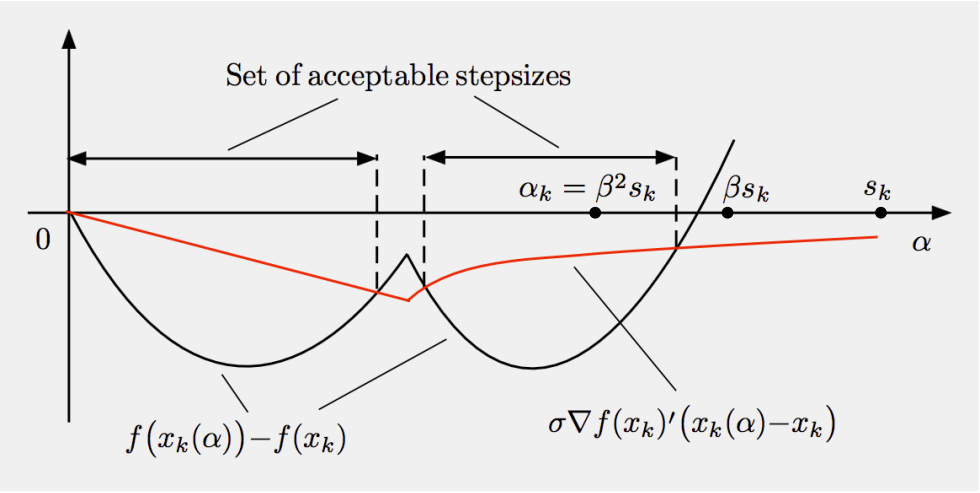
ARMIJO STEPSIZE RULES

- Search along the projection arc: $\alpha_k = \beta^{m_k} s$, where $s > 0$ and $\beta \in (0,1)$ are fixed scalars, and m_k is the first integer m such that

$$f(x_k) - f\big(x_k(\beta^m s)\big) \geq \sigma \nabla f(x_k)' \big(x_k - x_k(\beta^m s)\big),$$

with $\sigma \in (0,1)$ being some small constant, and

$$x_k(\alpha) = P_X\big(x_k - \alpha \nabla f(x_k)\big)$$



- Similar rule searches along the feasible direction

CONVERGENCE RATE - $\alpha_K \equiv 1/L$

• Assume f: convex, the Lipschitz condition, $X^* \neq \emptyset$, and the eventually constant stepsize rule. Denote $d(x_k) = \min_{x^* \in X^*} \|x_k - x^*\|$. Then

$$\lim_{k \to \infty} d(x_k) = 0, \qquad f(x_k) - f^* \leq \frac{L d(x_0)^2}{2k}$$

Proof: Let $x^* \in X^*$ be such that $\|x_0 - x^*\| = d(x_0)$. Using the descent lemma and the three-term inequality,

$$\begin{aligned} f(x_{k+1}) &\leq \ell(x_{k+1}; x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &\leq \ell(x^*; x_k) + \frac{L}{2}\|x^* - x_k\|^2 - \frac{L}{2}\|x^* - x_{k+1}\|^2 \\ &\leq f(x^*) + \frac{L}{2}\|x^* - x_k\|^2 - \frac{L}{2}\|x^* - x_{k+1}\|^2 \end{aligned}$$

Let $e_k = f(x_k) - f(x^*)$ and note that $e_k \downarrow$. Then

$$\frac{L}{2}\|x^* - x_{k+1}\|^2 \leq \frac{L}{2}\|x^* - x_k\|^2 - e_{k+1}$$

Use this relation with $k = k-1, k-2, \ldots$, and add

$$0 \leq \frac{L}{2}\|x^* - x_{k+1}\|^2 \leq \frac{L}{2} d(x_0)^2 - (k+1) e_{k+1}$$

GENERALIZATION - EVENTUALLY CONST. α_K

- Assume f: convex, the Lipschitz condition, $X^* \neq \emptyset$, and any stepsize rule such that

$$\alpha_k \downarrow \overline{\alpha},$$

for some $\overline{\alpha} > 0$, and for all k,

$$f(x_{k+1}) \leq \ell(x_{k+1}; x_k) + \frac{1}{2\alpha_k}\|x_{k+1} - x_k\|^2.$$

Denote $d(x_k) = \min_{x^* \in X^*} \|x_k - x^*\|$. Then

$$\lim_{k\to\infty} d(x_k) = 0, \qquad f(x_k) - f^* \leq \left(\frac{d(x_0)^2}{2\overline{\alpha}\, k}\right)$$

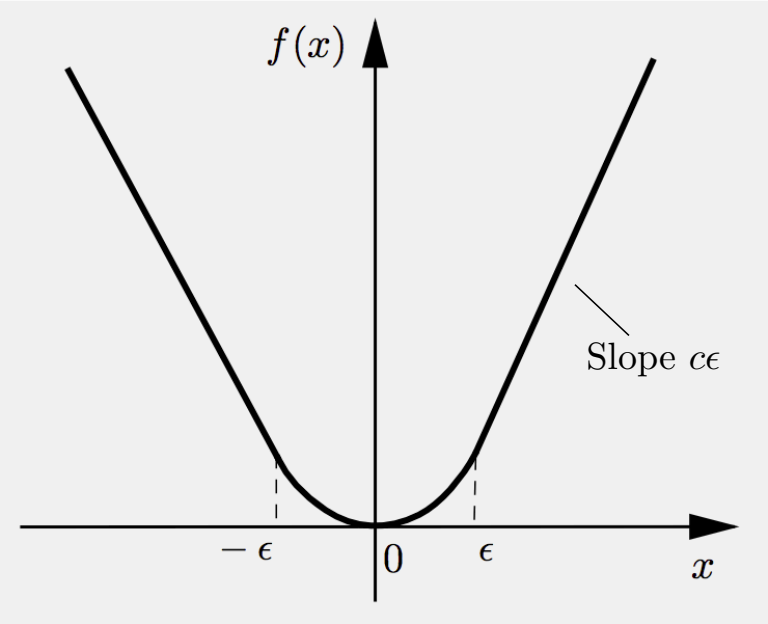
Proof: Show that

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2\alpha_k}\|x_{k+1} - x_k\|^2,$$

and generalize the preceding proof. **Q.E.D.**

- Applies to eventually constant stepsize rule.
- Error complexity $O(1/k)$, (k iterations produce $O(1/k)$ cost error), i.e., $\min_{\ell \leq k} f(x_\ell) \leq f^* + \frac{\text{const}}{k}$
- Iteration complexity $O(1/\epsilon)$, ($O(1/\epsilon)$ iterations produce ϵ cost error), i.e., $\min_{k \leq \frac{\text{const}}{\epsilon}} f(x_k) \leq f^* + \epsilon$

SHARPNESS OF COMPLEXITY ESTIMATE



- Unconstrained minimization of

$$f(x) = \begin{cases} \frac{c}{2}|x|^2 & \text{if } |x| \leq \epsilon, \\ c\epsilon|x| - \frac{c\epsilon^2}{2} & \text{if } |x| > \epsilon \end{cases}$$

- With stepsize $\alpha = 1/L = 1/c$ and any $x_k > \epsilon$,

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) = x_k - \frac{1}{c}c\,\epsilon = x_k - \epsilon$$

- The number of iterations to get within an ϵ-neighborhood of $x^* = 0$ is $|x_0|/\epsilon$.

- The number of iterations to get to within ϵ of $f^* = 0$ is proportional to $1/\epsilon$ for large x_0.

LECTURE 22

LECTURE OUTLINE

- Gradient projection method
- Iteration complexity issues
- Gradient projection with extrapolation
- Proximal gradient method

****

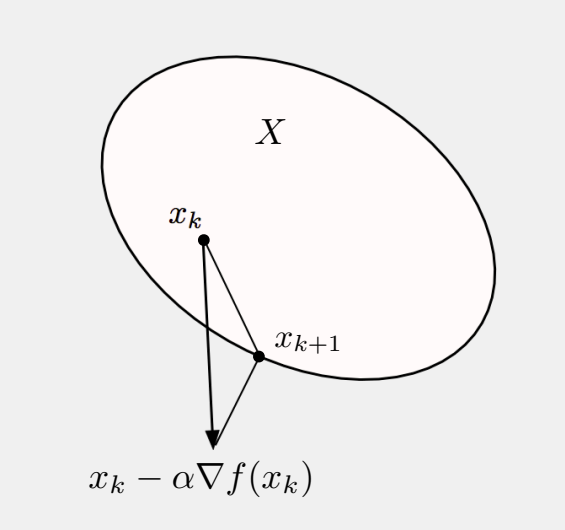
References:

- The on-line chapter of the textbook
- Beck, A., and Teboulle, M., 2010. "Gradient-Based Algorithms with Applications to Signal Recovery Problems, in Convex Optimization in Signal Processing and Communications (Y. Eldar and D. Palomar, eds.), Cambridge University Press, pp. 42-88.
- J. Lee, Y. Sun, M. Saunders, "Proximal Newton-Type Methods for Convex Optimization," NIPS, 2012.

REVIEW OF GRADIENT PROJECTION METHOD

- Let f be continuously differentiable, and X be closed convex.
- Gradient projection method:

$$x_{k+1} = P_X\big(x_k - \alpha_k \nabla f(x_k)\big)$$

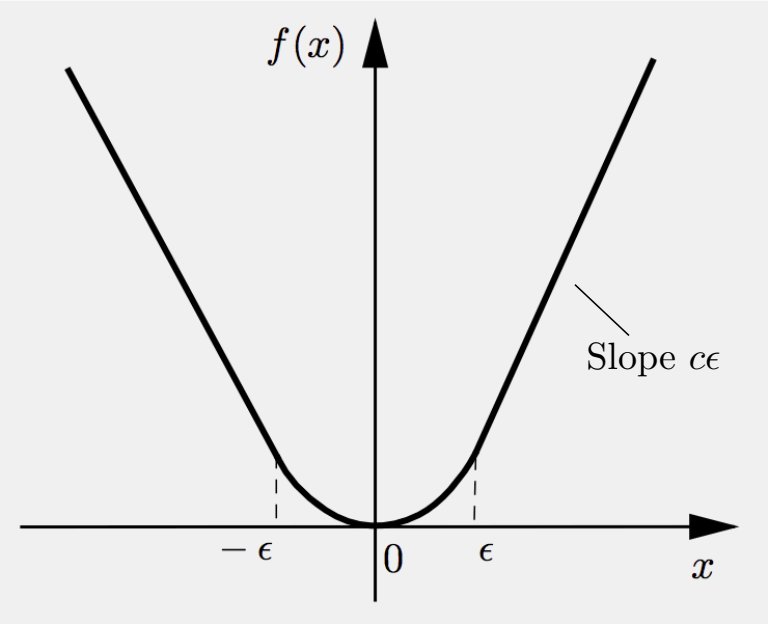


- α_k may be constant or chosen by cost descent-based stepsize rules
- Under gradient Lipschitz assumption

$$\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\,\|x - y\|, \qquad \forall\ x, y \in X$$

iteration complexity $O(1/\epsilon)$, ($O(1/\epsilon)$ iterations for ϵ cost error), i.e., $\min_{k \leq \frac{\text{const}}{\epsilon}} f(x_k) \leq f^* + \epsilon$

SHARPNESS OF COMPLEXITY ESTIMATE



- Unconstrained minimization of

$$f(x) = \begin{cases} \frac{1}{2}|x|^2 & \text{if } |x| \leq \epsilon, \\ \epsilon|x| - \frac{\epsilon^2}{2} & \text{if } |x| > \epsilon \end{cases}$$

- With stepsize $\alpha = 1/L = 1$ and any $x_k > \epsilon$,

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) = x_k - \epsilon$$

- The number of iterations to get within an ϵ-neighborhood of $x^* = 0$ is $|x_0|/\epsilon$.

- The number of iterations to get to within ϵ of $f^* = 0$ is proportional to $1/\epsilon$ for large x_0.

EXTRAPOLATION VARIANTS

- An old method for unconstrained optimization, known as the *heavy-ball* method or gradient method with *momentum*:

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}),$$

where $x_{-1} = x_0$ and β is a scalar with $0 < \beta < 1$.

- A variant for constrained problems separates the extrapolation and the gradient steps:

$$y_k = x_k + \beta(x_k - x_{k-1}), \quad \text{(extrapolation step)},$$
$$x_{k+1} = P_X\big(y_k - \alpha \nabla f(y_k)\big), \quad \text{(grad. projection step)}.$$

- When applied to the preceding example, the method converges to the optimum, and reaches a neighborhood of the optimum more quickly

- However, the method still has an $O(1/k)$ error complexity, since for $x_0 >> 1$, we have

$$x_{k+1} - x_k = \beta(x_k - x_{k-1}) - \epsilon$$

so $x_{k+1} - x_k \approx \epsilon/(1-\beta)$, and the number of iterations needed to obtain $x_k < \epsilon$ is $O\big((1-\beta)/\epsilon\big)$.

OPTIMAL COMPLEXITY ALGORITHM

- Surprisingly with a proper more vigorous extrapolation $\beta_k \to 1$ in the extrapolation scheme

$$y_k = x_k + \beta_k(x_k - x_{k-1}), \quad \text{(extrapolation step)},$$
$$x_{k+1} = P_X\Big(y_k - \frac{1}{L}\nabla f(y_k)\Big), \quad \text{(grad. projection step)},$$

the method has iteration complexity $O\big(\sqrt{L/\epsilon}\big)$. (Also with "eventually constant" rule for α.)

- Choices that work

$$\beta_k = \frac{\theta_k(1-\theta_{k-1})}{\theta_{k-1}}$$

where $\{\theta_k\}$ satisfies $\theta_0 = \theta_1 \in (0,1]$, and

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}, \qquad \theta_k \leq \frac{2}{k+2}$$

- One possible choice is

$$\beta_k = \begin{cases} 0 & \text{if } k = 0, \\ \frac{k-1}{k+2} & \text{if } k \geq 1, \end{cases} \qquad \theta_k = \begin{cases} 1 & \text{if } k = -1, \\ \frac{2}{k+2} & \text{if } k \geq 0. \end{cases}$$

- Highly unintuitive. Good practical performance reported.

EXTENSION TO NONDIFFERENTIABLE CASE

- Consider the nondifferentiable problem of minimizing convex function $f : \Re^n \mapsto \Re$ over a closed convex set X.
- "Smooth" f, i.e., approximate it with a differentiable function f_ϵ with Lipschitz constant $O(1/\epsilon)$ by using a proximal minimization scheme.
- The smoothed function satisfies

$$f_\epsilon(x) \leq f(x) \leq f_\epsilon(x) + O(\epsilon)$$

- Apply optimal complexity gradient projection method with extrapolation. Then an $O(1/\epsilon)$ complexity algorithm is obtained.
- Can be shown that this complexity bound is sharp.
- Improves on the subgradient complexity bound by an ϵ factor.
- Limited practical experience with such methods.

CRITIQUE OF THE OPTIMAL ALGORITHM

- Requires gradient Lipschitz assumption
- Chooses the stepsize α_k in the basis of the worst possible curvature information (same Lipschitz constant assumed in all directions).
- Compares well relative to competitors for some difficult problems (singular Hessian, but under Lipschitz gradient assumption).
- Not so well for other difficult problems (Lipschitz gradient assumption not holding) or easier problems (nonsingular Hessian) for which it has to compete with conjugate gradient and quasi-Newton methods
- A weak point: Cannot take advantage of special structure, e.g., there are no incremental versions.
- A strong point: Its favorable complexity estimate carries over to combinations with proximal algorithms.

PROXIMAL GRADIENT METHOD

- Minimize $f(x) + h(x)$ over $x \in X$, where X: closed convex, f, h: convex, f is differentiable.
- Proximal gradient method:

$$x_{k+1} \in \arg\min_{x \in X} \left\{ \ell(x; x_k) + h(x) + \frac{1}{2\alpha} \|x - x_k\|^2 \right\}$$

where $\ell(x; x_k) = f(x_k) + \nabla f(x_k)'(x - x_k)$

- Equivalent definition of proximal gradient:

$$z_k = x_k - \alpha \nabla f(x_k)$$

$$x_{k+1} \in \arg\min_{x \in X} \left\{ h(x) + \frac{1}{2\alpha} \|x - z_k\|^2 \right\}$$

- Simplifies the implementation of proximal, by using gradient iteration to deal with the case of an inconvenient component f
- Important example: h is the ℓ_1 norm - use the shrinkage operation to simplify the proximal
- The gradient projection and extrapolated variant analysis carries through, with the same iteration complexity

PROXIMAL GRADIENT METHOD ANALYSIS

- Recall descent lemma: For all $x, y \in X$

$$f(y) \leq \ell(y; x) + \frac{L}{2}\|y - x\|^2$$

where

$$\ell(y; x) = f(x) + \nabla f(x)'(y - x), \qquad \forall\ x, y \in \Re^n$$

- Recall three-term inequality: For all $y \in \Re^n$,

$$\begin{aligned}\|x_{k+1} - y\|^2 \leq\ & \|x_k - y\|^2 \\ & - 2\alpha_k\big(\ell(x_{k+1}; x_k) + h(x_{k+1}) - \ell(y; x_k) - h(y)\big) \\ & - \|x_k - x_{k+1}\|^2\end{aligned}$$

- Eventually constant stepsize rule: Keep using same α, as long as

$$f(x_{k+1}) \leq \ell(x_{k+1}; x_k) + \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2 \qquad (1)$$

- As soon as this condition is violated, reduce α by a certain factor, and repeat the iteration as many times as is necessary for Eq. (1) to hold.

RATE OF CONVERGENCE RESULT

• Assume ∇f satisfies the Lipschitz condition and the set of minima X^* of f over X is nonempty. If $\{x_k\}$ is a sequence generated by the proximal gradient method using any stepsize rule such that

$$\alpha_k \downarrow \overline{\alpha},$$

for some $\overline{\alpha} > 0$, and for all k,

$$f(x_{k+1}) \leq \ell(x_{k+1}; x_k) + \frac{1}{2\alpha_k} \|x_{k+1} - x_k\|^2,$$

then $\lim_{k\to\infty} d(x_k) = 0$, and

$$f(x_k) + h(x_k) - \min_{x\in X} \Big\{ f(x) + h(x) \Big\} \leq \frac{\overline{\alpha}\, d(x_0)^2}{2k}, \quad \forall\ k,$$

where

$$d(x) = \min_{x^*\in X^*} \|x - x^*\|, \qquad x \in \Re^n$$

SCALED PROXIMAL GRADIENT METHODS

- Idea: Instead of gradient, use scaled gradient, quasi-Newton, or Newton:

$$x_{k+1} \in \arg\min_{x \in X} \left\{ \ell(x; x_k) + h(x) + \tfrac{1}{2}(x - x_k)' H_k (x - x_k) \right\},$$

where H_k is a positive definite symmetric matrix.

- Can use $H_k = \nabla^2 f(x_k)$ (fast convergence) but the proximal minimization may become complicated.

- Lots of room for new methods ...

LECTURE 23

LECTURE OUTLINE

- Incremental methods
- Review of large sum problems
- Review of incremental gradient methods
- Incremental subgradient-proximal methods
- Convergence analysis
- Cyclic and randomized component selection

****

- References:

(1) D. P. Bertsekas, "Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey", Lab. for Information and Decision Systems Report LIDS-P-2848, MIT, August 2010.

(2) Published versions in Math. Programming J., and the edited volume "Optimization for Machine Learning," by S. Sra, S. Nowozin, and S. J. Wright, MIT Press, Cambridge, MA, 2012.

LARGE SUM PROBLEMS

- Minimize over $X \subset \Re^n$

$$f(x) = \sum_{i=1}^{m} f_i(x), \qquad m \text{ is very large,}$$

where X, f_i are convex. Some examples:

- Dual cost of a separable problem - Lagrangian relaxation, integer programming.

- Data analysis/machine learning: x is parameter vector of a model; each f_i corresponds to error between data and output of the model.
  - ℓ_1-regularization (least squares plus ℓ_1 penalty):

$$\min_x \gamma \sum_{j=1}^{n} |x^j| + \sum_{i=1}^{m} (c_i' x - d_i)^2$$

  - Classification (logistic regression, support vector machines)
  - Max-likelihood

- Min of an expected value $\min_x E\{F(x,w)\}$ - stochastic programming:

$$\min_x \left[F_1(x) + E_w \left\{ \min_y F_2(x,y,w) \right\} \right]$$

- More (many constraint problems, etc ...)

INCREMENTAL GRADIENT METHOD

- Problem: Minimization of $f(x) = \sum_{i=1}^{m} f_i(x)$ over a closed convex set X (f_i differentiable).
- Operates in cycles: If x_k is the vector obtained after k cycles, the vector x_{k+1} obtained after one more cycle is $x_{k+1} = \psi_{m,k}$, where $\psi_{0,k} = x_k$, and

$$\psi_{i,k} = P_X\big(\psi_{i-1,k} - \alpha_k \nabla f_{i,k}(\psi_{i-1,k})\big), \qquad i = 1, \ldots, m$$

- Does NOT compute the (expensive) gradient of f, which is $\sum_i \nabla f_i$.
- Interesting issues of ordering the processing of components.
- Randomization of selection of component f_i is possible. Connection with stochastic gradient method.
- Diminishing stepsize needed for convergence.
- Example: Consider

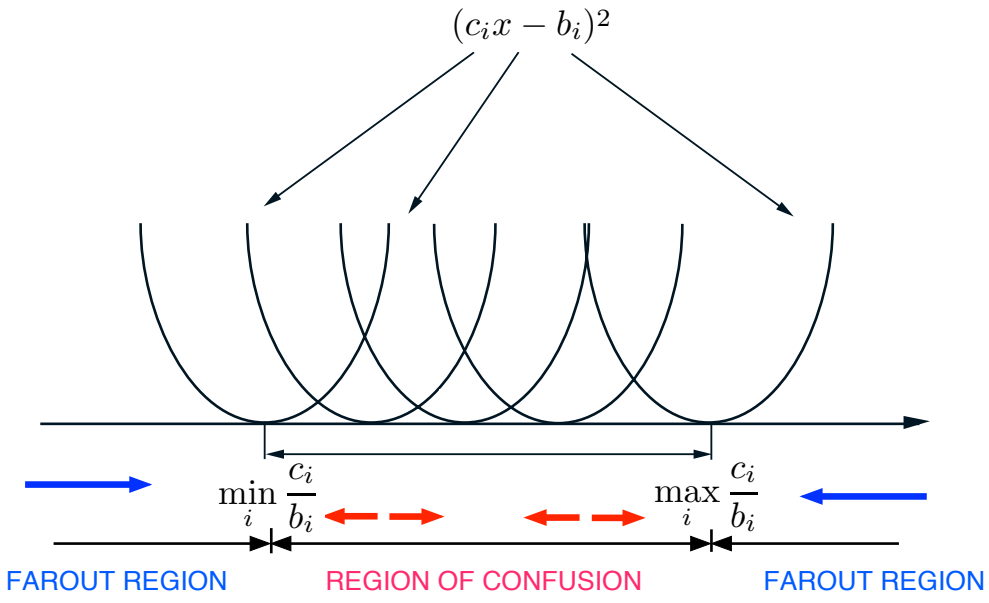
$$\min_{x \in \Re} \frac{1}{2}\big\{(1-x)^2 + (1+x)^2\big\}$$

For a constant stepsize the incremental gradient method oscillates.

COMPARE W/ NONINCREMENTAL GRADIENT

- Two complementary performance issues:
  - Progress when far from convergence. Here the incremental method can be much faster.
  - Progress when close to convergence. Here the incremental method can be inferior.
- Example: Scalar case

$$f_i(x) = \tfrac{1}{2}(c_i x - b_i)^2, \qquad x \in \Re$$



- Interesting issues of batching/shaping the region of confusion.
- Hybrids between incremental and nonincremental gradient methods. Aggregated gradient method.

INCREMENTAL SUBGRADIENT METHODS

- Problem: Minimize

$$f(x) = \sum_{i=1}^{m} f_i(x)$$

over a closed convex set X, where $f_i : \Re^n \mapsto \Re$ are convex, and possibly nondifferentiable.

- We first consider incremental subgradient methods which move x along a subgradient $\tilde{\nabla} f_i$ of a component function f_i.

- At iteration k select a component i_k and set

$$x_{k+1} = P_X\big(x_k - \alpha_k \tilde{\nabla} f_{i_k}(x_k)\big),$$

with $\tilde{\nabla} f_{i_k}(x_k)$ being a subgradient of f_{i_k} at x_k.

- Motivation is faster convergence. A cycle can make much more progress than a subgradient iteration with essentially the same computation.

CONVERGENCE: CYCLIC ORDER

- Algorithm

$$x_{k+1} = P_X\Big(x_k - \alpha_k \tilde{\nabla} f_{i_k}(x_k)\Big)$$

- Assume all subgradients generated by the algorithm are bounded: $\|\tilde{\nabla} f_{i_k}(x_k)\| \leq c$ for all k
- Assume components are chosen for iteration in cyclic order, and stepsize is constant within a cycle of iterations (for all k with $i_k = 1$ we have $\alpha_k = \alpha_{k+1} = \cdots = \alpha_{k+m-1}$)
- Key inequality: For all $y \in X$ and all k that mark the beginning of a cycle

$$\|x_{k+m} - y\|^2 \leq \|x_k - y\|^2 - 2\alpha_k\Big(f(x_k) - f(y)\Big) + \alpha_k^2 m^2 c^2$$

Progress if $-2\alpha_k\Big(f(x_k) - f(y)\Big) + \alpha_k^2 m^2 c^2 < 0$.

- Result for a constant stepsize $\alpha_k \equiv \alpha$:

$$\liminf_{k\to\infty} f(x_k) \leq f^* + \alpha\frac{m^2 c^2}{2}$$

- Convergence for $\alpha_k \downarrow 0$ with $\sum_{k=0}^{\infty} \alpha_k = \infty$.

CONVERGENCE: RANDOMIZED ORDER

- Algorithm

$$x_{k+1} = P_X\left(x_k - \alpha_k \tilde{\nabla} f_{i_k}(x_k)\right)$$

- Assume component i_k chosen for iteration in randomized order (independently with equal probability).
- Assume all subgradients generated by the algorithm are bounded: $\|\tilde{\nabla} f_{i_k}(x_k)\| \leq c$ for all k.
- Result for a constant stepsize $\alpha_k \equiv \alpha$:

$$\liminf_{k\to\infty} f(x_k) \leq f^* + \alpha \frac{mc^2}{2}$$

(with probability 1) - improvement by a factor m over the cyclic order case.

- Convergence for $\alpha_k \downarrow 0$ with $\sum_{k=0}^{\infty} \alpha_k = \infty$. (with probability 1). Use of the supermartingale convergence theorem.
- In practice, randomized stepsize and variations (such as randomization of the order within a cycle at the start of a cycle) often work much faster.

SUBGRADIENT-PROXIMAL CONNECTION

- Key Connection: The proximal iteration

$$x_{k+1} = \arg\min_{x \in X} \left\{ f(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}$$

can be written as

$$x_{k+1} = P_X \left(x_k - \alpha_k \tilde{\nabla} f(x_{k+1}) \right)$$

where $\tilde{\nabla} f(x_{k+1})$ is some subgradient of f at x_{k+1}.

- Consider an incremental proximal iteration for $\min_{x \in X} \sum_{i=1}^{m} f_i(x)$

$$x_{k+1} = \arg\min_{x \in X} \left\{ f_{i_k}(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}$$

- Motivation: Proximal methods are more "stable" than subgradient methods.

- Drawback: Proximal methods require special structure to avoid large overhead.

- This motivates a combination of incremental subgradient and proximal (split iteration, similar to proximal gradient).

INCR. SUBGRADIENT-PROXIMAL METHODS

- Consider the problem

$$\min_{x\in X} F(x) \stackrel{\text{def}}{=} \sum_{i=1}^{m} F_i(x)$$

where for all i,

$$F_i(x) = f_i(x) + h_i(x)$$

X, f_i and h_i are convex.

- Consider a combination of subgradient and proximal incremental iterations

$$z_k = \arg\min_{x\in X} \left\{ f_{i_k}(x) + \frac{1}{2\alpha_k}\|x - x_k\|^2 \right\}$$

$$x_{k+1} = P_X\left(z_k - \alpha_k \tilde{\nabla} h_{i_k}(z_k)\right)$$

- Idea: Handle "favorable" components f_i with the more stable proximal iteration; handle other components h_i with subgradient iteration.

- Variations:
  - Min. over $\Re^n$ (rather than X) in proximal
  - Do the subgradient without projection first and then the proximal.

CONVERGENCE: CYCLIC ORDER

- Assume all subgradients generated by the algorithm are bounded: $\|\tilde{\nabla} f_{i_k}(x_k)\| \leq c$, $\|\tilde{\nabla} h_{i_k}(x_k)\| \leq c$ for all k, plus mild additional conditions.

- Assume components are chosen for iteration in cyclic order, and stepsize is constant within a cycle of iterations.

- Key inequality: For all $y \in X$ and all k that mark the beginning of a cycle:

$$\|x_{k+m}-y\|^2 \leq \|x_k-y\|^2-2\alpha_k\Big(F(x_k)-F(y)\Big)+\beta\alpha_k^2 m^2 c^2$$

where β is the constant $\beta = 1/m + 4$.

- Result for a constant stepsize $\alpha_k \equiv \alpha$:

$$\liminf_{k\to\infty} f(x_k) \leq f^* + \alpha\beta\frac{m^2c^2}{2}$$

- Convergence for $\alpha_k \downarrow 0$ with $\sum_{k=0}^{\infty} \alpha_k = \infty$.

CONVERGENCE: RANDOMIZED ORDER

- Convergence and convergence rate results are qualitatively similar to incremental subgradient case.
- Result for a constant stepsize $\alpha_k \equiv \alpha$:

$$\liminf_{k\to\infty} f(x_k) \leq f^* + \alpha\beta\frac{mc^2}{2}$$

(with probability 1).

- Faster convergence for randomized stepsize rule - improvement by a factor m over the cyclic order case.
- Convergence for $\alpha_k \downarrow 0$ with $\sum_{k=0}^{\infty}\alpha_k = \infty$. (with probability 1). Use of the supermartingale convergence theorem.

EXAMPLE I

- ℓ_1-Regularization for least squares

$$\min_{x\in\Re^n}\left\{\gamma\,\|x\|_1+\frac{1}{2}\sum_{i=1}^{m}(c_i'x-d_i)^2\right\}$$

- Use incremental gradient or proximal on the quadratic terms.
- Use proximal on the $\|x\|_1$ term:

$$z_k=\arg\min_{x\in\Re^n}\left\{\gamma\,\|x\|_1+\frac{1}{2\alpha_k}\|x-x_k\|^2\right\}$$

- Decomposes into the n one-dimensional minimizations

$$z_k^j=\arg\min_{x^j\in\Re}\left\{\gamma\,|x^j|+\frac{1}{2\alpha_k}|x^j-x_k^j|^2\right\},$$

and can be done with the shrinkage operation

$$z_k^j=\begin{cases}x_k^j-\gamma\alpha_k & \text{if } \gamma\alpha_k\le x_k^j,\\ 0 & \text{if } -\gamma\alpha_k<x_k^j<\gamma\alpha_k,\\ x_k^j+\gamma\alpha_k & \text{if } x_k^j\le-\gamma\alpha_k.\end{cases}$$

- Note that "small" coordinates x_k^j are set to 0.

EXAMPLE II

- Incremental constraint projection methods for

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{m} f_i(x) \\ &\text{subject to} \quad x \in \cap_{i=1}^{m} X_i, \end{aligned} \tag{1}$$

- Convert to the problem

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{m} f_i(x) + c \sum_{i=1}^{m} \text{dist}(x; X_i) \\ &\text{subject to} \quad x \in \Re^n, \end{aligned} \tag{2}$$

where c is a positive penalty parameter.

- Then for f Lipschitz continuous and c sufficiently large, problems (1) and (2) are equivalent (their minima coincide).

- Apply incremental subgradient-proximal:

$$y_k = x_k - \alpha_k \tilde{\nabla} f_{i_k}(x_k),$$

$$x_{k+1} \in \arg\min_{x \in \Re^n} \left\{ c\, \text{dist}(x; X_{j_k}) + \frac{1}{2\alpha_k} \|x - y_k\|^2 \right\}.$$

The second iteration can be implemented in "closed form," using projection on X_{j_k}.

LECTURE 24

LECTURE OUTLINE

- Extensions of proximal and projection ideas
- Nonquadratic proximal algorithms
- Entropy minimization algorithm
- Exponential augmented Lagrangian method
- Entropic descent algorithm

****

References:

- On-line chapter on algorithms

- Bertsekas, D. P., 1999. Nonlinear Programming, Athena Scientific, Belmont, MA.

- Beck, A., and Teboulle, M., 2003. "Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization," Operations Research Letters, Vol. 31, pp. 167-175.

GENERALIZED PROXIMAL-RELATED ALGS

- Introduce a general regularization term D_k:

$$x_{k+1} \in \arg\min_{x \in X} \Big\{ f(x) + D_k(x, x_k) \Big\}$$

- All the ideas extend to the nonquadratic case (although the analysis may not be trivial).
- In particular we have generalizations as follows:
  - Dual proximal algorithms (based on Fenchel duality)
  - Augmented Lagrangian methods with nonquadratic penalty functions
  - Combinations with polyhedral approximations (bundle-type methods)
  - Proximal gradient method
  - Incremental subgradient-proximal methods
  - Gradient projection algorithms with "nonquadratic metric"
- We may look also at what happens when f is not convex.

SPECIAL CASE: ENTROPY REGULARIZATION

$$D_k(x,y) = \begin{cases} \frac{1}{c_k}\sum_{i=1}^n x^i\left(\ln\left(\frac{x^i}{y^i}\right)-1\right) & \text{if } x>0,\ y>0, \\ \infty & \text{otherwise} \end{cases}$$

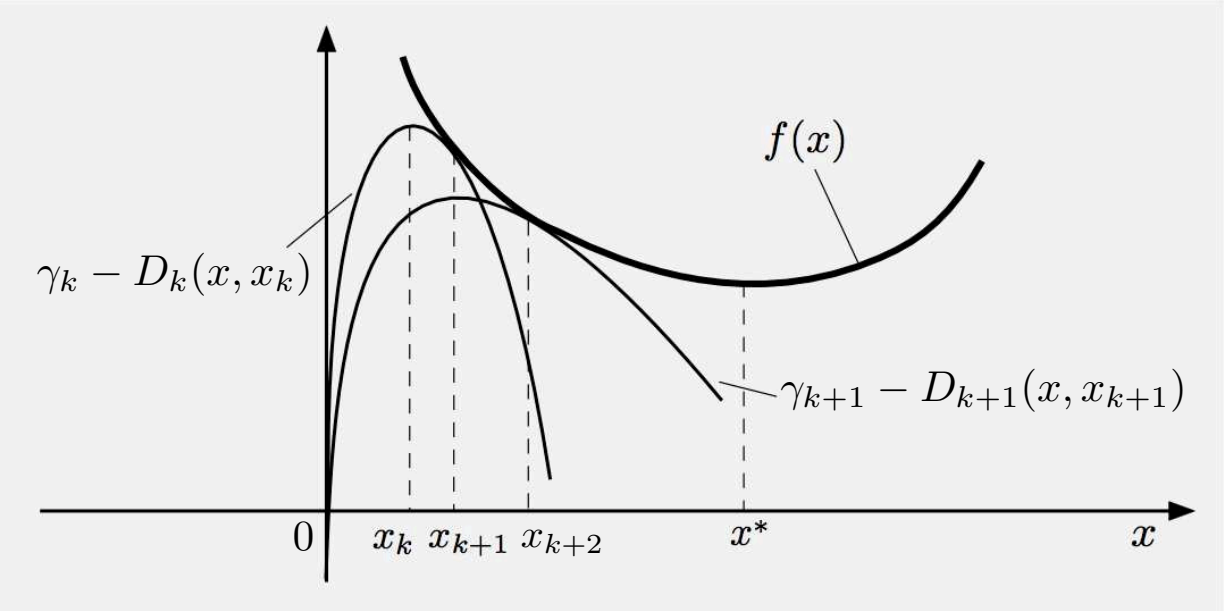
- Also written as

$$D_k(x,y) = \frac{1}{c_k}\sum_{i=1}^n y^i \phi_i\left(\frac{x^i}{y^i}\right),$$

where

$$\phi(x) = \begin{cases} x\big(\ln(x)-1\big) & \text{if } x>0, \\ 0 & \text{if } x=0, \\ \infty & \text{if } x<0. \end{cases}$$

- Proximal algorithm:

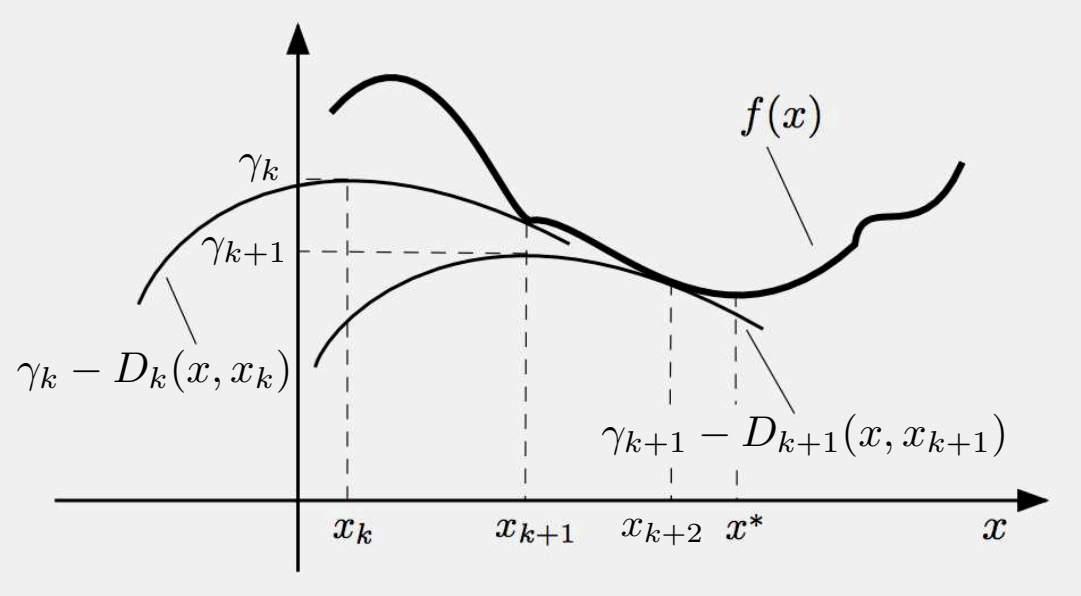


GENERALIZED PROXIMAL ALGORITHM

- Introduce a general regularization term D_k : $\Re^{2n} \mapsto (-\infty, \infty]$:

$$x_{k+1} \in \arg\min_{x \in \Re^n} \Big\{ f(x) + D_k(x, x_k) \Big\}$$

- Consider a general cost function f



- Assume attainment of min (but this is not automatically guaranteed)

- Complex/unreliable behavior when f is nonconvex

SOME GUARANTEES ON GOOD BEHAVIOR

- Assume "stabilization property"

$$D_k(x, x_k) \geq D_k(x_k, x_k), \qquad \forall\ x \in \Re^n,\ k \tag{1}$$

Then we have a cost improvement property:

$$\begin{aligned} f(x_{k+1}) &\leq f(x_{k+1}) + D_k(x_{k+1}, x_k) - D_k(x_k, x_k) \\ &\leq f(x_k) + D_k(x_k, x_k) - D_k(x_k, x_k) \\ &= f(x_k) \end{aligned} \tag{2}$$

- Assume algorithm stops only when x_k is in optimal solution set X^*, i.e.,

$$x_k \in \arg\min_{x \in \Re^n} \big\{ f(x) + D_k(x, x_k) \big\} \quad \Rightarrow \quad x_k \in X^*$$

- Then strict cost improvement for $x_k \notin X^*$ [the second inequality in (2) is strict].

- Guaranteed if f is convex and:

  (a) $D_k(\cdot, x_k)$ satisfies (1), and is convex and differentiable at x_k.

  (b) $\text{ri}\big(\text{dom}(f)\big) \cap \text{ri}\big(\text{dom}(D_k(\cdot, x_k))\big) \neq \emptyset$.

EXAMPLES

- Bregman distance function

$$D_k(x,y) = \frac{1}{c_k}\big(\phi(x) - \phi(y) - \nabla\phi(y)'(x-y)\big),$$

where $\phi : \Re^n \mapsto (-\infty, \infty]$ is a convex function, differentiable within an open set containing $\mathrm{dom}(f)$, and c_k is a positive penalty parameter. Special cases: quadratic and entropy functions.

- Majorization-Minimization algorithm:

$$D_k(x,y) = M_k(x,y) - M_k(y,y),$$

where M satisfies

$$M_k(y,y) = f(y), \qquad \forall\ y \in \Re^n,\ k = 0, 1,$$

$$M_k(x,x_k) \geq f(x_k), \qquad \forall\ x \in \Re^n,\ k = 0, 1, \ldots$$

- Example for case $f(x) = R(x) + \|Ax - b\|^2$, where R is a convex regularization function

$$M(x,y) = R(x) + \|Ax - b\|^2 - \|Ax - Ay\|^2 + \|x - y\|^2$$

- Expectation-Maximization (EM) algorithm (special context in inference, f nonconvex)

DUAL PROXIMAL MINIMIZATION

- The proximal iteration can be written in the Fenchel form: $\min_x\{f_1(x)+f_2(x)\}$ with

$$f_1(x)=f(x), \qquad f_2(x)=D_k(x;x_k)$$

- The Fenchel dual is

$$\begin{aligned} &\text{minimize} \quad f^\star(\lambda)+D_k^\star(\lambda;x_k)\\ &\text{subject to} \quad \lambda\in\Re^n \end{aligned}$$

where $D_k^\star(\cdot;x_k)$ is the conjugate of $D_k(\cdot;x_k)$:

$$D_k^\star(\lambda;x_k)=\sup_{x\in\Re^n}\Big\{-\lambda'x-D_k(x;x_k)\Big\}$$

- If $D_k(\cdot;x_k)$ or $D_k^\star(\cdot;x_k)$ is real-valued, there is no duality gap.
- Can use the Fenchel dual for a dual proximal implementation.

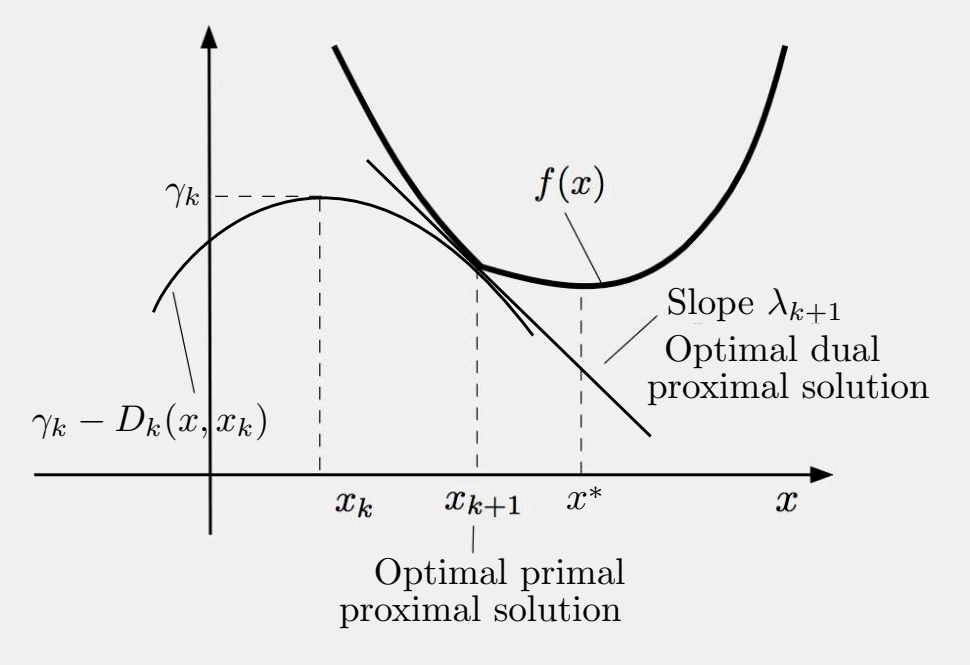
DUAL IMPLEMENTATION

- We can solve the Fenchel-dual problem instead of the primal at each iteration:

$$\lambda_{k+1} = \arg \min_{\lambda \in \Re^n} \{f^\star(\lambda) + D_k^\star(\lambda; x_k)\}$$

- Primal-dual optimal pair (x_{k+1}, λ_{k+1}) are related by the "differentiation" condition:

$$\lambda_{k+1} \in \partial D_k(x_{k+1}; x_k) \quad \text{or} \quad x_{k+1} \in \partial D_k^\star(\lambda_{k+1}; x_k)$$



- The primal and dual algorithms generate identical sequences $\{x_k, \lambda_k\}$.

- Special cases: Augmented Lagrangian methods with nonquadratic penalty functions.

ENTROPY/EXPONENTIAL DUALITY

- A special case involving entropy regularization:

$$x_{k+1} \in \arg\min_{x \in X} \left\{ f(x) + \frac{1}{c_k} \sum_{i=1}^{n} x^i \left(\ln\left(\frac{x^i}{x_k^i} \right) - 1 \right) \right\}$$

where $x_k > 0$.

- Fenchel duality $\Rightarrow$ Augmented Lagrangian method
- Note: The conjugate of the logarithmic

$$h(x) = \begin{cases} x\big(\ln(x) - 1\big) & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ \infty & \text{if } x < 0, \end{cases}$$

is the exponential $h^\star(y) = e^y$.

- The dual (augmented Lagrangian) problem is

$$u_{k+1} \in \arg\min_{u \in \Re^n} \left\{ f^\star(u) + \frac{1}{c_k} \sum_{i=1}^{n} x_k^i e^{c_k u^i} \right\}$$

The proximal/multiplier iteration is

$$x_{k+1}^i = x_k^i e^{c_k u_{k+1}^i}, \qquad i = 1, \ldots, n$$

EXPONENTIAL AUGMENTED LAGRANGIAN

- A special case for the convex problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad g_1(x) \leq 0, \ldots, g_r(x) \leq 0, \quad x \in X \end{aligned}$$

- Apply proximal to the (Langrange) dual problem. It consists of unconstrained minimizations

$$x_k \in \arg\min_{x \in X} \left\{ f(x) + \frac{1}{c_k} \sum_{j=1}^{r} \mu_k^j e^{c_k g_j(x)} \right\},$$

followed by the multiplier iterations

$$\mu_{k+1}^j = \mu_k^j e^{c_k^j g_j(x_k)}, \qquad j = 1, \ldots, r$$

- Note: We must have $\mu_0 > 0$, which implies $\mu_k > 0$ for all k.
- Theoretical convergence properties are similar to the quadratic augmented Lagrangian method.
- The exponential is twice differentiable, hence more suitable for Newton-like methods.

NONLINEAR PROJECTION ALGORITHM

- Subgradient projection with general regularization D_k:

$$x_{k+1} \in \arg\min_{x \in X} \left\{ f(x_k) + \tilde{\nabla} f(x_k)'(x - x_k) + D_k(x, x_k) \right\}$$

where $\tilde{\nabla} f(x_k)$ is a subgradient of f at x_k. Also called mirror descent method.

- Linearization of f simplifies the minimization.

- The use of nonquadratic linearization is useful in problems with special structure.

- Entropic descent method: Minimize $f(x)$ over the unit simplex $X = \left\{ x \geq 0 \mid \sum_{i=1}^n x^i = 1 \right\}$.

- Method:

$$x_{k+1} \in \arg\min_{x \in X} \sum_{i=1}^n \left(x^i \tilde{\nabla}_i f(x_k) + \frac{1}{\alpha_k} x^i \left(\ln \left(\frac{x^i}{x_k^i} \right) - 1 \right) \right)$$

where $\tilde{\nabla}_i f(x_k)$ are the components of $\tilde{\nabla} f(x_k)$.

- This minimization can be done in closed form:

$$x_{k+1}^i = \frac{x_k^i e^{-\alpha_k \tilde{\nabla}_i f(x_k)}}{\sum_{j=1}^n x_k^j e^{-\alpha_k \tilde{\nabla}_j f(x_k)}}, \qquad i = 1, \ldots, n$$

LECTURE 25

LECTURE OUTLINE

- Interior point methods
- Coordinate descent methods
- Distributed asynchronous fixed point computation

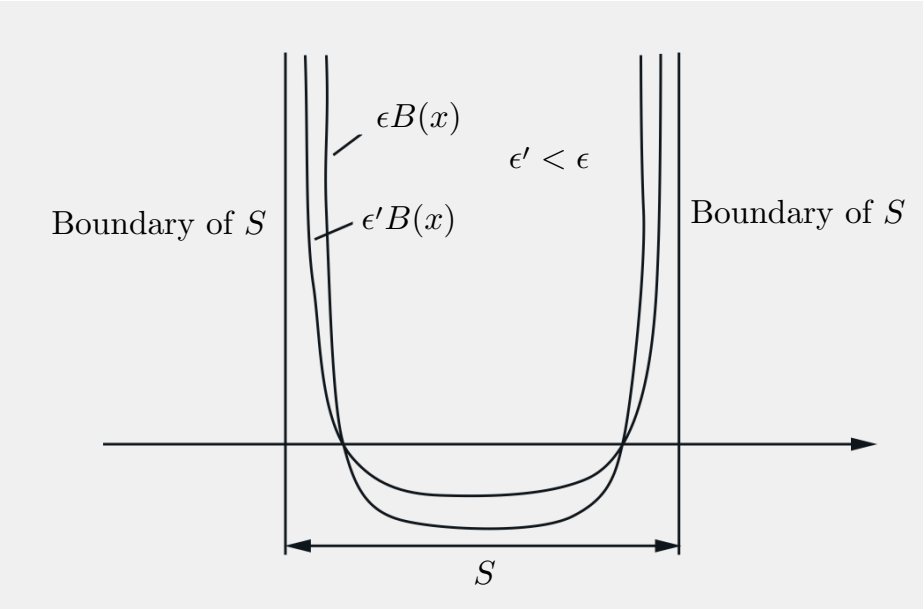
****

References:

- Boyd and Vanderbergue, 2004. Convex Optimization, Cambridge U. Press.
- Bertsekas, D. P., 1999. Nonlinear Programming, Athena Scientific, Belmont, MA.
- Bertsekas, D. P., and Tsitsiklis, J. N., 1989. Parallel and Distributed Algorithms: Numerical Methods, Prentice-Hall.

INTERIOR POINT METHODS

- Problem: $\min_{x \in X,\, g_j(x),\, j=1,\ldots,r} f(x)$
- Let $S = \{x \in X \mid g_j(x) < 0,\, j = 1, \ldots, r\}$ (assumed nonempty). A barrier function, is defined and is continuous on S, and goes to ∞ as any one of the constraints $g_j(x) \uparrow 0$.



- Examples:

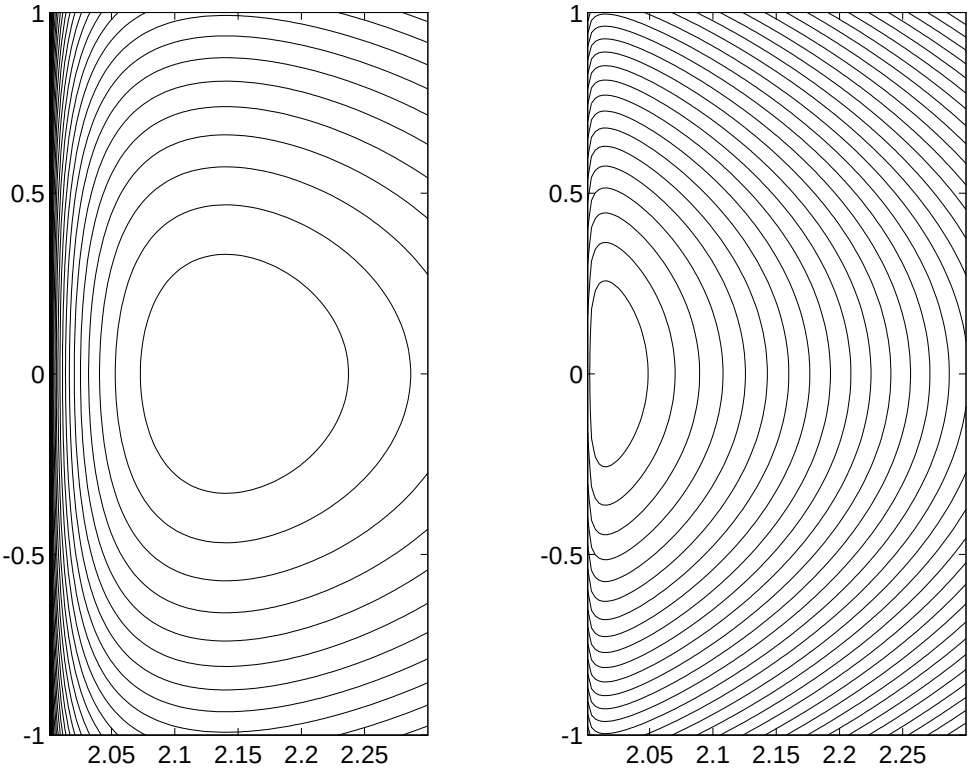
$$B(x) = -\sum_{j=1}^{r} \ln\bigl\{-g_j(x)\bigr\}, \qquad B(x) = -\sum_{j=1}^{r} \frac{1}{g_j(x)}.$$

- Barrier method: Generates

$$x_k = \arg\min_{x \in S} \bigl\{ f(x) + \epsilon_k B(x) \bigr\}, \qquad k = 0, 1, \ldots,$$

where $\epsilon_k \downarrow 0$.

BARRIER METHOD - EXAMPLE



$$\text{minimize} \quad f(x) = \tfrac{1}{2}\left((x^1)^2 + (x^2)^2\right)$$

$$\text{subject to} \quad 2 \leq x^1,$$

with optimal solution $x^* = (2, 0)$.

- Logarithmic barrier: $B(x) = -\ln(x^1 - 2)$
- We have $x_k = \left(1 + \sqrt{1 + \epsilon_k}\,, 0\right)$ from

$$x_k \in \arg\min_{x^1 > 2} \left\{ \tfrac{1}{2}\left((x^1)^2 + (x^2)^2\right) - \epsilon_k \ln(x^1 - 2) \right\}$$

- As ϵ_k is decreased, the unconstrained minimum x_k approaches the constrained minimum $x^* = (2, 0)$.
- As $\epsilon_k \to 0$, computing x_k becomes more difficult because of ill-conditioning (a Newton-like method is essential for solving the approximate problems).

CONVERGENCE

- Assume that X is closed convex, and f, and g_j are convex. Every limit point of a sequence $\{x_k\}$ generated by a barrier method is a minimum of the original constrained problem.

Proof: Let $\{\overline{x}\}$ be the limit of a subsequence $\{x_k\}_{k\in K}$. Since $x_k \in S$ and X is closed, $\overline{x}$ is feasible for the original problem.

If $\overline{x}$ is not a minimum, there exists a feasible x^* such that $f(x^*) < f(\overline{x})$ and therefore (by the Line Segment Principle) also an interior point $\tilde{x} \in S$ such that $f(\tilde{x}) < f(\overline{x})$. By the definition of x_k,

$$f(x_k) + \epsilon_k B(x_k) \leq f(\tilde{x}) + \epsilon_k B(\tilde{x}), \qquad \forall\ k,$$

so by taking limit

$$f(\overline{x}) + \liminf_{k\to\infty,\ k\in K} \epsilon_k B(x_k) \leq f(\tilde{x}) < f(\overline{x})$$

Hence $\liminf_{k\to\infty,\ k\in K} \epsilon_k B(x_k) < 0$.

If $\overline{x} \in S$, we have $\lim_{k\to\infty,\ k\in K} \epsilon_k B(x_k) = 0$, while if $\overline{x}$ lies on the boundary of S, we have by assumption $\lim_{k\to\infty,\ k\in K} B(x_k) = \infty$. Thus

$$\liminf_{k\to\infty} \epsilon_k B(x_k) \geq 0, \qquad \text{a contradiction.}$$

SECOND ORDER CONE PROGRAMMING

- Consider the SOCP

$$\begin{aligned} &\text{minimize} \quad c'x \\ &\text{subject to} \quad A_i x - b_i \in C_i, \ i = 1, \ldots, m, \end{aligned}$$

where $x \in \Re^n$, c is a vector in $\Re^n$, and for $i = 1, \ldots, m$, A_i is an $n_i \times n$ matrix, b_i is a vector in $\Re^{n_i}$, and C_i is the second order cone of $\Re^{n_i}$.

- We approximate this problem with

$$\begin{aligned} &\text{minimize} \quad c'x + \epsilon_k \sum_{i=1}^{m} B_i(A_i x - b_i) \\ &\text{subject to} \quad x \in \Re^n, \quad A_i x - b_i \in \text{int}(C_i), \ i = 1, \ldots, m, \end{aligned}$$

where B_i is the logarithmic barrier function:

$$B_i(y) = -\ln\left(y_{n_i}^2 - (y_1^2 + \cdots + y_{n_i-1}^2)\right), \quad y \in \text{int}(C_i),$$

and $\epsilon_k \downarrow 0$.

- Essential to use Newton's method to solve the approximating problems.

- Interesting complexity analysis

SEMIDEFINITE PROGRAMMING

- Consider the dual SDP

$$\begin{aligned} &\text{maximize} \quad b'\lambda \\ &\text{subject to} \quad D - (\lambda_1 A_1 + \cdots + \lambda_m A_m) \in C, \end{aligned}$$

where $b \in \Re^m$, $D, A_1, \ldots, A_m$ are symmetric matrices, and C is the cone of positive semidefinite matrices.

- The logarithmic barrier method uses approximating problems of the form

$$\text{maximize} \quad b'\lambda + \epsilon_k \ln\big(\det(D - \lambda_1 A_1 - \cdots - \lambda_m A_m)\big)$$

over all $\lambda \in \Re^m$ such that $D - (\lambda_1 A_1 + \cdots + \lambda_m A_m)$ is positive definite.

- Here $\epsilon_k \downarrow 0$.

- Furthermore, we should use a starting point such that $D - \lambda_1 A_1 - \cdots - \lambda_m A_m$ is positive definite, and Newton's method should ensure that the iterates keep $D - \lambda_1 A_1 - \cdots - \lambda_m A_m$ within the positive definite cone.

COORDINATE DESCENT

- Problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in X, \end{aligned}$$

where $f : \Re^n \mapsto \Re$ is a differentiable convex function, and

$$X = X_1 \times X_2 \times \cdots \times X_m,$$

where X_i is a closed convex subset $\Re^{n_i}$.

- Partition x into "block" components

$$x = (x^1, x^2, \ldots, x^m),$$

constrained by $x^i \in X_i$.

- (Block) Coordinate descent: At each iteration, minimized the cost w.r.t. each of the block components x_k^i, in cyclic order

$$x_{k+1}^i \in \arg\min_{\xi \in X_i} f(x_{k+1}^1, \ldots, x_{k+1}^{i-1}, \xi, x_k^{i+1}, \ldots, x_k^m)$$

COORDINATE DESCENT CONVERGENCE

- Proposition: Assume that f is convex and continuously differentiable. Assume also that for each $x = (x^1, \ldots, x^m) \in X$ and i,

$$f(x^1, \ldots, x^{i-1}, \xi, x^{i+1}, \ldots, x^m)$$

viewed as a function of ξ, attains a unique minimum over X_i. Let $\{x_k\}$ be the sequence generated by the block coordinate descent method. Then, every limit point of $\{x_k\}$ minimizes f over X.

- Variant to eliminate the uniqueness assumption:

$$x_{k+1}^i = \arg \min_{\xi \in X_i} f(x_{k+1}^1, \ldots, x_{k+1}^{i-1}, \xi, x_k^{i+1}, \ldots, x_k^m) + \frac{1}{2c} \|\xi - x_k^i\|^2,$$

where c is any fixed positive scalar.

- Justification: Apply the original method to minimization over $(x, y) \in X \times X$ of

$$f(x) + \frac{1}{2c} \|x - y\|^2$$

COORDINATE DESCENT EXTENSIONS

- When f is convex but nondifferentiable, the coordinate descent approach may fail in general (there may be nonoptimal points for which descent along all coordinate directions is impossible).
- Favorable special case, when the nondifferentiable portion of f is separable, i.e., f has the form

$$f(x) = F(x) + \sum_{i=1}^{n} G_i(x^i),$$

where F is convex and differentiable, and each $G_i : \Re \mapsto \Re$ is convex.

- A case of special interest is ℓ_1-regularization:

$$\sum_{i=1}^{n} G_i(x^i) = \gamma \|x\|_1$$

- It is possible to iterate the block components in an irregular even randomized order instead of a fixed cyclic order.
- Distributed asynchronous implementation.

ASYNCHRONOUS FIXED POINT ALGORITHMS

- Fixed point problem $x = F(x)$, where $x = (x^1, \ldots, x^m)$, to be solved with m processors.
- Asynchronous fixed point algorithm:

$$x^i_{t+1} = \begin{cases} F_i\left(x^1_{\tau_{i1}(t)}, \ldots, x^m_{\tau_{im}(t)}\right) & \text{if } t \in \mathcal{R}_i, \\ x^i_t & \text{if } t \notin \mathcal{R}_i. \end{cases} \quad (1)$$

$\mathcal{R}_i$ are the computation times of processor i and $t - \tau_{ij}(t)$ are the interprocessor communication delays.

- Some processors may execute more iterations than others, while the communication delays between processors may be unpredictable.
- Continuous Updating and Information Renewal Assumption:
  - The set of times $\mathcal{R}_i$ at which processor i updates x^i is infinite, for each $i = 1, \ldots, m$.
  - $\lim_{t\to\infty} \tau_{ij}(t) = \infty$ for all $i, j = 1, \ldots, m$.
- This is totally asynchronous operation.
- Can show that the algorithm works when F is a contraction with respect to a weighted sup-norm (special case of a more general theorem).

ASYNCHRONOUS CONVERGENCE THEOREM

- Let F have a unique fixed point x^*, and assume that there is a sequence of nonempty subsets $\{S(k)\} \subset \Re^n$ with

$$S(k+1) \subset S(k), \quad k = 0, 1, \ldots,$$

and is such that if $\{y_k\}$ is any sequence with $y_k \in S(k)$, for all $k \geq 0$, then $\{y_k\}$ converges to x^*. Assume further the following:

(1) Synchronous Convergence Condition: We have

$$F(x) \in S(k+1), \qquad \forall\ x \in S(k),\ k = 0, 1, \ldots.$$

(2) Box Condition: For all k, $S(k)$ is a Cartesian product of the form

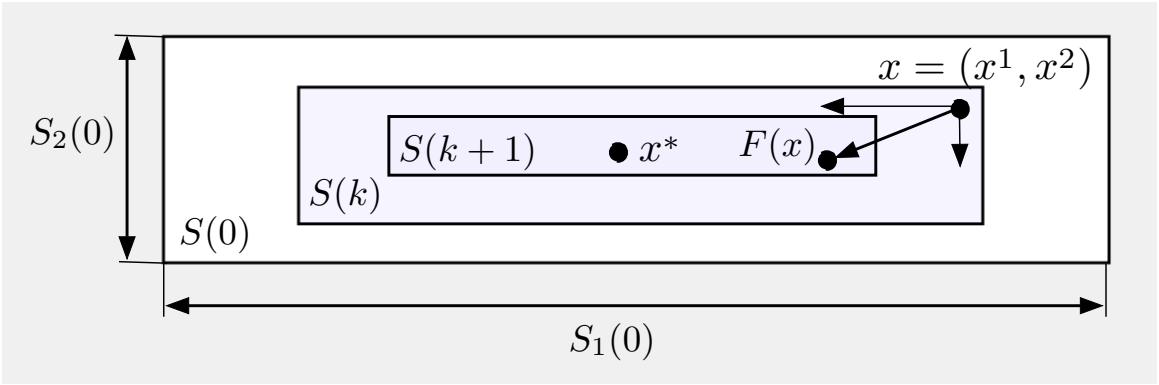
$$S(k) = S_1(k) \times \cdots \times S_m(k),$$

where $S_i(k)$ is a set of real-valued functions on X_i, $i = 1, \ldots, m$.

Then for every $x_0 \in S(0)$, the sequence $\{x_t\}$ generated by the asynchronous algorithm (1) converges to x^*.

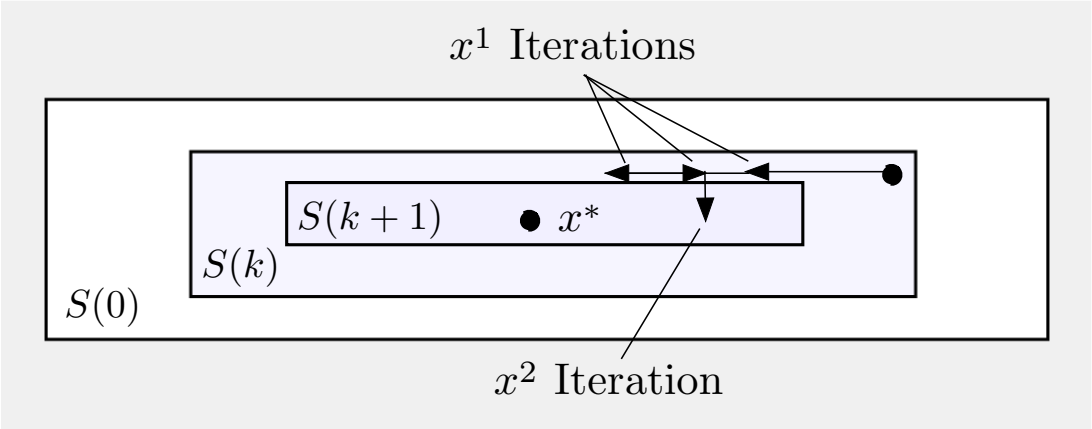
CONVERGENCE ANALYSIS

- Interpretation of assumptions:



A synchronous iteration from any x in $S(k)$ moves into $S(k+1)$ (component-by-component)

- Convergence mechanism:



Key: "Independent" component-wise improvement. An asynchronous component iteration from any x in $S(k)$ moves into the corresponding component portion of $S(k+1)$

LECTURE 26: REVIEW/EPILOGUE

LECTURE OUTLINE

CONVEX ANALYSIS AND DUALITY

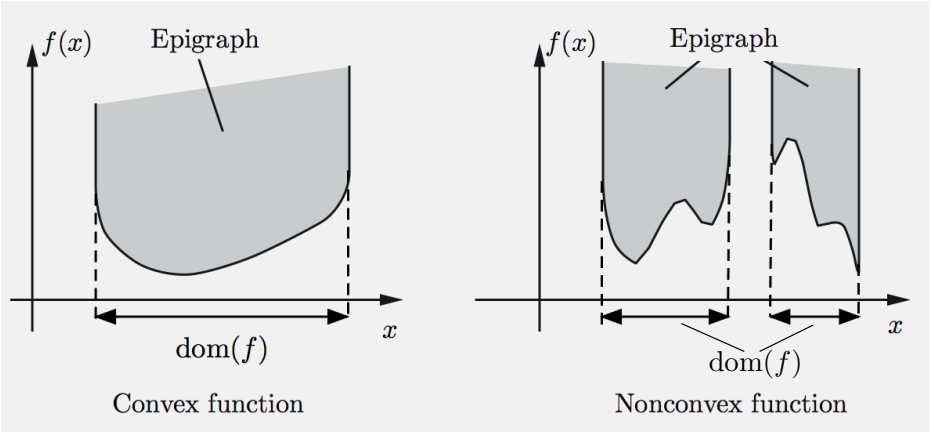
- Basic concepts of convex analysis
- Basic concepts of convex optimization
- Geometric duality framework - MC/MC
- Constrained optimization duality
- Subgradients - Optimality conditions

CONVEX OPTIMIZATION ALGORITHMS

- Special problem classes
- Subgradient methods
- Polyhedral approximation methods
- Proximal methods
- Dual proximal methods - Augmented Lagrangeans
- Optimal complexity methods
- Incremental methods
- Various combinations around proximal idea
- Interior point methods

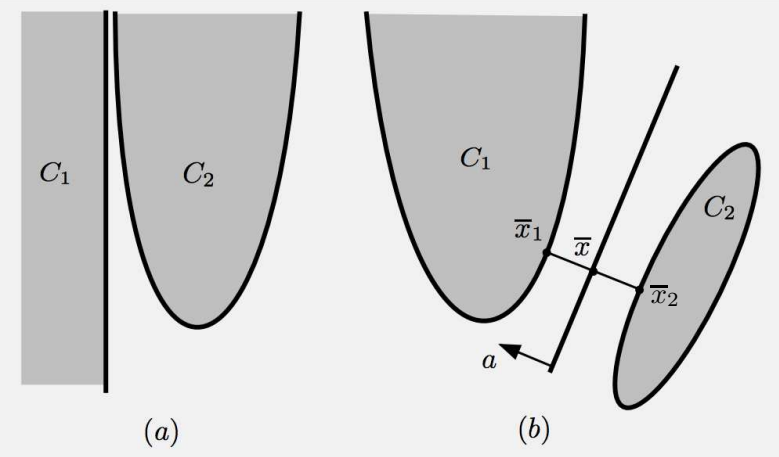
BASIC CONCEPTS OF CONVEX ANALYSIS

- Epigraphs, level sets, closedness, semicontinuity

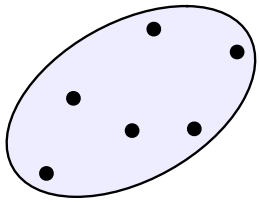


- Finite representations of generated cones and convex hulls - Caratheodory's Theorem.
- Relative interior:
  - Nonemptiness for a convex set
  - Line segment principle
  - Calculus of relative interiors
- Continuity of convex functions
- Nonemptiness of intersections of nested sequences of closed sets.
- Closure operations and their calculus.
- Recession cones and their calculus.
- Preservation of closedness by linear transformations and vector sums.

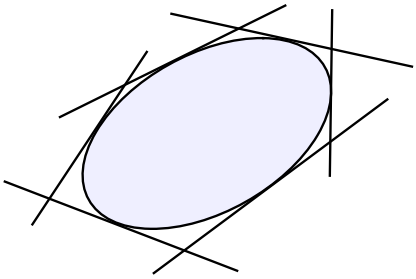
HYPERPLANE SEPARATION



- Separating/supporting hyperplane theorem.
- Strict and proper separation theorems.
- Dual representation of closed convex sets as unions of points and intersection of halfspaces.



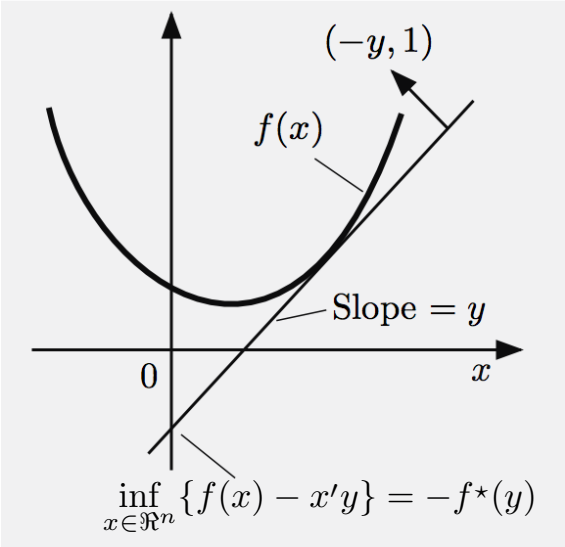
A union of points



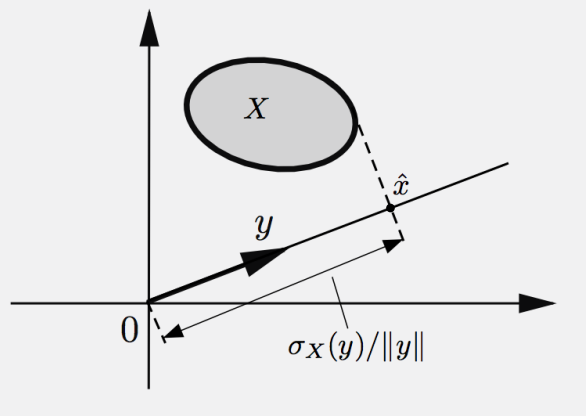
An intersection of halfspaces

- Nonvertical separating hyperplanes.

CONJUGATE FUNCTIONS



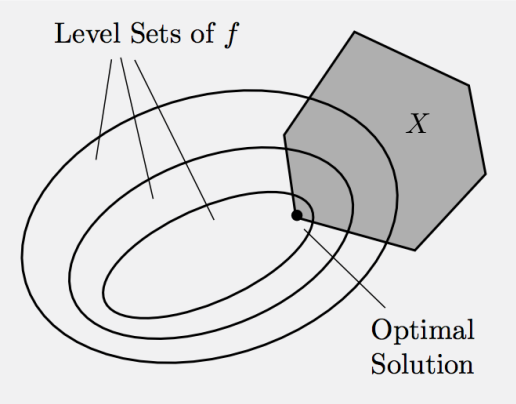
- Conjugacy theorem: $f = f^{\star\star}$
- Support functions



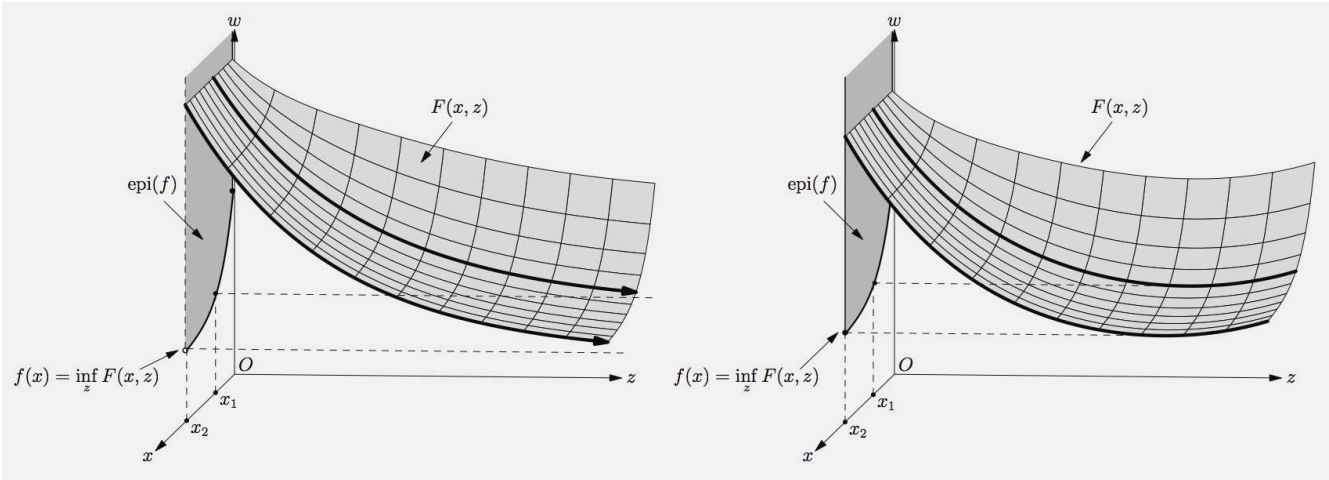
- Polar cone theorem: $C = C^{\star\star}$
  - Special case: Linear Farkas' lemma

BASIC CONCEPTS OF CONVEX OPTIMIZATION

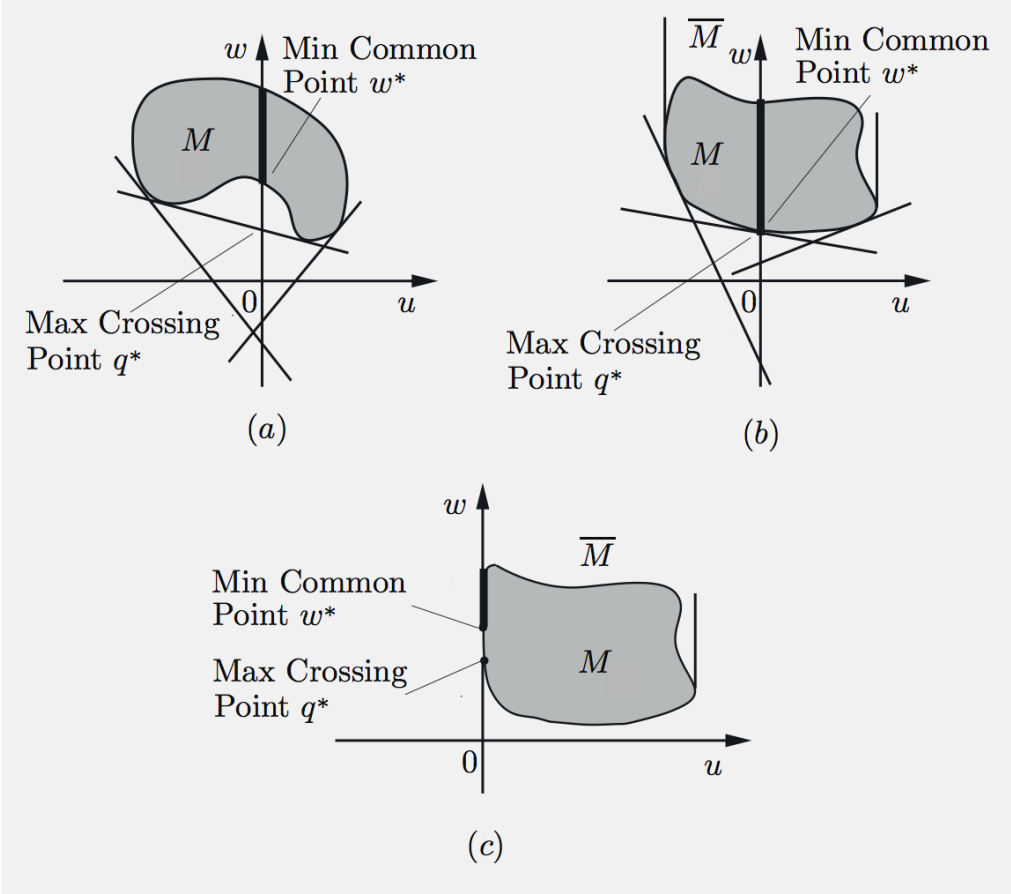
- Weierstrass Theorem and extensions.
- Characterization of existence of solutions in terms of nonemptiness of nested set intersections.



- Role of recession cone and lineality space.
- Partial Minimization Theorems: Characterization of closedness of $f(x) = \inf_{z \in \Re^m} F(x,z)$ in terms of closedness of F.



MIN COMMON/MAX CROSSING DUALITY



- Defined by a single set $M \subset \Re^{n+1}$.
- $w^* = \inf_{(0,w)\in M} w$
- $q^* = \sup_{\mu\in\Re^n} q(\mu) \stackrel{\triangle}{=} \inf_{(u,w)\in M}\{w + \mu' u\}$
- Weak duality: $q^* \leq w^*$
- Two key questions:
  - When does strong duality $q^* = w^*$ hold?
  - When do there exist optimal primal and dual solutions?

MC/MC THEOREMS ($\overline{M}$ CONVEX, $W^* < \infty$)

- MC/MC Theorem I: We have $q^* = w^*$ if and only if for every sequence $\{(u_k, w_k)\} \subset M$ with $u_k \to 0$, there holds

$$w^* \leq \liminf_{k \to \infty} w_k.$$

- MC/MC Theorem II: Assume in addition that $-\infty < w^*$ and that

$$D = \{u \mid \text{there exists } w \in \Re \text{ with } (u, w) \in \overline{M}\}$$

contains the origin in its relative interior. Then $q^* = w^*$ and there exists μ such that $q(\mu) = q^*$.

- MC/MC Theorem III: Similar to II but involves special polyhedral assumptions.

(1) $\overline{M}$ is a "horizontal translation" of $\tilde{M}$ by $-P$,

$$\overline{M} = \tilde{M} - \{(u, 0) \mid u \in P\},$$

where P: polyhedral and $\tilde{M}$: convex.

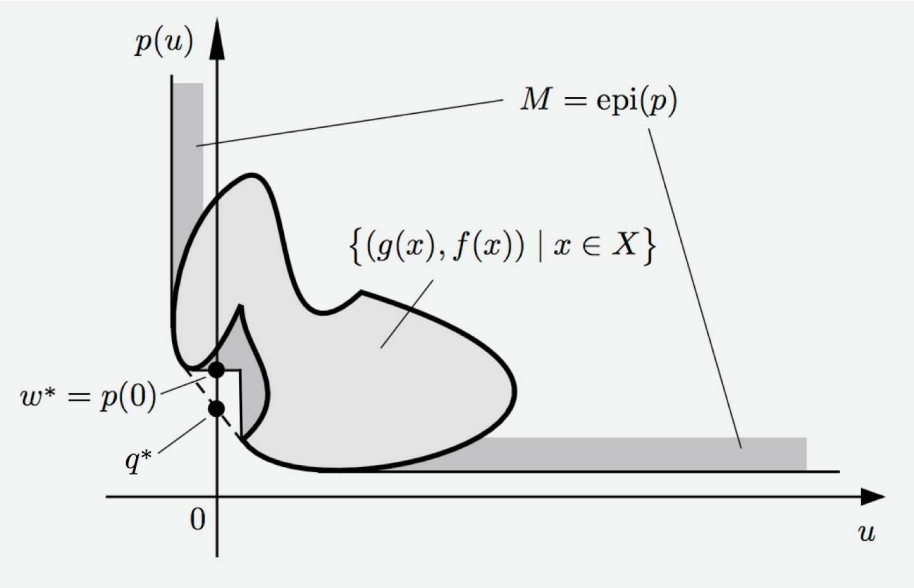
(2) We have $\text{ri}(\tilde{D}) \cap P \neq \emptyset$, where

$$\tilde{D} = \{u \mid \text{there exists } w \in \Re \text{ with } (u, w) \in \tilde{M}\}$$

IMPORTANT SPECIAL CASE

- Constrained optimization: $\inf_{x \in X,\, g(x) \leq 0} f(x)$
- Perturbation function (or primal function)

$$p(u) = \inf_{x \in X,\, g(x) \leq u} f(x),$$



- Introduce $L(x, \mu) = f(x) + \mu' g(x)$. Then

$$\begin{aligned} q(\mu) &= \inf_{u \in \Re^r} \left\{ p(u) + \mu' u \right\} \\ &= \inf_{u \in \Re^r,\, x \in X,\, g(x) \leq u} \left\{ f(x) + \mu' u \right\} \\ &= \begin{cases} \inf_{x \in X} L(x, \mu) & \text{if } \mu \geq 0, \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

NONLINEAR FARKAS' LEMMA

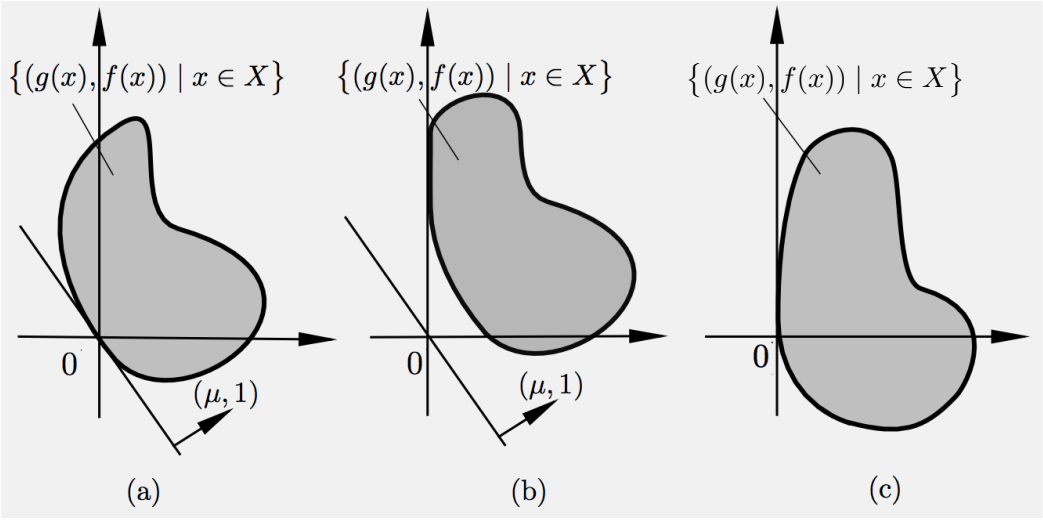
- Let $X \subset \Re^n$, $f : X \mapsto \Re$, and $g_j : X \mapsto \Re$, $j = 1, \ldots, r$, be convex. Assume that

$$f(x) \geq 0, \qquad \forall\ x \in X \text{ with } g(x) \leq 0$$

Let

$$Q^* = \big\{\mu \mid \mu \geq 0,\ f(x) + \mu' g(x) \geq 0,\ \forall\ x \in X\big\}.$$

- Nonlinear version: Then Q^* is nonempty and compact if and only if there exists a vector $\overline{x} \in X$ such that $g_j(\overline{x}) < 0$ for all $j = 1, \ldots, r$.



- Polyhedral version: Q^* is nonempty if g is linear $[g(x) = Ax - b]$ and there exists a vector $\overline{x} \in \text{ri}(X)$ such that $A\overline{x} - b \leq 0$.

CONSTRAINED OPTIMIZATION DUALITY

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

$$\text{subject to} \quad x \in X, \ g_j(x) \leq 0, \ j = 1, \ldots, r,$$

where $X \subset \Re^n$, $f : X \mapsto \Re$ and $g_j : X \mapsto \Re$ are convex. Assume f^*: finite.

- Connection with MC/MC: $M = \text{epi}(p)$ with $p(u) = \inf_{x \in X, \, g(x) \leq u} f(x)$
- Dual function:

$$q(\mu) = \begin{cases} \inf_{x \in X} L(x, \mu) & \text{if } \mu \geq 0, \\ -\infty & \text{otherwise} \end{cases}$$

where $L(x, \mu) = f(x) + \mu' g(x)$ is the Lagrangian function.

- Dual problem of maximizing $q(\mu)$ over $\mu \geq 0$.
- Strong Duality Theorem: $q^* = f^*$ and there exists dual optimal solution if one of the following two conditions holds:

(1) There exists $\overline{x} \in X$ such that $g(\overline{x}) < 0$.

(2) The functions g_j, $j = 1, \ldots, r$, are affine, and there exists $\overline{x} \in \text{ri}(X)$ such that $g(\overline{x}) \leq 0$.

OPTIMALITY CONDITIONS

- We have $q^* = f^*$, and the vectors x^* and μ^* are optimal solutions of the primal and dual problems, respectively, iff x^* is feasible, $\mu^* \geq 0$, and

$$x^* \in \arg\min_{x \in X} L(x, \mu^*), \qquad \mu_j^* g_j(x^*) = 0, \quad \forall\ j.$$

- For the linear/quadratic program

$$\text{minimize} \quad \tfrac{1}{2} x' Q x + c' x$$

$$\text{subject to} \quad Ax \leq b,$$

where Q is positive semidefinite, (x^*, μ^*) is a primal and dual optimal solution pair if and only if:

(a) Primal and dual feasibility holds:

$$Ax^* \leq b, \qquad \mu^* \geq 0$$

(b) Lagrangian optimality holds [x^* minimizes $L(x, \mu^*)$ over $x \in \Re^n$]. (Unnecessary for LP.)

(c) Complementary slackness holds:

$$(Ax^* - b)' \mu^* = 0,$$

i.e., $\mu_j^* > 0$ implies that the jth constraint is tight. (Applies to inequality constraints only.)

FENCHEL DUALITY

- Primal problem:

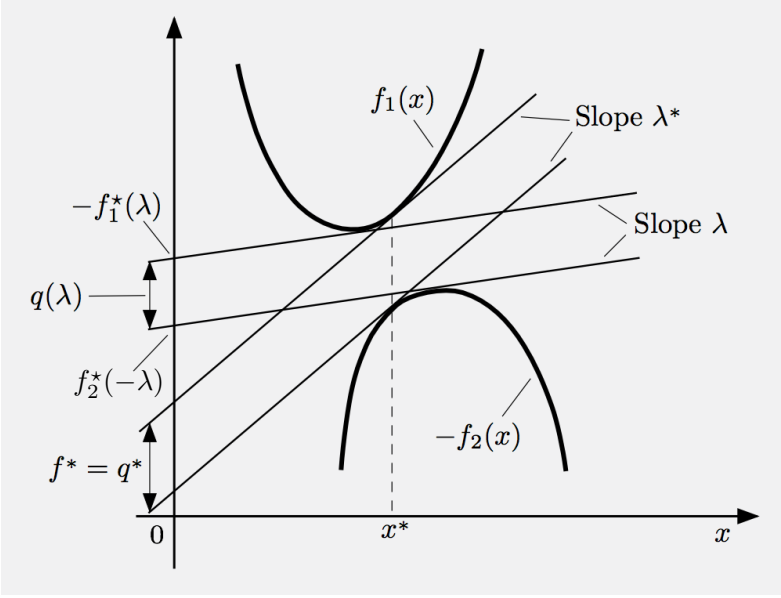
$$\begin{aligned} &\text{minimize} \quad f_1(x) + f_2(x) \\ &\text{subject to} \quad x \in \Re^n, \end{aligned}$$

where $f_1 : \Re^n \mapsto (-\infty, \infty]$ and $f_2 : \Re^n \mapsto (-\infty, \infty]$ are closed proper convex functions.

- Dual problem:

$$\begin{aligned} &\text{minimize} \quad f_1^\star(\lambda) + f_2^\star(-\lambda) \\ &\text{subject to} \quad \lambda \in \Re^n, \end{aligned}$$

where $f_1^\star$ and $f_2^\star$ are the conjugates.



CONIC DUALITY

- Consider minimizing $f(x)$ over $x \in C$, where $f : \Re^n \mapsto (-\infty, \infty]$ is a closed proper convex function and C is a closed convex cone in $\Re^n$.
- We apply Fenchel duality with the definitions

$$f_1(x) = f(x), \qquad f_2(x) = \begin{cases} 0 & \text{if } x \in C, \\ \infty & \text{if } x \notin C. \end{cases}$$

- Linear Conic Programming:

$$\begin{aligned} &\text{minimize} \quad c'x \\ &\text{subject to} \quad x - b \in S, \quad x \in C. \end{aligned}$$

- Equivalent dual linear conic problem:

$$\begin{aligned} &\text{minimize} \quad b'\lambda \\ &\text{subject to} \quad \lambda - c \in S^{\perp}, \quad \lambda \in \hat{C}. \end{aligned}$$

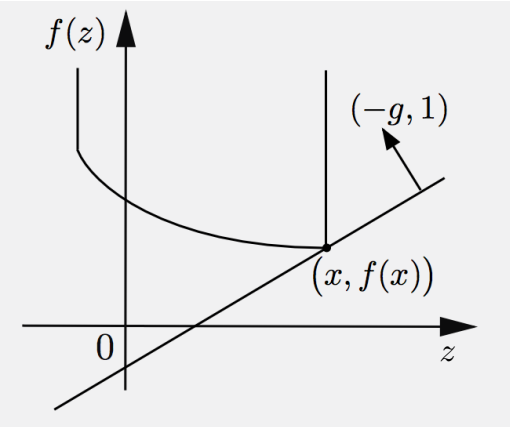
- Special Linear-Conic Forms:

$$\min_{Ax=b,\ x \in C} c'x \quad \Longleftrightarrow \quad \max_{c - A'\lambda \in \hat{C}} b'\lambda,$$

$$\min_{Ax-b \in C} c'x \quad \Longleftrightarrow \quad \max_{A'\lambda = c,\ \lambda \in \hat{C}} b'\lambda,$$

where $x \in \Re^n$, $\lambda \in \Re^m$, $c \in \Re^n$, $b \in \Re^m$, $A : m \times n$.

SUBGRADIENTS



- $\partial f(x) \neq \emptyset$ for $x \in \text{ri}\big(\text{dom}(f)\big)$.
- Conjugate Subgradient Theorem: If f is closed proper convex, the following are equivalent for a pair of vectors (x, y):

  (i) $x'y = f(x) + f^\star(y)$.

  (ii) $y \in \partial f(x)$.

  (iii) $x \in \partial f^\star(y)$.

- Characterization of optimal solution set $X^* = \arg\min_{x \in \Re^n} f(x)$ of closed proper convex f:

  (a) $X^* = \partial f^\star(0)$.

  (b) X^* is nonempty if $0 \in \text{ri}\big(\text{dom}(f^\star)\big)$.

  (c) X^* is nonempty and compact if and only if $0 \in \text{int}\big(\text{dom}(f^\star)\big)$.

CONSTRAINED OPTIMALITY CONDITION

- Let $f : \Re^n \mapsto (-\infty, \infty]$ be proper convex, let X be a convex subset of $\Re^n$, and assume that one of the following four conditions holds:

  (i) $\mathrm{ri}\big(\mathrm{dom}(f)\big) \cap \mathrm{ri}(X) \neq \emptyset$.

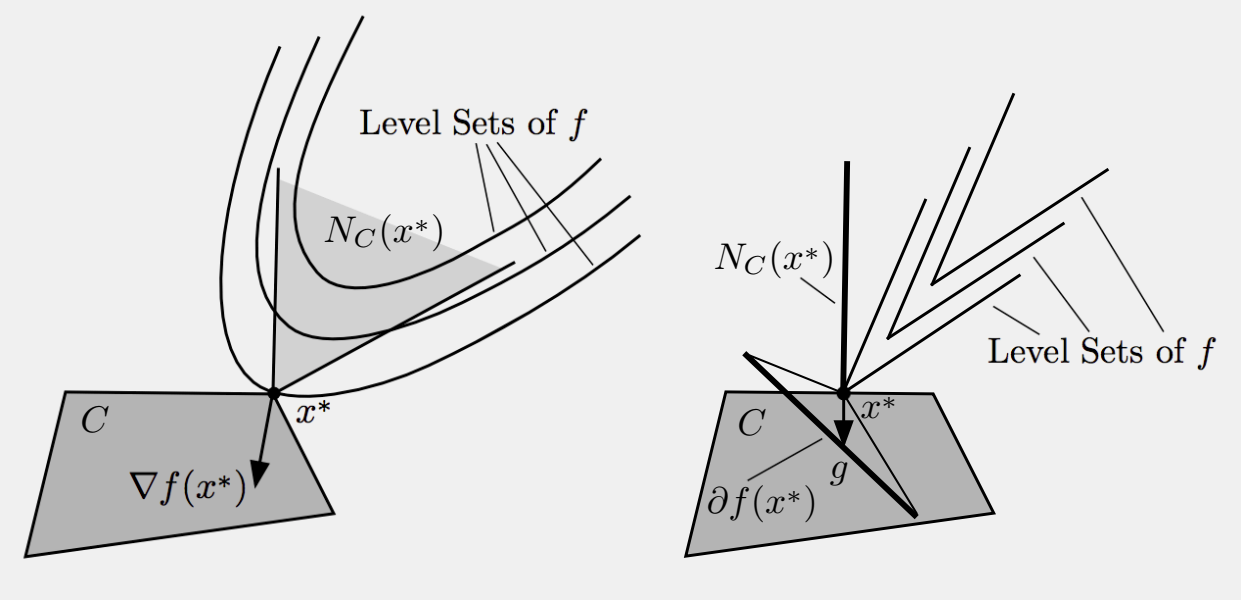
  (ii) f is polyhedral and $\mathrm{dom}(f) \cap \mathrm{ri}(X) \neq \emptyset$.

  (iii) X is polyhedral and $\mathrm{ri}\big(\mathrm{dom}(f)\big) \cap X \neq \emptyset$.

  (iv) f and X are polyhedral, and $\mathrm{dom}(f) \cap X \neq \emptyset$.

Then, a vector x^* minimizes f over X iff there exists $g \in \partial f(x^*)$ such that $-g$ belongs to the normal cone $N_X(x^*)$, i.e.,

$$g'(x - x^*) \geq 0, \qquad \forall\ x \in X.$$

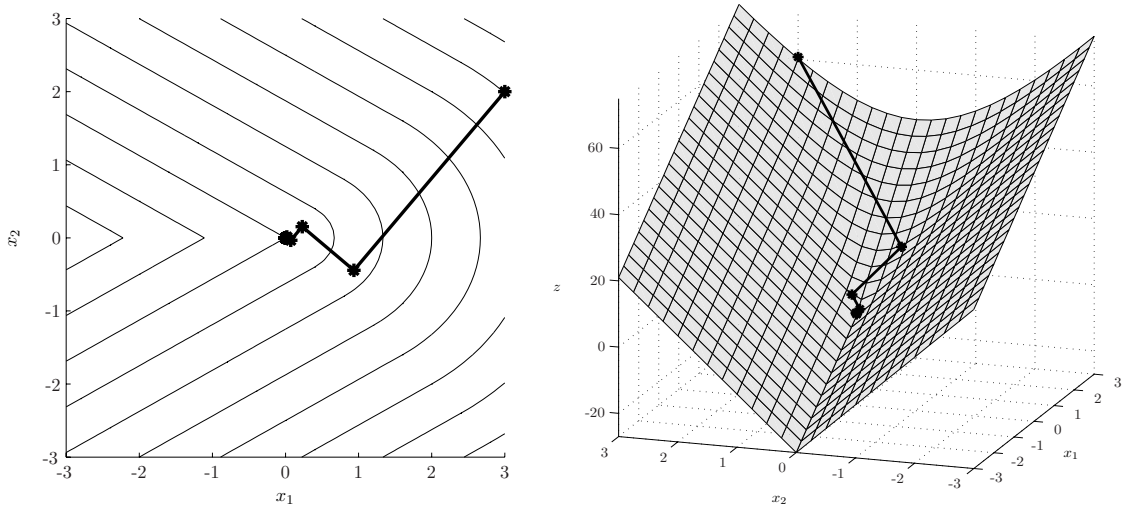


COMPUTATION: PROBLEM RANKING IN INCREASING COMPUTATIONAL DIFFICULTY

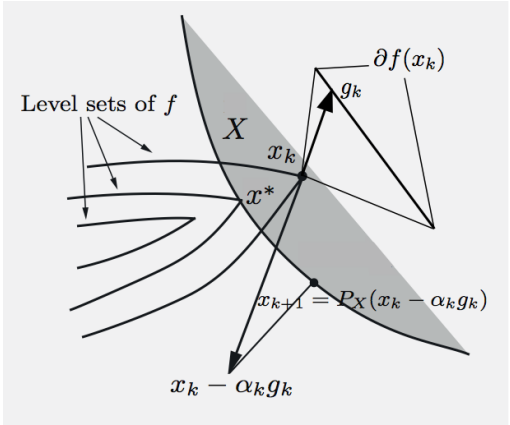
- Linear and (convex) quadratic programming.
  - Favorable special cases.
- Second order cone programming.
- Semidefinite programming.
- Convex programming.
  - Favorable cases, e.g., separable, large sum.
  - Geometric programming.
- Nonlinear/nonconvex/continuous programming.
  - Favorable special cases.
  - Unconstrained.
  - Constrained.
- Discrete optimization/Integer programming
  - Favorable special cases.
- Caveats/questions:
  - Important role of special structures.
  - What is the role of "optimal algorithms"?
  - Is complexity the right philosophical view to convex optimization?

DESCENT METHODS

- Steepest descent method: Use vector of min norm on $-\partial f(x)$; has convergence problems.



- Subgradient method:

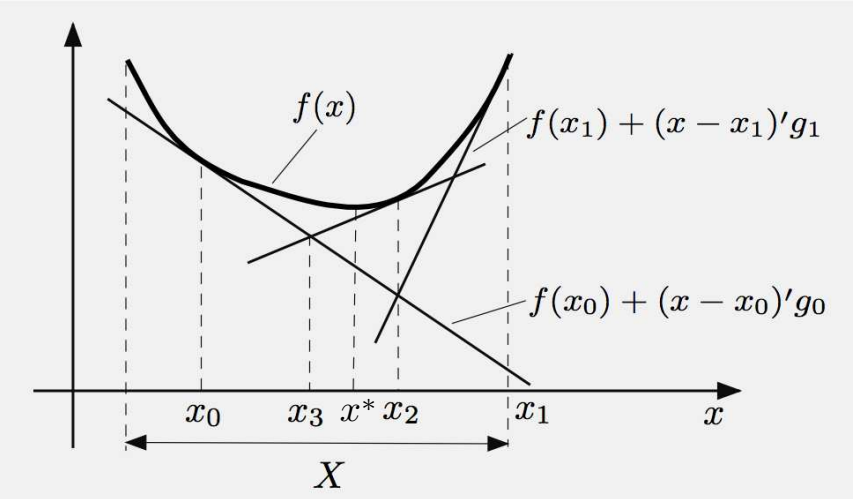


- ϵ-subgradient method (approx. subgradient)

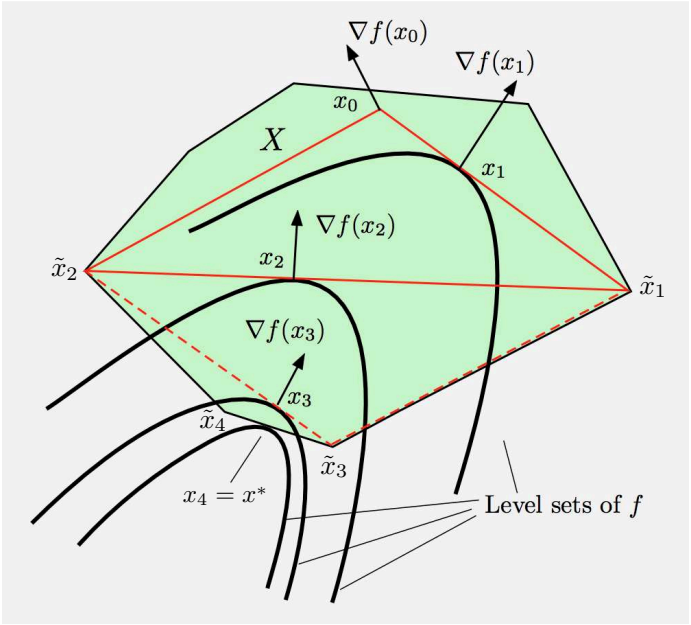
- Incremental (possibly randomized) variants for minimizing large sums (can be viewed as an approximate subgradient method).

OUTER AND INNER LINEARIZATION

- Outer linearization: Cutting plane



- Inner linearization: Simplicial decomposition



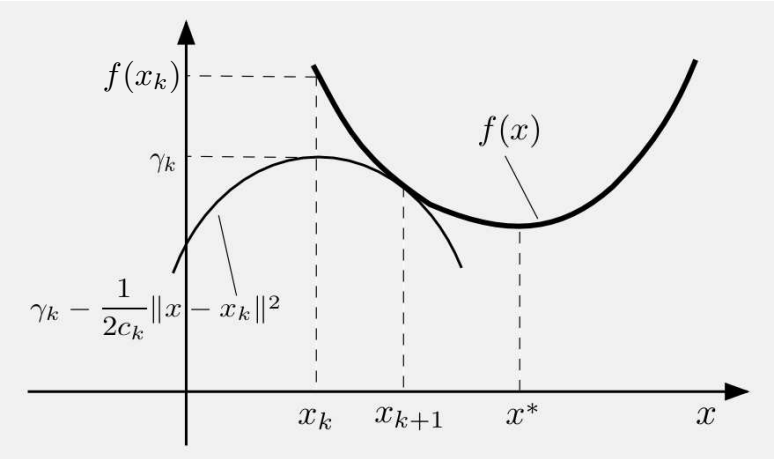
- Fenchel-like duality between outer and inner linearization.
  - Extended monotropic programming

PROXIMAL MINIMIZATION ALGORITHM

- A general algorithm for convex fn minimization

$$x_{k+1} \in \arg\min_{x \in \Re^n} \left\{ f(x) + \frac{1}{2c_k} \|x - x_k\|^2 \right\}$$

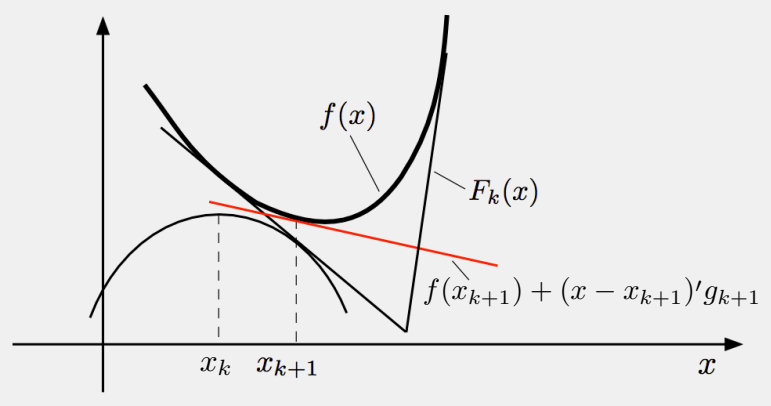
  - $f : \Re^n \mapsto (-\infty, \infty]$ is closed proper convex
  - c_k is a positive scalar parameter
  - x_0 is arbitrary starting point



- x_{k+1} exists because of the quadratic.
- Strong convergence properties
- Starting point for extensions (e.g., nonquadratic regularization) and combinations (e.g., with linearization)

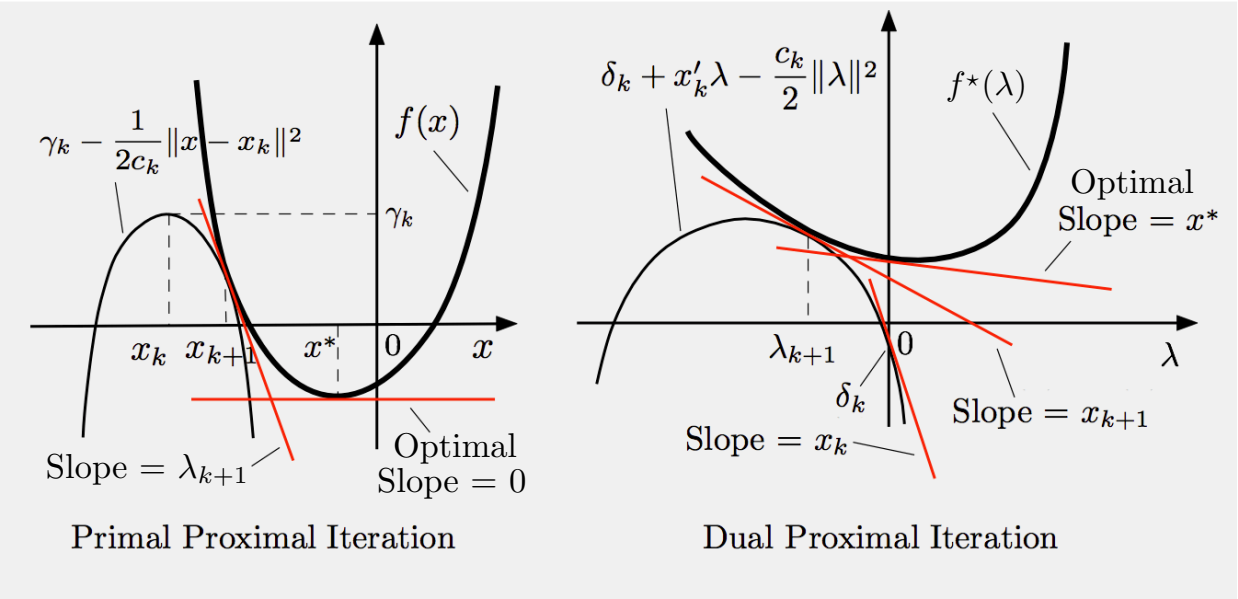
PROXIMAL-POLYHEDRAL METHODS

- Proximal-cutting plane method

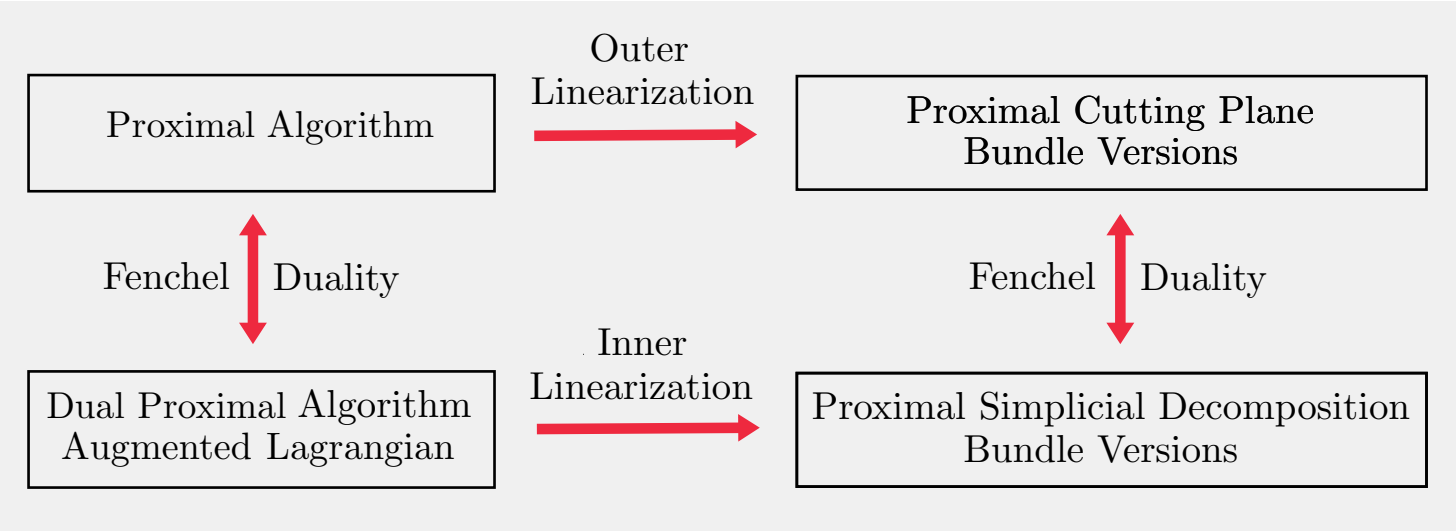


- Proximal-cutting plane-bundle methods: Replace f with a cutting plane approx. and/or change quadratic regularization more conservatively.

- Dual Proximal - Augmented Lagrangian methods: Proximal method applied to the dual problem of a constrained optimization problem.



DUALITY VIEW OF PROXIMAL METHODS



- Applies also to cost functions that are sums of convex functions

$$f(x) = \sum_{i=1}^{m} f_i(x)$$

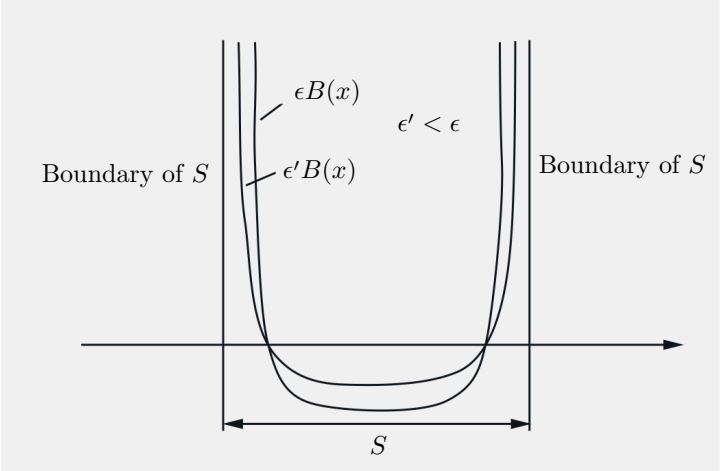
in the context of extended monotropic programming

INTERIOR POINT METHODS

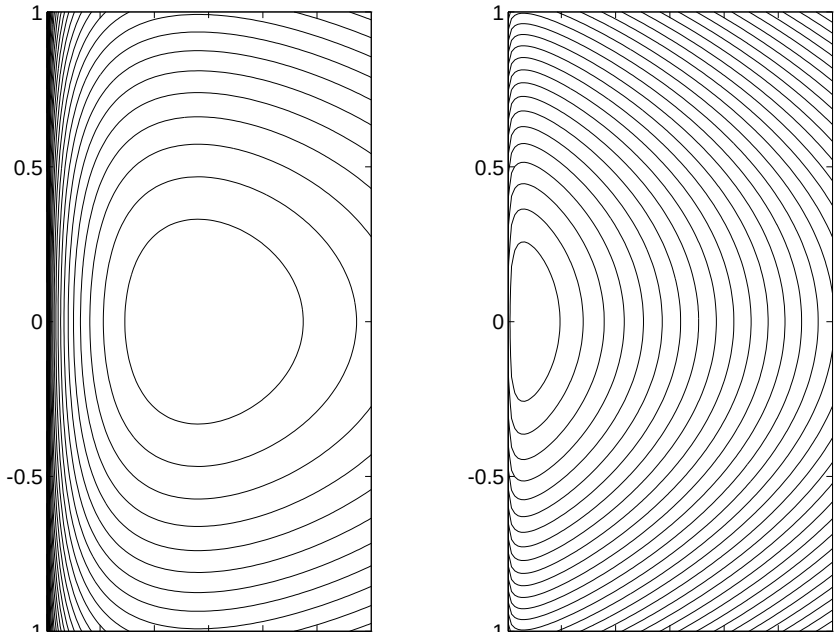
- Barrier method: Let

$$x_k = \arg\min_{x \in S} \Big\{ f(x) + \epsilon_k B(x) \Big\}, \qquad k = 0, 1, \ldots,$$

where $S = \{x \mid g_j(x) < 0,\ j = 1, \ldots, r\}$ and the parameter sequence $\{\epsilon_k\}$ satisfies $0 < \epsilon_{k+1} < \epsilon_k$ for all k and $\epsilon_k \to 0$.



- Ill-conditioning. Need for Newton's method



ADVANCED TOPICS

- Complexity view of first order algorithms
  - Gradient-projection for differentiable problems
  - Gradient-projection with extrapolation
  - Optimal iteration complexity version (Nesterov)
  - Extension to nondifferentiable problems by smoothing
- Proximal gradient method
- Incremental subgradient-proximal methods
- Useful extensions of proximal approach. General (nonquadratic) regularization - Bregman distance functions
  - Entropy-like regularization
  - Corresponding augmented Lagrangean method (exponential)
  - Corresponding proximal gradient method
  - Nonlinear gradient/subgradient projection (entropic minimization methods)
- Coordinate descent methods
- Distributed totally asynchronous methods