LECTURE SLIDES ON NONLINEAR PROGRAMMING

BASED ON LECTURES GIVEN AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

CAMBRIDGE, MASS

DIMITRI P. BERTSEKAS

These lecture slides are based on the book: "Nonlinear Programming," Athena Scientific, by Dimitri P. Bertsekas; see

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

for errata, selected problem solutions, and other support material.

LAST REVISED: Feb. 3, 2005

6.252 NONLINEAR PROGRAMMING

LECTURE 1: INTRODUCTION

LECTURE OUTLINE

- Nonlinear Programming
- Application Contexts
- Characterization Issue
- Computation Issue
- Duality
- Organization

NONLINEAR PROGRAMMING

$$\min_{x \in X} f(x),$$

where

- $f : \Re^n \mapsto \Re$ is a continuous (and usually differentiable) function of n variables
- $X = \Re^n$ or X is a subset of $\Re^n$ with a "continuous" character.

- If $X = \Re^n$, the problem is called unconstrained
- If f is linear and X is polyhedral, the problem is a linear programming problem. Otherwise it is a nonlinear programming problem
- Linear and nonlinear programming have traditionally been treated separately. Their methodologies have gradually come closer.

TWO MAIN ISSUES

- Characterization of minima
  - Necessary conditions
  - Sufficient conditions
  - Lagrange multiplier theory
  - Sensitivity
  - Duality
- Computation by iterative algorithms
  - Iterative descent
  - Approximation methods
  - Dual and primal-dual methods

APPLICATIONS OF NONLINEAR PROGRAMMING

- Data networks – Routing
- Production planning
- Resource allocation
- Computer-aided design
- Solution of equilibrium models
- Data analysis and least squares formulations
- Modeling human or organizational behavior

CHARACTERIZATION PROBLEM

- Unconstrained problems
  - Zero 1st order variation along all directions
- Constrained problems
  - Nonnegative 1st order variation along all feasible directions
- Equality constraints
  - Zero 1st order variation along all directions on the constraint surface
  - Lagrange multiplier theory
- Sensitivity

COMPUTATION PROBLEM

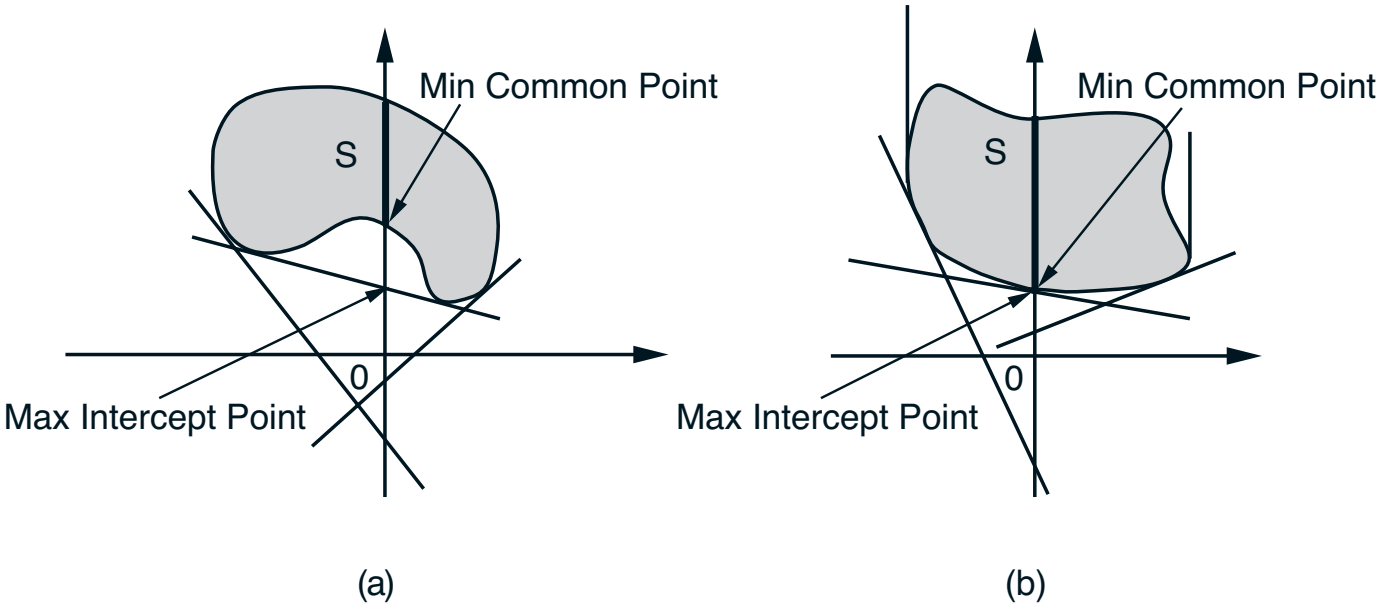
- Iterative descent
- Approximation
- Role of convergence analysis
- Role of rate of convergence analysis
- Using an existing package to solve a nonlinear programming problem

POST-OPTIMAL ANALYSIS

- Sensitivity
- Role of Lagrange multipliers as prices

DUALITY

- Min-common point problem / max-intercept problem duality



(a) (b)

Illustration of the optimal values of the min common point and max intercept point problems. In (a), the two optimal values are not equal. In (b), the set S, when "extended upwards" along the nth axis, yields the set

$$\bar{S} = \{\bar{x} \mid \text{for some } x \in S,\ \bar{x}_n \geq x_n,\ \bar{x}_i = x_i, i = 1, \ldots, n-1\}$$

which is convex. As a result, the two optimal values are equal. This fact, when suitably formalized, is the basis for some of the most important duality results.

6.252 NONLINEAR PROGRAMMING

LECTURE 2

UNCONSTRAINED OPTIMIZATION -

OPTIMALITY CONDITIONS

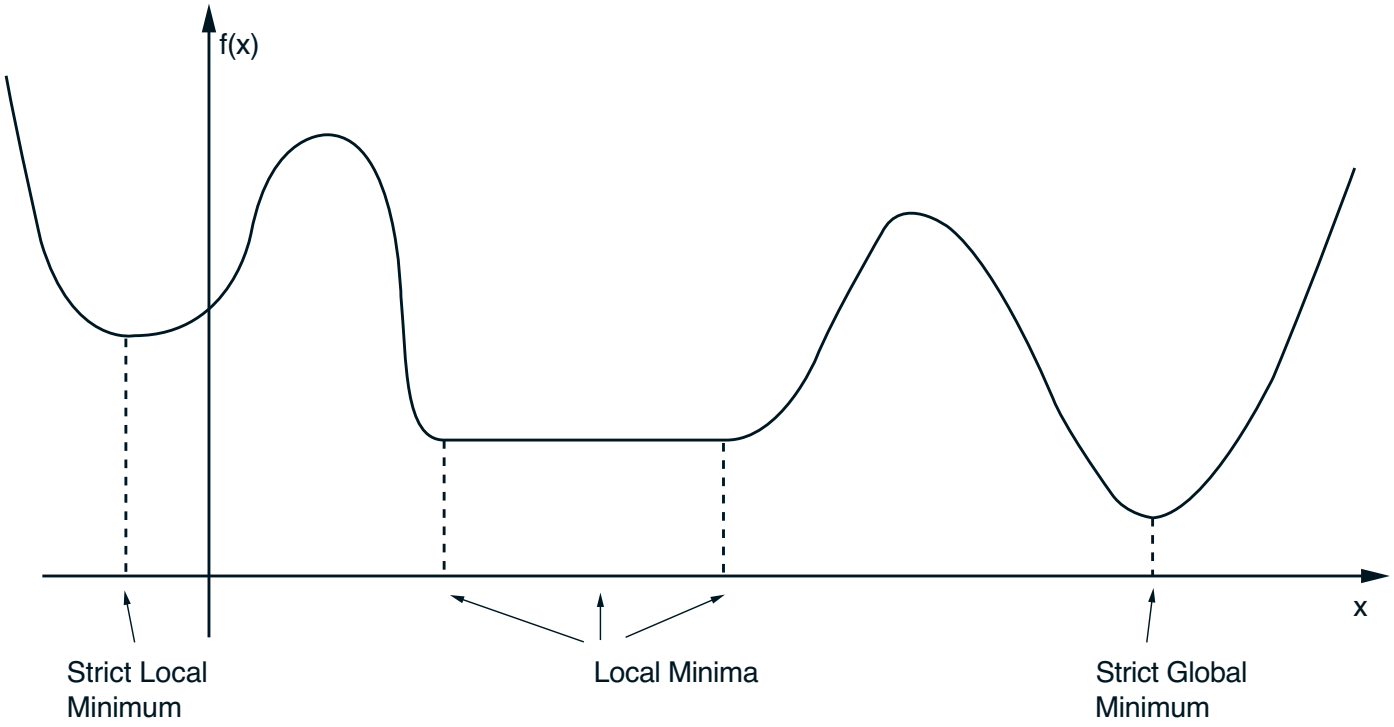
LECTURE OUTLINE

- Unconstrained Optimization
- Local Minima
- Necessary Conditions for Local Minima
- Sufficient Conditions for Local Minima
- The Role of Convexity

MATHEMATICAL BACKGROUND

- Vectors and matrices in $\Re^n$
- Transpose, inner product, norm
- Eigenvalues of symmetric matrices
- Positive definite and semidefinite matrices
- Convergent sequences and subsequences
- Open, closed, and compact sets
- Continuity of functions
- 1st and 2nd order differentiability of functions
- Taylor series expansions
- Mean value theorems

LOCAL AND GLOBAL MINIMA



Unconstrained local and global minima in one dimension.

NECESSARY CONDITIONS FOR A LOCAL MIN

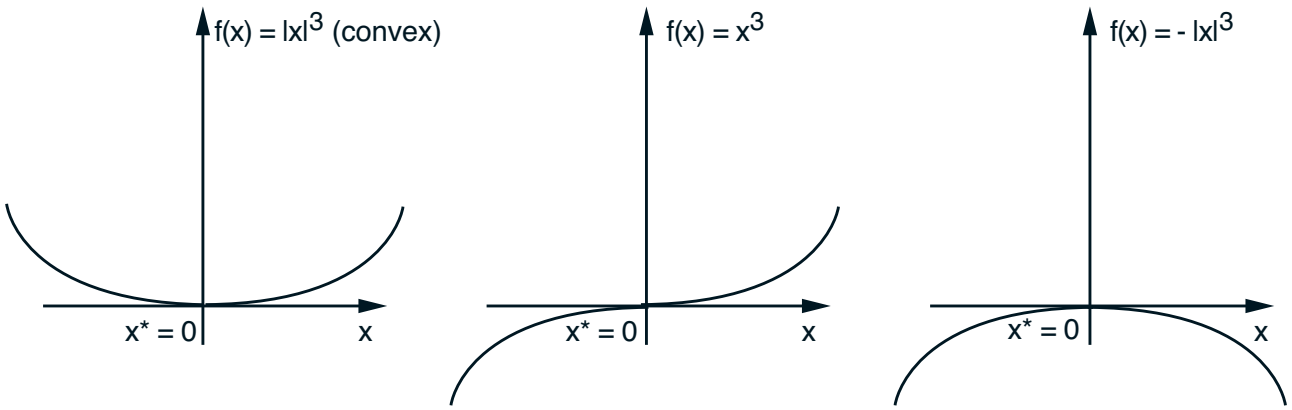
- **1st order condition:** Zero slope at a local minimum x^*

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

- **2nd order condition:** Nonnegative curvature at a local minimum x^*

$$\nabla^2 f(x^*) : \text{ Positive Semidefinite}$$

- There may exist points that satisfy the 1st and 2nd order conditions but are not local minima



First and second order necessary optimality conditions for functions of one variable.

PROOFS OF NECESSARY CONDITIONS

- **1st order condition $\nabla f(x^*) = 0$.** Fix $d \in \Re^n$. Then (since x^* is a local min), from 1st order Taylor

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

Replace d with $-d$, to obtain

$$d' \nabla f(x^*) = 0, \quad \forall \, d \in \Re^n$$

- **2nd order condition $\nabla^2 f(x^*) \geq 0$.** From 2nd order Taylor

$$f(x^*+\alpha d)-f(x^*) = \alpha \nabla f(x^*)' d + \frac{\alpha^2}{2} d' \nabla^2 f(x^*) d + o(\alpha^2)$$

Since $\nabla f(x^*) = 0$ and x^* is local min, there is sufficiently small $\epsilon > 0$ such that for all $\alpha \in (0, \epsilon)$,

$$0 \leq \frac{f(x^* + \alpha d) - f(x^*)}{\alpha^2} = \tfrac{1}{2} d' \nabla^2 f(x^*) d + \frac{o(\alpha^2)}{\alpha^2}$$

Take the limit as $\alpha \to 0$.

SUFFICIENT CONDITIONS FOR A LOCAL MIN

- **1st order condition:** Zero slope

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

- **1st order condition:** Positive curvature

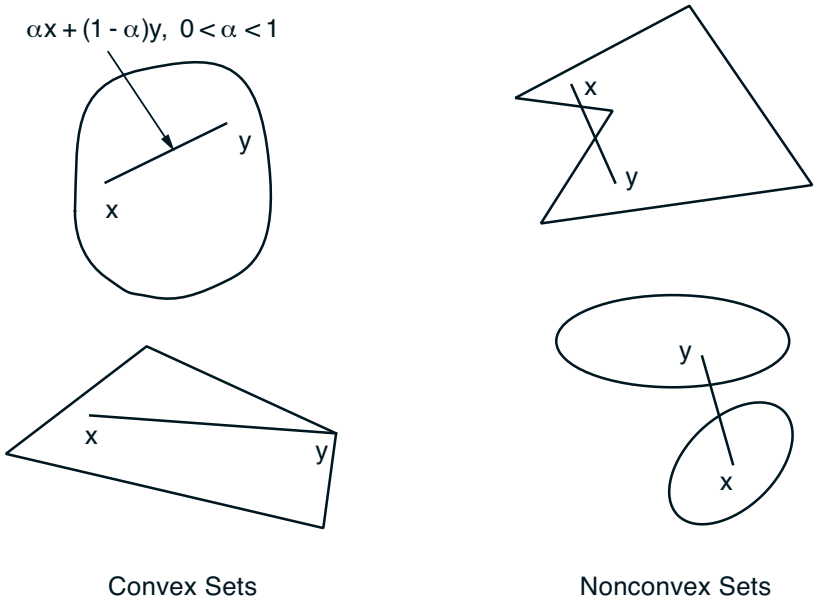
$$\nabla^2 f(x^*) : \text{Positive Definite}$$

- **Proof:** Let $\lambda > 0$ be the smallest eigenvalue of $\nabla^2 f(x^*)$. Using a second order Taylor expansion, we have for all d

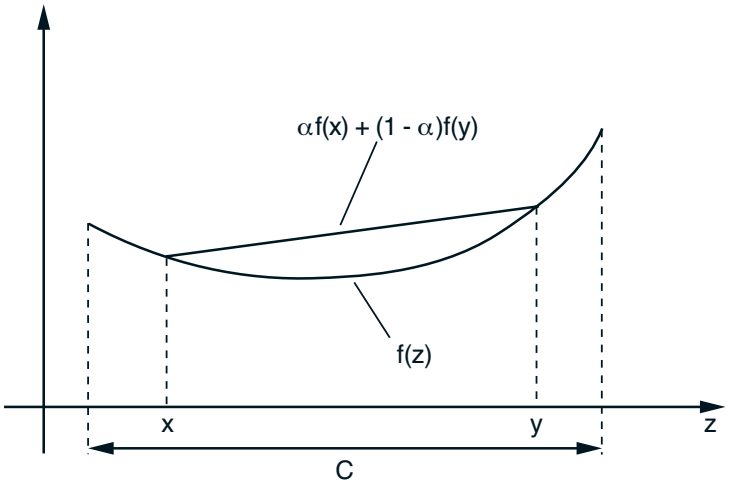
$$\begin{aligned} f(x^* + d) - f(x^*) &= \nabla f(x^*)' d + \frac{1}{2} d' \nabla^2 f(x^*) d \\ &\quad + o(\|d\|^2) \\ &\geq \frac{\lambda}{2} \|d\|^2 + o(\|d\|^2) \\ &= \left(\frac{\lambda}{2} + \frac{o(\|d\|^2)}{\|d\|^2} \right) \|d\|^2. \end{aligned}$$

For $\|d\|$ small enough, $o(\|d\|^2)/\|d\|^2$ is negligible relative to $\lambda/2$.

CONVEXITY



Convex and nonconvex sets.



A convex function. Linear interpolation underestimates the function.

MINIMA AND CONVEXITY

- Local minima are also global under convexity

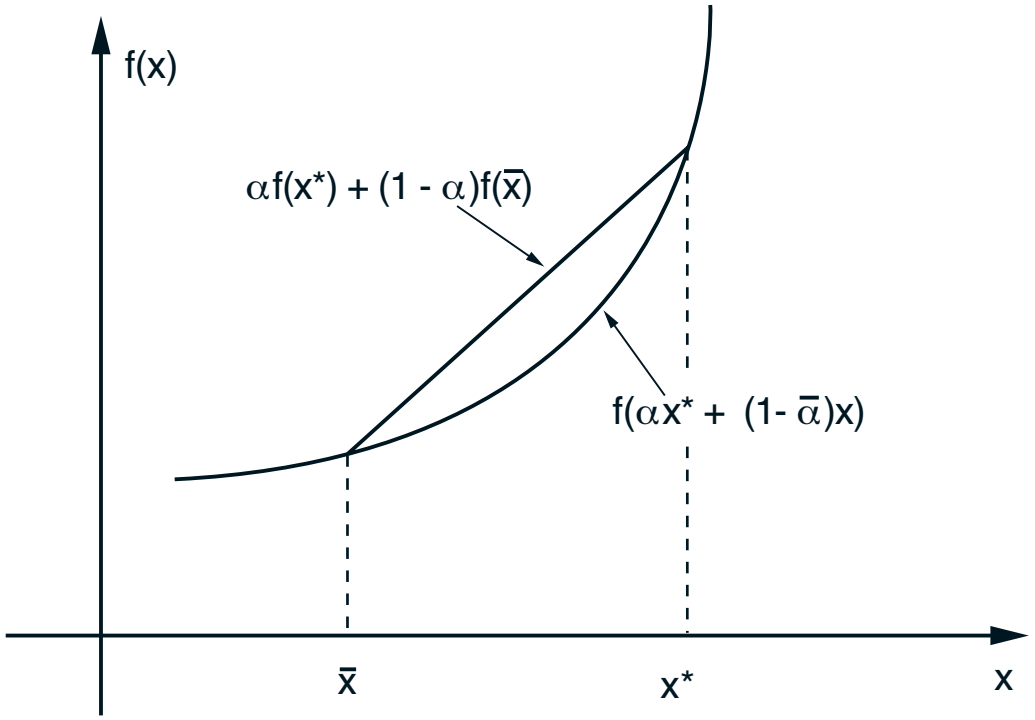


Illustration of why local minima of convex functions are also global. Suppose that f is convex and that x^* is a local minimum of f. Let $\overline{x}$ be such that $f(\overline{x}) < f(x^*)$. By convexity, for all $\alpha \in (0, 1)$,

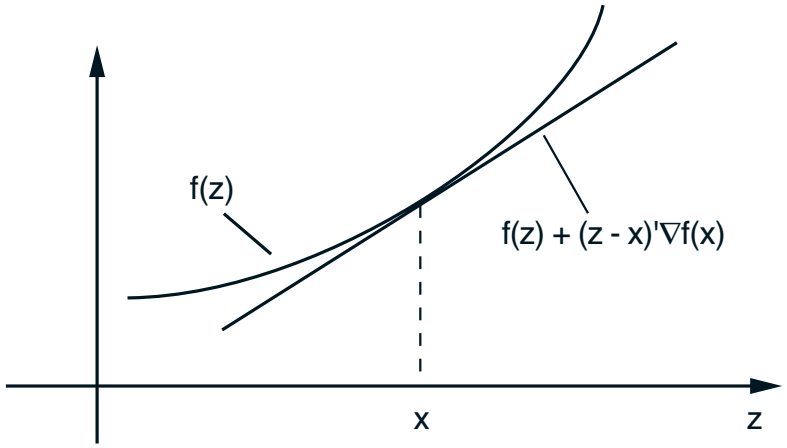
$$f\Big(\alpha x^* + (1-\alpha)\overline{x}\Big) \leq \alpha f(x^*) + (1-\alpha) f(\overline{x}) < f(x^*).$$

Thus, f takes values strictly lower than $f(x^*)$ on the line segment connecting x^* with $\overline{x}$, and x^* cannot be a local minimum which is not global.

OTHER PROPERTIES OF CONVEX FUNCTIONS

- f is convex if and only if the linear approximation at a point x based on the gradient, underestimates f:

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



  - Implication:

$$\nabla f(x^*) = 0 \quad \Rightarrow x^* \text{ is a global minimum}$$

- f is convex if and only if $\nabla^2 f(x)$ is positive semidefinite for all x

6.252 NONLINEAR PROGRAMMING

LECTURE 3: GRADIENT METHODS

LECTURE OUTLINE

- Quadratic Unconstrained Problems
- Existence of Optimal Solutions
- Iterative Computational Methods
- Gradient Methods - Motivation
- Principal Gradient Methods
- Gradient Methods - Choices of Direction

QUADRATIC UNCONSTRAINED PROBLEMS

$$\min_{x \in \Re^n} f(x) = \tfrac{1}{2} x'Qx - b'x,$$

where Q is $n \times n$ symmetric, and $b \in \Re^n$.

- Necessary conditions:

$$\nabla f(x^*) = Qx^* - b = 0,$$

$$\nabla^2 f(x^*) = Q \geq 0 \text{ : positive semidefinite.}$$

- $Q \geq 0 \ \Rightarrow \ f$: convex, nec. conditions are also sufficient, and local minima are also global
- Conclusions:
  - Q : not $\geq 0 \ \Rightarrow \ f$ has no local minima
  - If $Q > 0$ (and hence invertible), $x^* = Q^{-1}b$ is the unique global minimum.
  - If $Q \geq 0$ but not invertible, either no solution or ∞ number of solutions

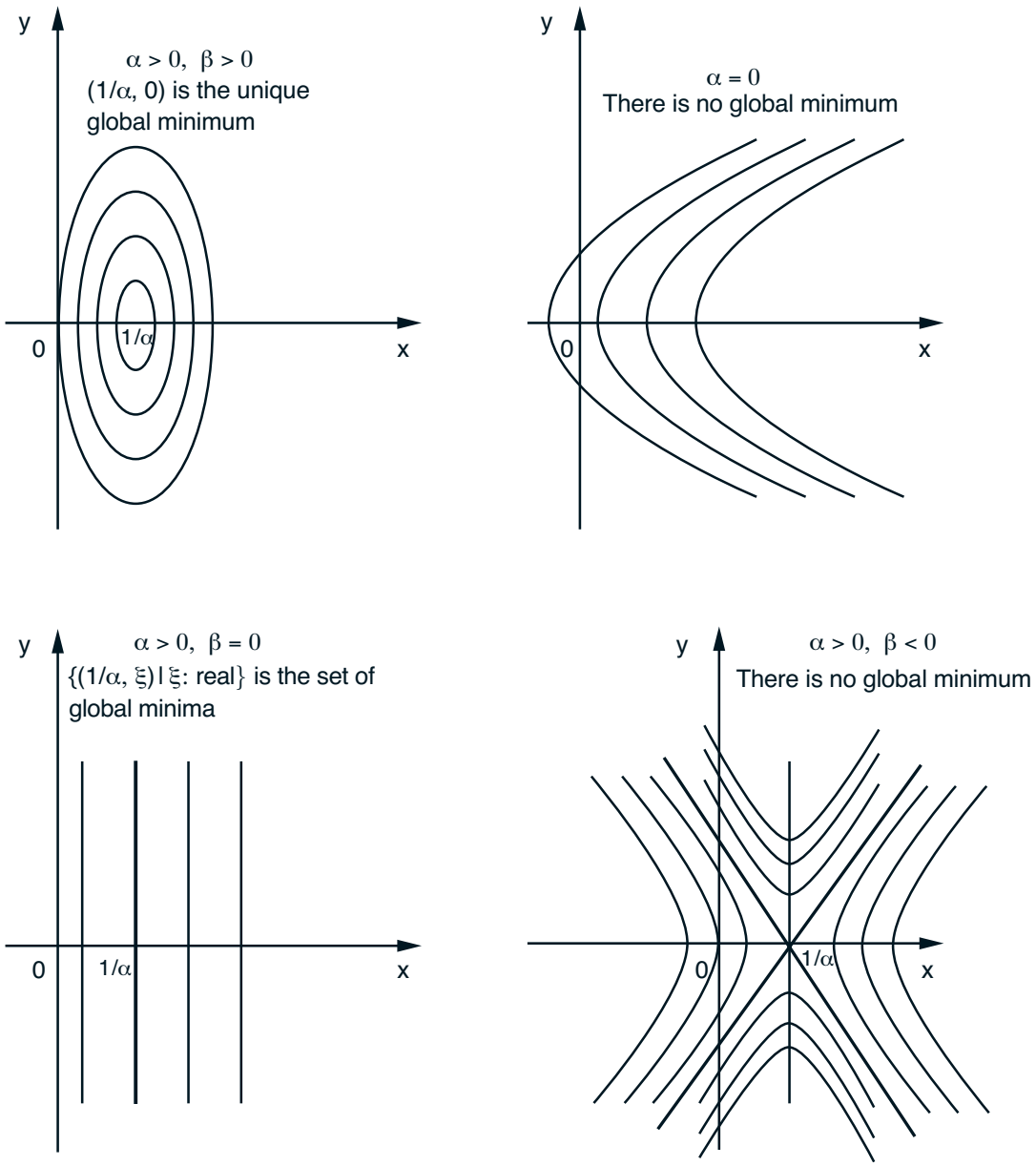


Illustration of the isocost surfaces of the quadratic cost function $f : \Re^2 \mapsto \Re$ given by

$$f(x, y) = \tfrac{1}{2}\left(\alpha x^2 + \beta y^2\right) - x$$

for various values of α and β.

EXISTENCE OF OPTIMAL SOLUTIONS

Consider the problem

$$\min_{x \in X} f(x)$$

- The set of optimal solutions is

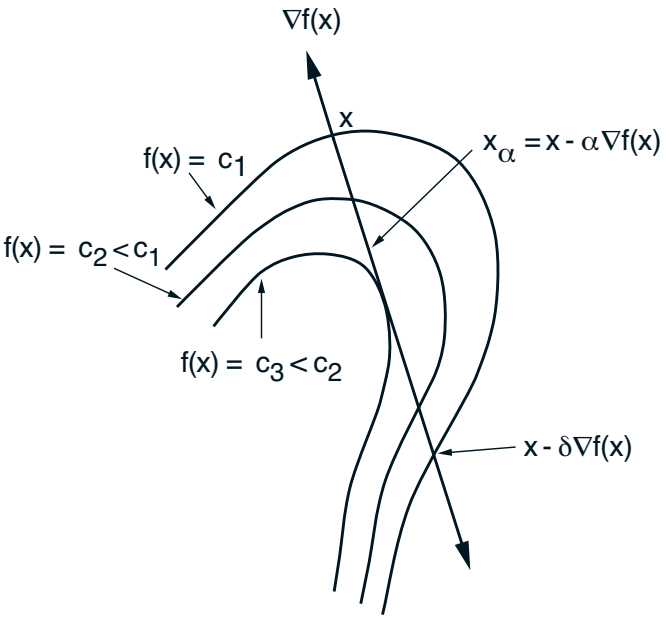
$$X^* = \cap_{k=1}^{\infty} \big\{ x \in X \mid f(x) \leq \gamma_k \big\}$$

where $\{\gamma_k\}$ is a scalar sequence such that $\gamma_k \downarrow f^*$ with

$$f^* = \inf_{x \in X} f(x)$$

- X^* is nonempty and compact if all the sets $\{x \in X \mid f(x) \leq \gamma_k\}$ are compact. So:
  - A global minimum exists if f is continuous and X is compact (Weierstrass theorem)
  - A global minimum exists if X is closed, and f is continuous and coercive, that is, $f(x) \to \infty$ when $\|x\| \to \infty$

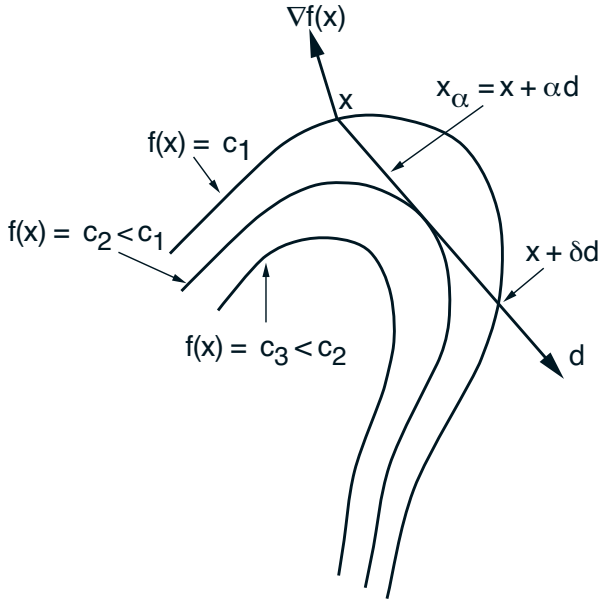
GRADIENT METHODS - MOTIVATION



If $\nabla f(x) \neq 0$, there is an interval $(0, \delta)$ of stepsizes such that

$$f\Big(x - \alpha\nabla f(x)\Big) < f(x)$$

for all $\alpha \in (0, \delta)$.



If d makes an angle with $\nabla f(x)$ that is greater than 90 degrees,

$$\nabla f(x)'d < 0,$$

there is an interval $(0, \delta)$ of stepsizes such that $f(x + \alpha d) < f(x)$ for all $\alpha \in (0, \delta)$.

PRINCIPAL GRADIENT METHODS

$$x^{k+1} = x^k + \alpha^k d^k, \qquad k = 0, 1, \ldots$$

where, if $\nabla f(x^k) \neq 0$, the direction d^k satisfies

$$\nabla f(x^k)' d^k < 0,$$

and α^k is a positive stepsize. Principal example:

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k),$$

where D^k is a positive definite symmetric matrix

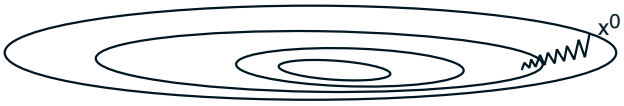
- Simplest method: $\mathbf{Steepest\ descent}$

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k), \qquad k = 0, 1, \ldots$$

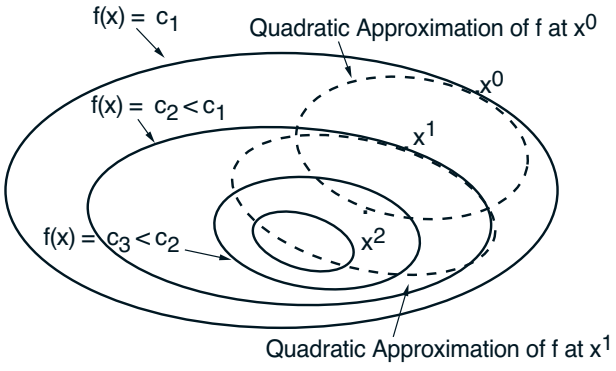
- Most sophisticated method: $\mathbf{Newton's\ method}$

$$x^{k+1} = x^k - \alpha^k \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k), \qquad k = 0, 1, \ldots$$

STEEPEST DESCENT AND NEWTON'S METHOD



Slow convergence of steepest descent



Fast convergence of Newton's method w/ $\alpha^k = 1$.

Given x^k, the method obtains x^{k+1} as the minimum of a quadratic approximation of f based on a second order Taylor expansion around x^k.

OTHER CHOICES OF DIRECTION

- **Diagonally Scaled Steepest Descent**

$$D^k = \text{ Diagonal approximation to } \left(\nabla^2 f(x^k)\right)^{-1}$$

- **Modified Newton's Method**

$$D^k = \left(\nabla^2 f(x^0)\right)^{-1}, \qquad k = 0, 1, \ldots,$$

- **Discretized Newton's Method**

$$D^k = \left(H(x^k)\right)^{-1}, \qquad k = 0, 1, \ldots,$$

where $H(x^k)$ is a finite-difference based approximation of $\nabla^2 f(x^k)$

- **Gauss-Newton method** for least squares problems: $\min_{x \in \Re^n} \frac{1}{2} \|g(x)\|^2$. Here

$$D^k = \left(\nabla g(x^k) \nabla g(x^k)'\right)^{-1}, \qquad k = 0, 1, \ldots$$

6.252 NONLINEAR PROGRAMMING

LECTURE 4

CONVERGENCE ANALYSIS OF GRADIENT METHODS

LECTURE OUTLINE

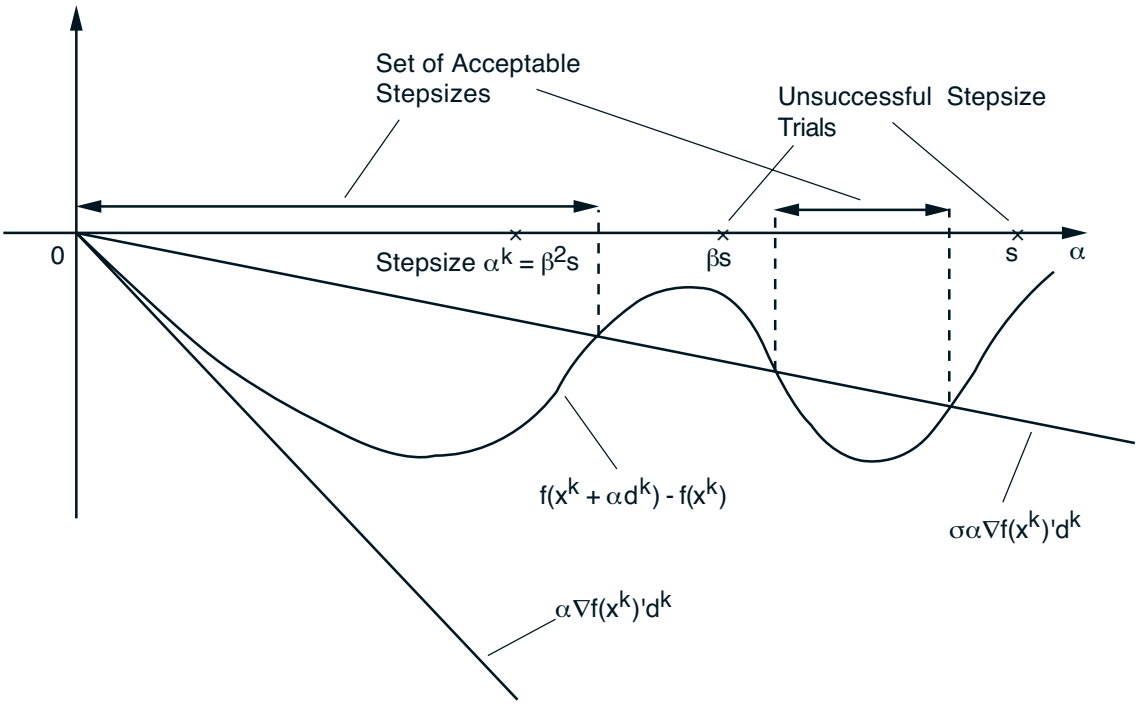
- Gradient Methods - Choice of Stepsize
- Gradient Methods - Convergence Issues

CHOICES OF STEPSIZE I

- Minimization Rule: α^k is such that

$$f(x^k + \alpha^k d^k) = \min_{\alpha \geq 0} f(x^k + \alpha d^k).$$

- Limited Minimization Rule: Min over $\alpha \in [0, s]$
- Armijo rule:



Start with s and continue with $\beta s, \beta^2 s, ...,$ until $\beta^m s$ falls within the set of α with

$$f(x^k) - f(x^k + \alpha d^k) \geq -\sigma\alpha\nabla f(x^k)' d^k.$$

CHOICES OF STEPSIZE II

- Constant stepsize: α^k is such that

$$\alpha^k = s : \text{ a constant}$$

- Diminishing stepsize:

$$\alpha^k \to 0$$

but satisfies the infinite travel condition

$$\sum_{k=0}^{\infty} \alpha^k = \infty$$

GRADIENT METHODS WITH ERRORS

$$x^{k+1} = x^k - \alpha^k(\nabla f(x^k) + e^k)$$

where e^k is an uncontrollable error vector

- Several special cases:
  - e^k small relative to the gradient; i.e., for all k, $\|e^k\| < \|\nabla f(x^k)\|$

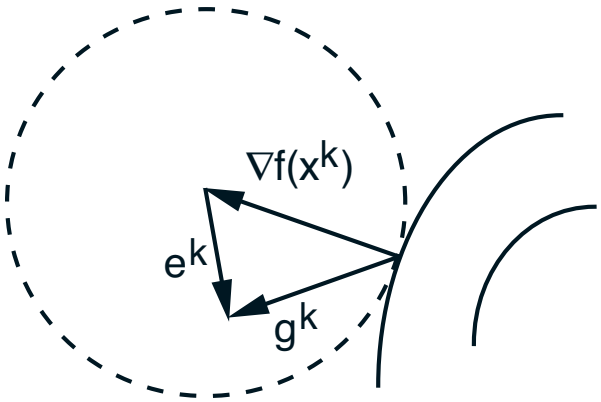


Illustration of the descent property of the direction $g^k = \nabla f(x^k) + e^k$.

  - $\{e^k\}$ is bounded, i.e., for all k, $\|e^k\| \leq \delta$, where δ is some scalar.
  - $\{e^k\}$ is proportional to the stepsize, i.e., for all k, $\|e^k\| \leq q\alpha^k$, where q is some scalar.
  - $\{e^k\}$ are independent zero mean random vectors

CONVERGENCE ISSUES

- Only convergence to stationary points can be guaranteed
- Even convergence to a single limit may be hard to guarantee (capture theorem)
- Danger of nonconvergence if directions d^k tend to be orthogonal to $\nabla f(x^k)$
- Gradient related condition:

For any subsequence $\{x^k\}_{k\in\mathcal{K}}$ that converges to a nonstationary point, the corresponding subsequence $\{d^k\}_{k\in\mathcal{K}}$ is bounded and satisfies

$$\limsup_{k\to\infty,\, k\in\mathcal{K}} \nabla f(x^k)' d^k < 0.$$

- Satisfied if $d^k = -D^k \nabla f(x^k)$ and the eigenvalues of D^k are bounded above and bounded away from zero

CONVERGENCE RESULTS

CONSTANT AND DIMINISHING STEPSIZES

Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, where $\{d^k\}$ is gradient related. Assume that for some constant $L > 0$, we have

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \qquad \forall\, x, y \in \Re^n,$$

Assume that either

(1) there exists a scalar ϵ such that for all k

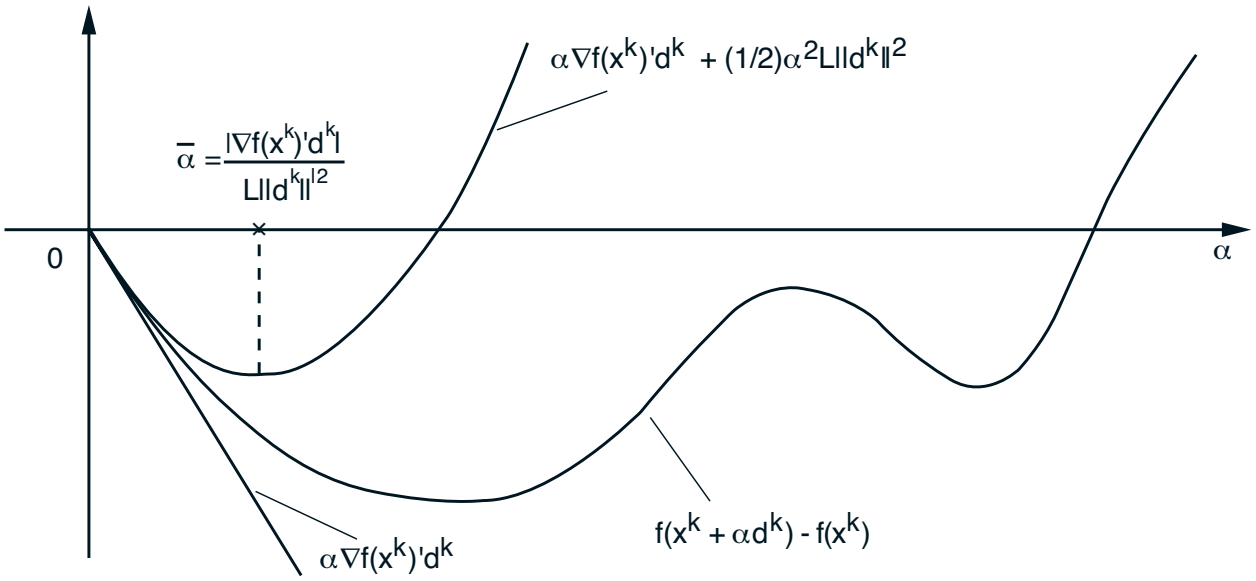
$$0 < \epsilon \leq \alpha^k \leq \frac{(2-\epsilon)|\nabla f(x^k)' d^k|}{L\|d^k\|^2}$$

or

(2) $\alpha^k \to 0$ and $\sum_{k=0}^{\infty} \alpha^k = \infty$.

Then either $f(x^k) \to -\infty$ or else $\{f(x^k)\}$ converges to a finite value and $\nabla f(x^k) \to 0$.

MAIN PROOF IDEA



The idea of the convergence proof for a constant stepsize. Given x^k and the descent direction d^k, the cost difference $f(x^k + \alpha d^k) - f(x^k)$ is majorized by $\alpha \nabla f(x^k)' d^k + \frac{1}{2}\alpha^2 L \|d^k\|^2$ (based on the Lipschitz assumption; see next slide). Minimization of this function over α yields the stepsize

$$\overline{\alpha} = \frac{|\nabla f(x^k)' d^k|}{L\|d^k\|^2}$$

This stepsize reduces the cost function f as well.

DESCENT LEMMA

Let α be a scalar and let $g(\alpha) = f(x + \alpha y)$. Have

$$
\begin{aligned}
f(x+y) - f(x) = g(1) - g(0) &= \int_0^1 \frac{dg}{d\alpha}(\alpha)\, d\alpha \\
&= \int_0^1 y' \nabla f(x + \alpha y)\, d\alpha \\
&\leq \int_0^1 y' \nabla f(x)\, d\alpha \\
&+ \left| \int_0^1 y' \big(\nabla f(x + \alpha y) - \nabla f(x) \big)\, d\alpha \right| \\
&\leq \int_0^1 y' \nabla f(x)\, d\alpha \\
&+ \int_0^1 \|y\| \cdot \|\nabla f(x + \alpha y) - \nabla f(x)\| d\alpha \\
&\leq y' \nabla f(x) + \|y\| \int_0^1 L\alpha \|y\|\, d\alpha \\
&= y' \nabla f(x) + \frac{L}{2} \|y\|^2.
\end{aligned}
$$

CONVERGENCE RESULT – ARMIJO RULE

Let $\{x^k\}$ be generated by $x^{k+1} = x^k + \alpha^k d^k$, where $\{d^k\}$ is gradient related and α^k is chosen by the Armijo rule. Then every limit point of $\{x^k\}$ is stationary.

Proof Outline: Assume $\overline{x}$ is a nonstationary limit point. Then $f(x^k) \to f(\overline{x})$, so $\alpha^k \nabla f(x^k)' d^k \to 0$.

- If $\{x^k\}_{\mathcal{K}} \to \overline{x}$, $\limsup_{k\to\infty,\, k\in\mathcal{K}} \nabla f(x^k)' d^k < 0$, by gradient relatedness, so that $\{\alpha^k\}_{\mathcal{K}} \to 0$.
- By the Armijo rule, for large $k \in \mathcal{K}$

$$f(x^k) - f\left(x^k + (\alpha^k/\beta) d^k\right) < -\sigma(\alpha^k/\beta)\nabla f(x^k)' d^k.$$

Defining $p^k = \frac{d^k}{\|d^k\|}$ and $\overline{\alpha}^k = \frac{\alpha^k \|d^k\|}{\beta}$, we have

$$\frac{f(x^k) - f(x^k + \overline{\alpha}^k p^k)}{\overline{\alpha}^k} < -\sigma \nabla f(x^k)' p^k.$$

Use the Mean Value Theorem and let $k \to \infty$. We get $-\nabla f(\overline{x})' \overline{p} \leq -\sigma \nabla f(\overline{x})' \overline{p}$, where $\overline{p}$ is a limit point of p^k – a contradiction since $\nabla f(\overline{x})' \overline{p} < 0$.

6.252 NONLINEAR PROGRAMMING

LECTURE 5: RATE OF CONVERGENCE

LECTURE OUTLINE

- Approaches for Rate of Convergence Analysis
- The Local Analysis Method
- Quadratic Model Analysis
- The Role of the Condition Number
- Scaling
- Diagonal Scaling
- Extension to Nonquadratic Problems
- Singular and Difficult Problems

APPROACHES FOR RATE OF CONVERGENCE ANALYSIS

- Computational complexity approach
- Informational complexity approach
- Local analysis
- Why we will focus on the local analysis method

THE LOCAL ANALYSIS APPROACH

- Restrict attention to sequences x^k converging to a local min x^*
- Measure progress in terms of an error function $e(x)$ with $e(x^*) = 0$, such as

$$e(x) = \|x - x^*\|, \qquad e(x) = f(x) - f(x^*)$$

- Compare the tail of the sequence $e(x^k)$ with the tail of standard sequences
- Geometric or linear convergence [if $e(x^k) \leq q\beta^k$ for some $q > 0$ and $\beta \in [0, 1)$, and for all k]. Holds if

$$\limsup_{k \to \infty} \frac{e(x^{k+1})}{e(x^k)} < \beta$$

- Superlinear convergence [if $e(x^k) \leq q \cdot \beta^{p^k}$ for some $q > 0$, $p > 1$ and $\beta \in [0, 1)$, and for all k].
- Sublinear convergence

QUADRATIC MODEL ANALYSIS

- Focus on the quadratic function $f(x) = (1/2)x'Qx$, with $Q > 0$.
- Analysis also applies to nonquadratic problems in the neighborhood of a nonsingular local min
- Consider steepest descent

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = (I - \alpha^k Q)x^k$$

$$\begin{aligned} \|x^{k+1}\|^2 &= x^{k\prime}(I - \alpha^k Q)^2 x^k \\ &\leq \big(\max \text{ eig. } (I - \alpha^k Q)^2\big)\|x^k\|^2 \end{aligned}$$

The eigenvalues of $(I - \alpha^k Q)^2$ are equal to $(1 - \alpha^k \lambda_i)^2$, where λ_i are the eigenvalues of Q, so

$$\max \text{ eig of } (I-\alpha^k Q)^2 = \max\big\{(1-\alpha^k m)^2, (1-\alpha^k M)^2\big\}$$

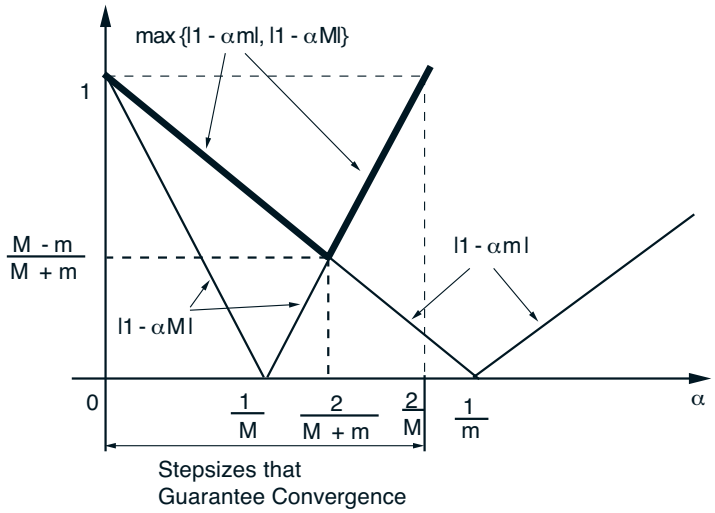
where m, M are the smallest and largest eigenvalues of Q. Thus

$$\frac{\|x^{k+1}\|}{\|x^k\|} \leq \max\big\{|1 - \alpha^k m|, |1 - \alpha^k M|\big\}$$

OPTIMAL CONVERGENCE RATE

- The value of α^k that minimizes the bound is $\alpha^* = 2/(M+m)$, in which case

$$\frac{\|x^{k+1}\|}{\|x^k\|} \leq \frac{M-m}{M+m}$$



- Conv. rate for minimization stepsize (see text)

$$\frac{f(x^{k+1})}{f(x^k)} \leq \left(\frac{M-m}{M+m}\right)^2$$

- The ratio M/m is called the $condition\ number$ of Q, and problems with M/m: large are called $ill\text{-}conditioned$.

SCALING AND STEEPEST DESCENT

- View the more general method

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k)$$

as a scaled version of steepest descent.

- Consider a change of variables $x = Sy$ with $S = (D^k)^{1/2}$. In the space of y, the problem is

$$\begin{aligned} &\text{minimize} \quad h(y) \equiv f(Sy) \\ &\text{subject to} \quad y \in \Re^n \end{aligned}$$

- Apply steepest descent to this problem, multiply with S, and pass back to the space of x, using $\nabla h(y^k) = S \nabla f(x^k)$,

$$y^{k+1} = y^k - \alpha^k \nabla h(y^k)$$

$$Sy^{k+1} = Sy^k - \alpha^k S \nabla h(y^k)$$

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k)$$

DIAGONAL SCALING

- Apply the results for steepest descent to the scaled iteration $y^{k+1} = y^k - \alpha^k \nabla h(y^k)$:

$$\frac{\|y^{k+1}\|}{\|y^k\|} \leq \max\{|1 - \alpha^k m^k|, |1 - \alpha^k M^k|\}$$

$$\frac{f(x^{k+1})}{f(x^k)} = \frac{h(y^{k+1})}{h(y^k)} \leq \left(\frac{M^k - m^k}{M^k + m^k}\right)^2$$

where m^k and M^k are the smallest and largest eigenvalues of the Hessian of h, which is

$$\nabla^2 h(y) = S \nabla^2 f(x) S = (D^k)^{1/2} Q (D^k)^{1/2}$$

- It is desirable to choose D^k as close as possible to Q^{-1}. Also if D^k is so chosen, the stepsize $\alpha = 1$ is near the optimal $2/(M^k + m^k)$.

- Using as D^k a diagonal approximation to Q^{-1} is common and often very effective. Corrects for poor choice of units expressing the variables.

NONQUADRATIC PROBLEMS

- Rate of convergence to a nonsingular local minimum of a nonquadratic function is very similar to the quadratic case (linear convergence is typical).

- If $D^k \to \left(\nabla^2 f(x^*)\right)^{-1}$, we asymptotically obtain optimal scaling and superlinear convergence

- More generally, if the direction $d^k = -D^k \nabla f(x^k)$ approaches asymptotically the Newton direction, i.e.,

$$\lim_{k\to\infty} \frac{\|d^k + \left(\nabla^2 f(x^*)\right)^{-1} \nabla f(x^k)\|}{\|\nabla f(x^k)\|} = 0$$

and the Armijo rule is used with initial stepsize equal to one, the rate of convergence is superlinear.

- Convergence rate to a singular local min is typically sublinear (in effect, condition number $= \infty$)

6.252 NONLINEAR PROGRAMMING

LECTURE 6

NEWTON AND GAUSS-NEWTON METHODS

LECTURE OUTLINE

- Newton's Method
- Convergence Rate of the Pure Form
- Global Convergence
- Variants of Newton's Method
- Least Squares Problems
- The Gauss-Newton Method

NEWTON'S METHOD

$$x^{k+1} = x^k - \alpha^k \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$$

assuming that the Newton direction is defined and is a direction of descent

- Pure form of Newton's method (stepsize $= 1$)

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$$

  - Very fast when it converges (how fast?)
  - May not converge (or worse, it may not be defined) when started far from a nonsingular local min
  - Issue: How to modify the method so that it converges globally, while maintaining the fast convergence rate

CONVERGENCE RATE OF PURE FORM

- Consider solution of nonlinear system $g(x) = 0$ where $g : \Re^n \mapsto \Re^n$, with method

$$x^{k+1} = x^k - \big(\nabla g(x^k)'\big)^{-1} g(x^k)$$

  - If $g(x) = \nabla f(x)$, we get pure form of Newton

- **Quick derivation:** Suppose $x^k \to x^*$ with $g(x^*) = 0$ and $\nabla g(x^*)$ is invertible. By Taylor

$$0 = g(x^*) = g(x^k) + \nabla g(x^k)'(x^* - x^k) + o\big(\|x^k - x^*\|\big).$$

Multiply with $\big(\nabla g(x^k)'\big)^{-1}$:

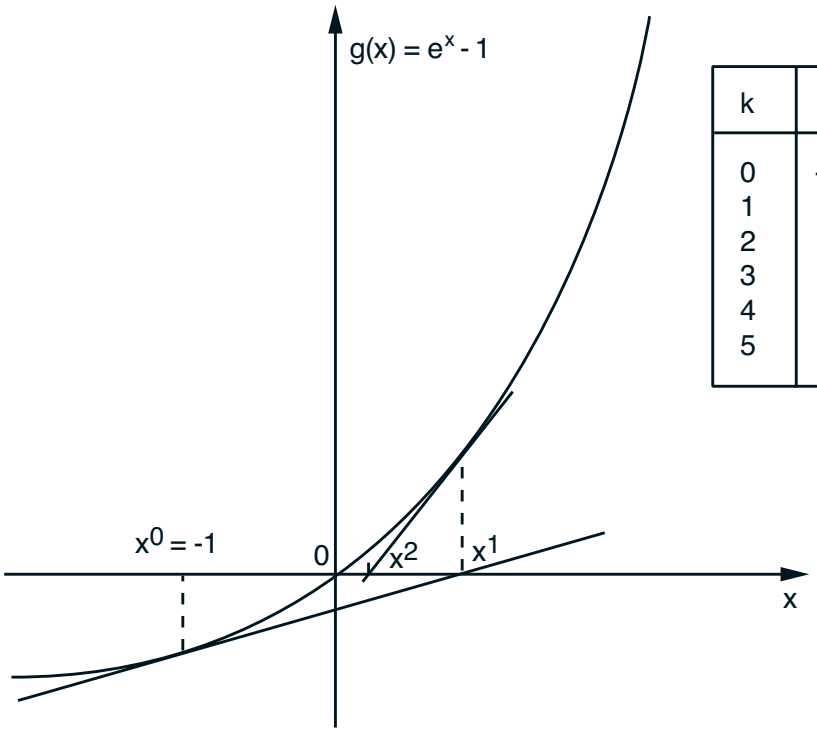
$$x^k - x^* - \big(\nabla g(x^k)'\big)^{-1} g(x^k) = o\big(\|x^k - x^*\|\big),$$

so

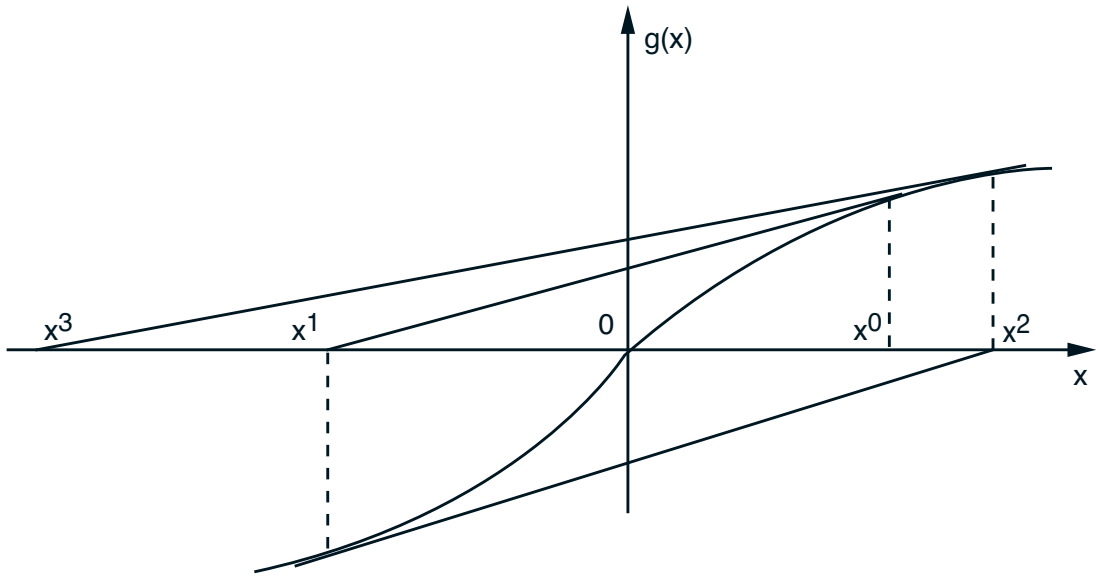
$$x^{k+1} - x^* = o\big(\|x^k - x^*\|\big),$$

implying superlinear convergence and capture.

CONVERGENCE BEHAVIOR OF PURE FORM



k	x^k	$g(x^k)$
0	- 1.00000	- 0.63212
1	0.71828	1.05091
2	0.20587	0.22859
3	0.01981	0.02000
4	0.00019	0.00019
5	0.00000	0.00000



MODIFICATIONS FOR GLOBAL CONVERGENCE

- Use a stepsize
- Modify the Newton direction when:
  - Hessian is not positive definite
  - When Hessian is nearly singular (needed to improve performance)
- Use

$$d^k = -\left(\nabla^2 f(x^k) + \Delta^k\right)^{-1} \nabla f(x^k),$$

whenever the Newton direction does not exist or is not a descent direction. Here Δ^k is a diagonal matrix such that

$$\nabla^2 f(x^k) + \Delta^k > 0$$

  - Modified Cholesky factorization
  - Trust region methods

LEAST-SQUARES PROBLEMS

$$\text{minimize} \quad f(x) = \tfrac{1}{2}\|g(x)\|^2 = \tfrac{1}{2}\sum_{i=1}^{m}\|g_i(x)\|^2$$

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

where $g = (g_1, \ldots, g_m)$, $g_i : \Re^n \to \Re^{r_i}$.

- Many applications:
  - Solution of systems of n nonlinear equations with n unknowns
  - Model Construction – Curve Fitting
  - Neural Networks
  - Pattern Classification

PURE FORM OF THE GAUSS-NEWTON METHOD

- Idea: Linearize around the current point x^k

$$\tilde{g}(x, x^k) = g(x^k) + \nabla g(x^k)'(x - x^k)$$

and minimize the norm of the linearized function $\tilde{g}$:

$$\begin{aligned} x^{k+1} &= \arg \min_{x \in \Re^n} \tfrac{1}{2} \|\tilde{g}(x, x^k)\|^2 \\ &= x^k - \big(\nabla g(x^k) \nabla g(x^k)'\big)^{-1} \nabla g(x^k) g(x^k) \end{aligned}$$

- The direction

$$-\big(\nabla g(x^k) \nabla g(x^k)'\big)^{-1} \nabla g(x^k) g(x^k)$$

is a descent direction since

$$\nabla g(x^k) g(x^k) = \nabla \big((1/2)\|g(x)\|^2\big)$$

$$\nabla g(x^k) \nabla g(x^k)' > 0$$

MODIFICATIONS OF THE GAUSS-NEWTON

- Similar to those for Newton's method:

$$x^{k+1} = x^k - \alpha^k \left(\nabla g(x^k) \nabla g(x^k)' + \Delta^k\right)^{-1} \nabla g(x^k) g(x^k)$$

where α^k is a stepsize and Δ^k is a diagonal matrix such that

$$\nabla g(x^k) \nabla g(x^k)' + \Delta^k > 0$$

- Incremental version of the Gauss-Newton method:
  - Operate in cycles
  - Start a cycle with ψ_0 (an estimate of x)
  - Update ψ using a $single$ component of g

$$\psi_i = \arg \min_{x \in \Re^n} \sum_{j=1}^{i} \|\tilde{g}_j(x, \psi_{j-1})\|^2, \ i = 1, \ldots, m,$$

where $\tilde{g}_j$ are the linearized functions

$$\tilde{g}_j(x, \psi_{j-1}) = g_j(\psi_{j-1}) + \nabla g_j(\psi_{j-1})'(x - \psi_{j-1})$$

MODEL CONSTRUCTION

- Given set of m input-output data pairs (y_i, z_i), $i = 1, \ldots, m$, from the physical system

- Hypothesize an input/output relation $z = h(x, y)$, where x is a vector of unknown parameters, and h is known

- Find x that matches best the data in the sense that it minimizes the sum of squared errors

$$\tfrac{1}{2} \sum_{i=1}^{m} \|z_i - h(x, y_i)\|^2$$

- Example of a linear model: Fit the data pairs by a cubic polynomial approximation. Take

$$h(x, y) = x_3 y^3 + x_2 y^2 + x_1 y + x_0,$$

where $x = (x_0, x_1, x_2, x_3)$ is the vector of unknown coefficients of the cubic polynomial.

NEURAL NETS

- Nonlinear model construction with multilayer perceptrons
- x of the vector of weights
- Universal approximation property

PATTERN CLASSIFICATION

- Objects are presented to us, and we wish to classify them in one of s categories $1, \ldots, s$, based on a vector y of their features.

- Classical maximum posterior probability approach: Assume we know

$$p(j|y) = P(\text{object w/ feature vector } y \text{ is of category } j)$$

Assign object with feature vector y to category

$$j^*(y) = \arg \max_{j=1,\ldots,s} p(j|y).$$

- If $p(j|y)$ are unknown, we can estimate them using functions $h_j(x_j, y)$ parameterized by vectors x_j. Obtain x_j by minimizing

$$\tfrac{1}{2} \sum_{i=1}^{m} \big(z_j^i - h_j(x_j, y_i)\big)^2,$$

where

$$z_j^i = \begin{cases} 1 & \text{if } y_i \text{ is of category } j, \\ 0 & \text{otherwise.} \end{cases}$$

6.252 NONLINEAR PROGRAMMING

LECTURE 7: ADDITIONAL METHODS

LECTURE OUTLINE

- Least-Squares Problems and Incremental Gradient Methods
- Conjugate Direction Methods
- The Conjugate Gradient Method
- Quasi-Newton Methods
- Coordinate Descent Methods

- Recall the least-squares problem:

$$\text{minimize} \quad f(x) = \tfrac{1}{2}\|g(x)\|^2 = \tfrac{1}{2}\sum_{i=1}^{m}\|g_i(x)\|^2$$

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

where $g = (g_1, \ldots, g_m)$, $g_i : \Re^n \to \Re^{r_i}$.

INCREMENTAL GRADIENT METHODS

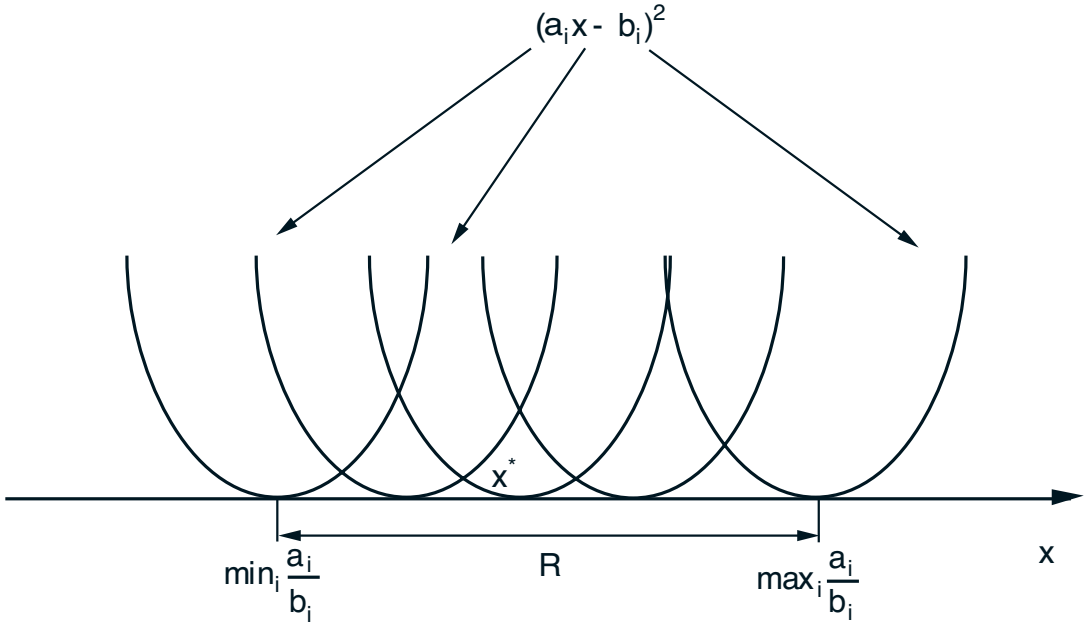
- Steepest descent method

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = x^k - \alpha^k \sum_{i=1}^{m} \nabla g_i(x^k) g_i(x^k)$$

- Incremental gradient method:

$$\psi_i = \psi_{i-1} - \alpha^k \nabla g_i(\psi_{i-1}) g_i(\psi_{i-1}), \quad i = 1, \ldots, m$$

$$\psi_0 = x^k, \qquad x^{k+1} = \psi_m$$



Advantage of incrementalism

VIEW AS GRADIENT METHOD W/ ERRORS

- Can write incremental gradient method as

$$x^{k+1} = x^k - \alpha^k \sum_{i=1}^{m} \nabla g_i(x^k) g_i(x^k)$$

$$+ \alpha^k \sum_{i=1}^{m} \big(\nabla g_i(x^k) g_i(x^k) - \nabla g_i(\psi_{i-1}) g_i(\psi_{i-1}) \big)$$

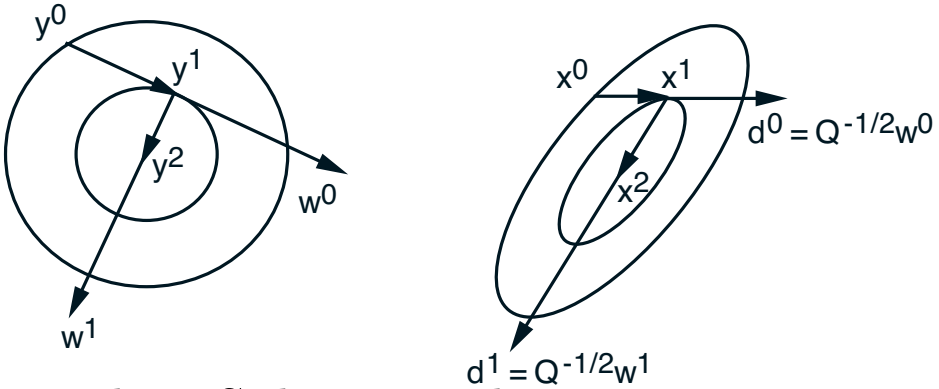
- Error term is proportional to stepsize α^k
- Convergence (generically) for a diminishing stepsize (under a Lipschitz condition on $\nabla g_i g_i$)
- Convergence to a "neighborhood" for a constant stepsize

CONJUGATE DIRECTION METHODS

- Aim to improve convergence rate of steepest descent, without the overhead of Newton's method.
- Analyzed for a quadratic model. They require n iterations to minimize $f(x) = (1/2)x'Qx - b'x$ with Q an $n \times n$ positive definite matrix $Q > 0$.
- Analysis also applies to nonquadratic problems in the neighborhood of a nonsingular local min.
- The directions $d^1, \ldots, d^k$ are $Q\text{-}conjugate$ if $d^{i\prime} Q d^j = 0$ for all $i \neq j$.
- Generic conjugate direction method:

$$x^{k+1} = x^k + \alpha^k d^k$$

where α^k is obtained by line minimization.



Expanding Subspace Theorem

GENERATING Q-CONJUGATE DIRECTIONS

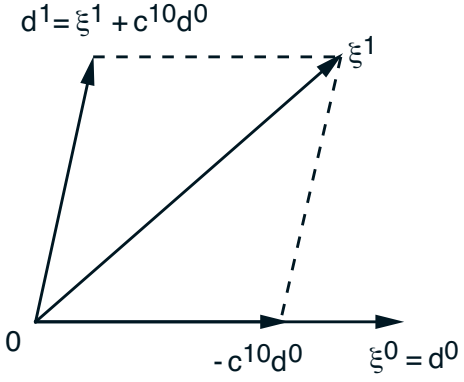
- Given set of linearly independent vectors $\xi^0, \ldots, \xi^k$, we can construct a set of Q-conjugate directions $d^0, \ldots, d^k$ s.t. $Span(d^0, \ldots, d^i) = Span(\xi^0, \ldots, \xi^i)$

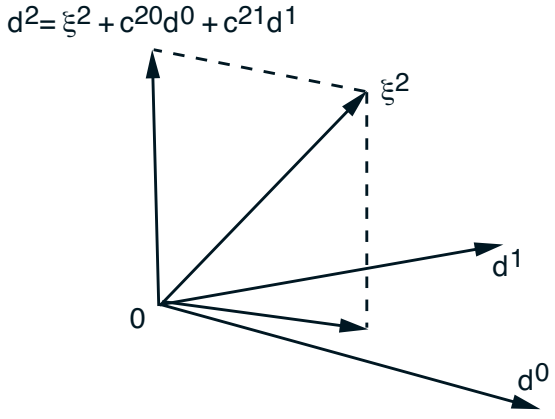
- *Gram-Schmidt procedure*. Start with $d^0 = \xi^0$. If for some $i < k$, $d^0, \ldots, d^i$ are Q-conjugate and the above property holds, take

$$d^{i+1} = \xi^{i+1} + \sum_{m=0}^{i} c^{(i+1)m} d^m;$$

choose $c^{(i+1)m}$ so d^{i+1} is Q-conjugate to $d^0, \ldots, d^i$,

$$d^{i+1\prime} Q d^j = \xi^{i+1\prime} Q d^j + \left(\sum_{m=0}^{i} c^{(i+1)m} d^m \right)' Q d^j = 0.$$





CONJUGATE GRADIENT METHOD

- Apply Gram-Schmidt to the vectors $\xi^k = -g^k = -\nabla f(x^k)$, $k = 0, 1, \ldots, n-1$. Then

$$d^k = -g^k + \sum_{j=0}^{k-1} \frac{g^{k\prime} Q d^j}{d^{j\prime} Q d^j} d^j$$

- **Key fact:** Direction formula can be simplified.

Proposition : The directions of the CGM are generated by $d^0 = -g^0$, and

$$d^k = -g^k + \beta^k d^{k-1}, \qquad k = 1, \ldots, n-1,$$

where β^k is given by

$$\beta^k = \frac{g^{k\prime} g^k}{g^{k-1\prime} g^{k-1}} \quad \text{or} \quad \beta^k = \frac{(g^k - g^{k-1})' g^k}{g^{k-1\prime} g^{k-1}}$$

Furthermore, the method terminates with an optimal solution after at most n steps.

- Extension to nonquadratic problems.

PROOF OF CONJUGATE GRADIENT RESULT

- Use induction to show that all gradients g^k generated up to termination are linearly independent. True for $k = 1$. Suppose no termination after k steps, and $g^0, \ldots, g^{k-1}$ are linearly independent. Then, $Span(d^0, \ldots, d^{k-1}) = Span(g^0, \ldots, g^{k-1})$ and there are two possibilities:
  - $g^k = 0$, and the method terminates.
  - $g^k \neq 0$, in which case from the expanding manifold property

$$g^k \text{ is orthogonal to } d^0, \ldots, d^{k-1}$$

$$g^k \text{ is orthogonal to } g^0, \ldots, g^{k-1}$$

    so g^k is linearly independent of $g^0, \ldots, g^{k-1}$, completing the induction.

- Since at most n lin. independent gradients can be generated, $g^k = 0$ for some $k \leq n$.
- Algebra to verify the direction formula.

QUASI-NEWTON METHODS

- $x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k)$, where D^k is an inverse Hessian approximation.

- Key idea: Successive iterates x^k, x^{k+1} and gradients $\nabla f(x^k)$, $\nabla f(x^{k+1})$, yield curvature info

$$q^k \approx \nabla^2 f(x^{k+1}) p^k,$$

$$p^k = x^{k+1} - x^k, \quad q^k = \nabla f(x^{k+1}) - \nabla f(x^k),$$

$$\nabla^2 f(x^n) \approx \left[q^0 \ \cdots \ q^{n-1} \right] \left[p^0 \ \cdots \ p^{n-1} \right]^{-1}$$

- Most popular Quasi-Newton method is a clever way to implement this idea

$$D^{k+1} = D^k + \frac{p^k p^{k\prime}}{p^{k\prime} q^k} - \frac{D^k q^k q^{k\prime} D^k}{q^{k\prime} D^k q^k} + \xi^k \tau^k v^k v^{k\prime},$$

$$v^k = \frac{p^k}{p^{k\prime} q^k} - \frac{D^k q^k}{\tau^k}, \quad \tau^k = q^{k\prime} D^k q^k, \quad 0 \leq \xi^k \leq 1$$

and $D^0 > 0$ is arbitrary, α^k by line minimization, and $D^n = Q^{-1}$ for a quadratic.

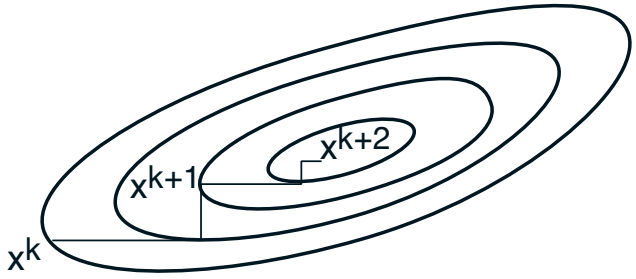
NONDERIVATIVE METHODS

- Finite difference implementations
- Forward and central difference formulas

$$\frac{\partial f(x^k)}{\partial x^i} \approx \frac{1}{h}\big(f(x^k + he_i) - f(x^k)\big)$$

$$\frac{\partial f(x^k)}{\partial x^i} \approx \frac{1}{2h}\big(f(x^k + he_i) - f(x^k - he_i)\big)$$

- Use central difference for more accuracy near convergence



- Coordinate descent. Applies also to the case where there are bound constraints on the variables.

- Direct search methods. Nelder-Mead method.

6.252 NONLINEAR PROGRAMMING

LECTURE 8

OPTIMIZATION OVER A CONVEX SET;

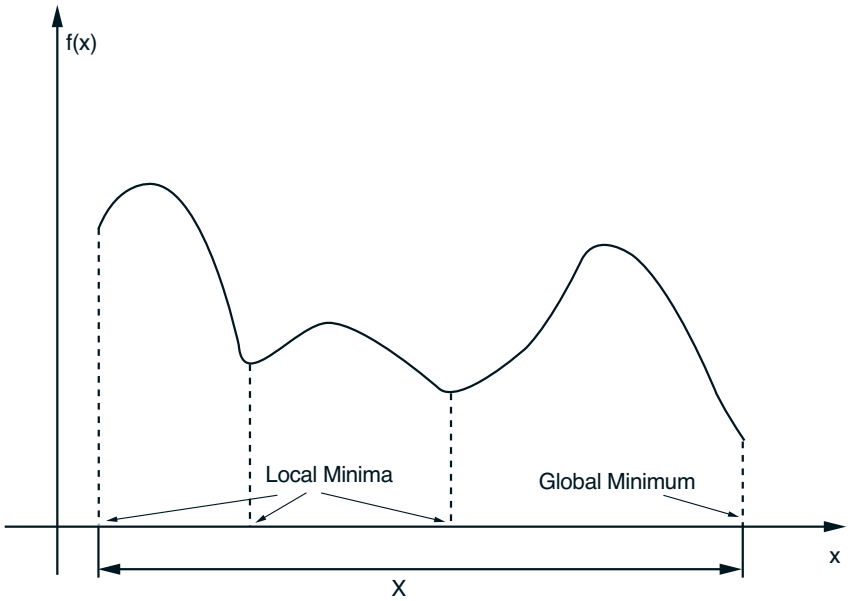
OPTIMALITY CONDITIONS

Problem: $\min_{x \in X} f(x)$, where:

(a) $X \subset \Re^n$ is nonempty, convex, and closed.

(b) f is continuously differentiable over X.

• Local and global minima. If f is convex local minima are also global.



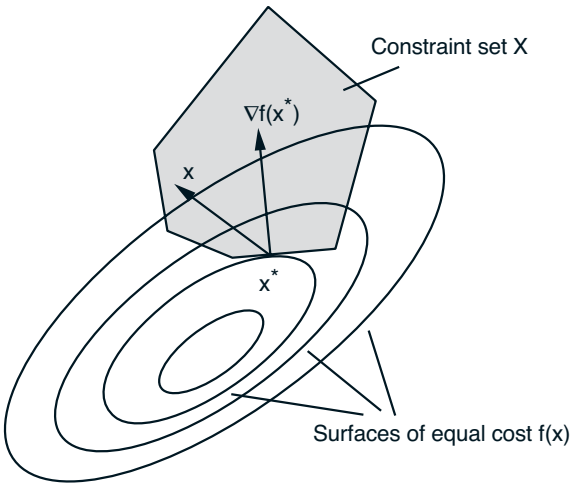
OPTIMALITY CONDITION

Proposition (Optimality Condition)

(a) If x^* is a local minimum of f over X, then

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

(b) If f is convex over X, then this condition is also sufficient for x^* to minimize f over X.



At a local minimum x^*, the gradient $\nabla f(x^*)$ makes an angle less than or equal to 90 degrees with all feasible variations $x - x^*$, $x \in X$.

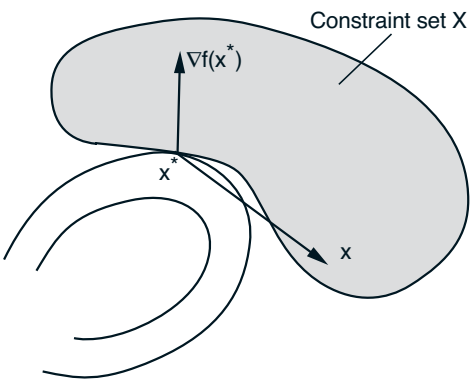


Illustration of failure of the optimality condition when X is not convex. Here x^* is a local min but we have $\nabla f(x^*)'(x - x^*) < 0$ for the feasible vector x shown.

PROOF

Proof: (a) By contradiction. Suppose that $\nabla f(x^*)'(x-x^*) < 0$ for some $x \in X$. By the Mean Value Theorem, for every $\epsilon > 0$ there exists an $s \in [0,1]$ such that

$$f\big(x^*+\epsilon(x-x^*)\big) = f(x^*)+\epsilon\nabla f\big(x^*+s\epsilon(x-x^*)\big)'(x-x^*).$$

Since ∇f is continuous, for suff. small $\epsilon > 0$,

$$\nabla f\big(x^* + s\epsilon(x - x^*)\big)'(x - x^*) < 0$$

so that $f\big(x^* + \epsilon(x - x^*)\big) < f(x^*)$. The vector $x^* + \epsilon(x - x^*)$ is feasible for all $\epsilon \in [0,1]$ because X is convex, so the optimality of x^* is contradicted.

(b) Using the convexity of f

$$f(x) \geq f(x^*) + \nabla f(x^*)'(x - x^*)$$

for every $x \in X$. If the condition $\nabla f(x^*)'(x-x^*) \geq 0$ holds for all $x \in X$, we obtain $f(x) \geq f(x^*)$, so x^* minimizes f over X. **Q.E.D.**

OPTIMIZATION SUBJECT TO BOUNDS

- Let $X = \{x \mid x \geq 0\}$. Then the necessary condition for $x^* = (x_1^*, \ldots, x_n^*)$ to be a local min is

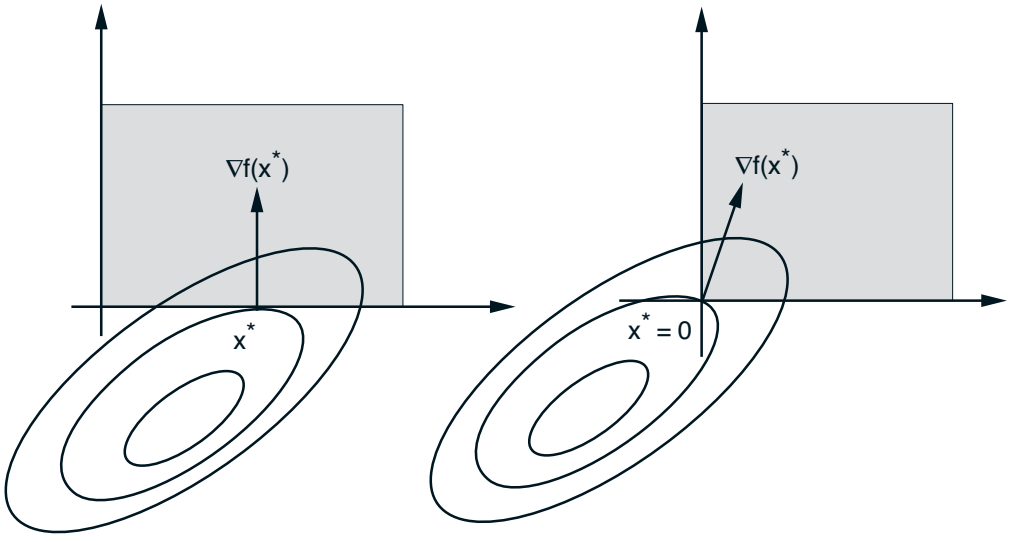
$$\sum_{i=1}^{n} \frac{\partial f(x^*)}{\partial x_i}(x_i - x_i^*) \geq 0, \qquad \forall\ x_i \geq 0,\ i = 1, \ldots, n.$$

- Fix i. Let $x_j = x_j^*$ for $j \neq i$ and $x_i = x_i^* + 1$:

$$\frac{\partial f(x^*)}{\partial x_i} \geq 0, \qquad \forall\ i.$$

- If $x_i^* > 0$, let also $x_j = x_j^*$ for $j \neq i$ and $x_i = \frac{1}{2}x_i^*$. Then $\partial f(x^*)/\partial x_i \leq 0$, so

$$\frac{\partial f(x^*)}{\partial x_i} = 0, \qquad \text{if } x_i^* > 0.$$



OPTIMIZATION OVER A SIMPLEX

$$X = \left\{ x \;\middle|\; x \geq 0, \sum_{i=1}^{n} x_i = r \right\}$$

where $r > 0$ is a given scalar.

- Necessary condition for $x^* = (x_1^*, \ldots, x_n^*)$ to be a local min:

$$\sum_{i=1}^{n} \frac{\partial f(x^*)}{\partial x_i}(x_i - x_i^*) \geq 0, \qquad \forall\, x_i \geq 0 \text{ with } \sum_{i=1}^{n} x_i = r.$$

- Fix i with $x_i^* > 0$ and let j be any other index. Use x with $x_i = 0$, $x_j = x_j^* + x_i^*$, and $x_m = x_m^*$ for all $m \neq i, j$:

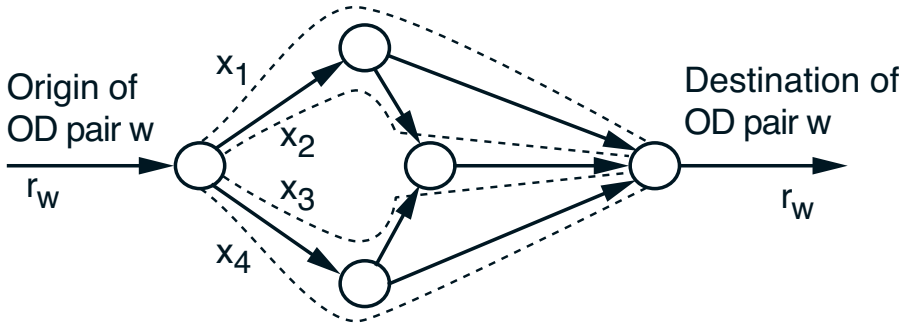
$$\left(\frac{\partial f(x^*)}{\partial x_j} - \frac{\partial f(x^*)}{\partial x_i} \right) x_i^* \geq 0,$$

$$x_i^* > 0 \quad \Longrightarrow \quad \frac{\partial f(x^*)}{\partial x_i} \leq \frac{\partial f(x^*)}{\partial x_j}, \qquad \forall\, j,$$

i.e., at the optimum, positive components have minimal (and equal) first cost derivative.

OPTIMAL ROUTING

- Given a data net, and a set W of OD pairs $w = (i, j)$. Each OD pair w has input traffic r_w.



- Optimal routing problem:

$$\text{minimize} \quad D(x) = \sum_{(i,j)} D_{ij} \left(\sum_{\substack{\text{all paths } p \\ \text{containing } (i,j)}} x_p \right)$$

$$\text{subject to} \quad \sum_{p \in P_w} x_p = r_w, \quad \forall\, w \in W,$$

$$x_p \geq 0, \quad \forall\, p \in P_w,\ w \in W$$

- Optimality condition

$$x_p^* > 0 \quad \Longrightarrow \quad \frac{\partial D(x^*)}{\partial x_p} \leq \frac{\partial D(x^*)}{\partial x_{p'}}, \qquad \forall\, p' \in P_w,$$

i.e., paths carrying > 0 flow are shortest with respect to first cost derivative.

TRAFFIC ASSIGNMENT

- Transportation network with OD pairs w. Each w has paths $p \in P_w$ and traffic r_w. Let x_p be the flow of path p and let $T_{ij}\left(\sum_{p:\text{ crossing }(i,j)} x_p\right)$ be the travel time of link (i,j).

- **User-optimization principle:** Traffic equilibrium is established when each user of the network chooses, among all available paths, a path of minimum travel time, i.e., for all $w \in W$ and paths $p \in P_w$,

$$x_p^* > 0 \quad \Longrightarrow \quad t_p(x^*) \leq t_{p'}(x^*), \qquad \forall\, p' \in P_w,\ \forall\, w \in W$$

where $t_p(x)$, is the travel time of path p

$$t_p(x) = \sum_{\substack{\text{all arcs } (i,j) \\ \text{on path } p}} T_{ij}(F_{ij}), \qquad \forall\ p \in P_w,\ \forall\ w \in W.$$

Identical with the optimality condition of the routing problem if we identify the arc travel time $T_{ij}(F_{ij})$ with the cost derivative $D'_{ij}(F_{ij})$.

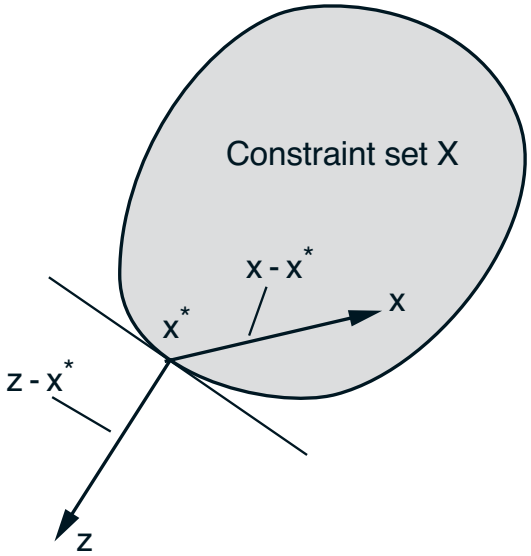
PROJECTION OVER A CONVEX SET

- Let $z \in \Re^n$ and a closed convex set X be given.

Problem:

$$\text{minimize} \quad f(x) = \|z - x\|^2$$
$$\text{subject to} \quad x \in X.$$

Proposition (Projection Theorem) Problem has a unique solution $[z]^+$ (the projection of z).



Necessary and sufficient condition for x^* to be the projection. The angle between $z - x^*$ and $x - x^*$ should be greater or equal to 90 degrees for all $x \in X$, or $(z - x^*)'(x - x^*) \leq 0$

- If X is a subspace, $z - x^* \perp X$.
- The mapping $f : \Re^n \mapsto X$ defined by $f(x) = [x]^+$ is continuous and nonexpansive, that is,

$$\|[x]^+ - [y]^+\| \leq \|x - y\|, \qquad \forall\, x, y \in \Re^n.$$

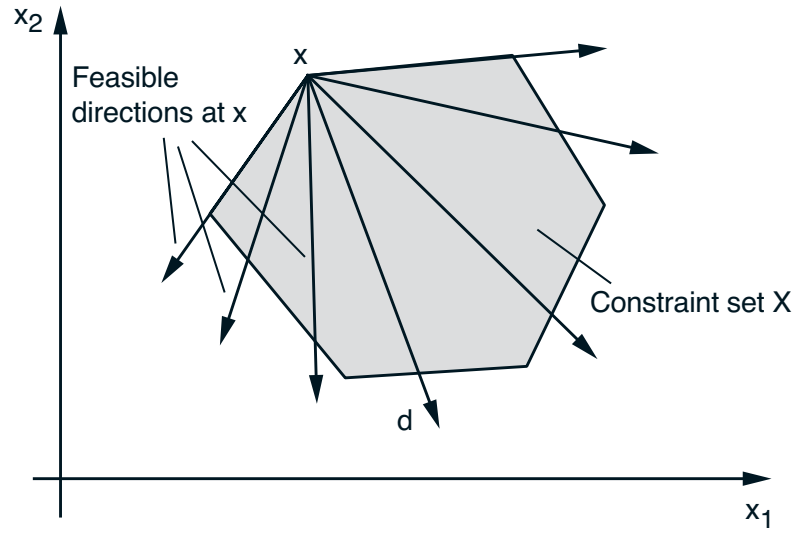
6.252 NONLINEAR PROGRAMMING

LECTURE 9: FEASIBLE DIRECTION METHODS

LECTURE OUTLINE

- Conditional Gradient Method
- Gradient Projection Methods

A $feasible\ direction$ at an $x \in X$ is a vector $d \neq 0$ such that $x + \alpha d$ is feasible for all suff. small $\alpha > 0$



- Note: the set of feasible directions at x is the set of all $\alpha(z - x)$ where $z \in X$, $z \neq x$, and $\alpha > 0$

FEASIBLE DIRECTION METHODS

- A *feasible direction method*:

$$x^{k+1} = x^k + \alpha^k d^k,$$

where d^k: feasible *descent* direction [$\nabla f(x^k)' d^k < 0$], and $\alpha^k > 0$ and such that $x^{k+1} \in X$.

- Alternative definition:

$$x^{k+1} = x^k + \alpha^k (\overline{x}^k - x^k),$$

where $\alpha^k \in (0, 1]$ and if x^k is nonstationary,

$$\overline{x}^k \in X, \qquad \nabla f(x^k)'(\overline{x}^k - x^k) < 0.$$

- Stepsize rules: Limited minimization, Constant $\alpha^k = 1$, Armijo: $\alpha^k = \beta^{m_k}$, where m_k is the first nonnegative m for which

$$f(x^k) - f\left(x^k + \beta^m (\overline{x}^k - x^k)\right) \geq -\sigma \beta^m \nabla f(x^k)'(\overline{x}^k - x^k)$$

CONVERGENCE ANALYSIS

- Similar to the one for (unconstrained) gradient methods.

- The direction sequence $\{d^k\}$ is *gradient related* to $\{x^k\}$ if the following property can be shown: *For any subsequence $\{x^k\}_{k\in\mathcal{K}}$ that converges to a nonstationary point, the corresponding subsequence $\{d^k\}_{k\in\mathcal{K}}$ is bounded and satisfies*

$$\limsup_{k\to\infty,\, k\in\mathcal{K}} \nabla f(x^k)'d^k < 0.$$

Proposition (Stationarity of Limit Points) Let $\{x^k\}$ be a sequence generated by the feasible direction method $x^{k+1} = x^k + \alpha^k d^k$. Assume that:

  - $\{d^k\}$ is gradient related
  - α^k is chosen by the limited minimization rule or the Armijo rule.

Then every limit point of $\{x^k\}$ is a stationary point.

- Proof: Nearly identical to the unconstrained case.

CONDITIONAL GRADIENT METHOD

- $x^{k+1} = x^k + \alpha^k(\overline{x}^k - x^k)$, where

$$\overline{x}^k = \arg\min_{x \in X} \nabla f(x^k)'(x - x^k).$$

- We assume that X is compact, so $\overline{x}^k$ is guaranteed to exist by Weierstrass.

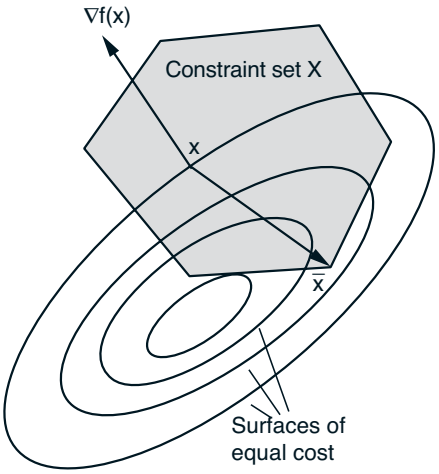
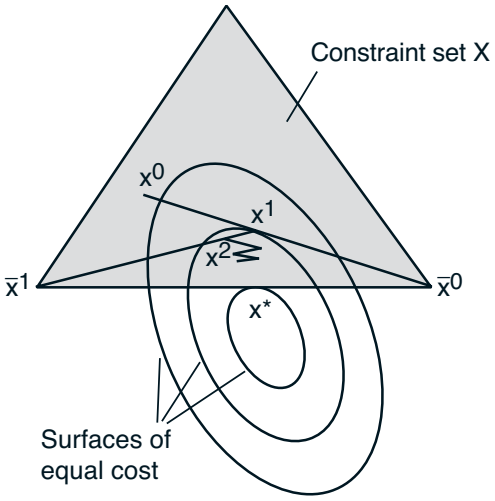


Illustration of the direction of the conditional gradient method.



Operation of the method. Slow (sublinear) convergence.

CONVERGENCE OF CONDITIONAL GRADIENT

- Show that the direction sequence of the conditional gradient method is gradient related, so the generic convergence result applies.
- Suppose that $\{x^k\}_{k\in K}$ converges to a nonstationary point $\tilde{x}$. We must prove that

$$\{x^k - x^k\}_{k\in K} : \text{bounded}, \quad \limsup_{k\to\infty,\, k\in K} \nabla f(x^k)'(\overline{x}^k - x^k) < 0$$

- 1st relation: Holds because $\overline{x}^k \in X$, $x^k \in X$, and X is assumed compact.
- 2nd relation: Note that by definition of $\overline{x}^k$,

$$\nabla f(x^k)'(\overline{x}^k - x^k) \le \nabla f(x^k)'(x - x^k), \qquad \forall\, x \in X$$

Taking limit as $k \to \infty,\ k \in K$, and min of the RHS over $x \in X$, and using the nonstationarity of $\tilde{x}$,

$$\limsup_{k\to\infty,\, k\in K} \nabla f(x^k)'(\overline{x}^k - x^k) \le \min_{x\in X} \nabla f(\tilde{x})'(x - \tilde{x}) < 0,$$

thereby proving the 2nd relation.

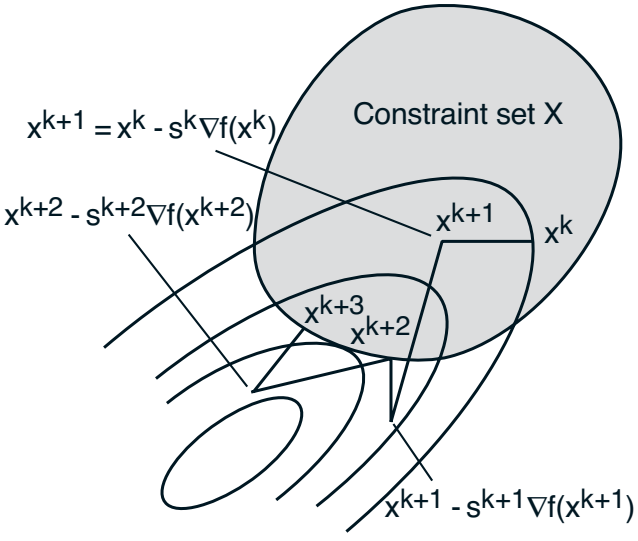
GRADIENT PROJECTION METHODS

• Gradient projection methods determine the feasible direction by using a quadratic cost subproblem. Simplest variant:

$$x^{k+1} = x^k + \alpha^k(\overline{x}^k - x^k)$$

$$\overline{x}^k = \left[x^k - s^k \nabla f(x^k)\right]^+$$

where, $[\cdot]^+$ denotes projection on the set X, $\alpha^k \in (0, 1]$ is a stepsize, and s^k is a positive scalar.



Gradient projection iterations for the case

$$\alpha^k \equiv 1, \quad x^{k+1} \equiv \overline{x}^k$$

If $\alpha^k < 1$, x^{k+1} is in the line segment connecting x^k and $\overline{x}^k$.

• Stepsize rules for α^k (assuming $s^k \equiv s$): Limited minimization, Armijo along the feasible direction, constant stepsize. Also, Armijo along the projection arc ($\alpha^k \equiv 1$, s^k: variable).

CONVERGENCE

- If α^k is chosen by the limited minimization rule or by the Armijo rule along the feasible direction, every limit point of $\{x^k\}$ is stationary.

- Proof: Show that the direction sequence $\{\overline{x}^k - x^k\}$ is gradient related. Assume $\{x^k\}_{k\in K}$ converges to a nonstationary $\tilde{x}$. Must prove

$$\{x^k - x^k\}_{k\in K} : \text{bounded}, \quad \limsup_{k\to\infty,\, k\in K} \nabla f(x^k)'(\overline{x}^k - x^k) < 0$$

1st relation holds because $\left\{\|\overline{x}^k - x^k\|\right\}_{k\in K}$ converges to $\|[\tilde{x} - s\nabla f(\tilde{x})]^+ - \tilde{x}\|$. By optimality condition for projections, $\left(x^k - s\nabla f(x^k) - \overline{x}^k\right)'(x - \overline{x}^k) \leq 0$ for all $x \in X$. Applying this relation with $x = x^k$, and taking limit,

$$\limsup_{k\to\infty,\, k\in K} \nabla f(x^k)'(\overline{x}^k - x^k) \leq -\frac{1}{s}\left\|\tilde{x} - \left[\tilde{x} - s\nabla f(\tilde{x})\right]^+\right\|^2 < 0$$

- Similar conclusion for constant stepsize $\alpha^k = 1$, $s^k = s$ (under a Lipschitz condition on ∇f).

- Similar conclusion for Armijo rule along the projection arc.

CONVERGENCE RATE – VARIANTS

- Assume $f(x) = \frac{1}{2}x'Qx - b'x$, with $Q > 0$, and a constant stepsize ($a^k \equiv 1$, $s^k \equiv s$). Using the nonexpansiveness of projection

$$\begin{aligned}\left\|x^{k+1} - x^*\right\| &= \left\| \left[x^k - s\nabla f(x^k)\right]^+ - \left[x^* - s\nabla f(x^*)\right]^+ \right\| \\ &\leq \left\| \left(x^k - s\nabla f(x^k)\right) - \left(x^* - s\nabla f(x^*)\right) \right\| \\ &= \left\| (I - sQ)(x^k - x^*) \right\| \\ &\leq \max\left\{ |1 - sm|, |1 - sM| \right\} \left\| x^k - x^* \right\|\end{aligned}$$

where m, M: min and max eigenvalues of Q.

- Scaled version: $x^{k+1} = x^k + \alpha^k(\overline{x}^k - x^k)$, where

$$\overline{x}^k = \arg\min_{x \in X} \left\{ \nabla f(x^k)'(x - x^k) + \frac{1}{2s^k}(x - x^k)'H^k(x - x^k) \right\},$$

and $H^k > 0$ (involves transformation $y^k = (H^k)^{1/2}x^k$. Since the minimum value above is negative when x^k is nonstationary, $\nabla f(x^k)'(\overline{x}^k - x^k) < 0$.

- Newton's method for $H^k = \nabla^2 f(x^k)$.

- Variants: Projecting on an expanded constraint set, projecting on a restricted constraint set, combinations with unconstrained methods, etc.

6.252 NONLINEAR PROGRAMMING

LECTURE 10

ALTERNATIVES TO GRADIENT PROJECTION

LECTURE OUTLINE

- Three Alternatives/Remedies for Gradient Projection
  - Two-Metric Projection Methods
  - Manifold Suboptimization Methods
  - Affine Scaling Methods

Scaled GP method with scaling matrix $H^k > 0$:

$$x^{k+1} = x^k + \alpha^k(\overline{x}^k - x^k),$$

$$\overline{x}^k = \arg\min_{x \in X} \left\{ \nabla f(x^k)'(x - x^k) + \frac{1}{2s^k}(x - x^k)' H^k (x - x^k) \right\}.$$

- The QP direction subproblem is complicated by:
  - Difficult inequality (e.g., nonorthant) constraints
  - Nondiagonal H^k, needed for Newton scaling

THREE WAYS TO DEAL W/ THE DIFFICULTY

- Two-metric projection methods:

$$x^{k+1} = \left[x^k - \alpha^k D^k \nabla f(x^k)\right]^+$$

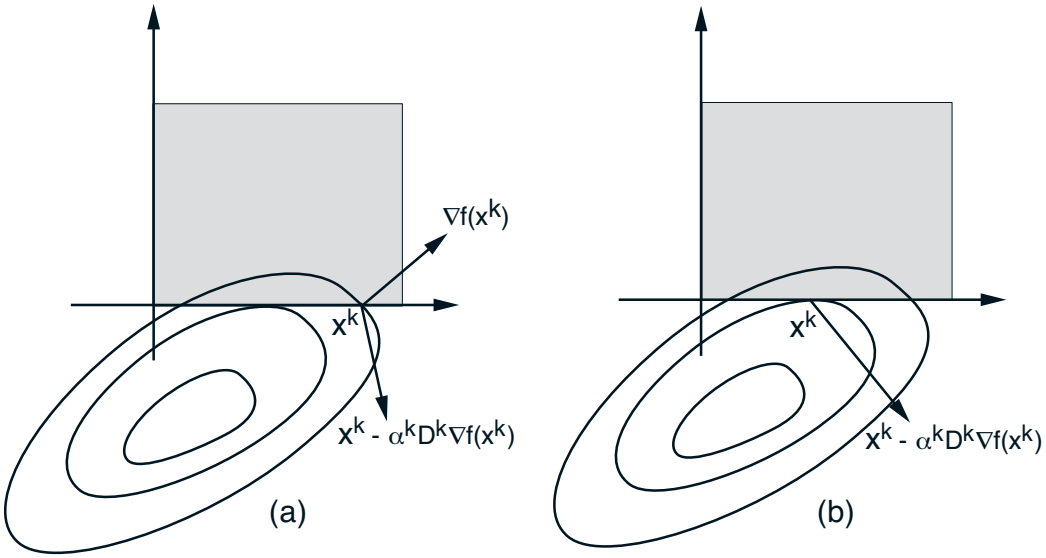
  - Use Newton-like scaling but use a standard projection
  - Suitable for bounds, simplexes, Cartesian products of simple sets, etc
- Manifold suboptimization methods:
  - Use (scaled) gradient projection on the manifold of active inequality constraints
  - Each QP subproblem is equality-constrained
  - Need strategies to cope with changing active manifold (add-drop constraints)
- Affine Scaling Methods
  - Go through the interior of the feasible set
  - Each QP subproblem is equality-constrained, AND we don't have to deal with changing active manifold

TWO-METRIC PROJECTION METHODS

- In their simplest form, apply to constraint: $x \geq 0$, but generalize to bound and other constraints
- Like unconstr. gradient methods except for $[\cdot]^+$

$$x^{k+1} = \left[x^k - \alpha^k D^k \nabla f(x^k)\right]^+, \qquad D^k > 0$$

- Major difficulty: Descent is not guaranteed for D^k: arbitrary



- Remedy: Use D^k that is diagonal w/ respect to indices that "are active and want to stay active"

$$I^+(x^k) = \left\{ i \;\middle|\; x_i^k = 0,\ \partial f(x^k)/\partial x_i > 0 \right\}$$

PROPERTIES OF 2-METRIC PROJECTION

- Suppose D^k is diagonal with respect to $I^+(x^k)$, i.e., $d^k_{ij} = 0$ for $i, j \in I^+(x^k)$ with $i \neq j$, and let

$$x^k(a) = \left[x^k - \alpha D^k \nabla f(x^k)\right]^+$$

  - If x^k is stationary, $x^k = x^k(\alpha)$ for all $\alpha > 0$.
  - Otherwise $f\big(x(\alpha)\big) < f(x^k)$ for all sufficiently small $\alpha > 0$ (can use Armijo rule).

- Because $I^+(x)$ is discontinuous w/ respect to x, to guarantee convergence we need to include in $I^+(x)$ constraints that are "ϵ-active" [those w/ $x^k_i \in [0, \epsilon]$ and $\partial f(x^k)/\partial x_i > 0$].

- The constraints in $I^+(x^*)$ eventually become active and don't matter.

- Method reduces to unconstrained Newton-like method on the manifold of active constraints at x^*.

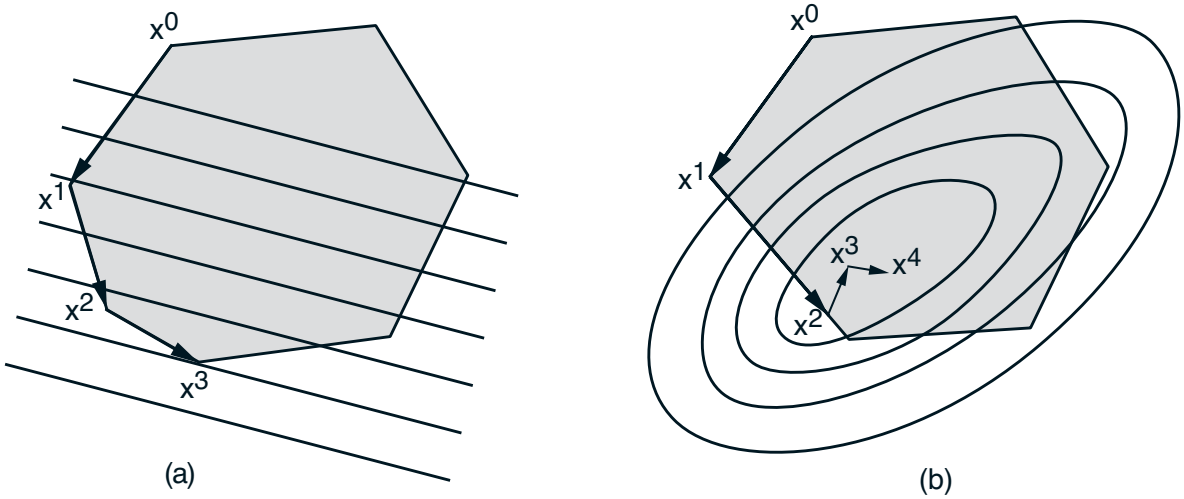
- Thus, superlinear convergence is possible w/ simple projections.

MANIFOLD SUBOPTIMIZATION METHODS

- Feasible direction methods for

$$\min f(x) \qquad \text{subject to } a_j' x \leq b_j, \quad j = 1, \ldots, r$$

- Gradient is projected on a linear manifold of active constraints rather than on the entire constraint set (linearly constrained QP).



- Searches through sequence of manifolds, each differing by at most one constraint from the next.

- Potentially many iterations to identify the active manifold; then method reduces to (scaled) steepest descent on the active manifold.

- Well-suited for a small number of constraints, and for quadratic programming.

OPERATION OF MANIFOLD METHODS

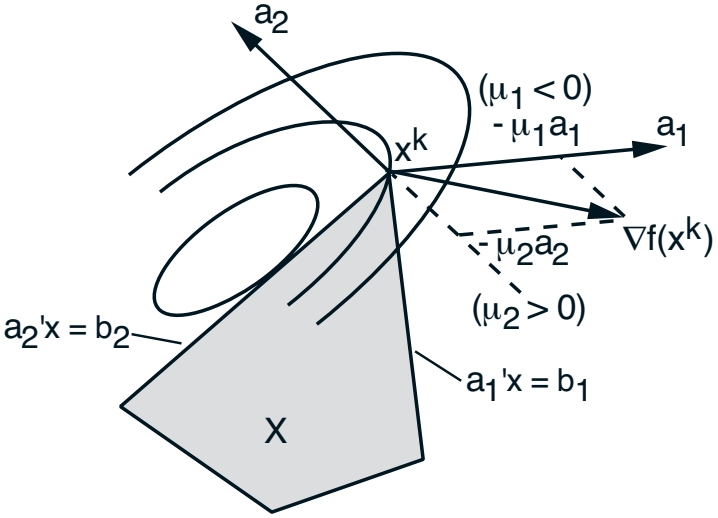
- Let $A(x) = \{j \mid a_j' x = b_j\}$ be the active index set at x. Given x^k, we find

$$d^k = \arg \min_{a_j' d = 0,\, j \in A(x^k)} \nabla f(x^k)' d + \tfrac{1}{2} d' H^k d$$

- If $d^k \neq 0$, then d^k is a feasible descent direction. Perform feasible descent on the current manifold.

- If $d^k = 0$, either (1) x^k is stationary or (2) we enlarge the current manifold (drop an active constraint). For this, use the scalars μ_j such that

$$\nabla f(x^k) + \sum_{j \in A(x^k)} \mu_j a_j = 0$$



If $\mu_j \geq 0$ for all j, x^k is stationary, since for all feasible x, $\nabla f(x^k)'(x - x^k)$ is equal to

$$- \sum_{j \in A(x^k)} \mu_j a_j'(x - x^k) \geq 0$$

Else, drop a constraint j with $\mu_j < 0$.

AFFINE SCALING METHODS FOR LP

- Focus on the LP $\min_{Ax=b,\, x\geq 0} c'x$, and the scaled gradient projection $x^{k+1} = x^k + \alpha^k(\overline{x}^k - x^k)$, with

$$\overline{x}^k = \arg \min_{Ax=b,\, x\geq 0} c'(x - x^k) + \frac{1}{2s^k}(x - x^k)' H^k (x - x^k)$$

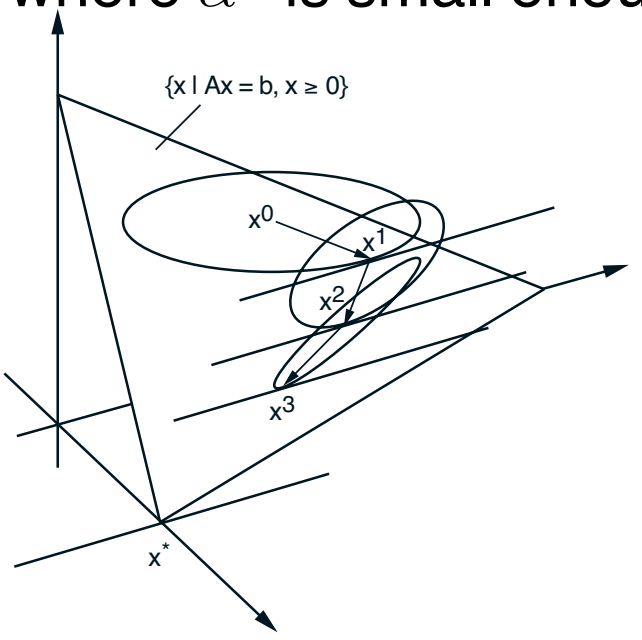
- If $x^k > 0$ then $\overline{x}^k > 0$ for s^k small enough, so $\overline{x}^k = x^k - s^k(H^k)^{-1}(c - A'\lambda^k)$ with

$$\lambda^k = \left(A(H^k)^{-1}A'\right)^{-1} A(H^k)^{-1}c$$

Lumping s^k into α^k:

$$x^{k+1} = x^k - \alpha^k(H^k)^{-1}(c - A'\lambda^k),$$

where α^k is small enough to ensure that $x^{k+1} > 0$



Importance of using time-varying H^k (should bend $\overline{x}^k - x^k$ away from the boundary)

AFFINE SCALING

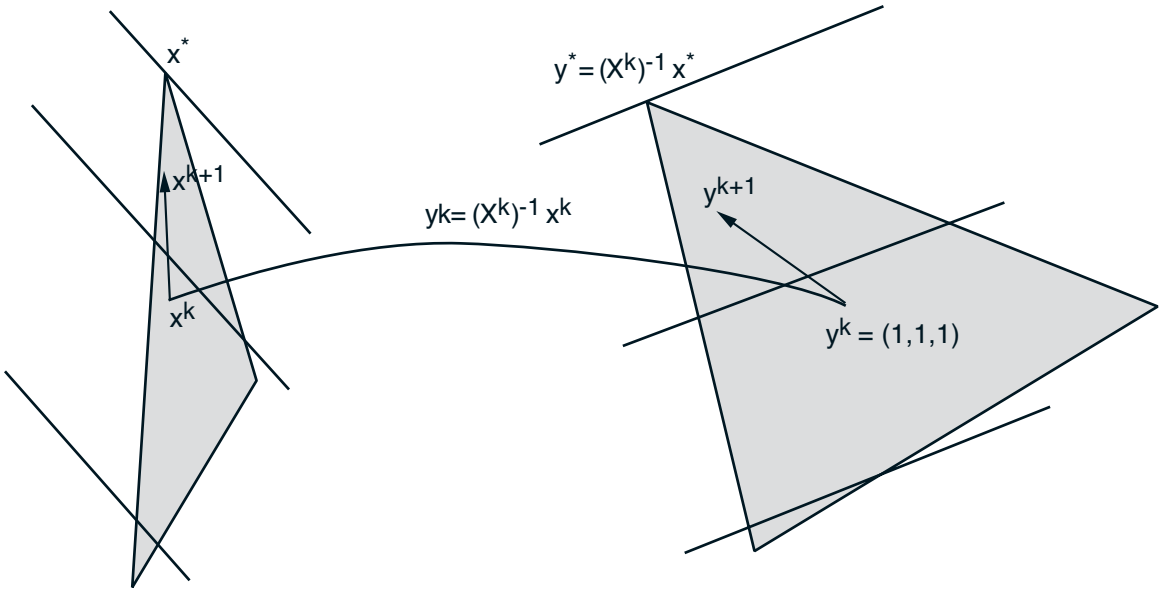
- Particularly interesting choice (affine scaling)

$$H^k = (X^k)^{-2},$$

where X^k is the diagonal matrix having the (positive) coordinates x_i^k along the diagonal:

$$x^{k+1} = x^k - \alpha^k (X^k)^2 (c - A'\lambda^k), \quad \lambda^k = \left(A(X^k)^2 A'\right)^{-1} A(X^k)^2 c$$

- Corresponds to unscaled gradient projection iteration in the variables $y = (X^k)^{-1}x$. The vector x^k is mapped onto the unit vector $y^k = (1, \ldots, 1)$.



- Extensions, convergence, practical issues.

6.252 NONLINEAR PROGRAMMING

LECTURE 11

CONSTRAINED OPTIMIZATION;

LAGRANGE MULTIPLIERS

LECTURE OUTLINE

- Equality Constrained Problems
- Basic Lagrange Multiplier Theorem
- Proof 1: Elimination Approach
- Proof 2: Penalty Approach

Equality constrained problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad h_i(x) = 0, \qquad i = 1, \ldots, m. \end{aligned}$$

where $f : \Re^n \mapsto \Re$, $h_i : \Re^n \mapsto \Re$, $i = 1, \ldots, m$, are continuously differentiable functions. (Theory also applies to case where f and h_i are cont. differentiable in a neighborhood of a local minimum.)

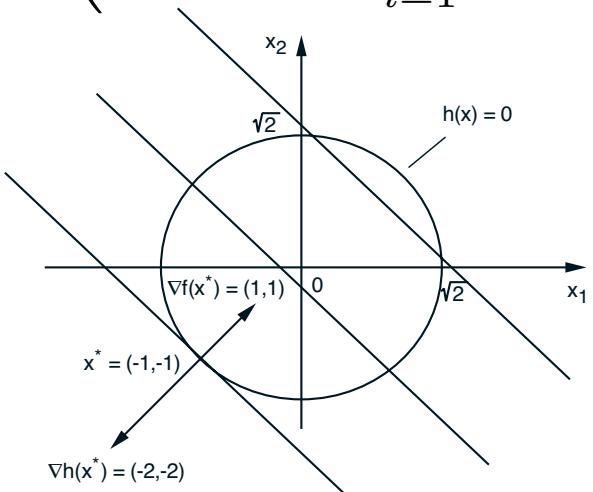
LAGRANGE MULTIPLIER THEOREM

- Let x^* be a local min and a regular point [$\nabla h_i(x^*)$: linearly independent]. Then there exist unique scalars $\lambda_1^*, \ldots, \lambda_m^*$ such that

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) = 0.$$

If in addition f and h are twice cont. differentiable,

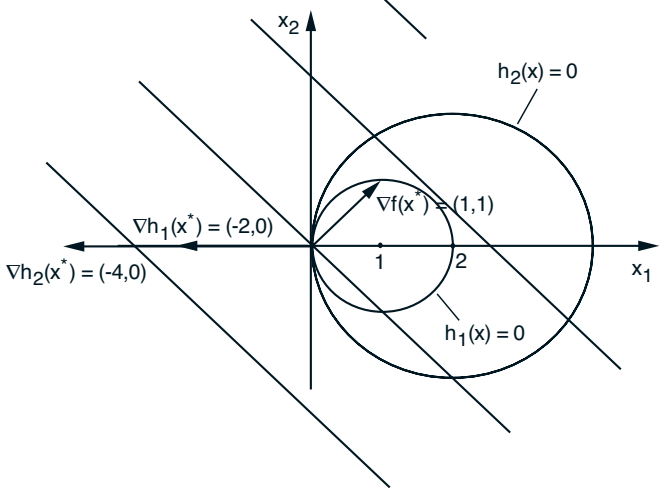
$$y' \left(\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x^*) \right) y \geq 0, \ \forall \, y \text{ s.t. } \nabla h(x^*)' y = 0$$



$$\text{minimize} \quad x_1 + x_2$$

$$\text{subject to} \quad x_1^2 + x_2^2 = 2.$$

The Lagrange multiplier is $\lambda = 1/2$.



$$\text{minimize} \quad x_1 + x_2$$

$$\text{s. t.} \ (x_1 - 1)^2 + x_2^2 - 1 = 0$$

$$(x_1 - 2)^2 + x_2^2 - 4 = 0$$

PROOF VIA ELIMINATION APPROACH

- Consider the linear constraints case

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

$$\text{subject to} \quad Ax = b$$

where A is an $m \times n$ matrix with linearly independent rows and $b \in \Re^m$ is a given vector.

- Partition $A = (\, B \quad R \,)$, where B is $m \times m$ invertible, and $x = (\, x_B \quad x_R \,)'$. Equivalent problem:

$$\text{minimize} \quad F(x_R) \equiv f\Big(B^{-1}(b - Rx_R), x_R\Big)$$

$$\text{subject to} \quad x_R \in \Re^{n-m}.$$

- Unconstrained optimality condition:

$$0 = \nabla F(x_R^*) = -R'(B')^{-1}\nabla_B f(x^*) + \nabla_R f(x^*) \quad (1)$$

By defining

$$\lambda^* = -(B')^{-1}\nabla_B f(x^*),$$

we have $\nabla_B f(x^*) + B'\lambda^* = 0$, while Eq. (1) is written $\nabla_R f(x^*) + R'\lambda^* = 0$. Combining:

$$\nabla f(x^*) + A'\lambda^* = 0$$

ELIMINATION APPROACH - CONTINUED

- Second order condition: For all $d \in \Re^{n-m}$

$$0 \leq d' \nabla^2 F(x_R^*) d = d' \nabla^2 \Big(f\Big(B^{-1}(b - Rx_R), x_R \Big) \Big) d. \quad (2)$$

- After calculation we obtain

$$\nabla^2 F(x_R^*) = \; R'(B')^{-1} \nabla^2_{BB} f(x^*) B^{-1} R$$
$$- R'(B')^{-1} \nabla^2_{BR} f(x^*) - \nabla^2_{RB} f(x^*) B^{-1} R + \nabla^2_{RR} f(x^*).$$

- Eq. (2) and the linearity of the constraints [implying that $\nabla^2 h_i(x^*) = 0$], yields for all $d \in \Re^{n-m}$

$$0 \leq d' \nabla^2 F(x_R^*) d = y' \nabla^2 f(x^*) y$$
$$= y' \left(\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x^*) \right) y,$$

where $y = (\, y_B \quad y_R \,)' = (\, -B^{-1} R d \quad d \,)'$.

- y has this form iff

$$0 = B y_B + R y_R = \nabla h(x^*)' y.$$

PROOF VIA PENALTY APPROACH

- Introduce, for $k = 1, 2, \ldots$, the cost function

$$F^k(x) = f(x) + \frac{k}{2}||h(x)||^2 + \frac{\alpha}{2}||x - x^*||^2,$$

where $\alpha > 0$ and x^* is a local minimum.

- Let $\epsilon > 0$ be such that $f(x^*) \leq f(x)$ for all feasible x in the *closed* sphere $S = \Big\{x \mid ||x - x^*|| \leq \epsilon\Big\}$, and let $x^k = \arg\min_{x \in S} F^k(x)$. Have

$$F^k(x^k) = f(x^k) + \frac{k}{2}||h(x^k)||^2 + \frac{\alpha}{2}||x^k - x^*||^2 \leq F^k(x^*) = f(x^*)$$

Hence, $\lim_{k \to \infty} ||h(x^k)|| = 0$, so for every limit point $\overline{x}$ of $\{x^k\}$, $h(\overline{x}) = 0$.

- Furthermore, $f(x^k) + (\alpha/2)||x^k - x^*||^2 \leq f(x^*)$ for all k, so by taking lim,

$$f(\overline{x}) + \frac{\alpha}{2}||\overline{x} - x^*||^2 \leq f(x^*).$$

Combine with $f(x^*) \leq f(\overline{x})$ [since $\overline{x} \in S$ and $h(\overline{x}) = 0$] to obtain $||\overline{x} - x^*|| = 0$ so that $\overline{x} = x^*$. Thus $\{x^k\} \to x^*$.

PENALTY APPROACH - CONTINUED

- Since $x^k \to x^*$, for large k, x^k is interior to S, and is an $unconstrained$ local minimum of $F^k(x)$.
- From 1st order necessary condition,

$$0 = \nabla F^k(x^k) = \nabla f(x^k) + k \nabla h(x^k) h(x^k) + \alpha(x^k - x^*). \quad (3)$$

Since $\nabla h(x^*)$ has rank m, $\nabla h(x^k)$ also has rank m for large k, so $\nabla h(x^k)' \nabla h(x^k)$: invertible. Thus, multiplying Eq. (3) w/ $\nabla h(x^k)'$

$$kh(x^k) = -\Big(\nabla h(x^k)' \nabla h(x^k)\Big)^{-1} \nabla h(x^k)' \Big(\nabla f(x^k) + \alpha(x^k - x^*)\Big).$$

Taking limit as $k \to \infty$ and $x^k \to x^*$,

$$\Big\{kh(x^k)\Big\} \to -\Big(\nabla h(x^*)' \nabla h(x^*)\Big)^{-1} \nabla h(x^*)' \nabla f(x^*) \equiv \lambda^*.$$

Taking limit as $k \to \infty$ in Eq. (3), we obtain

$$\nabla f(x^*) + \nabla h(x^*) \lambda^* = 0.$$

- 2nd order L-multiplier condition: Use 2nd order unconstrained condition for x^k, and algebra.

LAGRANGIAN FUNCTION

- Define the Lagrangian function

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x).$$

Then, if x^* is a local minimum which is regular, the Lagrange multiplier conditions are written

$$\nabla_x L(x^*, \lambda^*) = 0, \qquad \nabla_\lambda L(x^*, \lambda^*) = 0,$$

System of $n + m$ equations with $n + m$ unknowns.

$$y' \nabla^2_{xx} L(x^*, \lambda^*) y \geq 0, \qquad \forall\ y \text{ s.t. } \nabla h(x^*)' y = 0.$$

- Example

minimize $\frac{1}{2}\left(x_1^2 + x_2^2 + x_3^2\right)$

subject to $x_1 + x_2 + x_3 = 3.$

Necessary conditions

$$x_1^* + \lambda^* = 0, \quad x_2^* + \lambda^* = 0,$$

$$x_3^* + \lambda^* = 0, \quad x_1^* + x_2^* + x_3^* = 3.$$

EXAMPLE - PORTFOLIO SELECTION

- Investment of 1 unit of wealth among n assets with random rates of return e_i, and given means $\overline{e}_i$, and covariance matrix $Q = \Big[E\{(e_i - \overline{e}_i)(e_j - \overline{e}_j)\}\Big]$.
- If x_i: amount invested in asset i, we want to

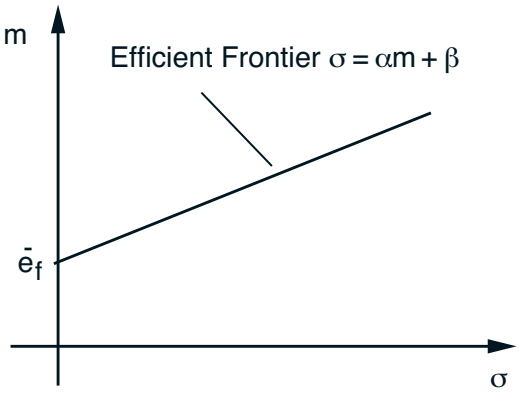
$$\text{minimize} \quad x'Qx \left(= \text{Variance of return } \sum_i e_i x_i \right)$$

$$\text{subject to } \sum_i x_i = 1, \text{ and a given mean } \sum_i \overline{e}_i x_i = m$$

- Let λ_1 and λ_2 be the L-multipliers. Have $2Qx^* + \lambda_1 u + \lambda_2 \overline{e} = 0$, where $u = (1, \ldots, 1)'$ and $\overline{e} = (\overline{e}_1, \ldots, \overline{e}_n)'$. This yields

$$x^* = mv + w, \quad \text{Variance of return} = \sigma^2 = (\alpha m + \beta)^2 + \gamma,$$

where v and w are vectors, and α, β, and γ are some scalars that depend on Q and $\overline{e}$.



For given m the optimal σ lies on a line (called "efficient frontier").

6.252 NONLINEAR PROGRAMMING

LECTURE 12: SUFFICIENCY CONDITIONS

LECTURE OUTLINE

- Equality Constrained Problems/Sufficiency Conditions
- Convexification Using Augmented Lagrangians
- Proof of the Sufficiency Conditions
- Sensitivity

Equality constrained problem

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && h_i(x) = 0, \qquad i = 1, \ldots, m. \end{aligned}$$

where $f : \Re^n \mapsto \Re$, $h_i : \Re^n \mapsto \Re$, are continuously differentiable. To obtain sufficiency conditions, assume that f and h_i are $twice$ continuously differentiable.

SUFFICIENCY CONDITIONS

Second Order Sufficiency Conditions: Let $x^* \in \Re^n$ and $\lambda^* \in \Re^m$ satisfy

$$\nabla_x L(x^*, \lambda^*) = 0, \qquad \nabla_\lambda L(x^*, \lambda^*) = 0,$$

$$y' \nabla^2_{xx} L(x^*, \lambda^*) y > 0, \quad \forall\ y \neq 0 \text{ with } \nabla h(x^*)' y = 0.$$

Then x^* is a strict local minimum.

Example: Minimize $-(x_1 x_2 + x_2 x_3 + x_1 x_3)$ subject to $x_1 + x_2 + x_3 = 3$. We have that $x_1^* = x_2^* = x_3^* = 1$ and $\lambda^* = 2$ satisfy the 1st order conditions. Also

$$\nabla^2_{xx} L(x^*, \lambda^*) = \begin{pmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{pmatrix}.$$

We have for all $y \neq 0$ with $\nabla h(x^*)' y = 0$ or $y_1 + y_2 + y_3 = 0$,

$$\begin{aligned} y' \nabla^2_{xx} L(x^*, \lambda^*) y &= -y_1(y_2 + y_3) - y_2(y_1 + y_3) - y_3(y_1 + y_2) \\ &= y_1^2 + y_2^2 + y_3^2 > 0. \end{aligned}$$

Hence, x^* is a strict local minimum.

A BASIC LEMMA

Lemma: Let P and Q be two symmetric matrices. Assume that $Q \geq 0$ and $P > 0$ on the nullspace of Q, i.e., $x'Px > 0$ for all $x \neq 0$ with $x'Qx = 0$. Then there exists a scalar $\bar{c}$ such that

$$P + cQ : \text{positive definite}, \qquad \forall\ c > \bar{c}.$$

Proof: Assume the contrary. Then for every k, there exists a vector x^k with $\|x^k\| = 1$ such that

$$x^{k\prime} P x^k + k x^{k\prime} Q x^k \leq 0.$$

Consider a subsequence $\{x^k\}_{k \in K}$ converging to some $\bar{x}$ with $\|\bar{x}\| = 1$. Taking the limit superior,

$$\bar{x}' P \bar{x} + \limsup_{k \to \infty,\ k \in K} (k x^{k\prime} Q x^k) \leq 0. \qquad (*)$$

We have $x^{k\prime} Q x^k \geq 0$ (since $Q \geq 0$), so $\{x^{k\prime} Q x^k\}_{k \in K} \to 0$. Therefore, $\bar{x}' Q \bar{x} = 0$ and using the hypothesis, $\bar{x}' P \bar{x} > 0$. This contradicts (*).

PROOF OF SUFFICIENCY CONDITIONS

Consider the $augmented\ Lagrangian$ function

$$L_c(x,\lambda) = f(x) + \lambda' h(x) + \frac{c}{2}\|h(x)\|^2,$$

where c is a scalar. We have

$$\nabla_x L_c(x,\lambda) = \nabla_x L(x,\tilde{\lambda}),$$

$$\nabla^2_{xx} L_c(x,\lambda) = \nabla^2_{xx} L(x,\tilde{\lambda}) + c\nabla h(x)\nabla h(x)'$$

where $\tilde{\lambda} = \lambda + ch(x)$. If (x^*,λ^*) satisfy the suff. conditions, we have using the lemma,

$$\nabla_x L_c(x^*,\lambda^*) = 0, \qquad \nabla^2_{xx} L_c(x^*,\lambda^*) > 0,$$

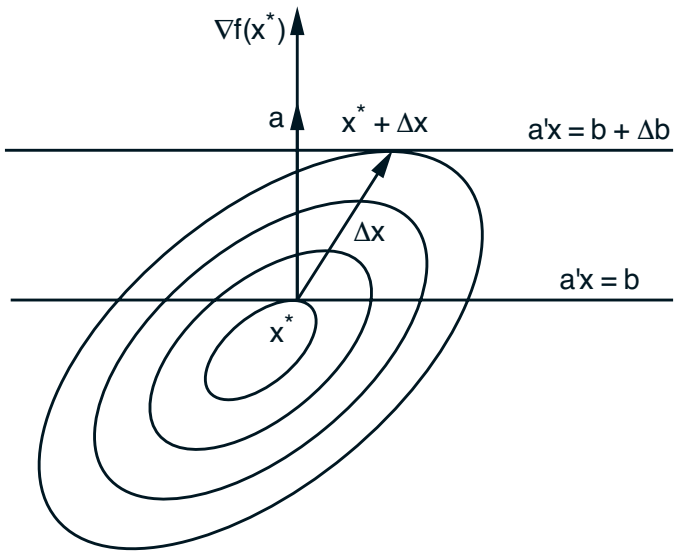
for suff. large c. Hence for some $\gamma > 0$, $\epsilon > 0$,

$$L_c(x,\lambda^*) \geq L_c(x^*,\lambda^*) + \frac{\gamma}{2}\|x-x^*\|^2, \quad \text{if } \|x-x^*\| < \epsilon.$$

Since $L_c(x,\lambda^*) = f(x)$ when $h(x) = 0$,

$$f(x) \geq f(x^*) + \frac{\gamma}{2}\|x-x^*\|^2, \quad \text{if } h(x) = 0,\ \|x-x^*\| < \epsilon.$$

SENSITIVITY - GRAPHICAL DERIVATION



Sensitivity theorem for the problem $\min_{a'x=b} f(x)$. If b is changed to $b+\Delta b$, the minimum x^* will change to $x^*+\Delta x$. Since $b+\Delta b = a'(x^*+\Delta x) = a'x^* + a'\Delta x = b + a'\Delta x$, we have $a'\Delta x = \Delta b$. Using the condition $\nabla f(x^*) = -\lambda^* a$,

$$\begin{aligned}\Delta\text{cost} &= f(x^*+\Delta x) - f(x^*) = \nabla f(x^*)'\Delta x + o(\|\Delta x\|) \\ &= -\lambda^* a'\Delta x + o(\|\Delta x\|)\end{aligned}$$

Thus $\Delta\text{cost} = -\lambda^*\Delta b + o(\|\Delta x\|)$, so up to first order

$$\lambda^* = -\frac{\Delta\text{cost}}{\Delta b}.$$

For multiple constraints $a_i'x = b_i$, $i = 1, \ldots, n$, we have

$$\Delta\text{cost} = -\sum_{i=1}^{m} \lambda_i^* \Delta b_i + o(\|\Delta x\|).$$

SENSITIVITY THEOREM

Sensitivity Theorem: Consider the family of problems

$$\min_{h(x)=u} f(x) \qquad (^*)$$

parameterized by $u \in \Re^m$. Assume that for $u = 0$, this problem has a local minimum x^*, which is regular and together with its unique Lagrange multiplier λ^* satisfies the sufficiency conditions.

Then there exists an open sphere S centered at $u = 0$ such that for every $u \in S$, there is an $x(u)$ and a $\lambda(u)$, which are a local minimum-Lagrange multiplier pair of problem (*). Furthermore, $x(\cdot)$ and $\lambda(\cdot)$ are continuously differentiable within S and we have $x(0) = x^*$, $\lambda(0) = \lambda^*$. In addition,

$$\nabla p(u) = -\lambda(u), \qquad \forall\ u \in S$$

where $p(u)$ is the *primal function*

$$p(u) = f\Big(x(u)\Big).$$

EXAMPLE

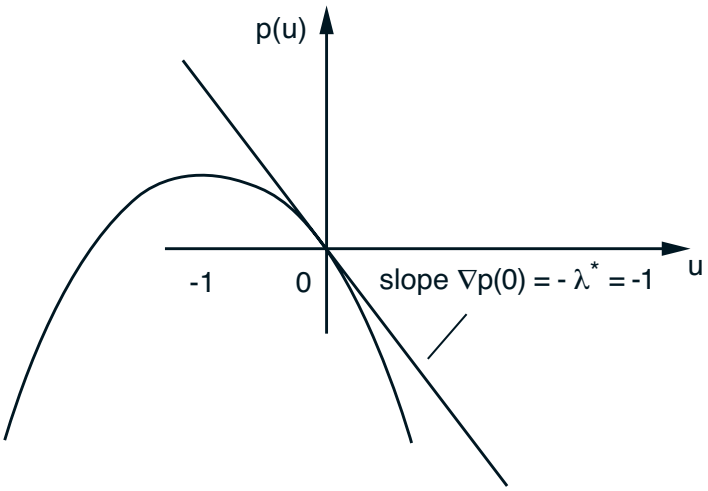


Illustration of the primal function $p(u) = f\Big(x(u)\Big)$ for the two-dimensional problem

$$\text{minimize} \quad f(x) = \tfrac{1}{2}\left(x_1^2 - x_2^2\right) - x_2$$

$$\text{subject to} \quad h(x) = x_2 = 0.$$

Here,

$$p(u) = \min_{h(x)=u} f(x) = -\tfrac{1}{2}u^2 - u$$

and $\lambda^* = -\nabla p(0) = 1$, consistently with the sensitivity theorem.

- Need for regularity of x^*: Change constraint to $h(x) = x_2^2 = 0$. Then $p(u) = -u/2 - \sqrt{u}$ for $u \geq 0$ and is undefined for $u < 0$.

PROOF OUTLINE OF SENSITIVITY THEOREM

Apply implicit function theorem to the system

$$\nabla f(x) + \nabla h(x)\lambda = 0, \qquad h(x) = u.$$

For $u = 0$ the system has the solution (x^*, λ^*), and the corresponding $(n+m) \times (n+m)$ Jacobian

$$J = \begin{pmatrix} \nabla^2 f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla^2 h_i(x^*) & \nabla h(x^*) \\ \nabla h(x^*)' & 0 \end{pmatrix}$$

is shown nonsingular using the sufficiency conditions. Hence, for all u in some open sphere S centered at $u = 0$, there exist $x(u)$ and $\lambda(u)$ such that $x(0) = x^*$, $\lambda(0) = \lambda^*$, the functions $x(\cdot)$ and $\lambda(\cdot)$ are continuously differentiable, and

$$\nabla f\Big(x(u)\Big) + \nabla h\Big(x(u)\Big)\lambda(u) = 0, \quad h\Big(x(u)\Big) = u.$$

For u close to $u = 0$, using the sufficiency conditions, $x(u)$ and $\lambda(u)$ are a local minimum-Lagrange multiplier pair for the problem $\min_{h(x)=u} f(x)$.

To derive $\nabla p(u)$, differentiate $h\Big(x(u)\Big) = u$, to obtain $I = \nabla x(u) \nabla h\Big(x(u)\Big)$, and combine with the relations $\nabla x(u) \nabla f\Big(x(u)\Big) + \nabla x(u) \nabla h\Big(x(u)\Big)\lambda(u) = 0$ and $\nabla p(u) = \nabla_u \Big\{ f\Big(x(u)\Big) \Big\} = \nabla x(u) \nabla f\Big(x(u)\Big)$.

6.252 NONLINEAR PROGRAMMING

LECTURE 13: INEQUALITY CONSTRAINTS

LECTURE OUTLINE

- Inequality Constrained Problems
- Necessary Conditions
- Sufficiency Conditions
- Linear Constraints

Inequality constrained problem

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

$$\text{subject to } h(x) = 0, \qquad g(x) \leq 0$$

where $f : \Re^n \mapsto \Re$, $h : \Re^n \mapsto \Re^m$, $g : \Re^n \mapsto \Re^r$ are continuously differentiable. Here

$$h = (h_1, ..., h_m), \qquad g = (g_1, ..., g_r).$$

TREATING INEQUALITIES AS EQUATIONS

- Consider the **set of active inequality constraints**

$$A(x) = \Big\{ j \mid g_j(x) = 0 \Big\}.$$

- If x^* is a local minimum:
  - The active inequality constraints at x^* can be treated as equations
  - The inactive constraints at x^* don't matter
- Assuming regularity of x^* and assigning zero Lagrange multipliers to inactive constraints,

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^{r} \mu_j^* \nabla g_j(x^*) = 0,$$

$$\mu_j^* = 0, \qquad \forall\ j \notin A(x^*).$$

- Extra property: $\mu_j^* \geq 0$ for all j.
- Intuitive reason: Relax jth constraint, $g_j(x) \leq u_j$. Since $\Delta\text{cost} \leq 0$ if $u_j > 0$, by the sensitivity theorem, we have

$$\mu_j^* = -(\Delta\text{cost due to } u_j)/u_j \geq 0$$

BASIC RESULTS

Kuhn-Tucker Necessary Conditions: Let x^* be a local minimum and a regular point. Then there exist unique Lagrange mult. vectors $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*)$, $\mu^* = (\mu_1^*, \ldots, \mu_r^*)$, such that

$$\nabla_x L(x^*, \lambda^*, \mu^*) = 0,$$

$$\mu_j^* \geq 0, \qquad j = 1, \ldots, r,$$

$$\mu_j^* = 0, \qquad \forall\ j \notin A(x^*).$$

If f, h, and g are twice cont. differentiable,

$$y' \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) y \geq 0, \qquad \text{for all } y \in V(x^*),$$

where

$$V(x^*) = \Big\{ y \mid \nabla h(x^*)' y = 0,\ \nabla g_j(x^*)' y = 0,\ j \in A(x^*) \Big\}.$$

- Similar sufficiency conditions and sensitivity results. They require strict complementarity, i.e.,

$$\mu_j^* > 0, \qquad \forall\ j \in A(x^*),$$

as well as regularity of x^*.

PROOF OF KUHN-TUCKER CONDITIONS

Use equality-constraints result to obtain all the conditions except for $\mu_j^* \geq 0$ for $j \in A(x^*)$. Introduce the penalty functions

$$g_j^+(x) = \max\Big\{0, g_j(x)\Big\}, \qquad j = 1, \ldots, r,$$

and for $k = 1, 2, \ldots$, let x^k minimize

$$f(x) + \frac{k}{2}||h(x)||^2 + \frac{k}{2}\sum_{j=1}^{r}\Big(g_j^+(x)\Big)^2 + \frac{1}{2}||x - x^*||^2$$

over a closed sphere of x such that $f(x^*) \leq f(x)$. Using the same argument as for equality constraints,

$$\lambda_i^* = \lim_{k\to\infty} kh_i(x^k), \qquad i = 1, \ldots, m,$$

$$\mu_j^* = \lim_{k\to\infty} kg_j^+(x^k), \qquad j = 1, \ldots, r.$$

Since $g_j^+(x^k) \geq 0$, we obtain $\mu_j^* \geq 0$ for all j.

GENERAL SUFFICIENCY CONDITION

Consider the problem

minimize $f(x)$

subject to $x \in X, \qquad g_j(x) \leq 0, \quad j = 1, \ldots, r.$

Let x^* be feasible and μ^* satisfy

$$\mu_j^* \geq 0, \quad j = 1, \ldots, r, \qquad \mu_j^* = 0, \quad \forall\ j \notin A(x^*),$$

$$x^* = \arg\min_{x \in X} L(x, \mu^*).$$

Then x^* is a global minimum of the problem.

Proof: We have

$$f(x^*) = f(x^*) + \mu^{*\prime} g(x^*) = \min_{x \in X} \Big\{ f(x) + \mu^{*\prime} g(x) \Big\}$$

$$\leq \min_{x \in X,\ g(x) \leq 0} \Big\{ f(x) + \mu^{*\prime} g(x) \Big\} \leq \min_{x \in X,\ g(x) \leq 0} f(x),$$

where the first equality follows from the hypothesis, which implies that $\mu^{*\prime} g(x^*) = 0$, and the last inequality follows from the nonnegativity of μ^*. **Q.E.D.**

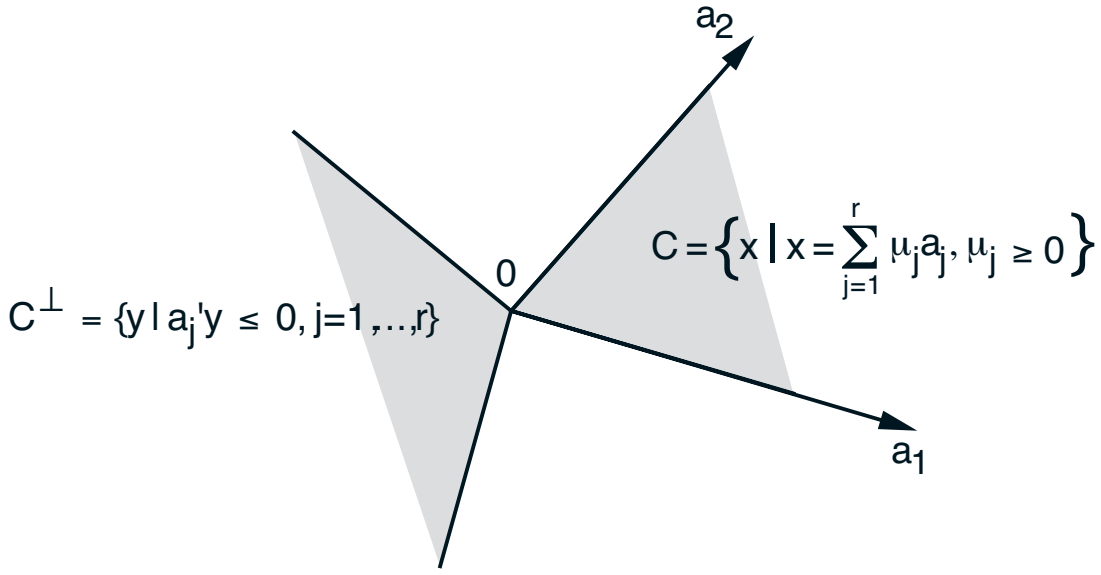
- Special Case: Let $X = \Re^n$, f and g_j be convex and differentiable. Then the 1st order Kuhn-Tucker conditions are also sufficient for global optimality.

LINEAR CONSTRAINTS

- Consider the problem $\min_{a_j' x \leq b_j,\, j=1,\ldots,r} f(x)$.
- Remarkable property: No need for regularity.
- Proposition: If x^* is a local minimum, there exist $\mu_1^*, \ldots, \mu_r^*$ with $\mu_j^* \geq 0$, $j = 1, \ldots, r$, such that

$$\nabla f(x^*) + \sum_{j=1}^{r} \mu_j^* a_j = 0, \qquad \mu_j^* = 0, \quad \forall\, j \notin A(x^*).$$

- The proof uses Farkas Lemma: Consider the cone C "generated" by a_j, $j \in A(x^*)$, and the "polar" cone $C^\perp$ shown below

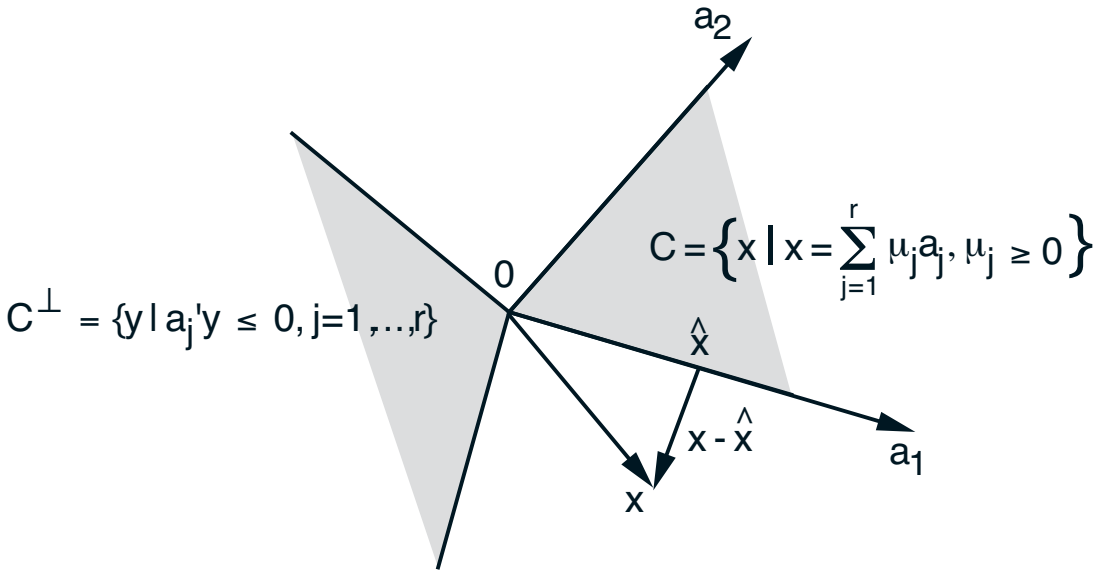


Then, $\left(C^\perp\right)^\perp = C$, i.e.,

$$x \in C \qquad \text{iff} \qquad x'y \leq 0, \quad \forall\, y \in C^\perp.$$

PROOF OF FARKAS LEMMA

$x \in C$ iff $x'y \leq 0, \quad \forall\ y \in C^{\perp}.$



Proof: First show that C is closed (nontrivial). Then, let x be such that $x'y \leq 0, \forall\ y \in C^{\perp}$, and consider its projection $\hat{x}$ on C. We have

$$x'(x - \hat{x}) = \|x - \hat{x}\|^2, \qquad (*)$$

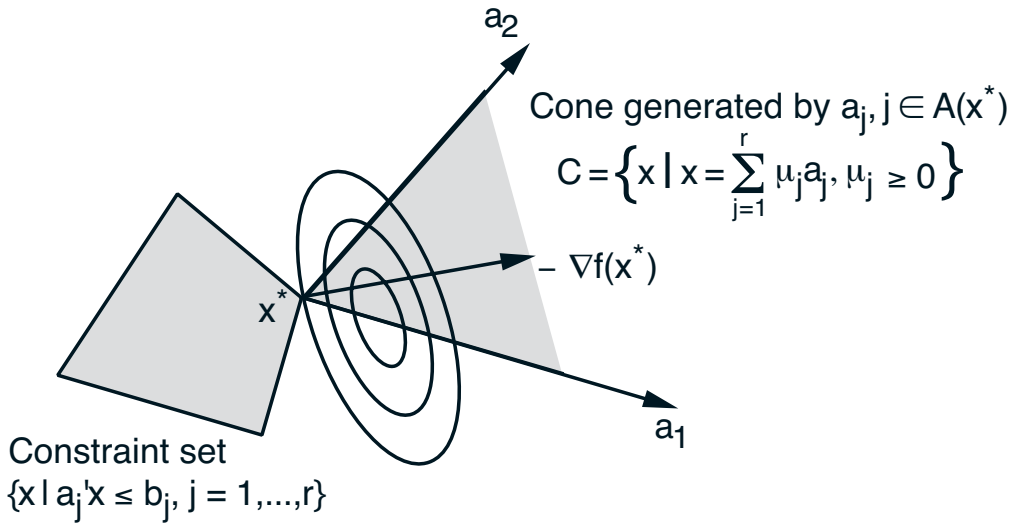
$$(x - \hat{x})'a_j \leq 0, \qquad \forall\ j.$$

Hence, $(x - \hat{x}) \in C^{\perp}$, and using the hypothesis,

$$x'(x - \hat{x}) \leq 0. \qquad (**)$$

From $(*)$ and $(**)$, we obtain $x = \hat{x}$, so $x \in C$.

PROOF OF LAGRANGE MULTIPLIER RESULT



The local min x^* of the original problem is also a local min for the problem $\min_{a_j' x \le b_j,\, j \in A(x^*)} f(x)$. Hence

$$\nabla f(x^*)'(x - x^*) \ge 0, \quad \forall\ x \text{ with } a_j' x \le b_j,\ j \in A(x^*).$$

Since a constraint $a_j' x \le b_j$, $j \in A(x^*)$ can also be expressed as $a_j'(x - x^*) \le 0$, we have

$$\nabla f(x^*)'y \ge 0, \qquad \forall\ y \text{ with } a_j' y \le 0,\ j \in A(x^*).$$

From Farkas' lemma, $-\nabla f(x^*)$ has the form

$$\sum_{j \in A(x^*)} \mu_j^* a_j, \quad \text{for some } \mu_j^* \ge 0,\ j \in A(x^*).$$

Let $\mu_j^* = 0$ for $j \notin A(x^*)$.

6.252 NONLINEAR PROGRAMMING

LECTURE 14: INTRODUCTION TO DUALITY

LECTURE OUTLINE

- Convex Cost/Linear Constraints
- Duality Theorem
- Linear Programming Duality
- Quadratic Programming Duality

Linear inequality constrained problem

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

where f is convex and continuously differentiable over $\Re^n$.

LAGRANGE MULTIPLIER RESULT

Let $J \subset \{1, \ldots, r\}$. Then x^* is a global min if and only if x^* is feasible and there exist $\mu_j^* \geq 0$, $j \in J$, such that $\mu_j^* = 0$ for all $j \in J \notin A(x^*)$, and

$$x^* = \arg \min_{\substack{a_j' x \leq b_j \\ j \notin J}} \left\{ f(x) + \sum_{j \in J} \mu_j^* (a_j' x - b_j) \right\}.$$

Proof: Assume x^* is global min. Then there exist $\mu_j^* \geq 0$, such that $\mu_j^*(a_j' x^* - b_j) = 0$ for all j and $\nabla f(x^*) + \sum_{j=1}^r \mu_j^* a_j = 0$, implying

$$x^* = \arg \min_{x \in \Re^n} \left\{ f(x) + \sum_{j=1}^r \mu_j^* (a_j' x - b_j) \right\}.$$

Since $\mu_j^*(a_j' x^* - b_j) = 0$ for all j,

$$f(x^*) = \min_{x \in \Re^n} \left\{ f(x) + \sum_{j=1}^r \mu_j^* (a_j' x - b_j) \right\}$$

$$\leq \min_{\substack{a_j' x \leq b_j \\ j \notin J}} \left\{ f(x) + \sum_{j=1}^r \mu_j^* (a_j' x - b_j) \right\}$$

Since $\mu_j^*(a_j' x - b_j) \leq 0$ if $a_j' x - b_j \leq 0$,

$$f(x^*) \leq \min_{\substack{a_j' x \leq b_j \\ j \notin J}} \left\{ f(x) + \sum_{j \in J} \mu_j^* (a_j' x - b_j) \right\} \leq f(x^*).$$

PROOF (CONTINUED)

Conversely, if x^* is feasible and there exist scalars μ_j^*, $j \in J$ with the stated properties, then x^* is a global min by the general sufficiency condition of the preceding lecture (where X is taken to be the set of x such that $a_j' x \leq b_j$ for all $j \notin J$). **Q.E.D.**

- **Interesting observation:** The same set of μ_j^* works for all index sets J.
- The flexibility to split the set of constraints into those that are handled by Lagrange multipliers (set J) and those that are handled explicitly comes handy in many analytical and computational contexts.

THE DUAL PROBLEM

- Consider the problem

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

where f is convex and cont. differentiable over $\Re^n$ and X is polyhedral.

- Define the *dual function* $q : \Re^r \mapsto [-\infty, \infty)$

$$q(\mu) = \inf_{x \in X} L(x, \mu) = \inf_{x \in X} \left\{ f(x) + \sum_{j=1}^{r} \mu_j (a_j' x - b_j) \right\}$$

and the *dual problem*

$$\max_{\mu \geq 0} q(\mu).$$

- If X is bounded, the dual function takes real values. In general, $q(\mu)$ can take the value $-\infty$. The "effective" constraint set of the dual is

$$Q = \left\{ \mu \mid \mu \geq 0,\ q(\mu) > -\infty \right\}.$$

DUALITY THEOREM

(a) If the primal problem has an optimal solution, the dual problem also has an optimal solution and the optimal values are equal.
(b) x^* is primal-optimal and μ^* is dual-optimal if and only if x^* is primal-feasible, $\mu^* \geq 0$, and

$$f(x^*) = L(x^*, \mu^*) = \min_{x \in X} L(x, \mu^*).$$

Proof: (a) Let x^* be a primal optimal solution. For all primal feasible x, and all $\mu \geq 0$, we have $\mu_j'(a_j'x - b_j) \leq 0$ for all j, so

$$\begin{aligned} q(\mu) &\leq \inf_{x \in X,\, a_j'x \leq b_j,\ j=1,\ldots,r} \left\{ f(x) + \sum_{j=1}^{r} \mu_j (a_j'x - b_j) \right\} \\ &\leq \inf_{x \in X,\, a_j'x \leq b_j,\ j=1,\ldots,r} f(x) = f(x^*). \end{aligned} \qquad (*)$$

By L-Mult. Th., there exists $\mu^* \geq 0$ such that $\mu_j^*(a_j'x^* - b_j) = 0$ for all j, and $x^* = \arg\min_{x \in X} L(x, \mu^*)$, so

$$q(\mu^*) = L(x^*, \mu^*) = f(x^*) + \sum_{j=1}^{r} \mu_j^*(a_j'x^* - b_j) = f(x^*).$$

PROOF (CONTINUED)

(b) If x^* is primal-optimal and μ^* is dual-optimal, by part (a)

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

which when combined with Eq. (*), yields

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

Conversely, the relation $f(x^*) = \min_{x \in X} L(x, \mu^*)$ is written as $f(x^*) = q(\mu^*)$, and since x^* is primal-feasible and $\mu^* \geq 0$, Eq. (*) implies that x^* is primal-optimal and μ^* is dual-optimal. **Q.E.D.**

- Linear equality constraints are treated similar to inequality constraints, except that the sign of the Lagrange multipliers is unrestricted:

Primal: $$\min_{x \in X,\, e_i' x = d_i,\, i=1,\ldots,m \;\; a_j' x \leq b_j,\;\; j=1,\ldots,r} f(x)$$

Dual: $$\max_{\lambda \in \Re^m,\, \mu \geq 0} q(\lambda, \mu) = \max_{\lambda \in \Re^m,\, \mu \geq 0} \inf_{x \in X} L(x, \lambda, \mu).$$

THE DUAL OF A LINEAR PROGRAM

- Consider the linear program

  minimize $c'x$

  subject to $e_i'x = d_i, \quad i = 1, \ldots, m, \qquad x \geq 0$

- Dual function

$$q(\lambda) = \inf_{x \geq 0} \left\{ \sum_{j=1}^{n} \left(c_j - \sum_{i=1}^{m} \lambda_i e_{ij} \right) x_j + \sum_{i=1}^{m} \lambda_i d_i \right\}.$$

- If $c_j - \sum_{i=1}^{m} \lambda_i e_{ij} \geq 0$ for all j, the infimum is attained for $x = 0$, and $q(\lambda) = \sum_{i=1}^{m} \lambda_i d_i$. If $c_j - \sum_{i=1}^{m} \lambda_i e_{ij} < 0$ for some j, the expression in braces can be arbitrarily small by taking x_j suff. large, so $q(\lambda) = -\infty$. Thus, the dual is

  maximize $\sum_{i=1}^{m} \lambda_i d_i$

  subject to $\sum_{i=1}^{m} \lambda_i e_{ij} \leq c_j, \qquad j = 1, \ldots, n.$

THE DUAL OF A QUADRATIC PROGRAM

- Consider the quadratic program

$$\text{minimize} \quad \tfrac{1}{2}x'Qx + c'x$$

$$\text{subject to} \quad Ax \leq b,$$

where Q is a given $n \times n$ positive definite symmetric matrix, A is a given $r \times n$ matrix, and $b \in \Re^r$ and $c \in \Re^n$ are given vectors.

- Dual function:

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

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

$$\text{minimize} \quad \tfrac{1}{2}\mu' P\mu + t'\mu$$

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

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

6.252 NONLINEAR PROGRAMMING

LECTURE 15: INTERIOR POINT METHODS

LECTURE OUTLINE

- Barrier and Interior Point Methods
- Linear Programs and the Logarithmic Barrier
- Path Following Using Newton's Method

Inequality constrained problem

minimize $f(x)$

subject to $x \in X, \qquad g_j(x) \leq 0, \ j = 1, \ldots, r,$

where f and g_j are continuous and X is closed. We assume that the set

$$S = \Big\{ x \in X \mid g_j(x) < 0, \, j = 1, \ldots, r \Big\}$$

is nonempty and any feasible point is in the closure of S.

BARRIER METHOD

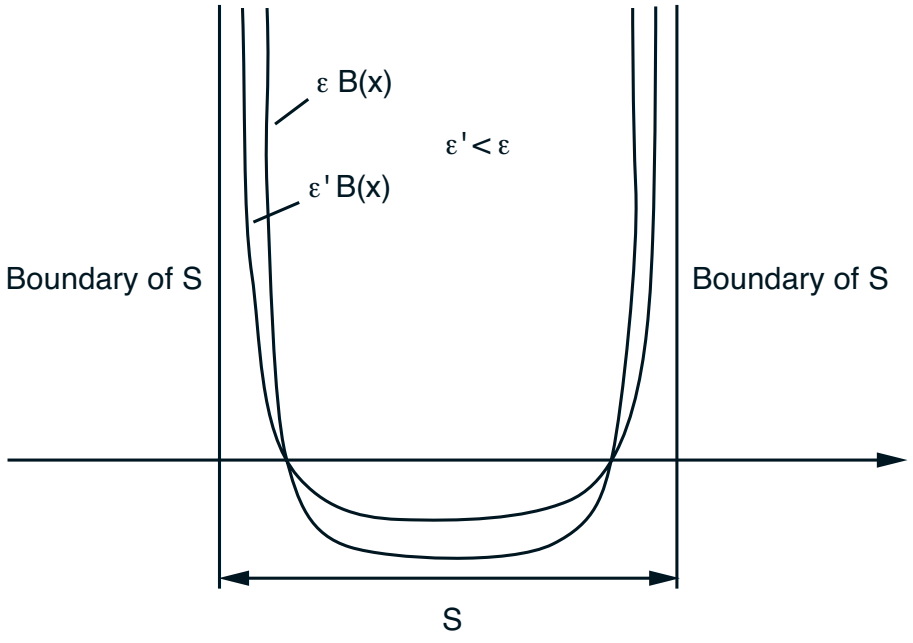
- Consider a *barrier function*, that is continuous and goes to ∞ as any one of the constraints $g_j(x)$ approaches 0 from negative values. Examples:

$$B(x) = -\sum_{j=1}^{r} \ln\bigl\{-g_j(x)\bigr\}, \qquad B(x) = -\sum_{j=1}^{r} \frac{1}{g_j(x)}.$$

- Barrier Method:

$$x^k = \arg\min_{x \in S} \Bigl\{ f(x) + \epsilon^k B(x) \Bigr\}, \qquad k = 0, 1, \ldots,$$

where the parameter sequence $\{\epsilon^k\}$ satisfies $0 < \epsilon^{k+1} < \epsilon^k$ for all k and $\epsilon^k \to 0$.



CONVERGENCE

Every limit point of a sequence $\{x^k\}$ generated by a barrier method is a global 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 global minimum, there exists a feasible x^* such that $f(x^*) < f(\overline{x})$ and therefore also an interior point $\tilde{x} \in S$ such that $f(\tilde{x}) < f(\overline{x})$. By the definition of x^k, $f(x^k)+\epsilon^k B(x^k) \leq f(\tilde{x}) + \epsilon^k B(\tilde{x})$ for all 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

$$\liminf_{k\to\infty} \epsilon^k B(x^k) \geq 0,$$

– a contradiction.

LINEAR PROGRAMS/LOGARITHMIC BARRIER

- Apply logarithmic barrier to the linear program

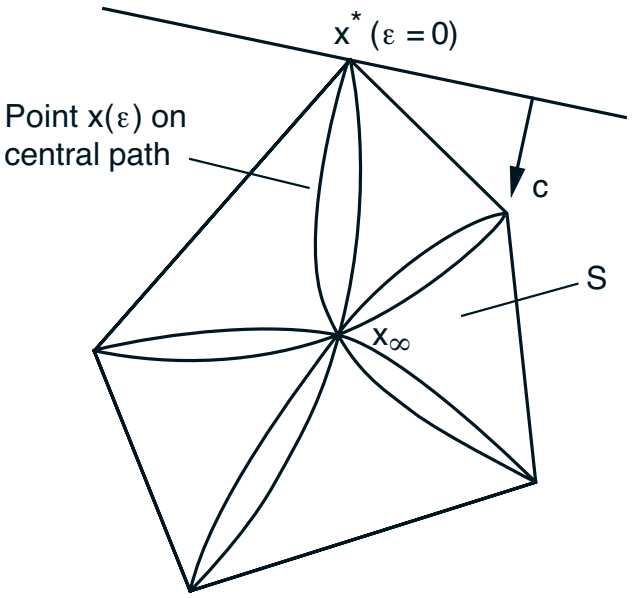
$$\begin{aligned} &\text{minimize} \quad c'x \\ &\text{subject to} \quad Ax = b, \qquad x \geq 0, \end{aligned} \tag{LP}$$

The method finds for various $\epsilon > 0$,

$$x(\epsilon) = \arg\min_{x \in S} F_\epsilon(x) = \arg\min_{x \in S} \left\{ c'x - \epsilon \sum_{i=1}^{n} \ln x_i \right\},$$

where $S = \{x \mid Ax = b,\ x > 0\}$. We assume that S is nonempty and bounded.

- As $\epsilon \to 0$, $x(\epsilon)$ follows the *central path*



All central paths start at the *analytic center*

$$x_\infty = \arg\min_{x \in S} \left\{ -\sum_{i=1}^{n} \ln x_i \right\},$$

and end at optimal solutions of (LP).

PATH FOLLOWING W/ NEWTON'S METHOD

- Newton's method for minimizing F_ϵ:

$$\tilde{x} = x + \alpha(\overline{x} - x),$$

where $\overline{x}$ is the pure Newton iterate

$$\overline{x} = \arg\min_{Az=b} \left\{ \nabla F_\epsilon(x)'(z-x) + \tfrac{1}{2}(z-x)'\nabla^2 F_\epsilon(x)(z-x) \right\}$$

- By straightforward calculation

$$\overline{x} = x - Xq(x,\epsilon),$$

$$q(x,\epsilon) = \frac{Xz}{\epsilon} - e, \quad e = (1 \dots 1)', \quad z = c - A'\lambda,$$

$$\lambda = (AX^2A')^{-1}AX\Big(Xc - \epsilon e\Big),$$

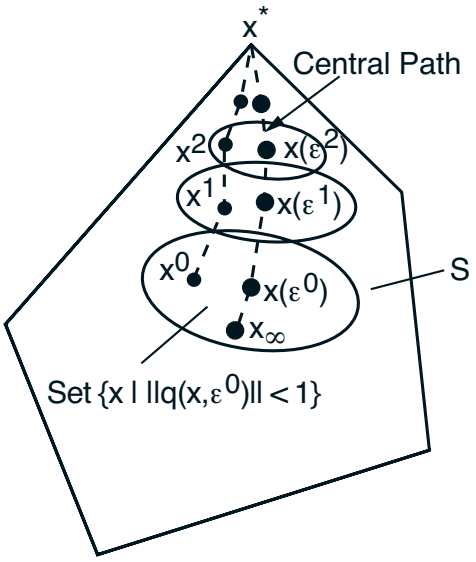
and X is the diagonal matrix with x_i, $i = 1, \dots, n$ along the diagonal.

- View $q(x,\epsilon)$ as the Newton increment $(x - \overline{x})$ transformed by X^{-1} that maps x into e.

- Consider $\|q(x,\epsilon)\|$ as a *proximity measure* of the current point to the point $x(\epsilon)$ on the central path.

KEY RESULTS

- It is sufficient to minimize F_ϵ approximately, up to where $\|q(x,\epsilon)\| < 1$.



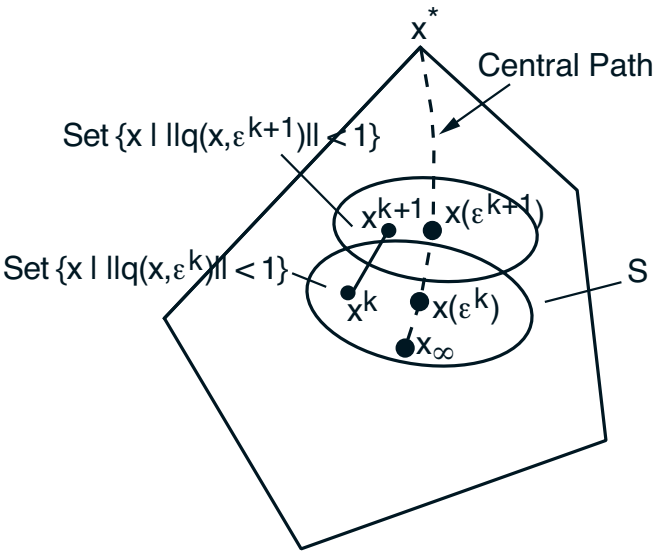
If $x > 0$, $Ax = b$, and $\|q(x,\epsilon)\| < 1$, then

$$c'x - \min_{Ay=b,\, y\geq 0} c'y \leq \epsilon\left(n+\sqrt{n}\right).$$

- The "termination set" $\left\{x \mid \|q(x,\epsilon)\| < 1\right\}$ is part of the region of quadratic convergence of the pure form of Newton's method. In particular, if $\|q(x,\epsilon)\| < 1$, then the pure Newton iterate $\overline{x} = x - Xq(x,\epsilon)$ is an interior point, that is, $\overline{x} \in S$. Furthermore, we have $\|q(\overline{x},\epsilon)\| < 1$ and in fact

$$\|q(\overline{x},\epsilon)\| \leq \|q(x,\epsilon)\|^2.$$

SHORT STEP METHODS



Following approximately the central path by using a single Newton step for each ϵ^k. If ϵ^k is close to ϵ^{k+1} and x^k is close to the central path, one expects that x^{k+1} obtained from x^k by a single pure Newton step will also be close to the central path.

Proposition Let $x > 0$, $Ax = b$, and suppose that for some $\gamma < 1$ we have $\|q(x, \epsilon)\| \leq \gamma$. Then if $\overline{\epsilon} = (1 - \delta n^{-1/2})\epsilon$ for some $\delta > 0$,

$$\|q(\overline{x}, \overline{\epsilon})\| \leq \frac{\gamma^2 + \delta}{1 - \delta n^{-1/2}}.$$

In particular, if

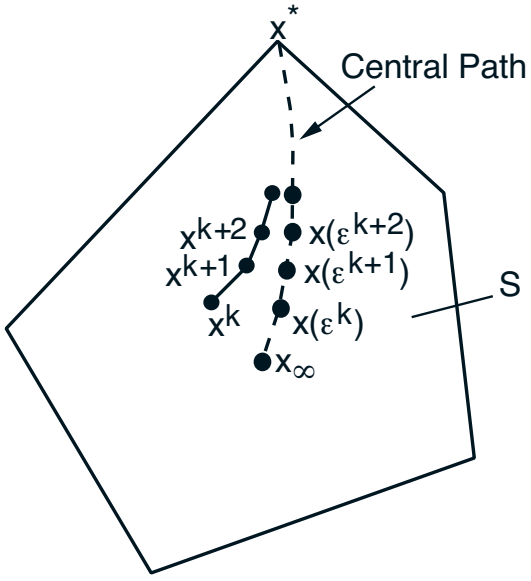
$$\delta \leq \gamma(1 - \gamma)(1 + \gamma)^{-1},$$

we have $\|q(\overline{x}, \overline{\epsilon})\| \leq \gamma$.

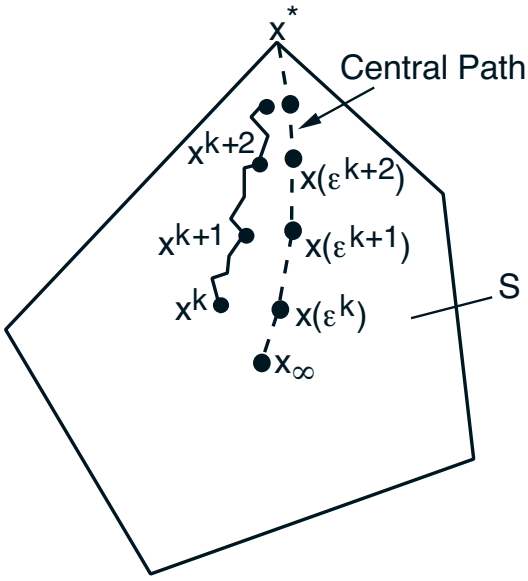
- Can be used to establish nice complexity results; but ϵ must be reduced VERY slowly.

LONG STEP METHODS

- Main features:
  - Decrease ϵ faster than dictated by complexity analysis.
  - Require more than one Newton step per (approximate) minimization.
  - Use line search as in unconstrained Newton's method.
  - Require much smaller number of (approximate) minimizations.



(a)



(b)

- The methodology generalizes to quadratic programming and convex programming.

6.252 NONLINEAR PROGRAMMING

LECTURE 16: PENALTY METHODS

LECTURE OUTLINE

- Quadratic Penalty Methods
- Introduction to Multiplier Methods

****

- Consider the equality constrained problem

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

$$\text{subject to} \quad x \in X, \qquad h(x) = 0,$$

where $f : \Re^n \to \Re$ and $h : \Re^n \to \Re^m$ are continuous, and X is closed.

- The quadratic penalty method:

$$x^k = \arg\min_{x \in X} L_{c^k}(x, \lambda^k) \equiv f(x) + \lambda^{k\prime} h(x) + \frac{c^k}{2} \|h(x)\|^2$$

where the $\{\lambda^k\}$ is a bounded sequence and $\{c^k\}$ satisfies $0 < c^k < c^{k+1}$ for all k and $c^k \to \infty$.

TWO CONVERGENCE MECHANISMS

- Taking λ^k close to a Lagrange multiplier vector
  - Assume $X = \Re^n$ and (x^*, λ^*) is a local min-Lagrange multiplier pair satisfying the 2nd order sufficiency conditions
  - For c suff. large, x^* is a strict local min of $L_c(\cdot, \lambda^*)$
- Taking c^k very large
  - For large c and any λ

$$L_c(\cdot, \lambda) \approx \begin{cases} f(x) & \text{if } x \in X \text{ and } h(x) = 0 \\ \infty & \text{otherwise} \end{cases}$$

- Example:

minimize $f(x) = \frac{1}{2}(x_1^2 + x_2^2)$

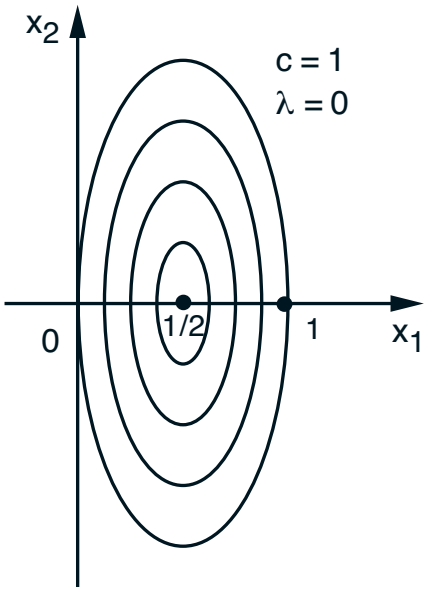
subject to $x_1 = 1$

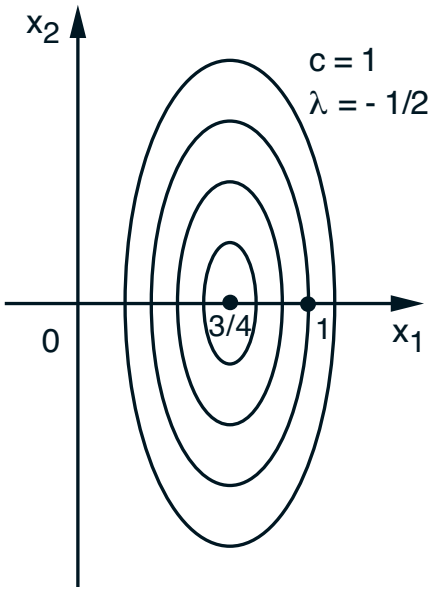
$$L_c(x, \lambda) = \tfrac{1}{2}(x_1^2 + x_2^2) + \lambda(x_1 - 1) + \frac{c}{2}(x_1 - 1)^2$$

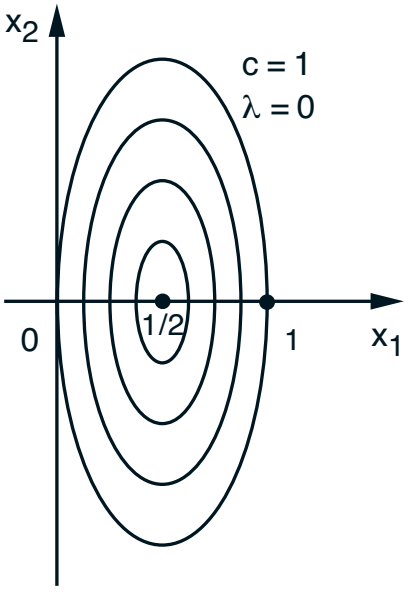
$$x_1(\lambda, c) = \frac{c - \lambda}{c + 1}, \qquad x_2(\lambda, c) = 0$$

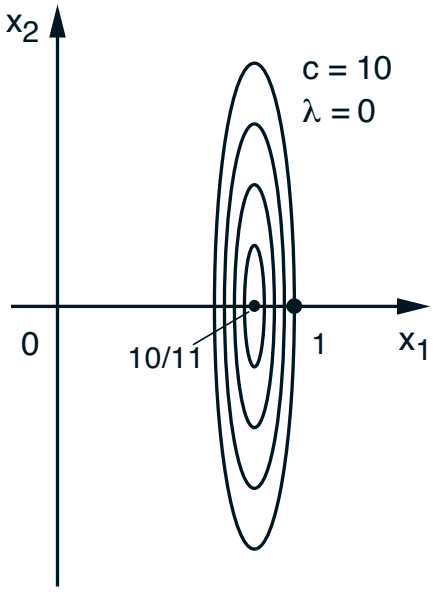
EXAMPLE CONTINUED

$$\min_{x_1=1} x_1^2 + x_2^2, \qquad x^* = 1, \quad \lambda^* = -1$$









GLOBAL CONVERGENCE

- Every limit point of $\{x^k\}$ is a global min.

Proof: The optimal value of the problem is $f^* = \inf_{h(x)=0,\, x\in X} L_{c^k}(x, \lambda^k)$. We have

$$L_{c^k}(x^k, \lambda^k) \leq L_{c^k}(x, \lambda^k), \qquad \forall\ x \in X$$

so taking the inf of the RHS over $x \in X$, $h(x) = 0$

$$L_{c^k}(x^k, \lambda^k) = f(x^k) + \lambda^{k\prime} h(x^k) + \frac{c^k}{2}\|h(x^k)\|^2 \leq f^*.$$

Let $(\bar{x}, \bar{\lambda})$ be a limit point of $\{x^k, \lambda^k\}$. Without loss of generality, assume that $\{x^k, \lambda^k\} \to (\bar{x}, \bar{\lambda})$. Taking the limsup above

$$f(\bar{x}) + \bar{\lambda}' h(\bar{x}) + \limsup_{k\to\infty} \frac{c^k}{2}\|h(x^k)\|^2 \leq f^*. \qquad (*)$$

Since $\|h(x^k)\|^2 \geq 0$ and $c^k \to \infty$, it follows that $h(x^k) \to 0$ and $h(\bar{x}) = 0$. Hence, $\bar{x}$ is feasible, and since from Eq. (*) we have $f(\bar{x}) \leq f^*$, $\bar{x}$ is optimal. **Q.E.D.**

LAGRANGE MULTIPLIER ESTIMATES

- Assume that $X = \Re^n$, and f and h are cont. differentiable. Let $\{\lambda^k\}$ be bounded, and $c^k \to \infty$. Assume x^k satisfies $\nabla_x L_{c^k}(x^k, \lambda^k) = 0$ for all k, and that $x^k \to x^*$, where x^* is such that $\nabla h(x^*)$ has rank m. Then $h(x^*) = 0$ and $\tilde{\lambda}^k \to \lambda^*$, where

$$\tilde{\lambda}^k = \lambda^k + c^k h(x^k), \qquad \nabla_x L(x^*, \lambda^*) = 0.$$

Proof: We have

$$\begin{aligned} 0 = \nabla_x L_{c^k}(x^k, \lambda^k) &= \nabla f(x^k) + \nabla h(x^k)\big(\lambda^k + c^k h(x^k)\big) \\ &= \nabla f(x^k) + \nabla h(x^k)\tilde{\lambda}^k. \end{aligned}$$

Multiply with

$$\Big(\nabla h(x^k)'\nabla h(x^k)\Big)^{-1}\nabla h(x^k)'$$

and take lim to obtain $\tilde{\lambda}^k \to \lambda^*$ with

$$\lambda^* = -\Big(\nabla h(x^*)'\nabla h(x^*)\Big)^{-1}\nabla h(x^*)'\nabla f(x^*).$$

We also have $\nabla_x L(x^*, \lambda^*) = 0$ and $h(x^*) = 0$ (since $\tilde{\lambda}^k$ converges).

PRACTICAL BEHAVIOR

- Three possibilities:
  - The method breaks down because an x^k with $\nabla_x L_{c^k}(x^k, \lambda^k) \approx 0$ cannot be found.
  - A sequence $\{x^k\}$ with $\nabla_x L_{c^k}(x^k, \lambda^k) \approx 0$ is obtained, but it either has no limit points, or for each of its limit points x^* the matrix $\nabla h(x^*)$ has rank $< m$.
  - A sequence $\{x^k\}$ with with $\nabla_x L_{c^k}(x^k, \lambda^k) \approx 0$ is found and it has a limit point x^* such that $\nabla h(x^*)$ has rank m. Then, x^* together with λ^* [the corresp. limit point of $\left\{\lambda^k + c^k h(x^k)\right\}$] satisfies the first-order necessary conditions.

- Ill-conditioning: The condition number of the Hessian $\nabla^2_{xx} L_{c^k}(x^k, \lambda^k)$ tends to increase with c^k.

- To overcome ill-conditioning:
  - Use Newton-like method (and double precision).
  - Use good starting points.
  - Increase c^k at a moderate rate (if c^k is increased at a fast rate, $\{x^k\}$ converges faster, but the likelihood of ill-conditioning is greater).

INEQUALITY CONSTRAINTS

- Convert them to equality constraints by using squared slack variables that are eliminated later.
- Convert inequality constraint $g_j(x) \leq 0$ to equality constraint $g_j(x) + z_j^2 = 0$.
- The penalty method solves problems of the form

$$\min_{x,z} \bar{L}_c(x, z, \lambda, \mu) = f(x) + \sum_{j=1}^{r} \left\{ \mu_j \left(g_j(x) + z_j^2 \right) + \frac{c}{2} |g_j(x) + z_j^2|^2 \right\},$$

for various values of μ and c.

- First minimize $\bar{L}_c(x, z, \lambda, \mu)$ with respect to z,

$$L_c(x, \lambda, \mu) = \min_z \bar{L}_c(x, z, \lambda, \mu) = f(x) + \sum_{j=1}^{r} \min_{z_j} \left\{ \mu_j \left(g_j(x) + z_j^2 \right) + \frac{c}{2} |g_j(x) + z_j^2|^2 \right\}$$

and then minimize $L_c(x, \lambda, \mu)$ with respect to x.

MULTIPLIER METHODS

- Recall that if (x^*, λ^*) is a local min-Lagrange multiplier pair satisfying the 2nd order sufficiency conditions, then for c suff. large, x^* is a strict local min of $L_c(\cdot, \lambda^*)$.
- This suggests that for $\lambda^k \approx \lambda^*$, $x^k \approx x^*$.
- Hence it is a good idea to use $\lambda^k \approx \lambda^*$, such as

$$\lambda^{k+1} = \tilde{\lambda}^k = \lambda^k + c^k h(x^k)$$

This is the (1st order) method of multipliers.

- Key advantages to be shown:
  - Less ill-conditioning: It is not necessary that $c^k \to \infty$ (only that c^k exceeds some threshold).
  - Faster convergence when λ^k is updated than when λ^k is kept constant (whether $c^k \to \infty$ or not).

6.252 NONLINEAR PROGRAMMING

LECTURE 17: AUGMENTED LAGRANGIAN METHODS

LECTURE OUTLINE

- Multiplier Methods

****

- Consider the equality constrained problem

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

$$\text{subject to } h(x) = 0,$$

where $f : \Re^n \to \Re$ and $h : \Re^n \to \Re^m$ are continuously differentiable.

- The (1st order) multiplier method finds

$$x^k = \arg \min_{x \in \Re^n} L_{c^k}(x, \lambda^k) \equiv f(x) + \lambda^{k\prime} h(x) + \frac{c^k}{2} \|h(x)\|^2$$

and updates λ^k using

$$\lambda^{k+1} = \lambda^k + c^k h(x^k)$$

CONVEX EXAMPLE

- Problem: $\min_{x_1=1}(1/2)(x_1^2+x_2^2)$ with optimal solution $x^*=(1,0)$ and Lagr. multiplier $\lambda^*=-1$.
- We have

$$x^k=\arg\min_{x\in\Re^n} L_{c^k}(x,\lambda^k)=\left(\frac{c^k-\lambda^k}{c^k+1},0\right)$$

$$\lambda^{k+1}=\lambda^k+c^k\left(\frac{c^k-\lambda^k}{c^k+1}-1\right)$$

$$\lambda^{k+1}-\lambda^*=\frac{\lambda^k-\lambda^*}{c^k+1}$$

- We see that:
  - $\lambda^k\to\lambda^*=-1$ and $x^k\to x^*=(1,0)$ for every nondecreasing sequence $\{c^k\}$. It is NOT necessary to increase c^k to ∞.
  - The convergence rate becomes faster as c^k becomes larger; in fact $\left\{|\lambda^k-\lambda^*|\right\}$ converges superlinearly if $c^k\to\infty$.

NONCONVEX EXAMPLE

- Problem: $\min_{x_1=1}(1/2)(-x_1^2+x_2^2)$ with optimal solution $x^*=(1,0)$ and Lagr. multiplier $\lambda^*=1$.
- We have

$$x^k=\arg\min_{x\in\Re^n} L_{c^k}(x,\lambda^k)=\left(\frac{c^k-\lambda^k}{c^k-1},0\right)$$

provided $c^k>1$ (otherwise the min does not exist)

$$\lambda^{k+1}=\lambda^k+c^k\left(\frac{c^k-\lambda^k}{c^k-1}-1\right)$$

$$\lambda^{k+1}-\lambda^*=-\frac{\lambda^k-\lambda^*}{c^k-1}$$

- We see that:
  - No need to increase c^k to ∞ for convergence; doing so results in faster convergence rate.
  - To obtain convergence, c^k must eventually exceed the threshold 2.

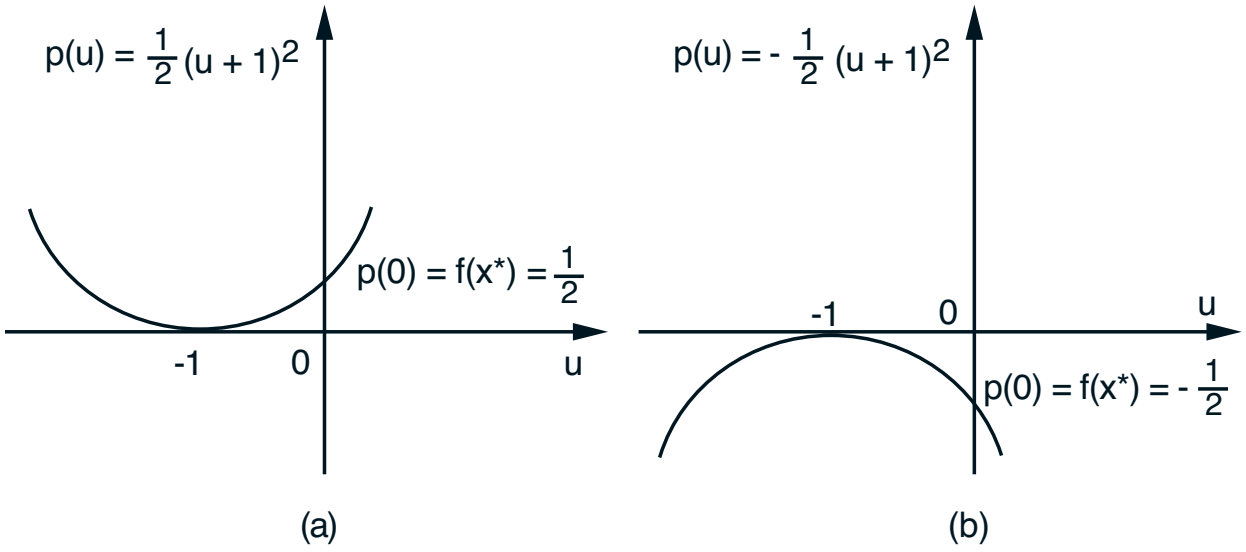
THE PRIMAL FUNCTIONAL

- Let (x^*, λ^*) be a regular local min-Lagr. pair satisfying the 2nd order suff. conditions are satisfied.
- The primal functional

$$p(u) = \min_{h(x)=u} f(x),$$

defined for u in an open sphere centered at $u = 0$, and we have

$$p(0) = f(x^*), \qquad \nabla p(0) = -\lambda^*,$$



$$p(u) = \min_{x_1 - 1 = u} \tfrac{1}{2}(x_1^2 + x_2^2), \qquad p(u) = \min_{x_1 - 1 = u} \tfrac{1}{2}(-x_1^2 + x_2^2)$$

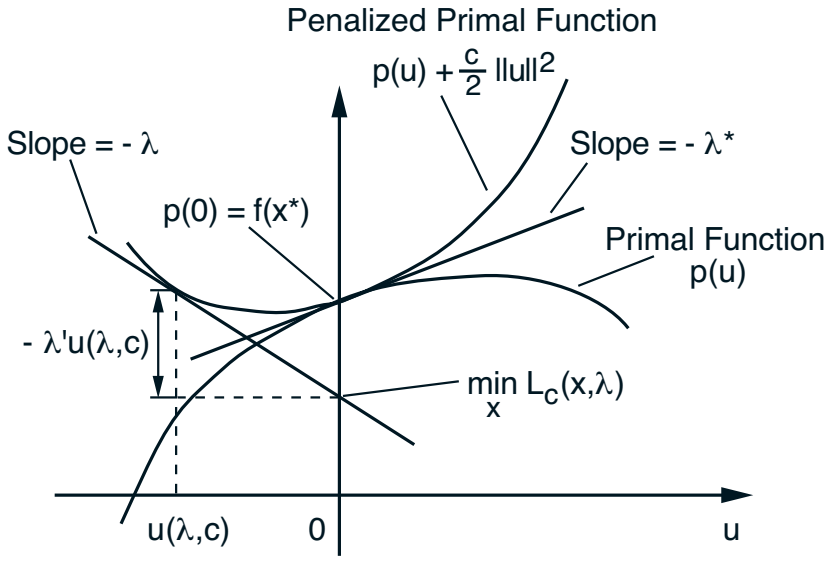
AUGM. LAGRANGIAN MINIMIZATION

- Break down the minimization of $L_c(\cdot, \lambda)$:

$$\min_x L_c(x,\lambda) = \min_u \min_{h(x)=u} \left\{ f(x) + \lambda' h(x) + \frac{c}{2}\|h(x)\|^2 \right\}$$
$$= \min_u \left\{ p(u) + \lambda' u + \frac{c}{2}\|u\|^2 \right\},$$

where the minimization above is understood to be local in a neighborhood of $u = 0$.

- Interpretation of this minimization:

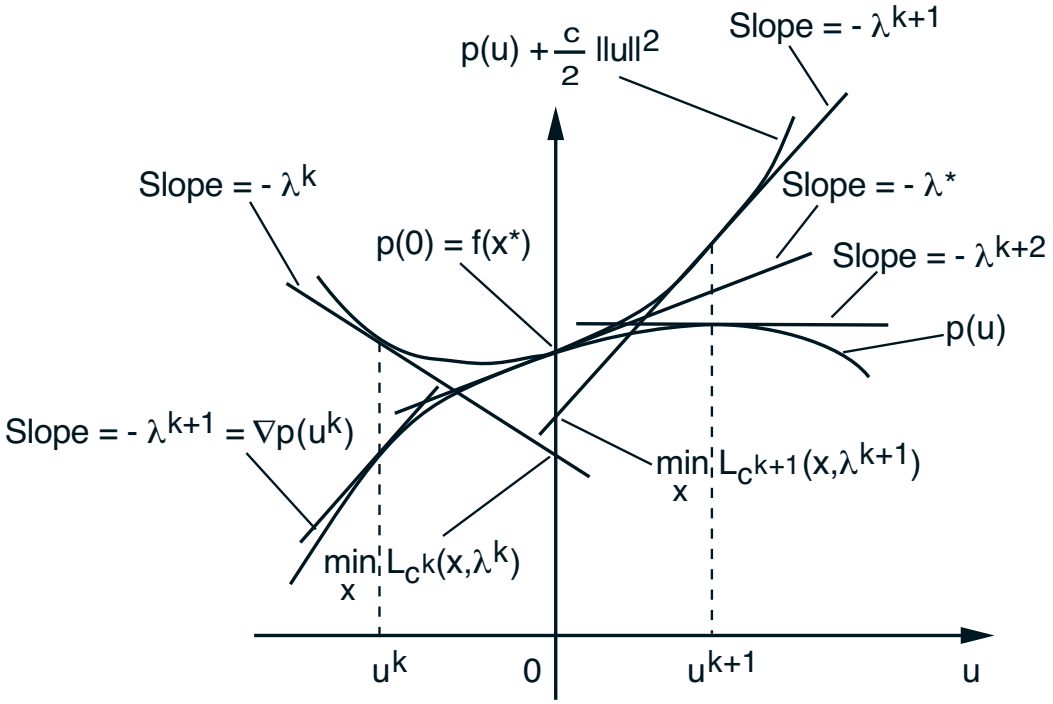


- If c is suf. large, $p(u) + \lambda' u + \frac{c}{2}\|u\|^2$ is convex in a neighborhood of 0. Also, for $\lambda \approx \lambda^*$ and large c, the value $\min_x L_c(x,\lambda) \approx p(0) = f(x^*)$.

INTERPRETATION OF THE METHOD

- Geometric interpretation of the iteration

$$\lambda^{k+1} = \lambda^k + c^k h(x^k).$$



- If λ^k is sufficiently close to λ^* and/or c^k is suf. large, λ^{k+1} will be closer to λ^* than λ^k.

- c^k need not be increased to ∞ in order to obtain convergence; it is sufficient that c^k eventually exceeds some threshold level.

- If $p(u)$ is linear, convergence to λ^* will be achieved in one iteration.

COMPUTATIONAL ASPECTS

- Key issue is how to select $\{c^k\}$.
  - c^k should eventually become larger than the "threshold" of the given problem.
  - c^0 should not be so large as to cause ill-conditioning at the 1st minimization.
  - c^k should not be increased so fast that too much ill-conditioning is forced upon the unconstrained minimization too early.
  - c^k should not be increased so slowly that the multiplier iteration has poor convergence rate.

- A good practical scheme is to choose a moderate value c^0, and use $c^{k+1} = \beta c^k$, where β is a scalar with $\beta > 1$ (typically $\beta \in [5, 10]$ if a Newton-like method is used).

- In practice the minimization of $L_{c^k}(x, \lambda^k)$ is typically inexact (usually exact asymptotically). In some variants of the method, only one Newton step per minimization is used (with safeguards).

DUALITY FRAMEWORK

- Consider the problem

$$\text{minimize} \quad f(x) + \frac{c}{2}\|h(x)\|^2$$

$$\text{subject to} \quad \|x - x^*\| < \epsilon, \qquad h(x) = 0,$$

where ϵ is small enough for a local analysis to hold based on the implicit function theorem, and c is large enough for the minimum to exist.

- Consider the dual function and its gradient

$$q_c(\lambda) = \min_{\|x-x^*\|<\epsilon} L_c(x,\lambda) = L_c\Big(x(\lambda,c),\lambda\Big)$$

$$\begin{aligned}\nabla q_c(\lambda) &= \nabla_\lambda x(\lambda,c)\nabla_x L_c\Big(x(\lambda,c),\lambda\Big) + h\Big(x(\lambda,c)\Big)\\ &= h\Big(x(\lambda,c)\Big).\end{aligned}$$

We have $\nabla q_c(\lambda^*) = h(x^*) = 0$ and $\nabla^2 q_c(\lambda^*) < 0$.

- The multiplier method is a steepest ascent iteration for maximizing q_{c^k}

$$\lambda^{k+1} = \lambda^k + c^k \nabla q_{c^k}(\lambda^k),$$

6.252 NONLINEAR PROGRAMMING

LECTURE 18: DUALITY THEORY

LECTURE OUTLINE

- Geometrical Framework for Duality
- Geometric Multipliers
- The Dual Problem
- Properties of the Dual Function
- Consider the problem

  minimize $f(x)$

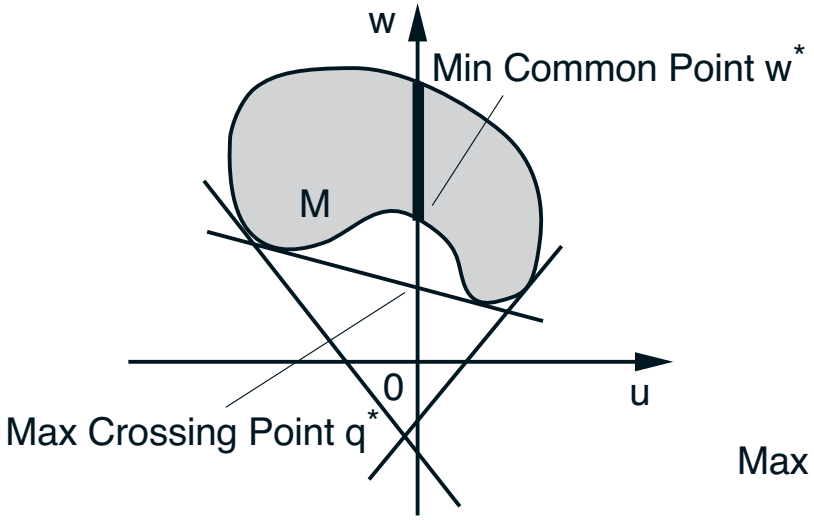
  subject to $x \in X, \qquad g_j(x) \leq 0, \qquad j = 1, \ldots, r,$

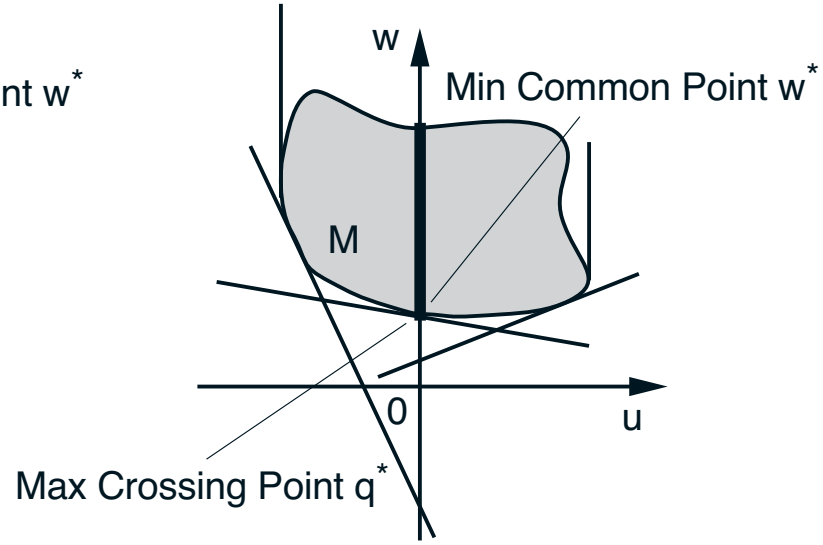
- We assume that the problem is feasible and the cost is bounded from below:

$$-\infty < f^* = \inf_{\substack{x \in X \\ g_j(x) \leq 0,\ j=1,\ldots,r}} f(x) < \infty$$

MIN COMMON POINT/MAX CROSSING POINT

- Let M be a subset of $\Re^n$:
- $Min\ Common\ Point\ Problem$: Among all points that are common to both M and the nth axis,find the one whose nth component is minimum.
- $Max\ Crossing\ Point\ Problem$: Among all hyperplanes that intersect the nth axis and support the set M from "below", find the hyperplane for which point of intercept with the nth axis is maximum.





- Note: We will see that the min common/max crossing framework applies to the problem of the preceding slide, i.e., $min_{x \in X,\, g(x) \leq 0} f(x)$, with the choice

$$M = \Big\{ (z, f(x)) \mid g(x) \leq z,\ x \in X \Big\}$$

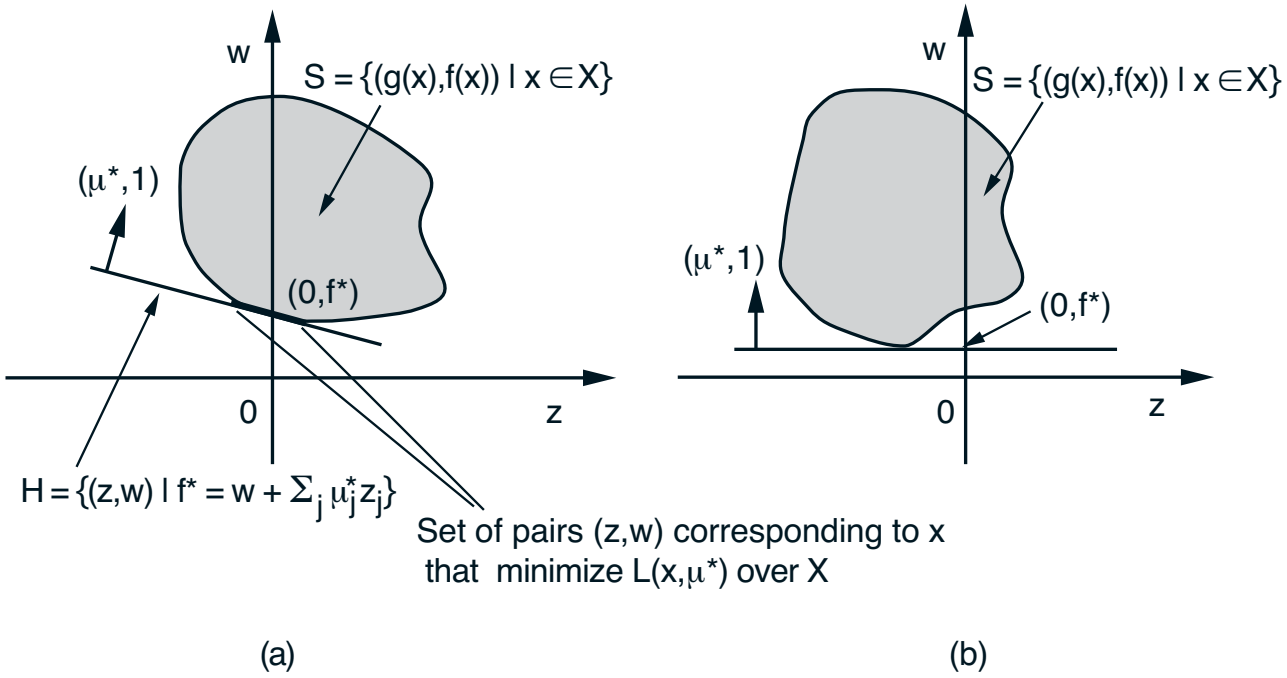
GEOMETRICAL DEFINITION OF A MULTIPLIER

- A vector $\mu^* = (\mu_1^*, \ldots, \mu_r^*)$ is said to be a *geometric multiplier* for the primal problem if

$$\mu_j^* \geq 0, \qquad j = 1, \ldots, r,$$

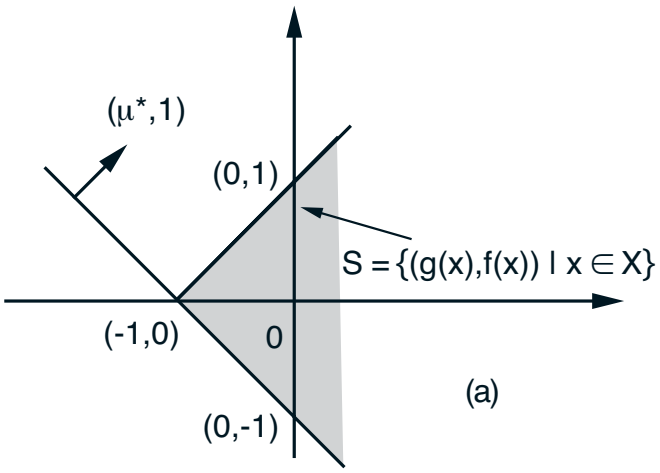
and

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

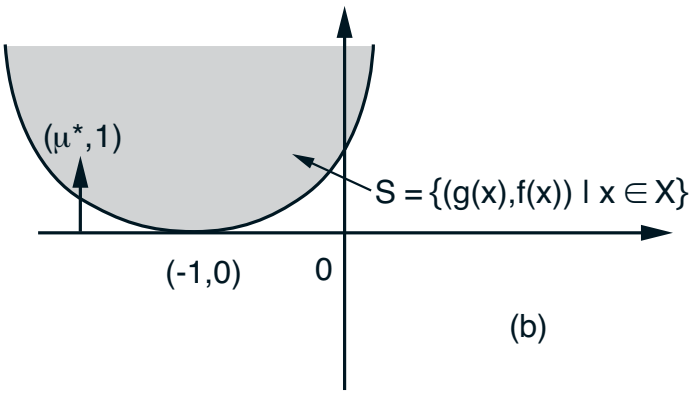


- Note that the definition differs from the one given in Chapter 3 ... but if $X = \Re^n$, f and g_j are convex and differentiable, and x^* is an optimal solution, the Lagrange multipliers corresponding to x^*, as per the definition of Chapter 3, coincide with the geometric multipliers as per the above definition.

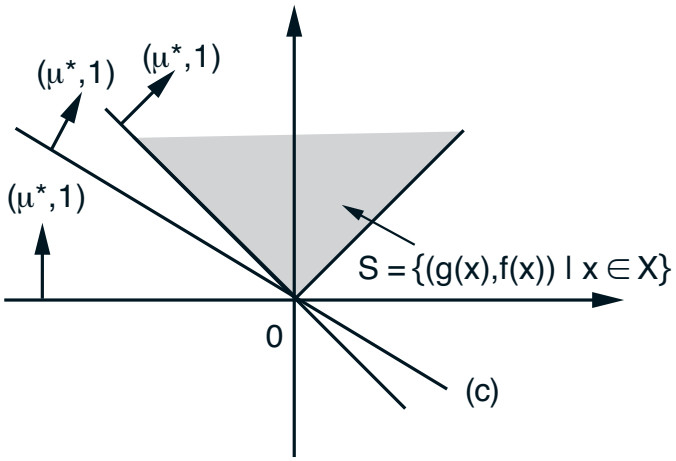
EXAMPLES: A G-MULTIPLIER EXISTS



$$\min\ f(x) = x_1 - x_2$$
$$\text{s.t. } g(x) = x_1 + x_2 - 1 \le 0$$
$$x \in X = \{(x_1, x_2) \mid x_1 \ge 0,\ x_2 \ge 0\}$$

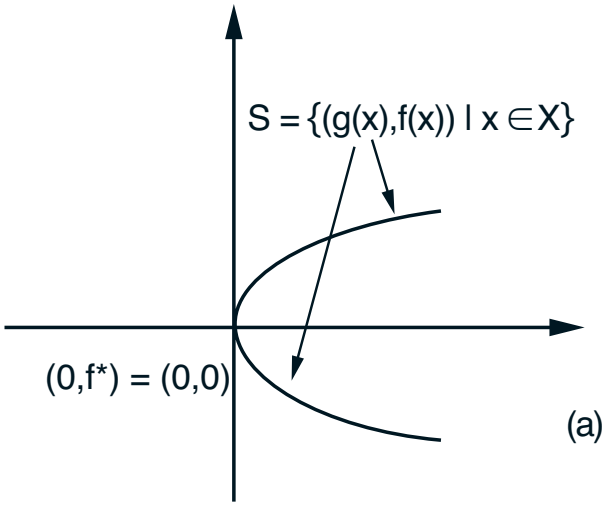


$$\min\ f(x) = (1/2)\,(x_1^2 + x_2^2)$$
$$\text{s.t. } g(x) = x_1 - 1 \le 0$$
$$x \in X = R^2$$



$$\min\ f(x) = |x_1| + x_2$$
$$\text{s.t. } g(x) = x_1 \le 0$$
$$x \in X = \{(x_1, x_2) \mid x_2 \ge 0\}$$

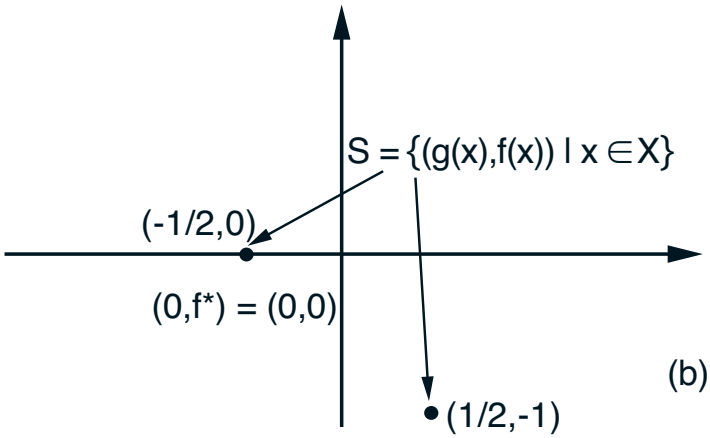
EXAMPLES: A G-MULTIPLIER DOESN'T EXIST



min f(x) = x

s.t. g(x) = $x^2 \le 0$

x ∈X = R



min f(x) = - x

s.t. g(x) = x - 1/2 ≤ 0

x ∈X = {0,1}

- Proposition: Let μ^* be a geometric multiplier. Then x^* is a global minimum of the primal problem if and only if x^* is feasible and

$$x^* = \arg\min_{x \in X} L(x, \mu^*), \qquad \mu_j^* g_j(x^*) = 0, \qquad j = 1, \ldots, r$$

THE DUAL FUNCTION AND THE DUAL PROBLEM

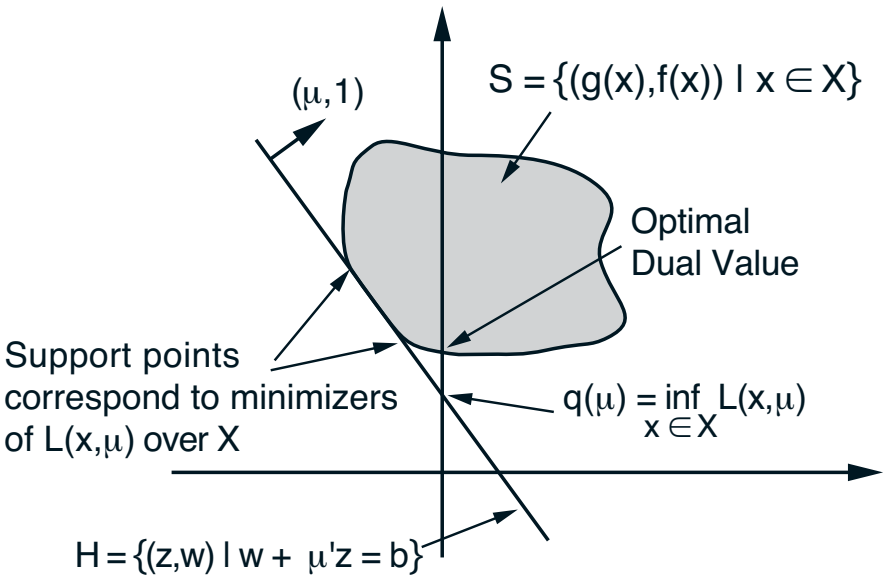
- The $dual\ problem$ is

$$\text{maximize } q(\mu)$$
$$\text{subject to } \mu \geq 0,$$

where q is the dual function

$$q(\mu) = \inf_{x \in X} L(x, \mu), \qquad \forall\ \mu \in \Re^r.$$

- Question: How does the optimal dual value $q^* = \sup_{\mu \geq 0} q(\mu)$ relate to f^*?



WEAK DUALITY

- The $domain$ of q is

$$D_q = \Big\{\mu \mid q(\mu) > -\infty\Big\}.$$

- Proposition: The domain D_q is a convex set and q is concave over D_q.
- Proposition: (Weak Duality Theorem) We have

$$q^* \leq f^*.$$

Proof: For all $\mu \geq 0$, and $x \in X$ with $g(x) \leq 0$, we have

$$q(\mu) = \inf_{z \in X} L(z, \mu) \leq f(x) + \sum_{j=1}^{r} \mu_j g_j(x) \leq f(x),$$

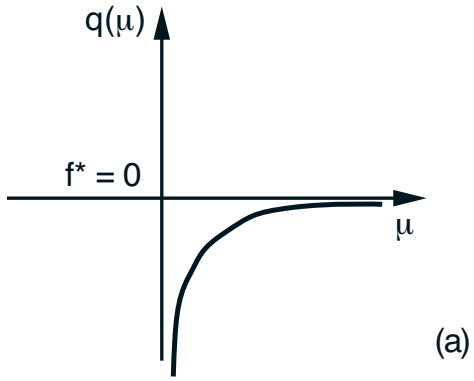
so

$$q^* = \sup_{\mu \geq 0} q(\mu) \leq \inf_{x \in X,\, g(x) \leq 0} f(x) = f^*.$$

DUAL OPTIMAL SOLUTIONS AND G-MULTIPLIERS

- Proposition: (a) If $q^* = f^*$, the set of geometric multipliers is equal to the set of optimal dual solutions. (b) If $q^* < f^*$, the set of geometric multipliers is empty.

Proof: By definition, a vector $\mu^* \geq 0$ is a geometric multiplier if and only if $f^* = q(\mu^*) \leq q^*$, which by the weak duality theorem, holds if and only if there is no duality gap and μ^* is a dual optimal solution. **Q.E.D.**

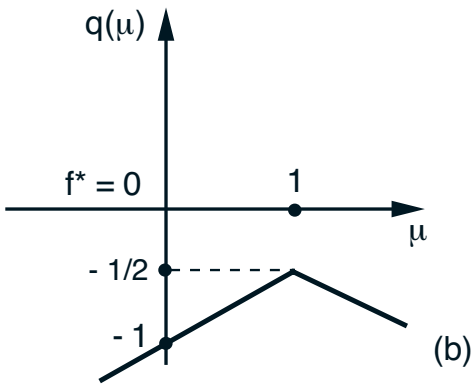


(a)

min f(x) = x

s.t. g(x) = $x^2 \leq 0$
x ∈ X = R

$$q(\mu) = \min_{x \in R} \{x + \mu x^2\} = \begin{cases} -1/(4\mu) & \text{if } \mu > 0 \\ -\infty & \text{if } \mu \leq 0 \end{cases}$$



(b)

min f(x) = - x

s.t. g(x) = x - 1/2 ≤ 0
x ∈ X = {0,1}

$$q(\mu) = \min_{x \in \{0,1\}} \{-x + \mu(x - 1/2)\} = \min\{-\mu/2, \mu/2 - 1\}$$

6.252 NONLINEAR PROGRAMMING

LECTURE 19: DUALITY THEOREMS

LECTURE OUTLINE

- Duality and G-multipliers (continued)
- Consider the problem

  minimize $f(x)$

  subject to $x \in X, \qquad g_j(x) \leq 0, \qquad j = 1, \ldots, r,$

assuming $-\infty < f^* < \infty$.

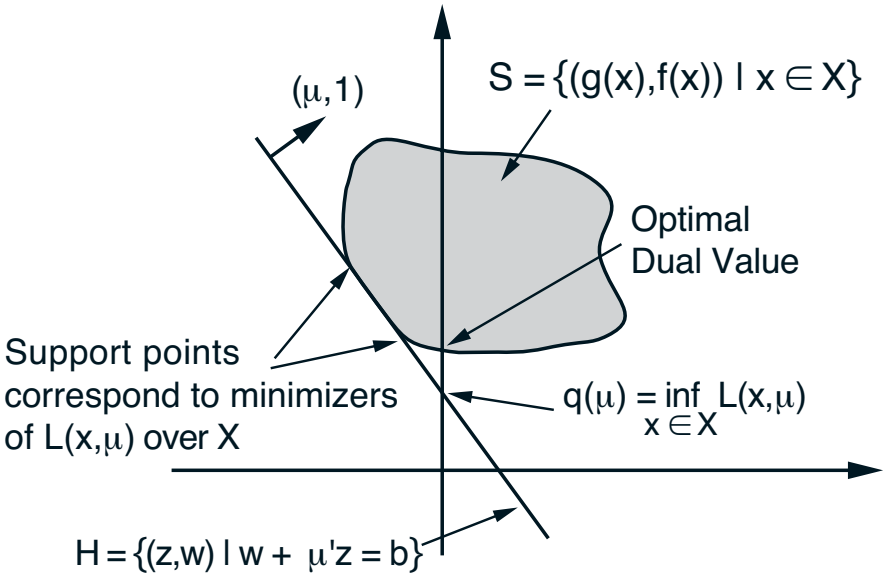
- μ^* is a geometric multiplier if $\mu^* \geq 0$ and $f^* = \inf_{x \in X} L(x, \mu^*)$.
- The *dual problem* is

  maximize $q(\mu)$

  subject to $\mu \geq 0,$

where q is the dual function $q(\mu) = \inf_{x \in X} L(x, \mu)$.

DUAL OPTIMALITY



- Weak Duality Theorem: $q^* \leq f^*$.
- Geometric Multipliers and Dual Optimal Solutions:
  - (a) If there is no duality gap, the set of geometric multipliers is equal to the set of optimal dual solutions.
  - (b) If there is a duality gap, the set of geometric multipliers is empty.

DUALITY PROPERTIES

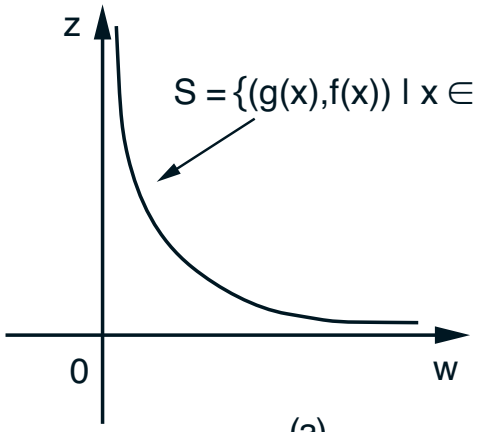
- Optimality Conditions: (x^*, μ^*) is an optimal solution-geometric multiplier pair if and only if

$$
\begin{aligned}
x^* \in X, \quad g(x^*) \leq 0, \quad & \text{(Primal Feasibility)}, \\
\mu^* \geq 0, \quad & \text{(Dual Feasibility)}, \\
x^* = \arg\min_{x \in X} L(x, \mu^*), \quad & \text{(Lagrangian Optimality)}, \\
\mu_j^* g_j(x^*) = 0, \quad j = 1, \ldots, r, \quad & \text{(Compl. Slackness)}.
\end{aligned}
$$

- Saddle Point Theorem: (x^*, μ^*) is an optimal solution-geometric multiplier pair if and only if $x^* \in X$, $\mu^* \geq 0$, and (x^*, μ^*) is a saddle point of the Lagrangian, in the sense that

$$
L(x^*, \mu) \leq L(x^*, \mu^*) \leq L(x, \mu^*), \qquad \forall\ x \in X,\ \mu \geq 0.
$$

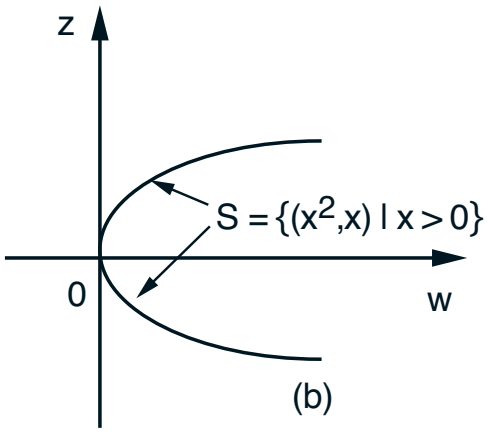
INFEASIBLE AND UNBOUNDED PROBLEMS



(a)

$$\min\ f(x) = 1/x$$
$$\text{s.t. } g(x) = x \le 0$$
$$x \in X = \{x \mid x > 0\}$$

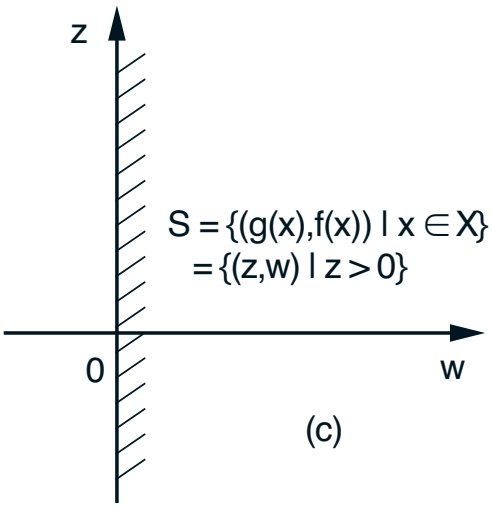
$$f^* = \infty, \quad q^* = \infty$$



(b)

$$\min\ f(x) = x$$
$$\text{s.t. } g(x) = x^2 \le 0$$
$$x \in X = \{x \mid x > 0\}$$

$$f^* = \infty, \quad q^* = 0$$



(c)

$$\min\ f(x) = x_1 + x_2$$
$$\text{s.t. } g(x) = x_1 \le 0$$
$$x \in X = \{(x_1, x_2) \mid x_1 > 0\}$$

$$f^* = \infty, \quad q^* = -\infty$$

EXTENSIONS AND APPLICATIONS

- Equality constraints $h_i(x) = 0,\ i = 1, \ldots, m$, can be converted into the two inequality constraints

$$h_i(x) \leq 0, \qquad -h_i(x) \leq 0.$$

- Separable problems:

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

$$\text{subject to} \quad \sum_{i=1}^{m} g_{ij}(x_i) \leq 0, \qquad j = 1, \ldots, r,$$

$$x_i \in X_i, \qquad i = 1, \ldots, m.$$

- Separable problem with a single constraint:

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

$$\text{subject to} \quad \sum_{i=1}^{n} x_i \geq A, \quad \alpha_i \leq x_i \leq \beta_i, \quad \forall\ i.$$

DUALITY THEOREM I FOR CONVEX PROBLEMS

- Strong Duality Theorem - Linear Constraints: Assume that the problem

minimize $f(x)$

subject to $x \in X, \quad a_i' x - b_i = 0, \quad i = 1, \ldots, m,$

$$e_j' x - d_j \leq 0, \quad j = 1, \ldots, r,$$

has finite optimal value f^*. Let also f be convex over $\Re^n$ and let X be polyhedral. Then there exists at least one geometric multiplier and there is no duality gap.

- Proof Issues
- Application to Linear Programming

COUNTEREXAMPLE

- A Convex Problem with a Duality Gap: Consider the two-dimensional problem

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

$$\text{subject to} \quad x_1 \leq 0, \qquad x \in X = \{x \mid x \geq 0\},$$

where

$$f(x) = e^{-\sqrt{x_1 x_2}}, \qquad \forall\ x \in X,$$

and $f(x)$ is arbitrarily defined for $x \notin X$.

- f is convex over X (its Hessian is positive definite in the interior of X), and $f^* = 1$.
- Also, for all $\mu \geq 0$ we have

$$q(\mu) = \inf_{x \geq 0} \left\{ e^{-\sqrt{x_1 x_2}} + \mu x_1 \right\} = 0,$$

since the expression in braces is nonnegative for $x \geq 0$ and can approach zero by taking $x_1 \to 0$ and $x_1 x_2 \to \infty$. It follows that $q^* = 0$.

DUALITY THEOREM II FOR CONVEX PROBLEMS

- Consider the problem

  minimize $f(x)$

  subject to $x \in X, \qquad g_j(x) \leq 0, \qquad j = 1, \ldots, r.$

- Assume that X is convex and the functions $f : \Re^n \mapsto \Re$, $g_j : \Re^n \mapsto \Re$ are convex over X. Furthermore, the optimal value f^* is finite and there exists a vector $\bar{x} \in X$ such that

$$g_j(\bar{x}) < 0, \qquad \forall\ j = 1, \ldots, r.$$

- Strong Duality Theorem: There exists at least one geometric multiplier and there is no duality gap.
- Extension to linear equality constraints.

6.252 NONLINEAR PROGRAMMING

LECTURE 20: STRONG DUALITY

LECTURE OUTLINE

- Strong Duality Theorem
- Linear Equality Constraints
- Fenchel Duality

****

- Consider the problem

minimize $f(x)$

subject to $x \in X, \qquad g_j(x) \leq 0, \qquad j = 1, \ldots, r,$

assuming $-\infty < f^* < \infty$.

- μ^* is a geometric multiplier if $\mu^* \geq 0$ and $f^* = \inf_{x \in X} L(x, \mu^*)$.

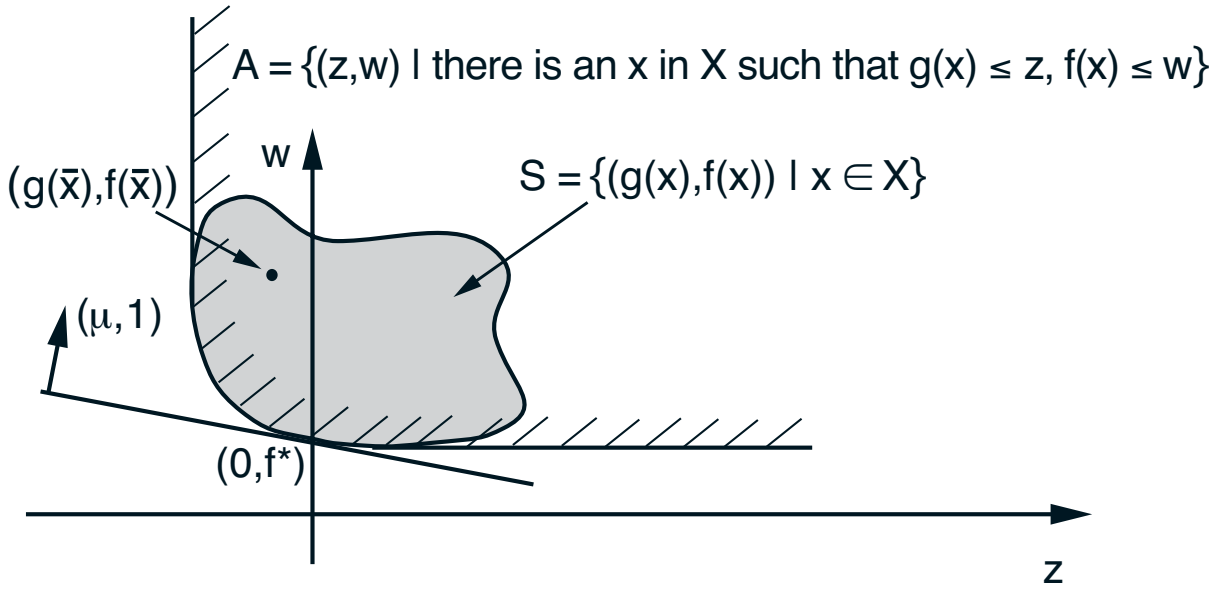
- Dual problem: Maximize $q(\mu) = \inf_{x \in X} L(x, \mu)$ subject to $\mu \geq 0$.

DUALITY THEOREM FOR INEQUALITIES

- Assume that X is convex and the functions $f : \Re^n \mapsto \Re$, $g_j : \Re^n \mapsto \Re$ are convex over X. Furthermore, the optimal value f^* is finite and there exists a vector $\bar{x} \in X$ such that

$$g_j(\bar{x}) < 0, \qquad \forall\ j = 1, \ldots, r.$$

- Strong Duality Theorem: There exists at least one geometric multiplier and there is no duality gap.



PROOF OUTLINE

- Show that A is convex. [Consider vectors $(z, w) \in A$ and $(\tilde{z}, \tilde{w}) \in A$, and show that their convex combinations lie in A.]
- Observe that $(0, f^*)$ is not an interior point of A.
- Hence, there is hyperplane passing through $(0, f^*)$ and containing A in one of the two corresponding halfspaces; i.e., a $(\mu, \beta) \neq (0, 0)$ with

$$\beta f^* \leq \beta w + \mu' z, \qquad \forall\ (z, w) \in A.$$

This implies that $\beta \geq 0$, and $\mu_j \geq 0$ for all j.

- Prove that the hyperplane is nonvertical, i.e., $\beta > 0$.
- Normalize ($\beta = 1$), take the infimum over $x \in X$, and use the fact $\mu \geq 0$, to obtain

$$f^* \leq \inf_{x \in X} \Big\{ f(x) + \mu' g(x) \Big\} = q(\mu) \leq \sup_{\overline{\mu} \geq 0} q(\overline{\mu}) = q^*.$$

Using the weak duality theorem, μ is a geometric multiplier and there is no duality gap.

LINEAR EQUALITY CONSTRAINTS

- Suppose we have the additional constraints

$$e_i' x - d_i = 0, \qquad i = 1, \ldots, m$$

- We need the notion of the $affine\ hull$ of a convex set X [denoted $aff(X)$]. This is the intersection of all hyperplanes containing X.

- The $relative\ interior$ of X, denoted $ri(X)$, is the set of all $x \in X$ s.t. there exists $\epsilon > 0$ with

$$\Big\{ z \mid \|z - x\| < \epsilon,\ z \in aff(X) \Big\} \subset X,$$

that is, $ri(X)$ is the interior of X relative to $aff(X)$.

- Every nonempty convex set has a nonempty relative interior.

DUALITY THEOREM FOR EQUALITIES

- Assumptions:
  - The set X is convex and the functions f, g_j are convex over X.
  - The optimal value f^* is finite and there exists a vector $\bar{x} \in ri(X)$ such that

$$g_j(\bar{x}) < 0, \qquad j = 1, \ldots, r,$$

$$e_i' \bar{x} - d_i = 0, \qquad i = 1, \ldots, m.$$

- Under the preceding assumptions there exists at least one geometric multiplier and there is no duality gap.

COUNTEREXAMPLE

- Consider

  minimize $f(x) = x_1$

  subject to $x_2 = 0, \quad x \in X = \Big\{(x_1, x_2) \mid x_1^2 \leq x_2\Big\}.$

- The optimal solution is $x^* = (0, 0)$ and $f^* = 0$.
- The dual function is given by

$$q(\lambda) = \inf_{x_1^2 \leq x_2} \{x_1 + \lambda x_2\} = \begin{cases} -\frac{1}{4\lambda}, & \text{if } \lambda > 0, \\ -\infty, & \text{if } \lambda \leq 0. \end{cases}$$

- No dual optimal solution and therefore there is no geometric multiplier. (Even though there is no duality gap.)
- Assumptions are violated (the feasible set and the relative interior of X have no common point).

FENCHEL DUALITY FRAMEWORK

- Consider the problem

$$\begin{aligned} &\text{minimize} \quad f_1(x) - f_2(x) \\ &\text{subject to} \quad x \in X_1 \cap X_2, \end{aligned}$$

where f_1 and f_2 are real-valued functions on $\Re^n$, and X_1 and X_2 are subsets of $\Re^n$.

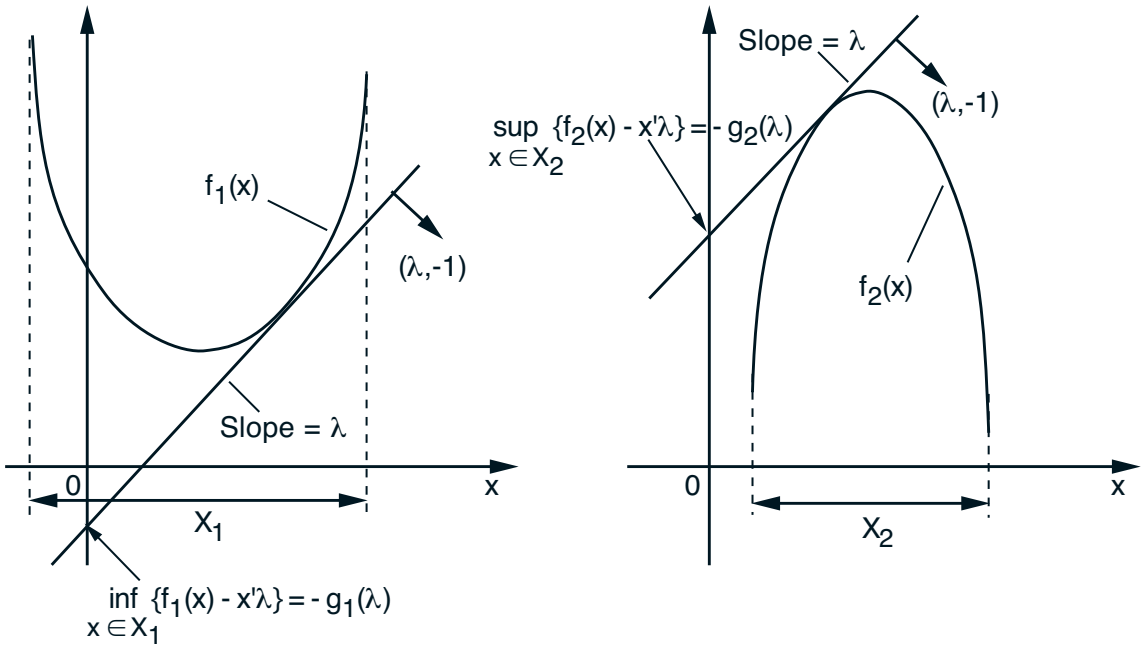
- Assume that $-\infty < f^* < \infty$.
- Convert problem to

$$\begin{aligned} &\text{minimize} \quad f_1(y) - f_2(z) \\ &\text{subject to} \quad z = y, \qquad y \in X_1, \qquad z \in X_2, \end{aligned}$$

and dualize the constraint $z = y$.

$$\begin{aligned} q(\lambda) &= \inf_{y \in X_1,\, z \in X_2} \Big\{ f_1(y) - f_2(z) + (z - y)'\lambda \Big\} \\ &= \inf_{z \in X_2} \Big\{ z'\lambda - f_2(z) \Big\} - \sup_{y \in X_1} \Big\{ y'\lambda - f_1(y) \Big\} \\ &= g_2(\lambda) - g_1(\lambda) \end{aligned}$$

DUALITY THEOREM



- Assume that
  - X_1 and X_2 are convex
  - f_1 and f_2 are convex and concave over X_1 and X_2, respectively
  - The relative interiors of X_1 and X_2 intersect
- The duality theorem for equalities applies and shows that

$$f^* = \max_{\lambda \in \Re^n} \Big\{ g_2(\lambda) - g_1(\lambda) \Big\}$$

and that the maximum above is attained.

OPTIMALITY CONDITIONS

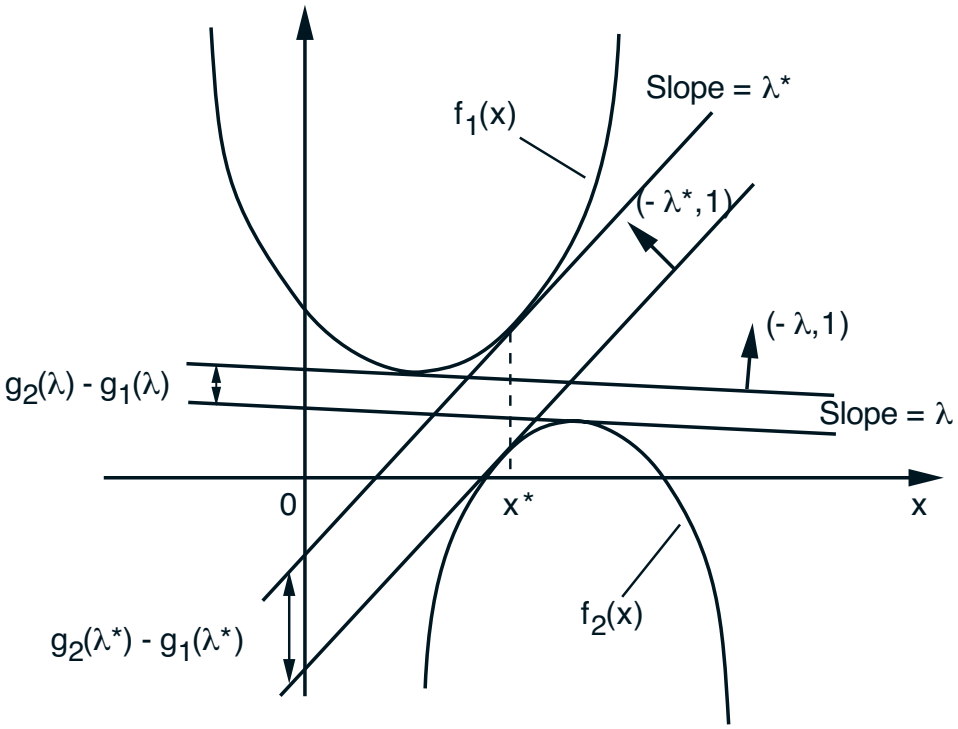
- There is no duality gap, while simultaneously, (x^*, λ^*) is an optimal primal and dual solution pair, if and only if

$$x^* \in X_1 \cap X_2, \qquad (\text{primal feasibility}),$$

$$\lambda^* \in \Lambda_1 \cap \Lambda_2, \qquad (\text{dual feasibility}),$$

$$x^* = \arg \max_{x \in X_1} \left\{ x' \lambda^* - f_1(x) \right\}$$

$$= \arg \min_{x \in X_2} \left\{ x' \lambda^* - f_2(x) \right\}, \quad (\text{Lagr. optimality}).$$



6.252 NONLINEAR PROGRAMMING

LECTURE 21: DISCRETE OPTIMIZATION

LECTURE OUTLINE

- Discrete Constraints and Integer Programming
- Examples of Discrete Optimization Problems
- Constraint Relaxation and Rounding
- Branch-and-Bound
- Lagrangian Relaxation

****

- Consider

  minimize $f(x)$

  subject to $x \in X, \qquad g_j(x) \leq 0, \qquad j = 1, \ldots, r,$

where X is a *finite* set.

- Example: 0-1 Integer programming:

$$X = \Big\{(x_1, \ldots, x_n) \mid x_i = 0 \text{ or } 1,\ i = 1, \ldots, n\Big\}.$$

EXAMPLES OF DISCRETE PROBLEMS

- Given a directed graph with set of nodes $\mathcal{N}$ and set of arcs $(i,j) \in \mathcal{A}$, the (integer constrained) minimum cost network flow problem is

$$\text{minimize} \quad \sum_{(i,j)\in\mathcal{A}} a_{ij}x_{ij}$$

subject to the constraints

$$\sum_{\{j|(i,j)\in\mathcal{A}\}} x_{ij} - \sum_{\{j|(j,i)\in\mathcal{A}\}} x_{ji} = s_i, \qquad \forall\ i \in \mathcal{N},$$

$$b_{ij} \le x_{ij} \le c_{ij}, \quad \forall\ (i,j) \in \mathcal{A}, \qquad x_{ij}:\ \text{integer},$$

where a_{ij}, b_{ij}, c_{ij}, and s_i are given scalars.

- Think of:
  - Nodes i with $s_i > 0$ and $s_i < 0$ as production and consumption points, respectively.
  - s_i supply or demand of node i.
  - Arcs (i,j) as transportation links with flow capacity c_{ij} and cost per unit flow a_{ij}
  - Problem is to accomplish a minimum cost transfer from the supply to the demand points.
- Important special cases: Shortest path, max-flow, transportation, assignment problems.

UNIMODULARITY PROPERTY

- The minimum cost flow problem has an interesting property: If the s_i and c_{ij} are integer, the optimal solutions of the integer-constrained problem also solve the $relaxed$ problem, obtained when the integer constraints are neglected.
- Great practical significance, since the relaxed problem can be solved using efficient linear (not integer) programming algorithms.
- This is special case of $unimodularity$:
  - A square matrix A with integer components is $unimodular$ if its determinant is 0, 1, or -1.
  - If A is invertible and unimodular, by Kramer's rule, the inverse matrix A^{-1} has integer components. Hence, the solution x of the system $Ax = b$ is integer for every integer vector b.
  - A rectangular matrix with integer components is called $totally\ unimodular$ if each of its square submatrices is unimodular.
- A polyhedron $\{x \mid Ex = d,\ b \leq x \leq c\}$ has integer extreme points if E is totally unimodular and b, c, and d have integer components.
- The matrix E corresponding to the minimum cost flow problem is totally unimodular.

EXAMPLES OF NONUNIMODULAR PROBLEMS

- Unimodularity is an exceptional property.
- Nonunimodular example (Traveling salesman problem): A salesman wants to find a minimum cost tour that visits each of N given cities exactly once and returns to the starting city.
- Let a_{ij}: cost of going from city i to city j, and let x_{ij} be a variable that takes the value 1 if the salesman visits city j immediately following city i, and the value 0 otherwise. The problem is

$$\text{minimize} \quad \sum_{i=1}^{N} \sum_{\substack{j=1,\ldots,N \\ j \neq i}} a_{ij} x_{ij}$$

$$\text{subject to} \quad \sum_{\substack{j=1,\ldots,N \\ j \neq i}} x_{ij} = 1, \qquad i = 1, \ldots, N,$$

$$\sum_{\substack{i=1,\ldots,N \\ i \neq j}} x_{ij} = 1, \qquad j = 1, \ldots, N,$$

plus the constraints $x_{ij} = 0$ or 1, and that the set of arcs $\{(i,j) \mid x_{ij} = 1\}$ forms a connected tour, i.e.,

$$\sum_{i \in S,\, j \notin S} (x_{ij} + x_{ji}) \geq 2, \quad \forall \text{ proper subsets } S \text{ of cities.}$$

APPROACHES TO INTEGER PROGRAMMING

- Enumeration of the finite set of all feasible (integer) solutions, and comparison to obtain an optimal solution (this is rarely practical).
- Constraint relaxation and heuristic rounding.
  - Neglect the integer constraints
  - Solve the problem using linear/nonlinear programming methods
  - If a noninteger solution is obtained, round it to integer using a heuristic
  - Sometimes, with favorable structure, clever problem formulation, and good heuristics, this works remarkably well
- Implicit enumeration (or branch-and-bound):
  - Combines the preceding two approaches
  - It uses constraint relaxation and solution of noninteger problems to obtain certain lower bounds that are used to discard large portions of the feasible set.
  - In principle it can find an optimal (integer) solution, but this may require unacceptable long time.
  - In practice, usually it is terminated with a heuristically obtained integer solution, often derived by rounding a noninteger solution.

PRINCIPLE OF BRANCH-AND-BOUND

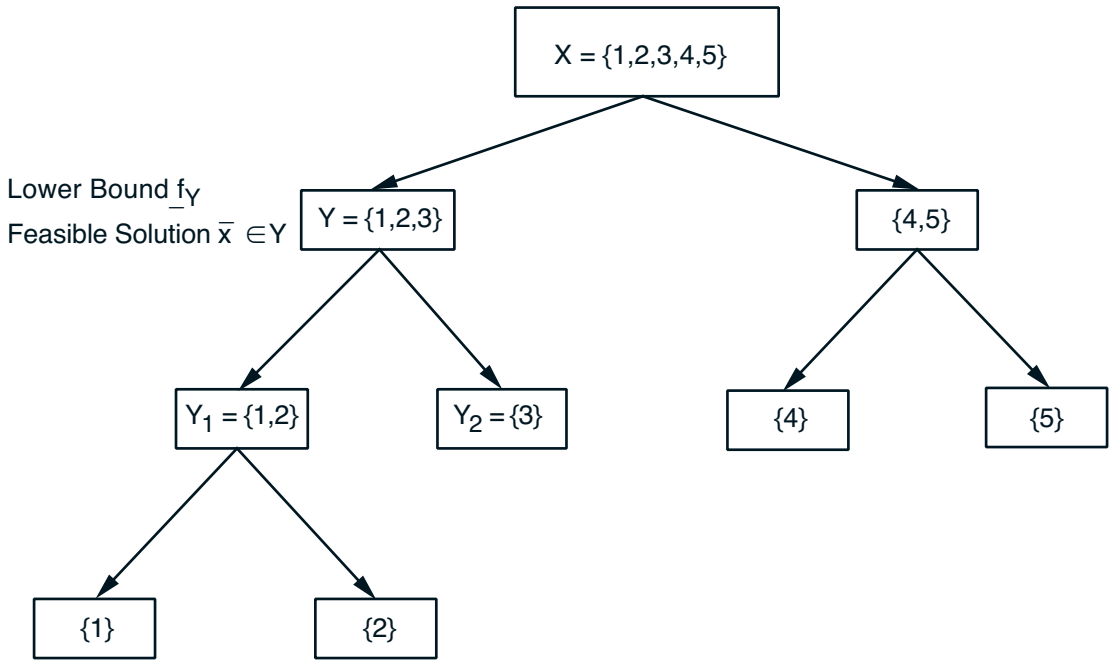
- **Bounding Principle:** Consider minimizing $f(x)$ over a finite set $x \in X$. Let Y_1 and Y_2 be two subsets of X, and suppose that we have bounds

$$\underline{f}_1 \leq \min_{x \in Y_1} f(x), \qquad \overline{f}_2 \geq \min_{x \in Y_2} f(x).$$

Then, if $\overline{f}_2 \leq \underline{f}_1$, the solutions in Y_1 may be disregarded since their cost cannot be smaller than the cost of the best solution in Y_2.

- The branch-and-bound method uses suitable upper and lower bounds, and the bounding principle to eliminate substantial portions of X.

- It uses a tree, with nodes that correspond to subsets of X, usually obtained by binary partition.



BRANCH-AND-BOUND ALGORITHM

- The algorithm maintains a node list called OPEN, and a scalar called UPPER, which is equal to the minimal cost over feasible solutions found so far. Initially, OPEN$= \{X\}$, and UPPER$= \infty$ or to the cost $f(\overline{x})$ of some feasible solution $\overline{x} \in X$.

- **Step 1:** Remove a node Y from OPEN. For each child Y_j of Y, do the following: Find the lower bound $\underline{f}_{Y_j}$ and a feasible solution $\overline{x} \in Y_j$. If

$$\underline{f}_{Y_j} < \text{ UPPER},$$

place Y_j in OPEN. If in addition

$$f(\overline{x}) < \text{ UPPER},$$

set UPPER $= f(\overline{x})$ and mark $\overline{x}$ as the best solution found so far.

Step 2: (Termination Test) If OPEN is nonempty, go to step 1. Otherwise, terminate; the best solution found so far is optimal.

- Termination with a global minimum is guaranteed, but the number of nodes to be examined may be huge. In practice, the algorithm is terminated when an ϵ-optimal solution is obtained.

- Tight lower bounds $\underline{f}_{Y_j}$ are important for quick termination.

LAGRANGIAN RELAXATION

- One method to obtain lower bounds in the branch-and-bound method is by constraint relaxation (e.g., replace $x_i \in \{0, 1\}$ by $0 \leq x_i \leq 1$)

- Another method, called *Lagrangian relaxation*, is based on weak duality. If the subproblem of a node of the branch-and-bound tree has the form

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

use as lower bound the optimal dual value

$$q^* = \max_{\mu \geq 0} q(\mu),$$

where

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

- Essential for applying Lagrangian relaxation is that the dual problem is easy to solve (e.g., the dual is a simple linear program, or it involves useful structure, such as separability).

6.252 NONLINEAR PROGRAMMING

LECTURE 22: DUAL COMPUTATIONAL METHODS

LECTURE OUTLINE

- Dual Methods
- Nondifferentiable Optimization

****

- Consider the primal problem

minimize $f(x)$

subject to $x \in X, \qquad g_j(x) \leq 0, \qquad j = 1, \ldots, r,$

assuming $-\infty < f^* < \infty$.

- Dual problem: Maximize

$$q(\mu) = \inf_{x \in X} L(x, \mu) = \inf_{x \in X} \{f(x) + \mu' g(x)\}$$

subject to $\mu \geq 0$.

PROS AND CONS FOR SOLVING THE DUAL

ADVANTAGES:

- The dual is concave.

- The dual may have smaller dimension and/or simpler constraints.

- If there is no duality gap and the dual is solved exactly for a geometric multiplier μ^*, all optimal primal solutions can be obtained by minimizing the Lagrangian $L(x, \mu^*)$ over $x \in X$.

- Even if there is a duality gap, $q(\mu)$ is a lower bound to the optimal primal value for every $\mu \geq 0$.

DISADVANTAGES:

- Evaluating $q(\mu)$ requires minimization of $L(x, \mu)$ over $x \in X$.

- The dual function is often nondifferentiable.

- Even if we find an optimal dual solution μ^*, it may be difficult to obtain a primal optimal solution.

STRUCTURE

- **Separability:** Classical duality structure (Lagrangian relaxation).
- **Partitioning:** The problem

$$\text{minimize} \quad F(x) + G(y)$$

$$\text{subject to} \quad Ax + By = c, \quad x \in X, \quad y \in Y$$

can be written as

$$\text{minimize} \quad F(x) + \inf_{By = c - Ax,\, y \in Y} G(y)$$

$$\text{subject to} \quad x \in X.$$

With no duality gap, this problem is written as

$$\text{minimize} \quad F(x) + Q(Ax)$$

$$\text{subject to} \quad x \in X,$$

where

$$Q(Ax) = \max_{\lambda} q(\lambda, Ax)$$

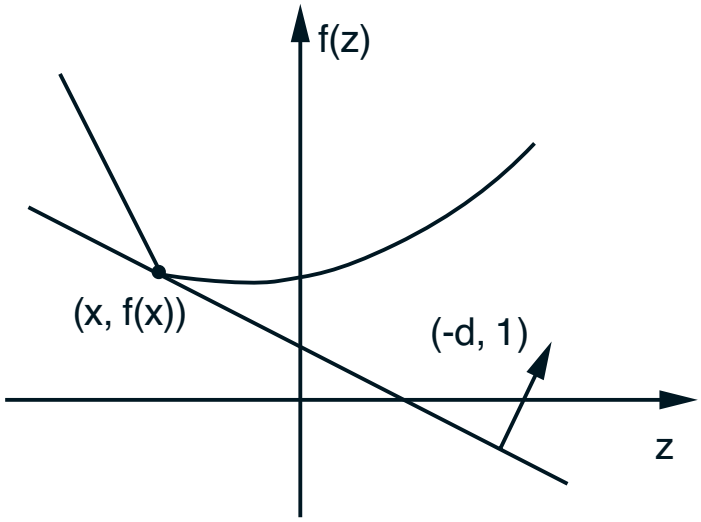
$$q(\lambda, Ax) = \inf_{y \in Y} \Big\{ G(y) + \lambda'(Ax + By - c) \Big\}$$

SUBGRADIENTS

- A vector d is said to be a $subgradient$ at x of a convex function $f : \Re^n \mapsto \Re$ if

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

The set of subgradients at x is called the $subdifferential$ and is denoted by $\partial f(x)$.



- If f is concave, d is said to be a $subgradient$ at x if

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

- **Danskin's Theorem:** Consider the function $f(x) = \max_{z \in Z} \phi(x, z)$, where $\phi : \Re^{n+m} \mapsto \Re$ is continuous, Z is compact, and $\phi(\cdot, z)$ is convex for each $z \in Z$. Then f is convex and

$$\partial f(x) = \text{Convex Hull}\Big\{\partial_x \phi(x, z) \mid z : \text{attains the max}\Big\}$$

DUAL DERIVATIVES

- Let

$$x_\mu = \arg\min_{x \in X} L(x, \mu) = \arg\min_{x \in X} \Big\{ f(x) + \mu' g(x) \Big\}.$$

Then for all $\tilde{\mu} \in \Re^r$,

$$\begin{aligned} q(\tilde{\mu}) &= \inf_{x \in X} \Big\{ f(x) + \tilde{\mu}' g(x) \Big\} \\ &\leq f(x_\mu) + \tilde{\mu}' g(x_\mu) \\ &= f(x_\mu) + \mu' g(x_\mu) + (\tilde{\mu} - \mu)' g(x_\mu) \\ &= q(\mu) + (\tilde{\mu} - \mu)' g(x_\mu). \end{aligned}$$

- Thus $g(x_\mu)$ *is a subgradient of* q *at* μ.

- **Proposition:** Let X be compact, and let f and g be continuous over X. Assume also that for every μ, $L(x, \mu)$ is minimized over $x \in X$ at a unique point x_μ. Then, q is everywhere continuously differentiable and

$$\nabla q(\mu) = g(x_\mu), \qquad \forall\ \mu \in \Re^r.$$

NONDIFFERENTIABLE DUAL

- If there exists a duality gap, the dual function is nondifferentiable at every dual optimal solution.
- Important nondifferentiable case: When q is polyhedral, that is,

$$q(\mu) = \min_{i \in I} \Big\{ a_i' \mu + b_i \Big\},$$

where I is a finite index set, and $a_i \in \Re^r$ and b_i are given (arises when X is a discrete set, as in integer programming).

- **Proposition:** Let q be polyhedral as above, and let I_μ be the set of indices attaining the minimum

$$I_\mu = \Big\{ i \in I \mid a_i' \mu + b_i = q(\mu) \Big\}.$$

The set of all subgradients of q at μ is

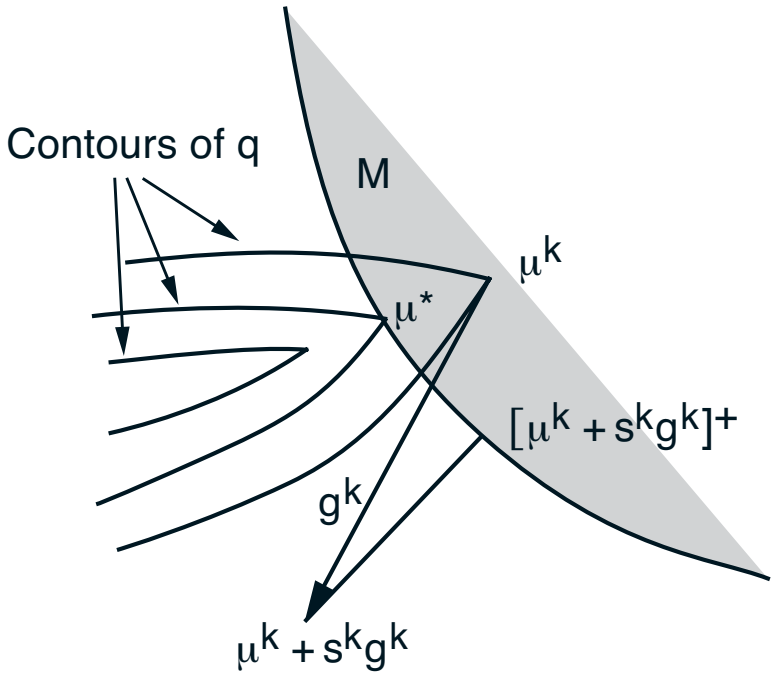
$$\partial q(\mu) = \left\{ g \;\middle|\; g = \sum_{i \in I_\mu} \xi_i a_i,\ \xi_i \geq 0,\ \sum_{i \in I_\mu} \xi_i = 1 \right\}.$$

NONDIFFERENTIABLE OPTIMIZATION

- Consider maximization of $q(\mu)$ over $M = \{| \ \mu \geq 0, \ q(\mu) > -\infty\}$
- Subgradient method:

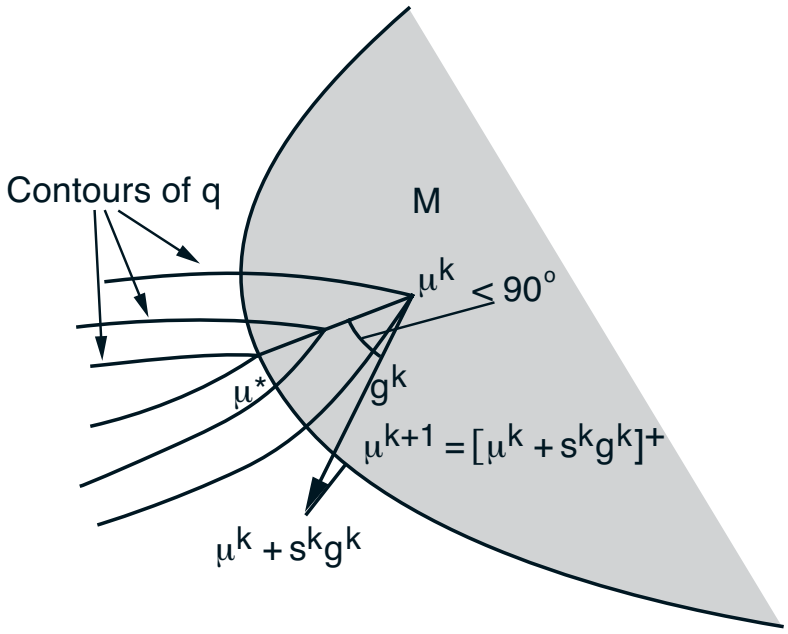
$$\mu^{k+1} = \left[\mu^k + s^k g^k\right]^+,$$

where g^k is the subgradient $g(x_{\mu^k})$, $[\cdot]^+$ denotes projection on the closed convex set M, and s^k is a positive scalar stepsize.



KEY SUBGRADIENT METHOD PROPERTY

- For a small stepsize it reduces the Euclidean distance to the optimum.



- **Proposition:** For any dual optimal solution μ^* and any nonoptimal μ^k, we have

$$\|\mu^{k+1} - \mu^*\| < \|\mu^k - \mu^*\|,$$

for all stepsizes s^k such that

$$0 < s^k < \frac{2\Big(q(\mu^*) - q(\mu^k)\Big)}{\|g^k\|^2}.$$

STEPSIZE RULES

- Diminishing stepsize is one possibility.
- More common method:

$$s^k = \frac{\alpha^k \left(q^k - q(\mu^k) \right)}{\|g^k\|^2},$$

where $q^k \approx q^*$ and

$$0 < \alpha^k < 2.$$

- Some possibilities:
  - q^k is the best known upper bound to q^*; $\alpha^0 = 1$ and α^k decreased by a certain factor every few iterations.
  - $\alpha^k = 1$ for all k and

$$q^k = \max_{0 \le i \le k} q(\mu^i) + \delta^k,$$

where δ^k represents an "aspiration" level that is adjusted depending on algorithmic progress of the algorithm.

6.252 NONLINEAR PROGRAMMING

LECTURE 23: ADDITIONAL DUAL METHODS

LECTURE OUTLINE

- Cutting Plane Methods
- Decomposition

****

- Consider the primal problem

minimize $f(x)$

subject to $x \in X, \qquad g_j(x) \leq 0, \qquad j = 1, \ldots, r,$

assuming $-\infty < f^* < \infty$.

- Dual problem: Maximize

$$q(\mu) = \inf_{x \in X} L(x, \mu) = \inf_{x \in X} \{f(x) + \mu' g(x)\}$$

subject to $\mu \in M = \{\mu \mid \mu \geq 0,\ q(\mu) > -\infty\}$.

CUTTING PLANE METHOD

- kth iteration, after μ^i and $g^i = g\left(x_{\mu^i}\right)$ have been generated for $i = 0, \ldots, k-1$: Solve

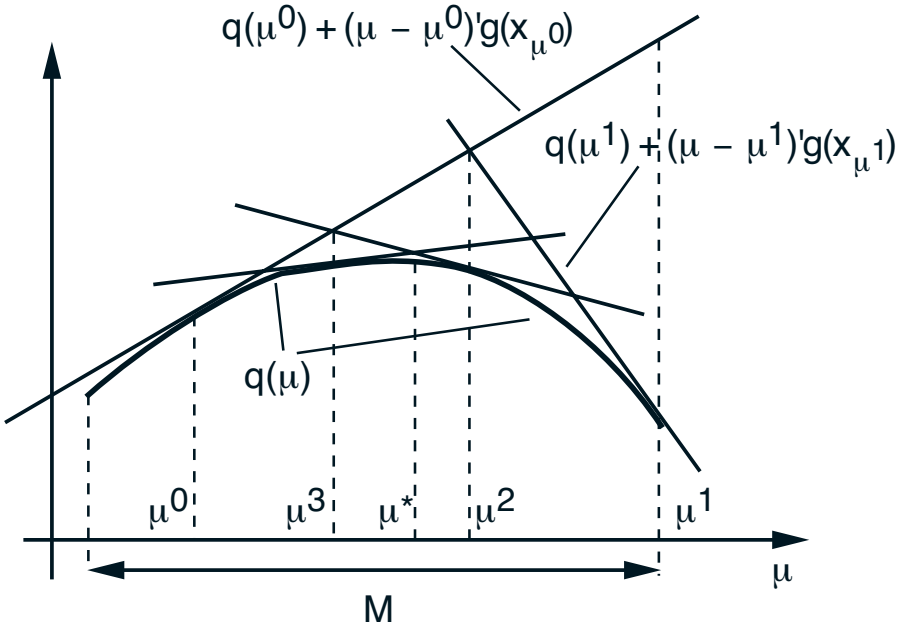
$$\max_{\mu \in M} Q^k(\mu)$$

where

$$Q^k(\mu) = \min_{i=0,\ldots,k-1} \left\{ q(\mu^i) + (\mu - \mu^i)' g^i \right\}.$$

Set

$$\mu^k = \arg\max_{\mu \in M} Q^k(\mu).$$

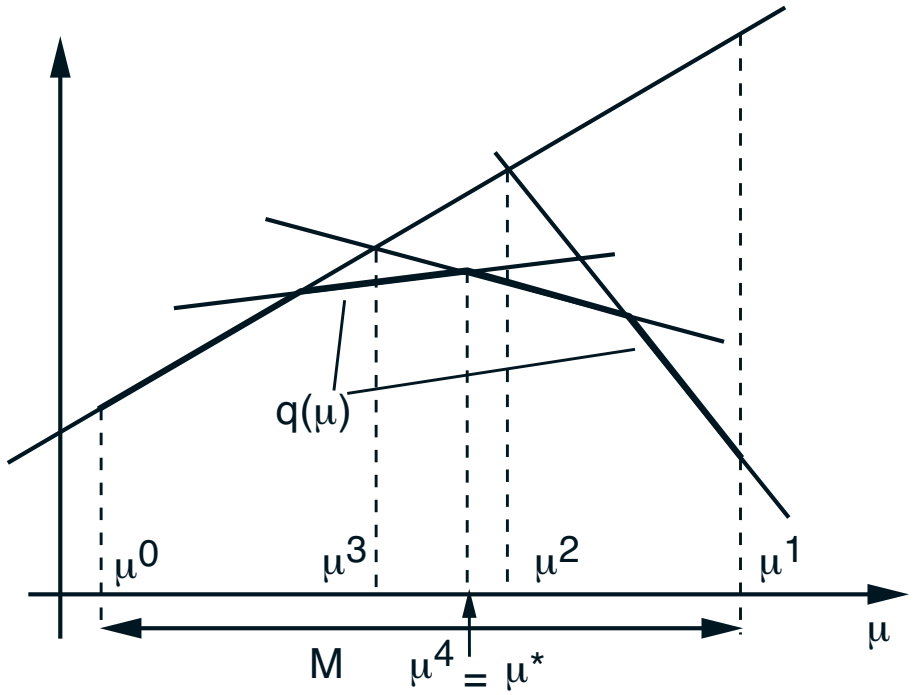


POLYHEDRAL CASE

$$q(\mu) = \min_{i \in I} \left\{ a_i' \mu + b_i \right\}$$

where I is a finite index set, and $a_i \in \Re^r$ and b_i are given.

- Then subgradient g^k in the cutting plane method is a vector a_{ik} for which the minimum is attained.
- Finite termination expected.



CONVERGENCE

- **Proposition:** Assume that the max of Q_k over M is attained and that q is real-valued. Then every limit point of a sequence $\{\mu^k\}$ generated by the cutting plane method is a dual optimal solution.

Proof: g^i is a subgradient of q at μ^i, so

$$q(\mu^i) + (\mu - \mu^i)'g^i \geq q(\mu), \qquad \forall\ \mu \in M,$$

$$Q^k(\mu^k) \geq Q^k(\mu) \geq q(\mu), \qquad \forall\ \mu \in M. \qquad (1)$$

- Suppose $\{\mu^k\}_K$ converges to $\bar{\mu}$. Then, $\bar{\mu} \in M$, and by Eq. (1) and continuity of Q^k and q (real-valued assumption), $Q^k(\bar{\mu}) \geq q(\bar{\mu})$. Using this and Eq. (1), we obtain for all k and $i < k$,

$$q(\mu^i) + (\mu^k - \mu^i)'g^i \geq Q^k(\mu^k) \geq Q^k(\bar{\mu}) \geq q(\bar{\mu}).$$

- Take the limit as $i \to \infty$, $k \to \infty$, $i \in K$, $k \in K$,

$$\lim_{k \to \infty,\ k \in K} Q^k(\mu^k) = q(\bar{\mu}).$$

Combining with (1), $q(\bar{\mu}) = \max_{\mu \in M} q(\mu)$.

LAGRANGIAN RELAXATION

- Solving the dual of the separable problem

  minimize $\sum_{j=1}^{J} f_j(x_j)$

  subject to $x_j \in X_j, \quad j = 1, \ldots, J, \quad \sum_{j=1}^{J} A_j x_j = b.$

- Dual function is

$$q(\lambda) = \sum_{j=1}^{J} \min_{x_j \in X_j} \left\{ f_j(x_j) + \lambda' A_j x_j \right\} - \lambda' b$$

$$= \sum_{j=1}^{J} \left\{ f_j\left(x_j(\lambda)\right) + \lambda' A_j x_j(\lambda) \right\} - \lambda' b$$

where $x_j(\lambda)$ attains the min. A subgradient at λ is

$$g_\lambda = \sum_{j=1}^{J} A_j x_j(\lambda) - b.$$

DANTSIG-WOLFE DECOMPOSITION

- D-W decomposition method is just the cutting plane applied to the dual problem $\max_\lambda q(\lambda)$.
- At the kth iteration, we solve the "approximate dual"

$$\lambda^k = \arg\max_{\lambda \in \Re^r} Q^k(\lambda) \equiv \min_{i=0,\ldots,k-1} \Big\{ q(\lambda^i) + (\lambda - \lambda^i)' g^i \Big\}.$$

- Equivalent linear program in v and λ

maximize v

subject to $v \le q(\lambda^i) + (\lambda - \lambda^i)' g^i, \quad i = 0, \ldots, k-1$

The dual of this (called $master\ problem$) is

$$\text{minimize} \quad \sum_{i=0}^{k-1} \xi^i \Big(q(\lambda^i) - \lambda^{i\prime} g^i \Big)$$

$$\text{subject to} \quad \sum_{i=0}^{k-1} \xi^i = 1, \qquad \sum_{i=0}^{k-1} \xi^i g^i = 0,$$

$$\xi^i \ge 0, \quad i = 0, \ldots, k-1,$$

DANTSIG-WOLFE DECOMPOSITION (CONT.)

- The master problem is written as

$$\text{minimize} \quad \sum_{j=1}^{J} \left(\sum_{i=0}^{k-1} \xi^i f_j\left(x_j(\lambda^i)\right) \right)$$

$$\text{subject to} \quad \sum_{i=0}^{k-1} \xi^i = 1, \qquad \sum_{j=1}^{J} A_j \left(\sum_{i=0}^{k-1} \xi^i x_j(\lambda^i) \right) = b,$$

$$\xi^i \geq 0, \quad i = 0, \ldots, k-1.$$

- The primal cost function terms $f_j(x_j)$ are approximated by

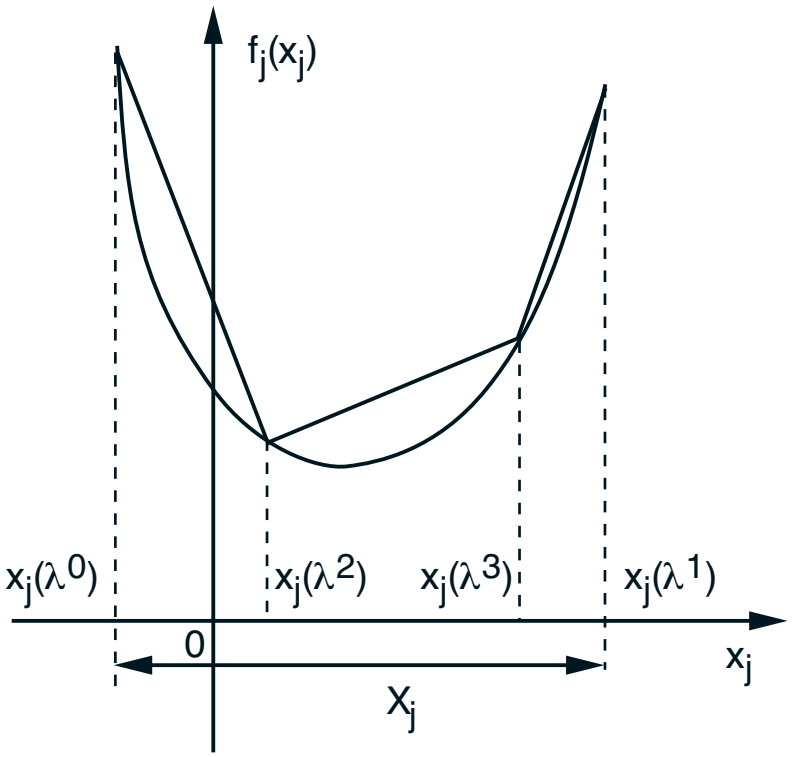
$$\sum_{i=0}^{k-1} \xi^i f_j\left(x_j(\lambda^i)\right)$$

- Vectors x_j are expressed as

$$\sum_{i=0}^{k-1} \xi^i x_j(\lambda^i)$$

GEOMETRICAL INTERPRETATION

- Geometric interpretation of the master problem (the dual of the approximate dual solved in the cutting plane method) is *inner linearization*.



- This is a "dual" operation to the one involved in the cutting plane approximation, which can be viewed as *outer linearization*.

6.252 NONLINEAR PROGRAMMING

LECTURE 24: EPILOGUE

LECTURE OUTLINE

- Review of three dominant themes from this course
  - Descent along feasible directions
  - Approximation of a "difficult" problem by a sequence of "easier" problems
  - Duality
- Discussion of an algorithmic approach that we did not cover (Sections 4.3, 4.4): Solving the necessary optimality conditions, viewed as a system of equations and inequalities
- More on duality: Relation of primal and dual functions
- Connection of constrained optimization duality and saddle point/game theory

THE DESCENT APPROACH

- Use in necessary optimality conditions: at a local minimum x^* of f over X, we have $\nabla f(x^*)'d \geq 0$ for all feasible directions d of X at x^*. Special cases:
  - $\nabla f(x^*) = 0$ when $X = \Re^n$
  - $\nabla f(x^*)'(x - x^*) \geq 0$ for all $x \in X$, when X is convex
- Use in sufficient optimality conditions under Hessian positive definiteness, or under convexity assumptions
- Use in algorithms:
  - Gradient-related methods for unconstrained optimization
  - Feasible direction algorithms
  - Subgradient methods (based on descent of the distance of the current iterate to the optimum)

THE APPROXIMATION APPROACH

- Use in Lagrange multiplier theory:
  - Introduce a penalized problem that "converges" to the original constrained problem as the penalty parameter goes to ∞
  - Take the limit in the optimality conditions for the penalized problem to obtain optimality conditions for the constrained problem
- Also use in sufficient optimality conditions using an augmented Lagrangian approach
- Use in algorithms:
  - Barrier/interior point methods
  - Penalty and augmented Lagrangian methods
  - Cutting plane methods

SOLVING THE NECESSARY CONDITIONS

- Another algorithmic approach for equality and inequality constrained problems (we did not cover it; see Sections 4.3, 4.4). It is based on:
  - Viewing the optimality (KKT) conditions as a system of (nonlinear) equations and inequalities to be solved for x and the multipliers
  - Solving them using some method for solving systems of equations and inequalities
- Principal solution methods are a number of variants of Newton's method
- Important issue: how to enlarge the region of convergence of Newton's method without destroying its fast convergence near a solution
- Principal tools: stepsize procedures and merit functions
- Important methods:
  - Sequential quadratic programming (Section 4.3)
  - Primal-dual interior point methods (Section 4.4)

DUALITY - MIN COMMON/MAX CROSSING

- The principal issues in constrained optimization duality are intuitively captured in the min common/max crossing framework, including:
  - The weak duality theorem
  - The need for convexity assumptions in order that there is no duality gap
  - The Slater condition, which guarantees the existence of a geometric multiplier
  - The pathologies that can result in a duality gap, even under convexity conditions
- For the problem $\min_{x \in X,\, g(x) \leq 0} f(x)$, an important concept is the primal function, defined by

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

- If X, f, and g_j are convex, then it can be shown that p is convex
- Assuming convexity of p
  - The set of geometric multipliers is equal to the set of subgradients of p at 0
  - Absence of a duality gap is equivalent to right continuity of p at 0, i.e., $p(0) = \lim_{u \downarrow 0} p(u)$

DUALITY OF PRIMAL AND DUAL FUNCTION

- The primal function p and the dual function q are intimately connected: For every $\mu \geq 0$, we have

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

$$= \inf_{\{(u,x) | x \in X,\, g_j(x) \leq u_j,\, j=1,\ldots,r\}} \left\{ f(x) + \sum_{j=1}^{r} \mu_j g_j(x) \right\}$$

$$= \inf_{\{(u,x) | x \in X,\, g_j(x) \leq u_j,\, j=1,\ldots,r\}} \left\{ f(x) + \sum_{j=1}^{r} \mu_j u_j \right\}$$

$$= \inf_{u \in \Re^r} \inf_{\substack{x \in X,\, g_j(x) \leq u_j \\ j=1,\ldots,r}} \left\{ f(x) + \sum_{j=1}^{r} \mu_j u_j \right\},$$

and finally $q(\mu) = \inf_{u \in \Re^r} \left\{ p(u) + \mu' u \right\}$ for all $\mu \geq 0$

- Thus,

$$q(\mu) = -h(-\mu), \qquad \forall\ \mu \geq 0,$$

where h is the conjugate convex function of p:

$$h(\nu) = \sup_{u \in \Re^r} \left\{ \nu' u - p(u) \right\}.$$

DUALITY AND MINIMAX THEORY

- Duality issues for the problem $\min_{x \in X,\, g(x) \leq 0} f(x)$ are closely connected to saddle point issues for the Lagrangian function

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

- We have

$$f^* = \inf_{x \in X,\, g(x) \leq 0} f(x) = \inf_{x \in X} \sup_{\mu \geq 0} L(x, \mu),$$

$$q^* = \sup_{\mu \geq 0} q(\mu) = \sup_{\mu \geq 0} \inf_{x \in X} L(x, \mu),$$

so no duality gap is equivalent to

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

- Also, we showed that (x, μ) is a global minimum-geometric multiplier pair if and only if it is a saddle point of $L(x, \mu)$ over $x \in X$ and $\mu \geq 0$

- Constrained optimization duality theory can be viewed as the special case of minimax theory where μ appears linearly and is constrained by $\mu \geq 0$; but general minimax theory does not shed much light on this special case

COMMON ROOT OF DUALITY AND MINIMAX

- Constrained optimization duality theory and minimax theory are not "equivalent" but they have a common geometrical root: the min common/max crossing structure
- Consider the issue whether $\inf_{x\in X}\sup_{\mu\in M}\phi(x,\mu) = \sup_{\mu\in M}\inf_{x\in X}\phi(x,\mu)$ and let

$$p(u) = \inf_{x\in X}\sup_{\mu\in M}\Big\{\phi(x,\mu) - u'\mu\Big\}$$

[If $\phi(x,\mu) = L(x,\mu)$, p is equal to the primal function.]

- Consider also the min common/max crossing framework for the set $\big\{(u,w) \mid p(u) \le w\big\}$. Then the min common value is $p(0) = \inf_{x\in X}\sup_{\mu\in M}\phi(x,\mu)$
- Under convexity/semicontinuity assumptions on X, M, $\phi(\cdot,\mu)$, and $-\phi(x,\cdot)$, it can be shown that the max crossing value is equal to $\sup_{\mu\in M}\inf_{x\in X}\phi(x,\mu)$
- Thus equality of the min common and max crossing values is equivalent to

$$\inf_{x\in X}\sup_{\mu\in M}\phi(x,\mu) = \sup_{\mu\in M}\inf_{x\in X}\phi(x,\mu)$$

- For an extensive analysis of all this, see the author's book "Convex Analysis and Optimization"

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